Algebraic Statistics

Seth Sullivant

North Carolina State University
E-mail address: smsulli2@ncsu.edu
Abstract. Algebraic statistics uses tools from algebraic geometry, commutative algebra, combinatorics, and their computational sides to address problems in statistics and its applications. The starting point for this connection is the observation that many statistical models are semialgebraic sets. The algebra/statistics connection is now over twenty years old—this book presents the first comprehensive and introductory treatment of the subject. After background material in probability, algebra, and statistics, the book covers a range of topics in algebraic statistics including algebraic exponential families, likelihood inference, Fisher’s exact test, bounds on entries of contingency tables, design of experiments, identifiability of hidden variable models, phylogenetic models, and model selection. The book is suitable for both classroom use and independent study, as it has numerous examples, references, and over 150 exercises.
# Contents

<table>
<thead>
<tr>
<th>Preface</th>
<th>xi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>§1.1. Discrete Markov Chain</td>
<td>2</td>
</tr>
<tr>
<td>§1.2. Exercises</td>
<td>9</td>
</tr>
<tr>
<td>Chapter 2. Probability Primer</td>
<td>11</td>
</tr>
<tr>
<td>§2.1. Probability</td>
<td>11</td>
</tr>
<tr>
<td>§2.2. Random Variables and their Distributions</td>
<td>17</td>
</tr>
<tr>
<td>§2.3. Expectation, Variance, and Covariance</td>
<td>22</td>
</tr>
<tr>
<td>§2.4. Multivariate Normal Distribution</td>
<td>29</td>
</tr>
<tr>
<td>§2.5. Limit Theorems</td>
<td>32</td>
</tr>
<tr>
<td>§2.6. Exercises</td>
<td>37</td>
</tr>
<tr>
<td>Chapter 3. Algebra Primer</td>
<td>39</td>
</tr>
<tr>
<td>§3.1. Varieties</td>
<td>39</td>
</tr>
<tr>
<td>§3.2. Ideals</td>
<td>43</td>
</tr>
<tr>
<td>§3.3. Gröbner Bases</td>
<td>47</td>
</tr>
<tr>
<td>§3.4. First Applications of Gröbner Bases</td>
<td>53</td>
</tr>
<tr>
<td>§3.5. Computational Algebra Vignettes</td>
<td>57</td>
</tr>
<tr>
<td>§3.6. Projective Space and Projective Varieties</td>
<td>63</td>
</tr>
<tr>
<td>§3.7. Exercises</td>
<td>66</td>
</tr>
<tr>
<td>Chapter 4. Conditional Independence</td>
<td>69</td>
</tr>
<tr>
<td>§4.1. Conditional Independence Models</td>
<td>70</td>
</tr>
</tbody>
</table>
## Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>§15.1</td>
<td>Trees and Splits</td>
<td>330</td>
</tr>
<tr>
<td>§15.2</td>
<td>Types of Phylogenetic Models</td>
<td>333</td>
</tr>
<tr>
<td>§15.3</td>
<td>Group-based Phylogenetic Models</td>
<td>341</td>
</tr>
<tr>
<td>§15.4</td>
<td>The General Markov Model</td>
<td>352</td>
</tr>
<tr>
<td>§15.5</td>
<td>The Allman-Rhodes-Draisma-Kuttler Theorem</td>
<td>358</td>
</tr>
<tr>
<td>§15.6</td>
<td>Exercises</td>
<td>362</td>
</tr>
<tr>
<td>Chapter 16</td>
<td>Identifiability</td>
<td>363</td>
</tr>
<tr>
<td>§16.1</td>
<td>Tools for Testing Identifiability</td>
<td>364</td>
</tr>
<tr>
<td>§16.2</td>
<td>Linear Structural Equation Models</td>
<td>371</td>
</tr>
<tr>
<td>§16.3</td>
<td>Tensor Methods</td>
<td>377</td>
</tr>
<tr>
<td>§16.4</td>
<td>State Space Models</td>
<td>382</td>
</tr>
<tr>
<td>§16.5</td>
<td>Exercises</td>
<td>388</td>
</tr>
<tr>
<td>Chapter 17</td>
<td>Model Selection and Bayesian Integrals</td>
<td>391</td>
</tr>
<tr>
<td>§17.1</td>
<td>Information Criteria</td>
<td>392</td>
</tr>
<tr>
<td>§17.2</td>
<td>Bayesian Integrals and Singularities</td>
<td>397</td>
</tr>
<tr>
<td>§17.3</td>
<td>The Real Log-Canonical Threshold</td>
<td>402</td>
</tr>
<tr>
<td>§17.4</td>
<td>Information Criteria for Singular Models</td>
<td>410</td>
</tr>
<tr>
<td>§17.5</td>
<td>Exercises</td>
<td>414</td>
</tr>
<tr>
<td>Chapter 18</td>
<td>MAP Estimation and Parametric Inference</td>
<td>415</td>
</tr>
<tr>
<td>§18.1</td>
<td>MAP Estimation General Framework</td>
<td>416</td>
</tr>
<tr>
<td>§18.2</td>
<td>Hidden Markov Models and the Viterbi Algorithm</td>
<td>418</td>
</tr>
<tr>
<td>§18.3</td>
<td>Parametric Inference and Normal Fans</td>
<td>424</td>
</tr>
<tr>
<td>§18.4</td>
<td>Polytope Algebra and Polytope Propogation</td>
<td>427</td>
</tr>
<tr>
<td>§18.5</td>
<td>Exercises</td>
<td>429</td>
</tr>
<tr>
<td>Chapter 19</td>
<td>Finite Metric Spaces</td>
<td>431</td>
</tr>
<tr>
<td>§19.1</td>
<td>Metric Spaces and The Cut Polytope</td>
<td>431</td>
</tr>
<tr>
<td>§19.2</td>
<td>Tree Metrics</td>
<td>439</td>
</tr>
<tr>
<td>§19.3</td>
<td>Finding an Optimal Tree Metric</td>
<td>445</td>
</tr>
<tr>
<td>§19.4</td>
<td>Toric Varieties Associated to Finite Metric Spaces</td>
<td>450</td>
</tr>
<tr>
<td>§19.5</td>
<td>Exercises</td>
<td>453</td>
</tr>
<tr>
<td>Bibliography</td>
<td></td>
<td>455</td>
</tr>
<tr>
<td>Index</td>
<td></td>
<td>475</td>
</tr>
</tbody>
</table>
Preface

Algebraic statistics is a relatively young field based on the observation that many questions in statistics are fundamentally problems of algebraic geometry. This observation is now at least twenty years old and the time seems ripe for a comprehensive book that could be used as a graduate textbook on this topic.

Algebraic statistics represents an unusual intersection of mathematical disciplines, and it is rare that a mathematician or statistician would come to work in this area already knowing both the relevant algebraic geometry and statistics. I have tried to provide sufficient background in both algebraic geometry and statistics so that a newcomer to either area would be able to benefit from using the book to learn algebraic statistics. Of course both statistics and algebraic geometry are huge subjects and the book only scratches the surface on either of these disciplines.

I made the conscious decision to introduce algebraic concepts alongside statistical concepts where they can be applied, rather than having long introductory chapters on algebraic geometry, statistics, combinatorial optimization, etc. that must be waded through first, or flipped back to over and over again, before all the pieces are put together. Besides the three introductory chapters on probability, algebra, and statistics (Chapters 2, 3, 5, respectively), this perspective is followed throughout the text. While this choice might make the book less useful as a reference book on algebraic statistics, I hope that it will make the book more useful as an actual textbook that students and faculty plan to learn from.

Here is a breakdown of material that appears in each chapter in the book.
Chapter 1 is an introductory chapter that shows how ideas from algebra begin to arise when considering elementary problems in statistics. These ideas are illustrated with the simple example of a Markov Chain. As statistical and algebraic concepts are introduced the chapter makes forward reference to other sections and chapters in the book where those ideas are highlighted in more depth.

Chapter 2 provides necessary background information in probability theory which is useful throughout the book. This starts from the axioms of probability, works through familiar and important examples of discrete and continuous random variables, and includes limit theorems that are useful for asymptotic results in statistics.

Chapter 3 provides necessary background information in algebra and algebraic geometry, with an emphasis on computational aspects. This starts from definitions of polynomial rings, their ideals, and the associated varieties. Examples are typically drawn from probability theory to begin to show how tools from algebraic geometry can be applied to study families of probability distributions. Some computational examples using computer software packages are given.

Chapter 4 is an in-depth treatment of conditional independence, an important property in probability theory that is essential for the construction of multivariate statistical models. To study implications between conditional independence models, we introduce primary decomposition, an algebraic tool for decomposing solutions of polynomial equations into constituent irreducible pieces.

Chapter 5 provides some necessary background information in statistics. It includes some examples of basic statistical models and hypothesis tests that can be performed in reference to those statistical models. This chapter has significantly fewer theorems than other chapters and is primarily concerned with introducing the philosophy behind various statistical ideas.

Chapter 6 provides a detailed introduction to exponential families, an important general class of statistical models. Exponential families are related to familiar objects in algebraic geometry like toric varieties. Nearly all models that we study in this book arise by taking semialgebraic subsets of the natural parameter space of some exponential family, making these models extremely important for everything that follows. Such models are called algebraic exponential families.

Chapter 7 gives an in-depth treatment of maximum likelihood estimation from an algebraic perspective. For many algebraic exponential families maximum likelihood estimation amounts to solving a system of polynomial equations. For a fixed model and generic data, the number of critical points
of this system is fixed and gives an intrinsic measure of the complexity of calculating maximum likelihood estimates.

**Chapter 8** concerns the geometry of the cone of sufficient statistics of an exponential family. This geometry is important for maximum likelihood estimation: maximum likelihood estimates exist in an exponential family if and only if the data lies in the interior of the cone of sufficient statistics. This chapter also introduces techniques from polyhedral and general convex geometry which are useful in subsequent chapters.

**Chapter 9** describes Fisher’s exact test, a hypothesis test used for discrete exponential families. A fundamental computational problem that arises is that of generating random lattice points from inside of convex polytopes. Various methods are explored including methods that connect the problem to the study of toric ideals. This chapter also introduces the hierarchical models, a special class of discrete exponential family.

**Chapter 10** concerns the computation of upper and lower bounds on cell entries in contingency tables given some lower dimensional marginal totals. One motivation for the problem comes from the sampling problem of Chapter 9: fast methods for computing bounds on cell entries can be used in sequential importance sampling, an alternate strategy for generating random lattice points in polytopes. A second motivation comes from certain data privacy problems associated with contingency tables. The chapter connects these optimization problems to algebraic methods for integer programming.

**Chapter 11** describes the exponential random graphs models, a family of statistical models used in the analysis of social networks. While these models fit in the framework of the exponential families introduced in Chapter 6, they present a particular challenge for various statistical analyses because they have a large number of parameters and the underlying sample size is small. They also present a novel area of study for application of Fisher’s exact test and studying the existence of maximum likelihood estimates.

**Chapter 12** concerns the use of algebraic methods for the design of experiments. Specific algebraic tools that are developed include the Gröbner fan of an ideal. Connections between the design of experiments and the toric ideals of Chapters 9 and 10 are also explored.

**Chapter 13** introduces the graphical statistical models. In graphical models, complex interactions between large collections of random variables are constructed using graphs to specify interactions between subsets of the random variables. A key feature of these models is that they can be specified either by parametric descriptions or via conditional independence constructions. This chapter compares these two perspectives via the primary decompositions from Chapter 4.
Chapter 14 provides a general introduction of statistical models with hidden variables. Graphical models with hidden variables are widely used in statistics, but the presence of hidden variables complicates their use. This chapter starts with some basic examples of these constructions, including mixture models. Mixture models are connected to secant varieties in algebraic geometry.

Chapter 15 concerns the study of phylogenetic models, certain hidden variable statistical models used in computational biology. The chapter highlights various algebraic issues involved with studying these models and their equations. The equations that define a phylogenetic model are known as phylogenetic invariants in the literature.

Chapter 16 concerns the identifiability problem for parametric statistical models. Identifiability of model parameters is an important structural feature of a statistical model. Identifiability is studied for graphical models with hidden variables and structural equation models. Tools are also developed for addressing identifiability problems for dynamical systems models.

Chapter 17 concerns the topic of model selection. This is a well-developed topic in statistics and machine learning, but becomes complicated in the presence of model singularities that arise when working with models with hidden variables. The mathematical tools to develop corrections come from studying the asymptotics of Bayesian integrals. The issue of non-standard asymptotics arises precisely at the points of parameter space where the model parameters are not identifiable.

Chapter 18 concerns the geometry of maximum a posteriori (MAP) estimation of the hidden states in a model. This involves performing computations in the tropical semiring. The related parametric inference problem studies how the MAP estimate changes as underlying model parameters change, and is related to problems in convex and tropical geometry.

Chapter 19 is a study of the geometry of finite metric spaces. Of special interest are the set of tree metrics and ultrametrics which play an important role in phylogenetics. More generally, the set of cut metrics are closely related to hierarchical models studied earlier in the book.

There are a number of other books that address topics in algebraic statistics. The first book on algebraic statistics was Pistone, Riccomagno, and Wynn’s text [PRW01] which is focused on applications of computational commutative algebra to the design of experiments. Pachter and Sturmfels’ book [PS05] is focused on applications of algebraic statistics to computational biology, specific phylogenetics and sequence alignment. Steueny’s book [Stu05] is specifically focused on the combinatorics of conditional independence structures. Aoki, Hara, and Takemura [AHT12] give a detailed study of Markov bases which we discuss in Chapter 9. Zwiernik
[Zwi16] gives an introductory treatment of tree models from the perspective of real algebraic geometry. Drton, Sturmfels, and I wrote a short book [DSS09] based on a week long short course we gave at the Mathematisches Forschungsinstitut Oberwolfach (MFO). While there are many books in algebraic statistics touching on a variety of topics, this is the first one that gives a comprehensive treatment. I have tried to add in sufficient background and provide many examples and exercises so that algebraic statistics might be picked up by a nonexpert.

My first attempt at a book on algebraic statistics was in 2007 with Mathias Drton. That project eventually led to the set of lecture notes [DSS09]. As part of Mathias’s and my first attempt at writing, we produced two background chapters on probability and algebra which were not used in [DSS09] and which Mathias has graciously allowed me to use here.

I am grateful to a number of readers who provided feedback on early drafts of the book. These include Elizabeth Allman, Daniel Bernstein, Jane Coons, Mathias Drton, Eliana Duarte, Elizabeth Gross, Serkan Hosten, David Kahle, Thomas Kahle, Kaie Kubjas, Christian Lehn, Colby Long, Frantisek Matuš, Nicolette Meshkat, John Rhodes, Elina Robeva, Anna Seigal, Rainer Sinn, Ruriko Yoshida, as well as four anonymous reviewers. Tim Römer at the University Osnabrück also organized a reading course on the material that provided extensive feedback. David Kahle also provided extensive help in preparing examples of statistical analyses using R. My research during the writing of this book has been generously supported by the David and Lucille Packard Foundation and the US National Science Foundation under grants DMS 0954865 and DMS 1615660.
Chapter 1

Introduction

Algebraic statistics advocates using tools from algebraic geometry, commutative algebra, combinatorics and related tools from symbolic computation to address problems in probability theory, statistics, and their applications. The connection between the algebra and statistics sides goes both directions and statistical problems with algebraic character help to focus new research directions in algebra.

While the area called algebraic statistics is relatively new, connections between algebra and statistics are old going back to the beginnings of statistics. For instance, the constructions of different types of combinatorial designs use group theory and finite geometries [Bos47], algebraic structures being used to describe central limit theorems in complex settings [Gre63], or representation theoretic methods in the analysis of discrete data [Dia88]. In spite of these older contact points, the term algebraic statistics has primarily been used for the connection between algebraic geometry and statistics which is the focus of this book, a topic that has developed since the mid 1990’s.

Historically, the main thread of algebraic statistics started with the work of Diaconis and Sturmfels [DS98] on conditional inference, establishing a connection between random walks on sets of contingency tables and generating sets of toric ideals. Inspired by [DS98], Pistone, Riccomagno, and Wynn explored connections between algebraic geometry and the design of experiments, describing their work in the monograph [PRW01], which coined the name algebraic statistics. Since then there has been an explosion of research in the area. The goal of this book is to illustrate and explain some of the advances in algebraic statistics highlighting the main areas of research directions since those first projects. Of course, it is impossible to highlight
everything, and new results are constantly being added, but we have tried to hit major points.

Whenever two fields come together, it is tempting to form a “dictionary” connecting the two areas. In algebraic statistics there are some concepts in statistics which directly correspond to objects in algebraic geometry, but there are many other instances where we only have a very rough correspondence. In spite of this, it can be useful to keep these correspondences in mind, some of which are illustrated in the table below.

<table>
<thead>
<tr>
<th>Probability/Statistics</th>
<th>Algebra/Geometry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability distribution</td>
<td>Point</td>
</tr>
<tr>
<td>Statistical Model</td>
<td>(Semi)Algebraic set</td>
</tr>
<tr>
<td>Exponential family</td>
<td>Toric variety</td>
</tr>
<tr>
<td>Conditional inference</td>
<td>Lattice points in polytopes</td>
</tr>
<tr>
<td>Maximum likelihood estimation</td>
<td>Polynomial optimization</td>
</tr>
<tr>
<td>Model selection</td>
<td>Geometry of singularities</td>
</tr>
<tr>
<td>Multivariate Gaussian model</td>
<td>Spectrahedral geometry</td>
</tr>
<tr>
<td>Phylogenetic model</td>
<td>Tensor networks</td>
</tr>
<tr>
<td>MAP estimates</td>
<td>Tropical geometry</td>
</tr>
</tbody>
</table>

The goal of this book is to illustrate these connections. In the remainder of this chapter we illustrate some of the ways that algebra arises when thinking about statistics problems by illustrating it with the example of a discrete Markov chain. This is only intended as a single illustrative example to highlight some of the insights and connections we make in algebraic statistics. The reader should not concern themselves if they do not understand everything. More elementary background material will appear in later chapters.

### 1.1. Discrete Markov Chain

Let $X_1, X_2, \ldots, X_m$ be a sequence of random variables on the same state space $\Sigma$, a finite alphabet. Since these are random variables there is a joint probability distribution associated with them, i.e. to each tuple $x_1, \ldots, x_m \in \Sigma$ there is a real number

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m)$$

which is the probability that $X_1 = x_1, X_2 = x_2, \ldots$, et cetera. To be a probability distribution, These numbers are between 0 and 1, and the sum over all $(x_1, \ldots, x_m) \in \Sigma^m$ is one.

**Definition 1.1.1.** The sequence $X_1, X_2, \ldots, X_m$ is called a Markov chain if for all $i = 3, \ldots, m$ and for all $x_1, \ldots, x_i \in \Sigma$

\[
P(X_i = x_i | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}) = P(X_i = x_i | X_{i-1} = x_{i-1}).
\]
In other words, in a Markov chain, the conditional distribution of \( X_i \)
given all predecessors only depends on the most recent predecessor. Or said
in another way, adding knowledge further back in the chain does not help
in predicting the next step in the chain.

Let us consider the model of a discrete Markov chain, that is for a fixed
\( m \) and \( \Sigma \), the set of all probability distributions consistent with the condition
in Definition 1.1.1. As for many statistical models, the underlying descrip-
tion of this model is fundamentally algebraic in character: in particular, it
is described implicitly by polynomial constraints on the joint distribution.
Here a specific example.

**Example 1.1.2** (Three Step Markov Chain). Let \( m = 3 \) and \( \Sigma = \{0, 1\} \) in
the Markov chain model. We can naturally associate a probability distri-
bution in this context with a point in \( \mathbb{R}^8 \). Indeed, the joint distribution is
determined by the 8 values \( P(X_1 = i, X_2 = j, X_3 = k) \) for \( i, j, k \in \{0, 1\} \). We
use the following shorthand for this probability distribution: \( p_{ijk} = P(X_1 = i, X_2 = j, X_3 = k) \). Then a joint probability distribution for three binary
random variables is a point
\[
(p_{000}, p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, p_{111}) \in \mathbb{R}^8.
\]
Computing conditional distributions gives rational expressions in terms of
the joint probabilities:
\[
P(X_3 = k | X_1 = i, X_2 = j) = \frac{p_{ijk}}{p_{ij+}}
\]
where the “+” in the subscript denotes a summation, e.g.
\[
p_{ij+} = \sum_{k \in \{0, 1\}} p_{ijk}.
\]
The condition from (1.1.1) for a distribution to come from a Markov chain
translates into the rational expression
\[
\frac{p_{ijk}}{p_{ij+}} = \frac{p_{+jk}}{p_{+j+}}
\]
for all \( i, j, k \in \{0, 1\} \). We can also represent this as
\[
\frac{p_{ijk}}{p_{ij+}} = \frac{p_{i'jk}}{p_{i'j+}}
\]
for all \( i, i', j, k \in \{0, 1\} \) by setting two different left hand sides of Equation
1.1.2 equal to each other for the same values of \( j \) and \( k \). Clearing denom-
inators to get polynomial expressions and expanding and simplifying these
expressions yields the following characterization of distributions that are
Markov chains:

A vector \( p = (p_{000}, p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, p_{111}) \in \mathbb{R}^8 \)
is the probability distribution from the Markov chain model if and only if it satisfies
(1) $p_{ijk} \geq 0$ for all $i, j, k \in \{0, 1\}$,
(2) $\sum_{i,j,k \in \{0, 1\}} p_{ijk} = 1$,
(3) $p_{000}p_{101} - p_{001}p_{100} = 0$, and
(4) $p_{010}p_{111} - p_{011}p_{110} = 0$.

Remark 1.1.3. Example 1.1.2 illustrates a theme that will recur throughout the book and is a hallmark of algebraic statistics: statistical models are semialgebraic sets, that is, they can be represented as the solutions sets of systems of polynomials equations and inequalities.

The Markov chain model is an example of a conditional independence model, that is, it is specified by conditional independence constraints on the random variables in the model. Equation 1.1.1 is equivalent to the conditional independence statement

$X_i \independent (X_1, \ldots, X_{i-2})|X_{i-1}$.

Conditional independence models will be studied in more detail in Chapter 4 and play an important role in the theory of graphical models in Chapter 13. In spite of the commonality of conditional independence models, it is more typical to specify models parametrically, and the Markov chain model also has a natural parametrization. To “discover” the parametrization, we use a standard factorization for joint distributions into their conditional distributions:

$P(X_1 = x_1, \ldots, X_m = x_m) = \prod_{i=1}^{m} P(X_i = x_i|X_1 = x_1, \ldots, X_{i-1} = x_{i-1})$.

Note that the $i = 1$ term in this product is just the marginal probability $P(X_1 = x_1)$. Then we can substitute in using Equation 1.1.1 to get

$P(X_1 = x_1, \ldots, X_m = x_m) = \prod_{i=1}^{m} P(X_i = x_i|X_{i-1} = x_{i-1})$.

To discover the parametrization, we just treat each of the individual conditional distributions as free parameters. Let $\pi_{j_1} = P(X_1 = j_1)$ and let $\alpha_{i,j,k} = P(X_i = k|X_{i-1} = j)$. So we have the Markov chain model described by a polynomial parametrization:

$p_{j_1,j_2,\ldots,j_m} = \pi_{j_1} \prod_{i=2}^{m} \alpha_{i,j_{i-1},j_i}$.

Remark 1.1.4. The Markov chain example illustrates another key feature of algebraic statistics highlighted throughout the book: many parametric statistical models have their parametric representations given by polynomial functions of the parameters.
As we have seen, algebra plays an important role in characterizing the distributions that belong to a statistical model. Various data analytic questions associated with using the statistical model also have an algebraic character. Here we discuss two such problems (which we state informally): fitting the model to the data and determining whether or not the model fits the data well. Both of these notions will be made more precise in later chapters.

There are many possible notions for fitting a model to data. Here we use the frequentist maximum likelihood estimate. First of all, there are many different types of data that one might receive that might be compatible with the analysis of a particular statistical model. Here we focus on the simplest which is independent and identically distributed data. That is, we assume there is some true unknown distribution \( p \) according to which all of our data are independent samples from this underlying distribution. Focusing on the case of the Markov chain model of length 3 on a binary alphabet \( \Sigma = \{0, 1\} \), the data then would be a collection of elements of \( \{0, 1\}^3 \) for example:

\[
000, 010, 110, 000, 110, 100, 010, 110, 111, 000, 000, 010
\]

would be a data set with 13 data points. Since we assume the data is generated independently from the same underlying distribution, the probability of observing the data really only depends on the vector of counts \( u \), whose entry \( u_{ijk} \) records the number of times the configuration \( ijk \) occurred in the data. In this case we have

\[
(u_{000}, u_{001}, u_{010}, u_{011}, u_{100}, u_{101}, u_{110}, u_{111}) = (4, 0, 3, 0, 1, 1, 3, 1).
\]

The probability of observing the dataset \( D \) given the true unknown probability distribution \( p \) is then

\[
P(\text{Data}) = \prod_{ijk} p_{ijk}^{u_{ijk}} = L(p \mid u).
\]

This probability is called the likelihood function. While the likelihood is just a probability calculation, from the standpoint of using it in statistical inference, we consider the data vector \( u \) as given to us and fixed, the probability \( p \) unknown parameters that we are trying to discover.

The maximum likelihood estimate is the probability distribution in our model that maximizes the likelihood function. In other words, the maximum likelihood estimate is the probability distribution in our model that makes the probability of the observed data as large as possible. For a given probabilistic model \( \mathcal{M} \), the maximum likelihood estimation problem is to compute

\[
\arg \max_{p \in \mathcal{M}} L(p \mid u) \quad \text{subject to } p \in \mathcal{M}.
\]
The argmax function asks for the maximizer rather than the maximum value of the function. For our discrete Markov chain model, we have representations of the model in both implicit and parametric forms, which means we can express this as either a constrained or unconstrained optimization problem, i.e.

\[
\arg \max \prod_{ijk} p_{ijk} \text{ subject to } p_{ijk} \geq 0, \sum_{ijk} p_{ijk} = 1, \text{ and } \\
p_{000}p_{101} - p_{001}p_{100} = p_{010}p_{111} - p_{011}p_{110} = 0.
\]

On the other hand, we can also express this as an unconstrained optimization problem, directly plugging in the parametrized form of the Markov chain.

\[
\arg \max \left( \pi_i \alpha_{ij} \beta_{jk} \right)^{u_{ijk}}.
\]

Note this optimization problem is not a completely unconstrained optimization problem, because we need \( \pi \) to be a probability distribution and \( \alpha \) and \( \beta \) to be conditional distribution. That is, we have the constraints

\[
\pi_i \geq 0, \quad \sum_i \pi_i = 1, \quad \alpha_{ij} \geq 0, \quad \sum_j \alpha_{ij} = 1, \quad \beta_{jk} \geq 0, \quad \sum_k \beta_{jk} = 1.
\]

A typical direct strategy to compute the maximum likelihood estimator is to compute the logarithm of the likelihood, and compute partial derivatives and set them equal to zero. Since our model is algebraic in nature, these score equations (or critical equations) give an algebraic system of polynomial equations which can be solved using symbolic techniques or numerically in specific instances. In this particular situation, there is a closed form rational expression for the maximum likelihood estimates. In terms of the parameters, these are

\[
(1.1.3) \quad \hat{\pi}_i = \frac{u_{i++}}{u_{+++}}, \quad \hat{\alpha}_{ij} = \frac{u_{ij+}}{u_{i++}}, \quad \hat{\beta}_{jk} = \frac{u_{+jk}}{u_{+j+}}
\]

where

\[
u_{i++} = \sum_{j,k} u_{ijk}, \quad u_{+++} = \sum_{i,j,k} u_{ijk}, \quad u_{ij+} = \sum_k u_{ijk}, \quad \text{etc.}
\]

In terms of probability distribution in the model we can multiply these together to get

\[
\hat{p}_{ijk} = \frac{u_{ij+} \cdot u_{+jk}}{u_{+++} \cdot u_{+j+}}.
\]

So for our example data set with data vector \( u = (4, 0, 3, 0, 1, 1, 3, 1) \) we arrive at the maximum likelihood estimate

\[
\hat{p} = \left( \frac{4 \cdot 5}{6 \cdot 13}, \frac{4 \cdot 1}{6 \cdot 13}, \frac{3 \cdot 6}{7 \cdot 13}, \frac{3 \cdot 1}{7 \cdot 13}, \frac{2 \cdot 5}{6 \cdot 13}, \frac{2 \cdot 1}{6 \cdot 13}, \frac{4 \cdot 6}{7 \cdot 13}, \frac{4 \cdot 1}{7 \cdot 13} \right) \\
\approx (0.256, 0.051, 0.198, 0.033, 0.128, 0.026, 0.264, 0.044)
\]
1.1. Discrete Markov Chain

In general, the score equations for a statistical model rarely have such simple closed form solutions—i.e. systems of polynomial equations often have multiple solutions. It is natural to ask: “For which statistical models does such a nice closed form solution arise?” “Can we say anything about the form of these closed form expressions?” Amazingly a complete, beautiful, yet still mysterious classification of models with rational formulas for maximum likelihood estimates exists and has remarkable connections to other areas in algebraic geometry. This was discovered by June Huh [Huh14] and will be explained in Chapter 7.

Remark 1.1.5. The classification of statistical models with rational maximum likelihood estimates illustrates a frequent point in algebraic statistics. In particular, trying to classify and identify statistical models that satisfy some nice statistical property lead to interesting classification theorems in algebraic geometry and combinatorics.

Now that we have fit the model to the data, we can ask: how well does the model fit the data? In other words, do we think that the data was generated by some distribution in the model? In this case one often might be interested in performing a hypothesis test, to either accept or reject the hypothesis that the data generating distribution belongs to the model. (Caveat: usually in statistics we might only come to a conclusion about either rejecting containment in the model or deciding that the test was inconclusive.)

Here is one typical way to try to test whether a given model fits the data. We instead pose the following question: Assuming that the true underlying distribution $\pi$ belongs to our model $\mathcal{M}$, among all data sets $v$ that could have been generated, what proportion of such data $v$ is more likely to have occurred than our observed data set $u$. From this we compute a $p$-value which might allow us to reject the hypothesis that the distribution $\pi$ actually belongs to the model $\mathcal{M}$. Essentially, we will reject data as coming from the model, if it had a (relatively) low probability of being generated from the model.

Of course, to measure the probability that a particular data vector $v$ appeared depends on the value of the (unknown) probability distribution $p$, and so it is impossible to directly measure the proportion from the previous paragraph. A key insight of Fisher’s, leading to Fisher’s Exact Test, is that for some statistical models, the dependence of the likelihood function on the data and the parameter values is only through a lower order linear function of the data vector. These lower order functions of the data are called sufficient statistics. Indeed, for a particular vector of counts $u$, for the two step Markov chain whose likelihood we viewed above, in terms of
the parameters we have
\[
P(u|\alpha, \beta, \gamma) = \prod_{i,j,k} (\pi_i \alpha_{ij} \beta_{jk})^{u_{ijk}}
\]
\[
= \binom{n}{u} \prod_{i,j} (\pi_i \alpha_{ij})^{u_{ij}+} \prod_{j,k} \beta_{jk}^{u_{+jk}}.
\]

The quantities \(u_{ij+}, u_{+jk}\) with \(i, j, k \in \Sigma\) are the sufficient statistics of the Markov chain model of length three. Note, since we have here written the probability of observing the vector of counts \(u\), rather than an ordered list of observations, we include the multinomial coefficient in the likelihood.

Fisher’s idea was that we can use this to pose a different question: among all possible data sets with the same sufficient statistics as the given data \(u\), what proportion are more likely to occur than \(u\)?

From the standpoint of computing a probability, we are looking at the conditional probability distribution
\[
P(v|\alpha, \beta, \gamma, v_{ij+} = u_{ij+}, v_{+jk} = u_{+jk}) = \frac{\binom{n}{u} \prod_{i,j} (\pi_i \alpha_{ij})^{u_{ij}+} \prod_{j,k} \beta_{jk}^{u_{+jk}}}{\sum_v \binom{n}{v} \prod_{i,j} (\pi_i \alpha_{ij})^{v_{ij}+} \prod_{j,k} \beta_{jk}^{v_{+jk}}}
\]
where the sum in the denominator is over all \(v \in \mathbb{N}^8\) such that \(v_{ij+} = u_{ij+}, v_{+jk} = u_{+jk}\). Since the term that depends on \(p\) in this expression is the same in every summand, and equals the product in the numerator, we see that in fact
\[
P(v|\alpha, \beta, \gamma, v_{ij+} = u_{ij+}, v_{+jk} = u_{+jk}) = \binom{n}{v} \sum_v \binom{n}{v}
\]
and in particular, this probability does not depend on the unknown distribution \(p\). The fact that this probability does not depend on unknown distribution \(p\) is the reason for adding the condition “with the same sufficient statistics as the given data \(u\)”. It allows us to perform a hypothesis test without worrying about this unknown “nuisance parameter”.

Let
\[
\mathcal{F}(u) = \{v \in \mathbb{N}^8 : v_{ij+} = u_{ij+}, v_{+jk} = u_{+jk} \text{ for all } i, j, k\},
\]
that is, the set of all vectors of counts with the same sufficient statistics as \(u\). To carry out Fisher’s exact test amounts to computing the ratio
\[
\frac{\sum_{v \in \mathcal{F}(u)} \binom{n}{v} 1_{(v)<(u)}(v)}{\sum_{v \in \mathcal{F}(u)} \binom{n}{v}}
\]
where \(1_{(v)<(u)}(v)\) denotes the indicator function of the event that \(\binom{n}{v} < \binom{n}{u}\). The resulting number, the p-value of Fisher’s exact test, gives the probability that a random table drawn from \(\mathcal{F}(u)\) according to the generalized
hypergeometric distribution (s.t. \( P(v) \propto \binom{v}{n} \)) has higher probability than the observed table \( u \). Of course, this raises all sorts of new questions, some of which we will explore in the present book. It is usually impossible to actually enumerate all the \( v \in \mathcal{F}(u) \) to calculate the sum, and one must find a way to generate them at random, using a sampling strategy. Note that the set \( \mathcal{F}(u) \) consists of the set of lattice points that satisfy a given system of linear equations and inequalities:

\[
\mathcal{F}(u) = \left\{ v \in \mathbb{Z}^n : v_{ijk} \geq 0, \sum_k v_{ijk} = u_{i+j}, \text{ and } \sum_i v_{ijk} = u_{+jk} \right\}.
\]

This leads to the theory of Markov bases, studied in Chapter 9, and algebraic and combinatorial methods for sampling random lattice points in polytopes that allows one to apply Fisher’s exact test.

### 1.2. Exercises

**Exercise 1.1.** Derive the formula for the maximum likelihood estimates in the discrete Markov chain model from Equation 1.1.3.

**Exercise 1.2.** Call a Markov chain *homogeneous* if it satisfies the following extra condition:

\[
P(X_i = a | X_{i-1} = b) = P(X_j = a | X_{j-1} = b)
\]

for all \( a, b \in \Sigma \) and \( i, j \in \{2, \ldots, n\} \).

1. For \( n = 3 \) and \( \Sigma = \{0, 1\} \), find an implicit algebraic representation of the homogeneous Markov chain model.
2. Find a parametric description of the model.
3. Find closed form formulas for the maximum likelihood estimates in this model.
4. Describe the minimal sufficient statistics of the model and the resulting polytope of data points satisfying the condition that those points have the same sufficient statistics as a given data point \( u \).
Probability Primer

Probability theory provides mathematical language to describe the observation of random quantities or quantities that arise in processes that are deterministic but too complicated to be predictable. The language of probability is fundamental to the design and evaluation of statistical methodology. In this chapter we review, at an elementary level, basic notions from probability theory such as random variables, expectation, and limit theorems. We also define the normal or Gaussian distribution that will play a prominent role throughout the entire book. The discussion of the topics in this chapter will be brief and the reader will certainly benefit by consulting in-depth treatments of probability theory as found in [Bil95], [Chu01] or [Sho00].

2.1. Probability

Phrases such as “there is an 80% chance of rain this coming Saturday” testify to the fact that the concept of probability is omnipresent in our daily lives. Nevertheless it is not entirely straightforward to give a dictionary-definition of probability. Certainly, the probability of a coin toss resulting in heads being equal to 50% can be understood as a long-run frequency. If the coin is tossed repeatedly, then roughly 50% of the tosses will show heads. However, thinking of probability as a long-run frequency need not always be natural. For example, if the above statement about the probability of rain is to be interpreted in this way, then, strictly speaking, we would have to be able to relive this coming Saturday over and over again. Hence, it can sometimes be more appropriate to understand probability simply as the quantification of our belief about the occurrence of the event in question. Placing a bet on the event happening provides a way to quantify this belief.
We will not dwell, any further, on the nature of probability and instead push towards mathematical abstraction, in which probability is turned into a mathematical object. This object is a map that assigns to each set of interest a non-negative real number, which is called the probability of the event represented by the set. Reassuringly, this framework will lead to a theorem, the Law of Large Numbers in Section 2.5, that recovers the interpretation in terms of long-run frequencies.

Before giving a first formal definition of probability, a clarification of language is in order. We will refer to a process that yields random outcomes as a random experiment. The set of all possible outcomes in such a random experiment is the sample space, denoted by $\Omega$. An event is simply a subset $A \subseteq \Omega$, and we say that the event $A$ occurs if the random experiment results in an outcome $\omega \in A$. The symbol $\mathcal{P}(\Omega)$ is used to denote the power set of the sample space $\Omega$.

**Definition 2.1.1.** A probability measure on a finite or countably infinite sample space $\Omega$ is a map  
$$P : \mathcal{P}(\Omega) \to [0, 1]$$  
$$A \mapsto P(A)$$  
such that $P(\Omega) = 1$ and  
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \quad \text{if} \quad A_i \cap A_j = \emptyset \quad \text{for all} \quad i \neq j.$$  

This definition involves countably infinite as opposed to finite collections of pairwise disjoint sets, which facilitates the later development of limit theorems. These theorems will be seen to be useful for approximating the behavior of procedures for statistical inference.

From Definition 2.1.1 a number of intuitive properties of probability measures can be derived. For example: the empty set has probability $P(\emptyset) = 0$; we have finite additivity  
$$P(A_1 \cup A_2) = P(A_1) + P(A_2) \quad \text{if} \quad A_1 \cap A_2 = \emptyset$$  
and inclusion-exclusion formulas of the type  
$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A \cap A_2);$$

and we can bound probabilities such as  
$$P(A_1) \leq P(A_2), \quad \text{if} \quad A_1 \subseteq A_2.$$  
Since we included countable collection of sets in Definition 2.1.1 we can also derive facts such as  
$$\lim_{n \to \infty} P(A_n) = P(\bigcup_{n=1}^{\infty} A_n)$$
for an increasing sequence of sets $A_1 \subseteq A_2 \subseteq \ldots$.

**Example 2.1.2** (Lottery). Suppose an urn contains 49 balls labelled by the integers 1 through 49. Six balls are drawn at random from this urn without replacement. What is the probability that exactly three of the balls show a number smaller than 10?

Assuming that the order in which the balls are drawn is not recorded, we can formalize the problem as follows. The sample space is the set of all six-element subsets of $[49] = \{1, \ldots, 49\}$, i.e.,
$$\Omega = \{W \subseteq [49] : \#W = 6\}.$$ Since a set $W = \{\omega_1, \ldots, \omega_6\}$ is unordered, we can assume that $\omega_1 < \omega_2 < \cdots < \omega_6$. The question then asks for the probability of the event
$$A = \{\{\omega_1, \ldots, \omega_6\} \in \Omega : \omega_i < 10, i = 1, 2, 3, \text{ and } w_i \geq 10, i = 4, 5, 6\}.$$ However, this question can be answered only if “drawing at random” is interpreted to specify a particular probability measure. As customary, we will interpret the statement to mean that the probability measure $P$ assigns equal probability to each one of the
$$\#\Omega = \binom{49}{6} = 13,983,816$$ singletons $\{\omega\}, \omega \in \Omega$. This implies that the desired probability equals
$$P(A) = \frac{\#A}{\#\Omega} = \frac{\binom{3}{3}\binom{46}{3}}{\binom{49}{6}} \approx 0.0593.$$

What if the order in which the balls are drawn was recorded? Clearly that would alter the sample space but does this change in sample space affect the probability of the considered event? The answer is no, and Exercise 2.1 asks you to go through the details of why the probability in question is still equal to the value found in (2.1.3).

Definition 2.1.1 was concerned with a finite or countably infinite sample space. However, when observing continuous quantities it is desirable to allow uncountably infinite sample spaces such as an interval. Unfortunately, it is then problematic to define probability measures on the power set $\mathcal{P}(\Omega)$.

**Proposition 2.1.3.** If the sample space in Definition 2.1.1 is allowed to be the uncountably infinite set $\Omega = [0, 1)$, then there does not exist a translation-invariant probability measure on $\mathcal{P}([0, 1))$.

**Proof.** Suppose $P$ is a translation-invariant probability measure on the power set $\mathcal{P}([0, 1))$. Define $x, y \in [0, 1)$ to be equivalent if there exists a rational $q \in \mathbb{Q}$ such that $x = y + q \mod 1$. Choose a set $A$ that contains
The nonexistence of a “uniform distribution” on $\mathcal{P}([0, 1))$ arises because an interval has too many subsets. The way around this problem is to define probability measures only on suitable subsets of $\mathcal{P}(\Omega)$.

**Definition 2.1.4.** A $\sigma$-algebra over a set $\Omega$ is a subset $\mathcal{A}$ of the power set $\mathcal{P}(\Omega)$ such that

1. $\Omega \in \mathcal{A}$,
2. for all $A \in \mathcal{A}$, the complement $A^c \in \mathcal{A}$, and
3. if $A_1, A_2, \ldots \in \mathcal{A}$ then the countable union $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

**Definition 2.1.5.** Let $\mathcal{A}$ be $\sigma$-algebra on a sample space $\Omega$. A **probability measure** on $(\Omega, \mathcal{A})$ is a map

$$P : \mathcal{A} \to [0, 1]$$

$$A \mapsto P(A)$$

such that $P(\Omega) = 1$ and

$$P \left( \bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} P(A_i) \quad \text{if} \quad A_i \cap A_j = \emptyset \quad \text{for all} \ i \neq j.$$

The triple $(\Omega, \mathcal{A}, P)$ is referred to as a **probability space**.

For finite or countably infinite sample spaces $\Omega$ the $\sigma$-algebra $\mathcal{A}$ can always be chosen to be the power set $\mathcal{P}(\Omega)$. The most important non-trivial $\sigma$-algebra is the **Borel $\sigma$-algebra** $\mathcal{B}_k$, which is the smallest $\sigma$-algebra over $\mathbb{R}^k$ that contains all rectangles of the form $(-\infty, b_1) \times \cdots \times (-\infty, b_k)$ with $b_i \in \mathbb{R}$. We will refer to the sets in the Borel $\sigma$-algebra, which include all open and all closed sets, as **measurable sets**. Measurable sets are of course a key ingredient to Lebesgue integration, which provides the following tool to define probability measures on $(\mathbb{R}^k, \mathcal{B}_k)$.

**Definition 2.1.6.** A **density function** is an integrable function $f : \mathbb{R}^k \to [0, \infty)$ with

$$\int_{\mathbb{R}^k} f(x) \, dx = 1.$$
Given a density function $f$, the assignment

$$P(A) = \int_A f(x) \, dx$$

defines a probability measure $P$ on $(\mathbb{R}^k, \mathcal{B}_k)$.

A density function of a probability measure is clearly not unique. If $N$ is a set of Lebesgue measure zero, then altering the function values $f(x)$ for $x \in N$ yields another density function. However, two density functions $f$ and $g$ for the same probability measure are almost surely equal in the sense that the set \{ $x \in \mathbb{R}^k : f(x) \neq g(x)$ \} must have Lebesgue measure zero.

**Example 2.1.7** (Normal distribution). Suppose $\mu \in \mathbb{R}$ and $\sigma > 0$. The probability measure $\mathcal{N}(\mu, \sigma^2)$ defined by the density function

$$\phi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp \left\{ \frac{(x - \mu)^2}{2\sigma^2} \right\}, \quad x \in \mathbb{R}$$

is called the normal distribution with mean $\mu$ and variance $\sigma^2$. Note that $\phi_{\mu,\sigma}(x) > 0$ for all $x \in \mathbb{R}$. The fact that $\phi_{\mu,\sigma}$ integrates to one can be seen as follows. First, using the substitution $x = \sigma y + \mu$, we may assume that $\mu = 0$ and $\sigma = 1$ in which case the probability measure $\mathcal{N}(0,1)$ is known as the standard normal distribution. Now consider the squared integral

$$\left( \int_{\mathbb{R}} \phi_{0,1}(x) \, dx \right)^2 = \int_{\mathbb{R}^2} \frac{1}{2\pi} \exp \left\{ -\frac{1}{2}(x^2 + y^2) \right\} \, dx \, dy.$$

Transforming to polar coordinates via $x = r \cos \theta$ and $y = r \sin \theta$ we find that

$$\int_{r=0}^{\infty} \int_{\theta=0}^{2\pi} \frac{1}{2\pi} \exp \left\{ -\frac{r^2}{2} \right\} r \, d\theta \, dr = \int_{r=0}^{\infty} \exp \left\{ -\frac{r^2}{2} \right\} r \, dr = 1.$$

Often one is concerned with not just a single event but two or more events, in which case it is interesting to ask questions of the type: what is the chance that event $A$ occurs given the information that event $B$ has occurred? This question could be addressed by redefining the sample space from $\Omega$ to $B$ and restricting the given probability measure $P$ to $B$. However, it is more convenient to adopt the following notion of conditional probability that avoids an explicit change of sample space.

**Definition 2.1.8.** Let $A, B \in \mathfrak{A}$ be two events such that the probability measure $P$ on $(\Omega, \mathfrak{A})$ assigns positive probability $P(B) > 0$ to event $B$. Then the conditional probability of event $A$ given event $B$ is the ratio

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$
The following rules for computing with conditional probabilities are immediate consequences of Definition 2.1.8.

**Lemma 2.1.9** (Total probability and Bayes’ rule). Suppose $P$ is a probability measure on $(\Omega, \mathcal{A})$ and $A_1, \ldots, A_n \in \mathcal{A}$ are pairwise disjoint sets that form a partition of $\Omega$. Then we have the two equalities

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i)$$

and

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{n} P(A \mid B_j)P(B_j)},$$

which are known as the law of total probability and Bayes’ rule, respectively.

**Example 2.1.10** (Urn). Suppose an urn contains two green and four red balls. In two separate draws two balls are taken at random and without replacement from the urn and the ordered pair of their colors is recorded. What is the probability that the second ball is red?

To answer this we can apply the law of total probability. Let $A = \{(r, r), (g, r)\}$ be the event of the second ball being red, and $B_r = \{(r, r), (r, g)\}$ and $B_g = \{(g, r), (g, g)\}$ the events defined by fixing the color of the first ball. Then we can compute

$$P(A) = P(A \mid B_r)P(B_r) + P(A \mid B_g)P(B_g) = \frac{3}{5} \cdot \frac{4}{6} + \frac{4}{5} \cdot \frac{2}{6} = \frac{2}{3}.$$  

What is the probability that the first ball is green given that the second ball is red?

Now Bayes’ rule yields

$$P(B_g \mid A) = \frac{P(A \mid B_g)P(B_g)}{P(A \mid B_r)P(B_r) + P(A \mid B_g)P(B_g)} = \frac{4}{5} \cdot \frac{2}{6} \cdot \frac{3}{2} = \frac{2}{5}.$$  

In the urn example, the conditional probability $P(B_g \mid A)$ is larger than the unconditional probability of the first ball being green, which is $P(B_g) = 1/3$. This change in probability under conditioning captures the fact that due to the draws occurring without replacement, knowing the color of the second ball provides information about the color of the first ball.

**Definition 2.1.11.** An event $A$ is independent of a second event $B$ if either $P(B) > 0$ and $P(A \mid B) = P(A)$ or $P(B) = 0$.

The following straightforward characterization of independence shows that the relation is symmetric.

**Lemma 2.1.12.** Two events $A$ and $B$ are independent if and only if

$$P(A \cap B) = P(A)P(B).$$
2.2. Random Variables and their Distributions

**Proof.** If \( P(B) = 0 \), then \( A \) and \( B \) are independent and also \( P(A \cap B) = 0 \). If \( P(B) > 0 \) and \( A \) is independent of \( B \), then

\[
P(A) = P(A \mid B) = \frac{P(A \cap B)}{P(B)} \iff P(A \cap B) = P(A)P(B).
\]

The product formulation of independence is often the starting point for discussing more advanced concepts like conditional independence that we will see in Chapter 4. However, Definition 2.1.11 gives the more intuitive notion that knowing that \( B \) happened does not give any further information about whether or not \( A \) happened. In light of Lemma 2.1.12 we use the following definition to express that the occurrence of any subset of these events provides no information about the occurrence of the remaining events.

**Definition 2.1.13.** The events \( A_1, \ldots, A_n \) are **mutually independent** or **completely independent** if

\[
P\left( \bigcap_{i \in I} A_i \right) = \prod_{i \in I} P(A_i), \quad \text{for all } I \subseteq [n].
\]

Under mutual independence of the events \( A_1, \ldots, A_n \), it holds that

\[
P\left( \bigcap_{i \in I} A_i \mid \bigcap_{j \in J} A_j \right) = P\left( \bigcap_{i \in I} A_i \right)
\]

for any pair of disjoint index sets \( I, J \subseteq [n] \). Exercise 2.4 discusses an example in which it becomes clear that it does not suffice to consider independence of all pairs \( A_i \) and \( A_j \) in a definition of mutual independence of several events.

### 2.2. Random Variables and their Distributions

Many random experiments yield numerical quantities as outcomes that the observer may modify using arithmetic operations. Random variables are a useful concept for tracking the change of probability measures that is induced by calculations with random quantities. At an intuitive level a random variable is simply a “random number” and the distribution of a random variable determines the probability with which such a random number belongs to a given interval. The concept is formalized as follows.

**Definition 2.2.1.** Let \( (\Omega, \mathcal{A}, P) \) be a probability space. A **random variable** is a function \( X : \Omega \to \mathbb{R} \) such that for every measurable set \( B \) it holds that the preimage \( X^{-1}(B) = \{ X \in B \} \in \mathcal{A} \). The **distribution** of a random variable \( X \) is the probability measure \( P^X \) on \((\mathbb{R}, \mathcal{B})\) that is defined by

\[
P^X(B) := P(X \in B) = P(X^{-1}(B)), \quad B \in \mathcal{B}.
\]
A random vector in $\mathbb{R}^m$ is a vector of random variables $X = (X_1, \ldots, X_m)$ that are all defined on the same probability space. The distribution of this random vector is defined as above but considering the set $B \in \mathcal{B}_m$. It is sometimes referred to as the joint distribution of the random vector $X$. For Cartesian products, $B = B_1 \times \cdots \times B_m$ we also write $P(X_1 \in B_1, \ldots, X_m \in B_m)$ to denote $P(X \in B)$.

All random variables appearing in this book will be of one of two types. Either the random variable takes values in a finite or countably infinite set, in which case it is called a discrete random variable, or its distribution has a density function, in which case we will call the random variable continuous. Note that in more advanced treatments of probability a continuous random variable would be referred to as absolutely continuous. For a random vector, we will use the terminology discrete and continuous if the component random variables are all discrete or all continuous, respectively.

Example 2.2.2 (Binomial distribution). Suppose a coin has chance $p \in (0, 1)$ to show heads, and that this coin is tossed $n$ times. It is reasonable to assume that the outcomes of the $n$ tosses define independent events. The probability of seeing a particular sequence $\omega = (\omega_1, \ldots, \omega_n)$ in the tosses is equal to $p^k(1-p)^{n-k}$ where $k$ is the number of heads among the $\omega_i$. If we define $X$ to be the discrete random variable

$$X : \omega \mapsto \sum_{i=1}^n 1_{\{\omega_i = \text{H}\}}$$

that counts the number of heads, then the distribution $P^X$ corresponds to the probability vector with components

$$(2.2.1) \quad p_k = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \ldots, n.$$ 

The distribution is known as the Binomial distribution $B(n,p)$.

Definition 2.2.3. For a random vector $X$ in $\mathbb{R}^m$, we define the cumulative distribution function (cdf) to be the function

$$F_X : \mathbb{R}^m \to [0, 1],$$

$$x \mapsto P(X \leq x) = P(X_1 \leq x_1, \ldots, X_m \leq x_m).$$

The fact that the Borel $\sigma$-algebra $\mathcal{B}_m$ is the smallest $\sigma$-algebra containing the rectangles $(-\infty, b_1) \times \cdots \times (-\infty, b_m)$ can be used to show that the cdf uniquely determines the probability distribution of a random vector. A cdf exhibits monotonicity and continuity properties that follow from (2.1.1) and (2.1.2). In particular, the cdf $F$ of a random variable is non-decreasing, right-continous function with $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$. For
2.2. Random Variables and their Distributions

continuous random vectors, the fundamental theorem of calculus yields the following relationship.

Lemma 2.2.4. If the random vector \( X = (X_1,\ldots,X_m) \) has density \( f \) that is continuous at \( x \in \mathbb{R}^m \), then

\[
f(x) = \frac{\partial^m}{\partial x_1 \cdots \partial x_m} F_X(x).
\]

Two natural probabilistic operations can be applied to a random vector \( X = (X_1,\ldots,X_m) \). On one hand, we can consider the distribution of a subvector \( X_I = (X_i \mid i \in I) \) for an index set \( I \subseteq [m] \). This distribution is referred to as the marginal distribution of \( X \) over \( I \). If \( X \) has density \( f \), then the marginal distribution over \( I \) has the marginal density

\[(2.2.2) \quad f_{X_I}(x_I) = \int_{\mathbb{R}^J} f(x_I, x_J) \, dx_J,
\]

where \( J = [m] \setminus I \). On the other hand, given disjoint index sets \( I, J \subseteq [m] \) and a vector \( x_J \in \mathbb{R}^J \), one can form the conditional distribution of \( X_I \) given \( X_J = x_J \). This distribution assigns the conditional probabilities \( P(X_I \in B \mid X_J = x_J) \) to the measurable sets \( B \in \mathcal{B}_I \). Clearly this definition of a conditional distribution is meaningful only if \( P(X_J = x_J) > 0 \). This seems to prevent us from defining conditional distributions for continuous random vectors as then \( P(X_J = x_J) = 0 \) for all \( x_J \). However, measure theory provides a way out that applies to arbitrary random vectors; see e.g. [Bil95, Chap. 6]. For random vectors with densities, this leads to intuitive ratios of densities.

Definition 2.2.5. Let the random vector \( X = (X_1,\ldots,X_m) \) have density \( f \) and consider two index sets \( I, J \subseteq [m] \) that partition \([m]\). For a vector \( x_J \in \mathbb{R}^J \), the conditional distribution of \( X_I \) given \( X_J = x_J \) is the probability distribution on \((\mathbb{R}^I, \mathcal{B}_I)\) that has density

\[
f(x_I \mid x_J) = \begin{cases} \frac{f(x_I, x_J)}{f_{X_J}(x_J)} & \text{if } f_{X_J}(x_J) > 0, \\ 0 & \text{otherwise.} \end{cases}
\]

The function \( f(x_I \mid x_J) \) is indeed a density as it is non-negative by definition and is easily shown to integrate to one by using (2.2.2).

When modeling random experiments using the language of random variables, an explicit reference to an underlying probability space can often be avoided by combining particular distributional assumptions with assumptions of independence.

Definition 2.2.6. Let \( X_1,\ldots,X_n \) be random vectors on the probability space \((\Omega, \mathcal{F}, P)\). Then \( X_1,\ldots,X_n \) are mutually independent or completely
independent if

\[
P(X_1 \in B_1, \ldots, X_n \in B_n) = \prod_{i=1}^{n} P(X \in B_i)
\]

for all measurable sets \(B_1, \ldots, B_n\) in the appropriate Euclidean spaces.

Note that we can see the definition of mutual independence of events given in Definition 2.1.13, can be seen as a special case of Definition 2.2.6. Indeed, to a set of events \(B_1, \ldots, B_n\) we can defined the indicator random variables \(X_1, \ldots, X_n\) such that \(X_i^{-1}(1) = B_i\) and \(X_i^{-1}(0) = B_i^c\). Then the mutual independence of the events \(B_1, \ldots, B_n\) is equivalent to the mutual independence of the random variables \(X_1, \ldots, X_n\).

**Example 2.2.7** (Sums of Bernoulli random variables). Suppose we toss the coin from Example 2.2.2 and record a one if heads shows and a zero otherwise. This yields a random variable \(X_1\) with \(P(X_1 = 1) = p\) and \(P(X_1 = 0) = 1 - p\). Such a random variable is said to have a Bernoulli\((p)\) distribution. If we now toss the coin \(n\) times and these tosses are independent, then we obtain the random variables \(X_1, \ldots, X_n\) that are independent and identically distributed (i.i.d.) according to Bernoulli\((p)\). We see that

\[
Y = \sum_{i=1}^{n} X_i
\]

has a Binomial distribution, in symbols, \(Y \sim B(n, p)\). More generally, if \(X \sim B(n, p)\) and \(Y \sim B(m, p)\) are independent, then \(X + Y \sim B(n + m, p)\). This fact is also readily derived from the form of the probability vector of the Binomial distribution. By Lemma 2.1.9,

\[
P(X + Y = k) = \sum_{j=0}^{m} P(X + Y = k \mid Y = j)P(Y = j)
\]

\[
= \sum_{j=0}^{m} P(X = k - j \mid Y = j)P(Y = j).
\]

Under the assumed independence \(P(X = k - j \mid Y = j) = P(X = k - j)\) and we find that

\[
P(X + Y = k) = \sum_{j=0}^{m} \binom{n}{k-j} p^{k-j}(1-p)^{n-k+j} \binom{m}{j} p^j (1-p)^{m-j}
\]

\[
= p^k (1-p)^{n+m-k} \sum_{j=0}^{m} \binom{n}{k-j} \binom{m}{j}.
\]
A simple combinatorial argument can be used to prove Vandermonde’s identity:

$$\sum_{j=0}^{m} \binom{n}{k-j} \binom{m}{j} = \binom{n+m}{k}$$

which shows that $X + Y \sim B(n + m, p)$.

The following result gives a simplified criterion for checking independence of random vectors. It can be established as a consequence of the Borel $\sigma$-algebra being defined in terms of rectangular sets.

**Lemma 2.2.8.** The random vectors $X_1, \ldots, X_n$ on $(\Omega, \mathcal{A}, P)$ are independent if and only if

$$F_{(X_1, \ldots, X_n)}(x) = \prod_{i=1}^{n} F_{X_i}(x_i)$$

for all vectors $x = (x_1, \ldots, x_n)$ in the appropriate Euclidean space.

If densities are available then it is often easier to recognize independence in a density factorization.

**Lemma 2.2.9.** Suppose the random vectors $X_1, \ldots, X_n$ on $(\Omega, \mathcal{A}, P)$ have density functions $f_1, \ldots, f_n$. If $X_1, \ldots, X_n$ are independent, then the random vector $(X_1, \ldots, X_n)$ has the density function

$$(2.2.3) \quad f_{X_1, \ldots, X_n}(x) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

If the random vector $(X_1, \ldots, X_n)$ has the density function $f_{X_1, \ldots, X_n}$, then $X_1, \ldots, X_n$ are independent if and only if the equality (2.2.3) holds almost surely.

**Proof.** The first claims follows by induction and the observation that

$$P(X_1 \in B_1)P(X_2 \in B_2) = \int_{B_1} \int_{B_2} f(x_1)f(x_2) \, dx_2 \, dx_1.$$

The second claim can be verified using this same observation for one direction and Lemma 2.2.8 in combination with the relation between cdf and density in Lemma 2.2.4 for the other direction.

As Example 2.2.7 shows, the changes in distribution when performing arithmetic operations with discrete random variables are not too hard to track. For a random vector with densities we can employ the change of variables theorem from analysis.
Theorem 2.2.10. Suppose the random vector $X = (X_1, \ldots, X_k)$ has density $f_X$ that vanishes outside the set $\mathcal{X}$, which may be equal to all of $\mathbb{R}^k$. Suppose further that $g : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}^k$ is a differentiable bijection whose inverse map $g^{-1}$ is also differentiable. Then the random vector $Y = g(X)$ has the density function

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) |Dg^{-1}(y)| & \text{if } y \in \mathcal{Y}, \\ 0 & \text{otherwise.} \end{cases}$$

Here, $Dg^{-1}(y)$ is the Jacobian matrix of $g^{-1}$ at $y$.

Remark 2.2.11. The differentiability, in fact the continuity, of $g$ in Theorem 2.2.10 is sufficient for the vector $Y = g(X)$ to be a random vector according to Definition 2.2.1 [Bil95, Thm. 13.2].

2.3. Expectation, Variance, and Covariance

When presented with a random variable it is natural to ask for numerical features of its distribution. These provide useful tools to summarize a random variable, or to compare two random variables to each other. The expectation, or in different terminology the expected value or mean, is a numerical description of the center of the distribution. The expectation can also be used to define the higher moments of a random variable.

Definition 2.3.1. If $X$ is a discrete random variable taking values in $\mathcal{X} \subseteq \mathbb{R}$ and $\sum_{x \in \mathcal{X}} |x|P(X = x) < \infty$, then the expectation of $X$ is defined as the weighted average

$$E[X] = \sum_{x \in \mathcal{X}} xP(X = x).$$

If $\sum_{x \in \mathcal{X}} |x|P(X = x)$ diverges, then we say that the expectation of $X$ does not exist. If $X$ is a continuous random variable with density $f$, then the expectation exists if $\int_{\mathbb{R}} |x|f(x) \, dx < \infty$, in which case it is defined as the integral

$$E[X] = \int_{\mathbb{R}} xf(x) \, dx.$$

For a random vector, we define the expectation as the vector of expectations of its components, provided they all exist.

Example 2.3.2 (Binomial expectation). If $X \sim B(n, p)$, then the expected value exists because $\mathcal{X} = \{0, 1, \ldots, n\}$ is finite. It is equal to

$$E[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} = np.$$
2.3. Expectation, Variance, and Covariance

The following result allows us to compute the expectation of functions of random variables without having to work out the distribution of the function of the random variable. It is easily proven for the discrete case and is a consequence of Theorem 2.2.10 in special cases of the continuous case. For the most general version of this result see [Bil95, pp. 274, 277].

**Theorem 2.3.3.** Suppose $X$ is a discrete random vector taking values in $\mathcal{X}$. If the function $g : \mathcal{X} \to \mathbb{R}$ defines a random variable $g(X)$, then the expectation of $g(X)$ is equal to

$$E[g(X)] = \sum_{x \in \mathcal{X}} g(x) P(X = x)$$

if it exists. It exists if the sum obtained by replacing $g(x)$ by $|g(x)|$ is finite.

Suppose $X$ is a continuous random vector with density $f$ that vanishes outside the set $\mathcal{X}$. If the function $g : \mathcal{X} \to \mathbb{R}$ defines a random variable $g(X)$, then the expectation of $g(X)$ is equal to

$$E[g(X)] = \int_{\mathcal{X}} g(x) f(x) \, dx$$

if it exists, which it does if the integral replacing $g(x)$ by $|g(x)|$ is finite.

**Corollary 2.3.4.** Suppose that the expectation of the random vector $X = (X_1, \ldots, X_k)$ exists. If $a_0 \in \mathbb{R}$ and $a \in \mathbb{R}^k$ are constants, then the expectation of an affine function $Y = a_0 + a^t X$ exists and is equal to

$$E[Y] = a_0 + a^t E[X] = a_0 + \sum_{i=1}^{k} a_i E[X_i].$$

**Proof.** The existence of $E[Y]$ follows from the triangle inequality and the form of $E[Y]$ follows from Theorem 2.3.3 and the fact that sums and integrals are linear operators. $\square$

**Example 2.3.5** (Binomial expectation revisited). If $X \sim B(n, p)$, we showed in Example 2.3.2 that $E[X] = np$. A more straightforward argument based on Corollary 2.3.4 is as follows. Let $Y_1, \ldots, Y_n$ be i.i.d. Bernoulli random variables with probability $p$. Then $X = Y_1 + \cdots + Y_n$. The expectation of a Bernoulli random variable is calculated as

$$E[Y_i] = 0 \cdot (1 - p) + 1 \cdot p = p.$$

Hence, $E[X] = E[Y_1] + \cdots + E[Y_n] = np$.

In the case of discrete random vectors the following corollary is readily obtained from Definition 2.2.6 and Theorem 2.3.3. In the case of densities it follows using Lemma 2.2.9.
Corollary 2.3.6. If the random variables \(X_1, \ldots, X_k\) are independent with finite expectations, then the expectation of their product exists and is equal to
\[
E[X_1X_2 \cdots X_k] = \prod_{i=1}^{k} E[X_i].
\]

Another natural measure for the distribution of a random variable is the spread around the expectation, which is quantified in the variance.

Definition 2.3.7. Suppose \(X\) is a random variable with finite expectation. The variance \(\text{Var}[X] = E[(X - E[X])^2]\) is the expectation of the squared deviation from \(E[X]\), if the expectation exists.

The variance can be computed using Theorem 2.3.3, but often it is easier to compute the variance as
\[
\]
Note that we employed Corollary 2.3.4 for the second equality and that \(E[X]\) is a constant. Using Corollary 2.3.4 we can also easily show that
\[
\text{Var}[a_1X + a_0] = a_1^2 \text{Var}[X].
\]

If a unit is attached to a random variable, then the variance is measured in the square of that unit. Hence, one often quotes the standard deviation \(\sqrt{\text{Var}[X]}\), which is measured in the original unit.

We next derive a simple but useful bound on the probability that a random variable \(X\) deviates from its expectation by a certain amount. This bound is informative if the variance of \(X\) is finite. One consequence of this bound is that if \(\text{Var}[X] = 0\), then \(X\) is equal to the constant \(E[X]\) with probability one.

Lemma 2.3.8 (Chebyshev’s inequality). Let \(X\) be a random variable with finite expectation \(E[X]\) and finite variance \(\text{Var}[X] < \infty\). Then
\[
P(\{|X - E[X]| \geq t\}) \leq \frac{\text{Var}[X]}{t^2} \quad \text{for all } t > 0.
\]

Proof. We will assume that \(X\) is continuous with density \(f\); the case of discrete \(X\) is analogous. Define \(Y = g(X) = (X - E[X])^2\) such that \(E[Y] = \text{Var}[X]\) and \(P(Y \geq 0) = 1\). Let \(f_Y\) be the density function of \(Y\). The fact that \(Y\) has a density follows from two separate applications of the substitution rule from analysis (compare Theorem 2.2.10). The two separate applications consider \(g\) over \((E[X], \infty)\) and \((-\infty, E[X])\).

Now,
\[
E[Y] = \int_{-\infty}^{\infty} y f_Y(y) \, dy \geq \int_{t^2}^{\infty} t^2 f_Y(y) \, dy = t^2 P(Y \geq t^2).
\]

Thus,
\[
P(Y \geq t^2) \leq \frac{E[Y]}{t^2} \leq \frac{\text{Var}[X]}{t^2} \quad \text{for all } t > 0.
\]

\[
P(\{|X - E[X]| \geq t\}) \leq \frac{\text{Var}[X]}{t^2} \quad \text{for all } t > 0.
\]
2.3. Expectation, Variance, and Covariance

It follows that
\[ P(|Y| \geq t) \leq \frac{E[Y]}{t^2}, \]
which is the claim if all quantities are re-expressed in terms of \( X \). \( \square \)

We now turn to a measure of dependence between two random variables.

**Definition 2.3.9.** Suppose \( X \) and \( Y \) are two random variables whose expectations exist. The **covariance** of \( X \) and \( Y \) is defined as
\[ \text{Cov}[X, Y] = E[(X - E[X])(Y - E[Y])]. \]

In analogy with (2.3.1) it holds that
\[ (2.3.3) \text{Cov}[X, Y] = E[XY] - E[X]E[Y]. \]

It follows from Corollary 2.3.6 that if \( X \) and \( Y \) are independent then they have zero covariance. The reverse implication is not true: one can construct random variables with zero covariance that are dependent (see Exercise 2.9).

By Corollary 2.3.4, if \( X \) and \( Y \) are two random vectors in \( \mathbb{R}^k \) and \( \mathbb{R}^m \), respectively, then
\[ (2.3.4) \text{Cov} \left[ \sum_{i=1}^{k} a_i X_i, \sum_{j=1}^{m} b_j Y_j \right] = \sum_{i=1}^{k} \sum_{j=1}^{m} a_i b_j \text{Cov}[X_i, Y_j]. \]

Since \( \text{Var}[X] = \text{Cov}[X, X] \) it follows that
\[ (2.3.5) \text{Var} \left[ \sum_{i=1}^{k} a_i X_i \right] = \sum_{i=1}^{k} a_i^2 \text{Var}[X_i] + 2 \sum_{i=1}^{k} \sum_{j=i+1}^{k} a_i a_j \text{Cov}[X_i, X_j]. \]

If \( X_1, \ldots, X_k \) are independent, then
\[ (2.3.6) \text{Var} \left[ \sum_{i=1}^{k} a_i X_i \right] = \sum_{i=1}^{k} a_i^2 \text{Var}[X_i]. \]

**Example 2.3.10 (Binomial variance).** Suppose \( X \sim B(n, p) \). Then we can compute \( \text{Var}[X] \) in a similar fashion as we computed \( E[X] \) in Example 2.3.2. However, it is easier to exploit that \( X \) has the same distribution as the sum of independent Bernoulli\((p)\) random variables \( X_1, \ldots, X_n \); recall Example 2.2.7. The expectation of \( X_1 \) is equal to \( E[X_1] = 0 \cdot (1 - p) + 1 \cdot p = p \) and thus the variance of \( X_1 \) is equal to
\[ E[X_1^2] = E[X_1]^2 = E[X_1] - E[X_1]^2 = p(1 - p). \]

By (2.3.6), we obtain that \( \text{Var}[X] = np(1 - p) \).

**Definition 2.3.11.** The **covariance matrix** \( \text{Var}[X] \) of a random vector \( X = (X_1, \ldots, X_k) \) is the \( k \times k \)-matrix with entries \( \text{Cov}[X_i, X_j] \).
Using the covariance matrix, we can rewrite (2.3.5) as

\[
\text{Var}[a_0 + a'tX] = a^T \text{Var}[X]a,
\]

which reveals more clearly the generalization of (2.3.2). One consequence is the following useful fact.

**Proposition 2.3.12.** For any random vector \(X\) for which the covariance matrix \(\text{Var}[X]\) is defined, the matrix \(\text{Var}[X]\) is positive semidefinite.

**Proof.** Recall that a symmetric matrix \(M \in \mathbb{R}^{m \times m}\) is positive semidefinite if for all \(a \in \mathbb{R}^m, a^T Ma \geq 0\). Since \(\text{Var}[a_0 + a'tX] \geq 0\) for any choice of \(a \in \mathbb{R}^k\), using (2.3.7) we see that \(\text{Var}[X]\) is positive semidefinite. \(\square\)

An inconvenient feature of covariance is that it is affected by changes of scales and thus its numerical value is not directly interpretable. This dependence on scales is removed in the measure of correlation.

**Definition 2.3.13.** The **correlation** between two random variables \(X\) and \(Y\) where all the expectations \(E[X], E[Y]\), variances \(\text{Var}[X], \text{Var}[Y]\), and covariance \(\text{Cov}[X, Y]\) exist is a standardized covariance defined as

\[
\text{Corr}[X, Y] = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X] \text{Var}[Y]}}.
\]

A correlation is, indeed, unaffected by (linear) changes of scale in that

\[
\text{Corr}[a_0 + a_1X, b_0 + b_1Y] = \text{Corr}[X, Y].
\]

As shown next, a correlation is never larger than one in absolute value and the extremes indicate a perfect linear relationship.

**Proposition 2.3.14.** Let \(\rho = \text{Corr}[X, Y]\). Then \(\rho \in [-1, 1]\). If \(|\rho| = 1\) then there exist \(a \in \mathbb{R}\) and \(b \neq 0\) with \(\text{sign}(b) = \text{sign}(\rho)\) such that \(P(Y = a + bX) = 1\).

**Proof.** Let \(\sigma_X = \sqrt{\text{Var}[X]}\) and \(\sigma_Y = \sqrt{\text{Var}[Y]}\). The claims follow from the inequality

\[
0 \leq \text{Var}\left[\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right] = 2(1 \pm \rho)
\]

and the fact that a random variable with zero variance is constant almost surely. \(\square\)

**Remark 2.3.15.** Proposition 2.3.14 shows that

\[
\text{Cov}[X_1, X_2]^2 \leq \text{Var}[X_1] \text{Var}[X_2],
\]

which implies that the covariance of two random variables exists if both their variances are finite.
2.3. Expectation, Variance, and Covariance

Considering conditional distributions can be helpful when computing expectations and (co-)variances. We will illustrate this in Example 2.3.18, but it also plays an important role in hidden variable models. If a random vector is partitioned into subvectors $X$ and $Y$ then we can compute the expectation of $X$ if $X$ is assumed to be distributed according to its conditional distribution given $Y = y$. This expectation, in general a vector, is referred to as the conditional expectation denoted $E[X \mid Y = y]$. Its components $E[X_i \mid Y = y]$ can be computed by passing to the marginal distribution of the vector $(X_i, Y)$ and performing the following calculation.

**Lemma 2.3.16.** Suppose a random vector is partitioned as $(X, Y)$ with $X$ being a random variable. Then the conditional expectation of $X$ given $Y = y$ is equal to

$$E[X \mid Y = y] = \begin{cases} \sum_x x P(X = x \mid Y = y) & \text{if } (X, Y) \text{ is discrete}, \\ \int_x x f(x \mid y) \, dx & \text{if } (X, Y) \text{ has density } f. \end{cases}$$

Similarly, we can compute conditional variances $\text{Var}[X \mid Y = y]$ and covariances $\text{Cov}[X_1, X_2 \mid Y = y]$. The conditional expectation is a function of $y$. Plugging in the random vector $Y$ for the value $y$ we obtain a random vector $E[X \mid Y]$.

**Lemma 2.3.17.** Suppose a random vector is partitioned as $(X, Y)$ where $X$ and $Y$ may be arbitrary subvectors. Then the expectation of $X$ is equal to

$$E[X] = E[E[X \mid Y]],$$

and the covariance matrix of $X$ can be computed as

$$\text{Var}[X] = \text{Var}[E[X \mid Y]] + E[\text{Var}[X \mid Y]].$$

**Proof.** Suppose that $(X, Y)$ has density $f$ on $\mathbb{R}^{k+1}$ and that $X$ is a random variable. (If $(X, Y)$ is discrete, then the proof is analogous.) Let $f_Y$ be the marginal density of $Y$; compare (2.2.2). Then $f(x, y) = f(x \mid y)f_Y(y)$ such that we obtain

$$E[E[X \mid Y]] = \int_{\mathbb{R}^k} \left( \int_{\mathbb{R}} x f(x \mid y) \, dx \right) f_Y(y) \, dy$$

$$= \int_{\mathbb{R}^{k+1}} x f(x, y) \, d(x, y)$$

$$= E[X].$$

Note that the last equality follows from Theorem 2.3.3. The vector-version of the equality $E[X] = E[E[X \mid Y]]$ follows by component-wise application of the result we just derived.
The claim about the covariance matrix follows if we can show that for a random vector \((X_1, X_2, Y)\) with random variable components \(X_1\) and \(X_2\) it holds that
\[
(2.3.8) \quad \text{Cov}[X_1, X_2] = \text{Cov} \left[ \mathbb{E}[X_1 \mid Y], \mathbb{E}[X_2 \mid Y] \right] + \mathbb{E} \left[ \text{Cov}[X_1, X_2 \mid Y] \right].
\]
For the first term in (2.3.8) we obtain by (2.3.3) and the result about the expectation that
\[
\text{Cov} \left[ \mathbb{E}[X_1 \mid Y], \mathbb{E}[X_2 \mid Y] \right] = \mathbb{E} \left[ \mathbb{E}[X_1 \mid Y] \cdot \mathbb{E}[X_2 \mid Y] \right] - \mathbb{E}[X_1] \mathbb{E}[X_2].
\]
Similarly, the second term in (2.3.8) equals
\[
\mathbb{E} \left[ \text{Cov}[X_1, X_2 \mid Y] \right] = \mathbb{E}[X_1 X_2] - \mathbb{E} \left[ \mathbb{E}[X_1 \mid Y] \cdot \mathbb{E}[X_2 \mid Y] \right].
\]
Hence, adding the two terms gives \(\text{Cov}[X_1, X_2]\) as claimed.

**Example 2.3.18** (Exam Questions). Suppose that a student player in a course on algebraic statistics is selected at random and takes an exam with two questions. Let \(X_1\) and \(X_2\) be the two Bernoulli random variables that indicate whether the students gets problem 1 and problem 2 correct. We would like to compute the correlation \(\text{Corr}[X_1, X_2]\) under the assumption that the random variable \(Y\) that denotes the probability of a randomly selected student getting a problem correct has density \(f(y) = 6y(1-y)^1_{[0,1]}(y)\).

Clearly, \(\text{Corr}[X_1, X_2]\) will not be zero as the two problems are attempted by the same student. However, it is reasonable to assume that conditional on the student’s skill level \(Y = y\), success in the first and success in the second problem are independent. This implies that \(\text{Cov}[X_1, X_2 \mid Y = y] = 0\) and thus the second term in (2.3.8) is zero. Since the expectation of a Bernoulli\((y)\) random variable is equal to \(y\) we obtain that \(\mathbb{E}[X_1 \mid Y] = Y\), which implies that the first term in (2.3.8) equals \(\text{Var}[Y]\). We compute
\[
\mathbb{E}[Y] = \int_0^1 6y^2(1-y) \, dy = 1/2
\]
and
\[
\mathbb{E}[Y^2] = \int_0^1 6y^3(1-y) \, dy = 3/10
\]
and obtain by (2.3.1) that
\[
\text{Cov}[X_1, X_2] = \text{Var}[Y] = 3/10 - (1/2)^2 = 1/20.
\]
Similarly, using the variance formula in Lemma 2.3.17 and the fact \(\text{Var}[X_i \mid Y] = Y(1 - Y)\), we find that
\[
\text{Var}[X_i] = \text{Var}[Y] + \mathbb{E}[Y(1 - Y)] = 1/20 + 1/2 - 3/10 = 1/4.
\]
The correlation thus equals
\[
\text{Corr}[X_1, X_2] = \frac{1/20}{1/4} = 1/5.
\]
We remark that the set-up of Example 2.3.18 does not quite fit in our previous discussion in the sense that we discuss a random vector \((X_1, X_2, Y)\) that has two discrete and one continuous component. However, all our results carry over to this mixed case if multiple integrals (or multiple sums) are replaced by a mix of summations and integrations.

Besides the expectation and variance of a random vector, it can also be worthwhile to look at higher moments of random variables. These are defined as following.

**Definition 2.3.19.** Let \(X \in \mathbb{R}^m\) be a random vector and \(u = (u_1, \ldots, u_m) \in \mathbb{N}^m\). The \(u\)-moment of \(X\) is the expected value
\[
\mu_u = \mathbb{E}[X^u] = \mathbb{E}[X_1^{u_1} X_2^{u_2} \cdots X_m^{u_m}]
\]
if this expected value exists. The order of the moment \(\mu_u\) is \(\|u\|_1 = u_1 + \cdots + u_m\).

Note that, from the standpoint of the information content, computing the expectation and variance of a random vector \(X\) is equivalent to computing all the first order and second order moments of \(X\). Higher order moments can be used to characterize random vectors, and can by conveniently packaged into the moment generating function \(M_X(t) = \mathbb{E}[\exp(tX)]\).

### 2.4. Multivariate Normal Distribution

In this section we introduce a multivariate generalization of the normal distribution \(\mathcal{N}(\mu, \sigma^2)\) from Example 2.1.7, where the parameters \(\mu\) and \(\sigma\) were called mean and standard deviation. Indeed it is an easy exercise to show that an \(\mathcal{N}(\mu, \sigma^2)\) random variable \(X\) has expectation \(\mathbb{E}[X] = \mu\) and variance \(\text{Var}[X] = \sigma^2\).

To begin, let us recall that the (univariate) standard normal distribution \(\mathcal{N}(0, 1)\) has as density function the Gaussian bell curve
\[
\phi(x) = \phi_{0,1}(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\}.
\]

Let \(I \in \mathbb{R}^{m \times m}\) be the \(m \times m\) identity matrix. Then we say that a random vector \(X = (X_1, \ldots, X_m)\) is distributed according to the *standard multivariate normal distribution* \(\mathcal{N}_m(0, I)\) if its components \(X_1, \ldots, X_k\) are independent and identically distributed (i.i.d.) as \(X_i \sim \mathcal{N}(0, 1)\). Thus, by Lemma 2.2.9, the distribution \(\mathcal{N}_m(0, I)\) has the density function
\[
\phi_{0,I}(x) = \prod_{i=1}^m \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x_i^2\right\} = \frac{1}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}x^T x\right\}.
\]
Let $\Lambda \in \mathbb{R}^{m \times m}$ be a matrix of full rank and let $\mu \in \mathbb{R}^m$. Suppose $X \sim N_m(0, I)$ and define $Y = \Lambda X + \mu$. By Theorem 2.2.10, the random vector $Y$ has the density function

$$\phi_Y(y) = \frac{1}{(2\pi)^{m/2}} \exp \left\{ -\frac{1}{2} (y - \mu)^T \Lambda^{-T} \Lambda^{-1} (y - \mu) \right\} |\Lambda^{-1}|,$$

where $|.|$ denotes the determinant, and $\Lambda^{-T}$ denotes the transpose of the inverse. Define $\Sigma = \Lambda \Lambda^T$. Then the density $\phi_Y$ is seen to depend only on $\mu$ and $\Sigma$ and can be rewritten as

$$\phi_{\mu, \Sigma}(y) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (y - \mu)^T \Sigma^{-1} (y - \mu) \right\}, \tag{2.4.3}$$

Note that since standard multivariate normal random vector has expectation $0$ and covariance matrix $I$, it follows from Corollary 2.3.4 and (2.3.4) that $Y$ has expectation $\Lambda 0 + \mu = \mu$ and covariance matrix $\Lambda I \Lambda^T = \Sigma$.

**Definition 2.4.1.** Suppose $\mu$ is a vector in $\mathbb{R}^m$ and $\Sigma$ is a symmetric positive definite $m \times m$-matrix. Then a random vector $X = (X_1, \ldots, X_m)$ is said to be distributed according to the multivariate normal distribution $N_m(\mu, \Sigma)$ with mean vector $\mu$ and covariance matrix $\Sigma$ if it has the density function in (2.4.3).

We also write $N(\mu, \Sigma)$ when the dimension is clear. The random vector $X$ is also called a Gaussian random vector. One of the many amazing properties of the multivariate normal distribution is that its marginal distributions as well as its conditional distributions are again multivariate normal.

**Theorem 2.4.2.** Let $X \sim N_m(\mu, \Sigma)$ and let $A, B \subseteq [m]$ be two disjoint index sets.

(i) The marginal distribution of $X$ over $B$, that is, the distribution of the subvector $X_B = (X_i \mid i \in B)$ is multivariate normal, namely,

$$X_B \sim N_{\#B}(\mu_B, \Sigma_{B,B}),$$

where $\mu_B$ is the respective subvector of the mean vector $\mu$ and $\Sigma_{B,B}$ is the corresponding principal submatrix of $\Sigma$.

(ii) For every $x_B \in \mathbb{R}^B$, the conditional distribution of $X_A$ given $X_B = x_B$ is the multivariate normal distribution

$$N_{\#A} \left( \mu_A + \Sigma_{A,B} \Sigma_{B,B}^{-1} (x_B - \mu_B), \Sigma_A - \Sigma_{A,B} \Sigma_{B,B}^{-1} \Sigma_{B,A} \right).$$

The matrix $\Sigma_{A,B} \Sigma_{B,B}^{-1}$ appearing in Theorem 2.4.2 is the matrix of regression coefficients. The positive definite matrix

$$\Sigma_{A,A} = \Sigma_A - \Sigma_{A,B} \Sigma_{B,B}^{-1} \Sigma_{B,A} \tag{2.4.4}$$
is the conditional covariance matrix. We also note that \((X_A \mid X_B = x_B)\) is not well-defined as a random vector. It is merely a way of denoting the conditional distribution of \(X_A\) given \(X_B = x_B\).

**Proof of Theorem 2.4.2.** Suppose that \(A\) and \(B\) are disjoint and \(A \cup B = [m]\). Then we can partition the covariance matrix as

\[
\Sigma = \begin{pmatrix} \Sigma_{A,A} & \Sigma_{A,B} \\ \Sigma_{B,A} & \Sigma_{B,B} \end{pmatrix}.
\]

Using the same partitioning, \(\Sigma^{-1}\) can be factored as

\[
\Sigma^{-1} = \begin{pmatrix} I & 0 \\ -\Sigma^{-1}_{B,B}\Sigma_{B,A} & I \end{pmatrix} \left( \begin{pmatrix} \Sigma_{A,A}^{-1} & 0 \\ 0 & \Sigma_{B,B}^{-1} \end{pmatrix} \begin{pmatrix} I & -\Sigma_{A,B}\Sigma^{-1}_{B,B} \\ 0 & I \end{pmatrix} \right)
\]

Plugging this identity into (2.4.3) we obtain that

\[
\phi_{\mu, \Sigma}(x) = \phi_{\mu_A + \Sigma_{A,B}\Sigma^{-1}_{B,B}(x_B - \mu_B), \Sigma_{A,A}}(x_A) \phi_{\mu_B, \Sigma_{B,B}}(x_B).
\]

Integrating out \(x_A\), we see that \(X_B \sim \mathcal{N}_{\#_B}(\mu_B, \Sigma_{B,B})\), which concludes the proof of claim (i).

By Definition 2.2.5, it also follows that (ii) holds if, as we have assumed, \(A\) and \(B\) partition \([m]\). If this does not hold we can use (i) to pass the marginal distribution of \(X_{A:B}\) first. Hence, we have also shown (ii). \(\square\)

The multivariate normal distribution is the distribution of a non-singular affine transformation of standard normal random vector. Consequently, further affine transformations remain multivariate normal.

**Lemma 2.4.3 (Affine transformation).** If \(X \sim \mathcal{N}_k(\mu, \Sigma)\), \(A\) is a full rank \(m \times k\)-matrix with \(m \leq k\) and \(b \in \mathbb{R}^m\), then \(AX + b \sim \mathcal{N}_m(A\mu + b, A \Sigma A^T)\).

**Proof.** If \(m = k\) the claim follows from Theorem 2.2.10. If \(m < k\) then complete \(A\) to a full rank \(k \times k\)-matrix and \(b\) to a vector in \(\mathbb{R}^k\), which allows to deduce the result from Theorem 2.4.2(i). \(\square\)

In Section 2.3, we noted that uncorrelated random variables need not be independent in general. However, this is the case for random variables that form a multivariate normal random vector.

**Proposition 2.4.4.** If \(X \sim \mathcal{N}_k(\mu, \Sigma)\) and \(A, B \subseteq [m]\) are two disjoint index sets, then the subvectors \(X_A\) and \(X_B\) are independent if and only if \(\Sigma_{A,B} = 0\).

**Proof.** Combining Lemma 2.2.9 and Definition 2.2.5, we see that independence holds if and only if the conditional density of \(X_A\) given \(X_B = x_B\) does not depend on \(x_B\). By Theorem 2.4.2, the latter condition holds if and only if \(\Sigma_{A,B}\Sigma^{-1}_{B,B} = 0\) if and only if \(\Sigma_{A,B} = 0\) because \(\Sigma_{B,B}\) is of full rank. \(\square\)
An important distribution intimately related to the multivariate normal distribution is the chi-squared distribution.

**Definition 2.4.5.** Let \( Z_1, Z_2, \ldots, Z_k \) be i.i.d. \( \mathcal{N}(0,1) \) random variables. Then the random variable \( Q = \sum_{i=1}^{k} Z_i^2 \) has a chi-square distribution with \( k \) degrees of freedom. This is denoted \( Q \sim \chi^2_k \).

The chi-square distribution is rarely used to model natural phenomena. However, it is the natural distribution that arises in limiting phenomena in many hypothesis tests, since it can be interpreted as the distribution of squared distances in many contexts. A typical example is the following.

**Proposition 2.4.6.** Let \( X \in \mathbb{R}^m \sim \mathcal{N}(0,I) \), let \( L \) be a linear space of dimension \( d \) through the origin, and let \( Z = \min_{y \in L} \|y - X\|_2 \). Then \( Z \sim \chi^2_{m-d} \).

**Proof.** There is an orthogonal change of coordinates \( \Lambda \) so that \( L \) becomes a coordinate subspace. In particular, we can apply a rotation so that transforms \( L \) into the subspace

\[
L_d = \{(x_1, \ldots, x_m) \in \mathbb{R}^m : x_1 = \ldots = x_{m-d} = 0\}.
\]

The standard normal distribution is rotationally invariant so \( Z \) has the same distribution as

\[
Z' = \min_{y \in L_d} \|y - X\|_2^2 = \sum_{i=1}^{m-d} X_i^2
\]

so \( Z \sim \chi^2_{m-d} \). \( \square \)

## 2.5. Limit Theorems

According to Definition 2.2.1, a random vector is a map from a sample space \( \Omega \) to \( \mathbb{R}^k \). Hence, when presented with a sequence of random vectors \( X_1, X_2, \ldots \) all defined on the same probability space \( (\Omega, \mathfrak{F}, P) \) and all taking values in \( \mathbb{R}^k \), then we can consider their pointwise convergence to a limiting random vector \( X \). Such a limit \( X \) would need to satisfy that

\[
X(\omega) = \lim_{n \to \infty} X_n(\omega)
\]

for all \( \omega \in \Omega \). While pointwise convergence can be a useful concept in analysis it presents an unnaturally strong requirement for probabilistic purposes. For example, it is always possible to alter the map defining a continuous random vector at countably many elements in \( \Omega \) without changing the distribution of the random vector. A more appropriate concept is almost sure convergence, denoted by \( X_n \to_{a.s.} X \), in which (2.5.1) is required to hold for all \( \omega \) in a set with probability one.
Example 2.5.1 (Dyadic intervals). Suppose that $\Omega = (0, 1]$ and that $\mathcal{A}$ is the $\sigma$-algebra of measurable subsets of $(0, 1]$. Let $P$ be the uniform distribution on $\Omega$, that is, $P$ has density function $f(\omega) = 1_{(0,1]}(\omega)$. Define

$$X_{2^j+i}(\omega) = 1_{(i(1-1)/2^j,i/2^j]}(\omega), \quad j = 1, 2, \ldots, i = 1, \ldots, 2^j.$$ 

Then for all $\omega \in \Omega$ it holds that

$$\lim \inf_{n \to \infty} X_n(\omega) = 0 \neq 1 = \lim \sup_{n \to \infty} X_n(\omega)$$

such that the sequence $(X_n)_{n \geq 1}$ has no limit under almost sure convergence.

In Example 2.5.1, the probability $P(X_{2^j+i} \neq 0) = 2^{-j}$ converges to zero as $j$ grows large, from which it follows that the sequence $(X_n)_{n \geq 1}$ converges to the constant 0 in the weaker sense of the next definition.

Definition 2.5.2. Let $X$ be a random vector and $(X_n)_{n}$ a sequence of random vectors on $(\Omega, \mathcal{A}, P)$ that all take values in $\mathbb{R}^k$. We say that $X_n$ converges to $X$ in probability, in symbols $X_n \to_p X$ if for all $\epsilon > 0$ it holds that

$$\lim_{n \to \infty} P(\|X_n - X\|_2 > \epsilon) = 0.$$ 

Equipped with the appropriate notions of convergence of random variables, we can give a long-run average interpretation of expected values.

Theorem 2.5.3 (Weak law of large numbers). Let $(X_n)_{n}$ be a sequence of random variables with finite expectation $E[X_n] = \mu$ and finite variance $\text{Var}[X_n] = \sigma^2 < \infty$. Suppose further that the random variables are pairwise uncorrelated, that is, $\text{Cov}[X_n, X_m] = 0$ for $n \neq m$. Then, as $n \to \infty$,

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \to_p \mu.$$ 

Proof. By Corollary 2.3.4 and (2.3.6), $E[\bar{X}_n] = \mu$ and $\text{Var}[\bar{X}_n] = \sigma^2/n$. Thus by Lemma 2.3.8,

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0.$$  

Suppose $X_n$ is the Bernoulli($p$) indicator for the heads in the $n$-th toss of a coin. Then $\bar{X}_n$ is the proportion of heads in $n$ tosses. Assuming independence of the repeated tosses, Theorem 2.5.3 implies that the proportion converges to the probability $p$. This is the result promised in the beginning of this chapter.

The version of the law of large numbers we presented is amenable to a very simple proof. With more refined techniques, the assumptions made can be weakened. For example, it is not necessary to require finite variances.
Moreover, if independence instead of uncorrelatedness is assumed then it is also possible to strengthen the conclusion from convergence in probability to almost sure convergence in which case one speaks of the strong law of large numbers. We state without proof one version of the strong law of large numbers; compare Theorem 22.1 in [Bil95].

**Theorem 2.5.4** (Strong law of large numbers). *If the random variables in the sequence \((X_n)_n\) are independent and identically distributed with finite expectation \(E[X_n] = \mu\), then \(\bar{X}_n\) converges to \(\mu\) almost surely.*

**Example 2.5.5** (Cauchy distribution). An example to which the law of large numbers does not apply is the Cauchy distribution, which has density function

\[
f(x) = \frac{1}{\pi(1 + x^2)}, \quad x \in \mathbb{R}.
\]

If \(X_1, X_2, \ldots\) are i.i.d. Cauchy random variables, then Theorem 2.2.10 can be used to show that \(\bar{X}_n\) has again a Cauchy distribution and thus does not converge in probability or almost surely to a constant. The failure of the law of large numbers to apply is due to the fact that the expectation of a Cauchy random variable does not exist, i.e., \(E[|X_1|] = \infty\).

All convergence concepts discussed so far require that the sequence and its limiting random variable are defined on the same probability space. This is not necessary in the following definition.

**Definition 2.5.6.** Let \(X\) be a random vector and \((X_n)_n\) a sequence of random vectors that all take values in \(\mathbb{R}^k\). Denote their cdfs by \(F_X\) and \(F_{X_n}\). Then \(X_n\) converges to \(X\) in distribution, in symbols \(X_n \to_d X\) if for all points \(x \in \mathbb{R}^k\) at which \(F_X\) is continuous it holds that

\[
\lim_{n \to \infty} F_{X_n}(x) = F(x).
\]

Exercise 2.7 explains the necessity of excluding the points of discontinuity of \(F_X\) when defining convergence of distribution. What is shown there is that without excluding the points of discontinuity convergence in distribution may fail for random variables that converge to a limit in probability and almost surely. We summarize the relationship among the three convergence concepts we introduced.

**Theorem 2.5.7.** Almost sure convergence implies convergence in probability which in turn implies convergence in distribution.

**Proof.** We prove the implication from convergence in probability to convergence in distribution; for the other implication see Theorem 5.2 in [Bil95]. Suppose that \(X_n \to_p X\) and let \(\epsilon > 0\). Let \(x\) be point of continuity of the cdf \(F_X\), which implies that \(P(X = x) = 0\). If \(X_n \leq x\) and \(|X_n - x| \leq \epsilon\),
then $X \leq x + \epsilon$. Hence, $X > x + \epsilon$ implies that $X_n > x$ or $|X_n - X| > \epsilon$. Consequently,

$$1 - P(X \leq x + \epsilon) \leq 1 - P(X_n \leq x) + P(|X_n - X| > \epsilon)$$

from which we deduce that

$$P(X_n \leq x) \leq F_X(x + \epsilon) + P(|X_n - X| > \epsilon).$$

Starting the same line of reasoning with the fact that $X \leq x - \epsilon$ and $|X_n - X| \leq \epsilon$ imply that $X_n \leq x$ we find that

$$P(X_n \leq x) \geq F_X(x - \epsilon) - P(|X_n - X| > \epsilon).$$

Combining the two inequalities and letting $n$ tend to infinity we obtain that

$$F_X(x - \epsilon) \leq \liminf_{n \to \infty} P(X_n \leq x) \leq \limsup_{n \to \infty} P(X_n \leq x) \leq F_X(x + \epsilon).$$

Now letting $\epsilon$ tend to zero, the continuity of $F_X$ at $x$ yields

$$\lim_{n \to \infty} P(X_n \leq x) = F_X(x),$$

which is what we needed to show.

\[\square\]

**Remark 2.5.8.** In general none of the reverse implications in Theorem 2.5.7 hold. However, if $c$ is a constant and $X_n$ converges to $c$ in distribution, then one can show that $X_n$ also converges to $c$ in probability.

Convergence in distribution is clearly the relevant notion if one is concerned with limits of probabilities of events of interest. The next result is a powerful tool that allows to approximate difficult to evaluate probabilities by probabilities computed under a normal distribution.

**Theorem 2.5.9** (Central limit theorem). Let $(X_n)_n$ be a sequence of independent and identically distributed random vectors with existent expectation $E[X_n] = \mu \in \mathbb{R}^k$ and positive definite covariance matrix $\text{Var}[X_n] = \Sigma$. Then, as $n \to \infty$,

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} \mathcal{N}_k(0, \Sigma).$$

The literature offers two types of proofs for central limit theorems. The seemingly simpler approach employs so-called characteristic functions and a Taylor-expansion; compare Sections 27 and 29 in [Bil95]. However, this argument relies on the non-trivial fact that characteristic functions uniquely determine the probability distribution of a random vector or variable. The second approach known as Stein’s method is perhaps more elementary and is discussed for example in [Sho00]. Instead of attempting to reproduce a general proof of the central limit theorem here, we will simply illustrate its conclusion in the example that is at the origin of the modern result.
Corollary 2.5.10 (Laplace DeMoivre Theorem). Let \( X_1, X_2, \ldots \) be independent Bernoulli(\( p \)) random variables. Then

\[
S_n = \sum_{i=1}^{n} X_i
\]

follows the Binomial \( B(n, p) \) distribution. Let \( Z \sim N(0, p(1-p)) \). Then it holds for all \( a < b \) that

\[
\lim_{n \to \infty} P \left( a < \frac{S_n - np}{\sqrt{n}} \leq b \right) = P(a < Z \leq b).
\]

We conclude this review of probability by stating two useful results on transformations, which greatly extend the applicability of asymptotic approximations.

Lemma 2.5.11. Suppose the random vectors \( X_n \) converge to \( X \) either almost surely, in probability, or in distribution. If \( g \) is a continuous map that defines random vectors \( g(X_n) \) and \( g(X) \), then the sequence \( g(X_n) \) converges to \( g(X) \) almost surely, in probability, or in distribution, respectively.

The claim is trivial for almost sure convergence. For the proofs for convergence in probability and in distribution see Theorem 25.7 in [Bil95] and also Theorem 2.3 in [vdV98].

Lemma 2.5.12 (Delta method). Let \( X_n \) and \( X \) be random vectors in \( \mathbb{R}^k \) and \((a_n)_n\) a diverging sequence of real numbers with \( \lim_{n \to \infty} a_n = \infty \). Suppose there exists a vector \( b \in \mathbb{R}^k \) such that

\[
a_n(X_n - b) \longrightarrow_d X.
\]

Let \( g : \mathbb{R}^k \to \mathbb{R}^m \) be a map defining random vectors \( g(X_n) \) and \( g(X) \).

(i) If \( g \) is differentiable at \( b \) with total derivative matrix \( D = Dg(b) \in \mathbb{R}^{m \times k} \), then

\[
a_n\left(g(X_n) - g(b)\right) \longrightarrow_d DX.
\]

(ii) If \( g \) is a function, i.e., \( m = 1 \), that is twice differentiable at \( b \) with gradient \( Dg(b)^t \in \mathbb{R}^k \) equal to zero and Hessian matrix \( H = Hg(b) \in \mathbb{R}^{k \times k} \), then

\[
a_n\left(g(X_n) - g(b)\right) \longrightarrow_d \frac{1}{2} X^tHX.
\]

Proof sketch. A rigorous proof is given for Theorem 3.1 in [vdV98]. We sketch the idea of the proof. Apply a Taylor expansion to \( g \) at the point \( b \) and multiply by \( a_n \) to obtain that

\[
a_n\left(g(X_n) - g(b)\right) \approx D \cdot a_n(X_n - b).
\]
By Lemma 2.5.11, the right hand side converges to $DX$ in distribution, which is claim (i). For claim (ii) look at the order two term in the Taylor expansion of $g$. 

We remark that if the limiting random vector $X$ in Lemma 2.5.12 has a multivariate normal distribution $N_k(\mu, \Sigma)$ and the derivative matrix $D$ is of full rank then it follows from Lemma 2.4.3 that $DX \sim N_m(D\mu, D\Sigma D^t)$. If $D$ is not of full rank then $DX$ takes its values in the linear subspace given by the image of $D$. Such distributions are often called singular normal distributions [Rao73, Chap. 8a].

2.6. Exercises

Exercise 2.1. An urn contains 49 balls labelled by the integers 1 through 49. Six balls are drawn at random from this urn without replacement. The balls are drawn one by one and the numbers on the balls are recorded in the order of the draw. Define the appropriate sample space and show that the probability of exactly three balls having a number smaller than 10 is equal to (2.1.3).

Exercise 2.2. (i) Show that a $\sigma$-algebra contains the empty set and is closed under countable intersections. (ii) Show that the Borel $\sigma$-algebra $\mathcal{B}$ on $\mathbb{R}$ contains all open subsets of $\mathbb{R}$. (iii) Show that the Borel $\sigma$-algebra $\mathcal{B}_k$ on $\mathbb{R}^k$ is not the $k$-fold Cartesian product of $\mathcal{B}$.

Exercise 2.3. The exponential distribution is the probability measure $P$ on $(\mathbb{R}, \mathcal{B})$ defined by the density function $f(x) = e^{-x}1_{\{x > 0\}}$. For $t, s > 0$, show that $P((0, s]) = P((t, t + s] \mid (t, \infty))$.

Exercise 2.4. A fair coin showing heads and tails with equal probability is tossed twice. Show that the any two of the events Heads with the first toss, Heads with exactly one toss and Heads with the second toss are independent but that the three events are not mutually independent.

Exercise 2.5. Show that if $X \sim \mathcal{N}(0, 1)$, then the sign of $X$ and the absolute value of $X$ are independent.

Exercise 2.6. Find the density of $X + Y$ for $X$ and $Y$ being independent random variables with an exponential distribution (cf. Exercise 2.3). 

(Hint: One way of approaching this problem is to consider the map $g : (x, y) \mapsto (x + y, y)$ and employ Theorem 2.2.10.)

Exercise 2.7. Let the discrete random variable $X$ have a Poisson($\lambda$) distribution, i.e.,

$$P(X = k) = e^{-\lambda}\lambda^k/k!, \quad k = 0, 1, 2, \ldots$$

where the parameter $\lambda$ is a positive real. Show that $E[X] = \text{Var}[X] = \lambda$.
Exercise 2.8. Suppose the random vector \((X_1, X_2, X_3)\) follows the multivariate normal distribution \(N_3(0, \Sigma)\) with
\[
\Sigma = \begin{pmatrix} 3/2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 3/2 \end{pmatrix}.
\]
Verify that the conditional distribution of \(X_1\) given \((X_2, X_3) = (x_2, x_3)\) does not depend on \(x_3\). Compute the inverse \(\Sigma^{-1}\); do you see a connection?

Exercise 2.9. Let \(X\) and \(Y\) be discrete random variables each with state space \([3]\). Suppose that
\[
P(X = 1, Y = 2) = P(X = 2, Y = 1) = P(X = 2, Y = 2) = \frac{1}{5}
\]
\[
P(X = 2, Y = 3) = P(X = 3, Y = 2) = \frac{1}{5}
\]
\[
P(X = 1, Y = 1) = P(X = 1, Y = 3) = 0
\]
\[
P(X = 3, Y = 1) = P(X = 3, Y = 3) = 0.
\]
Show that \(\text{Cov}[X, Y] = 0\) but \(X\) and \(Y\) are dependent.

Exercise 2.10. Let \((X_n)\) be a sequence of independent discrete random variables that take the values \(\pm 1\) with probability 1/2 each. Let \(F_{\bar{X}_n}\) denote the cdf of \(\bar{X}_n\). By the weak law of large numbers it holds that \(\bar{X}_n\) converges to zero in probability. Let \(F_X(x) = 1_{[0,\infty)}(x)\) be the cdf of the constant zero. Show that \(F_{\bar{X}_n}(x)\) converges to \(F_X(x)\) if and only if \(x \neq 0\). What is the limit if \(x = 0\)?

Exercise 2.11. Let \((X_n)\) be a sequence of independent and identically distributed random variables that follow a uniform distribution on \((0, 1]\), which has density \(f(x) = 1_{(0,1]}(x)\). What does the central limit theorem say about the convergence of \(\bar{X}_n\)? Consider the parametrized family of functions \(g_t(x) = (x - t)^2\) with parameter \(t \in (0, 1]\). Use the delta method from Lemma 2.5.12 to find a convergence in distribution result for \(g_t(\bar{X}_n)\). Contrast the two cases \(t \neq 1/2\) and \(t = 1/2\).

Exercise 2.12. Suppose \((X_n)\) is a sequence of independent and identically distributed random variables that follow an exponential distribution with rate parameter \(\lambda = 1\). Find the cdf of \(M_n = \max\{X_1, \ldots, X_n\}\). Show that \(M_n - \log n\) converges in distribution and find the limiting distribution.
Chapter 3

Algebra Primer

The basic geometric objects that are studied in algebraic geometry are algebraic varieties. An algebraic variety is just the set of solutions to a system of polynomial equations. On the algebraic side, varieties will form one of the main objects of study in this book. In particular, we will be interested in ways to use algebraic and computational techniques to understand geometric properties of varieties. This chapter will be devoted to laying the groundwork for algebraic geometry. We introduce algebraic varieties, and the tools used to study them, in particular, their vanishing ideals, and Gröbner bases of these ideals. In Section 3.4 we will describe some applications of Gröbner bases to computing features of algebraic varieties. Our view is also that being able to compute key quantities in the area is important, and we illustrate the majority of concepts with computational examples using computer algebra software.

Our aim in this chapter is to provide a concise introduction to the main tools we will use from computational algebra and algebraic geometry. Thus, it will be highly beneficial for the reader to also consult some texts that go into more depth on these topics, for instance [CLO07] or [Has07].

3.1. Varieties

Let $K$ be a field. For the purposes of this book, there are three main fields of interest: the rational numbers $\mathbb{Q}$, which are useful for performing computations, the real numbers $\mathbb{R}$, which are fundamental because probabilities are real, and the complex number $\mathbb{C}$, which we will need in some circumstances for proving precise theorems in algebraic geometry.
Let $p_1, p_2, \ldots, p_r$ be polynomial variables or indeterminates (we will generally use the word indeterminate instead of variable to avoid confusion with random variables). A monomial in these indeterminates is an expression of the form

$$p^u := p_1^{u_1} p_2^{u_2} \cdots p_r^{u_r}$$

where $u = (u_1, u_2, \ldots, u_r) \in \mathbb{N}^r$ is a vector of nonnegative integers. The set of integers is denoted by $\mathbb{Z}$ and the set of nonnegative integers is $\mathbb{N}$. The notation $p^u$ is monomial vector notation for the monomial in Equation 3.1.1.

A polynomial in indeterminates $p_1, p_2, \ldots, p_r$ over $K$ is a linear combination of finitely many monomials, with coefficients in $K$. That is, a polynomial can be represented as a sum

$$f = \sum_{u \in A} c_u p^u$$

where $A$ is a finite subset of $\mathbb{N}^r$, each each coefficient $c_u \in K$. Alternately, we could say that $f = \sum_{u \in \mathbb{N}^r} c_u p^u$ where only finitely many of the coefficients $c_u$ are nonzero elements of the field $K$.

For example, the expression $p_1^2 p_3^4$ is a monomial in the indeterminates $p_1, p_2, p_3$, with $u = (2, 0, 4)$, whereas $p_1^2 p_3^{-4}$ is not a monomial because there is a negative in the exponent of indeterminate $p_3$. An example of a polynomial is $f = p_1^2 p_3^4 + 6p_1^2 p_2 - \sqrt{2}p_1^2 p_2 p_3^2$ which has three terms. The power series

$$\sum_{i=0}^{\infty} p_1^i$$

is not a polynomial because it has infinitely many terms with nonzero coefficients.

Let

$$K[p] := K[p_1, p_2, \ldots, p_r]$$

be the set of all polynomials in the indeterminates $p_1, p_2, \ldots, p_r$, with coefficients in $K$. This set of polynomials forms a ring, because the sum of two polynomials is a polynomial, the product of two polynomials is a polynomials, and these two operations follow all the usual properties of a ring, (e.g. multiplication distributes over addition of polynomials). Each polynomial is naturally considered as a multivariate function from $\mathbb{K}^r \to \mathbb{K}$, by evaluation. For this reason, we will often refer to the ring $K[p]$ as the ring of polynomial functions.

**Definition 3.1.1.** Let $S \subseteq K[p]$ be a set of polynomials. The variety defined by $S$ is the set

$$V(S) = \{a \in \mathbb{K}^r : f(a) = 0 \text{ for all } f \in S\}.$$ 

The variety $V(S)$ is also called the zero set of $S$. 
Note that the set of points in the variety depends on which field \( K \) we are considering. Most of the time, the particular field of interest should be clear from the context. In nearly all situations explored in this book, this will either be the real numbers \( \mathbb{R} \) or the complex numbers \( \mathbb{C} \). When we wish to call attention to the nature of the particular field under consideration, we will use the notation \( V_{\mathbb{K}}(S) \) to indicate the variety in the \( \mathbb{K} \).

**Example 3.1.2 (Some Varieties in \( \mathbb{C}^2 \)).** The variety \( V(\{(p_1 - p_2^2)\}) \subset \mathbb{R}^2 \) is the familiar parabolic curve, facing towards the right, whereas the variety \( V(\{(p_1^2 + p_2^2 - 1)\}) \subset \mathbb{R}^2 \) is a circle centered at the origin. The variety \( V(\{(p_1 - p_2^2, p_1^2 + p_2^2 - 1)\}) \) is the set of points that satisfy both of the equations

\[
p_1 - p_2^2 = 0 \quad p_1^2 + p_2^2 - 1 = 0
\]

and thus is the intersection \( V(\{(p_1 - p_2^2)\}) \cap V(\{(p_1^2 + p_2^2 - 1)\}) \). Thus, the variety \( V(\{(p_1 - p_2^2, p_1^2 + p_2^2 - 1)\}) \) is a finite number of points. It makes a difference which field we consider this variety over, and the number of points varies with the field. Indeed,

\[
V_{\mathbb{Q}}(\{(p_1 - p_2^2, p_1^2 + p_2^2 - 1)\}) = \emptyset
\]

\[
V_{\mathbb{R}}(\{(p_1 - p_2^2, p_1^2 + p_2^2 - 1)\}) = \left\{ \left( \frac{-1 + \sqrt{5}}{2}, \pm \sqrt{\frac{-1 + \sqrt{5}}{2}} \right) \right\}
\]

\[
V_{\mathbb{C}}(\{(p_1 - p_2^2, p_1^2 + p_2^2 - 1)\}) = \left\{ \left( \frac{-1 \pm \sqrt{5}}{2}, \pm \sqrt{\frac{-1 \pm \sqrt{5}}{2}} \right) \right\}
\]

In many situations, it is natural to consider polynomial equations like \( p_1^2 + 4p_2 = -17 \), that is, polynomial equations that are not of the form \( f(p) = 0 \), but instead of the form \( f(p) = g(p) \). These can, of course, be transformed into the form discussed previously, by moving all terms to one side of the equation, i.e. \( f(p) - g(p) = 0 \). Writing equations in this standard form is often essential for exploiting the connection to algebra which is at the heart of algebraic geometry.

In many situations in algebraic statistics, varieties are presented as parametric sets. That is, there is some polynomial map \( \phi : \mathbb{K}^d \rightarrow \mathbb{K}^r \), such that \( V = \{ \phi(t) : t \in \mathbb{K}^d \} \) is the image of the map. We will also use \( \phi(\mathbb{K}^d) \) as shorthand to denote this image.

**Example 3.1.3 (Twisted Cubic).** Consider the polynomial map

\[
\phi : \mathbb{C} \rightarrow \mathbb{C}^3, \quad \phi(t) = (t, t^2, t^3)
\]

The image of this map is the twisted cubic curve in \( \mathbb{C}^3 \). Every point on the twisted cubic satisfies the equations \( p_1^2 = p_2 \) and \( p_1^3 = p_3 \), and conversely every point satisfying these two equations is in the image of the map \( \phi \).
Example 3.1.4 (Binomial Random Variables). Consider the polynomial map \( \phi : \mathbb{C} \to \mathbb{C}^{r+1} \) whose \( i \)th coordinate function is

\[
\phi_i(t) = \binom{r}{i} t^i (1-t)^{r-i}, \quad i = 0, 1, \ldots, r.
\]

For a real parameter \( \theta \in [0, 1] \), the value \( \phi_i(\theta) \) is the probability \( P(X = i) \) where \( X \) is a binomial random variable with success probability \( \theta \). That is, if \( \theta \) is the probability of getting a head in one flip of a biased coin, \( \phi_i(\theta) \) is the probability of seeing \( i \) heads in \( r \) independent flips of the same biased coin. Hence, the vector \( \phi(\theta) \) denotes the probability distribution of a binomial random variable.

The image of the real interval, \([0, 1]\), is a curve inside the probability simplex

\[
\Delta_r = \left\{ p \in \mathbb{R}^{r+1} : \sum_{i=0}^{r} p_i = 1, p_i \geq 0 \text{ for all } i \right\}
\]

which consists of all possible probability distributions of a binomial random variable. For example, in the case of \( r = 2 \), the curve \( \phi([0, 1]) \) is the curve inside the probability simplex in Figure 3.1.1.

The image of the complex parametrization \( \phi(\mathbb{C}) \) is a curve in \( \mathbb{C}^{r+1} \) that is also an algebraic variety. In particular, it is the set of points satisfying the trivial equation \( \sum_{i=0}^{r} p_i = 1 \) together with the vanishing of all \( 2 \times 2 \) subdeterminants of the following \( 2 \times r \) matrix:

\[
\begin{pmatrix}
\left(\frac{p_0}{\binom{r}{0}}\right) & \left(\frac{p_1}{\binom{r}{1}}\right) & \left(\frac{p_2}{\binom{r}{2}}\right) & \cdots & \left(\frac{p_{r-1}}{\binom{r}{r-1}}\right) \\
\left(\frac{p_1}{\binom{r}{1}}\right) & \left(\frac{p_2}{\binom{r}{2}}\right) & \left(\frac{p_3}{\binom{r}{3}}\right) & \cdots & \left(\frac{p_r}{\binom{r}{r}}\right)
\end{pmatrix}.
\]

In applications to algebraic statistics, it is common to consider not only maps whose coordinates functions are polynomial functions, but also maps whose coordinate functions are rational functions. Such, functions have the form \( f/g \) where \( f \) and \( g \) are polynomials. Hence, a rational map is a
map \( \phi : \mathbb{C}^d \rightarrow \mathbb{C}^r \), where each coordinate function \( \phi_i = f_i/g_i \), where each \( f_i, g_i \in \mathbb{K}[t] := \mathbb{K}[t_1, \ldots, t_d] \) are polynomials. Note that the map \( \phi \) is not defined on all of \( \mathbb{C}^d \), it is only well defined off of the hypersurface \( V(\{ \prod_i g_i \}) \).

If we speak of the image of \( \phi \), we mean the image of the set \( \mathbb{C}^d \setminus V(\{ \prod_i g_i \}) \), where the rational map \( \phi \) is well-defined. The symbol \( \mapsto \) is used in the notational definition of a rational map to indicate that the map need not be defined everywhere.

### 3.2. Ideals

In the preceding section, we showed how to take a collection of polynomials (an algebraic object) to produce their common zero set (a geometric object). This construction can be reversed: from a geometric object (subset of \( \mathbb{K}^r \)) we can construct the set of polynomials that vanish on it. The sets of polynomials that can arise in this way have the algebraic structure of a radical ideal. This interplay between varieties and ideals is at the heart of algebraic geometry. We describe the ingredients in this section.

**Definition 3.2.1.** Let \( W \subseteq \mathbb{K}^r \). The set of polynomials

\[
I(W) = \{ f \in \mathbb{K}[p] : f(a) = 0 \text{ for all } a \in W \}
\]

is the **vanishing ideal** of \( W \) or the **defining ideal** of \( W \).

Recall that a subset \( I \) of a ring \( R \) is an **ideal** if \( f + g \in I \) for all \( f, g \in I \) and \( hf \in I \) for all \( f \in I \) and \( h \in R \). As the name implies, the set \( I(W) \) is an ideal (e.g. Exercise 3.1).

A collection of polynomials \( \mathcal{F} \) is said to **generate** an ideal \( I \), if it is the case that every polynomial \( f \in I \) can be written as a finite linear combination of the form \( f = \sum_{i=1}^{k} h_i f_i \) where the \( f_i \in \mathcal{F} \) and the \( h_i \) are arbitrary polynomials in \( \mathbb{K}[p] \). The number \( k \) might depend on \( f \). In this case we use the notation \( \langle \mathcal{F} \rangle = I \) to denote that \( I \) is generated by \( \mathcal{F} \).

**Theorem 3.2.2 (Hilbert Basis Theorem).** For every ideal \( I \) in a polynomial ring \( \mathbb{K}[p_1, \ldots, p_r] \) there exists a finite set of polynomials \( \mathcal{F} \subset I \) such that \( I = \langle \mathcal{F} \rangle \). That is, every ideal in \( \mathbb{K}[p] \) has a finite generating set.

Note that the Hilbert Basis Theorem only holds for polynomial rings in finitely many variables. In general, a ring that satisfies the property that every ideal has a finite generating set is called a Noetherian ring.

When presented with a set \( W \subseteq \mathbb{K}^r \), we are often given the task of finding a finite set of polynomials in \( I(W) \) that generate the ideal. The special case where the set \( W = \phi(\mathbb{K}^d) \) for some rational map \( \phi \) is called an **implicitization problem** and many problems in algebraic statistics can be interpreted as implicitization problems.
Example 3.2.3 (Binomial Random Variables). Let $\phi : \mathbb{C} \to \mathbb{C}^{r+1}$ be the map from Example 3.1.4. Then the ideal $I(\phi(\mathbb{C})) \subseteq \mathbb{C}[p_0, \ldots, p_r]$ is generated by the polynomials $\sum_{i=0}^{r} p_i - 1$ plus the $2 \times 2$ minors of the matrix from Equation (3.1.2).

In the special case where $r = 2$, we have

$$I(\phi(\mathbb{C})) = \langle p_0 + p_1 + p_2 - 1, 4p_0p_2 - p_1^2 \rangle.$$ 

So the two polynomials $p_0 + p_1 + p_2 - 1$ and $4p_0p_2 - p_1^2$ solve the implicitization problem in this instance. We will describe computational techniques for solving the implicitization problem in Section 3.4.

If we are given an ideal, we can iteratively compute the variety defined by the vanishing of that ideal, and then compute the vanishing ideal of the resulting variety. We always have $I \subseteq I(V(I))$. However, this containment need not be an equality, as the example $I = (p_1^2)$ indicates. The ideals that can arise as vanishing ideals are forced to have some extra structure.

Definition 3.2.4. An ideal $I$ is called radical if $f^k \in I$ for some polynomial $f$ and positive integer $k$ implies that $f \in I$. The radical of $I$, denoted $\sqrt{I}$, is the smallest radical ideal that contains $I$ and consists of the polynomials:

$$\sqrt{I} = \left\{ f \in \mathbb{K}[p] : f^k \in I \text{ for some } k \in \mathbb{N} \right\}.$$ 

The following straightforward proposition explains the importance of radical ideals in algebraic geometry. Its proof is left to the exercises.

Proposition 3.2.5. For any field and any set $W \subseteq \mathbb{K}^r$, the ideal $I(W)$ is a radical ideal.

A much more significant result shows that the process of taking vanishing ideals of varieties gives the radical of an underlying ideal, in the case of an algebraically closed field. (Recall that a field is algebraically closed if every univariate polynomial with coefficients in that field has a root in the field.)

Theorem 3.2.6 (Nullstellensatz). Suppose that $\mathbb{K}$ is algebraically closed (e.g. $\mathbb{C}$). Then the vanishing ideal of the variety of an ideal is the radical of the ideal: $I(V(I)) = \sqrt{I}$.

See [CLO07] or [Has07] for a proof of Theorem 3.2.6. When $\mathbb{K}$ is algebraically closed, the radical provides a closure operation for ideals, by determining all the functions that vanish on the underlying varieties $V(I)$. Computing $V(I(W))$ provides the analogous algebraic closure operations for subsets of $\mathbb{K}^r$. 
Proposition 3.2.7. The set \( V(I(W)) \) is called the Zariski closure of \( W \). It is the smallest algebraic variety that contains \( W \).

See Exercise 3.1 for the proof of Proposition 3.2.7. A set \( W \subseteq \mathbb{K}^r \) is called Zariski closed if it is a variety, that is of the form \( W = V(S) \) for some \( S \subseteq \mathbb{K}[p] \). Note that since polynomial functions are continuous in the standard Euclidean topology, the Zariski closed sets are also closed in the standard Euclidean topology.

Example 3.2.8 (A Simple Zariski Closure). As an example of a Zariski closure let \( W \subset \mathbb{R}^2 \) consists of the \( p_1 \) axis with the origin \((0,0)\) removed. Because polynomials are continuous, any polynomial that vanishes on \( W \) vanishes on the origin as well. Thus, the Zariski closure contains the \( p_1 \) axis. On the other hand, the \( p_1 \) axis is the variety defined by the ideal \( (p_2) \), so the Zariski closure of \( W \) is precisely the \( p_1 \) axis.

Note that if we consider the same set \( W \) but over a finite field \( \mathbb{K} \), the set \( W \) will automatically be Zariski closed (since every finite set is always Zariski closed).

Example 3.2.9 (Binomial Random Variable). Let \( \phi : \mathbb{C} \to \mathbb{C}^{r+1} \) be the map from Example 3.1.4. The image of the interval \([0,1]\) in \( \Delta_r \) is not a Zariski closed set. The Zariski closure of \( \phi([0,1]) \) equals \( \phi(\mathbb{C}) \). This is because \( \phi(\mathbb{C}) \) is Zariski closed, and the Zariski closure of \([0,1] \subseteq \mathbb{C}\) is all of \( \mathbb{C}\).

Note that the Zariski closure of \([0,1]\) is not \( \mathbb{R} \subseteq \mathbb{C} \). While \( \mathbb{R} = \{z : z = \overline{z}\} \), \( \mathbb{R} \) is not a variety when considered as a subset of \( \mathbb{C} \), since complex conjugation is not a polynomial operation.

If we only care about subsets of \( \Delta_r \), then \( \phi([0,1]) \) is equal to the Zariski closure intersected with the probability simplex. In this case, the variety gives a good approximation to the statistical model of binomial random variables.

The complement of a Zariski closed sets is called a Zariski open set. The Zariski topology is the topology whose closed sets are the Zariski closed sets. The Zariski topology is a much coarser topology that the standard topology, as the following example illustrates.

Example 3.2.10 (Mixture of Binomial Random Variables). Let \( \mathcal{P} \subseteq \Delta_{r-1} \) be a family of probability distributions (i.e. a statistical model) and let \( s > 0 \) be an integer. The \( s \)th mixture model \( \text{Mix}^s(\mathcal{P}) \) consists of all probability distributions which arises a convex combinations of \( s \) distributions in \( \mathcal{P} \).
That is,

\[ \text{Mixt}^s(P) = \left\{ \sum_{j=1}^{s} \pi_j p^{(j)} : \pi \in \Delta_{s-1} \text{ and } p^{(j)} \in P \text{ for all } j \right\}. \]

Mixture models allow to build a complex model from a simple model \( P \). The motivation for modeling with a mixture model will be explained in Chapter 14.

The special case of mixtures of binomial random variables arises in Bayesian statistics through its connection to exchangeability via De Finetti’s Theorem (see e.g. [DF80]). Focusing on the case \( r = 2 \) and \( s = 2 \), we see that \( \text{Mixt}^2(P) \) is precisely the region below the curve in Figure 3.1.1. Since \( \text{Mixt}^2(P) \) is a 2-dimensional subset of \( \Delta_2 \), we see that the Zariski closure of \( \text{Mixt}^2(P) \) is the entire plane \( V(p_0 + p_1 + p_2 - 1) \). Thus, the intersection of the Zariski closure and the probability simplex \( \Delta_2 \) is not the mixture model. This example illustrates the fact that in some situations in algebraic statistics, it might not be sufficient to only consider varieties. Indeed, sometimes constraints using inequalities are necessary. In this example, the mixture model is the set

\[ \Delta_2 \cap \left\{ p : \det \begin{pmatrix} 2p_0 & p_1 \\ p_1 & 2p_2 \end{pmatrix} \geq 0 \right\}. \]

We will return to the issues of real algebraic geometry in later chapters.

Building on the Nullstellensatz, and the notion of Zariski closures, there is a natural correspondence between radical ideals on the one hand, and varieties on the other.

**Theorem 3.2.11 (Ideal-Variety Correspondence).** Let \( K \) be an algebraically closed field. Then the maps

\[ V : \left\{ I \subseteq K[p] : I = \sqrt{I} \right\} \rightarrow \left\{ W \subseteq K^r : W = V(J) \text{ for some } J \subseteq K[p] \right\} \]

\[ I : \left\{ W \subseteq K^r : W = V(J) \text{ for some } J \subseteq K[p] \right\} \rightarrow \left\{ I \subseteq K[p] : I = \sqrt{I} \right\} \]

are inclusion-reversing bijections between the set of radical ideals and the set of varieties.

Note that the condition that \( K \) is algebraically closed is necessary (see Exercise 3.4).

**Proof.** The fact that \( V(I(V)) = V \) for \( V \) a variety follows from the definition of variety. On the other hand, the Nullstellensatz shows that \( V(I(V)) = I \) if \( I \) is a radical ideal and \( K \) is algebraically closed. This shows that \( V \) and \( I \) are bijections and inverses of each other. To see that these maps are
inclusion reversing, note that if \( I \subseteq J \) then \( V(J) \subseteq V(I) \), since in the definition of variety \( V(J) \) there are more constraints than in \( V(I) \), and hence fewer points can satisfy these conditions.

Theorem 3.2.11 is the Rosetta stone for translating between algebra and geometry. Natural operations on the ideal or variety side lead to natural related operations on the opposing side. For example consider the sum of two ideals:

\[
I + J := \{ f + g : f \in I \text{ and } g \in J \}.
\]

If \( I = \langle \mathcal{F} \rangle \) and \( J = \langle \mathcal{G} \rangle \), then \( I + J = \langle \mathcal{F} \cup \mathcal{G} \rangle \). This observation immediately leads to a geometric consequence on the level of varieties.

**Proposition 3.2.12.** Let \( I \) and \( J \) be ideals of \( \mathbb{K}[p] \). Then

\[
V(I + J) = V(I) \cap V(J).
\]

If \( \mathbb{K} \) is an algebraically closed field and \( V \) and \( W \) are arbitrary subsets of \( \mathbb{K}^r \) then

\[
I(V \cap W) = \sqrt{I(V)} + I(W).
\]

**Proof.** It suffices to show the first equation, since the second follows by the ideal-variety correspondence. Suppose that \( a \in V(I + J) \). Then \( f(a) = 0 \) for all \( f \in I \) and for all \( f \in J \). This implies that \( a \in V(I) \cap V(J) \) and implies the containment \( V(I + J) \subseteq V(I) \cap V(J) \). On the other hand, if \( a \in V(I) \cap V(J) \), it satisfies all polynomials in \( I \) and in \( J \), and hence, all polynomials in \( I + J \). □

An analogous result holds connecting intersections of ideals and unions of varieties. The proof is left as an exercise.

**Proposition 3.2.13.** Let \( I \) and \( J \) be ideals of \( \mathbb{K}[p] \). Then

\[
V(I \cap J) = V(I) \cup V(J).
\]

If \( V \) and \( W \) are arbitrary subsets of \( \mathbb{K}^r \) then

\[
I(V \cup W) = I(V) \cap I(W).
\]

### 3.3. Gröbner Bases

Gröbner bases are the algebraic tool for studying ideals computationally. The most basic use for Gröbner bases to answer the ideal membership problem; that is, to decide whether or not a given polynomial \( f \) belongs to a particular ideal \( I \). The fact that Gröbner bases solve the ideal membership problem allows for a great many further applications. We highlight some of these uses in this chapter and defer other applications until later in the book.
Definition 3.3.1. A term order $\prec$ is a total order on the set of monomials in the polynomial ring $\mathbb{K}[p]$ which satisfies the two conditions

1. $1 = p^0 \preceq p^u$ for all $u \in \mathbb{N}^r$
2. $p^u \prec p^v$ implies that $p^w \cdot p^u \prec p^w \cdot p^v$ for all $w \in \mathbb{N}^r$.

Three common examples of terms orders are the lexicographic, reverse lexicographic, and weight term orders.

Example 3.3.2 (Lexicographic Term Order). The lexicographic term order $\prec_{\text{lex}}$ is defined by the rule $p^u \prec_{\text{lex}} p^v$ if and only if the leftmost nonzero element of $v - u$ is positive. For instance, the following is a chain of valid inequalities in the lexicographic order:

$p_1^3 \succ_{\text{lex}} p_2^2 p_2^3 \succ_{\text{lex}} p_1^2 p_3^3 \succ_{\text{lex}} p_2 \succ_{\text{lex}} p_3^3$.

The lexicographic order we have defined is with respect to a particular ordering of the indeterminates. In particular, we are assuming that $p_1 \succ_{\text{lex}} p_2 \succ_{\text{lex}} \cdots \succ_{\text{lex}} p_r$. However, any permutation of the indeterminates yields a different lexicographic term order.

Example 3.3.3 (Reverse Lexicographic Term Order). The reverse lexicographic term order $\prec_{\text{rev}}$ is defined by the rule $p^u \prec_{\text{rev}} p^v$ if and only if

1. $\sum_{i=1}^r u_i < \sum_{i=1}^r v_i$, or
2. $\sum_{i=1}^r u_i = \sum_{i=1}^r v_i$ and the rightmost nonzero entry of $v - u$ is negative.

For instance, the following is a chain of valid inequalities in the reverse lexicographic term order:

$p_1^3 \succ_{\text{rev}} p_2^2 p_2^3 \succ_{\text{rev}} p_1^2 p_3^3 \succ_{\text{rev}} p_2 \succ_{\text{rev}} p_3^3 \succ_{\text{rev}} p_1$.

As with the lexicographic order, we can get many different reverse lexicographic term orders, by permuting the order of the indeterminates.

Example 3.3.4 (Weight Orders). The weight orders are an important class of term orders that will play a role in Chapter 10. Let $c \in \mathbb{R}^r$ be a real vector, and $\prec$ an arbitrary term order. The weight order $\prec_c$ is defined by the rule $p^u \prec_c p^v$ if and only if

1. $c^T u < c^T v$ or
2. $c^T u = c^T v$ and $p^u \prec p^v$.

The interpretation of the weight orders that will be useful in Chapter 10 is that the vector $c$ is a cost vector. Then monomials are larger in the weight order if they have larger cost with respect to the linear cost function $c^T u$, and ties are broken with the tie breaking term order $\prec$. Note that as we
have defined term orders, the weight order is only a term order if $c_i \geq 0$ for all $i$, an issue that we will remedy later.

Besides being a useful tool for its connections to integer programming that we will see in Chapter 10, the weight orders also provide a useful mnemonic for remembering the definitions of the lexicographic and reverse lexicographic term orders. While neither of these term orders can be realized as weight orders, the idea of weight of an indeterminate is natural in this context. Indeed, for the lexicographic order, the most “expensive” indeterminates ($p_1$ then $p_2$, etc.) are so expensive that they make any monomial containing them large in the term order $\prec_{\text{lex}}$. On the other hand, for the reverse lexicographic order, the cheap variables ($p_r$, then $p_{r-1}$, etc.) are so cheap that they make any monomial containing them small in the term order $\prec_{\text{rev}}$.

**Definition 3.3.5.** Given a term order $\prec$ and a polynomial $f \in K[p]$, the initial monomial, initial term, or leading term of $f$ with respect to $\prec$ is the largest monomial $p^\alpha$ with respect to the ordering $\prec$ among all the monomials appearing in $f$ with nonzero coefficient. The initial monomial of $f$ is denoted $\text{in}_\prec(f)$.

For example, if $f = p_2^3 - 3p_2^2p_2 + 6p_2^2p_3 + \frac{x}{y}p_2^4 + p_1^2p_2^2$. Then $\text{in}_{\text{lex}}(f) = p_1^2p_2$ whereas $\text{in}_{\text{rev}}(f) = p_3^4$. Initial terms extend naturally from polynomials to ideals.

**Definition 3.3.6.** The initial ideal $\text{in}_\prec(I)$ of an ideal $I \subseteq K[p]$ with respect to the term order $\prec$ is the ideal generated by the initial monomials of all polynomials in $I$:

$$\text{in}_\prec(I) = \langle \text{in}_\prec(f) : f \in I \rangle.$$

Note that the initial ideal is not, in general, generated by the initial terms of a generating set of an ideal. For instance, consider the ideal $I = \langle p_1^2, p_1p_2 + p_2^2 \rangle$. Then $\text{in}_{\text{ex}}$ contains both the monomials $p_1^2$ and $p_1p_2$. These two monomials do not generate the initial ideal because the polynomial $p_2 \cdot p_1^2 - (p_1 - p_2) \cdot (p_1p_2 + p_2^2) = p_2^3$ belongs to $I$, but its initial term $p_2^3$ does not belong to the monomial ideal $\langle p_1^2, p_1p_2 \rangle$.

**Definition 3.3.7.** A Gröbner basis of the ideal $I$ with respect to the term order $\prec$ is a finite collection of polynomials $\mathcal{G} \subseteq I$ such that the initial monomials of the elements of $\mathcal{G}$ generate the initial ideal:

$$\langle \text{in}_\prec(g) : g \in \mathcal{G} \rangle = \text{in}_\prec(I).$$

An equivalent definition of a Gröbner basis is a finite set of polynomials $\mathcal{G} \subseteq I$ such that for all nonzero $f \in I$, there is a $g \in \mathcal{G}$ such that $\text{in}_\prec(g)$
divides in_{\prec}(f)$. Note that (finite) Gröbner bases always exist for any ideal and any term order by applying the Hilbert basis theorem: since in_{\prec}(I) is an ideal with finitely many (monomial) generators, we can choose a finite set $\mathcal{G} \subseteq I$ whose initial terms generate in_{\prec}(I). The goal of the remainder of this section is to describe how to compute Gröbner bases and initial ideals of ideals. While the definition of Gröbner basis might seem mysterious at first glance, the motivation comes from its many applications and by its fundamental connection to the multivariate division algorithm.

Algorithm 3.3.8 (Multivariate Division Algorithm). Input: Let $\mathcal{G} = \{g_1, \ldots, g_k\}$ be a finite set of polynomials, $f$ another polynomial, and $\prec$ a term order. Output: A representation $f = \sum_{i=1}^{k} h_i g_i + r$ such that no term of $r$ is divisible by an initial term of any of the polynomials in $\mathcal{G}$.

1. Set $h_i = 0$ for all $i$ and $r = f$.
2. While $r$ has a term $c g^p$ divisible by a leading term of some $g_i$, replace $h_i$ by $h_i + c a p^a / \text{in}_{\prec}(g_i)$ and $r$ by $r - c a p^a / \text{in}_{\prec}(g_i) \cdot g_i$.

The division algorithm terminates because a term ordering is a well ordering: there is no infinite decreasing sequence of monomials with respect to a term order.

The remainder $r$ upon dividing $f$ by $\mathcal{G}$ is called the normal form of $f$, and is denoted by the shorthand notation $\text{NF}_\mathcal{G}(f)$. Of course, one must be careful with this definition because the normal form is not uniquely determined by the $\mathcal{G}$ and $f$ in our loose definition of the division algorithm. However, uniqueness can be enforced by eliminating choices in the division procedure (e.g. always dividing the smallest available initial term into the large available term).

In the special case where $\mathcal{G}$ consists of a single polynomial $g$, the normal form computation directly allows us to decide whether or not the polynomial $f$ belongs to the ideal $(g)$. In particular, in this case $\text{NF}_\mathcal{G}(f) = 0$ if and only if $f \in (g)$. The division algorithm also gives a method for addressing the ideal membership problem for arbitrary ideals, but one must work with a Gröbner basis. Indeed, this gives an alternate characterization of a Gröbner basis.

Theorem 3.3.9. Let $I$ be an ideal, $\prec$ a term order, and $\mathcal{G}$ a finite subset of $I$. Then $\text{NF}_\mathcal{G}(f) = 0$ for all $f \in I$ if and only if $\mathcal{G}$ is a Gröbner basis for $I$ with respect to $\prec$.

Proof. Suppose that $\text{NF}_\mathcal{G}(f) = 0$ for all $f \in I$. Then if $f \in I$, there must be some $g \in \mathcal{G}$ such that in_{\prec}(g) divides in_{\prec}(f) so that in_{\prec}(f)$ could be
eventually cancelled out in the long division process. This implies that $G$ is a Gröbner basis for $I$.

On the other hand, if $G$ is a Gröbner basis for $I$ and $f \in I$ and $r = NF_G(f)$, then the division algorithm produces $h_1, \ldots, h_k$ such that $f = \sum_{i=1}^k h_ig_i + r$. Rearranging this expression we have

$$r = f - \sum_{i=1}^k h_ig_i$$

so the normal form $r = NF_G(f) \in I$. But in the division algorithm, $\omega(r)$ is not divisible by any leading term of $G$. Since $G$ is a Gröbner basis, this means that $r = 0$. \hfill \Box

One particular application of Theorem 3.3.9 is we can directly check for ideal member once we have a Gröbner basis. In particular, if $G$ is a Gröbner basis for $I$ then $f \in I$ if and only if $NF_G(f) = 0$. Another consequence of the definition of Gröbner basis is that a Gröbner basis of an ideal is always a generating set for that ideal.

**Corollary 3.3.10.** Let $I$ be an ideal $\prec$ a term order, and $G$ a Gröbner basis for $I$ with respect to $\prec$. Then $G$ is a generating set of $I$.

**Proof.** Let $f \in I$ and $G = \{g_1, \ldots, g_k\}$. Theorem 3.3.9 shows that $NF_G(f) = 0$. This, in turn, shows that there are $h_1, \ldots, h_k$ in $K[p]$ such that $f = h_1g_1 + \cdots + h_ig_k$. In other words, every element of $f$ can be expressed as a polynomial combination of the elements of $G$. Hence, $G$ is a generating set of $I$. \hfill \Box

While we have seen that Gröbner bases always exist, it is not obvious at first glance how to compute a Gröbner basis from a generating set. One simple ideal is to try to generate new polynomials that are in the same ideal, but whose initial terms might not be in the ideal generated by the initial terms of our given generating set. One way to do this is based on canceling initial terms. In particular, let $f$ and $g$ be two polynomials with $\prec$ leading terms $p^\alpha$ and $p^\beta$ respectively. We can form the polynomial

$$S(f, g) = \frac{p^\alpha}{\gcd(p^\alpha, p^\beta)} f - \frac{p^\beta}{\gcd(p^\alpha, p^\beta)} g$$

which is called the $S$-polynomial of $f$ and $g$. Note that the $S$-polynomial cancels the leading terms of $f$ and $g$ in the most economical way possible, and $S(f, g)$ gives a candidate for a new polynomial which might need to appear in a Gröbner basis. If $f$ and $g$ were part of a Gröbner basis $G$ of $I$ with respect to $\prec$, then $NF_G(S(f, g)) = 0$, by Theorem 3.3.9. The converse is also true.
Theorem 3.3.11 (Buchberger’s Criterion). Let $I$ be an ideal, $\prec$ a term order and $\mathcal{G}$ a finite generating set of $I$. Then $\mathcal{G}$ is a Gröbner basis of $I$ with respect to $\prec$ if and only if $NF_G(S(f, g)) = 0$ for all $f, g \in \mathcal{G}$.

See [CLO07] or [Has07] for a proof. Buchberger’s criterion leads to a straightforward algorithm for computing Gröbner bases of ideals from their generating sets.

Algorithm 3.3.12 (Buchberger’s Algorithm).

**Input:** Finite generating set $F$ of $I$, and a term order $\prec$.

**Output:** Gröbner basis $G$ for $I$ with respect to $\prec$.

2. While $G$ does not satisfy Buchberger’s criterion, do:
   - There exists a pair $f, g \in G$ such that $NF_G(S(f, g)) = h \neq 0$.
   - Set $G := G \cup \{h\}$.
3. Output: $G$ is a Gröbner basis for $I$ with respect to $\prec$.

Note that the algorithm is guaranteed to terminate by the Hilbert basis theorem. At each step we have a candidate initial ideal $J_i = \langle \text{in}_\prec(f) : f \in G \rangle$.

These ideals form an ascending chain $J_1 \subset J_2 \subset J_3 \subset \cdots$.

Then Exercise 3.5 guarantees these ideals, and hence the set $G$, must stabilize. Algorithms which are implemented to compute Gröbner bases are all based on Buchberger’s algorithm. There are many refinements and extensions of Buchberger’s algorithm which can be used to speed up the computation and are implemented in practice.

Example 3.3.13. [The simplest Gröbner basis] Consider the ideal $I = \langle p_1^2, p_1p_2 + p_2^3 \rangle$ and the lexicographic term order. We will describe how to compute a Gröbner basis using Buchberger’s algorithm. We compute the $S$-polynomial

$$S(p_1^2, p_1p_2 + p_2^3) = p_2 \cdot p_1^2 - p_1 \cdot (p_1p_2 + p_2^3) = -p_1p_2^3.$$ 

Performing long-division by $p_1^2, p_1p_2 + p_2^3$ produces the normal form

$$NF_G(S(p_1^2, p_1p_2 + p_2^3)) = p_2^3,$$

which is nonzero. Hence, our new candidate Gröbner basis becomes $G = \{p_1^2, p_1p_2 + p_2^3\}$. Computing the normal forms of the $S$-polynomials for all pairs in $G$ reveals that $G$ is a Gröbner basis for $I$ with respect to the lexicographic term order.
3.4. First Applications of Gröbner Bases

The Buchberger algorithm and its generalizations are the major workhorse of computational algebra packages, because many useful invariants of an ideal \( I \), and hence, a variety, can be read off from properties of the initial ideal. Furthermore, Gröbner bases can be used to solve the implicitization problem for rational maps, which we describe in this section.

To begin, we show how to determine the dimension of a variety from a Gröbner basis calculation. For a variety, the dimension is simply the dimension as a topological space. For an ideal \( I \), we define the dimension to simply be the dimension of the variety \( V(I) \). For this to make sense, it is essential that we work over an algebraically closed field. When referring to the dimension of an ideal when the underlying field is not algebraically closed, we will mean the dimension over the algebraic closure (which corresponds to more general algebraic definitions of dimension, for instance, in [Eis95]). It turns out that this fundamental invariant of a variety can be computed directly from any initial ideal.

**Proposition 3.4.1.** Let \( K \) be an algebraically closed field. Then \( \dim I = \dim \text{in}_<(I) \) for any term order \( < \). Furthermore, for a monomial ideal \( M \), the dimension is equal to the cardinality of the largest subset \( S \subseteq [m] \) such that the monomial \( \prod_{i \in S} t_i^k \) is not in \( M \) for all \( k \in \mathbb{N} \).

See [CLO07] or [Has07] for a proof. Proposition 3.4.1 gives a straightforward method to compute the dimension of a variety, and is the standard method for computing dimensions in computational algebra packages.

**Example 3.4.2** (A zero dimensional variety). Consider the ideal \( I \) from Example 3.3.13. A Gröbner basis of \( I \) with respect to the lexicographic term order consists of polynomials \( G = \{ p_1^2, p_1 p_2 + p_2^2, p_3^2 \} \). The initial ideal of \( I \) is the monomial ideal

\[
M = \text{in}_{\text{lex}}(I) = \langle p_1^2, p_1 p_2, p_3^2 \rangle.
\]

In this example we can see that there is no nonempty set of indices \( S \) such that \( \prod_{i \in S} t_i^k \) is not in \( M \) for all \( k \in \mathbb{N} \). This implies that \( V(I) \) is zero dimensional. In fact \( V(I) \) consists of just a single point, the origin.

Among the most important geometric objects for algebraic statistic are parametrized algebraic sets, which will play a major role throughout the book.

That is, suppose we are given a polynomials \( \phi_1, \phi_2, \ldots, \phi_r \in K[t] \), \( K = \mathbb{K}[t_1, t_2, \ldots, t_d] \). Consider the map

\[
\phi : \mathbb{K}^d \to \mathbb{K}^r, \quad (\theta_1, \ldots, \theta_d) \mapsto (\phi_1(\theta), \ldots, \phi_r(\theta)).
\]
If \( \Theta \subseteq \mathbb{K}^d \)
\[
\phi(\Theta) = \{ \phi(\theta) : \theta \in \Theta \}
\]
is the image of the parametrization \( \phi \). More generally, we could consider this problem in the case of rational of a rational map \( \phi : \Theta \to \mathbb{K}' \). The \textit{implicitization problem} asks to take the sets \( \Theta \) and the rational map \( \phi \) and compute the vanishing ideal \( I(\phi(\Theta)) \).

As we saw with the example of the twisted cubic, the image of a parametrization can be a variety. Note that this need not be the case in general.

**Example 3.4.3** (A non-closed image). Consider the parametrization \( \phi : \mathbb{C}^2 \to \mathbb{C}^2 \) given by
\[
\phi(\theta_1, \theta_2) = (\theta_1, \theta_1 \theta_2).
\]
The image of this parametrization \( \phi(\mathbb{C}^2) \) consists of all points \((a_1, a_2)\) in \(\mathbb{C}^2\) such that \(a_1 \neq 0\) together with the origin. The set \(\phi(\mathbb{C}^2)\) is not a variety: the Zariski closure is equal to all of \(\mathbb{C}^2\), since any polynomial that vanishes on a dense subset of \(\mathbb{C}^2\) is the zero polynomial.

While the image of a parametrization need not be a variety, in the case that the ground field \(\mathbb{K}\) is algebraically closed, it is close to being a variety, as highlighted by the following theorem.

**Theorem 3.4.4.** Let \(\mathbb{K}\) be algebraically closed, let \(V \subseteq \mathbb{K}^{r_2}\) be a variety and \(\phi\) a rational map \(\phi : V \to \mathbb{K}^{r_1}\). Then there is a finite sequence of varieties \(W_1 \supset W_2 \supset \cdots \supset W_k\) in \(\mathbb{K}^{r_1}\) such that
\[
\phi(V) = W_1 \setminus (W_2 \setminus (\cdots \setminus W_k \cdots )).
\]

The sets that arise in Theorem 3.4.4, that arises as iterated set differences of varieties, are called \textit{constructible sets}. See Chapter 3 of [Har95] for a proof. Note that Theorem 3.4.4 only holds when \(\mathbb{K}\) is algebraically closed. For example consider the map \(\phi : \mathbb{R} \to \mathbb{R}, \phi(\theta) = \theta^2\). The image \(\phi(\mathbb{R})\) is the half-ray \([0, \infty)\) which is not constructible. The images of real algebraic varieties under rational maps do have special structure and are the subject of Tarski-Seidenberg theorem, which we will return to in Chapter 6.

To understand how to solve the implicitization problem for general rational maps, we must first get a handle on linear maps. After a change of coordinates, we can assume such a linear map is a coordinate projection. The next proposition and theorem show that the vanishing ideal of a coordinate projection can be computed using Gröbner bases.

**Proposition 3.4.5.** Let \(\pi : \mathbb{K}^{r_1+r_2} \to \mathbb{K}^{r_1}\) be the coordinate projection
\[
(a_1, \ldots, a_{r_1}, b_1, \ldots, b_{r_2}) \mapsto (a_1, \ldots, a_{r_1}).
\]
Let \(V \subseteq \mathbb{K}^{r_1+r_2}\) be a variety and
\[
I = I(V) \subseteq \mathbb{K}[p, q] := \mathbb{K}[p_1, \ldots, p_{r_1}, q_1, \ldots, q_{r_2}]
\]
be its vanishing ideal. Then the vanishing ideal of the image \( \pi(V) \) is the ideal

\[
I(\pi(V)) = I \cap \mathbb{K}[p].
\]

Note that in computational algebra, if we have an ideal \( I \subseteq \mathbb{K}[p, q] \), the ideal \( I \cap \mathbb{K}[p] \) is called an elimination ideal because the variables \( q_1, \ldots, q_{r_2} \) have been eliminated from consideration.

**Proof.** A polynomial \( f \) is in the ideal \( I(\pi(V)) \) if and only if it is in \( \mathbb{K}[p] \) and it vanishes on all points of \( \pi(V) \). But to say that a polynomial in \( \mathbb{K}[p] \) vanishes on \( \pi(V) \) is the same as saying that it is in \( I(V) \cap \mathbb{K}[p] \).

Generators for the elimination ideals can be determined by computing a Gröbner basis with an appropriate ordering of the variables.

**Theorem 3.4.6.** Let \( I \subseteq \mathbb{K}[p, q] \) and let \( \prec \) be a lexicographic term order such that

\[
q_1 \succ \cdots \succ q_{r_2} \succ p_1 \succ \cdots \succ p_{r_1}.
\]

If \( \mathcal{G} \) is a Gröbner basis of \( I \) with respect to \( \prec \) then

\[
I \cap \mathbb{K}[p] = (\mathcal{G} \cap \mathbb{K}[p])
\]

and furthermore \( \mathcal{G} \cap \mathbb{K}[p] \) is a Gröbner basis for the elimination ideal \( I \cap \mathbb{K}[p] \).

**Proof.** It suffices to prove that \( \mathcal{G} \cap \mathbb{K}[p] \) is a Gröbner basis for the elimination ideal \( I \cap \mathbb{K}[p] \). Let \( f \in I \cap \mathbb{K}[p] \). Since \( \mathcal{G} \) is a Gröbner basis for \( I \) there is a \( g \in \mathcal{G} \) such that \( \text{in}_\prec(g) \) divides \( \text{in}_\prec(f) \). Since \( f \in I \cap \mathbb{K}[p] \), \( \text{in}_\prec(f) \in \mathbb{K}[p] \) and hence \( \text{in}_\prec(g) \in \mathbb{K}[p] \), as well. By the properties of the lexicographic order, all terms of \( g \) must also belong to \( \mathbb{K}[p] \). This implies that \( g \in \mathbb{K}[p] \), and hence \( g \in \mathcal{G} \cap \mathbb{K}[p] \). This implies that \( \mathcal{G} \cap \mathbb{K}[p] \) is a Gröbner basis for \( I \cap \mathbb{K}[p] \).

Aside from the lexicographic term order used in Theorem 3.4.6, any other elimination term ordering can be used. An elimination ordering is a term ordering where any monomial with a \( q \) indeterminate appearing to any nonzero power is larger than every monomial only in the \( p \) indeterminates.

From linear projections, it is only a single step to computing the images of general rational maps. We first consider the case of polynomial maps. Suppose that \( \phi : \mathbb{K}^{r_2} \to \mathbb{K}^{r_1} \) is a polynomial map (each component function is a polynomial). The graph of the map is the set of points

\[
\Gamma_\phi = \{(\phi(\theta), \theta) : \theta \in \mathbb{K}^{r_2}\} \subset \mathbb{K}^{r_1 + r_2}.
\]

If \( \mathbb{K} \) is an infinite field the vanishing ideal of the graph is

\[
I(\Gamma_\phi) = \langle p_1 - \phi_1(q), \ldots, p_{r_1} - \phi_{r_1}(q) \rangle.
\]
Proposition 3.4.7. Let $I \subseteq K[q]$ be the vanishing ideal of a variety $V \subseteq K^{r_2}$. Let $\phi$ be a polynomial map $\phi : K^{r_2} \to K^{r_1}$. Then the vanishing ideal of the Zariski closure is the elimination ideal

$$I(\phi(V)) = (I + I(\Gamma_\phi)) \cap K[p].$$

The ideal $I + I(\Gamma_\phi)$ has as its generating set all the polynomial generators of $I \subseteq K[q]$ plus the $r_1$ polynomials $p_1 - \phi_1(q), \ldots, p_{r_1} - \phi_{r_1}(q)$. Hence the implicitization problem can be solved computationally by computing a Gröbner basis.

Proof. It suffices to show that $I + I(\Gamma_\phi)$ is the vanishing ideal of the variety

$$\Gamma_\phi(V) = \{(a, b) \in K^{r_1+r_2} : b \in V, \phi(b) = a\}.$$ 

Now we can choose a lexicographic term order such that the $p$ variables are larger than the $q$ variables. It is not difficult to see that that a Gröbner basis for $I + I(\Gamma_\phi)$ will be obtained by taking the polynomials $p_1 - \phi_1(q), \ldots, p_{r_1} - \phi_{r_1}(q)$ together with a Gröbner basis for the ideal $I$. Now let $f(p, q) \in I(\Gamma_\phi(V))$. Performing long division by the polynomials $p_1 - \phi_1(q), \ldots, p_{r_1} - \phi_{r_1}(q)$ produces a remainder polynomial $r(q)$ that does not depend on $p$. This polynomial is in $I(\Gamma_\phi(V)) \cap K[q] = I$. This shows that $f(p, q) \in I + I(\Gamma_\phi)$. \hfill \Box

Example 3.4.8 (Implicitization for Binomial Random Variable). Consider the parametrization of the binomial random variable model with 2 trials from Example 3.1.4. To compute the implicit description of this model, we can start with the vanishing ideal $I$ of the probability simplex $\Delta_1$, which is the ideal

$$I = (s + t - 1) \subseteq \mathbb{R}[s, t].$$

The parametrization has the form

$$\phi(s, t) = (s^2, 2st, t^2).$$

We form the ideal of the graph

$$I + I(\Gamma_\phi) = (s + t - 1, p_0 - s^2, p_1 - 2st, p_2 - t^2) \subseteq \mathbb{R}[s, t, p_0, p_1, p_2].$$

Computing a lexicographic Gröbner basis for this ideal with $s \gg t \gg p_0 \gg p_1 \gg p_2$ produces the vanishing ideal of the model

$$(p_0 + p_1 + p_2 - 1, 4p_0p_2 - p_1^2).$$

To handle rational maps requires only a small modification to the construction above. In particular, we need a way to handle the denominators that appear in the rational functions $\phi_i = f_i / g_i$. 


Proposition 3.4.9. Let $I \subset \mathbb{K}[q]$ be the vanishing ideal of a variety $V \subset \mathbb{K}^2$. Let $\phi$ be a rational map $\phi : \mathbb{K}^2 \to \mathbb{K}^1$, with coordinate functions $\phi_i = f_i/g_i$. Then the vanishing ideal of the Zariski closure is the elimination ideal

$$I(\phi(V)) = (I + (g_1p_1 - f_1, \ldots, g_rp_r - f_r, zg_1 \cdots g_r - 1)) \cap \mathbb{K}[p].$$

Note that the ideal in parentheses on the right hand side of the equality in Proposition 3.4.9, belongs to the ring $\mathbb{K}[p, q, z]$, where $z$ is an extra indeterminate. This extra indeterminate is the added feature which is necessary to clear the denominators. The polynomial $zg_1 \cdots g_r - 1$ guarantees that we only consider points in the graph where none of the denominators are zero. The proof of Proposition 3.4.9 is similar to that of Proposition 3.4.7 and is left as an exercise.

Besides solving the implicitization problem, lexicographic Gröbner bases can also be useful for solving systems of polynomial equations, in the event that there are only finitely many solutions. The projection strategy described above can be used to compute all possible values for one coordinate, and then back substitution can be used to find all solutions. While this is the most straightforward way to use Gröbner bases to solve a system of polynomial equations, in practice, a lexicographic Gröbner basis can be expensive to compute, and there are faster methods for solving zero-dimensional systems based on Gröbner bases and eigenvalue of associated “multiplication matrices”. See [CLO05] for details.

3.5. Computational Algebra Vignettes

Gröbner bases are one of the main tools for computing with algebraic varieties. Fortunately, algorithms using and based on Gröbner bases have been implemented in most software for performing symbolic computation. Many computations in these systems which use Gröbner bases can be performed directly without understanding how the Buchberger algorithm works, nor how certain computations are translated into Gröbner basis calculations. We illustrate some of the basic things that can be done with these computational algebra software packages in this section. Two pieces of software we illustrate here are Macaulay2 [GS] and Singular [DGPS16]. Throughout the book we will return to these two pieces of computational algebraic geometry software to illustrate various points and to demonstrate examples that go beyond what can easily be computed by hand. We stress, however, that in practice in algebraic statistics, we use the computational algebra software to gain insight by studying small examples. Then we usually try to prove theorems that generalize the small examples to more general settings.
Although we do not emphasize it in this book, another strategy for computing with algebraic varieties is based on numerical computations. This forms part of the area of *numerical algebraic geometry* represented by the software Bertini [BHSW] among others. This can be especially useful for polynomial solving systems with many roots. See [BHSW13].

Macauley 2 is a computational algebraic geometry software developed by Dan Grayson and Mike Stillman with contributions of packages from many mathematicians. We display some elementary computations with Gröbner bases here.

```plaintext
R1 = QQ[p0,p1,p2,p3,p4, MonomialOrder => Lex];

This command declares the ring R1. Note that the monomial order is a fixed feature of the ring.

i2 : M = matrix{
    {p0, p1/4, p2/6, p3/4},
    {p1/4, p2/6, p3/4, p4}};

o2 : Matrix R1 <--- R1

i3 : I = ideal(p0 + p1 + p2 + p3 + p4 -1) + minors(2, M);

o3 : Ideal of R1

i4 : gb I

o4 = GroebnerBasis[status: done; S-pairs encountered up to degree 4]

o4 : GroebnerBasis

i5 : gens gb I

o5 = | p3^4+16p3^3p4+96p3^2p4^2+256p3p4^3+256p4^4-256p4^3 |
       ---------------------------------------------------------------
       8p2p4-3p3^2 2p2p3^2+6p3^3+36p3^2p4+96p3p4^2+96p4^3-96p4^2 |
       ---------------------------------------------------------------
       2p2^2+12p2p3+72p3p4+72p4^2-72p4 6p1p4-p2p3 |
       ---------------------------------------------------------------
       3p1p3+8p2p3+18p3^2+48p3p4+48p4^2-48p4 |
       ---------------------------------------------------------------
       p1p2-10p2p3-30p3^2-90p3p4-6p3-96p4^2+96p4 |
       ---------------------------------------------------------------
       3p1^2+40p2p3-8p2+135p3^2+432p3p4+48p3+480p4^2-480p4 |
       ---------------------------------------------------------------
       p0+p1+p2+p3+p4-1 |
```
This command creates the vanishing ideal of the binomial random variable from Example 3.1.4, for \( r = 4 \). Note that the command `gb I` computes the Gröbner basis of the ideal \( I \) with respect to the lexicographic order declared in the ring \( R1 \). However, to see the result of the computation, use the command `gens gb I`. The operator `%` computes the normal form with respect to the Gröbner basis. The calculation `dim I` shows that the variety of this ideal is one dimensional.

The next collection of computations shows how to compute the vanishing ideal of the mixture model \( \text{Mixt}_2(P) \) where \( P \subseteq \Delta_4 \) is the model of a binomial random variable with \( r = 4 \) flips of a biased coin (which we have been computing with, above). To be able to use Proposition 3.4.7 we realize the mixture model as the image of the parametrization:

\[
\phi : P \times P \times \Delta_1 \rightarrow \mathbb{R}^5, \quad (p^{(1)}, p^{(2)}, \pi) \mapsto \pi p^{(1)} + (1 - \pi)p^{(2)}.
\]

This means we can compute the vanishing ideal \( I(\text{Mixt}_2(P)) \) by performing a Gröbner basis computation in 16 indeterminates. This is presented here:
60

3. Algebra Primer

o3 : Ideal of R2

i4 : M2 = matrix{
    {q20, q21/4, q22/6, q23/4},
    {q21/4, q22/6, q23/4, q24}};

o4 : Matrix R2 <--- R2

i5 : I2 = ideal(q20 + q21 + q22 + q23 + q24 -1) + minors(2, M2);

o5 : Ideal of R2

i6 : I3 = ideal(
    p0 - p*q10 - (1-p)*q20,
    p1 - p*q11 - (1-p)*q21,
    p2 - p*q12 - (1-p)*q22,
    p3 - p*q13 - (1-p)*q23,
    p4 - p*q14 - (1-p)*q24);

o6 : Ideal of R2

i7 : J = I1 + I2 + I3;

o7 : Ideal of R2

i8 : betti gens gb J

0 1
o8 = total: 1 77
   0: 1 3
   1: . 16
   2: . 22
   3: . 18
   4: . 17
   5: . 1

o8 : BettiTally

i9 : Mixt2I = ideal selectInSubring(1, gens gb J);

o9 : Ideal of R2

i10 : gens Mixt2I

o10 = | p0+p1+p2+p3+p4-1 2p2^2-9p1p2p3-27p1p3^2-27p2p3^2-27p3^3+27p1 |
\[ 2p4+72p1p2p4+72p2p3p4-27p3^2p4+72p2p4^2+27p3^2-72p2p4 \]

\[
\begin{array}{c|cc}
1 & 2 \\
\hline
\end{array}
\]

Note that we have declared our ring with an elimination term ordering that eliminates the first 11 indeterminates in the ring. The command \texttt{betti gens gb J} gives the number in each degree of the elements in the \text{Gröbner} basis of the ideal of the graph \( J \), with respect to the elimination ordering. However, the row indices of this array are shifted by 1 because of the connection to Betti numbers of free resolutions. So the table displays that there are 3 polynomials of degree 1, 16 of degree 2, \ldots, down to 1 of degree 6. The command \texttt{selectInSubring(1, gens gb J)} retrieves only those elements which are in the \( p_i \) indeterminates.

The vanishing ideal \( I(\text{Mixt}^2(\mathcal{P})) \) is generated by the trivial linear polynomial \( p_0 + p_1 + p_2 + p_3 + p_4 - 1 \) and a complicated looking cubic. Just as the model of binomial random variables had a vanishing ideal generated by \( 2 \times 2 \) minors, its mixture model has vanishing ideal generated by a \( 3 \times 3 \) minor, as the following computation verifies.

\begin{verbatim}
i11 : M3 = matrix{
    {p0, p1/4, p2/6},
    {p1/4, p2/6, p3/4},
    {p2/6, p3/4, p4}};

i12 : K = ideal(p0 + p1 + p2 + p3 + p4 - 1) + minors(3, M3);

i13 : Mixt2I == K
\end{verbatim}

An alternate strategy for computing the vanishing ideal of this mixture model is to use the fact that the underlying model \( \mathcal{P} \) is itself the image of a parametrization. Indeed, it is the image of a parametrization

\[
\phi : \Delta_1 \times \Delta_1 \times \Delta_1 \to \mathbb{R}^5
\]

\[
\phi_i(\theta_1, \theta_2, \pi) = \binom{4}{i} \pi (1 - \theta_1)^{4-i} \theta_1^i + \binom{4}{i} (1 - \pi) (1 - \theta_2)^{4-i} \theta_2^i.
\]
This induces the pullback ring homomorphism
\[ \phi^* : \mathbb{K}[p_0, p_1, p_2, p_3, p_4] \rightarrow \mathbb{K}[\pi, \theta_1, \theta_2] \]
\[ \phi^*_i(p_i) = \left( \frac{4}{i} \right) \pi(1-\theta_1)^{4-i}\theta_1^i + \left( \frac{4}{i} \right)(1-\pi)(1-\theta_2)^{4-i}\theta_2^i. \]

The vanishing ideal \( I(\text{Mixt}^2(P)) \) is the kernel of \( \phi^* \). The following code realizes the computation of the vanishing ideal in this manner.

```plaintext
i1 : S = QQ[a1,a2,p];
i2 : R = QQ[p0,p1,p2,p3,p4];
i3 : f = map(S,R,{
p*(1-a1)^4*a1^0 + (1-p) * (1-a2)^4*a2^0,
4*p*(1-a1)^3*a1^1 + 4*(1-p) * (1-a2)^3*a2^1,
6*p*(1-a1)^2*a1^2 + 6*(1-p) *(1-a2)^2*a2^2,
4*p*(1-a1)^1*a1^3 + 4*(1-p) *(1-a2)^1*a2^3,
p*(1-a1)^0*a1^4 + (1-p) *(1-a2)^0*a2^4});
o3 : RingMap S <--- R

i4 : K = kernel f

o4 : Ideal of R
```

These are just some of the rich range of algebraic computations that can be performed in Macaulay2. Further examples are presented throughout the book.

A second major computer algebra software for working with algebraic varieties is the system Singular. The range of tasks that Macaulay2 and Singular can perform are similar, though each has their own range of problems that they are especially good at solving. While the syntax of Macaulay2 is intended to be similar to the way that mathematics is written by mathematicians, Singular uses syntax that is more similar to other more traditional programming languages.

```plaintext
> ring R = 0, (p0,p1,p2,p3), dp;
```
3.6. Projective Space and Projective Varieties

\[ \text{matrix } m\{2\}\{3\} = p_0, \frac{p_1}{3}, \frac{p_2}{3}, \frac{p_1}{3}, \frac{p_2}{3}, p_3; \]
\[ \text{ideal } I = \text{minor}(m,2), p_0+p_1+p_2+p_3 -1; \]
\[ I; \]
\[ I[1]=-1/9*p_2^2+1/3*p_1*p_3 \]
\[ I[2]=-1/9*p_1*p_2+p_0*p_3 \]
\[ I[3]=1/9*p_1^2-1/3*p_0*p_2 \]
\[ I[4]=p_0+p_1+p_2+p_3-1 \]
\[ \text{groebner}(I); \]
\[ _[1]=p_0+p_1+p_2+p_3-1 \]
\[ _[2]=p_2^2-3*p_1*p_3 \]
\[ _[3]=p_1*p_2+9*p_1*p_3+9*p_2*p_3+9*p_3^2-9*p_3 \]
\[ _[4]=p_1^2-18*p_1*p_3-24*p_2*p_3-27*p_3^2-3*p_2+27*p_3 \]

These commands show the declaration of the ideal \( I \) that is the vanishing ideal of the binomial model with 3 trials, as we have seen in Example 3.1.4. Note that the ring \( R \) is declared to be \( \mathbb{Q} \) by the first part of the declaration (which establishes characteristic zero). The final part of the ring declaration \( dp \) establishes the term order for this ring which is the (degree) reverse lexicographic ordering. One important feature to note in \texttt{Singular} is that it is important to always use a Gröbner basis when plugging into a function in \texttt{Singular}, as sometimes it is possible to get the wrong answer. For example, to compute the dimension of the ideal above, the following computation is recommend.

\[ \text{dim(groebner}(I)); \]
1

In \texttt{Singular} it is possible to directly solve systems of polynomial equations with the \texttt{solve} command. For example, if we wanted to find numerical approximations to all points in the variety of the ideal

\[ J = I + \langle p_0^2 + p_1^2 + p_2^2 + p_3^2 - 1 \rangle \]

we would use the sequence of commands

\[ \text{ideal } J = I, p_0^2 + p_1^2 + p_2^2 + p_3^2 - 1; \]
\[ \text{LIB } "\texttt{solve.lib}"; \]
\[ \text{solve(groebner}(J)); \]

The output is long so we suppress it here. It consists of numerical approximations to the six points in the variety \( V(J) \). This only touches on some of the capabilities of \texttt{Singular}, other examples will be given later in the text.

3.6. Projective Space and Projective Varieties

In some problems in algebraic statistics, it is useful to consider varieties as varieties in projective space, rather than as varieties in affine space. While,
to newcomers, the passage to projective space seems to make life more complicated, the primary reason for considering varieties in projective space is that it makes certain things simpler. In algebraic statistics we will have two reasons for passing to projective space to simplify life.

First, projective space is compact (unlike the affine space \( \mathbb{K}^r \)) and various theorems in algebraic geometry require this compactness to be true. A simple example from algebraic geometry is Bezout’s theorem, which says that the intersection of two curves of degree \( d \) and \( e \) in the plane without common components consists of exactly \( de \) points, when counted with multiplicity. This theorem is only true when considered in the projective plane (and over an algebraically closed field). For example, the intersection of two circles in the plane always has 4 intersection points when counted with multiplicity, but two of these are always complex points “at infinity”.

A second reason for passing to projective varieties is aesthetic. When considering statistical models for discrete random variables in projective space, the vanishing ideal tends to look simpler and more easy to interpret. This is especially true when staring at the output of symbolic computation software.

**Definition 3.6.1.** Let \( \mathbb{K} \) be a field. The **projective space** \( \mathbb{P}^{r-1} \) is the set of all lines through the origin in \( \mathbb{K}^r \). Equivalently, it is equal to the set of equivalence classes in \( \mathbb{K}^r \setminus \{0\} \) under the equivalence relation \( a \sim b \) if there is \( \lambda \in \mathbb{K}^\times \) such that \( a = \lambda b \).

The lines or equivalence classes in the definition are the points of the projective space. The equivalence classes that arise in the definition of projective space could be represented by any element of that equivalence class. Note, however, that it does not make sense to evaluate a polynomial \( f \in \mathbb{K}[p] \) at an equivalence class, since, in general \( f(\lambda a) \neq f(a) \) for \( \lambda \in \mathbb{K}^\times \). The exception to this comes when considering homogeneous polynomials.

**Definition 3.6.2.** A polynomial is **homogeneous** if all of its terms have the same degree. An ideal is homogeneous if it has a generating set consisting of homogeneous polynomials.

For example, the polynomial \( p_1^3 p_2 - 17 p_1 p_3 p_4^2 + 12 p_4^4 \) is homogeneous of degree 4. For a homogeneous polynomial, we always have \( f(\lambda a) = \lambda^\deg f f(a) \). Hence, it is well-defined to say that a homogeneous polynomial evaluates to zero at a point of projective space.

**Definition 3.6.3.** Let \( \mathcal{F} \subseteq \mathbb{K}[p] \) be a set of homogeneous polynomials. The **projective variety** defined by \( \mathcal{F} \) is the set

\[
V(\mathcal{F}) = \{ a \in \mathbb{P}^{r-1} : f(a) = 0 \text{ for all } f \in \mathcal{F} \}.
\]
If $I \subseteq \mathbb{K}[p]$ is a homogeneous ideal, then

$$V(I) = \{a \in \mathbb{P}^{r-1} : f(a) = 0 \text{ for all homogeneous } f \in I\}.$$  

If $S \subseteq \mathbb{P}^{r-1}$ its homogeneous vanishing ideal is

$$I(S) = \{f \in \mathbb{K}[p] \text{ homogeneous } : f(a) = 0 \text{ for all } a \in S\}.$$  

The projective space $\mathbb{P}^{r-1}$ is naturally covered by patches that are each isomorphic to the affine space $\mathbb{K}^{r-1}$. Each patch is obtained by taking a nonzero homogeneous linear polynomial $\ell \in \mathbb{K}[p]$, and taking all points in $\mathbb{P}^{r-1}$ where $\ell \neq 0$. This is realized as an affine space, because it is isomorphic to the linear affine variety $V(\ell - 1) \subseteq \mathbb{K}^r$. This suggests that if we are presented with a variety $V \subseteq \mathbb{K}^r$ whose vanishing ideal $I(V)$ contains a linear form $\ell - 1$, we should realize that we are merely looking at an affine piece of a projective variety in $\mathbb{P}^{r-1}$. This issue arises in algebraic statistics when considering statistical models $\mathcal{P}$ for discrete random variables. Since such a model is contained in the probability simplex $\Delta_{r-1}$, and hence $I(\mathcal{P})$ contains the linear polynomial $p_1 + \cdots + p_r - 1$, we should really consider this model as (part of) a patch of a projective variety in $\mathbb{P}^{r-1}$.

Given a radical ideal $I \subseteq \mathbb{K}[p]$, containing a linear polynomial $\ell - 1$, and hence defining a patch of a projective variety, we can use Gröbner bases to compute the homogeneous vanishing ideal of the Zariski closure of $V(I)$ (in projective space).

**Proposition 3.6.4.** Let $I \subseteq \mathbb{K}[p]$ be a radical ideal containing a linear form $\ell - 1$. The Zariski closure of $V(I) \mathbb{P}^{r-1}$ in projective space has homogeneous vanishing ideal

$$J = (I(q) + \langle p_i - tq_i : i = 1, \ldots, r \rangle) \cap \mathbb{K}[p].$$

The notation $I(q)$ means to take the ideal $I \subseteq \mathbb{K}[p]$ and rewrite it in terms of the new variables $q_1, \ldots, q_r$, substituting $q_1$ for $p_1$, $q_2$ for $p_2$, etc. Note that the ideal in parentheses on the right hand side of the equation in Proposition 3.6.4 is in the ring $\mathbb{K}[p, q, t]$ in $2r + 1$ indeterminates.

**Proof.** From our discussion in Section 3.4 on polynomial maps and the connection to elimination ideals, we can see that the ideal

$$I(q) + \langle p_i - tq_i : i = 1, \ldots, r \rangle$$

defines the graph of the map

$$\phi : V(I) \times \mathbb{K} \rightarrow \mathbb{K}^r, (q, t) \mapsto t \cdot q.$$ 

This parametrizes all points on lines through the points in $V(I)$ and the origin. In affine space this describes the cone over variety $V(I)$. The resulting homogeneous ideal $J$ is the vanishing ideal of a projective variety.
When a variety is given parametrically and the image would contain a linear polynomial \( f - 1 \), that parametrization can be homogenized as well, to compute the homogeneous vanishing ideal of the Zariski closure in projective space. This is illustrated below using Macaulay 2, to compute the ideal of the projective closure of the mixture model \( \text{Mixt}^2(\mathcal{P}) \) where \( \mathcal{P} \subseteq \Delta_4 \) is the model for a binomial random variable with 4 trials.

\[
\begin{align*}
i1 & : S = QQ[a1,a2,p,t]; \\
i2 & : R = QQ[p0,p1,p2,p3,p4]; \\
i3 & : f = \text{map}(S,R, \\
& \quad t*(p*(1-a1)^4*a1^0 + (1-p)* (1-a2)^4*a2^0), \\
& \quad t*(4*p*(1-a1)^3*a1^1 + 4*(1-p)* (1-a2)^3*a2^1), \\
& \quad t*(6*p*(1-a1)^2*a1^2 + 6*(1-p) *(1-a2)^2*a2^2), \\
& \quad t*(4*p*(1-a1)^1*a1^3 + 4*(1-p) *(1-a2)^1*a2^3), \\
& \quad t*(p*(1-a1)^0*a1^4 + (1-p) *(1-a2)^0*a2^4));
\end{align*}
\]

\[o3 : \text{RingMap } S \leftarrow R\]

\[i4 : K = \text{kernel } f\]

\[o4 = \text{ideal}(2p2 - 9p1*p2*p3 + 27p0*p3 + 27p1 p4 - 72p0*p2*p4)\]

We see that the generator of this homogeneous ideal is the determinant of the matrix

\[
\begin{pmatrix}
p0 & p1/4 & p2/6 \\
p1/4 & p2/6 & p3/4 \\
p2/6 & p3/4 & p4
\end{pmatrix}.
\]

The passage to projective space and homogeneous polynomials is natural for many problems in algebraic statistics. This perspective will allow us in many situations to ignore the “trivial invariant” \( p_1 + \cdots + p_n - 1 \) from our ideal computations. For example in Chapter 4, it will make things much simpler to work without the trivial invariant as we study conditional independence ideals.

3.7. Exercises

Exercise 3.1. (1) Verify that \( I(W) \) is an ideal for any \( W \subseteq \mathbb{K}^m \).

(2) Show that \( \sqrt{\mathcal{I}} \) is an ideal and that it is radical.

(3) Prove Proposition 3.2.5.
3.7. Exercises

(4) Prove Proposition 3.2.7.

Exercise 3.2. Let $M$ be an $m \times m$ nonsingular matrix of integers. Define an order relation $\prec_M$ on the set of monomials in $\mathbb{K}[p]$ by the rule $p^u \prec_M p^v$ if and only if the first nonzero entry in $M(v - u)$ is positive.

1. Show that if the first row of $M$ consists of nonnegative integers, then $\prec_M$ is a term order.
2. Show that every term order $\prec$ arises as a term order $\prec_M$ for some $M$.
3. Determine the matrices $M_{\text{lex}}$ and $M_{\text{rev}}$ which realize the lexicographic and reverse lexicographic term orders.

Exercise 3.3. (1) Show that if $\text{in}_\prec(I)$ is prime, then $I$ is also prime.
2. Show that if $\text{in}_\prec(I)$ is radical, then $I$ is also radical.
3. Find examples to show that a prime ideal need not have any prime initial ideals and that a radical ideal need not have any radical initial ideals.

Exercise 3.4. (1) Show that the duality between radical ideals and varieties fails if the underlying field $\mathbb{K}$ is not algebraically closed, by exhibiting two radical ideals with the same zero set.
2. An ideal $I$ in $\mathbb{R}[p]$ such that $I(V(I)) = I$ is called a real radical ideal. Show that $I$ and $V$ are inclusion reversing bijections between the set of real radical ideals and the set of real algebraic varieties.

Exercise 3.5. Show that the Hilbert basis theorem implies that for any infinite ascending chain of ideals

$$I_1 \subseteq I_2 \subseteq I_3 \subseteq \cdots \subseteq \mathbb{K}[p]$$

there exist a $j \geq 1$ such that $I_i = I_j$ for all $i > j$. A ring with the property that every infinite ascending sequence of ideals stabilizes is called a Noetherian ring.


Exercise 3.7. Prove Proposition 3.4.9.

Exercise 3.8. (1) Show that the vanishing ideal of the twisted cubic curve is $I = \langle p_1^2 - p_2, p_3^2 - p_5 \rangle$.
2. Show that every point in $V(I)$ is in the image of the parametrization of the twisted cubic.
3. Determine Gröbner bases for $I$ with respect to the lexicographic term order with $p_1 < p_2 < p_3$ and $p_3 < p_2 < p_1$. 
Exercise 3.9.  (1) A Gröbner basis $\mathcal{G}$ is *minimal* if no leading term
of $g \in \mathcal{G}$ divides the leading term of a different $g' \in \mathcal{G}$, and all
the leading coefficients are equal to 1. Explain how to construct a
minimal Gröbner basis of an ideal.

(2) A Gröbner basis $\mathcal{G}$ is *reduced* if it is minimal and no leading term
of $g \in \mathcal{G}$ divides any term of a different $g' \in \mathcal{G}$. Explain how to
construct a reduced Gröbner basis of an ideal.

(3) Show that with a fixed term order, the reduced Gröbner basis of
an ideal $I$ is uniquely determined.

(4) Show that if $I$ is a homogeneous ideal, then any reduced Gröbner
basis of $I$ consists of homogeneous polynomials.
Conditional Independence

Conditional independence constraints are simple and intuitive restrictions on probability densities that allow one to express the notion that two sets of random variables are unrelated, typically given knowledge of the values of a third set of random variables. A conditional independence model is the family of probability distributions that satisfy all the conditional independence constraints in a given set of constraints. In the case where the underlying random variables are discrete or follow a jointly normal distribution, conditional independence constraints are algebraic (polynomial) restrictions on the density function. In particular, they give constraints directly on the joint distribution in the discrete case, and on the covariance matrix in the Gaussian case. A common source of conditional independence models is from graphical modeling. Families of conditional independence constraints that are associated to graphs will be studied in Chapter 13.

In this chapter, we introduce the basic theory of conditional independence. The study of conditional independence structures is a natural place to begin to study the connection between probabilistic concepts and methods from algebraic geometry. In particular, we will tie conditional independence to algebraic statistics through the conditional independence ideals. Section 4.1 introduces conditional independence models, the conditional independence ideals, and the notion of conditional independence implication. In Section 4.2, the algebraic concept of primary decomposition is introduced. The primary decomposition provides an algebraic method to decompose a variety into its irreducible subvarieties. We put special emphasis on the case of primary decomposition of binomial ideals, as this important case arises
in a number of different examples of conditional independence ideals. In Section 4.3, the primary decomposition is applied to study conditional independence models. It is shown in a number of examples how conditional independence implications can be detected using primary decomposition.

4.1. Conditional Independence Models

Let \( X = (X_1, \ldots, X_m) \) be an \( m \) dimensional random vector that takes its values on a Cartesian product space \( \mathcal{X} = \prod_{i=1}^{m} \mathcal{X}_i \). We assume throughout that the joint probability distribution of \( X \) has a density function \( f(x) = f(x_1, \ldots, x_n) \) with respect to a product measure \( \nu \) on \( \mathcal{X} \), and that \( f \) is continuous on \( \mathcal{X} \). In the case where \( \mathcal{X} \) is a finite set, the continuity assumption places no restrictions on the resulting probability distribution.

For each subset \( A \subseteq [m] \), let \( A_n \in A \) be the subvector of \( X \) indexed by the elements of \( A \). Similarly \( \mathcal{X}_A = \prod_{a \in A} \mathcal{X}_a \). Given a partition \( A_1 \mid \cdots \mid A_k \) of \([m]\), we frequently use notation like \( f(x_{A_1}, \cdots, x_{A_k}) \) to denote the function \( f(x) \), but with some variables grouped together. Recall the definitions of marginal and conditional densities from Chapter 2:

**Definition 4.1.1.** Let \( A \subseteq [m] \). The marginal density \( f_A(x_A) \) of \( X_A \) is obtained by integrating out \( x_{[m]\setminus A} \):

\[
f_A(x_A) := \int_{x_{[m]\setminus A}} f(x_A, x_{[m]\setminus A}) \, d\nu_{[m]\setminus A}(x_{[m]\setminus A}), \quad x_A \in \mathcal{X}_A.
\]

Let \( A, B \subseteq [m] \) be disjoint subsets and \( x_B \in \mathcal{X}_B \). The conditional density of \( X_A \) given \( X_B = x_B \) is defined as:

\[
f_{A|B}(x_A \mid x_B) := \begin{cases} \frac{f_{A\cup B}(x_A, x_B)}{f_B(x_B)} & \text{if } f_B(x_B) > 0 \\ 0 & \text{otherwise}. \end{cases}
\]

Note that we use interchangeably various notation that can denote the same marginal or conditional densities, e.g. \( f_{A\cup B}(x_A, x_B) = f_{A\cup B}(x_A; B) \).

**Definition 4.1.2.** Let \( A, B, C \subseteq [m] \) be pairwise disjoint. The random vector \( X_A \) is conditionally independent of \( X_B \) given \( X_C \) if and only if

\[
f_{A; B|C}(x_A, x_B \mid x_C) = f_{A|C}(x_A \mid x_C) \cdot f_{B|C}(x_B \mid x_C)
\]

for all \( x_A, x_B, \) and \( x_C \). The notation \( X_A \perp \!\!\!\!\perp X_B \mid X_C \) is used to denote that the random vector \( X \) satisfies the conditional independence statement that \( X_A \) is conditionally independent of \( X_B \) given \( X_C \). This is often further abbreviated to \( A \perp B \mid C \).

Suppose we have some values \( x_B \) and \( x_C \) such that \( f_{B|C}(x_B \mid x_C) \) is defined and positive, and the conditional independence statement \( X_A \perp \!\!\!\!\perp X_B \mid X_C \)
4.1. Conditional Independence Models

holds. Then we can compute the conditional density $f_{A|B,C}$ as

$$f_{A|B,C}(x_A | x_B, x_C) = \frac{f_{A,B,C}(x_A, x_B, x_C)}{f_{B|C}(x_B | x_C)} = f_{A|C}(x_A | x_C).$$

This consequence of the definition of conditional independence gives a useful way to think about the conditional independence constraint. It can be summarized in plain English as saying "given $X_C$, knowing $X_B$ does not give any information about $X_A$.” This gives us a useful down-to-earth interpretation of conditional independence, which can be easier to digest than the abstract definition.

**Example 4.1.3 (Marginal Independence).** A statement of the form

$$X_A \perp \!\!\!\!\!\!\perp X_B : = X_A \perp X_B | X_C$$

is called a *marginal independence statement*, since it involves no conditioning. It corresponds to a density factorization

$$f_{A,B}(x_A, x_B) = f_A(x_A)f_B(x_B),$$

which should be recognizable as a familiar definition of independence of random variables, as we saw in Chapter 2.

Suppose that a random vector $X$ satisfies a list of conditional independence statements. What other constraints must the same random vector also satisfy? Here we do not assume that we know the density function of $X$ (in which case, we could simply test all conditional independence constraints), rather, we want to know which implications hold regardless of the distribution. Finding such implications is, in general, a challenging problem. It can depend on the type of random variables under consideration (for instance, an implication might be true for all jointly normal random variables but fail for discrete random variables). In general, it is known that it is impossible to find a finite set of axioms from which all conditional independence relations can be deduced [Stu92]. On the other hand, there are a number of easy conditional independence implications which follow directly from the definitions and which are often called the *conditional independence axioms* or *conditional independence inference rules*.

**Proposition 4.1.4.** Let $A, B, C, D \subseteq [m]$ be pairwise disjoint subsets. Then

(i) *(symmetry)* $X_A \perp X_B | X_C \implies X_B \perp X_A | X_C$

(ii) *(decomposition)* $X_A \perp X_{B|D} | X_C \implies X_A \perp X_B | X_C$

(iii) *(weak union)* $X_A \perp X_{B\cup D} | X_C \implies X_A \perp X_B | X_{C\cup D}$

(iv) *(contraction)* $X_A \perp X_B | X_{C\cup D}$ and $X_A \perp X_D | X_C \implies X_A \perp X_{B\cup D} | X_C$
4. Conditional Independence

**Proof.** The proofs of the symmetry axiom follows directly from the commutativity of multiplication. For the decomposition axiom, suppose that the CI statement $X_A \perp X_{B \cup D} \mid X_C$ holds. This is equivalent to the factorization of densities

\[(4.1.1) \quad f_{A \cup B \cup C \cup D}(x_A, x_B, x_D \mid x_C) = f_{A \mid C}(x_A \mid x_C) \cdot f_{B \cup D \mid C}(x_B, x_D \mid x_C).\]

Marginalizing this expression over $X_D$ (that is, integrating both sides of the equation over $X_D$) produces the definition of the conditional independence statement $X_A \perp X_B \mid X_C$ (since $f_{A \mid C}(x_A \mid x_C)$ will factor out of the integral). Similarly, for the weak union axiom, dividing through (4.1.1) by $f_{B \mid C}(x_D \mid x_C)$ (i.e., conditioning on $X_D$) produces the desired conclusion.

For the proof of the contraction axiom, let $x_C$ be such that $f_C(x_C) > 0$. By $X_A \perp X_B \mid X_{C \cup D}$, we have that

\[f_{A \cup B \cup C \cup D}(x_A, x_B \mid x_C, x_D) = f_{A \mid C \cup D}(x_A \mid x_C, x_D) \cdot f_{B \mid C \cup D}(x_B \mid x_C, x_D).\]

Multiplying by $f_{C \cup D}(x_C, x_D)$ we deduce that

\[f_{A \cup B \cup C \cup D}(x_A, x_B, x_C, x_D) = f_{A \mid C \cup D}(x_A, x_C, x_D) \cdot f_{B \mid C \cup D}(x_B \mid x_C, x_D).\]

Dividing by $f_C(x_C) > 0$ we obtain

\[f_{A \cup B \cup D}(x_A, x_B, x_D \mid x_C) = f_{A \mid D}(x_A, x_D \mid x_C) \cdot f_{B \mid D}(x_B \mid x_C, x_D).\]

Applying the conditional independence statement $X_A \perp X_D \mid X_C$, we deduce

\[f_{A \cup B \cup D}(x_A, x_B, x_D \mid x_C) = f_{A \mid C}(x_A \mid x_C) f_{D \mid C}(x_D \mid x_C) f_{B \mid D}(x_B \mid x_C, x_D) = f_{A \mid C}(x_A \mid x_C) \cdot f_{B \mid D}(x_B, x_D \mid x_C),\]

which means that $X_A \perp X_{B \cup D} \mid X_C$.

In addition to the four basic conditional independence “axioms” of Proposition 4.1.4, there is a further fifth “axiom”, which, however, does not hold for every probability distribution but only in special cases.

**Proposition 4.1.5 (Intersection axiom).** Suppose that $f(x) > 0$ for all $x$. Then

\[X_A \perp X_B \mid X_{C \cup D} \text{ and } X_A \perp X_C \mid X_{B \cup D} \implies X_A \perp X_{B \cup C} \mid X_D.\]

**Proof.** The first and second conditional independence statements imply

\[(4.1.2) \quad f_{A \mid B \mid C \cup D}(x_A, x_B \mid x_C, x_D) = f_{A \mid C \cup D}(x_A \mid x_C, x_D) f_{B \mid C \cup D}(x_B \mid x_C, x_D),\]

\[(4.1.3) \quad f_{A \mid C \mid B \cup D}(x_A, x_C \mid x_B, x_D) = f_{A \mid B \mid D}(x_A \mid x_B, x_D) f_{C \mid B \cup D}(x_C \mid x_B, x_D).\]
4.1. Conditional Independence Models

Multiplying (4.1.2) by \( f_{C \cup D}(x_C, x_D) \) and (4.1.3) by \( f_{B \cup D}(x_B, x_D) \), we obtain that

(4.1.4) 
\[
  f_{A \cup B \cup C \cup D}(x_A, x_B, x_C, x_D) = f_{A \cup C \cup D}(x_A \mid x_C, x_D) f_{B \cup C \cup D}(x_B, x_C, x_D),
\]

(4.1.5) 
\[
  f_{A \cup B \cup C \cup D}(x_A, x_B, x_C, x_D) = f_{A \mid B \cup D}(x_A \mid x_B, x_D) f_{B \cup C \cup D}(x_B, x_C, x_D).
\]

Equating (4.1.4) and (4.1.5) and dividing by \( f_{B \cup C \cup D}(x_B, x_C, x_D) \) (which is allowable since \( f(x) > 0 \)) we deduce that

\[
  f_{A \cup C \cup D}(x_A \mid x_C, x_D) = f_{A \mid B \cup D}(x_A \mid x_B, x_D).
\]

Since the right hand side of this expression does not depend on \( x_C \), we conclude

\[
  f_{A \cup C \cup D}(x_A \mid x_C, x_D) = f_{A \mid D}(x_A \mid x_D).
\]

Plugging this into (4.1.4) and conditioning on \( X_D \) gives

\[
  f_{A \cup B \cup C \mid D}(x_A, x_B, x_C \mid x_D) = f_{A \mid D}(x_A \mid x_D) f_{B \cup C \mid D}(x_B, x_C \mid x_D)
\]

and implies that \( X_A \indep X_{B \cup C} \mid X_D \). \( \Box \)

The condition that \( f(x) > 0 \) for all \( x \) is much stronger than necessary for the intersection property to hold. In the proof of Proposition 4.1.5, we only used the condition that \( f_{B \cup C \cup D}(x_B, x_C, x_D) > 0 \). However, it is possible to weaken this condition considerably. In the discrete case, it is possible to give a precise characterization of the conditions on the density which guarantee that the intersection property holds. This will be done in Section 4.3, based on work of Fink [Fin11].

Now we focus on the cases where an algebraic perspective is most useful, namely the case of discrete and Gaussian random variables. In both cases, conditional independence constraints on a random variable correspond to algebraic constraints on the distribution.

Let \( X = (X_1, \ldots, X_m) \) be a vector of discrete random variables. Returning to the notation used in previous chapters, we let \([r_j]\) be the set of values taken by \( X_j \). Then \( X \) takes its values in \( R = \prod_{j=1}^m [r_j] \). For \( A \subseteq [m] \), \( R_A = \prod_{a \in A} [r_a] \). In this discrete setting, a conditional independence constraint translates into a system of quadratic polynomial equations in the joint probability distribution.

**Proposition 4.1.6.** If \( X \) is a discrete random vector, then the conditional independence statement \( X_A \indep X_B \mid X_C \) holds if and only if

(4.1.6) 
\[
  p_{i_A \cdot B \cdot i_C} \cdot p_{j_A \cdot j_B \cdot i_C} - p_{i_A \cdot i_B \cdot i_C} \cdot p_{j_A \cdot j_B \cdot i_C} = 0
\]

for all \( i_A, j_A \in R_A \), \( i_B, j_B \in R_B \), and \( i_C \in R_C \).
4. Conditional Independence

The notation $p_{i_A, i_B, i_C}$ denotes the probability $P(X_A = i_A, X_B = i_B, X_C = i_C)$. As a polynomial expression in terms of indeterminates $p_i$ where $i \in \mathcal{R}$, this can be written as

$$p_{i_A, i_B, i_C} = \sum_{j[m] \setminus A \cup B \cup C \in \mathcal{R}[m] \setminus A \cup B \cup C} p_{i_A, i_B, i_C \setminus j[m] \setminus A \cup B \cup C}.$$ 

Here we use the usual convention that the string of indices $i$ has been regrouped into the string $(i_A, i_B, i_C \setminus j[m] \setminus A \cup B \cup C)$, for convenience.

**Proof.** By marginalizing, we may assume that $A \cup B \cup C = [m]$. By conditioning, we may assume that $C$ is empty. By aggregating the states indexed by $\mathcal{R}_A$ and $\mathcal{R}_B$ respectively, we can treat this as a conditional independence statement $X_1 \perp \! \! \! \perp X_2$. In this setting, the conditional independence constraints in (4.1.6) are equivalent to saying that all $2 \times 2$ minors of the matrix

$$P = \begin{pmatrix}
p_{11} & p_{12} & \cdots & p_{1r_2} \\
p_{21} & p_{22} & \cdots & p_{2r_2} \\
\vdots & \vdots & \ddots & \vdots \\
p_{r_11} & p_{r_12} & \cdots & p_{r_1r_2}
\end{pmatrix}$$

are zero. Since $P$ is not the zero matrix, this means that $P$ has rank 1. Thus $P = \alpha^T \beta$ for some vectors $\alpha \in \mathbb{R}^{r_1}$ and $\beta \in \mathbb{R}^{r_2}$. These vectors can be taken to be nonnegative vectors, and such that $\|\alpha\|_1 = \|\beta\|_1 = 1$. This implies that $p_{ij} = p_i + p_j$ for all $i$ and $j$, and hence $X_1 \perp \! \! \! \perp X_2$. Conversely, if $X_1 \perp \! \! \! \perp X_2$, $P$ must be a rank one matrix. \hfill \square

**Definition 4.1.7.** The conditional independence ideal $I_{A \perp \! \! \! \perp B | C} \subseteq \mathbb{R}[p]$ is generated by all the quadratic polynomials in (4.1.6). The discrete conditional independence model $\mathcal{M}_{A \perp \! \! \! \perp B | C} = V_\Delta(I_{A \perp \! \! \! \perp B | C}) \subseteq \Delta_\mathbb{R}$ is the set of all probability distributions in $\Delta_\mathbb{R}$ satisfying the quadratic polynomial constraints (4.1.6).

In fact, it can be shown that the ideal $I_{A \perp \! \! \! \perp B | C}$ is a prime ideal, which we will prove in Chapter 9.

**Example 4.1.8 (Simplest Conditional Independence Ideals).** Let $m = 2$, and consider the (ordinary) independence statement $X_1 \perp \! \! \! \perp X_2$. Then

$$I_{1 \perp 2} = \langle p_{i_1 j_1} p_{i_2 j_2} - p_{i_1 j_2} p_{i_2 j_1} : i_1, i_2 \in [r_1], j_1, j_2 \in [r_2] \rangle.$$ 

This is the ideal of 2 minors in the proof of Proposition 4.1.6.

If $m = 3$, and we consider the marginal independence statement $X_1 \perp \! \! \! \perp X_2$, and suppose that $r_3 = 2$, then

$$I_{1 \perp 2} = \langle (p_{i_1 j_1} + p_{i_1 j_2}) (p_{j_1 j_2} + p_{j_1 j_2}) - (p_{i_1 j_2} + p_{i_1 j_2}) (p_{j_1 i_2} + p_{j_1 i_2}) : i_1, j_1 \in [r_1], i_2, j_2 \in [r_2] \rangle.$$
If \( m = 3 \), and we consider the conditional independence statement \( X_2 \perp X_3 | X_1 \), then
\[
I_{2 \perp 3 | 1} = (p_{i_1 i_2 i_3} p_{i_1 j_2 j_3} - p_{i_1 i_2 j_3} p_{i_1 j_2 j_3} : i_1 \in [r_1], i_2, j_2 \in [r_2], i_3, j_3 \in [r_3]).
\]

If \( C = \{ A_1 \perp B_1 | C_1, A_2 \perp B_2 | C_2, \ldots \} \) is a set of conditional independence statements, we construct the conditional independence ideal
\[
I_C = \sum_{A \perp B | C \in C} I_{A \perp B | C}
\]
which is generated by all the quadratic polynomials appearing in any of the \( I_{A_i \perp B_i | C_i} \). The model \( M_C := V_\Delta (I_C) \subseteq \Delta_R \) consists of all probability distributions satisfying all the conditional independence constraints arising from \( C \).

Conditional independence for Gaussian random vectors is also an algebraic condition. In particular, the conditional independence statement \( X_A \perp X_B | X_C \) gives a rank condition on the covariance matrix \( \Sigma \). Since rank conditions are algebraic (described by vanishing determinants), the resulting conditional independence models can also be analyzed using algebraic techniques.

**Proposition 4.1.9.** The conditional independence statement \( X_A \perp X_B | X_C \) holds for a multivariate normal random vector \( X \sim \mathcal{N}(\mu, \Sigma) \) if and only if the submatrix \( \Sigma_{A \setminus C, B \setminus C} \) of the covariance matrix \( \Sigma \) has rank \( \# C \).

**Proof.** If \( X \sim \mathcal{N}(\mu, \Sigma) \) follows a multivariate normal distribution, then the conditional distribution of \( X_{A \setminus B} \) given \( X_C = x_c \) is the multivariate normal distribution
\[
\mathcal{N} \left( \mu_{A \setminus B} + \Sigma_{A \setminus B, C} \Sigma_{C, C}^{-1} (x_C - \mu_C), \Sigma_{A \setminus B, A \setminus B} - \Sigma_{A \setminus B, C} \Sigma_{C, C}^{-1} \Sigma_{C, A \setminus B} \right),
\]
by Theorem 2.4.2. The statement \( X_A \perp X_B | X_C \) holds if and only if
\[
(\Sigma_{A \setminus B, A \setminus B} - \Sigma_{A \setminus B, C} \Sigma_{C, C}^{-1} \Sigma_{C, A \setminus B})_{A \setminus B} = \Sigma_{A, B} - \Sigma_{A, C} \Sigma_{C, C}^{-1} \Sigma_{C, B} = 0
\]
by Proposition 2.4.4. The matrix \( \Sigma_{A, B} - \Sigma_{A, C} \Sigma_{C, C}^{-1} \Sigma_{C, B} \) is the Schur complement of the matrix
\[
\Sigma_{A \setminus C, B \setminus C} = \begin{pmatrix} \Sigma_{A, B} & \Sigma_{A, C} \\ \Sigma_{C, B} & \Sigma_{C, C} \end{pmatrix}.
\]
Since \( \Sigma_{C, C} \) is always invertible (it is positive definite), the Schur complement is zero if and only if the matrix \( \Sigma_{A \setminus C, B \setminus C} \) has rank equal to \( \# C \). \( \square \)

The set of partially symmetric matrices with rank \( \leq k \) is an irreducible variety, defined by the vanishing of all \( (k+1) \times (k+1) \) subdeterminants. The ideal generated by these subdeterminants is a prime ideal \([\text{Con94}]\). Hence, we obtain nice families of conditional independence ideals.
Definition 4.1.10. Fix pairwise disjoint subsets $A, B, C$ of $[m]$. The Gaussian conditional independence ideal $J_{A \perp B \mid C}$ is the following ideal in $\mathbb{R}[\sigma_{ij}, 1 \leq i \leq j \leq m]$:

$$J_{A \perp B \mid C} = \langle (\#C + 1) \times (\#C + 1) \text{ minors of } \Sigma_{A \cup C, B \cup C} \rangle$$

Similarly as with the discrete case, we can take a collection $C$ of conditional independence constraints, and get a conditional independence ideal $J_C$.

$$J_C = \sum_{A \perp B \mid C \in C} J_{A \perp B \mid C}.$$ 

The resulting Gaussian conditional independence model is a subset of the cone $PD_m$ of $m \times m$ symmetric positive definite matrices:

$$\mathcal{M}_C := V(J_C) \cap PD_m.$$ 

We use the notation $PD_m$ to denote the set of all $m \times m$ symmetric positive definite matrices.

Our goal in the remains of the chapter is to show how algebra can be used to study conditional independence models, and in particular, conditional independence implications. To illustrate the basic idea (and avoiding any advanced algebra), consider the following example.

Example 4.1.11 (Conditional and Marginal Independence). Let $C = \{1 \perp 3, 1 \perp 3 | 2\}$. The Gaussian conditional independence ideal $J_C$ is generated by two polynomials

$$J_C = \langle \sigma_{13}, \det \Sigma_{\{1,2\},\{2,3\}} \rangle.$$ 

The conditional independence model consists of all covariance matrices $\Sigma \in PD_3$ satisfying

$$\sigma_{13} = 0 \quad \text{and} \quad \sigma_{13}\sigma_{22} - \sigma_{12}\sigma_{23} = 0.$$ 

This is equivalent to the system

$$\sigma_{13} = 0 \quad \text{and} \quad \sigma_{12}\sigma_{23} = 0,$$

which splits as the union of two linear spaces:

$$L_1 = \{ \Sigma : \sigma_{13} = \sigma_{12} = 0 \} \quad \text{and} \quad L_2 = \{ \Sigma : \sigma_{13} = \sigma_{23} = 0 \}.$$ 

In other words, $V(J_C) = L_1 \cup L_2$. The first component $L_1$ consists of Gaussian random vectors satisfying $X_1 \perp (X_2, X_3)$, whereas the second component consists of Gaussian random vectors satisfying $X_3 \perp (X_1, X_2)$. It follows that the implication

$$X_1 \perp X_3 \mid X_2 \quad \text{and} \quad X_1 \perp X_3 \implies X_1 \perp (X_2, X_3) \quad \text{or} \quad (X_1, X_2) \perp X_3,$$

holds for multivariate normal random vectors.
4.2. Primary Decomposition

In Example 4.1.11, in the previous section, we saw the solution set to a system of polynomial equations break up into two simpler pieces. In this section, we discuss the algebraic concept which generalizes this notion. On the geometric side, any variety will break up as the union of finitely many irreducible subvarieties. On the more algebraic level of ideals, any ideal can be written as the intersection of finitely many primary ideals. This representation is called a primary decomposition. We provide a brief introduction to this material here. Further details on primary decomposition, including missing proofs of results in this section can be found in standard texts on algebraic geometry (e.g. [CLO07], [Has07]).

Definition 4.2.1. A variety $V$ is called reducible if there are two proper subvarieties $V_1, V_2$ of $V$ such that $V = V_1 \cup V_2$. A variety $V$ that is not reducible is called irreducible. An ideal $I \subseteq \mathbb{K}[p]$ is prime if there are no two elements $f_1, f_2 / I$ such that $f_1 f_2 \in I$.

Proposition 4.2.2. A variety $V$ is irreducible if and only if the corresponding ideal $I(V)$ is a prime ideal.

Proof. Suppose $V$ is reducible and let $V = V_1 \cup V_2$ where $V_1$ and $V_2$ are proper subvarieties of $V$. Since $V_1 \neq V_2$, there are polynomials $f_1 \in I(V_1) \setminus I(V_2)$ and $f_2 \in I(V_2) \setminus I(V_1)$. Then $I(V)$ is not prime since $f_1, f_2 \notin I(V) = I(V_1 \cup V_2) = I(V_1) \cap I(V_2)$, but $f_1 f_2 \in I(V)$. Conversely, if $I(V)$ is not prime, with $f_1, f_2 \notin I(V)$ but $f_1 f_2 \in I(V)$, then $V = V(I(V) + (f_1)) \cup V(I(V) + (f_2))$ is a proper decomposition of $V$.

Irreducible varieties are the basic building blocks for solution sets of polynomial equations, and every variety has a finite decomposition into irreducibles.

Proposition 4.2.3. Every variety $V$ has a unique decomposition $V = V_1 \cup \cdots \cup V_k$ into finitely many irreducible varieties $V_i$, such that no $V_i$ can be omitted from the decomposition. This decomposition is called the irreducible decomposition of the variety $V$.

Proof. To show existence of such a decomposition, consider the following decomposition “algorithm” for a variety. Suppose at some stage we have a representation $V = V_1 \cup \cdots \cup V_k$ as the union of varieties. If all the $V_i$ are irreducible, we have an irreducible decomposition. Suppose, on the other hand, that some component, say $V_1$ is reducible, i.e. it has a proper decomposition as $V_1 = V'_1 \cup V'_2$. Then we augment the decomposition as $V = V'_1 \cup V'_2 \cup V_2 \cup \cdots \cup V_k$ and repeat. This algorithm either terminates in finitely many steps with an irreducible decomposition (which can be subsequently
made irredundant), or there is an infinite sequence of varieties among those produced along the way such that

\[ W_1 \supset W_2 \supset \cdots . \]

This, in turn, implies an infinite ascending chain of ideals

\[ \mathcal{I}(W_1) \subset \mathcal{I}(W_2) \subset \cdots , \]

violating the Noetherianity of the polynomial ring (see Exercise 3.5 for the definition of a Noetherian ring). Uniqueness of the decomposition is deduced by comparing two different irreducible decompositions. \( \square \)

By the inclusion-reversing property of passing from ideals to varieties, we are naturally led to define an irreducible ideal to be one which has no proper decomposition \( I = I_1 \cap I_2 \). The same argument as in the proof of Proposition 4.2.3 shows that every polynomial ideal has a decomposition into irreducible ideals (uniqueness of the decomposition is not guaranteed however). The irreducible decomposition however, is not as useful as another decomposition, the primary decomposition, which we describe now.

**Definition 4.2.4.** An ideal \( Q \) is called primary if for all \( f \) and \( g \) such that \( fg \in Q \), either \( f \in Q \) or \( g^k \in Q \) for some \( k \).

The next proposition contains some basic facts about primary ideals.

**Proposition 4.2.5.**

1. Every irreducible ideal is primary.
2. If \( Q \) is primary, then \( \sqrt{Q} = P \) is a prime ideal, called the associated prime of \( Q \).
3. If \( Q_1 \) and \( Q_2 \) are primary, and \( \sqrt{Q_1} = \sqrt{Q_2} = P \), then \( Q_1 \cap Q_2 \) is also primary, with associated prime \( P \).

**Proof.** The proof of part (1) depends on the notion of quotient ideals, defined later in this section, and will be left as an exercise.

For part (2), suppose that \( Q \) is primary, and let \( P = \sqrt{Q} \), and suppose that \( fg \in P \). Then \( (fg)^n \in Q \) for some \( n \) since \( P \) is the radical of \( Q \). Since \( Q \) is primary either \( f^n \in Q \) or there is a \( k \) such that \( g^k \in Q \). But this implies that either \( f \in P \) or \( g \in P \), hence \( P \) is prime.

For part (3), suppose that \( Q_1 \) and \( Q_2 \) are primary with \( \sqrt{Q_1} = \sqrt{Q_2} = P \) and let \( fg \in Q_1 \cap Q_2 \) and suppose that \( f \notin Q_1 \cap Q_2 \). Without loss of generality we can suppose that \( f \notin Q_1 \). Since \( Q_1 \) is primary, there is an \( n \) such that \( g^n \in Q_1 \). This implies that \( g \in P \). Then there exists an \( m \) such that \( g^m \in Q_2 \), so \( g^{\text{max}(n,m)} \in Q_1 \cap Q_2 \), which is primary. Finally, \( \sqrt{Q_1 \cap Q_2} = \sqrt{Q_1} \cap \sqrt{Q_2} \) for any ideals, which implies that \( P \) is the associated prime of \( Q_1 \cap Q_2 \). \( \square \)
Definition 4.2.6. Let $I \subseteq \mathbb{K}[p]$ be an ideal. A primary decomposition of $I$ is a representation $I = Q_1 \cap \cdots \cap Q_r$ of $I$ as the intersection of finitely primary ideals. The decomposition is irredundant if no $Q_i \supseteq \cap_{j \neq i} Q_j$ and all the associated primes $\sqrt{Q_i}$ are distinct.

By Proposition 4.2.5 part one and the existence of irreducible decompositions of ideals, we see that every ideal $I$ has a finite irredundant primary decomposition. A primary decomposition is not unique in general, but it does have many uniqueness properties which we outline in the following theorem.

Theorem 4.2.7. Let $I$ be an ideal in a polynomial ring.

1. Then $I$ has an irredundant primary decomposition $I = Q_1 \cap \cdots \cap Q_k$ such that the prime ideals $\sqrt{Q_i}$ are all distinct.

2. The set of prime ideals $\text{Ass}(I) := \{\sqrt{Q_i} : i \in [k]\}$ is uniquely determined by $I$, and does not depend on the particular irredundant primary decomposition. They are called the associated primes of $I$.

3. The minimal elements of $\text{Ass}(I)$ with respect to inclusion are called minimal primes of $I$. The primary components $Q_i$ associated to minimal primes of $I$ are uniquely determined by $I$.

4. The minimal primes of $I$ are precisely the minimal elements among all prime ideals that contain the ideal $I$.

5. Associated primes of $I$ that are not minimal are called embedded primes. The primary components $Q_i$ associated to embedded primes of $I$ are not uniquely determined by $I$.

6. $\sqrt{I} = \cap \sqrt{Q_i}$. If the intersection is taken over the minimal primes of $I$, this gives a primary (in fact, prime) decomposition of $\sqrt{I}$.

7. The decomposition $V(I) = \cup V(\sqrt{Q_i})$, where the union is taken over the minimal primes of $I$ is the irreducible decomposition of the variety $V(I)$.

A proof of the many parts of Theorem 4.2.7 can be found in [Has07].

We saw in Example 4.1.11 a simple irreducible decomposition of a variety which corresponds to a primary decomposition of an ideal. In particular:

$$ J_C = \langle \sigma_{13}, \det \Sigma_{\{1,2\},\{2,3\}} \rangle = \langle \sigma_{12}, \sigma_{13} \rangle \cap \langle \sigma_{13}, \sigma_{23} \rangle. $$

Note that both primary components are prime in this case, in fact, they are the minimal primes, and so all components are uniquely determined.

For an example with embedded primes consider the ideal:

$$ \langle p_1^2, p_1 p_2 \rangle = \langle p_1 \rangle \cap \langle p_1^2, p_2 \rangle. $$
Both ideals in the decomposition are primary, and the decomposition is irredundant. The associated primes are \( \langle p_1 \rangle \) and \( \langle p_1, p_2 \rangle \), the first is minimal, and the second is embedded. Furthermore, the decomposition is not unique since,
\[
\langle p_1^2, p_1p_2 \rangle = \langle p_1 \rangle \cap \langle p_1^2, p_2 + cp_1 \rangle
\]
is also a primary decomposition for any \( c \in \mathbb{C} \).

Important operations related to primary decompositions are the ideal quotient and the saturation.

**Definition 4.2.8.** Let \( I, J \) be ideals in \( \mathbb{K}[p] \). The quotient ideal (or colon ideal) is the ideal
\[
I : J := \{ f \in \mathbb{K}[p] : fg \in I \text{ for all } g \in J \}.
\]
The saturation of \( I \) by \( J \) is the ideal
\[
I : J^\infty := \bigcup_{k=1}^\infty I : J^k.
\]

If \( J = \langle f \rangle \) is a principal ideal, then we use the shorthand \( I : f \) to denote \( I : \langle f \rangle \) and \( I : f^\infty \) to denote \( I : \langle f \rangle^\infty \). The quotient operations allows to detect the associated primes of an ideal.

**Proposition 4.2.9.** Let \( I \subseteq \mathbb{K}[p] \) be an ideal. Then for each prime \( P \in \text{Ass}(I) \) there exists a polynomial \( f \in \mathbb{K}[p] \) such that \( P = I : f \).

See [CLO07] or [Has07] for a proof. For example, for the ideal \( \langle p_1^2, p_1p_2 \rangle \) we have
\[
\langle p_1^2, p_1p_2 \rangle : p_1 = \langle p_1, p_2 \rangle \quad \text{and} \quad \langle p_1^2, p_1p_2 \rangle : p_2 = \langle p_1 \rangle.
\]
The geometric content of the quotient ideal and saturation is that they correspond to removing components.

**Proposition 4.2.10.** Let \( V \) and \( W \) be varieties in \( \mathbb{K}^r \). Then
\[
I(V) : I(W) = I(V \setminus W).
\]
If \( I \) and \( J \) are ideals of \( \mathbb{K}[p] \), then
\[
V(I : J^\infty) = \overline{V(I) \setminus V(J)}.
\]
The proof of Proposition 4.2.10 is left as an exercise. For reasonably sized examples, it is possible to compute primary decompositions in computer algebra packages. Such computations can often provide insight in specific examples, and give ideas on how to prove theorems about the decomposition in complicated families. We will see examples in Section 4.3.

The best possible case for finding a primary decomposition of an ideal \( I \) is when \( I \) is a radical ideal. This is because all the information in the
primary decomposition is geometric (only the minimal primes come into play). Sometimes, it might be possible to prove directly that an ideal is radical, then use geometric arguments to discover all minimal primes, and thus deduce the prime decomposition of the ideal. For example, the following result on initial ideals can be used in some instances:

**Proposition 4.2.11.** Let $I$ be an ideal and $<$ a term order, so that $\text{in}_<(I)$ is a radical ideal. Then $I$ is also a radical ideal.

**Proof.** Suppose that in$_<(I)$ is radical and suppose that $f \in \mathbb{K}[p]$ is a polynomial such that $f \notin I$ but $f^k \in I$ for some $k > 1$. We can further suppose that the leading term of $f$ is minimal in the term order $<$, among all $f$ with this property. Since in$_<(I)$ is radical, there is a $g \in I$ with in$_<(f) = \text{in}_<(g)$. Then $f - g$ has the property that in$_<(f - g)$ is smaller than in$_<(f)$, but $(f - g)^k \in I$, so $f - g \in I$ by our minimality assumption on in$_<(f)$. But since $f - g$ and $g$ are in $I$, so is $f$ contradicting our assumption that $f \notin I$. $\square$

An important and relatively well-understood class of ideals from the standpoint of primary decompositions is the class of binomial ideals. A binomial is a polynomial of the form $p^u - p^v$ where $\mathbb{K}$. An ideal is a binomial ideal if it has a generating set consisting of binomials. Binomial ideals make frequent appearances in algebraic statistics, especially because of their connections to random walks, as discussed in Chapter 9. The primary decomposition of binomial ideals are among the best behaved.

**Theorem 4.2.12.** Let $I$ be a binomial ideal and suppose that $\mathbb{K}$ is an algebraically closed field. Then $I$ has an irredundant primary decomposition such that all primary components are binomial ideals. Furthermore, the radical of $I$ and all associated prime ideals of $I$ are binomial ideals.

Theorem 4.2.12 was first proved in [ES96], which also includes algorithms for computing the primary decomposition of binomial ideals. However, it remains an active research problem to find better algorithms for binomial primary decomposition, and to actually make those algorithms usable in practice. See [DMM10, Kah10, KM14, KMO16] for recent work on this problem.

In some problems in algebraic statistics, we are only interested in the geometric content of the primary decomposition, and hence, on the minimal primes. In the case of binomials ideals, it is often possible to describe this quite explicitly in terms of lattices.

**Definition 4.2.13.** A pure difference binomial is a binomial of the form $p^u - p^v$. A pure difference binomial ideal is generated by pure difference binomials. Let $\mathcal{L} \subseteq \mathbb{Z}^r$ be a lattice. The lattice ideal $I_\mathcal{L}$ associated to $\mathcal{L}$ is
Conditional Independence

the ideal 
\[ (p^u - p^v : u - v \in \mathcal{L}). \]

The special case where \( \mathcal{L} \) is a saturated lattice (that is \( \mathbb{Z}^r / \mathcal{L} \) has no torsion) is called a toric ideal.

Toric ideals and lattice ideals will be elaborated upon in more detail in Chapters 6 and 9. Note that toric ideals are prime over any field. Lattice ideals are usually not prime if the corresponding lattice is not saturated. In the case where \( K \) is an algebraically closed field of characteristic 0 (e.g. \( \mathbb{C} \)), lattice ideals are always radical, and they have a simple decomposition into prime ideals in terms of the torsion subgroup of \( \mathbb{Z}^r / \mathcal{L} \). Because of this property, when we want to find the minimal primes of a pure difference binomial ideal over \( \mathbb{C} \), it is useful to look at certain associated lattice ideals (or rather, ideals that are the sum of a lattice ideal and prime monomial ideal). We will call these associated lattice ideals, for simplicity.

Let \( \mathcal{B} \) be a set of pure difference binomials and \( I = (\mathcal{B}) \). The lattice ideals that contain \( I \) and can be associated to \( I \) have a restricted form.

Consider the polynomial ring \( \mathbb{C}[p] = \mathbb{C}[p_1, \ldots, p_r] \). Let \( S \subseteq [r] \). Let \( \mathbb{C}[p_S] := \mathbb{C}[p_s : s \in S] \). Form the ideal

\[ I_S := \langle p_s : s \in S \rangle + \langle p^u - p^v : p^u - p^v \in \mathcal{B} \cap \mathbb{C}[p_{[r] \setminus S}] \rangle : (\prod_{s \in [r] \setminus S} p_s)^\infty. \]

We say that \( S \) is a candidate set for a minimal associated lattice ideal if for each binomial generator of \( I \), \( p^u - p^v \) either \( \langle p^u - p^v : p^u - p^v \in \mathcal{B} \cap \mathbb{C}[p_{[r] \setminus S}] \rangle \) or there are \( s, s' \in S \) such that \( p_s | p^u \) and \( p_{s'} | p^v \). Note that the ideal

\[ \langle p^u - p^v : p^u - p^v \in \mathcal{B} \cap \mathbb{C}[p_{[r] \setminus S}] \rangle : (\prod_{s \in [r] \setminus S} p_s)^\infty \]

is equal to the lattice ideal \( I_\mathcal{L} \) for the lattice \( \mathcal{L} = \mathbb{Z}\{u - v : p^u - p^v \in \mathcal{B} \cap \mathbb{C}[p_{[r] \setminus S}]\} \).

**Proposition 4.2.14.** Every minimal associated lattice ideal of the pure difference binomial ideal \( I \) has the form \( I_S \) for a candidate set \( S \). In particular, every minimal prime of \( I \) is the minimal prime of one of the ideals \( I_S \), for some \( S \).

**Proof Sketch.** The description of components in Proposition 4.2.14 is essentially concerned with finding which components of the variety \( V(I) \) are contained in each coordinate subspace. Indeed, it can be shown the if \( S \subseteq [r] \), then

\[ I_S = I(V(I) \cap V(\langle p_s : s \in S \rangle)) \]

This result is obtained, in particular, by using geometric properties of the saturation as in Proposition 4.2.10. \( \square \)
4.2. Primary Decomposition

Note that the candidate condition is precisely the condition that guarantees that \( I \subseteq I_\mathcal{L} \). The minimal associated lattice ideals are just the minimal elements among the candidate associated lattice ideals. Hence, solving the geometric problem of finding irreducible components is a combinatorial problem about when certain lattice ideals contain others (plus the general problem of decomposing lattice ideals).

**Example 4.2.15 (Adjacent Minors).** Let \( I \) be the pure difference binomial ideal

\[
I = \langle p_{11}p_{22} - p_{12}p_{21}, p_{12}p_{23} - p_{13}p_{22} \rangle.
\]

Note that \( I \) is generated by two of the three \( 2 \times 2 \) minors of the \( 2 \times 3 \) matrix

\[
\begin{pmatrix}
p_{11} & p_{12} & p_{13} \\
p_{21} & p_{22} & p_{23}
\end{pmatrix}.
\]

The candidate sets of variables must consist of sets of variables that intersect both the leading and trailing term of any binomial in which it has variables in common. The possible candidate sets for \( I \) are:

\[
\emptyset, \{p_{11}, p_{21}\}, \{p_{12}, p_{22}\}, \{p_{13}, p_{23}\}, \{p_{11}, p_{12}, p_{13}\}, \{p_{21}, p_{22}, p_{23}\}
\]

and any set that can be obtained from these by taking unions. The candidate ideal \( I_\emptyset \) is equal to

\[
I : (p_1 \cdots p_{23})^\infty = \langle p_{11}p_{22} - p_{12}p_{21}, p_{12}p_{23} - p_{13}p_{22}, p_{11}p_{23} - p_{13}p_{21} \rangle.
\]

The candidate ideal \( I_{\{p_{12}, p_{22}\}} \) is the monomial ideal

\[
I_{\{p_{12}, p_{22}\}} = \langle p_{12}, p_{22} \rangle.
\]

The ideal \( I_{\{p_{11}, p_{21}\}} \) is the ideal

\[
I_{\{p_{11}, p_{21}\}} = \langle p_{11}, p_{21}, p_{12}p_{23} - p_{13}p_{22} \rangle.
\]

Note, however, that \( I_\emptyset \subseteq I_{\{p_{11}, p_{21}\}} \), so \( I_{\{p_{11}, p_{21}\}} \) is not a minimal prime of \( I \). In fact,

\[
I = I_\emptyset \cap I_{\{p_{12}, p_{22}\}} = \langle p_{11}p_{22} - p_{12}p_{21}, p_{12}p_{23} - p_{13}p_{22}, p_{11}p_{23} - p_{13}p_{21} \rangle \cap \langle p_{12}, p_{22} \rangle
\]

is an irredundant primary decomposition of \( I \), showing that all of the other candidate sets are superfluous.

The variety of a lattice ideal \( I_\mathcal{L} \) in the interior of the positive orthant \( \mathbb{R}^r_{>0} \) is easy to describe. It consists of all points \( p \in \mathbb{R}^r_{>0} \), such that \( (\log p)^T u = 0 \) for all \( u \in \mathcal{L} \). Hence, Proposition 4.2.14 says that the variety of a pure difference binomial ideal in the probability simplex \( \Delta_{r-1} \) is the union of exponential families (see Chapter 6).
4.3. Primary Decomposition of CI Ideals

In this section, we compute primary decompositions of the conditional independence ideals $I_C$ and $J_C$ and use them to study conditional independence implications. First we illustrate the idea with some small examples, explaining how these examples can be computed in some computational algebra software. Then we show a number of case studies where theoretical descriptions of the primary decompositions (or information about the primary decomposition, without actually obtaining the entire decomposition) can be used to study conditional independence implications. We try to exploit the binomial structure of these conditional independence ideals whenever possible. We begin with two basic examples that illustrates some of the subtleties that arise.

Example 4.3.1 (Gaussian Contraction Axiom). Let $C = \{1 \perp 2 | 3, 2 \perp 3\}$. By the contraction axiom these two statements imply that $2 \perp \{1, 3\}$. Let us consider this in the Gaussian case. We have

$$J_C = \langle \det \Sigma_{\{1,3\},\{2,3\}}, \sigma_{23} \rangle$$

$$= \langle \sigma_{12} \sigma_{33} - \sigma_{13} \sigma_{23}, \sigma_{23} \rangle$$

$$= \langle \sigma_{12} \sigma_{33}, \sigma_{23} \rangle$$

$$= \langle \sigma_{12}, \sigma_{23} \rangle \cap \langle \sigma_{33}, \sigma_{23} \rangle.$$ 

In particular, the variety defined by the ideal $J_C$ has two components. However, only one of these components is interesting from the probabilistic standpoint. Indeed, the second component $V(\langle \sigma_{33}, \sigma_{23} \rangle)$ does not intersect the cone of positive definite matrices. It is the first component which contains the conclusion of the contraction axiom and which is relevant from a probabilistic standpoint.

Example 4.3.2 (Binary Contraction Axiom). Let $C = \{1 \perp 2 | 3, 2 \perp 3\}$. By the contraction axiom these two statements imply that $2 \perp \{1, 3\}$. Let us consider this in the case of binary random variables. We have

$$I_C = \langle p_{111}p_{221} - p_{121}p_{211}, p_{112}p_{222} - p_{122}p_{212},$$

$$(p_{111} + p_{211})(p_{122} + p_{222}) - (p_{112} + p_{212})(p_{121} + p_{221}) \rangle.$$

The following code computes the primary decomposition of this example in Singular [DGPS16]:

```plaintext
ring R = 0,(p111,p112,p121,p122,p211,p212,p221,p222),dp;
ideal I =
p111*p221 - p121*p211, p112*p222 - p122*p212,
(p111 + p211)*(p122 + p222) - (p112 + p212)*(p121 + p221);
LIB "primdec.lib";
primdecGTZ(I);```

The computation reveals that the conditional independence ideal $I_C$ is radical with three prime components:

$$I_C = I_{2 \perp \{1,3\}} \cap (p_{112} + p_{212}, p_{111} + p_{211}, p_{121} + p_{221})$$

Note that while the last two components do intersect the probability simplex, their intersection with the probability simplex is contained within the first component’s intersection with the probability simplex. For instance, we consider the second component

$$K = (p_{112} + p_{212}, p_{111} + p_{211}, p_{121} + p_{221})$$

since all probabilities are nonnegative, the requirements that, $p_{111} + p_{211} = 0$ and $p_{121} + p_{221} = 0$ forces both $p_{111} = 0$, $p_{211} = 0$, $p_{121} = 0$ and $p_{221} = 0$. Hence

$$V(K) \cap \Delta_7 \subseteq V(I_{2 \perp \{1,3\}}) \cap \Delta_7.$$ 

Hence, the primary decomposition and some simple probabilistic reasoning verifies the conditional independence implication.

A complete analysis of the ideal of the discrete contraction axiom was carried out in [GPSS05]. In both of the preceding examples it is relatively easy to derive these conditional independence implications directly, without using algebra. The point of the examples is familiarity with the idea of using primary decomposition. The conditional independence implications and their proofs become significantly more complicated once we move to more involved examples in the following subsections.

4.3.1. Failure of the Intersection Axiom. Here we follow Section 6.6 of [DSS09] and [Fin11] to analyze the failure of the intersection axiom for discrete random variables. Looked at another way, we give the strongest possible conditions based solely on the zero pattern in the distribution which imply the conclusion of the intersection axiom.

In the discrete case, it suffices to analyze the intersection axiom in the special case where we consider the two conditional independence statements $X_1 \perp X_2 | X_3$ and $X_1 \perp X_3 | X_2$. The traditional statement of the intersection axiom in this case is that if the density $f(x_1, x_2, x_3)$ is positive then also $X_1 \perp (X_2, X_3)$ holds. As mentioned after the proof of Proposition 4.1.5, the proof of this conditional independence implication actually only requires that the marginal density $f_{\{2,3\}}(x_2, x_3)$ is positive. This can, in fact, be further weakened.

To state the strongest possible version, we associate to a density function $f(x_1, x_2, x_3)$ a graph, $G_f$. The graph $G_f$ has a vertex for every $(i_2, i_3) \in$
[r_2] \times [r_3] such that f_{12,3}(x_2, x_3) > 0. A pair of vertices \((i_2, i_3)\) and \((j_2, j_3)\) is connected by an edge if either \(i_2 = j_2\) or \(i_3 = j_3\).

**Theorem 4.3.3.** Let \(X\) be a discrete random variable with density \(f\), satisfying the conditional independence constraints \(X_1 \perp\!\!\!\!\perp X_2 | X_3\) and \(X_1 \perp\!\!\!\!\perp X_3 | X_2\). If \(G_f\) is a connected graph, then the conditional independence constraint \(X_1 \perp\!\!\!\!\perp (X_2, X_3)\) also holds.

On the other hand, let \(H\) be any graph with vertex set in \([r_2] \times [r_3]\), and with connectivity between vertices as described above and such that \(H\) is disconnected. Then there is a discrete random vector \(X\) with density function \(f\) such that \(G_f = H\), \(X\) satisfies \(X_1 \perp\!\!\!\!\perp X_2 | X_3\) and \(X_1 \perp\!\!\!\!\perp X_3 | X_2\), but does not satisfy \(X_1 \perp\!\!\!\!\perp (X_2, X_3)\).

To prove Theorem 4.3.3, we compute the primary decomposition of the conditional independence ideal \(I_C = I_{1 \perp\!\!\!\!\perp 2\mid 3} + I_{1 \perp\!\!\!\!\perp 3\mid 2}\), and analyze the components of the ideal. Note that both conditional independence statements \(1 \perp\!\!\!\!\perp 2\mid 3\) and \(1 \perp\!\!\!\!\perp 3\mid 2\) translate to rank conditions on the \(r_1 \times (r_2 r_3)\) flattened matrix

\[
P = \begin{pmatrix}
  p_{111} & p_{112} & \cdots & p_{1r_2 r_3} \\
p_{211} & p_{212} & \cdots & p_{2r_2 r_3} \\
  \vdots & \vdots & \ddots & \vdots \\
p_{r_111} & p_{r_112} & \cdots & p_{r_1 r_2 r_3}
\end{pmatrix}.
\]

Columns in the flattening are indexed by pairs \((i_2, i_3)\) \in \([r_2] \times [r_3]\). In particular, the CI statement \(1 \perp\!\!\!\!\perp 2\mid 3\) says that pairs of columns with indices \((i_2, i_3)\) and \((j_2, j_3)\) are linearly dependent. The CI statement \(1 \perp\!\!\!\!\perp 3\mid 2\) says that pairs of columns with indices \((i_2, i_3)\) and \((i_3, j_3)\) are linearly dependent. The conclusion of the intersection axiom is that any pair of columns is linearly dependent, in other words, the matrix \(P\) has rank 1.

Now we explain the construction of the primary components, which can also be interpreted as imposing rank conditions on subsets of columns of the matrix \(P\). Let \(A = A_1 \cdots A_l\) be an ordered partition of \([r_2]\) and \(B = B_1 \cdots B_l\) be an ordered partition of \([r_3]\) with the same number of parts as \(A\). We assume that none of the \(A_k\) or \(B_k\) are empty. Form an ideal \(I_{A,B}\) with the following generators:

\[p_{i_1 i_2 j_3} p_{j_1 j_2 j_3} - p_{i_1 j_2 j_3} p_{j_1 i_2 j_3} \text{ for all } k \in [l], \ i_2, j_2 \in A_k \text{ and } i_3, j_3 \in B_k\]

\[p_{i_1 i_2 j_3} \text{ for all } k_1, k_2 \in [l] \text{ with } k_1 \neq k_2, \ i_2 \in A_{k_1}, j_2 \in B_{k_2}\].

**Theorem 4.3.4.** Let \(I_C = I_{1 \perp\!\!\!\!\perp 2\mid 3} + I_{1 \perp\!\!\!\!\perp 3\mid 2}\) be the conditional independence ideal of the intersection axiom. Then

\[I_C = \cap_{A,B} I_{A,B}\]
where the intersection runs over all partitions $A$ and $B$ with the same number of components (but of arbitrary size). This decomposition is an irredundant primary decomposition of $I_C$ after removing the redundancy arising from the permutation symmetry (simultaneously permuting the parts of $A$ and $B$).

Theorem 4.3.4 is proven in [Fin11]. A generalization to continuous random variables appears in [Pet14]. We will give a proof of the geometric version of the theorem, that is, show that the ideals $I_{A,B}$ give all minimal primes of $I_C$. To achieve this, we exploit to special features of this particular problem.

**Proof.** The proof of the geometric version of Theorem 4.3.4 is based on two observations. First, since both the conditional independence statements $X_1 \perp X_2 | X_3$ and $X_1 \perp X_3 | X_2$ are saturated (that is, they involve all three random variables), the ideal $I_C$ is a binomial ideal. Second, the special structure of the conditional independence statements in this particular problem allow us to focus on interpreting the conditional independence statements as rank conditions on submatrices of one particular matrix.

Consider the flattened matrix

$$P = \begin{pmatrix}
p_{111} & p_{112} & \cdots & p_{1r_2r_3} 
p_{211} & p_{212} & \cdots & p_{2r_2r_3} 
\vdots & \vdots & \ddots & \vdots 
p_{r_111} & p_{r_112} & \cdots & p_{r_1r_2r_3}
\end{pmatrix},$$

whose rows are indexed by $i_1 \in [r_1]$ and whose columns are indexed by tuples $(i_2, i_3) \in [r_2] \times [r_3]$. The conditional independence statement $X_1 \perp (X_2, X_3)$ says that the $r_3$ submatrices of $P$ of size $r_1 \times r_2$ of the form $(p_{i_1i_2j})_{i_1 \in [r_1]}, i_2 \in [r_2]$ have rank 1. The conditional independence statement $X_1 \perp X_3 | X_2$ says that the $r_2$ submatrices of $P$ of size $r_1 \times r_3$ of the form $(p_{i_1ij})_{i_1 \in [r_1], i_3 \in [r_3]}$ have rank 1.

The conditional independence constraint $X_1 \perp (X_2, X_3)$ says that the matrix $P$ has rank 1. Each of the components $I_{A,B}$ says that certain columns of $P$ are zero, while other submatrices have rank 1. We prove now that if a point $P \in V(I_C)$ then it must belong to one of $V(I_{A,B})$ for some pair of partitions $A$ and $B$.

Construct a graph $G_P$ as follows. The vertices of $G_P$ consists of all pairs $(i_2, i_3)$ such that not all $p_{i_1i_2i_3}$ are zero. A pair of vertices $(i_2, i_3)$ and $(j_2, j_3)$ will form an edge if $i_2 = j_2$ or $i_3 = j_3$. Let $G_1, \ldots, G_k$ be the connected components of $G_P$. For $l = 1, \ldots, k$ let $A_l = \{i_2 : (i_2, i_3) \in V(G_l) \text{ for some } i_3\}$ and let $B_l = \{i_3 : (i_2, i_3) \in V(G_l) \text{ for some } i_2\}$. If there is some $i_2 \in [r_2]$ which does not appear in any $A_l$, add it to $A_l$, and similarly, if there is some $i_3 \in [r_3]$ which does not appear in any $B_l$ add it to $B_l$. Clearly
\[ A_1 \cdots |A_k \text{ and } B_1 \cdots |B_k \text{ are partitions of } [r_2] \text{ and } [r_3] \text{ respectively, of the same size. We claim that } P \in V(I_{A,B}). \]

It is not difficult to see that \( I_C \subseteq I_{A,B} \), and that each ideal \( I_{A,B} \) is prime. Finally, for two different partitions \( A, B \) and \( A', B' \), which do not differ by a simultaneous permutation of the partition blocks we do not have \( I_{A,B} \subseteq I_{A',B'} \). These facts together prove that the varieties \( \bigcup V(I_{A,B}) = V(I_C) \) is an irredundant irreducible decomposition.

From the geometric version of Theorem 4.3.4, we deduce Theorem 4.3.3.

**Proof of Theorem 4.3.3.** Suppose that \( P \in V(I_C) \) and that the associated graph \( G_P \) is connected. This implies, by the argument in the proof of Theorem 4.3.4 that the flattened matrix \( P \) has rank 1. Hence \( P \in V(I_{[r_2],[r_3]}) \). The component \( V(I_{[r_2],[r_3]}) \) is the component where the flattening has rank 1, and hence \( I_{[r_2],[r_3]} = I_{2,3} \).

On the other hand, let \( H \) be a disconnected, induced subgraph of the complete bipartite graph \( K_{r_2,r_3} \). This induces a pair of partitions \( A, B \) of \([r_2] \) and \([r_3] \) (adding indices that do not appear in any of the vertices of \( H \) to any part of \( A \) and \( B \), respectively). Since \( H \) is disconnected, \( A \) and \( B \) will each have at least 2 parts. The resulting component \( V(I_{A,B}) \) contains distributions which have graph \( H \). Furthermore, by the rank conditions induced by the \( I_{A,B} \) only require that the matrix \( P \) have rank 1 on submatrices with column indices given by \( A_i, B_i \) pairs. Hence, we can construct elements in \( P \in V_{\Delta}(I_{A,B}) \) such that rank Flat\(_{123}(P) = \min(r_1, c(H)) \) where \( c(H) \) is the number of components of \( H \). In particular, such a \( P \) is not in \( V(I_{2,3}) \).

4.3.2. Pairwise Marginal Independence. In this section, we study the pairwise conditional independence model for three random variables, that is, three dimensional discrete random vectors satisfying the conditional independence constraints \( X_1 \perp X_2, X_1 \perp X_3, \text{ and } X_2 \perp X_3 \). This example and generalizations of it were studied by Kirkup [Kir07] from an algebraic perspective. In the standard presentation of this conditional independence ideal, it is not a binomial ideal. However, a simple change of coordinates reveals that in a new coordinate system it is binomial, and that new coordinate system can be exploited to get at the probability distributions that come from the model. The same change of coordinates will also work for many other marginal independence models.

We introduce a linear change of coordinates on \( \mathbb{R}^R \rightarrow \mathbb{R}^R \) as follows. Instead of indexing coordinates by tuples in \([r_1] \times [r_2] \times [r_3] \), we use the set with the same cardinality

\[ [r_1 - 1] \cup \{+\} \times [r_2 - 1] \cup \{+\} \times [r_3 - 1] \cup \{+\}. \]
4.3. Primary Decomposition of CI Ideals

So the new coordinates on $\mathbb{R}^R$ looks like $p_{i_1+\ldots+i_r}$. The plus denotes marginalization, and this is how the map between coordinates is defined:

\[
\begin{align*}
p_{+++} & := \sum_{j_1 \in [r_1], j_2 \in [r_2], j_3 \in [r_3]} p_{j_1j_2j_3} \\
p_{++i_3} & := \sum_{j_1 \in [r_1], j_2 \in [r_2]} p_{j_1j_2i_3} \\
p_{+i_2+} & := \sum_{j_1 \in [r_1], j_3 \in [r_3]} p_{j_1j_2j_3} \\
p_{i_1++} & := \sum_{j_2 \in [r_2], j_3 \in [r_3]} p_{i_1j_2j_3}
\end{align*}
\]

where $i_1 \in [r_1 - 1], i_2 \in [r_2 - 1], i_3 \in [r_3 - 1]$. This transformation is square and also in triangular form, hence it is invertible (essentially via Möbius inversion).

In the new coordinate system, any marginal independence statement can be generated by binomial equations. For example, the independence statement $X_1 \perp X_2$ is equivalent to the matrix

\[
\begin{pmatrix}
p_{11+} & p_{12+} & \cdots & p_{1r_2 - 1+} & p_{1++} \\
p_{21+} & p_{22+} & \cdots & p_{2r_2 - 1+} & p_{2++} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
p_{r_1 - 11+} & p_{r_1 - 12+} & \cdots & p_{r_1 - 1r_2 - 1+} & p_{r_1 - 1++} \\
p_{++1} & p_{++2} & \cdots & p_{++r_2 - 1+} & p_{+++}
\end{pmatrix}
\]

having rank 1. This follows because the matrix whose rank we usually ask to have rank 1 in the marginal independence statement is related to this one via elementary row and column operations, which preserve rank.

In the new coordinate system, the ideal $I_C$ is binomial. It is:

\[
I_C = \langle p_{i_1i_2+j_1j_2} + p_{i_1j_2+j_1i_2} - p_{i_1j_1}p_{j_1j_2} - p_{i_1+j_1}p_{j_1i_2} + i_1, j_1 \in [r_1 - 1] \cup \{+\}, \quad i_2, j_2 \in [r_2 - 1] \cup \{+\}, i_3, j_3 \in [r_3 - 1] \cup \{+\} \rangle.
\]

The following Macaulay 2 code computes the primary decomposition of this ideal in the case where $r_1 = r_2 = r_3 = 2$.

\[
R = \text{QQ\{p000,p001,p010,p011,p100,p101,p110,p111\};}
\]
\[
I = \text{ideal\{p000*p011 - p001*p010, p000*p101 - p001*p100, p000*p110 - p010*p100\};}
\]
\[
\text{primaryDecomposition } I
\]

Note that 0 takes the place of the symbol “+” in the Macaulay 2 code. We see that in this case of binary random variables, $I_C$ has primary decomposition consisting of four ideals:

\[
I_C = \langle \{ p_{++1}p_{1++} - p_{++1}p_{11+} + p_{++1}p_{1++} - p_{++1}p_{11+} \}, \quad p_{++1}p_{1++} - p_{++1}p_{11+} + p_{++1}p_{1++} - p_{++1}p_{11+} \rangle
\]

\[
\cap \{ p_{++1}p_{1++} - p_{++1}p_{11+} \} \cap \{ p_{++1}p_{1++} - p_{++1}p_{11+} \}
\]

\[
\cap \{ p_{++1}p_{1++} - p_{++1}p_{11+} \} \cap \{ p_{++1}p_{1++} - p_{++1}p_{11+} \}
\]
\( \cap \{ p_{1+++}, p_{+++1}, p_{+++} \} \)

Note that the three linear components each contain \( p_{+++} \) as a generator. There are no probability distributions that satisfy \( p_{+++} = 0 \). Since we are only interested in solutions to the equation system that are in the probability simplex, we see that the three linear components are extraneous do not contribute to our understanding of this conditional independence model. Our computation for binary variables suggests the following theorem.

**Theorem 4.3.5.** [Kir07] Let \( C = \{1 \perp 2, 1 \perp 3, 2 \perp 3\} \) be the set of all pairwise marginal independencies for three discrete random variables. Then the variety \( V(\mathcal{I}_C) \) has exactly one component that intersects the probability simplex nontrivially. The probability distributions in that unique component consist of all probability distributions which can be written as a sum of a rank 1 3-way tensor plus a three way tensor with all zero 2-way margins.

**Proof.** The ideal \( \mathcal{I}_C \) homogeneous, so we could consider it as defining a projective variety. To this end, it makes sense to investigate that variety on its affine coordinate charts. In particular, we consider the chart where \( p_{+++} = 1 \). After adding this condition, and eliminating extraneous binomials, we arrive at the ideal

\[
J = \langle p_{1i2+} - p_{1+++}p_{i2++}, p_{1i+3} - p_{1+++}p_{1i+3}, p_{++i3} - p_{++i2}p_{+++}, \\
i_1 \in [r_1 - 1], i_2 \in [r_2 - 1], i_3 \in [r_3 - 1] \rangle.
\]

This ideal is prime, since it has an initial ideal which is linear and hence prime (for example, take any lexicographic order that makes any indeterminate with one + bigger than any indeterminate with two +’s). This implies that besides the prime component \( J^0 \), (the homogenization of \( J \) with respect to the indeterminate \( p_{+++} \)) every other minimal prime of \( \mathcal{I}_C \) contains \( p_{+++} \). Since \( p_{+++} = 1 \) for probability distributions, there is only one component which intersects the probability simplex nontrivially.

The description of \( J \) also gives us a parametrization of this component after homogenization. It is given parametrically as

\[
p_{+++} = t; \quad p_{1+++} = \theta_{i_1}^{(1)} t; \quad p_{++i2} = \theta_{i_2}^{(2)} t; \quad p_{+++i3} = \theta_{i_3}^{(3)} t; \\
p_{1i2+} = \theta_{i_1}^{(1)} \theta_{i_2}^{(2)} t; \quad p_{1i+3} = \theta_{i_1}^{(1)} \theta_{i_3}^{(3)} t; \quad p_{++i23} = \theta_{i_2}^{(2)} \theta_{i_3}^{(3)} t; \quad p_{1i2i3} = \theta_{i_1i_2i_3}^{(123)},
\]

where \( i_1 \in [r_1 - 1], i_2 \in [r_2 - 1], i_3 \in [r_3 - 1] \). This gives a dense parametrization of this component because all the equations in \( J \) have the form that some coordinate is equal to the product of some other coordinates. Hence, we can treat some coordinates as free parameters (namely, \( p_{1i1i3}, p_{11++}, p_{++i2}, p_{+++i3} \), for \( i_1 \in [r_1 - 1], i_2 \in [r_2 - 1], i_3 \in [r_3 - 1] \)) and the remaining
coordinates as functions of those parameters. Note that we can modify the parametrization to have

\[ p_{i_1i_2i_3} = \theta_{i_1}^{(1)} \theta_{i_2}^{(2)} \theta_{i_3}^{(3)} t + \theta_{i_1i_2i_3}^{(123)} \]

without changing the image of the parametrization. This is because the \( p_{i_1i_2i_3} \) coordinates do not appear in any generators of \( I_C \), so they are completely free, hence adding expressions in other independent parameters will not change the image. This latter variation on the parametrization realizes the second paragraph of the theorem, as we can break the parametrization into the sum of two parts. The first part, with

\[ p_{i_1i_2i_3} = \theta_{i_1}^{(1)} \theta_{i_2}^{(2)} \theta_{i_3}^{(3)} t \]

gives the parametrization of a rank one tensor. This will still be a rank one tensor after we apply the Möbius transform to recover all the \( p_{i_1i_2i_3} \). The second part, with

\[ p_{i_1i_2i_3} = \begin{cases} \theta_{i_1i_2i_3}^{(123)} & i_1 \in [r_1 - 1], i_2 \in [r_2 - 1], i_3 \in [r_3 - 1], \\ 0 & \text{otherwise} \end{cases} \]

gives the parametrization of 3-way arrays with all margins equal to zero (since all the terms involving a + do not have this term).

4.3.3. No Finite Set of Axioms for Gaussian Independence. A motivating problem in the theoretical study of conditional independence was whether it was possible to give a complete axiomatization of all conditional independence implications. The conditional independence “axioms” given at the beginning of this chapter, were an attempt at this. Eventually, Studený [Stu92] showed there there is no finite list of axioms from which all other conditional independence implications can be deduced. Studený’s result applies to conditional independence implications that hold for all random variables, but one might ask if perhaps there were still a finite list of conditional independence axioms that hold in restricted settings, for example restricting to regular Gaussian random variables. This was resolved in the negative in [Sul09] using primary decomposition of conditional independence ideals. We illustrate the idea in this section.

Consider this cyclic set of conditional independence statements:

\[ X_1 \perp X_2 | X_3, \ldots, X_{m-1} \perp X_m | X_1, X_m \perp X_1 | X_2. \]

We will show that for Gaussian random variables these statements imply the marginal independent statements:

\[ X_1 \perp X_2, \ldots, X_{m-1} \perp X_m, X_m \perp X_1. \]
The conditional independence ideal for the first CI statement is the binomial ideal:

\[ J_C = (\sigma_{12}\sigma_{33} - \sigma_{13}\sigma_{23}, \ldots, \sigma_{1n}\sigma_{22} - \sigma_{2n}\sigma_{12}). \]

The ideal \( J_C \) is a pure difference binomial ideal, so its minimal associated lattice ideals have a prescribed form given by Proposition 4.2.14. The complete description is given by the following result.

**Theorem 4.3.6.** Assume \( m \geq 4 \) and let \( C = \{1 \nmid 2|3, 2 \nmid 3|4, \ldots, m \nmid 1|2\} \). Then the Gaussian CI ideal \( J_C \) has exactly two minimal primes which are

\[ (\sigma_{12}, \sigma_{23}, \ldots, \sigma_{m-1,m}, \sigma_{m1}) \]

and the toric component

\[ (J_C)_0 = J_C : \prod \sigma_{ij}^{\infty}. \]

The irreducible component \( V((J_C)_0) \) does not intersect the positive definite cone. Hence, for jointly normal random variables

\[ X_1 \nmid X_2 \nmid X_3, \ldots, X_{m-1} \nmid X_m | X_1, X_m \nmid X_1 | X_2 \implies X_1 \nmid X_2 \ldots, X_{m-1} \nmid X_m, X_m \nmid X_1. \]

**Proof Sketch.** The ideal \( J_C \) involves polynomials only in the variables \( \sigma_{ii}, \sigma_{i,i+1} \) and \( \sigma_{i,i+2} \), for \( i \in [m] \) (with indices considered modulo \( m \)). Note that the variables of the type \( \sigma_{ii} \) and \( \sigma_{i,i+2} \) each only appear in one of the binomial generators of \( J_C \). The results of \([HS00]\) imply that those variables could not appear as variables of a minimal associated lattice ideal. To get a viable candidate set of variables among the remaining variables, we must take the entire remaining set \( \{\sigma_{i,i+1} : i \in [m]\} \). This gives the ideal \( (\sigma_{12}, \sigma_{23}, \ldots, \sigma_{m-1,m}, \sigma_{m1}) \). The empty set is always a valid candidate set, and this produces the ideal \( (J_C)_0 \) from the statement of the theorem. The fact that \( V((J_C)_0) \) does not intersect the positive definite cone (proven below), and that the other component does, and that \( V((J_C)_0) \) contains points with all nonzero entries, while the other component does not, ensures that these two components do not contain each other.

To see that \( V((J_C)_0) \) does not intersect the positive definite cone, we will show that \( (J_C)_0 \) contains the polynomial

\[ f = \prod_{i=1}^{m} \sigma_{ii} - \prod_{i} \sigma_{i,i+2}. \]

For a positive definite matrix, we have for each \( i \) that \( \sigma_{ii}\sigma_{i+2,i+2} > \sigma_{i,i+2}^2 \), which implies that \( f > 0 \) for any positive definite matrix. Now, the polynomial obtained by multiplying the leading and trailing terms of each of the
generators of $J_C$, belongs to $J_C$, which is the polynomial

$$g = \prod_{i=1}^{m} \sigma_{i+1, i+2} \sigma_{i, i+1} - \prod_{i=1}^{m} \sigma_{i+1, i+2} \sigma_{i, i+2}$$

(where indices are again considered modulo $m$). The polynomial $g$ is divisible by $\prod_{i=1}^{m} \sigma_{i, i+1}$, and hence $f = g/\prod_{j} \sigma_{i, i+1}$ is in $(J_C)_{\emptyset} = J_C : \prod_{j}^{\infty} \sigma_{i, i+1}$.

The resulting conditional independence implication follows since the linear ideal $\langle \sigma_{12}, \sigma_{23}, \ldots, \sigma_{m1} \rangle$ is the CI ideal for $D = \{1 \sqcup 2, \ldots, m \sqcup 1\}$. $\square$

Note that by carefully choosing a term order, we can arrange to have the binomial generators of $J_C$ with squarefree and relatively prime initial terms. This implies that $J_C$ is radical and hence equal to the intersection of the two primes from Theorem 4.3.6.

The statement that heads the section, that there are no finite list of axioms for Gaussian conditional independence follows from Exercise 4.11 below, which shows that no subset of the given CI statements imply any of the marginal constraints $i \sqcup i + 1$. Hence, this shows that there can be Gaussian conditional independence implications that involve arbitrarily large sets of conditional independence constraints but cannot be deduced from conditional independence implications of smaller size.

4.4. Exercises

Exercise 4.1. The conditional independence implication

$$A \perp B | c \cup D \text{ and } A \perp B | D \implies A \perp B \cup c | D \text{ or } A \cup c \perp B | D$$

is called the Gaussoid axiom, where $c$ denotes a singleton.

1. Show that if $X$ is a jointly normal random variable, it satisfies the Gaussoid axiom.
2. Show that if $X$ is discrete with finitely many states, then $X$ satisfies the special case of the Gaussoid axiom with $D = \emptyset$ and $r_c = 2$.
3. Show that in the discrete case if we eliminate either the condition $D = \emptyset$ or $r_c = 2$ we have distributions that do not satisfy the Gaussoid axiom.

Exercise 4.2. (1) Show that every prime ideal is radical.

2. Show that the radical of a primary ideal is prime.

3. True or False: If $\sqrt{Q}$ is a prime ideal, then $Q$ is primary.

Exercise 4.3. Prove part (1) of Proposition 4.2.5 as follows.

1. Show that for any $I$ and $J$, there exists an $N$ such that $I : J^N = I : J^{N+1} = I : J^{N+2} \ldots$.
4. Conditional Independence

(2) Suppose that $I$ is an ideal, $fg \in I$, and $f \notin I$. Let $N$ be the integer such that $I : g^N = I : g^{N+1}$. Show that $I = (I + (f)) \cap (I + (g^N))$.

(3) Show that if $I$ is irreducible then $I$ is primary.

Exercise 4.4. Let $M \subseteq \mathbb{K}[p]$ be a monomial ideal.

(1) Show that $M$ is prime if and only if $M$ is generated by a subset of the variables.

(2) Show that $M$ is radical if and only if $M$ has a generating set consisting of monomials $p^u$ where each $u$ is a $0/1$ vector. (Radical monomial ideals are often called squarefree monomial ideals for this reason.

(3) Show that $M$ is irreducible if and only if $M$ has the form $\langle p_1^{\alpha_1} : i \in S \rangle$ for some set $S \subseteq [r]$ and $\alpha_i \in \mathbb{N}$.

(4) Show that $M$ is primary if and only if for each minimal generator $p^u$ of $M$, if $u_i > 0$, then there is another generator of $M$ of the form $p_i^{\alpha_i}$ for some $\alpha_i$.

Exercise 4.5. Show that if $I$ is a radical ideal every ideal in its primary decomposition is prime, and this decomposition is unique.

Exercise 4.6. Prove Proposition 4.2.10.

Exercise 4.7. For four binary random variables, consider the conditional independence model $C = \{1 \perp|2, 4 \perp|1, 3\}$. Compute the primary decomposition of $I_C$ and describe the components.

Exercise 4.8. Let $C = \{A \perp|B|C, B \perp|C\}$ be the set of conditional independence statements from the contraction axiom. Determine the primary decomposition of the Gaussian conditional independence ideal $J_C$ and use it to deduce the conclusion of the contraction axiom in this case.

Exercise 4.9. In the marginal independence model $C = \{1 \perp|2, 1 \perp|3, 2 \perp|3\}$ from Theorem 4.3.5, find the generators of the ideal of the unique component of $I_C$ that intersects the probability simplex.

Exercise 4.10. Consider the marginal independence model for 4 binary random variables $C = \{1 \perp|2, 2 \perp|3, 3 \perp|4, 1 \perp|4\}$.

(1) Compute the primary decomposition of $I_C$.

(2) Show that there is exactly one component of $V(I_C)$ that intersects the probability simplex.

(3) Give a parametrization of that component.

Exercise 4.11. Let $m \geq 4$ and consider the Gaussian conditional independence model

$C = \{1 \perp|2, 2 \perp|3, 3 \perp|4, \ldots, m - 1 \perp|m|1\}$. 

(1) Show that $J_C$ is a binomial ideal.

(2) Show that $J_C$ is a radical ideal (Hint: Show that the given binomial generators are a Gröbner basis with squarefree initial terms and then use Proposition 4.2.11.)

(3) Show that $J_C$ is prime by analyzing the candidate primes of Proposition 4.2.14. (To complete this problem will require some familiarity with the results of [HS00]. In particular, first show that $J_C$ is a lattice basis ideal. Then use Theorem 2.1 in [HS00] to show that $\emptyset$ is the only viable candidate set.)

**Exercise 4.12.** Consider the Gaussian conditional independence model

$$C = \{1 \perp 2|3, 1 \perp 3|4, 1 \perp 4|2\}.$$  

Compute the primary decomposition of $J_C$ and use this to determine what, if any, further conditional independence statements are implied.
Statistics Primer

Probability primarily concerns random variables and studying what happens to those random variables under various transformations or when repeated measurements are made. Implicit in the study is the fact that the underlying probability distribution is fixed and known (or described implicitly) and we want to infer properties of that distribution.

Statistics in large part turns this picture around. We have a collection of data, which are assumed to be random samples from some underlying, but unknown, probability distribution. We would like to develop methods to infer properties of that distribution from the data. Furthermore, we would like to perform hypothesis tests, that is, determine whether or not the data is likely to have been generated from a certain family of distributions, and measure how confident we are in that assessment.

Statistics is a rich subject and we only briefly touch on many topics in that area. Our focus for much of the book will be on frequentist methods of statistics, which is concerned with the development of statistical procedures based on the long term behavior of those methods under repeated trials or large sample sizes. We also discuss Bayesian methods in Section 5.5, where statistical inference is concerned with how data can be used to update our prior beliefs about an unknown parameter. Algebraic methods for Bayesian statistics will appear in Chapters 17 and 18.

5.1. Statistical Models

Much of statistics is carried out with respect to statistical models. A statistical model is just a family of probability distributions. Often these models
are parametric. In this section, we give the basic definitions of statistical models and give a number of important examples.

**Definition 5.1.1.** A *statistical model* $\mathcal{M}$ is a collection of probability distributions or density functions. A *parametric statistical model* $\mathcal{M}_\Theta$ is a mapping from a finite dimensional parameter space $\Theta \subseteq \mathbb{R}^d$ to a space of probability distributions or density functions, i.e.

$$p_\bullet : \Theta \to \mathcal{M}_\Theta, \quad \theta \mapsto p_\theta.$$  

The model is the image of the map $p_\bullet$, $\mathcal{M}_\Theta = \{p_\theta : \theta \in \Theta\}$.

A statistical model is called *identifiable* if the mapping $p_\bullet$ is one-to-one; that is, $p_{\theta_1} = p_{\theta_2}$ implies that $\theta_1 = \theta_2$. We will return to discussions of identifiability in Chapter 16.

We have already seen examples of statistical models in the previous chapters, which we remind the reader of, and put into context here.

**Example 5.1.2 (Binomial Random Variable).** Let $X$ be a discrete random variable with $r + 1$ states, labeled $0, 1, \ldots, r$. Let $\Theta = [0, 1]$ and for $\theta \in \Theta$ consider the probability distribution

$$P_\theta(X = i) = \binom{r}{i} \theta^i (1 - \theta)^{r-i}$$

which is the distribution of a binomial random variable with $r$ samples and parameter $\theta$. The binomial random variable model $\mathcal{M}_\Theta$ consists of all the probability distribution that arise this way. It sits as a subset of the probability simplex $\Delta_r$.

**Example 5.1.3 (Multivariable Normal Random Vector).** Let $X \in \mathbb{R}^m$ be an $m$-dimensional real random vector. Let $\Theta = \mathbb{R}^m \times PD_m$, where $PD_m$ is the cone of symmetric $m \times m$ positive definite matrices. For $\theta = (\mu, \Sigma) \in \Theta$ let

$$p_\theta(x) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp \left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

be the density of a jointly normal random variable with mean vector $\mu$ and covariance matrix $\Sigma$. The model $\mathcal{M}_\Theta$ consists of all such density functions for an $m$-dimensional real random vector.

Statistical models with infinite dimensional parameter spaces are called *nonparametric* or *semiparametric* depending on the context. Note that this does not necessarily mean that we are considering a statistical model which does not have a parametrization, as so-called nonparametric models might still have nice parametrizations. On the other hand, models which are defined via constraints on probability distributions or densities we will refer to as *implicit statistical models*. Note, however, that in the statistics literature
such models defined by implicit constraints on a finite dimensional parameter space are still called parametric models. We have already seen examples of implicit statistical models when looking at independence constraints.

**Example 5.1.4 (Independent Random Variables).** Let $X_1$ and $X_2$ be two discrete random variables with state spaces $[r_1]$ and $[r_2]$ respectively, with $\mathcal{R} = [r_1] \times [r_2]$. The model of independence $\mathcal{M}_{X_1 \perp X_2}$ consists of all distributions $p \in \Delta_{\mathcal{R}}$ such that

$$P(X_1 = i_1, X_2 = i_2) = P(X_1 = i_1)P(X_2 = i_2)$$

for all $i_1 \in [r_1]$ and $i_2 \in [r_2]$. Thus, the model of independence is an implicit statistical model. As discussed in Section 4.1, this model consists of all joint distribution matrices which are rank 1 matrices. From that description we can also realize this as a parametric statistical model. Indeed, let $\Theta = \Delta_{r_1-1} \times \Delta_{r_2-1}$, and given $\theta = (\alpha, \beta) \in \Theta$, define the probability distribution

$$P_\theta(X_1 = i_1, X_2 = i_2) = \alpha_{i_1}\beta_{i_2}.$$

Our main example of implicit statistical model in this book comes from looking at conditional independence constraints. If

$$\mathcal{C} = \{ A_1 \perp B_1 | C_1, A_2 \perp B_2 | C_2, \ldots \}$$

is a collection of conditional independence statements, for each fixed set of states $r_1, r_2, \ldots, r_m$ we get a discrete conditional independence model

$$\mathcal{M}_\mathcal{C} = V(I_{A_1 \perp B_1 | C_1} + I_{A_2 \perp B_2 | C_2} + \cdots) \cap \Delta_{\mathcal{R}}$$

which is an implicit statistical model. Similarly, we get a gaussian conditional independence model

$$\mathcal{M}_\mathcal{C} = V(J_{A_1 \perp B_1 | C_1} + J_{A_2 \perp B_2 | C_2} + \cdots) \cap PD_m.$$

An important concept in the study of statistical models is the notion of sufficiency of a statistic. Sufficient statistics explain the way that the state of a variable enters into the calculation of the probability of that state in a model, and the interaction of the state of the random variable with the parameters.

**Definition 5.1.5.** A statistic is a function from the state space of a random variable to some other set. For a parametric statistical model $\mathcal{M}_\theta$, a statistic $T$ is sufficient for the model if

$$P(X = x \mid T(X) = t, \theta) = P(X = x \mid T(X) = t).$$

Equivalently, the joint distribution $p_\theta(x)$ should factorize as

$$p_\theta(x) = f(x)g(T(x), \theta)$$
where $f$ is a function that does not depend on $\theta$. A statistic $T$ is \textit{minimal sufficient} if every other sufficient statistic is a function of $T$.

Examples of sufficiency will appear in the next section as we discuss different types of data.

5.2. Types of Data

Statistical models are used to analyze data. If we believe our data follows a distribution that belongs to the model, we would like to estimate the model parameters from the data and use the resulting quantities to make some conclusions about the underlying process that generated the data. Alternately, we might not be sure whether or not a given model is a good explanation of the data. To this end, we would perform a hypothesis test, to decide whether or not the model is a good fit. Before we can talk about any of these statistical settings, however, we need to discuss what we mean by data, and, in some sense, different types of models. In particular, natural classes of data generating procedures include: independent and identically distributed data, exchangeable data, and time-series and spatial data. The techniques from algebraic statistics that can be applied depend on what was the underlying data generation assumption.

The most common type of data generation we will encounter in this book is the case of \textit{independent and identically distributed} data. In this setting, we assume that we have a sample of $n$ data $D = X^{(1)}, X^{(2)}, \ldots, X^{(n)}$, each distributed identically and independently like the distribution $p(\theta)(X)$, for some distribution $p_\theta$ in our model. If our random variables are discrete, the probability of observing the particular sequence of data, given $\theta$ is

$$p_\theta(D) = \prod_{i=1}^{n} p_\theta(X^{(i)}).$$

For continuous distributions, the density function is represented similarly in product form. Note that in the discrete case, where our random variable has state space $[r]$, we can compute the vector of counts $u \in \mathbb{N}^r$, defined by

$$u_j = \# \{ i : X^{(i)} = j \}$$

in which case

$$p_\theta(D) = \prod_{j=1}^{r} p_\theta(j)^{u_j}$$

so that the vector of counts gives sufficient statistics for a model under the i.i.d. assumption. Note that for discrete random variables, the vector of counts $u$ is a summary of the data from which it is possible to compute the probability of observing the particular sequence $i_1, \ldots, i_n$. When working
5.2. Types of Data

with such models and discussing data for such models, we usually immediately pass to the vector of counts as our data for such models.

A second type of assumption that arises in Bayesian statistics is known as exchangeable random variables. Discrete random variables $X^{(1)}, \ldots, X^{(n)}$ are exchangeable if and only if

$$P(X_1 = i_1, \ldots, X_n = i_n) = P(X_1 = i_{\sigma(1)}, \ldots, X_n = i_{\sigma(n)})$$

for any permutation $\sigma$ of the index set $[n]$. For continuous random variables, we require the same permutation condition on the density function. Note that independent identically distributed random variables are exchangeable, but the reverse is not true. For example, sampling balls from an urn without replacement gives a distribution on sequences of balls that is exchangeable but not i.i.d. Although sampling without replacement is a common sampling mechanism, we will typically assume i.i.d. sampling in most contexts. We will discuss exchangeable random variables again in Section 14.1 where De Finetti’s theorem is discussed.

The first two data types we have discussed have the property that the order that the data arrives does not matter in terms of the joint distribution of all the data. Only the vector of counts is important in both i.i.d. and exchangeable random variables. A third type of data to consider is when the order of the variables does matter. This type of data typical arises in the context of time series data, or spatial data.

A typical example of a time series model, where the resulting order of the observations matters, is a Markov chain (see Chapter 1). In a Markov chain, we have a sequence of random variables $X_1, \ldots, X_n$ with the same state space $[r]$ such that

$$P(X_j = i_j \mid X_{j-1} = i_{j-1}, \ldots, X_1 = i_1) = P(X_j = i_j \mid X_{j-1} = i_{j-1})$$

for $j = 2, \ldots, n$, $i_1, \ldots, i_n \in [r]^n$. Note that by applying the multiplication rule of conditional probability and using the first condition of being a Markov chain, we see that the joint probability factorizes as

$$P(X_1 = i_1, \ldots, X_n = i_n) = P(X_1 = i_1)P(X_2 = i_2 \mid X_1 = i_1) \cdots P(X_n = i_n \mid X_{n-1} = i_{n-1}).$$

The further assumption of homogeneity requires that

$$P(X_j = i' \mid X_{j-1} = i) = P(X_2 = i' \mid X_1 = i)$$

for all $j = 3, \ldots, n$ and $i', i \in [r]$. From this we see that the sufficient statistics for a homogeneous Markov chain model are $X_1$ and the table of counts $u$ such that

$$u_{i'i} = \#\{j : X_j = i', X_{j-1} = i\}.$$
Note that \( u \) counts the number of each type of transition that occurred in the sequences \( X_1, X_2, \ldots, X_n \).

For i.i.d. samples, we typically observe many replicates from the same underlying model. When we use a time series or spatial model, the usual way data arrives is as a single sample from that model, whose length or size might not be a priori specified. For these models to be useful in practice, we need them to be specified with a not very large set of parameters, so that as the data grows (i.e. as the sequence gets longer) we have a hope of being able to estimate the parameters. Of course, it might be the situation that for a time-series or spatial model we have not just one sample, but i.i.d. data. For instance, in Chapter 1 during our discussion of the maximum likelihood estimation in a very short Markov chain, we analyzed the case where we received many data sets of the chain of length 3, where each was an i.i.d. sample from the same underlying distribution.

5.3. Parameter Estimation

Given a parametric statistical model and some data, a typical problem in statistics is to estimate some of, or all of, the parameters of the model based on the data. At this point we do not necessarily assume that the model accurately models the data. The problem of testing whether or not a model actually fits the data is the subject of the next section, on hypothesis testing.

Ideally, we would like a procedure which, as more and more data arrives, if the underlying distribution that generated the data comes from the model, the parameter estimate converges to the true underlying parameter. Such an estimator is called a consistent estimator.

Definition 5.3.1. Let \( \mathcal{M}_\theta \) be a parametric statistical model with parameter space \( \Theta \). A parameter of a statistical model is a function \( s : \Theta \to \mathbb{R} \). An estimator of \( s \), is a function from the data space \( D \) to \( \mathbb{R} \), \( \hat{s} : D \to \mathbb{R} \). The estimator \( \hat{s} \) is consistent if \( \hat{s} \to_p s \) as the sample size tends to infinity.

Among the simplest examples of estimators are the plug-in estimators. As the name suggests, a plug-in estimator is obtained by plugging in values obtained from the data to estimate parameters.

Example 5.3.2. As a simple example, consider the case of binomial random variable, with \( r + 1 \) states, \( 0, 1, \ldots, r \). The model consists of all distributions of the form

\[
\left\{ \binom{r}{\theta^r, (1-\theta)^{r-1}}, \ldots, (1-\theta)^r : \theta \in [0,1] \right\}.
\]

Under i.i.d. sampling, data consists of \( n \) repeated draws \( X^{(1)}, \ldots, X^{(n)} \) from an underlying distribution \( p_\theta \) in this model. The data is summarized by a
vector of counts $u = (u_0, \ldots, u_r)$, where $u_i = \#\{j : X(j) = i\}$. We would like to estimate the parameter $\theta$ from the data of counts $u$. The value $p_0(0) = \theta^r$, hence, if we had a consistent estimator of $p_0(0)$, we could obtain a plug-in estimate for $\theta$ by extracting the $r$-th root. For example, the formula

$$\sqrt[n]{\frac{1}{n} \sum_{i=1}^{n} 1_{X(i) = 0}} = \sqrt[n]{\frac{u_0}{n}}$$

gives a consistent plug-in estimator of the parameter $\theta$.

Intuitively, the plug-in estimator from Example 5.3.2 is unlikely to be a very useful estimator, since it only uses very little information from the data to obtain an estimate of the parameter $\theta$. When choosing a consistent plug-in estimator, we would generally like to use one whose variance rapidly tends to zero as $n \to \infty$. The estimator from Example 5.3.2 has high variance and so is an inefficient estimator of the parameter $\theta$.

Another natural choice for an estimator is a method of moments estimator. The idea of the method of moments is to choose the probability distribution in the model whose moments match the empirical moments of the data.

Definition 5.3.3. Given a random vector $X \in \mathbb{R}^m$, and an integer vector $\alpha \in \mathbb{N}^m$, the $\alpha$-th moment is

$$\mu_\alpha = E[X_1^{\alpha_1} \cdots X_m^{\alpha_m}].$$

Given i.i.d. data $X^{(1)}, \ldots, X^{(n)}$, their $\alpha$-th empirical moment is the estimate

$$\hat{\mu}_\alpha = \frac{1}{n} \sum_{i=1}^{n} (X_1^{(i)})^{\alpha_1} \cdots (X_m^{(i)})^{\alpha_m}.$$

So in method of moments estimation, we find formulas for some of the moments of the random vector $X \sim p_\theta \in \mathcal{M}_\theta$ in terms of the parameter vector $\theta$. If we calculate enough such moments, we can find a probability distribution in the model whose moments match the empirical moments. For many statistical models, formulas for the moments in terms of the parameters are given by polynomial or rational formulas in terms of the parameters. So finding the method of moments estimator will turn into the problem of solving a system of polynomial equations.

Example 5.3.4 (Binomial Random Variables). Consider the example of the model of binomial random variable from Exercise 5.3.2. Given a binomial random variable $X \sim \text{Bin}(\theta, r)$, the first moment $E[X] = r\theta$. The empirical first moment of $X^{(1)}, \ldots, X^{(n)}$ is the sample mean $\bar{X}$. Hence the method of moments estimate for $\theta$ in the binomial model is

$$\hat{\theta} = \frac{1}{r} \bar{X}.$$
The method of moments estimators can often lead to interesting algebraic problems \cite{AFS16}. One potential drawback to the method of moments estimators is that the empirical higher moments tend to have high variability, so there can be a lot of noise in the estimates.

Among the many possible estimators of a parameter, one of the most frequently used is the maximum likelihood estimator (MLE). The MLE is one of the most commonly used estimators in practice, both for its intuitive appeal, and for useful theoretical properties associated with it. In particular, it is usually a consistent estimator of the parameters and with certain smoothness assumptions on the model, it is asymptotically normally distributed. We will return to these properties in Chapter 7.

**Definition 5.3.5.** Let $D$ be data from some model with parameter space $\Theta$. The **likelihood function**

$$L(\theta \mid D) := p_\theta(D)$$

in the case of discrete data and

$$L(\theta \mid D) := f_\theta(D)$$

in the case of continuous data. Here $p_\theta(D)$ is the probability of observing the data given the parameter $\theta$ in the discrete case, and $f_\theta(D)$ is the density function evaluated at the data in the continuous case. The **maximum likelihood estimate** (MLE) $\hat{\theta}$ is maximizer of the likelihood function:

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta \mid D).$$

Note that we consider the likelihood function as a function of $\theta$ with the data $D$ fixed. This contrasts the interpretation of the probability distribution where the parameter is considered fixed and the random variable is the unknown (random) quantity.

In the case of i.i.d. sampling, so $D = X^{(1)}, \ldots, X^{(n)}$, the likelihood function factorizes as

$$L(\theta \mid D) = L(\theta \mid X^{(1)}, \ldots, X^{(n)}) = \prod_{i=1}^{n} L(\theta \mid X^{(i)}).$$

In the case of discrete data, this likelihood function is thus only a function of the vector of counts $u$, so that

$$L(\theta \mid X^{(1)}, \ldots, X^{(n)}) = \prod_{i} p_\theta(j)^{u_i}.$$
In the common setting in which we treat the vector of counts itself as the data, we need to multiply this quantity by an appropriate multinomial coefficient

\[ L(\theta \mid u) = \binom{n}{u} \prod_j p_\theta(j)^{u_j}, \]

which does not change the maximum likelihood estimate but will change the value of the likelihood function when evaluated at the maximizer.

It is common to replace the likelihood function with the log-likelihood function, which is defined as

\[ \ell(\theta \mid D) = \log L(\theta \mid D). \]

In the case of i.i.d. data, this has the advantage of turning a product into a sum. Since the logarithm is a monotone function both the likelihood and log-likelihood have the same maximizer, which is the maximum likelihood estimate.

**Example 5.3.6** (Maximum Likelihood of Binomial Random Variable). Consider the model of a binomial random variable with \( r \) trials. The probability \( p_\theta(i) = \binom{r}{i} \theta^i (1 - \theta)^{r-i} \). Given a vector of counts \( u \), the log-likelihood function is

\[
\ell(\theta, u) = C + \sum_{i=0}^{r} u_i \log(\binom{r}{i} \theta^i (1 - \theta)^{r-i})
\]

\[ = C + \sum_{i=0}^{r} (iu_i \log \theta + (r - i)u_i \log(1 - \theta)), \]

where \( C \) is a constant involving logarithms of binomial coefficients but does not depend on the parameter \( \theta \). To calculate the maximum likelihood estimate, we differentiate the log-likelihood function with respect to \( \theta \) and set it equal to zero, arriving at:

\[
\frac{\sum_{i=0}^{r} iu_i}{\theta} - \frac{\sum_{i=0}^{r} (r - i)u_i}{1 - \theta} = 0.
\]

Hence, the maximum likelihood estimator, \( \hat{\theta} \) is given by

\[
\hat{\theta} = \frac{\sum_{i=0}^{r} iu_i}{rn}.
\]

Note that \( \frac{1}{n} \sum_{i=0}^{r} iu_i \) is the sample mean \( \bar{X} \), so the maximum likelihood estimate of \( \theta \) is the same as the method of moments estimate of \( \theta \).

The gradient of the log-likelihood function is called the **score function**. Since the gradient of a function is zero at a global maximum of a differentiable function, the equations obtained by setting the score function to zero are called the **score equations** or the critical equations. In many cases,
these equations are algebraic and the algebraic nature of the equations will be explored in later chapters.

For some of the most standard statistical models, there are well-known closed formulas for the maximum likelihood estimates of parameters.

**Proposition 5.3.7.** For a multivariate normal random variable, the maximum likelihood estimates for the mean and covariance matrix are:

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X^{(i)}, \quad \hat{\Sigma} = \frac{1}{n} (X^{(i)} - \hat{\mu})(X^{(i)} - \hat{\mu})^T.
\]

**Proof.** The log-likelihood function has the form

\[
\log(\mu, \Sigma | D) = -\frac{1}{2} \sum_{i=1}^{n} \left( m \log(2\pi) + \log |\Sigma| + (X^{(i)} - \mu)^T \Sigma^{-1} (X^{(i)} - \mu) \right).
\]

The trace trick is useful for rewriting this log-likelihood as

\[
\log(\mu, \Sigma | D) = -\frac{1}{2} \left( nm \log(2\pi) + n \log |\Sigma| + \text{tr} \left( \sum_{i=1}^{n} ((X^{(i)} - \mu)(X^{(i)} - \mu)^T) \Sigma^{-1} \right) \right).
\]

Differentiating with respect to \( \mu \) and setting equal to zero yields the equation

\[-\frac{1}{2} \sum_{i=1}^{n} \Sigma^{-1} (X^{(i)} - \mu) = 0.
\]

From this we deduce that \( \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X^{(i)} \), the sample mean.

To find the maximum likelihood estimate for the covariance matrix we substitute \( K = \Sigma^{-1} \) and differentiate the log-likelihood with respect to an entry of \( K \). One makes use of the classical adjoint formula for the inverse of a matrix to see that

\[
\frac{\partial}{\partial k_{ij}} \log |K| = (1 + \delta_{ij}) \sigma_{ij}
\]

where \( \delta_{ij} \) is the Dirac delta function. Similarly,

\[
\frac{\partial}{\partial k_{ij}} \text{tr} \left( \sum_{i=1}^{n} ((X^{(i)} - \mu)(X^{(i)} - \mu)^T) K \right) = n(1 + \delta_{ij}) s_{ij}
\]

where \( S = \frac{1}{n} \sum_{i=1}^{n} ((X^{(i)} - \mu)(X^{(i)} - \mu)^T) \) is the sample covariance matrix. Putting these pieces together with our solution that the maximum likelihood estimate of \( \mu \) is \( \hat{\mu} \) gives that \( \hat{\Sigma} = \frac{1}{n} (X^{(i)} - \hat{\mu})(X^{(i)} - \hat{\mu})^T \). \( \square \)
Proposition 5.3.8. Let $\mathcal{M}_{1 \| 2}$ be the model of independence of two discrete random variables, with $r_1$ and $r_2$ states respectively. Let $u \in \mathbb{N}^{r_1 \times r_2}$ be the table of counts for this model obtained from i.i.d. samples from the model. Let $u_{i_1+} = \sum_{i_2} u_{i_1i_2}$ and $u_{+i_2} = \sum_{i_1} u_{i_1i_2}$ be the table marginals, and $n = \sum_{i_1,i_2} u_{i_1i_2}$ the sample size. Then the maximum likelihood estimate for a distribution $p \in \mathcal{M}_{1 \| 2}$ given the data $u$ is

$$\hat{p}_{i_1i_2} = \frac{u_{i_1+}u_{+i_2}}{n^2}.$$ 

Proof. A distribution $p \in \Delta_R$ belongs to the independence model if and only if we can write $p_{i_1i_2} = \alpha_{i_1}\beta_{i_2}$ for some $\alpha \in \Delta_{r_1-1}$ and $\beta \in \Delta_{r_2-1}$. We solve the likelihood equations in terms of $\alpha$ and $\beta$ and use them to find $\hat{p}$. Given a table of counts $u$, the log-likelihood function for a discrete random variable has the form

$$\ell(\alpha, \beta \mid u) = \sum_{i_1,i_2 \in R} u_{i_1i_2} \log p_{i_1i_2} = \sum_{i_1,i_2 \in R} u_{i_1i_2} \log \alpha_{i_1}\beta_{i_2} = \sum_{i_1 \in [r_1]} u_{i_1+} \log \alpha_{i_1} + \sum_{i_2 \in [r_2]} u_{+i_2} \log \beta_{i_2}.$$ 

From the last line, we see that we have two separate optimization problems that are independent of each other: maximizing with respect to $\alpha$ and maximizing with respect to $\beta$. Remembering that $\alpha_{r_1} = 1 - \sum_{i_1=1}^{r_1-1} \alpha_{i_1}$ and computing partial derivatives to optimize shows that $\hat{\alpha}_{i_1} = \frac{u_{i_1+}}{n}$. Similarly, $\hat{\beta}_{i_2} = \frac{u_{+i_2}}{n}$. 

Unlike the three preceding examples, most statistical models do not possess closed form expressions for their maximum likelihood estimates. The algebraic geometry of solving the critical equations will be discussed in later chapters, as will some numerical hill-climbing methods for approximating solutions.

5.4. Hypothesis Testing

A hypothesis test is a procedure given data for deciding whether or not a statistical hypothesis might be true. Typically, statistical hypotheses are phrased in terms of statistical models: for example, does the unknown distribution, about which we have collected i.i.d. samples, belong to a given model, or not. Note that statisticians are conservative so we rarely say that we accept the null hypothesis after performing a hypothesis test, only that
the test has not provided evidence against a particular hypothesis. The introduction of \( p\)-values provides more information about the hypothesis test besides simply whether or not we reject the null hypothesis.

As a simple example of a hypothesis test, suppose we have two random variables \( X \) and \( Y \) on the real line, and we suspect that they might be the same distribution. Hence the null hypothesis \( H_0 \) is that \( X = Y \) and the alternative hypothesis \( H_a \) is that \( X \neq Y \). This is stated as saying we are testing:

\[
H_0 : X = Y \text{ vs. } H_a : X \neq Y.
\]

We collect samples \( X^{(1)}, \ldots, X^{(n_1)} \) and \( Y^{(1)}, \ldots, Y^{(n_2)} \). If \( X = Y \) then these two empirical distributions should be similar, in particular, they should have nearly the same mean. Of course, the sampling means

\[
\bar{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X^{(i)}, \quad \bar{\mu}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y^{(i)}
\]

will not be exactly equal, and so we need a way to measure that these are close to each other up to the error induced by sampling.

The following permutation test (also called an exact test) gives a method, free from any distributional assumptions on the variables \( X \) and \( Y \), to test the null hypothesis versus the alternative hypothesis:

For the given data \( X^{(1)}, \ldots, X^{(n_1)} \) and \( Y^{(1)}, \ldots, Y^{(n_2)} \) calculate the absolute difference of means \( d = |\bar{\mu}_1 - \bar{\mu}_2| \). Now, group all of these samples together into one large set of samples \( Z_1, \ldots, Z_{n_1+n_2} \). For each partition \( A \mid B \) of \([n_1 + n_2]\) into a set \( A \) of size \( n_1 \) and a set \( B \) of size \( n_2 \), we compute the difference of sample means

\[
d(A, B) = \left| \frac{1}{n_1} \sum_{a \in A} Z_a - \frac{1}{n_2} \sum_{b \in B} Z_b \right|.
\]

If \( d(A, B) \) is larger than \( \bar{d} \) for few partitions \( A \mid B \), we are likely to conclude that \( X \neq Y \), since the test seems to indicate that the sample means are farther apart, more than would be expected to sampling error.

Of course, the crucial word “few” in the last sentence needs clarification. The proportion of partitions \( A \mid B \) such that \( d(A, B) \geq \bar{d} \) is called the \( p\)-value of the permutation test. A large \( p\)-value indicates that the permutation test provides no evidence against the null hypothesis, whereas a small \( p\)-value provides evidence against the null hypothesis. Fixing a level \( \alpha \), if the \( p\)-value is less than \( \alpha \) we reject the null hypothesis at level \( \alpha \). Otherwise, we are unable to reject the null hypothesis. Tradition in statistics sets the usual rejection level at \( \alpha = .05 \).
We will discuss a number of other hypothesis tests throughout the book, including the likelihood ratio test and Fisher’s exact test. In each case, we must define the \( p \)-value in a way that is appropriate to the given null hypothesis. This is stated most generally as follows. Most generally, when testing a specific hypothesis, we have a test statistic \( T \) which we evaluate on the data, which evaluates to zero if the data exactly fits the null hypothesis \( H_0 \), and which increases moving away from the null hypothesis. The \( p \)-value is then the probability under the null hypothesis that the test statistic has a larger value than the given observed value. In the case of the permutation test, we calculate the probability that a random partition of the data has differences of means above the value calculated from the data. Under the null hypothesis \( X = Y \), so the sample mean of each part of the partition converges to the common mean by the law of large numbers, so the test statistic is zero for perfect data that satisfies the null hypothesis.

**Example 5.4.1** (Jackal Mandible Lengths). The following dataset gives the mandible lengths in millimeters for 10 male and 10 female golden jackals in the collection of the British museum (the data set is taken from [Man07]).

<table>
<thead>
<tr>
<th>Males</th>
<th>120</th>
<th>107</th>
<th>110</th>
<th>116</th>
<th>114</th>
<th>111</th>
<th>113</th>
<th>117</th>
<th>114</th>
<th>112</th>
</tr>
</thead>
<tbody>
<tr>
<td>Females</td>
<td>110</td>
<td>111</td>
<td>107</td>
<td>108</td>
<td>110</td>
<td>105</td>
<td>107</td>
<td>106</td>
<td>111</td>
<td>111</td>
</tr>
</tbody>
</table>

A basic question about this data is whether the distribution of mandible lengths is the same or different for males versus females of this species. A quick look at the data suggests that the males tend to have larger mandibles than females, but is there enough data that we can really conclude this? Let \( \mu_m \) denote the mean of the male population and \( \mu_f \) denote the mean of the female population. According to our discussion above, we can perform the hypothesis test to compare

\[
H_0 : \mu_m = \mu_f \quad \text{vs.} \quad H_a : \mu_m \neq \mu_f.
\]

An exact calculation yields the \( p \)-value of 0.00333413, suggesting we can reject the null hypothesis that the means are equal. We run the following code in R [R C16] to perform the hypothesis test.

```r
m <- c(120, 107, 110, 116, 114, 111, 113, 117, 114, 112)
f <- c(110, 111, 107, 108, 110, 105, 107, 106, 111, 111)
observed_tstat <- abs(mean(m) - mean(f))
resampled_tstats <- replicate(1e6, {
  ndcs <- sample(length(c(m, f)), length(m))
  abs(mean(c(m, f)[ndcs]) - mean(c(m, f)[-ndcs]))
})
mean(resampled_tstats >= observed_tstat)
```
In this case the $p$-value is approximated by taking random samples, rather than computing all partitions. The approximate $p$-value on one run with 1000000 samples was 0.003422.

An important remark about hypothesis testing is that we rarely calculate the $p$-value exactly. Even in the permutation test, it is not possible to give a closed form formula for the $p$-value nor is it possible for large sample sizes to compute the sum over all partitions $A|B$. In that case, we can randomly sample partitions to approximate the $p$-value. That is, let $A_1|B_1, \ldots, A_k|B_k$ be random partitions of the data of size where $A_i$ has $n_1$ elements and $B_i$ has $n_2$ elements. An approximation of the $p$-value of the permutation test is

$$\frac{1}{k} \sum_{i=1}^{k} 1_{d(A_i|B_i) > \bar{d}}$$

which converges to the exact $p$-value as $k$ tends to infinity by the law of large numbers.

**Example 5.4.2** (Calculator random digits). A sequence of 300 pseudo-random digits was generated on a Casio fx-3600p calculator [DDt87]. The table displays the number of times that each digit occurred.

<table>
<thead>
<tr>
<th>Digit</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>1</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>35</td>
</tr>
<tr>
<td>5</td>
<td>31</td>
</tr>
<tr>
<td>6</td>
<td>27</td>
</tr>
<tr>
<td>7</td>
<td>33</td>
</tr>
<tr>
<td>8</td>
<td>32</td>
</tr>
<tr>
<td>9</td>
<td>35</td>
</tr>
</tbody>
</table>

Our null hypothesis in this case is that the digits are uniformly random, and we would like to perform a hypothesis test to determine if we have evidence against this hypothesis. There are a number of ways to test this hypothesis. All involve comparing a test statistic evaluated on the data to the same test statistic on random samples of data generated under the null hypothesis. Under the null hypothesis, the expected frequency of each bit is $\hat{u}_i = 30$. A natural choice is to use the “distance” between the expected frequency vector $\hat{u}$ a data vector $u$ as a natural test statistic. The distance could be a standard Euclidean distance or some other measure of the dissimilarity of the observation. For example, a distance frequently used in hypothesis tests is the Pearson’s $X^2$ statistic, which in this case is

$$X^2(u) = \sum_{i=0}^{9} \frac{(\hat{u}_i - u_i)^2}{\hat{u}_i}.$$  

Note that because $\hat{u}_i = 30$ is constant in this example, both hypothesis tests will produce the same result. We generated a random sample of 100000 data sets $v$ from the multinomial distribution with $n = 300$ and $p_i = 1/10$, and compared $X^2(v)$ to $X^2(u)$. The following R [R C16] code performs this hypothesis test:
compute_pearson_x2 <- function(o, e) sum((o - e)^2 / e)
( observed_x2 <- compute_pearson_x2(data, 300 * rep(1/10, 10)) )
samples <- rmultinom(1e6, 300, prob = rep(1/10, 10))
sampled_x2s <- apply(samples, 2, compute_pearson_x2, e = 300 * rep(1/10, 10))
mean(sampled_x2s >= observed_x2)

We found a p-value of approximately .896, which suggests that we cannot reject the null hypothesis that the samples are uniformly random.

In other contexts (e.g. the likelihood ratio test), a standard strategy is to assume that the sample size tends to infinity. In this case, it is possible to derive the asymptotic distribution of the test statistic. For smooth models, those asymptotic distributions are familiar distributions of probability theory like the chi-square distribution, and this allows us to compute asymptotic p-values. However, for models that are not smooth, nonstandard limiting distributions for the asymptotic p-values can arise. This is discussed in more detail in Section 7.4, but we illustrate the basic idea first.

**Example 5.4.3** (Calculator random digits). Suppose that we have i.i.d. samples from some unknown distribution \( q \in \Delta_{r-1} \) and we want to test the hypothesis that the distribution is equal to some fixed distribution \( p \). Assuming that all coordinates of \( p \) are nonzero, it can be shown that if the null hypothesis is true, as \( n \to \infty \), the distribution of the Pearson \( X^2 \) statistic converges to a chi-square distribution with \( r - 1 \) degrees of freedom. This yields the basis of an asymptotic test, since we can calculate the empirical \( X^2 \) value of our data vector \( u \) and determine the asymptotic p-value

\[
P(Q > X^2(u))
\]

where \( Q \sim \chi^2_{r-1} \). For the calculator random digit data set from Example 5.4.2 this is straight-forward to check in R \([R \ C16]\) with the command:

chisq.test(data, NULL, rep(1/10, 10))

This returns the asymptotic p-value of 0.893 which is very close to the exact test p-value calculated in Example 5.4.2.

5.5. Bayesian Statistics

Likelihood inference, hypothesis tests, and p-values are all standard tools from frequentist statistics. Frequentist statistics is typically concerned with the long term behavior of estimators, in particular, how they behave as more and more data comes in for analyzing a problem. In frequentist statistics approaches to parametric statistics models, it is assumed that there is a
true underlying parameter vector, and we seek to find information about that parameter by collecting data.

A complementary approach in statistics is called Bayesian statistics. In Bayesian statistics, the point of collecting data is to update our beliefs about unknown parameters. Typically we have prior beliefs about the parameters, represented by a prior distribution $\pi$ on the parameter space $\Theta$. Data $D$ allows us to compute an update on the distribution on $\Theta$ which gives us the posterior distribution. The posterior distribution is computed via the formula

$$p(\theta \mid D) = \frac{p(D \mid \theta)\pi(\theta)}{\int_{\theta} p(D \mid \theta)\pi(\theta) \, d\theta}.$$  

Note that the quantity in the denominator $\int_{\theta} p(D \mid \theta)\pi(\theta) \, d\theta$ is the marginal probability of the data, and is called the marginal likelihood integral. This application of Bayes theorem in the calculation of the posterior distribution is the reason for referring to these methods as Bayesian.

As with many statistical methods, the main difficulty in using Bayesian methods in practice are computational. There are usually no closed form for the posterior distribution and they must be approximated, often using Monte Carlo algorithms. Similarly, the marginal likelihood is usually not easy to calculate and must itself be approximated.

Bayesian methods produce a distribution over the parameter space, which can be thought of as our posterior belief about the parameter. When the parameter space is high dimensional, it can be useful or necessary to compute a point estimate of the parameter. A natural choice in Bayesian statistics is the maximum a posteriori estimate or MAP estimate. This is simply the parameter value that maximizes the posterior distribution.

**Definition 5.5.1.** Let $\mathcal{M}_\Theta$ be a parametric statistical model, $\pi$ a prior distribution over $\Theta$, $D$ a collection of data, and $p(\theta \mid D)$ the posterior distribution of $\theta$ given the data. The maximum a posteriori estimator of $\theta$, (MAP) is

$$\hat{\theta} = \arg \max_{\theta \in \Theta} p(\theta \mid D).$$

Note the similarity of the definition of the MAP estimate to the MLE. These estimates can be vastly different depending on the underlying prior distribution, though increasing amounts of data usually (as one would hope) eliminates the dependence on the prior in the limit. One advantage of the MAP estimate is that it is not necessary to compute the marginal likelihood integral $\int_{\theta} p(D \mid \theta)\pi(\theta) \, d\theta$ to obtain it.

**Example 5.5.2** (Prior for Discrete Random Variable). Consider a discrete random variable with $r$ states. We will consider a natural prior distribution,
the *Dirichlet distribution*, which is commonly used as a prior distribution on the parameter space in models with discrete random variables.

Let \( \alpha \in \mathbb{R}^r_{>0} \) be a nonnegative real parameter vector. A random variable \( X \in \Delta_{r-1} \) has Dirichlet distribution with parameter \( \alpha \), (denoted \( X \sim \text{Dir}(\alpha) \)) if the density function \( f_\alpha(x) \propto \prod_{i=1}^r x_i^{\alpha_i-1} \). The proportionality constant \( B(\alpha) \) is the multivariate beta function
\[
B(\alpha) = \frac{\prod_{i=1}^r \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^r \alpha_i)}.
\]
If \( \alpha_i > 1 \) for all \( i \), then Dirichlet distribution has peak (both mean and mode) at \( \alpha/\|\alpha\|_1 \), and the peak becomes narrower for larger values of \( \|\alpha\|_1 \). As a prior distribution for the multinomial distribution, the Dirichlet distribution gives concentration around the mode \( \alpha/\|\alpha\|_1 \) with \( \|\alpha\|_1 \) giving a measure of certainty about the mode.

Now suppose we receive \( n \) i.i.d. samples from a discrete random variable \( Y \) with state space \([r]\). Since we have i.i.d. samples, we can use the vector of counts \( u \) to record the data since it is a sufficient statistic. Given a sample size \( n \), the conditional distribution of \( u \) given \( X \) is multinomial, so the joint density of \( u \) and \( X \) is
\[
f(x, u) = B(\alpha) \prod_{i=1}^r x_i^{\alpha_i-1} \times \left( \frac{n}{u} \right) \prod_{i=1}^r x_i^{u_i}.
\]
Whatever the marginal distribution of \( u \) is, we see that the conditional density of \( X \) given \( u \) is
\[
f(x \mid u) \propto \prod_{i=1}^r x_i^{u_i+\alpha_i-1}.
\]
That is, the posterior distribution of the parameter given the data \( u \) is a Dirichlet distribution with parameter \( \alpha + u \). So assuming that all \( u_i > 0 \), the maximum a posterior estimate of the parameter given the data \( u \) is the vector \( (u + \alpha)/\|u + \alpha\|_1 \).

Now suppose that there is some true underlying distribution \( p \) which generates the samples. As the sample size \( n \) gets large, the law of large numbers shows that \( u/n \) will converge to \( p \). Since \( \alpha \) was some a priori fixed value, this will imply that \( (u + \alpha)/\|u + \alpha\|_1 \) also converges to \( p \) as \( n \) goes to infinity. Hence, regardless of what particular \( \alpha \) we chose at the start of our analysis, eventually we end up with a good estimate of the true underlying parameter \( p \).

An interesting fact about this example is that the posterior distribution has the same functional form as the prior distribution. This does not always happen, but happens here because the Dirichlet distribution is the *conjugate prior* to the multinomial distribution.
Bayesian and frequentist statistics are sometimes seen as being at odds with each other. We do not propose to “take sides” in the present work, rather seeing them both as useful methods which make sense for different situations, and, from the standpoint of algebraic statistics, both are rich with connections to algebraic geometry. Bayesian approaches make sense when we wish to incorporate our prior beliefs about parameters into our analysis of data. Frequentist perspectives make sense when we have lots of data arriving and we can analyze asymptotic properties of estimators. We will see specific discussion of Bayesian methods in the chapters on model selection and parametric inference.

5.6. Exercises

Exercise 5.1. An exponential random variable $X$ with parameter $\lambda > 0$ has density function

$$f_\lambda(x) = \begin{cases} \lambda \exp(-\lambda x) & x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Let $X^{(1)}, \ldots, X^{(n)}$ be i.i.d. exponential random variables. What is the maximum likelihood estimate of $\lambda$? What is the method of moments estimate of $\lambda$?

Exercise 5.2. Consider the two lists of, respectively, 15 and 12 real numbers.

$A = \left\{1.04, -0.12, 0.22, 0.08, -0.06, -0.11, 0.06, 1.06, \\ 1.73, -0.58, 0.67, 0.06, 1.10, -0.93, 0.34 \right\}$

$B = \left\{3.36, 1.00, 1.15, -0.57, 0.84, 0.18, 0.28, 0.52, \\ 2.44, -4.22, 2.95, -1.83 \right\}$

The elements of $A$ are i.i.d. samples from a probability distributions $p$ and the elements of $B$ are i.i.d. samples from a probability distribution $q$. Do you think $p = q$? Formulate and apply a hypothesis test, and calculate $p$-values.

Exercise 5.3. A gamma random variable with parameters $\alpha > 0, \beta > 0$ has density function

$$f(x | \alpha, \beta) = \begin{cases} \Gamma(\alpha)^{-1}\beta^\alpha x^{\alpha-1} \exp(-\beta x) & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x)dx$ denotes the gamma function. Suppose that we use the gamma distribution with parameters $\alpha, \beta$ as a prior on $\lambda$ in an exponential random variable. We observe i.i.d. samples $X^{(1)}, \ldots, X^{(n)}$. What is the posterior distribution of $\lambda$ given the data?
Chapter 6

Exponential Families

In this chapter we introduce exponential families. This class of statistical models plays an important role in modern statistics because it provides a broad framework for describing statistical models. The most commonly studied families of probability distributions are exponential families, including the families of jointly normal random variables, the exponential, Poisson, and multinomial models. Exponential families have been well-studied in the statistics literature. Some commonly used references include [BN78, Bro86, Che72, Let92].

We will see a large number of statistical models that are naturally interpreted as submodels of regular exponential families. The way in which the model is a submodel is in two ways: first, that the parameter of the model sit as a subset of the parameter space, and that the sufficient statistics of the exponential family then maps the data into the cone of sufficient statistics of the larger exponential families. This allows the mathematical analysis of many interesting and complicated statistical models beyond the regular exponential families.

The statistical literature has largely focused on the case where those submodels of exponential families are given by smooth submanifolds of the parameter space. These are known as curved exponential families in the literature. We focus here on the situation where the submodel is a (not-necessarily smooth) semi-algebraic subset of the natural parameter space of the exponential family which are called algebraic exponential families. Semi-algebraic sets are the most fundamental objects of real algebraic geometry and we will review their definitions and properties in Section 6.4. Algebraic exponential families will play a major role throughout the remainder of the text.
6. Exponential Families

Besides being a useful framework for generalizing and unifying families of statistical models for broad analysis, the class of exponential families also satisfies useful properties that make them pleasant to work with in statistical analyses. For example, all regular exponential families have concave likelihood functions which implies that hill-climbing approaches can be used to find the maximum likelihood estimates given data (see Section 7.3). These models also have nice conjugate prior distributions, which make them convenient to use in Bayesian analysis.

6.1. Regular Exponential Families

Consider a sample space \( \mathcal{X} \) with \( \sigma \)-algebra \( \mathcal{A} \) on which is defined a \( \sigma \)-finite measure \( \nu \). Let \( T : \mathcal{X} \to \mathbb{R}^k \) be a statistic, i.e., a measurable map, and \( h(x) : \mathcal{X} \to \mathbb{R}_{>0} \) a measurable function. Define the natural parameter space

\[
N = \left\{ \eta \in \mathbb{R}^k : \int_{\mathcal{X}} h(x) \cdot e^{\eta^T(x)} d\nu(x) < \infty \right\}.
\]

For \( \eta \in N \), we can define a probability density \( p_\eta \) on \( \mathcal{X} \) as

\[
p_\eta(x) = h(x)e^{\eta^T(x) - \phi(\eta)},
\]

where

\[
\phi(\eta) = \log \int_{\mathcal{X}} h(x)e^{\eta^T(x)} d\nu(x).
\]

Let \( P_\eta \) be the probability measure on \((\mathcal{X}, \mathcal{A})\) that has \( \nu \)-density \( p_\eta \). Define \( \nu^T = \nu \circ T^{-1} \) to be the measure that the statistic \( T \) induces on the Borel \( \sigma \)-algebra of \( \mathbb{R}^k \). The support of \( \nu^T \) is the intersection of all closed sets \( A \subseteq \mathbb{R}^k \) that satisfy \( \nu^T(\mathbb{R}^k \setminus A) = 0 \). The function \( \phi(\eta) \) can be interpreted as the logarithm of the Laplace transform of \( \nu^T \). Recall that the affine dimension of \( A \subseteq \mathbb{R}^k \) is the dimension of the linear space spanned by all differences \( x - y \) of two vectors \( x, y \in A \).

Definition 6.1.1. Let \( k \) be a positive integer. The probability distributions \((P_\eta : \eta \in N)\) form a regular exponential family of order \( k \) if \( N \) is an open set in \( \mathbb{R}^k \) and the affine dimension of the support of \( \nu^T \) is equal to \( k \). The statistic \( T(x) \) that induces the regular exponential family is called a canonical sufficient statistic.

The order of a regular exponential family is unique and if the same family is represented using two different canonical sufficient statistics then those two statistics are non-singular affine transforms of each other [Bro86, Thm. 1.9].

Regular exponential families comprising families of discrete distributions have been the subject of much of the work on algebraic statistics.
Example 6.1.2 (Discrete random variables). Let the sample space $\mathcal{X}$ be the set of integers $[r]$. Let $\nu$ be the counting measure on $\mathcal{X}$, i.e., the measure $\nu(A)$ of $A \subseteq \mathcal{X}$ is equal to the cardinality of $A$. Consider the statistic $T : \mathcal{X} \to \mathbb{R}^{r-1}$

$$T(x) = (1_{\{1\}}(x), \ldots, 1_{\{r-1\}}(x))^t,$$

whose zero-one components indicate which value in $\mathcal{X}$ the argument $x$ is equal to. In particular, when $x = r$, $T(x)$ is the zero vector. Let $h(x) = 1$ for all $x \in [r]$. The induced measure $\nu^T$ is a measure on the Borel $\sigma$-algebra of $\mathbb{R}^{r-1}$ with support equal to the $r$ vectors in $\{0, 1\}^{r-1}$ that have at most one non-zero component. The differences of these $r$ vectors include all canonical basis vectors of $\mathbb{R}^{r-1}$. Hence, the affine dimension of the support of $\nu^T$ is equal to $r - 1$.

It holds for all $\eta \in \mathbb{R}^{r-1}$ that

$$\phi(\eta) = \log \int_{\mathcal{X}} e^{\eta^T(x)} d\nu(x) = \log \left( 1 + \sum_{x=1}^{r-1} e^{\eta_x} \right) < \infty.$$ 

Hence, the natural parameter space $N$ is equal to all of $\mathbb{R}^{r-1}$ and in particular is open. The $\nu$-density $p_{\eta}$ is a probability vector in $\mathbb{R}^r$. The components $p_{\eta}(x)$ for $1 \leq x \leq r - 1$ are positive and given by

$$p_{\eta}(x) = \frac{e^{\eta_x}}{1 + \sum_{x=1}^{r-1} e^{\eta_x}}.$$ 

The last component of $p_{\eta}$ is also positive and equals

$$p_{\eta}(r) = 1 - \sum_{x=1}^{r-1} p_{\eta}(x) = \frac{1}{1 + \sum_{x=1}^{r-1} e^{\eta_x}}.$$ 

The family of induced probability distribution $(P_{\eta} : \eta \in \mathbb{R}^{r-1})$ is a regular exponential family of order $r - 1$. The natural parameters $\eta_x$ can be interpreted as log-odds ratios because $p_{\eta}$ is equal to a given positive probability vector $(p_1, \ldots, p_r)$ if and only if $\eta_x = \log(p_x/p_r)$ for $x = 1, \ldots, r - 1$. This establishes a correspondence between the natural parameter space $N = \mathbb{R}^{r-1}$ and the interior of the $r - 1$ dimensional probability simplex $\Delta_{r-1}$.

The other distributional framework that has seen application of algebraic geometry is that of multivariate normal distributions.

Example 6.1.3 (Multivariate normal random variables). Let the sample space $\mathcal{X}$ be Euclidean space $\mathbb{R}^m$ equipped with its Borel $\sigma$-algebra and Lebesgue measure $\nu$. Consider the statistic $T : \mathcal{X} \to \mathbb{R}^m \times \mathbb{R}^{m(m+1)/2}$ given by

$$T(x) = (x_1, \ldots, x_m, -x_1^2/2, \ldots, -x_m^2/2, -x_1 x_2, \ldots, -x_{m-1} x_m)^t.$$
The polynomial functions that form the components of $T(x)$ are linearly independent and thus the support of $\nu^T$ has the full affine dimension $m + m(m + 1)/2$. Let $h(x) = 1$ for all $x \in \mathbb{R}^m$.

If $\eta \in \mathbb{R}^m \times \mathbb{R}^{m(m+1)/2}$, then write $\eta_{[m]} \in \mathbb{R}^m$ for the vector of the first $p$ components $\eta_i$, $1 \leq i \leq m$. Similarly, write $\eta_{[m \times m]}$ for the symmetric $m \times m$-matrix formed from the last $m(m + 1)/2$ components $\eta_{ij}$, $1 \leq i \leq j \leq m$. The function $x \mapsto e^{\eta^T T(x)}$ is $\nu$-integrable if and only if $\eta_{[m \times m]}$ is positive definite. Hence, the natural parameter space $\mathcal{N}$ is equal to the Cartesian product of $\mathbb{R}^m$ and the cone of positive definite $m \times m$-matrices. If $\eta$ is in the open set $\mathcal{N}$, then

$$\phi(\eta) = -\frac{1}{2} \left( \log \det(\eta_{[m \times m]}) - \eta_{[m]}^T \eta_{[m]} - m \log(2\pi) \right).$$

The Lebesgue densities $p_\eta$ can be written as

$$p_\eta(x) = \frac{1}{\sqrt{(2\pi)^m \det(\eta_{[m \times m]})}} \times \exp \left\{ \eta_{[m]}^T x - \frac{1}{2} \eta_{[m]}^T \text{tr}(x^T \eta_{[m \times m]} x) \right\}.$$

Here we again use the trace trick, since $x^T \eta_{[m \times m]} x = \text{tr}(x^T \eta_{[m \times m]} x)$ allows us to express the quadratic form in $x$ as a dot product between $\eta_{[m \times m]}$ and the vector of sufficient statistics. Setting $\Sigma = \eta_{[m \times m]}^{-1}$ and $\mu = \eta_{[m \times m]}^{-1} \eta_{[m]}$, we find

$$p_\eta(x) = \frac{1}{\sqrt{(2\pi)^m \det(\Sigma)}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}$$

is the density of the multivariate normal distribution $\mathcal{N}_m(\mu, \Sigma)$. Hence, the family of all multivariate normal distributions on $\mathbb{R}^m$ with positive definite covariance matrix is a regular exponential family of order $m + m(m + 1)/2$.

The inverse of the covariance matrix, $\Sigma$, plays an important role for Gaussian models, as a natural parameter of the exponential family, and through its appearance in saturated conditional independence constraints (as in Chapter 4). For this reason, the inverse covariance matrix receives a special name: the concentration matrix or the precision matrix. The concentration matrix is often denoted by $K = \Sigma^{-1}$.

In some circumstances it is useful to consider the closure of an exponential family in a suitable representation. This occurs most commonly in the case of discrete exponential families, where we saw that the representation of the multinomial model gives the interior of the probability simplex. Taking the closure yields the entire probability simplex. The set of all distributions which lie in the closure of a regular exponential family is called
an extended exponential family. The extended exponential family might involve probability distributions that do not possess densities. For example, in the Gaussian case the closure operation yields covariance matrices that are positive semidefinite, but singular. These yield probability distributions that are supported on lower dimensional linear spaces, and hence do not have densities. See [CM08, Mat15].

Our goal in this chapter is to describe statistical models which naturally can be interpreted as submodels of exponential families through their natural parameter spaces or transformed versions of those parameter spaces. Before moving to the general setup of algebraic exponential families in Section 6.5, we focus on the special case where the submodel is itself a regular exponential family. Such a submodel is realized as a linear subspace of the natural parameter space. In the case of discrete random variables, this leads to the study of toric varieties, which we discuss in Section 6.2. In the case of Gaussian random variables, this yields inverse linear spaces, which are described in Section 6.3.

6.2. Discrete Regular Exponential Families

In this section we explore regular exponential families for discrete random variables with a finite number of states, and how they are related to objects in algebra. Since $X$ is assumed to be discrete with state space $X = [r]$, there are a number of simplifications that can be made to the representation of the regular exponential family from the presentation in the previous section. First of all, the statistic $T(x)$ gives a measurable map from $[r]$ to $\mathbb{R}^k$. This amounts to associating a vector $T(x) \in \mathbb{R}^k$ to each $x \in [r]$. Similarly, the measurable function $h : [r] \to \mathbb{R}_{>0}$ merely associates a nonnegative real number to each $x \in [r]$. Any nonnegative function is measurable in this setting and any list of vectors $T(1), \ldots, T(r)$ could be used. The Laplace transform $\int_X h(x) e^{\eta T(x)} d\nu$ is the sum

$$Z(\eta) = \sum_{x \in [r]} h(x) e^{\eta T(x)}.$$

Now we look at the individual elements of the product. Let us call the vector $T(x) = a_x$, written explicitly as $a_x = (a_{1x}, \ldots, a_{kx})^t$. Similarly, we define the vector $\eta = (\eta_1, \ldots, \eta_r)^t$ and $\theta_i = \exp(\eta_i)$. For notational simplicity, we denote $h(x) = h_x$ so that $h = (h_1, \ldots, h_r)$ is a vector in $\mathbb{R}_{>0}$. This means that

$$p_\eta(x) = h(x) e^{\eta T(x) - \phi(\eta)}$$

could also be rewritten as

$$p_\eta(x) = \frac{1}{Z(\theta)} h_x \prod_j \theta_j^{a_{jx}}.$$
where
\[ Z(\theta) = \sum_{x \in [r]} h_x \prod_{j} \theta_{j}^{a_{jx}}. \]

In particular, if the \( a_{jx} \) are integers for all \( j \) and \( x \), then the discrete exponential family is a parametrized family of probability distribution, where the parametrizing functions are rational functions.

A third and equivalent representation of the model in this discrete setting arises by asking to determine whether a given probability distribution \( p \in \text{int}(\Delta_{r-1}) \) belongs to the discrete exponential family associated to \( A = (a_{jx})_{j \in [k], x \in [r]} \in \mathbb{Z}^{k \times r} \) and \( h \in \mathbb{R}_{>0}^{r} \). Indeed, taking the logarithm of the exponential family representation, we arrive at
\[
\log p_{\theta}(x) = \log h(x) + \eta^{T}(x) - \phi(\eta)
\]
or equivalently
\[
\log p_{\theta}(x) = \log(h_x) + \log(\theta)^{T}A - \log Z(\theta).
\]

If we make the assumption, which we will do throughout, that the matrix \( A \) contains the vector \( 1 = (1, 1, \ldots, 1) \) in its rowspan, then this is equivalent to requiring that \( \log p \) belong to the affine space \( \log(h) + \text{rowspan}(A) \). For this reason, discrete exponential families are also called log-affine models. This connection also explains the interpretation of these models as arising from taking linear subspaces in the natural parameter space of the full regular exponential family for discrete random variables.

**Definition 6.2.1.** Let \( A \in \mathbb{Z}^{k \times r} \) be a matrix of integers such that \( 1 \in \text{rowspan}(A) \) and let \( h \in \mathbb{R}_{>0}^{r} \). The **log-affine model** associated to these data is the set of probability distributions
\[
\mathcal{M}_{A,h} := \{ p \in \text{int}(\Delta_{r-1}) : \log p \in \log(h) + \text{rowspan}(A) \}.
\]
If \( h = 1 \) then \( \mathcal{M}_A = \mathcal{M}_{A,1} \) is called a **log-linear model**.

In addition to the names exponential families and log-linear models, these models are also known in the algebraic statistics community as a **toric models**. This comes from their relation to toric varieties, as we now explain.

**Definition 6.2.2.** Let \( h \in \mathbb{R}_{\geq 0}^{r} \) be a vector of nonnegative numbers and \( A = (a_{ij}) \in \mathbb{Z}^{k \times r} \) a \( k \times r \) matrix of integers and suppose that \( 1 \in \text{rowspan}(A) \). The monomial map associated to this data is the rational map
\[ \phi^{A,h} : \mathbb{R}^{k} \to \mathbb{R}^{r}, \quad \phi^{A,h}_{j}(\theta) = h_{j} \prod_{i=1}^{k} \theta_{i}^{a_{ij}}. \]
6.2. Discrete Regular Exponential Families

Note that in the definition of the rational map \( \phi^{A,h} \) we have removed the normalizing constant \( Z(\theta) \). This has the effect of homogenizing the parametrization so that we can compute the homogeneous vanishing ideal of the model.

**Definition 6.2.3.** Let \( h \in \mathbb{R}_{\geq 0}^r \) be a vector of nonnegative numbers and \( A = (a_{ij}) \in \mathbb{Z}^{k \times r} \) a \( k \times r \) matrix. The ideal

\[
I_{A,h} := I(\phi^{A,h}(\mathbb{R}^k)) \subseteq \mathbb{R}[p]
\]

is called the toric ideal associated to the pair \( A \) and \( h \). In the special case that \( h = 1 \) we denote this as \( I_A := I_{A,1} \).

Note that generators for the toric ideal \( I_{A,h} \) are easily obtained from generators of the toric ideal \( I_A \), by globally making the substitution \( p_j \mapsto p_j/h_j \). Hence, it is sufficient to focus on the case of the toric ideal \( I_A \). These ideals turn out to be binomial ideals.

**Proposition 6.2.4.** Let \( A \in \mathbb{Z}^{k \times r} \) be a \( k \times r \) matrix of integers. Then the toric ideal \( I_A \) is a binomial ideal and

\[
I_A = \langle p^u - p^v : u, v \in \mathbb{N}^r \text{ and } Au = Av \rangle.
\]

If \( 1 \in \operatorname{rowspan}(A) \) then \( I_A \) is homogeneous.

**Proof.** First of all, it is clear that any binomial \( p^u - p^v \) such that \( Au = Av \) is in the toric ideal \( I_A \). To show that \( I_A \) is generated by all these binomials, we will show something stronger: in fact, these binomials span the ideal \( I_A \) as a vector space. Indeed, let \( f(p) \in I_A \) and \( c_u p^u \) any monomial appearing in \( f \) with nonzero coefficient. For any choice of \( \theta \), we must have \( f(\phi^A(\theta)) = 0 \). This means, that \( f(\phi^A(\theta)) = 0 \) as a polynomial in \( \theta \). Plugging in \( \phi^A(\theta) \) into the monomial \( c_u p^u \) yields \( c_u \theta^{Au} \). For this to cancel, there must be some other monomial \( c_v p^v \) such that plugging in \( \phi^A(\theta) \) yields the same monomial (with possibly different coefficient) \( c_v \theta^{Au} \), so \( Au = Av \). Forming the new polynomial \( f - c_u(p^u - p^v) \) yields a polynomial with fewer terms. The result is then achieved by induction on the number of terms.

To see that \( I_A \) is homogeneous, note that \( 1 \in \operatorname{rowspan}(A) \) forces that \( Au = Av \) implies that \( 1u = 1v \). Thus all of the generating binomials \( p^u - p^v \in I_A \) are homogeneous. \( \square \)

Note that toric ideals are also lattice ideals, as introduced in Chapter 4. Indeed, \( I_A = I_L \) where \( L \) is the lattice \( L = \ker_{\mathbb{Z}} A \). Note further that if \( h = 1 \), then the ideal generators can always be chosen to be binomials whose coefficients are \( \pm 1 \), so the coefficient field of the ring does not matter.
Example 6.2.5 (Twisted Cubic). Let $A = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 3 & 2 & 1 & 0 \end{pmatrix}$. The toric ideal $I_A$ is the vanishing ideal of the parametrization $p_1 = \theta_2^3$, $p_2 = \theta_1 \theta_2^2$, $p_3 = \theta_1^2 \theta_2$, $p_4 = \theta_1^3$. The toric ideal is generated by three quadratic binomials $I_A = \langle p_1 p_3 - p_2^2, p_1 p_4 - p_2 p_3, p_2 p_4 - p_3^2 \rangle$.

For example, the first binomial $p_1 p_3 - p_2^2$ is $p^u - p^v$ where $u = (1,0,1,0)^T$ and $v = (0,2,0,0)^T$. One checks that $Au = Av$ in this case. The sufficient statistic of this model, obtained from the vector of counts $u$, is the vector $Au$. The variety $V(I_A) \subseteq \mathbb{P}^3$ is the twisted cubic curve, that we first saw in Example 3.1.3 in its affine form.

Note that if we take $h = (1,3,3,1)$, then $I_{A,h}$ gives the vanishing ideal of the model of a binomial random variable with three trials.

Example 6.2.6 (Discrete Independent Random Variables). Any time we have a variety parametrized by a monomial functions of parameters the vanishing ideal is a toric ideal. For example, consider the parametrization

$$p_{ij} = \alpha_i \beta_j$$

where $i \in [r_1]$, $j \in [r_2]$ and $\alpha_i$ and $\beta_j$ are independent parameters. This is, of course, the (homogeneous) parametrization of two discrete independent random variables. The matrix $A$ representing the toric ideal is a $(r_1 + r_2) \times (r_1 r_2)$ matrix, with one row for each parameter and one column for each state of the pair of random variables. The $i,j$-th column of $A$, yielding the exponent vector for $p_{ij} = \alpha_i \beta_j$, is the vector $e_i \oplus e_j'$, where $e_i$ denotes the $i$th standard unit vector in $\mathbb{R}^{r_1}$ and $e_j'$ denotes the $j$th standard unit vector in $\mathbb{R}^{r_2}$. For example, for $r_1 = 3$ and $r_2 = 4$, we get the matrix

$$\begin{pmatrix}
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.$$ 

Note that the matrix $A$ acts as a linear operator that computes the row and column sums of the $r_1 \times r_2$ array $u$. The toric ideal $I_A$ is generated by all binomials $p^u - p^v$ such that $u, v$ have the same sufficient statistics under the model. This means that the arrays $u$ and $v$ have the same row and column sums. Note that $I_A = I_{\mathbb{R}_{r_1}, \mathbb{R}_{r_2}}$, so we already know that $I_A$ is generated by the $2 \times 2$ minors of a generic matrix

$$I_A = \langle p_{i_1j_1} p_{i_2j_2} - p_{i_1j_2} p_{i_2j_1} : i_1, i_2 \in [r_1], j_1, j_2 \in [r_2] \rangle.$$
Example 6.2.7 (Complete Independence). Let $A$ be the $6 \times 8$ matrix

$$A = \begin{pmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 1 \\
1 & 1 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 1
\end{pmatrix}. $$

The following Singular code computes a Gröbner basis for the toric ideal $I_A$.

```plaintext
ring r = 0, (p1,p2,p3,p4,p5,p6,p7,p8),dp;
intmat A[6][8] = 1,1,1,1,0,0,0,0,  
0,0,0,1,1,1,1,  
1,1,0,0,1,1,0,0,  
0,0,1,1,0,1,1,  
1,0,1,0,1,0,1,0,  
0,1,0,1,0,1,0,1;
intvec v = 1,1,1,1,1,1,1,1;
LIB "toric.lib";
ideal I = toric_ideal(A,"hs",v);
std(I);
```

Note that Singular has a built-in library for computing with toric ideals. Macaulay2 also has such functionality, using the package 4ti2 [HH03], which we will illustrate in Example 9.2.6.

The ideal $I_A$ has nine homogeneous quadratic generators,

$$\langle p_6p_7 - p_5p_8, \ldots, p_2p_3 - p_1p_4 \rangle.$$

Renaming the variables, $p_{111}, \ldots, p_{222}$, we see that the ideal $I_A$ is equal to the conditional independence ideal $I_C$ of complete independence of three binary random variables, that is, with

$$C = \{1 \perp \{2, 3\}, 2 \perp \{1, 3\}, 3 \perp \{1, 2\}\}.$$

6.3. Gaussian Regular Exponential Families

In this section we describe the regular exponential families that arise as subfamilies of the regular multivariate Gaussian statistical models. To make such a model, we should choose a statistic $T(x)$ that maps $x \in \mathbb{R}^m$ to a vector of degree 2 polynomials in $x$ with no constant term. Equivalently, such a model is obtained by taking a linear subspace of the natural parameter space of the regular Gaussian exponential family.

In principle, we could consider any set of the form $L \cap (\mathbb{R}^m \times PD_m)$ where $L$ is a linear space in the natural parameter space of pairs $(\Sigma^{-1}\mu, \Sigma^{-1})$ where $\Sigma$ is the covariance matrix of the multivariate normal vector.
The most commonly occurring case in applications is where the linear space $L$ is actually a direct product $L = L_1 \times L_2$ with $L_1 \subseteq \mathbb{R}^m$ and $L_2 \subseteq \mathbb{R}^{(m+1)m/2}$. Often $L_1$ is either $\{0\}$ or $\mathbb{R}^m$. For the purposes of investigating the vanishing ideal of such a model, this assumption means that the vanishing ideal breaks into the ideal sum of two pieces, a part corresponding to $L_1$ and a part corresponding to $L_2$.

When we want to consider the vanishing ideal of such a model we consider it in the ring

$$\mathbb{R}[\mu, \sigma] := \mathbb{R}[\mu_i, \sigma_{ij} : 1 \leq i \leq j \leq m].$$

We will restrict to the case that either $L_1 = 0$ or $L_1 = \mathbb{R}^m$. In the first case, the vanishing ideal of such a Gaussian exponential family has the form

$$\langle \mu_1, \ldots, \mu_m \rangle + I_2$$

where $I_2$ is an ideal in $\mathbb{R}[\sigma]$. In the second case, the vanishing ideal is generated by polynomials in $\mathbb{R}[\sigma]$. In either case, we can focus on the vanishing ideal of associated to the linear space $L_2$.

Note that the model of a regular exponential subfamily is a linear space in the space of concentration matrices. Since we are often interested in describing Gaussian models in the space of covariance matrices, this leads to the following definition.

**Definition 6.3.1.** Let $L \subseteq \mathbb{R}^{(m+1)m/2}$ be a linear space such that $L \cap \text{PD}_m$ is nonempty. The *inverse linear space* $L^{-1}$ is the set of positive definite matrices

$$L^{-1} = \{ K^{-1} : K \in L \cap \text{PD}_m \}.$$

Note that the map $L \rightarrow L^{-1}$ that takes a matrix and returns its inverse is a rational map (using the adjoint formula for the inverse of a matrix). This means that exponential families in the Gaussian case have interesting ideals in $\mathbb{R}[\sigma]$.

An important example, that we will see again later, concerns the case where $L$ is a coordinate subspace (with no zero entries on the diagonal to ensure the existence of positive definite matrices).

**Proposition 6.3.2.** If $K$ is a concentration matrix for a Gaussian random variable, a zero entry $k_{ij} = 0$ is equivalent to a conditional independence statement $i \indep j | [m] \setminus \{i, j\}$.

**Proof.** Let $C = [m] \setminus \{i, j\}$. From the adjoint formula for the inverse of a matrix, we see that $k_{ij} = (-1)^{i+j} \det \Sigma_{j \cup C, i \cup C} / \det \Sigma$. So $k_{ij} = 0$ if and only if $\det \Sigma_{j \cup C, i \cup C} = 0$. This is equivalent to the conditional independence statement $i \indep j | [m] \setminus \{i, j\}$ by Proposition 4.1.9.

$\square$
Hence, a linear space obtained by setting many of the concentration parameters equal to zero gives a conditional independence model where every conditional independence statement is a saturated conditional independence statement. (Recall that a conditional independence statement is saturated if it involves all the random variables under consideration.) The conditional independence ideals that arise from taking all the conditional independence statements implied by zeroes in the concentration matrix might not be prime in general. However, the linear space $L$ in the concentration coordinates is clearly irreducible, and this can allow us to parametrize the main component of interest.

**Example 6.3.3** (Gaussian Independence Model). Consider the Gaussian regular exponential family for $m = 3$ random variables described by the linear space of concentration matrices

$$L = \{ K \in PD_3 : k_{12} = 0, k_{13} = 0 \}.$$ 

This corresponds to the conditional independence statements $1 \perp 2|\{3\}$ and $1 \perp 3|2$. The intersection axiom implies that $1 \perp \{2, 3\}$ in this case, though this is not immediately clear from the Gaussian conditional independence ideal

$$J_C = (\sigma_{12} \sigma_{33} - \sigma_{13} \sigma_{23}, \sigma_{13} \sigma_{22} - \sigma_{12} \sigma_{23}).$$

which does not contain any linear polynomials. One could use a computer algebra system to compute the primary decomposition, and extract the component that corresponds to the exponential family. Another approach is to use the parametrization of the model to compute the vanishing ideal of the Gaussian exponential family. Here is an example of the parametrization computation in Macaulay2.

```plaintext
R = QQ[k11,k22,k23,k33, s11,s12,s13,s22,s23,s33]
K = matrix {{k11,0,0},{0,k22,k23},{0,k23,k33}}
S = matrix {{s11,s12,s13},{s12,s22,s23},{s13,s23,s33}}
I = ideal (K*S - identity(1)) -- ideal of entries
J = eliminate ({k11,k22,k23,k33}, I)
```

In this case, the resulting ideal $J$ is generated by the polynomials $\sigma_{12}$ and $\sigma_{13}$ realizing the conditional independence statement $1 \perp \{2, 3\}$.

We will return to examples such as this when we discuss graphical models in Chapter 13. More complex examples need not have the property that the vanishing ideal of the inverse linear space is generated by determinantal constraints associated with conditional independence statements.
6.4. Real Algebraic Geometry

Our eventual goal in this chapter is to introduce algebraic exponential families. To do this requires the language of real algebraic geometry. The difference between algebraic geometry over the reals versus the algebraic geometry over algebraically closed fields like the complex numbers (which has primarily been discussed thus far), is that the real numbers are ordered and this allows for inequalities. These inequalities are an important part of the study, even if one is only interested in zero sets of polynomials and maps between zero sets. Over the complex numbers, the topological closure of the image of a variety under a rational map is a variety. But over the reals, the topological closure of the image of a real algebraic variety under a rational map need not be a real algebraic variety. For instance the projection of the real variety \( V(p_1^2 + p_2^2 - 1) \) to the \( p_1 \) axis is the interval \([1, 1]\) which is not a real variety. Hence, the theory of real algebraic geometry requires more complicated objects than merely varieties. These more complicated objects are the semialgebraic sets. We provide a brief introduction here. More detailed background can be found in the texts \[BPR06, BCR98\].

**Definition 6.4.1.** Let \( \mathcal{F}, \mathcal{G} \subseteq \mathbb{R}[p_1, \ldots, p_r] \) be finite sets of polynomials. The basic semialgebraic set defined by \( \mathcal{F} \) and \( \mathcal{G} \) is the set

\[
\{ a \in \mathbb{R}^r : f(a) = 0 \text{ for all } f \in \mathcal{F} \text{ and } g(a) > 0 \text{ for all } g \in \mathcal{G} \}.
\]

A semialgebraic set is any set that can be written as a finite union of basic semialgebraic sets. In the case where \( \mathcal{G} = \emptyset \) the resulting basic semialgebraic set is called a real algebraic variety.

**Example 6.4.2 (Probability Simplex).** The interior of the probability simplex is a basic semialgebraic set with \( \mathcal{F} = \{ p_1 + \cdots + p_r - 1 \} \) and \( \mathcal{G} = \{ p_i : i \in [r] \} \). The probability simplex \( \Delta_{r-1} \) is a semialgebraic set, obtained as the union of \( 2^r - 1 \) basic semialgebraic sets which are the relative interiors of the faces of \( \Delta_{r-1} \).

In general, if \( \mathcal{F} \) and \( \mathcal{G} \) are finite sets of polynomials, a set of the form

\[
\{ a \in \mathbb{R}^r : f(a) = 0 \text{ for all } f \in \mathcal{F} \text{ and } g(a) \geq 0 \text{ for all } g \in \mathcal{G} \}
\]

usually called a basic closed semialgebraic set. An important example of this type of semialgebraic set is a polyhedron, where both sets \( \mathcal{F} \) and \( \mathcal{G} \) consist of linear polynomials. Polyhedra will be introduced and discussed in more detail in Chapter 8. Note that if we are interested only in closed semialgebraic sets, it is not sufficient to only consider semialgebraic sets of this type, as the following example illustrates.

**Example 6.4.3 (Non-basic Semialgebraic Set).** Let \( \Theta = D \cup R \) be the union of the disk \( D = \{ p \in \mathbb{R}^2 : p_1^2 + p_2^2 \leq 1 \} \) and the rectangle \( R = \{ p \in \mathbb{R}^2 : 0 \leq
6.4. Real Algebraic Geometry

This semialgebraic set cannot be written as a set of the form \( \{ p \in \mathbb{R}^2 : g(p) \geq 0 \text{ for all } g \in \mathcal{G} \} \) for a finite set of polynomials \( \mathcal{G} \). Indeed, since a half of the circle \( p_1^2 + p_2^2 = 1 \) forms part of the boundary of \( \Theta \), the polynomial \( g = 1 - p_1^2 - p_2^2 \) would need to appear in the set \( \mathcal{G} \) or as a factor of some polynomial in \( \mathcal{G} \). Not every point in \( \Theta \) satisfies \( g \geq 0 \), and similarly for any polynomial of the form \( fg \).

Most of the models in algebraic statistics arise from taking a parameter space \( \Theta \) which is a polyhedron and the model will be the image of \( \Theta \) under a rational map. As we will see, such sets are always semialgebraic sets.

**Theorem 6.4.4 (Tarski–Seidenberg).** Let \( \Theta \subseteq \mathbb{R}^{r+1} \) be a semialgebraic set, and let \( \pi : \mathbb{R}^{r+1} \to \mathbb{R}^r \) be a coordinate projection. Then \( \pi(\Theta) \subseteq \mathbb{R}^r \) is also a semialgebraic set.

The Tarski–Seidenberg theorem is the basis of useful results in real algebraic geometry. In particular, by constructing the graph of a rational map and repeatedly applying the Tarski-Seidenberg theorem, we deduce:

**Theorem 6.4.5.** Let \( \phi : \mathbb{R}^{r_1} \to \mathbb{R}^{r_2} \) be a rational map and \( \Theta \subseteq \mathbb{R}^{r_1} \) be a semialgebraic set. Then the image of \( \Theta \) under \( \phi \)

\[
\phi(\Theta) := \{ \phi(\theta) : \theta \in \Theta \text{ and } \phi(\theta) \text{ is defined} \}
\]

is a semialgebraic subset of \( \mathbb{R}^{r_2} \).

**Proof.** Let \( \Gamma_\phi := \{ (\theta, \phi(\theta)) : \theta \in \Theta \} \) be the graph of \( \phi \). Since \( \phi \) is given by rational functions \( \phi_i = f_i/g_i \), this is a semialgebraic set (which comes from taking the semialgebraic set representation of \( \Theta \) and adding the constraints that \( g_i(\phi(\theta)) - f_i(\theta) = 0 \), and then intersecting with the semialgebraic set \( \prod g_i(\theta) \neq 0 \)). The image \( \phi(\Theta) \) is the coordinate projection of the graph \( \Gamma_\phi \).

When we move on to likelihood inference for algebraic exponential families, we will see one nice property of semialgebraic sets is that they have well-defined tangent cones at points, which have the same dimension as the semialgebraic set at that point. This plays an important role in the geometric study of the likelihood ratio test in Section 7.4.

**Definition 6.4.6.** The tangent cone \( TC_{\theta_0}(\Theta) \) of a semialgebraic set \( \Theta \subseteq \mathbb{R}^k \) at the point \( \theta_0 \in \Theta \) is the set of limits of sequences \( \alpha_n(\theta_n - \theta_0) \) where \( \alpha_n \) are positive reals and \( \theta_n \in \Theta \) converge to \( \theta_0 \). The elements of the tangent cone are called tangent vectors. We say \( \theta_0 \) is a smooth point of \( \Theta \) if \( TC_{\theta_0}(\Theta) \) is a linear space.

The tangent cone is a closed set, and it is indeed a cone, that is multiplying a tangent vector by an arbitrary nonnegative real number yields another
tangent vector. Decomposing a semialgebraic set into simpler semialgebraic sets can be useful for determining the tangent cone at a point. For instance, if $\Theta = \Theta_1 \cup \Theta_2$ and $\theta_0 = \Theta_1 \cap \Theta_2$ then $\text{TC}_{\theta_0}(\Theta) = \text{TC}_{\theta_0}(\Theta_1) \cup \text{TC}_{\theta_0}(\Theta_2)$. If, on the other hand, $\theta \in \Theta_1 \setminus \Theta_2$ then $\text{TC}_{\theta_0}(\Theta) = \text{TC}_{\theta_0}(\Theta_1)$.

Over other fields and working solely with varieties, it is natural to give a definition of the tangent cone based purely on algebraic considerations. We give the standard definition over algebraically closed fields in Definition 6.4.7. To make this precise, let $f$ be a polynomial and expand this as a sum of homogeneous pieces $f = f_i + f_{i+1} + \cdots$ where $f_i$ is the smallest degree piece that appears in this polynomial. Let $\min(f) = f_i$, the homogeneous piece of lowest order.

**Definition 6.4.7.** Let $\mathbb{K}$ be an algebraically closed field, and $I \subseteq \mathbb{K}[p]$ an ideal. Suppose that the origin is in the affine variety $V(I)$. Then the algebraic tangent cone at the origin is defined to be $\text{TC}_0(V(I)) = V(\min(I))$ where $\min(I) := \langle \min(f) : f \in I \rangle$.

The algebraic tangent cone can be defined at any point of the variety, simply by taking the local Taylor series expansion of every polynomial in the ideal around that point (or, equivalently, applying an affine transformation to bring the point of interest to the origin). Over the complex numbers, the algebraic tangent cone at a point can be computed directly using Gröbner bases. The idea is a generalization of the following observation: let $f \in I(\Theta)$ and $\theta_0 \in \Theta$. Expand $f$ as a power series around the point $\theta_0$. Then the lowest order term of $f$ in this expansion vanishes on the tangent cone $\text{TC}_{\theta_0}(\Theta)$.

**Example 6.4.8 (Tangent Cone at a Simple Node).** Consider the polynomial $f = p_2^2 - p_1^2(p_1 + 1)$. The variety $V(f)$ is the nodal cubic. The polynomial $\min(f) = p_2^2 - p_1^2$, from which we see that the tangent cone at the origin of $V(f)$ is the union of two lines.

To compute the ideal $\min(I)$ and, hence, the algebraic tangent cone requires the use of Gröbner bases.

**Proposition 6.4.9.** Let $I \subseteq \mathbb{K}[p]$ an ideal. Let $I^h \subseteq \mathbb{K}[p_0, p]$ be the homogenization of $I$ with respect to the linear polynomial $p_0$. Let $\prec$ be an elimination order that eliminates $p_0$, and let $\mathcal{G}$ be a Gröbner basis for $I^h$ with respect to $\prec$. Then

$$\min(I) = \langle \min(g(1, p)) : g \in \mathcal{G} \rangle.$$

See [CLO07] for a proof of Proposition 6.4.9. To compute the algebraic tangent cone at a general point $\theta$, use Proposition 6.4.9 after first shifting the desired point of interest to the origin.
Example 6.4.10 (Singular Point on a Toric Variety). Consider the monomial parametrization $\phi^A$ given by the matrix
\[
A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 2 & 3 & 4 \end{pmatrix}.
\]
The point $(1, 0, 0, 0)$ is on the variety $V_A$ and is a singular point. The following Macaulay2 code computes the tangent cone at this point.

```plaintext
S = QQ[t1,t2];
R = QQ[p1,p2,p3,p4];
f = map(S,R,{t1, t1*t2^2, t1*t2^3,t1*t2^4});
g = map(R,R,{p1+1, p2,p3,p4});
h = map(R,R,{p1-1, p2,p3,p4});
I = kernel f
Ishift = g(I)
T = tangentCone Ishift
Itangentcone = h(T)
```

The maps $g$ and $h$ were involved in shifting $(1, 0, 0, 0)$ to the origin. The resulting ideal is $(p_3^2, p_1)$. So the tangent cone is a two-dimensional plane but with multiplicity 2.

Note that the algebraic tangent cone computation over the complex numbers only provides some information about the real tangent cone. Indeed, if $\Theta \subseteq \mathbb{R}^k$ is a semialgebraic set, $I(\Theta)$ its vanishing ideal, and $0 \in \Theta$, then the real algebraic variety $V(\min(I(\Theta)))$ contains the tangent cone $TC_0(\Theta)$ but need not be equal to it in general.

Example 6.4.11 (Tangent Cone at a Cusp). Let $f = p_2^2 - p_1^3$, and $V(f)$ be the cuspidal cubic in the plane. Then $\min(f) = p_2^2$ and hence the complex tangent cone is the $p_1$-axis. On the other hand, every point in $V_R(f)$ satisfies, $p_1 \geq 0$, and hence the real tangent cone satisfies this restriction as well. Thus, the tangent cone of the cuspidal cubic at the origin is the nonnegative part of the $p_1$-axis.

Generalizing Example 6.4.11, we can gain information about the tangent cone of semialgebraic sets at more general points, by looking at polynomials that are nonnegative at the given points.

Proposition 6.4.12. Let $\Theta$ be a semialgebraic set, suppose $0 \in \Theta$ and let $f \in \mathbb{R}[p]$ be a polynomial such that $f(\theta) \geq 0$ for all $\theta \in \Theta$. Then $\min(f)(\theta) \geq 0$ for all $\theta \in TC_0(\Theta)$.

Proof. If $f(\theta) \geq 0$ for all $\theta \in \Theta$ then as $\theta \to 0$, then $\min(f)(\theta) > 0$. This follows because $\min(f)(\theta)$ dominates the higher order terms of $f$ when $\theta$ is close to the origin. \qed
6.5. Algebraic Exponential Families

Now that we have the basic notions of real algebraic geometry in place, we can discuss algebraic exponential families.

Definition 6.5.1. Let \{P_\eta : \eta \in N\} be a regular exponential family of order \(k\). The subfamily induced by the set \(M \subseteq N\) is an algebraic exponential family if there exists an open set \(\tilde{N} \subseteq \mathbb{R}^k\), a diffeomorphism \(g : N \rightarrow \tilde{N}\), and a semi-algebraic set \(A \subseteq \mathbb{R}^k\) such that \(M = g^{-1}(A \cap \tilde{N})\).

We often refer to either the set \(M\), the set \(A \cap \tilde{N}\), or the resulting family of probability distributions \(\{P_\eta : \eta \in M\}\) as the algebraic exponential family. The notion of an algebraic exponential family is quite broad and covers nearly all of the statistical models we will consider in this book. The two most common examples that we will see in this book concern the discrete and multivariate Gaussian random variables.

Example 6.5.2 (Discrete Random Variables). Let \(\tilde{N} = \text{int}(\Delta_{k-1})\) and consider the diffeomorphism

\[
\eta \mapsto \left(\exp(\eta_1), \ldots, \exp(\eta_k)\right)/\sum_{i=1}^k \exp(\eta_i)
\]

where \(\eta_k = 1 - \eta_1 - \cdots - \eta_{k-1}\). This shows that every semialgebraic subset of the interior of the probability simplex is an algebraic exponential family.

The example of the discrete random variable shows that we cannot take only rational functions \(g\) in the definition, we might need a more complicated diffeomorphism when introducing an algebraic exponential family. Note that it can be useful and worthwhile to also take the closure of the set \(\tilde{N}\) and the model \(M \subseteq \tilde{N}\). The reason for considering open sets here is for applications to likelihood ratio tests in Chapter 7, but it is not needed in many other settings.

Example 6.5.3 (Gaussian Random Variables). Let \(N = \tilde{N} = \mathbb{R}^m \times PD_m\), the cone of positive definite matrices, and consider the diffeomorphism

\[
\mathbb{R}^m \times PD_m \rightarrow \mathbb{R}^m \times PD_m, \ (\eta, K) \mapsto (K^{-1}\eta, K^{-1}).
\]

This shows that any semialgebraic subset of \(\mathbb{R}^m \times PD_m\) of mean and covariance matrices yields an algebraic exponential family. Note that since the map \(K \mapsto K^{-1}\) is a rational map, we could have semialgebraic subsets of mean and concentration matrices: the Tarski–Seidenberg theorem guarantees that semialgebraic sets in one regime are also semialgebraic in the other.
The need for considering statistical models as semialgebraic sets is clear as soon as we begin studying hidden variable models, a theme that will begin to see in detail in Chapter 14. An important example is the case of mixture models, which we will address in more in Section 14.1. For example, the mixture of binomial random variables, Example 3.2.10, yields the semialgebraic set in $\mathbb{R}^2$ consisting of all distributions satisfying the polynomial constraint $4p_0p_2 - p_1^2 \geq 0$.

**Example 6.5.4 (Factor Analysis).** The factor analysis model $F_{m,s}$ is the family of multivariate normal distributions $N_m(\mu, \Sigma)$ on $\mathbb{R}^m$ whose mean vector $\mu$ is an arbitrary vector in $\mathbb{R}^m$ and whose covariance matrix lies in the set

$$F_{m,s} = \{ \Sigma = \Omega + \Lambda \Lambda^T : \Omega \succ 0 \text{ diagonal}, \Lambda \in \mathbb{R}^{m \times s} \}.$$ 

Here the notation $A \succ 0$ means that $A$ is a positive definite. Note that $F_{m,s}$ is a semialgebraic set since it is the image of the semialgebraic set $\mathbb{R}_{>0}^m \times \mathbb{R}^{m \times s}$ under a rational map. Hence, the factor analysis model is an algebraic statistical model. It is not, in general, of the form $PD_m \cap V(I)$ for any ideal $I$. In particular, the semialgebraic structure really matters for this model.

As an example, consider the case of $m = 3, s = 1$, a 1-factor model with three covariates. In this case, the model consists of all $3 \times 3$ covariance matrices of the form

$$\Sigma = \Omega + \Lambda \Lambda^T = \begin{pmatrix}
\omega_{11} + \lambda_1^2 & \lambda_1 \lambda_2 & \lambda_1 \lambda_3 \\
\lambda_1 \lambda_2 & \omega_{22} + \lambda_2^2 & \lambda_2 \lambda_3 \\
\lambda_1 \lambda_3 & \lambda_2 \lambda_3 & \omega_{33} + \lambda_3^2
\end{pmatrix}$$

where $\omega_{11}, \omega_{22}, \omega_{33} > 0$ and $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$.

Since this model has six parameters and is inside of six dimensional space, we suspect, and a computation verifies, that there are no polynomials that vanish on the model. That is, $I(F_{3,1}) = \{0\}$. On the other hand, not every $\Sigma \in PD_3$ belongs to $F_{3,1}$, since, for example, it is easy to see that $\Sigma \in F_{3,1}$ must satisfy $\sigma_{12}\sigma_{13}\sigma_{23} \geq 0$. In addition, it is not possible that exactly one of the entries $\sigma_{12}, \sigma_{13}, \sigma_{23}$ is zero, since such a zero would force one of the $\lambda_i$ to be equal to 0, which would introduce further zeros in $\Sigma$. Finally, if $\sigma_{jk} \neq 0$, then plugging in that $\sigma_{ii} = \omega_{ii} + \lambda_i^2$ and $\sigma_{ij} = \lambda_i \lambda_j$, we have the formula

$$\omega_{ii} = \sigma_{ii} + \frac{\sigma_{ij}\sigma_{ik}}{\sigma_{jk}}.$$ 

Note that $\omega_{ii}$ must be positive so we deduce that if $\sigma_{jk} \neq 0$ then

$$\sigma_{ii} - \frac{\sigma_{ij}\sigma_{ik}}{\sigma_{jk}} > 0.$$
Multiplying through by the $\sigma_{jk}^2$ we deduce the inequality

$$\sigma_{jk}(\sigma_{ii}\sigma_{jk} - \sigma_{ij}\sigma_{ik}) \geq 0$$

which now holds for all $\Sigma \in F_{3,1}$ (not only if $\sigma_{jk} \neq 0$). It can be shown that these conditions characterize the covariance matrices in $F_{3,1}$.

### 6.6. Exercises

**Exercise 6.1.** Recall that for $\lambda > 0$, a random variable $X$ has a Poisson distribution with parameter $\lambda$ if its state space is $\mathbb{N}$ and

$$P_{\lambda}(X = k) = \frac{\lambda^k}{k!} \exp(-\lambda), \quad \text{for } k \in \mathbb{N}.$$  

Show that the family of Poisson random variables is a regular exponential family. What is the natural parameter space?

**Exercise 6.2.** Consider the vector $h = (1, 1, 1, 2, 2, 2)$ and the matrix $A = \begin{pmatrix} 2 & 0 & 0 & 1 & 1 & 0 \\ 0 & 2 & 0 & 1 & 0 & 1 \\ 0 & 0 & 2 & 0 & 1 & 1 \end{pmatrix}$.

1. Compute generators for the toric ideals $I_A$ and $I_{A,h}$.
2. What familiar statistical model is the discrete exponential family $\mathcal{M}_{A,h}$?

**Exercise 6.3.** Consider the monomial parametrization

$$p_{ijk} = \alpha_{ij}\beta_{ik}\gamma_{jk}$$

for $i \in [r_1]$, $j \in [r_2]$, and $k \in [r_3]$. Describe the matrix $A$ associated to this monomial parametrization. How does $A$ act as a linear transformation on 3-way arrays? Compute the vanishing ideal $I_A$ for $r_1 = r_2 = r_3 = 3$.

**Exercise 6.4.** Consider the linear space $L$ of $4 \times 4$ symmetric matrices defined by

$$L = \{ K \in \mathbb{R}^{10} : k_{11} + k_{12} + k_{13} = k_{23} + k_{34} = 0 \}.$$  

Determine generators of the vanishing ideal $I(L^{-1}) \subseteq \mathbb{R}[\sigma]$.

**Exercise 6.5.** The convex hull of a set $S \subseteq \mathbb{R}^k$ is the smallest convex set that contains $S$. It denoted $\text{conv}(S)$ and is equal to

$$\text{conv}(S) = \{ \sum_{i=0}^{k} \lambda_i s_i : s_i \in S \text{ and } \lambda \in \Delta_k \}.$$  

1. Show that the convex hull of a semialgebraic set is a semialgebraic set.
(2) Give an example of a real algebraic curve in the plane (that is a set of the form $C = V(f) \subseteq \mathbb{R}^2$) such that $\text{conv}(S)$ is not a basic closed semialgebraic set.

**Exercise 6.6.** Show that the set $PSD_m$ of real positive semidefinite $m \times m$ symmetric matrices is a basic closed semialgebraic set.

**Exercise 6.7.** Characterize as completely as possible the semialgebraic set $F_{m,1} \subseteq PD_m$ of positive definite covariance matrices for the 1-factor models. In particular, give necessary and sufficient conditions on a positive definite matrix $\Sigma$ which imply that $\Sigma \in F_{m,1}$. 
Chapter 7

Likelihood Inference

This chapter is concerned with an in-depth treatment of maximum likelihood estimation from an algebraic perspective. In many situations, the maximum likelihood estimate is one of the critical points of the score equations. If those equations are rational, techniques from computational algebra can be used to find the maximum likelihood estimate, among the real and complex solutions of the score equations. An important observation is that the number of critical points of rational score equations is constant for generic data. This number is called the ML-degree of the model.

Section 7.1 concerns the maximum likelihood degree for discrete and Gaussian algebraic exponential families, which always have rational score equations. When parametrized statistical models are not identifiable, direct methods for solving the score equations in the parameters encounter difficulties, with many extraneous and repeated critical points.

One strategy to speed up and simplify computations is to work with statistical models in implicit form. In Section 7.2, we explain how to carry out the algebraic solution of the maximum likelihood equations in implicit form using Lagrange multipliers. This perspective leads to the study of the ML-degree as a geometric invariant of the underlying algebraic variety. It is natural to ask for a geometric explanation of this invariant. This connection is explained in Section 7.2. In particular, the beautiful classification of models with ML-degree one (that is, models that have rational formulas for their maximum likelihood estimates) is given. This is based on work of Huh [Huh13, Huh14].

The fact that the ML-degree is greater than one in many models means that there is usually no closed-form expression for the maximum likelihood
7. Likelihood Inference

estimate and that hill-climbing algorithms for finding the maximum likelihood estimate can potentially get stuck at a local but not global optimum. In Section 7.3, we illustrate algorithms for computing maximum likelihood estimates in exponential families and other models where the ML-degree can be bigger than 1, but there is still a unique local maximum of the likelihood function. Such models, which usually have concave likelihood functions, appear frequently in practice and specialized algorithms can be used to find their MLEs.

Section 7.4 concerns the asymptotic behavior of the likelihood ratio test statistic. If the data was generated from a true distribution that is a smooth point of a submodel, the likelihood ratio test statistic asymptotically approaches a chi-square distribution, which is the basis for p-value calculations in many hypothesis tests. However, many statistical models are not smooth, and the geometry of the singularities informs the way that the asymptotic p-value should be corrected at a singular point.

### 7.1. Algebraic Solution of the Score Equations

Let $\mathcal{M}_\Theta$ be a parametric statistical model with $\Theta \subset \mathbb{R}^d$ an open full dimensional parameter set. Let $D$ be data and let $\ell(\theta \mid D)$ be the log-likelihood function.

**Definition 7.1.1.** The *score equations* or *critical equations* of the model $\mathcal{M}_\Theta$ are the equations obtained by setting the gradient of the log-likelihood function to zero:

$$\frac{\partial}{\partial \theta_i} \ell(\theta \mid D) = 0, \quad i = 1, \ldots, d.$$ 

The point of this chapter is to understand the algebraic structure of this system of equations, one of whose solutions is the maximum likelihood estimate of the parameters given the data $D$. Note that because we take an open parameter space, the maximum likelihood estimate might not exist (e.g. if there is a sequence of model points approaching the boundary with increasing log-likelihood value). On the other hand, if we were to consider a closed parameter space, we can have situations where the maximum likelihood estimate is not a solution to the score equations.

First of all, we restrict our attention to algebraic exponential families, with i.i.d. data. In this section we assume that the statistical model is parametrized and we restrict to the discrete and Gaussian case, as these are situations where the score equations are always rational, giving a well-defined notion of maximum likelihood degree.

First we consider statistical models in the probability simplex $\Delta_{r-1}$, which are parametrized by a full dimensional subset $\Theta \subset \mathbb{R}^d$ and such that
the map \( p : \Theta \rightarrow \Delta_{r-1} \) is a rational map. Under the assumption of i.i.d. data, we collect a sequence of samples \( X^{(1)}, \ldots, X^{(n)} \) such that each \( X^{(i)} \sim p \) for some unknown distribution \( p \). This data is summarized by the vector of counts \( u \in \mathbb{N}^r \), given by \( u_j = \# \{ i : X^{(i)} = j \} \). The log-likelihood function of the model in this case becomes
\[
\ell(\theta \mid u) = \sum_{j=1}^r u_j \log p_j(\theta)
\]
in which case the score equations are the rational equations
\[
\sum_{j=1}^r \frac{u_j}{p_j} \frac{\partial p_j(\theta)}{\partial \theta_i} = 0, \quad i = 1, \ldots, d.
\]

**Theorem 7.1.2.** Let \( \mathcal{M}_\Theta \subseteq \Delta_{r-1} \) be a statistical model. For generic data, the number of solutions to the score equations (7.1.1) is independent of \( u \).

In algebraic geometry, the term generic means that the condition holds except possibly on a proper algebraic subvariety. This requires a word of comment since \( u \in \mathbb{N}^r \), and \( \mathbb{N}^r \) is not an algebraic variety. The score equations make sense for any \( u \in \mathbb{C}^r \). In the statement of Theorem 7.1.2 “generic” refers to a condition that holds off a proper subvariety of \( \mathbb{C}^r \). Since no proper subvariety of \( \mathbb{C}^r \) contains all of \( \mathbb{N}^r \), we have a condition that holds for generic data vectors \( u \).

To prove Theorem 7.1.2, requires us to use the notion of saturation and ideal quotient are used to remove extraneous components.

**Definition 7.1.3.** Let \( I, J \subseteq \mathbb{K}[p_1, \ldots, p_r] \) be ideals. The quotient ideal or colon ideal is the ideal
\[
I : J = \langle g \in \mathbb{K}[p_1, \ldots, p_r] : g \cdot J \subseteq I \rangle.
\]
The saturation of \( I \) with respect to \( J \) is the ideal,
\[
I : J^\infty = \bigcup_{k=1}^\infty I : J^k.
\]

Note that if \( J = \langle f \rangle \) is a principal ideal, then we use the shorthand \( I : f = I : \langle f \rangle \) and \( I : f^\infty = I : \langle f \rangle^\infty \). The geometric content of the saturation is that \( V(I : J^\infty) = V(I) \setminus V(J) \); that is, the saturation removes all the components of \( V(I) \) contained in \( V(J) \). When computing solutions to the score equations, we need to remove the extraneous components that arise from clearing denominators.

**Proof of Theorem 7.1.2.** To get a generic statement, it suffices to work in the coefficient field \( \mathbb{C}(u) := \mathbb{C}(u_1, \ldots, u_r) \). This allows us to multiply and
divide by arbitrary polynomials in \(u_1, \ldots, u_r\), and the nongeneric set will be the union of the zero sets of all of those functions.

Write \(p_j(\theta) = \frac{f_j(\theta)}{g_j(\theta)}\). To this end, consider the ideal

\[
J = \left( \prod_{j=1}^r f_j g_j \sum_{j=1}^r \frac{u_j}{p_j} \frac{\partial p_j}{\partial \theta_i}(\theta) \right) \cdot \left( \prod_{j=1}^r f_j g_j \right)^\infty \subseteq \mathbb{C}(u)[\theta].
\]

Multiplying the score equations by \(\prod_{j=1}^r f_j g_j\) yields a set of polynomial equations. This follows since,

\[
\frac{\partial}{\partial \theta_i} \log \frac{f_j}{g_j} = \frac{1}{f_j} \frac{\partial f_j}{\partial \theta_i} - \frac{1}{g_j} \frac{\partial g_j}{\partial \theta_i},
\]

so all the denominators are cleared. However, after clearing denominators, we might have introduced extraneous solutions, which lie where some of the \(f_j\) or \(g_j\) are zero. These are removed by computing the saturation by the product \(\prod_{j=1}^r f_j g_j\) (see Definition 7.1.3).

The resulting ideal \(J \in \mathbb{C}(u)[\theta]\) has zero set consisting of all solutions to the score equations, considered as elements of \(\overline{\mathbb{C}(u)}^d\), (where \(\overline{\mathbb{C}(u)}\) denotes the algebraic closure of \(\mathbb{C}(u)\)). This number is equal to the degree of the ideal \(J\), if it is a zero dimensional ideal, and infinity if \(J\) has dimension greater than zero.

**Definition 7.1.4.** The number of solutions to the score equations (7.1.1) for generic \(u\) is called the maximum likelihood degree (ML-degree) of the parametric discrete statistical model \(p : \Theta \to \Delta_{r-1}\).

Note that the maximum likelihood degree will be equal to the number of complex points for generic \(u\) in the variety \(V(J) \in \mathbb{C}^d\), where \(J\) is the ideal

\[
J = \left( \prod_{j=1}^r f_j g_j \sum_{j=1}^r \frac{u_j}{p_j} \frac{\partial p_j}{\partial \theta_i}(\theta) \right) \cdot \left( \prod_{j=1}^r f_j g_j \right)^\infty.
\]

that was computed in the proof of Theorem 7.1.2, but with \(u\) now treated as fixed numerical values, rather than as elements of the coefficient field.

**Example 7.1.5** (Random censoring). We consider families of discrete random variables that arise from randomly censoring exponential random variables. This example describes a special case of a censored continuous time conjunctive Bayesian network; see [BS09] for more details and derivations.

A random variable \(T\) is exponentially distributed with rate parameter \(\lambda > 0\) if it has the (Lebesgue) density function

\[
f(t) = \lambda \exp(-\lambda t) \cdot 1_{(t\geq 0)}, \quad t \in \mathbb{R}.
\]
Let $T_1, T_2, T_3$ be independent exponentially distributed random variables with rate parameters $\lambda_1, \lambda_2, \lambda_3$, respectively. Suppose that instead of observing the times $T_1, T_2, T_3$ directly, we can only observe for $i \in \{1, 2\}$, whether $T_i$ occurs before or after $T_3$. In other words, we observe a discrete random variable with the four states $\emptyset, \{1\}, \{2\}$ and $\{1, 2\}$, which are the elements $i \in \{1, 2\}$ such that $T_i \leq T_3$. This induces a rational map $g$ from the parameter space $(0, \infty)^3$ into the probability simplex $\Delta_3$. The coordinates of $g$ are the functions
\begin{align*}
g_0(\lambda_1, \lambda_2, \lambda_3) &= \frac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} \\
g_{(1)}(\lambda_1, \lambda_2, \lambda_3) &= \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} \cdot \frac{\lambda_3}{\lambda_2 + \lambda_3} \\
g_{(2)}(\lambda_1, \lambda_2, \lambda_3) &= \frac{\lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} \cdot \frac{\lambda_3}{\lambda_1 + \lambda_3} \\
g_{(1,2)}(\lambda_1, \lambda_2, \lambda_3) &= \frac{\lambda_1}{\lambda_1 + \lambda_3} \cdot \frac{\lambda_2}{\lambda_2 + \lambda_3} \cdot \frac{\lambda_1 + \lambda_2 + 2\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3},
\end{align*}
where $g_0(\lambda)$ is the probability that $T_1 > T_3$ and $T_2 > T_3$, $g_{(1)}(\lambda)$ $T_1 < T_3$ and $T_2 > T_3$, and so on. Given counts $u_0, u_1, u_2$, and $u_{12}$, the log-likelihood function is
\begin{align*}
\ell(\lambda) &= (u_1 + u_{12}) \log \lambda_1 + (u_2 + u_{12}) \log \lambda_2 + (u_0 + u_1 + u_2) \log \lambda_3 \\
&\quad + u_{12} \log(\lambda_1 + \lambda_2 + 2\lambda_3) \\
&\quad - (u_2 + u_{12}) \log(\lambda_1 + \lambda_3) - (u_1 + u_{12}) \log(\lambda_2 + \lambda_3) \\
&\quad - (u_0 + u_1 + u_2 + u_{12}) \log(\lambda_1 + \lambda_2 + \lambda_3).
\end{align*}
Since the parametrization involves rational functions of degree zero, we can set $\lambda_3 = 1$, and then solve the likelihood equations in $\lambda_1$ and $\lambda_2$. These likelihood equations are
\begin{align*}
\frac{u_1 + u_{12}}{\lambda_1} + \frac{u_{12}}{\lambda_1 + \lambda_2 + 2} - \frac{u_2 + u_{12}}{\lambda_1 + 1} - \frac{u_0 + u_1 + u_2 + u_{12}}{\lambda_1 + \lambda_2 + 1} &= 0 \\
\frac{u_2 + u_{12}}{\lambda_2} + \frac{u_{12}}{\lambda_1 + \lambda_2 + 2} - \frac{u_1 + u_{12}}{\lambda_2 + 1} - \frac{u_0 + u_1 + u_2 + u_{12}}{\lambda_1 + \lambda_2 + 1} &= 0.
\end{align*}
In this case, clearing denominators always introduces three extraneous solutions $(\lambda_1, \lambda_2) = (-1, -1), (0, -1), (-1, 0)$. The following code for the software Singular illustrates how to compute all solutions to the equations with the denominators cleared, and how to remove the extraneous solutions via the command sat. The particular counts used for illustration are specified in the third line.

\begin{verbatim}
LIB "solve.lib";
ring R = 0,(l1,l2),dp;
in u0 = 3; int u1 = 5; int u2 = 7; int u12 = 11;
\end{verbatim}
ideal I = (u1+u12)*(11+12+2)*(11+1)*(11+12+1) +
    (u12)*11*11*(11+1)*11+12+1) -
    (u2+u12)*11*(11+12+2)*(11+12+1) -
    (u0+u1+u2+u12)*11*(11+12+2)*(11+1),
    (u2+u12)*(11+12+2)*(12+1)*(11+12+1) +
    (u12)*12*(12+1)*(11+12+1) -
    (u1+u12)*12*(11+12+2)*(11+12+1) -
    (u0+u1+u2+u12)*12*(11+12+2)*(12+1);  
ideal K = 11*12*(11+1)*(12+1)*(11+12+1)*(11+12+2);  
ideal J = sat(I,K)[1];  
solve(J);  

In particular, there are three solutions to the likelihood equations, and the
ML-degree of this model is 3. In general, the maximum likelihood estimate
of \( \lambda_2 \) is a root of the cubic polynomial

\[
f(\lambda_2) = (u_0 + 2u_1)(u_1 - u_2)\lambda_2^3
+ (u_0u_1 + u_1^2 - 3u_0u_2 - 6u_1u_2 + u_2^2 - 2u_0u_{12} - 5u_1u_{12} + u_2u_{12})\lambda_2^2
+ (2u_0 + 3u_1 - 3u_2 - 3u_{12})(u_2 + u_{12})\lambda_2
+ 2(u_2 + u_{12})^2.
\]

This polynomial was computed using a variation of the above code, working
over the field \( \mathbb{Q}(u_0, u_1, u_2, u_{12}) \) that can be defined in *Singular* by

\[
\text{ring } R = (0,u_0,u_1,u_2,u_{12}),(11,12),dp;
\]

The leading coefficient reveals that the degree of \( f(\lambda_2) \) is three for all non-zero vectors of counts \( u \) away from the hyperplane defined by \( u_1 = u_2 \). \( \square \)

In addition to the algebraic models for discrete random variables, the
score equations are also rational for algebraic exponential families arising as
submodels of the parameter space of the general multivariate normal model.
Here we have parameter space \( \Theta \subseteq \mathbb{R}^m \times PD_m \) which gives a family of
probability densities

\[
P_\Theta = \{ N(\mu, \Sigma) : (\mu, \Sigma) \in \Theta \}.
\]

Given data, \( X^{(1)}, \ldots, X^{(n)} \), the minimal sufficient statistics for the saturated
Gaussian model are

\[
\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X^{(i)} \quad \text{and} \quad \hat{\Sigma} = S = \frac{1}{n} \sum_{i=1}^{n} (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^T.
\]

Since \( \bar{X} \) and \( S \) are minimal sufficient statistics for the saturated model,
they are sufficient statistics for any submodel, and we can rewrite the log-
likelihood for any Gaussian model in terms of these quantities. Indeed, the
log-likelihood function becomes:
7.1. Algebraic Solution of the Score Equations

(7.1.2) \[ \ell(\mu, \Sigma \mid \bar{X}, S) = \]
\[ -\frac{n}{2} (\log \det \Sigma + m \log 2\pi) - \frac{n}{2} \text{tr}(S\Sigma^{-1}) - \frac{n}{2} (\bar{X} - \mu)^T \Sigma^{-1} (\bar{X} - \mu). \]

See Proposition 5.3.7 for derivation.

Often we can ignore the term \(-\frac{nm}{2} \log(2\pi)\) since this does not affect the location of the maximum likelihood estimate, and is independent of restrictions we might place on \(\mu\) and \(\Sigma\). Two special cases are especially worthy of attention, and where the maximum-likelihood estimation problem reduces to a simpler problem.

**Proposition 7.1.6.** Let \(\Theta = \Theta_1 \times Id_m \subseteq \mathbb{R}^m \times PD_m\) be the parameter space for a Gaussian statistical model. Then the maximum likelihood estimation for \(\Theta\) is equivalent to the least-squares point on \(\Theta_1\).

**Proof.** Since \(\Sigma = Id_m\), the first two terms in the likelihood function are fixed, and we must maximize
\[ -\frac{n}{2} (\bar{X} - \mu)^T Id_m^{-1} (\bar{X} - \mu) \]
which is equivalent to minimizing \(|\bar{X} - \mu|^2\) over \(\Theta_1\). \(\square\)

In the mathematical optimization literature, the number of critical points of the critical equations to the least squares closest point on a variety is called the Euclidean distance degree \([DHO+16]\). Hence the ML-degree of a Gaussian model with identity covariance matrix, and where the means lie on a prescribed variety is the same as the ED-degree of that variety.

**Example 7.1.7** (Euclidean distance degree of cuspidal cubic). Continuing the setup of Proposition 7.1.6, let \(\Theta_1 = \{(t^3, t^4) : t \in \mathbb{R}\}\), consisting of the points on a cuspidal cubic curve. Given the observed mean vector \(\bar{X}\) we wish to compute the maximum likelihood estimate for \(\mu \in \Theta_1\). The relevant part of the likelihood function has the form
\[ (\bar{X}_1 - t^2)^2 + (\bar{X}_2 - t^3)^2. \]
Differentiating with respect to \(t\) yields the quintic equation
\[ t(2\bar{X}_1 + 3\bar{X}_2 t - 2t^2 - 3t^4) = 0. \]
Note that this degree 5 equation always has the origin \(t = 0\) as a solution, and the degree 4 factor is irreducible. In this case, although this equation has degree 5, we still say that the maximum likelihood degree is 4, since finding the critical point only involves an algebraic function of degree 4 in general. On the other hand, note that if \(\bar{X}\) lies to the left of the curve:
\[ W = V(2x + 32x^2 + 128x^3 + y^2 + 72xy^2 + 27y^4), \]
the MLE will be the origin. The curve $V(y^2 - x^3)$ and the discriminantal curve $W$ (that separates the regions where the optimum is on the smooth locus and the singular point at the origin) are illustrated in Figure 7.1.1.

We see in the previous example an important issue regarding maximum likelihood estimation: singular points matter, and for a large region of data space the MLE might be forced to lie on a lower dimensional stratum of the model. These sorts of issues can also arise when considering closed parameter spaces with boundary. The maximum likelihood estimate might lie on the boundary for a large region of the data space.

Example 7.1.8 (Euclidean distance degree of nodal cubic). Let $\Theta_1 = \{(t^2 - 1, t(t^2 - 1)) : t \in \mathbb{R}\}$, consisting of points on a nodal cubic. Note that the node arises when $t = \pm 1$. Given the observed mean vector $\bar{X}$ we wish to compute the maximum likelihood estimate for $\mu_1 \in \Theta_1$. The relevant part of the likelihood function has the form

$$\bar{X}_1^2 (t - 1)^2 + \bar{X}_2 (t^3 - t)).$$

Differentiating with respect to $t$ yields the irreducible quintic equation

$$-\bar{X}_2 + (2\bar{X}_1 + 1)t + 2\bar{X}_2 t^2 + 2t^3 - 3t^5 = 0.$$

In this case, the ML-degree is five. While both this and the previous are parametrized by cubics, the ML-degree drops in the cuspidal cubic because of the nature of the singularity there.

Proposition 7.1.9. Let $\Theta = \mathbb{R}^m \times \Theta_1 \subseteq \mathbb{R}^m \times PD_m$ be the parameter space for a Gaussian statistical model. Then the maximum likelihood estimation
7.1. Algebraic Solution of the Score Equations

problem gives $\mu = \bar{X}$, and reduces to maximizing

$$\ell(\Sigma \mid S) = -\frac{n}{2} \log \det \Sigma - \frac{n}{2} \text{tr}(S\Sigma^{-1}).$$

**Proof.** Since $\Sigma$ is positive definite, the expression

$$-\frac{n}{2}(\bar{X} - \mu)^T \Sigma^{-1}(\bar{X} - \mu)$$

in the likelihood function is always $\leq 0$, and equal to zero if and only if $(\bar{X} - \mu) = 0$. Since we are trying to maximize the likelihood function, choosing $\mu = \bar{X}$ makes this term as large as possible and does not affect the other terms of the likelihood function.

A similar result holds in the case of centered Gaussian models, that is, models where the mean is constrained to be zero.

**Proposition 7.1.10.** Let $\Theta = \{0\} \times \Theta_1 \subseteq \mathbb{R}^m \times PD_m$ be the parameter space for a centered Gaussian statistical model. Then the maximum likelihood estimation problem gives $\mu = \bar{X}$, and reduces to maximizing

$$\ell(\Sigma \mid S) = -\frac{n}{2} \log \det \Sigma - \frac{n}{2} \text{tr}(S\Sigma^{-1})$$

where $S = \sum_{i=1}^n X^{(i)}(X^{(i)})^T$ is the centered sample covariance matrix.

**Example 7.1.11 (Gaussian Marginal Independence).** Consider the Gaussian statistical model $\Theta = \mathbb{R}^m \times \Theta_1$ where

$$\Theta_1 = \{\Sigma \in PD_4 : \sigma_{12} = \sigma_{21} = 0, \sigma_{34} = \sigma_{43} = 0\}.$$  

This model is a Gaussian independence model with the two marginal independence constraints $X_1 \perp X_2$ and $X_3 \perp X_4$. Since the likelihood function involves both $\log \det \Sigma$ and $\Sigma^{-1}$, we will need to multiply all equations by $(\det \Sigma)^2$ to clear denominators, which introduces extraneous solutions. In this example, there is a six dimensional variety of extraneous solutions. Saturating with respect to this determinant removes that space of extraneous solutions to reveal 17 solutions to the score equations. The following code in Macaulay2 shows that there are 17 solutions for a particular realization of the sample covariance matrix $S$.

```plaintext
R = QQ[s11,s13,s14,s22,s23,s24,s33,s44];  
D = matrix{{19, 3, 5, 7}, {3,23,11,13},  
{5,11,31,-1}, {7,13,-1,37}};  
S = matrix{{s11,0 ,s13,s14}, {0 ,s22,s23,s24},  
{s13,s23,s33,0 },{s14,s24,0 ,s44}};  
W = matrix{{0,0,0,1},{0,0,-1,0},{0,1,0,0},{-1,0,0,0}};  
adj = W*exteriorPower(3,S)*W;  
L = trace(D*adj);  
f = det S;
```
J = ideal apply(flatten entries vars R,
      i -> f*diff(i,f) - f*diff(i,L) + diff(i,f)*L);
K = saturate(J,ideal(f));
dim J, dim K, degree K

Further study of maximum likelihood estimation in Gaussian marginal
independence models that generalize this example appear in [DR08]. In
particular, a characterization of when these models have ML-degree 1 can
be given.

7.2. Likelihood Geometry

The goal of this section is to explain some of the geometric underpinnings
of the maximum likelihood estimate and the solutions to the score equa-
tions. The maximum likelihood degree of a variety is intimately connected
to geometric features of the underlying variety. The title of this section
comes from the paper [HS14], which provides an in-depth introduction to
this topic with numerous examples.

To explain the geometric connection, we need to phrase the maximum
likelihood estimation problem in the language of projective geometry. Also
the most geometrically significant content arises when considering models
implicitly, and so we restrict attention to this case. Let $V \subseteq \mathbb{P}^{r-1}$ be an irre-
ducible projective variety. Let $u \in \mathbb{N}^r$. The maximum likelihood estimation
problem on the variety $V$ is to maximize the function

$$L_u(p) = \frac{p_1^{u_1} p_2^{u_2} \cdots p_r^{u_r}}{(p_1 + \cdots + p_r)^{u_1 + u_2 + \cdots + u_r}}$$

over the variety $V$. The function $L_u(p)$ is just the usual likelihood function if
we were to strict to the intersection of $V$ with the hyperplane where the sum
of coordinates is 1. Since we work with projective varieties, the denominator
is needed to make a well-defined degree zero rational function on projective
space.

Let $u_+ = u_1 + \cdots + u_r$ and let $p_+ = p_1 + \cdots + p_r$. Suppose that $I(V) =
\langle f_1, \ldots, f_k \rangle$. The maximizer of the likelihood over $V$ can be determined by
using Lagrange multipliers. In particular, consider the Lagrangian of the
log-likelihood function

$$\ell_u(p, \lambda) = \sum_{i=1}^r u_i \log p_i - u_+ \log p_+ + \sum_{j=1}^k \lambda_j f_j.$$  

The resulting critical equations are

$$0 = \frac{\partial \ell}{\partial p_i} = \frac{u_i}{p_i} - \frac{u_+}{p_+} + \sum_{j=1}^k \lambda_j \frac{\partial f_j}{\partial p_i} \quad \text{for } i \in [r]$$
7.2. Likelihood Geometry

Let
\[ \operatorname{Jac}(I(V)) = \begin{pmatrix} \frac{\partial f_1}{\partial p_1} & \cdots & \frac{\partial f_1}{\partial p_r} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_k}{\partial p_1} & \cdots & \frac{\partial f_k}{\partial p_r} \end{pmatrix} \]

be the Jacobian matrix of the vanishing ideal of \( V \). Then the resulting critical points of the log-likelihood function restricted to \( V \) are precisely the points \( p \in V \) such that
\[ \left( \frac{u_1}{p_1}, \ldots, \frac{u_r}{p_r} \right) = \frac{u_p}{p_r} (1, \ldots, 1) \in \text{rowspan} \left( \operatorname{Jac}(I(V)) \right). \]

Since the log-likelihood function becomes undefined when \( p \) belongs to the hyperplane arrangement \( \mathcal{H} = \{ p \in \mathbb{P}^{r-1} : p_1 \cdots p_r p_+ = 0 \} \), we remove such points from consideration in algebraic maximum likelihood estimation. Another issue that we need to deal with when considering maximum likelihood estimates with Lagrange multipliers is singular points. The definition begins with the following basic fact relating the rank of the Jacobian matrix to the dimension of the variety.

**Proposition 7.2.1.** Let \( V \) be an irreducible projective variety. Then at a generic point of \( V \), the rank of the Jacobian matrix \( \operatorname{Jac}(I(V)) \) is equal to the codimension of \( V \).

See [Sha13, Page 93] for a proof.

**Definition 7.2.2.** Let \( V \) be an irreducible variety. A point \( p \in V \) is singular if the rank of the Jacobian matrix \( \operatorname{Jac}(I(V)) \) is less than the codimension of \( V \). The set \( V_{\text{sing}} \subseteq V \) is the set of all singular points of \( V \) and is called the singular locus of \( V \). A point \( p \in V \setminus V_{\text{sing}} \) is called a regular point or a nonsingular point. The set of all regular points is denoted \( V_{\text{reg}} \). If \( V_{\text{reg}} = V \) then \( V \) is smooth.

For a variety that is reducible, the singular locus is the union of the singular loci of the component varieties, and the points on any intersections between components.

Since the rank of the Jacobian changes on the singular locus, it is typical to avoid these points when computing maximum likelihood estimates using algebraic techniques. These conditions can be restated as follows:

**Proposition 7.2.3.** A vector \( p \in V(I) \setminus V(p_1 \cdots p_r (p_1 + \cdots + p_r)) \) is a critical point of the Lagrangian if and only if the data vector \( u \) lies in the row span of
the augmented Jacobian matrix of \( f \), where the augmented Jacobian matrix of \( f \) is the matrix

\[
J(p) = \begin{pmatrix}
p_1 f_1 & p_2 f_1 & \cdots & p_r f_1 \\
p_1 f_2 & p_2 f_2 & \cdots & p_r f_2 \\
\vdots & \vdots & \ddots & \vdots \\
p_1 f_r & p_2 f_r & \cdots & p_r f_r 
\end{pmatrix}
\]

Proposition 7.2.3 can be used to develop algorithms for computing the maximum likelihood estimates in implicit models. Here is an algorithm taken from [HKS05] which can be used to compute solutions to the score equations for implicit models.

**Algorithm 7.2.4** (Solution to Score Equations of Implicit Model).

Input: A homogeneous prime ideal \( P = \langle f_1, \ldots, f_k \rangle \subseteq \mathbb{R}[p_1, \ldots, p_r] \) and a data vector \( u \in \mathbb{N}^r \).

Output: The likelihood ideal \( I_u \) for the model \( V(P) \cap \Delta_{r-1} \).

Step 1: Compute \( c = \text{codim}(P) \) and let \( Q \) be the ideal generated by the \( c + 1 \) minors of the Jacobian matrix \( \left( \frac{\partial f_i}{\partial p_j} \right) \).

Step 2: Compute the kernel of \( J(p) \) over the coordinate ring \( \mathbb{R}[V] := \mathbb{R}[p_1, \ldots, p_r]/P \).

This kernel is a submodule, \( M \), of \( \mathbb{R}[V]^r \).

Step 3: Let \( I'_u \) be the ideal in \( \mathbb{R}[V] \) generated by the polynomials \( \sum_{i=1}^r u_i \phi_i \) running over all the generators \( (\phi_1, \ldots, \phi_r) \) for the module \( M \).

Step 4: The likelihood ideal \( I_u \) equals

\[
I_u = I'_u : (p_1 \cdots p_r (p_1 + \cdots + p_r) Q)^\infty.
\]

**Example 7.2.5** (Random Censoring). Here we illustrate Algorithm 7.2.4 on the random censoring model from Example 7.1.5. Algorithms for implicitization from Chapter 3 can be used to show that the homogeneous vanishing ideal is

\[
P = \langle 2p_0 p_1 p_2 + p_1^2 p_2 + p_1 p_2^2 - p_0^2 p_1 p_2 + p_1 p_2 p_1^2 \rangle \subseteq \mathbb{R}[p_0, p_1, p_2, p_1^2].
\]

We solve the score equations in the implicit case using the following code in Singular.

```plaintext
LIB "solve.lib";
ring R = 0, (p0, p1, p2, p12), dp;
ideal P = 2*p0*p1*p2 + p1^2*p2 + p1*p2^2 - p0^2*p1*p2 + p1*p2*p1^2;
matrix u[1][4] = 4,2, 11, 15;
```
matrix J = jacob(sum(maxideal(1)) + P) * diag(maxideal(1));
matrix I = diag(P, nrows(J));
module M = modulo(J, I);
ideal Iprime = u * M;
int c = nvars(R) - dim(groebner(P));
ideal Isat = sat(Iprime, minor(J, c+1))[1];
ideal IP = Isat, sum(maxideal(1))-1;
solve(IP);

The computation shows that there are three solutions to the score equations
so this model has maximum likelihood degree 3, which agrees with our earlier
computations in the parametric setting.

The fact that we can calculate maximum likelihood estimates in the im-

ciplicit setting is useful not only as an alternate computation strategy, but
because it provides geometric insight into the maximum likelihood degree
and to likelihood geometry. In this area, a key role is played by the likelihood
correspondence. This is a special case of an incidence correspondence that
is important in algebraic geometry for studying how solutions to systems of
equations vary as parameters in the problem variety. In likelihood geome-
try, the likelihood correspondence ties together the data vectors and their
maximum likelihood estimates into one geometric object.

To explain the likelihood correspondence, we now think of the data vec-
tor $u$ as a coordinate on a different projective space $\mathbb{P}^{r-1}$. Note that the
critical equations do not depend on which representative for an element $u$
we choose that gives the same point in $\mathbb{P}^{r-1}$. This likelihood correspondence
is a variety in $\mathbb{P}^{r-1} \times \mathbb{P}^{r-1}$, whose coordinates correspond to pairs $(p, u)$ of
data points $u$ and corresponding critical points of the score equations.

**Definition 7.2.6.** Let $V \subseteq \mathbb{P}^{r-1}$ be an irreducible algebraic variety. Let
$\mathcal{H} = \{ p \in \mathbb{P}^{r-1} : p_1 \cdots p_r p_{r+1} = 0 \}$. The likelihood correspondence $\mathcal{L}_V$ is the

\[
\{ (p, u) : p \in V_{reg} \setminus \mathcal{H} \text{ and } \text{rank } J(P) \leq \text{codim } V \}.
\]

One key theorem in this area is the following.

**Theorem 7.2.7.** The likelihood correspondence $\mathcal{L}_V$ of any irreducible vari-

ety $V$ is an irreducible variety of dimension $r - 1$ in $\mathbb{P}^{r-1} \times \mathbb{P}^{r-1}$. The map
$\text{pr}_1 : \mathcal{L}_V \to \mathbb{P}^{r-1}$ that projects onto the first factor is a projective bundle
over $V_{reg} \setminus \mathcal{H}$. The map $\text{pr}_2 : \mathcal{L}_V \to \mathbb{P}^{r-1}$ that projects onto the second factor
is generically finite-to-one.

Given a point $u \in \mathbb{P}^{r-1}$, the preimage $\text{pr}_2^{-1}(u)$ is exactly the set of pairs
of points $(p, u)$ such that $p$ is a critical point for the score equations with
data vector \( u \). Hence, from the finite-to-one results, we see that there is a well-defined notion of maximum likelihood degree for statistical models that are defined implicitly. In particular, the degree of the map \( \text{pr}_2 \) is exactly the maximum likelihood degree in the implicit representation of the model. See [HS14] for a proof of Theorem 7.2.7. The mathematical tools for a deeper study of the ML-degree involve characteristic classes (Chern-Schwartz-Macpherson classes) that are beyond the scope of this book.

One consequence of the likelihood correspondence is that it allows a direct geometric interpretation of the ML-degree in the case of smooth very affine varieties.

**Definition 7.2.8.** The algebraic torus is the set \( (\mathbb{A}^r)^r \). A very affine variety is a set of the form \( V \cap (\mathbb{A}^r)^r \) where \( V \subseteq \mathbb{A}^r \) is a variety.

Note that if we can consider any projective varieties \( V \subseteq \mathbb{P}^{r-1} \), we obtain a very affine variety by taking the image of the rational map

\[
\phi : V \setminus \mathcal{H} \to (\mathbb{C}^r)^r, (p_1 : \ldots : p_r) \mapsto (p_1, \ldots, p_r)/p_+.
\]

The maximum likelihood estimation problem on the projective variety \( V \) is equivalent to the maximum likelihood estimation problem on \( \phi(V) \).

**Theorem 7.2.9 ([Huh13]).** Let \( V \subseteq (\mathbb{C}^r)^r \) be a very affine variety of dimension \( d \). If \( V \) is smooth then the ML-degree of \( V \) is \( (-1)^d \chi(V) \), where \( \chi(V) \) denotes the topological Euler characteristic.

As we have seen in our discussion of algebraic exponential families many statistical models are not smooth. Note, however, that Theorem 7.2.9 only requires smoothness away from the coordinate hyperplanes, and so this theorem applies to any discrete exponential family.

**Example 7.2.10 (Generic Curve in the Plane).** Let \( V \) be a generic curve of degree \( d \) in \( \mathbb{P}^2 \). Such a curve has Euler characteristic \( -d^2 + 3d \) and is smooth. Note that the hyperplane arrangement \( \mathcal{H} = \{ p : p_1p_2p_3(p_1 + p_2 + p_3) = 0 \} \) is a curve of degree 4. The intersection of a generic curve of degree \( d \), with the hyperplane curve of degree \( d \), consist of \( 4d \) reduced points, by Bezout’s theorem. Thus, the Euler characteristic of \( V \setminus \mathcal{H} \) is \( -d^2 + 3d - 4d = -d^2 - d \), so the ML-degree of \( V \) is \( d^2 + d \).

Example 7.2.10 can be substantially generalized to cover the case of generic complete intersections. Recall that a variety \( V \subseteq \mathbb{P}^{r-1} \) is a complete intersection if there exist polynomials \( f_1, \ldots, f_k \) such that \( I(V) = \langle f_1, \ldots, f_k \rangle \) and \( k = \text{codim} V \). A version of this result first appeared in [HKS05].
Theorem 7.2.11. Suppose that $V \subseteq \mathbb{P}^{r-1}$ is a generic complete intersection with $I(V) = \langle f_1, \ldots, f_k \rangle$, $k = \text{codim } V$, and $d_i = \deg f_i$, $d_k = \deg f_k$. Then the ML-degree of $V$ is $D d_1 \cdots d_k$ where

$$D = \sum_{i_1 + \cdots + i_r \leq r-1-k} d_1^{i_1} \cdots d_k^{i_r}.$$ 

So in the case of Example 7.2.10, we have $r = 3$ and $k = 1$, in which case $D = 1 + d$ and the ML-degree is $d^2 + d$, as calculated in Example 7.2.10.

Example 7.2.12 (Euler Characteristic of Independence Model). Consider the statistical model of independence for two binary random variables. As a projective variety, this is given by the equation $V(p_{11} p_{22} - p_{12} p_{21})$. As a very affine variety this is isomorphic to $(\mathbb{C}^*)^2$. Furthermore, this is a discrete exponential family, so is nonsingular as a very affine variety. To see this note that this very affine variety can be parametrized via the map $\phi : (\mathbb{C}^*)^2 \to (\mathbb{C}^*)^4$

$$(\alpha, \beta) \mapsto \left( \frac{\alpha}{1 + \alpha + \beta}, \frac{\beta}{1 + \alpha + \beta}, \frac{1}{1 + \alpha + \beta}, \frac{1}{1 + \alpha + \beta} \right)$$

which is undefined when $\alpha$ or $\beta$ are $-1$. Note that $\mathbb{C}^* - \{ -1 \}$ is a sphere minus three points, and has Euler characteristic $-1$. Thus, the ML-degree is $(-1) \times (-1) = 1$. On the other hand, Theorem 7.2.11 predicts that the ML-degree of a generic surface in $\mathbb{P}^3$ of degree 2 is $(1 + 2 + 2^2)/2 = 14$. So although the independence model is a complete intersection in this case, it is not generic, and the ML-degree is lower.

The independence model is a rather special statistical model, among other exceptional properties, it has maximum likelihood degree one. Most statistical models do not have ML-degree one. When a model does have ML-degree one, we immediately see that that model must be (part of) a unirational variety, since the map from data to maximum likelihood estimate will be a rational function that parametrizes the model. Huh [Huh14] has shown that the form of the resulting parametrization is extremely restricted.

Theorem 7.2.13. The following are equivalent conditions on a very affine variety.

(1) $V \subseteq (\mathbb{C}^*)^r$ has ML-degree 1.
(2) There is a vector \( h = (h_1, \ldots, h_r) \in (\mathbb{C}^*)^r \) an integral matrix \( B \in \mathbb{Z}^{n \times r} \) with all columns sums equal to zero, and a map \( \psi : (\mathbb{C})^r \rightarrow (\mathbb{C}^*)^r \)

\[
\Psi_k(u_1, \ldots, u_r) = h_k \cdot \prod_{i=1}^{n} \left( \sum_{j=1}^{r} b_{ij} u_j \right)^{b_{ik}}
\]

such that \( \Psi \) maps dominantly onto \( V \).

The map \( \Psi \) above which parametrizes any variety of ML-degree one is called the Horn uniformization. This shows that the varieties with ML-degree one are precisely the \( A \)-discriminants (see \cite{gkz94}).

Example 7.2.14 (Horn Uniformization of Independence Model). Let

\[
B = \begin{pmatrix}
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 \\
-2 & -2 & -2 & -2
\end{pmatrix}
\]

and \( h = (4, 4, 4, 4) \).

Then

\[
\Psi_{ij}(u_{11}, u_{12}, u_{21}, u_{22}) = \frac{(u_{11} + u_{22})(u_{1j} + u_{2j})}{(u_{11} + u_{12} + u_{21} + u_{22})^2}.
\]

In particular, the Horn uniformization parametrizes the model of independence of two discrete random variables.

7.3. Concave Likelihood Functions

Solving the score equations is a fundamental problem of computational statistics. Numerical strategies have been developed for finding solutions to the score equations, which are generally heuristic in nature. That is, algorithms for finding the maximum likelihood estimate might not be guaranteed to return the maximum likelihood estimate, but at least converge to a local maximum, or critical point, of the likelihood function. In special cases, in particular for exponential families and for discrete linear models, the likelihood function is concave and such hill-climbing algorithms converge to the maximum-likelihood estimate.

Definition 7.3.1. A set \( S \subseteq \mathbb{R}^d \) is convex if it is closed and for all \( x, y \in S \), \( (x + y)/2 \in S \) as well. If \( S \) is a convex set, a function \( f : S \rightarrow \mathbb{R} \) is convex if \( f((x + y)/2) \leq (f(x) + f(y))/2 \) for all \( x, y \in S \). A function \( f : S \rightarrow \mathbb{R} \) is concave if \( f((x + y)/2) \geq (f(x) + f(y))/2 \) for all \( x, y \in S \).

Proposition 7.3.2. Suppose that \( S \) is convex, and \( f : S \rightarrow \mathbb{R} \) is a concave function. Then the set \( U \subseteq S \) where \( f \) attains its maximum value is a
7.3. Concave Likelihood Functions

convex set. If \( f \) is strictly concave \((f((x + y)/2) > (f(x) + f(y))/2 \text{ for all } x \neq y \in S)\) then \( f \) has a unique global maximum, if a maximum exists.

**Proof.** Suppose that \( x, y \in U \) and let \( \alpha \) be the maximum value of \( f \) on \( S \). Then

\[
f((x + y)/2) \geq (f(x) + f(y))/2 = (\alpha + \alpha)/2 = \alpha \geq f((x + y)/2).
\]

This shows that the set of maximizers is convex. A strict inequality leads to a contradiction if there are two maximizers. \( \Box \)

**Definition 7.3.3.** Let \( f_1(\theta), \ldots, f_r(\theta) \) be nonzero linear polynomials in \( \theta \) such that \( \sum_{i=1}^r f_i(\theta) = 1 \). Let \( \Theta \subseteq \mathbb{R}^d \) be the set such that \( f_i(\theta) > 0 \) for all \( \theta \in \Theta \), and suppose that \( \dim \Theta = d \). The model \( M = f(\Theta) \subseteq \Delta_{r-1} \) is a called a discrete linear model.

In [PS05], the discrete linear models were called linear models. We choose not to use that name because “linear model” in statistics usually refers to models for continuous random variables where the random variables are related to each other by linear relations.

Note that our assumption that \( \sum_{i=1}^r f_i(\theta) = 1 \) guarantees that \( \Theta \subseteq \mathbb{R}^d \) is bounded. This follows since if \( \Theta \) were unbounded, there would be a point \( \theta_0 \in \Theta \) and a direction \( \nu \) such that the ray \( \{\theta_0 + t\nu : t \geq 0\} \) is contained in \( \Theta \). If this were true, then for at least one of the functions \( f_i \), \( f_i(\theta_0 + t\nu) \) increases without bound as \( t \to \infty \). Since \( \sum_{i=1}^r f_i(\theta) = 1 \), some other \( f_j(\theta_0 + t\nu) \) would go to negative infinity as \( t \to \infty \). This contradicts the fact that \( \Theta = \{\theta \in \mathbb{R}^d : f_i(\theta) > 0 \text{ for all } i\} \).

**Proposition 7.3.4.** Let \( M \) be a discrete linear model. Then the likelihood function \( \ell(\theta \mid u) \) is concave. The set

\[
\Theta' = \{\theta \in \mathbb{R}^d : f_i(\theta) > 0 \text{ for all } i \text{ such that } u_i > 0\}
\]

is bounded if and only if the likelihood function has a unique local maximum in \( \Theta' \). If \( u_i > 0 \) for all \( i \) then \( \Theta' = \Theta \) is bounded and there is a unique local maximum of the likelihood function.

Note that when some of the \( u_i \) are zero, the region \( \Theta' \) might be strictly larger than \( \Theta \). Hence, there might be a unique local maximizer of the likelihood function on \( \Theta' \) but no local maximizer of the likelihood function on \( \Theta \).

**Proof.** Any linear function is concave, and the logarithm of a concave function is concave. The sum of concave functions is concave. This implies that
the log-likelihood function

$$\ell(\theta | u) = \sum_{i=1}^{d} u_i \log f_i(\theta)$$

is a concave function over the region $\Theta'$. If $\Theta'$ is unbounded, by choosing a point $\theta_0 \in \Theta'$ and a direction $\nu$ such that the ray $\{\theta_0 + t\nu : t \geq 0\}$ is contained in $\Theta'$, we see that $\ell(\theta_0 + t\nu | u)$ increases without bound as $t \to \infty$, since each of the linear functions $f_i(\theta_0 + t\nu | u)$ must be nondecreasing as a function of $t$, and at least one is unbounded. Note that if the log-likelihood $\ell(\theta | u)$ is not strictly concave on $\Theta'$, then since all the functions $f_i(\theta)$ are linear, there is a linear space contained in $\Theta'$ on which $\ell$ is not strictly concave. So if $\Theta'$ is bounded, the log-likelihood is strictly concave, and there is a unique local maximum on $\Theta'$, by Proposition 7.3.2.

A similar argument as the proof of Proposition 7.3.4 shows that, for $u$ with $u_i > 0$ for all $i$, the likelihood function $L(\theta | u)$ has a real critical point on each bounded region of the arrangement complement

$$\mathcal{H}(f) = \{ \theta \in \mathbb{R}^d : \prod_{i=1}^{r} f_i(\theta) \neq 0 \}.$$ 

Remarkably, this lower bound on the maximum likelihood degree is tight.

**Theorem 7.3.5.** [Var95] The maximum likelihood degree of the discrete linear model parametrized by $f_1, \ldots, f_r$ equals the number of regions in the arrangement complement $\mathcal{H}(f)$.

**Example 7.3.6** (Mixture of Known Distributions). A typical situation where discrete linear models arise is when considering mixture distributions of known distributions. For example, consider the three probability distributions

$P_1 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}), \quad P_2 = (\frac{1}{5}, \frac{2}{5}, 0, \frac{2}{5}), \quad P_3 = (\frac{2}{5}, \frac{1}{5}, \frac{1}{5}, 0),$

and consider the statistical model $\mathcal{M} \subseteq \Delta_3$ that arises from taking all possible mixtures of these three distributions. This will consist of all distributions of the form

$$\theta_1(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}) + \theta_2(\frac{1}{5}, \frac{2}{5}, 0, \frac{2}{5}) + (1 - \theta_1 - \theta_2)(\frac{2}{5}, \frac{1}{5}, \frac{1}{5}, 0)$$

with $0 \leq \theta_1, 0 \leq \theta_2$, and $\theta_1 + \theta_2 \leq 1$. This is a subset of a discrete linear model, with linear parametrizing functions given by

$$f_1(\theta_1, \theta_2) = \frac{2}{3} - \frac{5}{12} \theta_1 - \frac{7}{15} \theta_2, \quad f_2(\theta_1, \theta_2) = \frac{1}{6} + \frac{1}{12} \theta_1 + \frac{7}{30} \theta_2,$$

$$f_3(\theta_1, \theta_2) = \frac{1}{6} + \frac{1}{12} \theta_1 - \frac{1}{6} \theta_2, \quad f_4(\theta_1, \theta_2) = \frac{1}{4} \theta_1 + \frac{2}{5} \theta_2.$$. 
7.3. Concave Likelihood Functions

Figure 7.3.1. The line arrangement from Example 7.3.6. The gray triangle is the region of the parameter space corresponding to statistically relevant parameters from the standpoint of the mixture model.

Figure 7.3.1 displays the configuration of lines given by the zero sets of each of the polynomials $f_1, \ldots, f_4$. The gray triangle $T$ denotes the region of $\theta$ space corresponding to viable parameters from the standpoint of the mixture model.

The polyhedral region containing $T$ is the region $\Theta$ on which $f_i(\theta_1, \theta_2) > 0$ for $i = 1, 2, 3, 4$. Note that the likelihood function will have a unique maximum in $T$ if all $u_i > 0$. This follows because $\ell$ will be strictly concave on $\Theta$ by Proposition 7.3.4 and since $T$ is a closed bounded convex subset of $\Theta$, $\ell$ must have a unique local maximum.

Note that in Figure 7.3.1 we see that the arrangement complement $\mathcal{H}(f)$ has three bounded regions so the maximum likelihood degree of this discrete linear model is three. Note that the maximum likelihood degree might drop if the data is such that the maximum likelihood estimate lies on the boundary of the parameter space. For example, this happens when the maximum likelihood estimate is a vertex of $T$.

Consider the data vector $u = (3, 5, 7, 11)$. The following code in R

```r
u <- c(3, 5, 7, 11)
p1 <- c(1, 1, 1, 1) / 4
p2 <- c(1, 2, 0, 2) / 5
p3 <- c(4, 1, 1, 0) / 6
l <- function(th) sum( u*log( th[1]*p1 + th[2]*p2 + (1-th[1]-th[2])*p3 ) )
```

can be used to compute the maximum likelihood estimate of the parameters in the discrete mixture model given this data.
ui <- matrix(c(
1, 0,
0, 1,
-1, -1
), nrow = 3, byrow = TRUE)
ci <- c(0, 0, -1)
constrOptim(
  c(1/3, 1/3), l, method = "Nelder-Mead",
  ui = ui, ci =
ci, control = list(fnscale = -1))

This returns the maximum likelihood estimates \( \hat{\theta} = (0.85323030, 0.1467697) \).
Note that this point lies on the boundary of parameter space since \( \hat{\theta}_1 + \hat{\theta}_2 = 1 \). If we wanted to consider maximum likelihood estimate over the entire region \( \Theta \) where all the \( f_i(\theta_1, \theta_2) \) are positive, we change the constraint set in the optimization to the following.

ui <- matrix(c(
-5/12, -7/15,
1/12, 7/30,
1/2, -1/6,
1/4, 2/5
), nrow = 4, byrow = TRUE)
ci <- c(-2/3, -1/6, -1/6, 0)

Now the maximum likelihood estimate is \( \hat{\theta} = (1.35733323, -0.01072592) \).

A second situation where the log-likelihood function is concave and there is a direct iterative procedure to compute the maximum likelihood estimate is for canonical exponential families. A standard algorithm for computing the maximum likelihood estimate in this case is called iterative proportional fitting (also iterative proportional scaling is used).

To prove that the likelihood function is concave for canonical exponential families uses some basic properties of these models.

**Proposition 7.3.7.** Let \( \mathcal{M} \) be a canonical exponential family with sufficient statistics \( T(x) \) and natural parameter \( \eta \), with density \( p_\eta(x) = h(x) \exp(\eta T(x) - \phi(\eta)) \). Then

\[
\frac{\partial}{\partial \eta_i} \phi(\eta) = \mathbb{E}_\eta[T_i(X)] \quad \text{and} \quad \frac{\partial^2}{\partial \eta_i \eta_j} \phi(\eta) = \text{Cov}_\eta[T_i(X), T_j(X)].
\]
Proof. Recall that \( \phi(\eta) = \log \int_{x \in X} h(x) \exp(\eta T(x)) dx \). Differentiating with respect to \( \eta_i \) yields
\[
\frac{\partial}{\partial \eta_i} \phi(\eta) = \frac{\int_{x \in X} T_i(x) h(x) \exp(\eta T(x)) dx}{\int_{x \in X} h(x) \exp(\eta T(x)) dx} = E_\eta[T_i(X)].
\]
Let \( Z(\eta) = \int_{x \in X} h(x) \exp(\eta T(x)) dx \). Differentiating this expression again with respect to \( \eta_j \) and using the quotient rule for derivatives yields
\[
\frac{\partial^2}{\partial \eta_i \partial \eta_j} \phi(\eta) = \frac{\int_{x \in X} T_i(x) T_j(x) h(x) \exp(\eta T(x)) dx}{Z(\eta)} = E_\eta[T_i(X) T_j(X)] - E_\eta[T_i(X)] E_\eta[T_j(X)] = \text{Cov}_\eta[T_i(X), T_j(X)].
\]

From Proposition 7.3.7 we can easily see that the likelihood function of a canonical exponential family must be concave, as shown in the following corollary.

**Corollary 7.3.8.** Let \( \mathcal{M} \) be a canonical exponential family with minimal sufficient statistics \( T(x) \) and natural parameter \( \eta \), with density \( p_\eta(x) = h(x) \exp(\eta T(x) - \phi(\eta)) \). Then the likelihood function is strictly concave. Furthermore the maximum likelihood estimate, if it exists, is the solution to the equation
\[
(7.3.1) \quad T(x) = E_\eta[T(X)]
\]
where \( x \) denotes the data vector.

**Proof.** We compute the log-likelihood given the data \( x \), \( \ell(\eta \mid x) = \log h(x) + \eta T(x) - \phi(\eta) \). Differentiating with respect to \( \eta_i \) and setting equal to zero yields the score equation \( T_i(x) = E_\eta[T_i(X)] \), from which the formula for the maximum likelihood estimate \( T(x) = E_\eta[T(X)] \) follows. Computing second partial derivatives and using Proposition 7.3.7 we see that the matrix of second partials is the negative of the covariance matrix of the vector \( T(X) \). Since \( T(X) \) are the minimal sufficient statistics of an exponential family, the covariance matrix is positive definite. This Hessian matrix is thus negative definite, which implies the log-likelihood function is strictly concave. \( \square \)

Note that Equation 7.3.1 implies that the maximum likelihood estimate in an exponential family can also be thought of as a method of moments estimator (where the moments are exactly the expected values of the sufficient statistics). This explains why a number of our previous computations of maximum likelihood estimate formulas in Chapter 5 agreed with the method of moments estimators.
It is useful to consider how to represent Equation 7.3.1 in various concrete situations. Note, in particular, that this immediately applies to maximum likelihood estimation in exponential families with i.i.d. data. This is because the taking i.i.d. samples \(X^{(1)}, \ldots, X^{(n)}\) immediately yields a new exponential family with the same parameter \(\eta\), sufficient statistic
\[
T_n(X^{(1)}, \ldots, X^{(n)}) = \sum_{i=1}^{n} T(X^{(i)})
\]
and with
\[
h_n(X^{(1)}, \ldots, X^{(n)}) = \prod_{i=1}^{n} h(X^{(i)}).
\]
We illustrate this for discrete and Gaussian exponential families and the resulting algebraic expressions of the score equations in implicit and explicit form. This is summarized in the next two corollaries. Then we explain the iterative proportional fitting algorithm for these models.

**Corollary 7.3.9.** Let \(A \in \mathbb{Z}^{k \times r}\) such that \(1 \in \text{rowspan}(A)\), let \(h \in \mathbb{R}_{\geq 0}^{r}\), and let \(u\) be the vector of counts from \(n\) i.i.d. samples. Then the maximum likelihood estimate in the log-linear model \(M_{A,h}\) given the data \(u\) is the unique solution, if it exists, to the equations:
\[
Au = nAp \quad \text{and} \quad p \in M_{A,h}.
\]

**Proof.** The sufficient statistic vector of the data \(u\) is \(Au\). Given a distribution \(p \in M_{A,h}\) and a sample of size \(n\), the expected value of the sufficient statistics is \(nAp\). Thus, the result follows from Equation 7.3.1. \(\square\)

The proof of the analogous result for Gaussian exponential families is similar and left to the reader.

**Corollary 7.3.10.** Let \(L\) be a linear space in \(\mathbb{R}^{m(m+1)/2}\) such that \(L \cap PD_m\) is not empty, and let \(\mathbb{R}^{m} \times M_{L^{-1}}\) be the corresponding parameter space of the canonical Gaussian exponential family. Let \(X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^{m}\) be i.i.d. samples and let \(\bar{X}\) and \(S\) be the corresponding sample mean and sample covariance matrix. Then the maximum likelihood estimate for \((\mu, \Sigma) \in \mathbb{R}^{m} \times M_{L^{-1}}\) is \((\bar{X}, \hat{S})\) where \(\hat{S}\) is the unique solution, if it exists, to the equations
\[
\pi(S) = \pi(\hat{S}) \quad \text{and} \quad \hat{S} \in M_{L^{-1}}
\]
where \(\pi\) denotes the orthogonal projection onto \(L\).

Here we explain how to use the iterative proportional fitting algorithm to compute maximum likelihood estimates for discrete exponential families \(M_{A,h}\). The basic idea of the method is as follows. We start the algorithm with any positive distribution \(p \in M_{A,h}\). This \(p\) is probably not the maximum likelihood estimate, in particular it does not satisfy \(Ap = A_{u}^{\frac{n}{n}}\). To make
Table 7.3.1. Plum root cuttings

<table>
<thead>
<tr>
<th>Length</th>
<th>Time</th>
<th>Dead</th>
<th>Alive</th>
</tr>
</thead>
<tbody>
<tr>
<td>long</td>
<td>at once</td>
<td>84</td>
<td>156</td>
</tr>
<tr>
<td></td>
<td>in spring</td>
<td>156</td>
<td>84</td>
</tr>
<tr>
<td>short</td>
<td>at once</td>
<td>133</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td>in spring</td>
<td>209</td>
<td>31</td>
</tr>
</tbody>
</table>

$p$ fit the data more closely we rescale the distribution by values related to this failure of the matching of the sufficient statistics to get a new distribution that is closer to having the correct sufficient statistics. We suppose that $A$ has the property that all column have the same sum $a$.

Algorithm 7.3.11 (Iterative Proportional Fitting). Input: The matrix $A \in \mathbb{N}^{k \times r}$, a vector $h \in \mathbb{R}_{>0}$, a vector of counts $u \in \mathbb{N}$ and a tolerance $\epsilon > 0$.

Output: Maximum likelihood estimate $\hat{p} \in \mathcal{M}_{A,h}$ given data $u$.

Step 1: Initialize a distribution $p \in \mathcal{M}_{A,h}$. For example, take $p_i = h_i / (\sum_j h_j)$.

Step 2: While $\|Ap - \frac{u}{n}\| > \epsilon$ do:

For all $i \in [r]$ set $p_i := p_i \cdot \left( \frac{A_{i,h}(Au/n)}{\phi_i^{A,h}(Ap)} \right)^{1/a}$

Step 3: Output $\hat{p} = p$.

The proof of convergence of iterative proportional scaling for discrete exponential families is due to Darroch and Ratcliff [DR72]. It is also possible to derive similar algorithms for canonical exponential families that are not discrete.

Example 7.3.12 (Plum root cuttings). Consider the model of no 3-way interaction. This is an example of a hierarchical model, to be introduced in more detail in Section 9.3. This model is a special type of discrete exponential family, that can be expressed in the following monomial parametrization.

$$p_{ijk} = \alpha_{ij} \beta_{ik} \gamma_{jk}.$$ 

For binary random variables, the homogeneous vanishing ideal is the principal ideal

$$\langle p_{111}p_{122}p_{212}p_{221} - p_{112}p_{121}p_{211}p_{222} \rangle.$$ 

We consider the dataset from [Bar35] displayed in Table 7.3.1. An experiment was designed to investigate the effect of cutting length and planting time on the survival of plum root cuttings. 240 cuttings for each of the 4 combinations of short/long and at once/in spring were made and then at a later time their survival was recorded. The maximum likelihood estimate
of the parameters in the no-3-way interaction model can be calculated with
the following R code.

```r
tab <- array(c(84, 156, 156, 84, 133, 209, 107, 31),
             dim = c(2, 2, 2))
mod <- loglin(tab, list(c(1,2), c(1,3), c(2,3)), fit = TRUE)
mod$fit
```

This produces the maximum likelihood estimate of the table of counts
under the model of no-3-way interaction.

<table>
<thead>
<tr>
<th>Survival Length</th>
<th>Time</th>
<th>Dead</th>
<th>Alive</th>
</tr>
</thead>
<tbody>
<tr>
<td>long</td>
<td>at once</td>
<td>78.908</td>
<td>161.095</td>
</tr>
<tr>
<td>in spring</td>
<td></td>
<td>161.092</td>
<td>78.905</td>
</tr>
<tr>
<td>short</td>
<td>at once</td>
<td>138.097</td>
<td>101.901</td>
</tr>
<tr>
<td>in spring</td>
<td></td>
<td>203.903</td>
<td>36.097</td>
</tr>
</tbody>
</table>

Note that the function `loglin` is computing the maximum likelihood esti-
mates via iterative proportional fitting. However, the implementation works
only in the case of hierarchical models where the specification of the facets
of the corresponding simplicial complex is required.

7.4. Likelihood Ratio Tests

After we have computed the maximum likelihood estimate of the model pa-
rameters given data, a natural next step is to try to address the question of
how well the model fits the data. A standard approach to this problem is
based on the likelihood ratio test. We typically have two models $M_0 \subseteq M_1$,
we compute the maximum likelihood estimate in each model and compare
the evaluation of the likelihood function at these estimates. If the ratio of
likelihood scores is larger than one would expect by chance, we reject the
hypothesis that the true underlying parameter belongs to the smaller model.
Theoretical study of this process involves understanding the asymptotic dis-
tribution of the likelihood ratio test statistic, and algebraic geometry enters
the picture because the asymptotic distribution of the likelihood ratio test
statistic changes at singular points of the models. Our presentation follows
that of [Drt09b], which contains extensive details and examples.

Let $M_0 = \{P_\theta : \theta \in \Theta\}$ be a statistical model where $\Theta \subseteq \mathbb{R}^k$ is the
parameter space which we assume is open and has dimension $k$. Consider
a submodel $M_{\Theta_0} \subseteq M_\Theta$ obtained by considering a subset of the parameter
space $\Theta_0 \subseteq \Theta$. We obtain independent random samples $X^{(1)}, \ldots, X^{(n)}$ which
we assume are distributed according to some distribution $P_\theta \in M_\Theta$. We
would like to test the hypothesis:

$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta \setminus \Theta_0.$

(7.4.1)
7.4. Likelihood Ratio Tests

The likelihood ratio test is one method for performing this hypothesis test. We assume that the distributions in \( M \times \) have densities \( p_\theta(x) \) with respect to a fixed measure on the sample space. Thus the log-likelihood function of the model \( M_\Theta \) has the form

\[
\ell_n(\theta) = \sum_{i=1}^{n} \log p_\theta(X^{(i)}).
\]

**Definition 7.4.1.** The **likelihood ratio statistic** for the hypothesis test (7.4.1) is

\[
\lambda_n = 2 \left( \sup_{\theta \in \Theta} \ell_n(\theta) - \sup_{\theta \in \Theta_0} \ell_n(\theta) \right).
\]

Since \( \Theta_0 \subset \Theta \), \( \lambda_n \geq 0 \). The likelihood ratio test assesses whether \( \lambda_n \) is too large for what one would expect if \( \theta \in \Theta_0 \), and rejects the null hypothesis if that is the case. Our goal is to try to understand what “too large” means. Note that we explicitly express the dependence of the log-likelihood function and the likelihood ratio statistic on the sample size, \( n \), because we will often need to understand this expression as \( n \) goes to infinity. To evaluate \( \lambda_n \) requires computing maximum likelihood estimates in \( M_\Theta \) and \( M_{\Theta_0} \) which can be achieved using methods from previous sections of this chapter.

To illustrate the main idea, we will first focus on a very special case: models of Gaussian random variables with restricted mean parameter. In particular let \( \Theta = \mathbb{R}^m \) and let the model \( M_\Theta \) be the family of normal distributions

\[
\{ \mathcal{N}(\mu, Id) : \mu \in \mathbb{R}^m \}
\]

where \( Id \) is the \( m \times m \) identity matrix. Let \( \bar{X} \) denote the sample mean of \( X^{(1)}, \ldots, X^{(n)} \). Evaluating (7.1.2) with \( \Sigma = Id \) gives the log-likelihood function

\[
\ell_n(\mu) = -\frac{n}{2} \| \bar{X} - \mu \|^2 - \frac{1}{2} \sum_{i=1}^{n} \| X^{(i)} - \bar{X} \|^2 - C_n
\]

where \( C_n \) is a constant, depending on \( n \) but not the distribution, that will drop out when we compute the likelihood ratio statistic. Since the maximum likelihood estimate of \( \mu \) over \( M_\Theta \) is the sample mean \( \bar{X} \), the likelihood ratio statistic becomes

\[
\lambda_n = n \cdot \inf_{\mu \in \Theta_0} \| \bar{X} - \mu \|^2.
\]

As we discussed in Section 7.1, maximum likelihood estimation in this case amounts to computing the least squares closest point in the model \( \Theta_0 \) (see Proposition 7.1.6). The likelihood ratio statistic is the squared Euclidean distance to \( \Theta_0 \), scaled by the sample size \( n \).

To know when to reject the null hypothesis via the likelihood ratio test, we need to understand when \( \lambda_n \) is large assuming the data was generated
from a $\mu_0 \in \Theta_0$. In particular, we must understand what probability distribution $\lambda_n$ has. To do this, we rewrite

$$\lambda_n = \inf_{\mu \in \Theta_0} \| \sqrt{n}(\bar{X} - \mu_0) - \sqrt{n}(\mu - \mu_0) \|^2.$$ 

We need to analyze the distribution of $\lambda_n$. Since $X^{(1)}, \ldots, X^{(n)}$ are i.i.d. $\mathcal{N}(\mu_0, Id)$ random variables, the term $\sqrt{n}(\bar{X} - \mu_0)$ is distributed as a standard normal random variable $Z \sim \mathcal{N}(0, Id)$, by the central limit theorem. So we can rewrite the likelihood ratio statistic as a random variable as

$$\lambda_n = \inf_{h \in \sqrt{n}(\Theta_0 - \mu_0)} \| Z - h \|^2.$$ 

In the special case where $\Theta_0$ is a linear space, we have an exact characterization of this distribution regardless of $n$.

**Proposition 7.4.2.** Let $\Theta_0 \subseteq \Theta = \mathbb{R}^m$ be a linear space of dimension $d$, $\mu_0 \in \Theta_0$, and consider the hypothesis test (7.4.1) for the models

$$\mathcal{M}_{\Theta_0} \subseteq \mathcal{M}_\Theta = \{ \mathcal{N}(\mu, Id) : \mu \in \mathbb{R}^m \}.$$ 

Then the likelihood ratio statistics $\lambda_n$ has a $\chi^2_{m-d}$-distribution.

**Proof.** If $\Theta_0$ is a linear space then $L = \sqrt{n}(\Theta_0 - \mu_0)$ is a linear space passing through the origin. As $Z \sim \mathcal{N}(0, Id)$ is rotationally invariant, we can assume that $L$ is a coordinate subspace where the first $d$ coordinates are nonzero. The squared distance between a point $Z = (Z_1, \ldots, Z_m)$ and $L$ is $\sum_{i=d+1}^m Z_i^2$ which has a $\chi^2_{m-d}$-distribution by definition. \qed

This leads to an explicit $p$-value calculation for the likelihood ratio test in the case that $\Theta_0 \subseteq \mathbb{R}^m$ is a linear space of dimension $d$. Namely we should reject the null hypothesis at level $\alpha$ if

$$P(\chi^2_{m-d} \geq \lambda_{obs}) \leq \alpha.$$ 

Here $\lambda_{obs}$ denotes the observed value of $\lambda_n$ obtained from a specific data set.

If $\Theta_0$ is not a linear space, the set $\sqrt{n}(\Theta_0 - \mu_0)$ will not be a fixed set as $n$ varies, and hence the distribution of $\lambda_n$ will change as a function of $n$. To deal with this problem, we assume we have a large sample size and take the limit as $n \to \infty$. If $\mu_0 \in \Theta_0$, as $n \to \infty$, the limiting object

$$\lim_{n \to \infty} \sqrt{n}(\Theta_0 - \mu_0)$$

is the tangent cone of $\Theta_0$ at $\mu_0$, $TC_{\mu_0}(\Theta_0)$ (see Definition 6.4.6). In other words:

**Proposition 7.4.3.** Let $\Theta_0 \subseteq \Theta = \mathbb{R}^m$, and $\mu_0 \in \Theta_0$. Then as $n \to \infty$,

$$\lambda_n \overset{D}{\to} \inf_{h \in TC_{\mu_0}(\Theta_0)} \| Z - h \|^2.$$
where $Z \sim \mathcal{N}(0, \text{Id})$. That is, the limiting distribution of $\lambda_n$ is the distribution of a standard normal random variable to the tangent cone of $\Theta_0$ at $\mu_0$.

Suppose that we remain in the case of Gaussian models with $\Theta_0 \subset \Theta = \mathbb{R}^m$ a smooth manifold of dimension $d$. Each tangent cone $TC_{\mu_0}(\Theta_0)$ is a linear space of dimension $d$. From Proposition 7.4.3 we deduce that, regardless of what the true underlying parameter $\mu_0 \in \Theta_0$ is, the limiting distribution of $n$ as $n \to \infty$ is a $\chi^2_{m-d}$ distribution. This leads to the asymptotic likelihood ratio test for smooth models: namely, take the asymptotic/approximate $p$-value

\begin{equation}
(7.4.2) \quad p = \lim_{n \to \infty} P(\lambda_n \geq \lambda_{\text{obs}}) = P(\chi^2_{m-d} \geq \chi^2_{\text{obs}})
\end{equation}

and reject at level $\alpha$ if this is less than $\alpha$. Such a limit can be used for any model that is a smooth submodel of a regular exponential family, and using this asymptotic approximation to a smooth model is usually what statisticians refer to when they talk about using a likelihood ratio test. Of course, as we have seen in Chapter 6, statistical models need not be smooth manifolds. The distribution of the distance from a standard normal to the tangent cone will usually not be a chi-square distribution in this case.

**Example 7.4.4** (Likelihood ratio statistic of cuspidal cubic). Let $\Theta_0 \subset \mathbb{R}^2$ be the cuspidal cubic curve

$$\Theta_0 = \{(\mu_1, \mu_2) \in \mathbb{R}^2 : \mu_2^2 = \mu_1^3\}$$

see Example 7.1.7 and Figure 7.1.1. The only singularity of $\Theta_0$ is at the origin, $\mu_0 = (0,0)$, and the tangent cone is the positive part of the $\mu_1$-axis:

$$TC_{(0,0)}(\Theta_0) = \{(\mu_1,0) : \mu_1 \geq 0\}.$$ 

Taking $Z \sim \mathcal{N}(0, \text{Id})$, we see that the squared Euclidean distance from $Z$ to $TC_{(0,0)}(\Theta_0)$ is either $Z_2^2$ if $Z_1 \geq 0$, or is $Z_1^2 + Z_2^2$ if $Z_1 \leq 0$. Since $P(Z_1 \leq 0) = P(Z_1 \geq 0) = \frac{1}{2}$ it follows that $\lambda_n$ converges to a mixture of chi-square distributions:

$\lambda_n \xrightarrow{D} \frac{1}{2} \chi^2_1 + \frac{1}{2} \chi^2_2$.

Since $P(\chi^2_1 \geq t) < P(\chi^2_2 \geq t)$ for all $\alpha > 0$, we see that if the true parameter $\mu_0$ were the origin, a typical asymptotic $p$-value computation based on the $\chi^2_1$ distribution would yield a $p$-value that is too small, relative to the actual distribution. See Figure 7.4.1.

**Example 7.4.5** ( Likelihood ratio statistic of nodal cubic). On the other hand, suppose $\Theta_0 \subset \mathbb{R}^2$ is the nodal cubic curve

$$\Theta_0 = \{(\mu_1, \mu_2) \in \mathbb{R}^2 : \mu_2^2 = \mu_1^3 + \mu_1^3\}.$$
This model also has a single singularity at $\mu_0 = (0, 0)$ and the tangent cone is

$$TC_{(0,0)}(\Theta_0) = \{(\mu_1, \mu_2) \in \mathbb{R}^2 : \mu_2^2 = \mu_1^2\}.$$ 

This tangent cone is the union of two perpendicular lines that cross at the origin. After applying a rotation, we can treat these two lines as the coordinate axes. If $Z \sim N(0, I)$ the squared Euclidean distance to the closest coordinate axis is $\min(Z_1^2, Z_2^2)$. In other words,

$$\lambda_n \xrightarrow{D} \min(\chi_1^2, \chi_1^2).$$

Since the $P(\min(Z_1^2, Z_2^2) < t) < P(\min(Z_1^2 < t)$ for all $t$, we see that an asymptotic $p$-value that was approximated using just the $\chi_1^2$ distribution would be too large compared to the true asymptotic distribution.

In Examples 7.4.4 and 7.4.5 we saw relatively clean expressions for the limiting distribution that arise in the likelihood ratio test. In general these limiting distributions need not have simple closed expressions. For example, replacing Example 7.4.5 with the nodal cubic

$$\Theta_0 = \{(\mu_1, \mu_2) \in \mathbb{R}^2 : \mu_2^2 = \mu_1^3 + 2\mu_1^2\}$$

yields a similar singularity but the limiting distribution does not have a simple expression as the minimum of two $\chi_1^2$ distributions. In general, one probably needs to run simulations to determine a histogram for the asymptotic distribution of the likelihood ratio test statistic as a total to approximate the asymptotic $p$-values.
So far we have only discussed very special models that arise as submodels of $\mathbb{R}^m$ as a restricted set of possible mean vectors for multivariate normal distribution, with identity as the covariance matrix. But the same general principles apply to arbitrary statistical models that are semialgebraic submodels of regular exponential families.

**Definition 7.4.6.** Let $\mathcal{M}_\Theta$ be a statistical model where $\Theta \subseteq \mathbb{R}^k$, and the $\theta \in \Theta$. The Fisher information matrix $I(\theta)$ is the $k \times k$ matrix with entries

$$I(\theta)_{ij} = \mathbb{E} \left[ \frac{\partial}{\partial \theta_i} \log p_\theta(X) \left( \frac{\partial}{\partial \theta_j} \log p_\theta(X) \right) \right].$$

The expectation is calculated assuming that $X \sim P_\theta$.

In the case that the model $\mathcal{M}_\Theta$, consists of multivariate normal distributions with covariance matrix equal to the identity, the Fisher information matrix is just the identity matrix. Note that the Fisher information will always be a symmetric positive semi-definite matrix, so $I(\theta)$ always has a matrix square root. In more general situations than independent normal random variables, the tangent cone gets perturbed by the Fisher information matrix. This is one of the main results of [Drt09b].

**Theorem 7.4.7.** Suppose that the model $\mathcal{M}_\Theta$ is a regular exponential family with parameter space $\Theta \subseteq \mathbb{R}^k$ and let $\theta_0$ be a semi-algebraic subset of $\Theta$. If the true parameter $\theta_0$ is in $\Theta_0$ and $n \to \infty$, then the likelihood ratio statistic converges in distribution to the distribution of the squared Euclidean distance

$$\min_{h \in T C_{\Theta_0}(\theta_0)} \left\| Z - I(\theta_0)^{1/2}h \right\|^2_2$$

between $Z \sim \mathcal{N}(0, I_d)$ and the transformed tangent cone $I(\theta_0)^{1/2}T C_{\Theta_0}(\Theta_0)$. Here $I(\theta_0)^{1/2}$ can be any matrix square root of $I(\theta_0)$.

Given the difficulties that can arise when models have singularities, it is natural to wonder what we should actually do when attempting to use the likelihood ratio tests with real data. In practice we do not know whether or not the true parameter $\theta_0$ belongs to $\Theta_0$ nor whether or not $\theta_0$ is a singular point of $\Theta_0$. So we do not know which distribution we should use when deciding to calculate a $p$-value, since we do not know the true value of $\theta$. Perhaps the most conservative choice is to use all the singularity information from the model $\Theta_0$ to produce a range of $p$-values and use this range to make some decision on whether to reject the null hypothesis.

### 7.5. Exercises

**Exercise 7.1.** Maximum likelihood estimation induces a map from the probability simplex, into the parameters. In Example 7.3.6, that function
$ML : \Delta_3 \to T$ is sometimes an algebraic function of degree 3, and sometimes an algebraic function of lower degree, depending on which face of $T$ the function $ML$ maps to. Determine the decomposition of $\Delta_3$ induced by the $ML$ map in this case.

**Exercise 7.2.** Let $M$ be the model of binomial random variables $\text{Bin}(2, \theta)$,

$$M = \{(\theta^2, 2\theta(1-\theta), \theta^2) \in \Delta_2 : \theta \in (0, 1)\}$$

and $\text{Mixt}^2(M)$, its mixture model. Compute the maximum likelihood estimates in both models given for the two data points $u = (8, 6, 5)$ and $v = (4, 20, 8)$.

**Exercise 7.3.** Consider the log-linear model described parametrically via the formula

$$p_{ijkl} = \frac{1}{Z} \alpha_{ij} \beta_{jk} \gamma_{kl}$$

for discrete random variables $X_1, X_2, X_3, X_4$, each with two states.

1. Show that this model has ML-degree one.
2. Realize this model via a Horn uniformization.

**Exercise 7.4.** Let $A \in \mathbb{Z}^{r \times r}$ with rank $A = r - 1$ and suppose that $1$ is in the row space of $A$, so that $I_A$ is a principal ideal generated by a single homogeneous binomial. Let $d$ be the degree of this binomial. Show that the maximum likelihood degree of $M_A$ is $d$ if $d$ is odd and $d - 1$ if $d$ is even. What happens in the models $M_{A,h}$ (again, assuming that $I_{A,h}$ is a principal ideal)?

**Exercise 7.5.** Calculate the Fisher information matrix when $\Theta = \Delta_{k-1}$ is the probability simplex for a discrete random variable with $k$ states.

**Exercise 7.6.** Let $\Theta_0 = \{(\mu_1, \mu_2) \in \mathbb{R}^2 : \mu_2^2 = \mu_1^3 + 2\mu_1^2\} \subseteq \mathbb{R}^2$ and consider the statistical model $M_{\Theta_0} = \{N(\mu, Id) : \mu \in \Theta_0\}$. Consider the ten data points $X^{(1)}, \ldots, X^{(10)}$ which are the columns of the following $2 \times 10$ matrix:

$$
\begin{pmatrix}
-0.71 & -0.76 & -1.37 & -0.68 & 0.35 & 0.36 & 0.23 & -0.44 & 1.35 & -0.99 \\
1.28 & -0.22 & 0.20 & -0.29 & 0.95 & -0.35 & -0.82 & -0.63 & 0.67
\end{pmatrix}
$$

Use the conservative asymptotic likelihood ratio test outlined at the end of Section 7.4 to decide if this data set was generated from a point on the model.
Chapter 8

The Cone of Sufficient Statistics

For a fixed exponential family \( \mathcal{M} \), the set of all sufficient statistics of (unnormalized, real) data that can arise forms a convex cone. For discrete exponential families this is a convex polyhedral cone, and for Gaussian exponential families it is a cone inside the cone of positive semidefinite matrices. We introduce and study these cones in this chapter. This will also provide us with the introduction to convex and polyhedral geometry, whose study will also play an important role in other chapters in this book. The question of when there exists a maximum likelihood estimate for an exponential family is also related and will be discussed.

8.1. Polyhedral Geometry

In this section we review various notions from convex and polyhedral geometry. Most of the main results are stated without proof, and can be found, for instance, in [Zie95] or [Tho06]. Recall the definition of a convex set from Chapter 7.

**Definition 8.1.1.** A set \( S \subseteq \mathbb{R}^d \) is called convex if for all \( p^1, p^2 \in S \) and all \( \lambda \in [0, 1] \subset \mathbb{R} \), \( \lambda p^1 + (1 - \lambda)p^2 \in S \).

Any linear or affine space is convex. Let \( \mathbf{a} \in \mathbb{R}^d \) and \( b \in \mathbb{R} \). The set \( H = \{ p \in \mathbb{R}^d : \mathbf{a}^T p \leq b \} \) is called a half-space, and is an important example of a convex set. Note that we include \( \mathbf{a} = \mathbf{0} \) so that \( \mathbb{R}^d \) and \( \emptyset \) are both examples of half-spaces. The intersection of any collection of convex sets is convex.
Definition 8.1.2. For an arbitrary set $S \subseteq \mathbb{R}^d$ the convex hull of $S$, denoted $\text{conv}(S)$ is the smallest convex set that contains $S$.

The convex hull of a set $S$ is well-defined since it can be explicitly taken to be the intersection of all convex sets that contain $S$. These definitions are illustrated in Figure 8.1.1.

Definition 8.1.3. A set $P \subseteq \mathbb{R}^d$ that is the intersection of finitely many half-spaces is called a polyhedron. A bounded polyhedron is called a polytope.

The dimension of a polyhedron $P$ is defined to be the dimension of the smallest affine space containing it, and is denoted $\dim P$.

Example 8.1.4 (Hypercube and Simplex). The set of points

$$C_d := \{ p \in \mathbb{R}^d : 0 \leq p_i \leq 1 \text{ for all } i \}$$

is a polytope called the standard $d$-dimensional hypercube. The set of points

$$\Delta_{d-1} := \{ p \in \mathbb{R}^d : 0 \leq p_i, \sum_{i=1}^d p_i = 1 \}$$

is the standard $(d-1)$-dimensional simplex, often called the probability simplex up to this point.

Note that any equality constraint $a^T p = b$ can be realized as the intersection of the two half-spaces $a^T p \leq b$ and $a^T p \geq b$. So combinations of finitely many linear inequality and equality constraints yield polyhedra.

Definition 8.1.5. A face of a polyhedron $P$ is a set of the form $\{ p : a^T p = b \} \cap P$ where the half-space $H = \{ p : a^T p \leq b \}$ contains $P$.

Technically we allow $a$ to be the zero vector in this definition, so that $P$ and $\emptyset$ are always both faces of $P$. The faces of a polyhedron are themselves polyhedra. The 0-dimensional faces are called vertices, the 1-dimensional
faces are called *edges*. If \( \dim P = d \), then the \( d - 1 \)-dimensional faces are called *facets*. For example, the \( d \)-dimensional cube \( C_d \) has \( 2d \) facets, and \( 2^d \) vertices. The \( d - 1 \)-dimensional simplex has \( \binom{d}{i} \) \( (i - 1) \)-dimensional faces for each \( i \). Note that the empty set is considered a \(-1\)-dimensional face.

Every polyhedron has finitely many faces. This follows from the fact that every face must be obtained by taking a subset of the half-space defining inequalities of \( P \) and making some of them equalities. So if \( P \) has \( m \) facets, it has at most \( 2^m \) faces. This upper bound on the number of faces is attained only for the simplex. Usually, there are many fewer faces than \( 2^m \).

**Definition 8.1.6.** Two polyhedra \( P \subseteq \mathbb{R}^d \) and \( Q \in \mathbb{R}^e \) are *affinely isomorphic* if there is an affine transformation \( \phi : \mathbb{R}^d \to \mathbb{R}^e \) such that \( \phi(P) = Q \), and an affine transformation \( \psi : \mathbb{R}^e \to \mathbb{R}^d \) such that \( \psi(Q) = P \). The polyhedra \( P \) and \( Q \) are *combinatorially equivalent* if there is a bijection between the faces of \( P \) and the faces of \( Q \) which preserves the incidences between faces.

Note that any polytope combinatorially equivalent to the standard simplex is affinely isomorphic to the standard simplex. On the other hand, a polytope combinatorially equivalent to the standard cube need not be affinely isomorphic to it.

The definition of polyhedra we have given so far is in terms of half-spaces. Such a polyhedron is sometimes called an \( \mathcal{H} \)-polyhedron (the \( \mathcal{H} \) for “half-space”). Polyhedra can also be described in a dual approach called a \( \mathcal{V} \)-polyhedron (the \( \mathcal{V} \) for vertex). In the special case of polytopes, the \( \mathcal{V} \)-description is easy to state.

**Theorem 8.1.7.** The convex hull of any finite set is a polytope. Every polytope is equal to the convex hull of its set of vertices.

Hence, the \( \mathcal{V} \)-description of a polytope is a presentation \( P = \text{conv}(V) \) for some finite set \( V \). Theorem 8.1.7 will follow from the generalization Theorem 8.1.11. To explain this generalization of this result to arbitrary polyhedra, we need a few more definitions.

**Definition 8.1.8.** A *polyhedral cone* \( C \) is a set of the form

\[
C = \{ p \in \mathbb{R}^d : Ap \leq 0 \}
\]

where \( A \in \mathbb{R}^{k \times d} \) is a matrix. For two vectors, \( p, q \in \mathbb{R}^k \), \( p \leq q \) means that the inequality holds for all coordinates.

Equivalently, a polyhedral cone is a polyhedron where all the defining half-spaces pass through the origin. Because of this, a polyhedral cone can have at most one vertex, namely, the origin. In the case where the polyhedral
cone $C$ does have the origin as a vertex, it is called a pointed cone. A cone being pointed is equivalent to it not containing any lines.

**Definition 8.1.9.** Let $V = \{v_1, \ldots, v_k\} \subseteq \mathbb{R}^d$ be a finite set of vectors. The set
\[
\text{cone}(V) = \{\lambda_1 v_1 + \cdots + \lambda_k v_k : \lambda_1, \ldots, \lambda_k \geq 0\}
\]
is the cone generated by $V$.

Definition 8.1.9 gives the $V$-description of a polyhedral cone.

**Definition 8.1.10.** Let $S, T \subseteq \mathbb{R}^d$ be two sets. The Minkowski sum is the set $S + T = \{p + q : p \in S \text{ and } q \in T\}$.

Note that if $S$ and $T$ are convex sets, then so is the Minkowski sum, $S + T$. In general, a $V$-polyhedron is a set that can be expressed in the form: $\text{conv}(V_1) + \text{cone}(V_2)$ for two finite sets $V_1$ and $V_2 \subseteq \mathbb{R}^d$. The main theorem relating $H$-polyhedra and $V$-polyhedra is the following.

**Theorem 8.1.11 (Minkowski-Weyl).** Every $H$-polyhedron is a $V$-polyhedron, and every $V$-polyhedron is a $H$-polyhedron. For every polyhedron $P$, if $P = C + Q$ where $C$ is a cone and $Q$ is a polytope, then $C$ is uniquely determined by $P$ and is called the recession cone of $P$. If the polyhedron $P$ contains no lines, $Q$ can be taken to be the convex hull of the set of vertices of $P$.

A proof of Theorem 8.1.11 is a standard result in polyhedral geometry and can be found, for instance in [Zie95]. For a pointed polyhedral cone $C$, there is a unique minimal set of vectors (up to scaling) $V$ such that $\text{cone}(V) = C$. The set $V$ are called the extreme rays of $C$.

**Example 8.1.12 (Small Minkowski Sum).** Let $P$ be the polyhedron
\[
P = \{(x, y) \in \mathbb{R}^2 : x \geq 0, y \geq 1, x + y \geq 2\}.
\]
Then $P = \text{cone}(V) + \text{conv}(W)$ where
\[
V = \{(1, 0), (0, 1)\} \quad \text{and} \quad W = \{(1, 1), (0, 2)\}.
\]
The decomposition is illustrated in Figure 8.1.2.

### 8.2. Discrete Exponential Families

Polyhedral geometry arises when studying exponential families since for a discrete exponential family, the set of all sufficient statistics for all (unnormalized nonnegative real) data is a polyhedral cone, the cone of sufficient statistics.

Recall the setup from Section 6.2. Let $A \in \mathbb{Z}^{k \times r}$ with $\mathbf{1} \in \text{rowspan}(A)$. Let $\mathcal{M}_A$ be the regular exponential family defined by $A$. Let $X^{(1)}, \ldots, X^{(n)}$
be i.i.d. samples from a distribution \( p \in \mathcal{M}_A \). The probability that \( X^{(i)} = j \) only depends on the parameter vector \( \theta \) through the vector \( a_j \).

Indeed, let \( u \in \mathbb{N}^r \) be the vector of counts, \( u_j = \# \{ i : X^{(i)} = j \} \). Then in the discrete log-linear model \( \mathcal{M}_A \) we have

\[
p_{\theta}(X^{(1)} = j_1, \ldots, X^{(n)} = j_n) = \frac{1}{Z(\theta)^n} \prod_{i=1}^{n} \theta^{a_{j_i}} = \frac{1}{Z(\theta)^n} \theta^{Au}.
\]

Since the probability only depends on the vector of sufficient statistics \( Au \), it makes sense to study the set of all such vectors of sufficient statistics, over all possible data sets. This is the **affine semigroup**

\[ NA := \{ Au : u \in \mathbb{N}^r \}. \]

This is a complicated discrete object, which we will discuss in more detail later. As a first approximation, we can instead study the polyhedral cone

\[ \text{cone}(A) := \{ Au : u \in \mathbb{R}^r_{\geq 0} \}. \]

One reason for studying the polyhedral cone \( \text{cone}(A) \) is because of its connection to maximum likelihood estimation as explained in the following theorem.

**Theorem 8.2.1.** Let \( A \in \mathbb{Z}^{k \times r} \) be the integer matrix with \( 1 \in \text{rowspace}(A) \) let \( h \in \mathbb{R}^r_{>0} \), and let \( \mathcal{M}_{A,h} \) be the corresponding regular exponential family. Let \( u \in \mathbb{N}^r \) be the vector of counts for i.i.d. data. Then a maximum likelihood estimate of parameters given the data \( u \) exists in the model \( \mathcal{M}_{A,h} \) if and only if \( b = Au \) lies in the relative interior of \( \text{cone}(A) \).

**Proof.** According to Corollary 7.3.9, the maximum likelihood estimate, if it exists, is the solution to the equation system

\[
A \frac{u}{n} = Ap \quad \text{and} \quad p \in \mathcal{M}_{A,h}.
\]
By the nature of the log-linear model $\mathcal{M}_{A,h}$ every distribution $p \in \mathcal{M}_{A,h}$ must have all coordinates of $p$ positive. If such a $p$ exists then $A_h p = A p$ could not lie on the boundary of $\text{cone}(A)$ because it is a positive combination of the columns of $A$.

On the other hand, suppose that $b = A u$ lies in the relative interior of $\text{cone}(A)$. Then there is a positive real vector $v$ such that $b = A v$. By the property of the sufficient statistics, maximizing the likelihood $\prod_{i=1}^n p_i^{u_i}$ over $p \in \mathcal{M}_{A,h}$ is equivalent to maximizing the likelihood $\prod_{i=1}^n p_i^{v_i}$ over $p \in \mathcal{M}_{A,h}$. Since all coordinates of $v_i$ are positive, this ensures that $\prod_{i=1}^n p_i^{v_i} = 0$ for any $p \in \mathcal{M}_{A,h}$ that has a zero coordinate. On the other hand, $\prod_{i=1}^n p_i^{v_i} > 0$ for any $p \in \mathcal{M}_{A,h}$. Since the likelihood function is convex, and is zero on the boundary, it must have a maximum in $\mathcal{M}_{A,h}$.

Example 8.2.2 (Non-existent MLE). Consider a modification of the dataset in Example 7.3.12 which is displayed in Table 8.2.1. Note the introduction of two zeros. Analyze this data from the point of view of the no-3-way interaction model. Let $A$ be the associated matrix. Since $\ker A$ is one dimensional, it is straightforward to see that the set

$$\{v \in \mathbb{R}^8_{\geq 0} : A v = A u\}$$

consists of a single point, namely the given data point. From this, we see that $A u = b$ must lie on the boundary of $\text{cone}(A)$, and so the maximum likelihood estimate does not exist.

Suppose we run the following R code to compute the maximum likelihood estimates in the model of no-3-way interaction:

```r
tab <- array(c(0, 156, 156, 84, 133, 209, 107, 0),
dim = c(2, 2, 2))
mod <- loglin(tab, list(c(1,2), c(1,3), c(2,3)), fit = TRUE,
eps = 1e-6, iter = 1000)
mod$fit
```

The output contains the warning message

Warning message:
In loglin(tab, list(c(1, 2), c(1, 3), c(2, 3)), fit = TRUE,
8.2. Discrete Exponential Families

\[
\text{eps} = 1e-06, \ : \\
\text{algorithm did not converge}
\]

The output of the computation is displayed in the following table.

<table>
<thead>
<tr>
<th>Survival Length</th>
<th>Time</th>
<th>Dead</th>
<th>Alive</th>
</tr>
</thead>
<tbody>
<tr>
<td>long at once</td>
<td>0.043</td>
<td>155.990</td>
<td></td>
</tr>
<tr>
<td>in spring</td>
<td>155.957</td>
<td>84.010</td>
<td></td>
</tr>
<tr>
<td>short at once</td>
<td>133.002</td>
<td>106.975</td>
<td></td>
</tr>
<tr>
<td>in spring</td>
<td>208.998</td>
<td>0.0247</td>
<td></td>
</tr>
</tbody>
</table>

This failure to converge is a common problem when \( Au \) lies on the boundary of cone(A) or is near to the boundary.

The condition of Theorem 8.2.1 can be checked using linear programing and the following proposition.

**Proposition 8.2.3.** Suppose that \( A \in \mathbb{Z}^{d \times r} \) satisfies \( 1 \in \text{rowspace}(A) \). Let \( u \) be a vector of counts, and let \( b = Au \) be the vector of sufficient statistics, and let

\[
Q(b) = \{ c \in \mathbb{R}^d : cA \leq 1 \cdot \frac{1}{n} cb \}.
\]

Then \( b \) is in the relative interior of cone(A) if and only if \( Q(u) \) is equal to the orthogonal complement of \( \text{span}(A) \).

The proof of Proposition 8.2.3 is left as an exercise. In practice, checking the condition of Proposition 8.2.3 might involve studying the geometry of a polytope with exponentially many inequalities, for example if \( r \) has exponential size. In many instances, if the model has some underlying combinatorial structure, it might be possible to set up an equivalent system using significantly smaller numbers of variables to decide whether or not \( b \) lies in the relative interior of cone(A). See [EFRS06].

A more direct, but also expensive, approach to deciding the existence of maximum likelihood estimates is to simply list all facet defining inequalities of the cone cone(A). The point \( b \) lies in the relative interior of cone(A) if and only if none of the facet defining inequalities are tight. A similar strategy is based on the fact that whether or not \( b = Au \) is on the boundary of cone(A) can be determined by looking at the pattern of zeros in \( u \). Then one could try to catalog all problematic patterns of zeros and check if a given dataset is of one of these bad types. See [EFRS06] and [FR07] for more details.

Note that if our data set does lie on the boundary of the cone of sufficient statistic, it might still be possible to estimate a generalized maximum likelihood estimate corresponding to distributions on the boundary of the exponential family. This is explored in [CM08, Mat15].
Figure 8.2.1. The semigroup from Example 8.2.4.

A second interesting problem concerning the cone of sufficient statistics for a discrete exponential family concerns the integer points inside that cone. After all, each data vector $u$ is a vector of counts, so the vectors of sufficient statistics $Au$ is also an integral vector. The set of all possible sufficient statistics vectors is a semigroup denoted

$$NA := \{Au : u \in \mathbb{N}^r\}.$$  

Similarly let $ZA := \{Au : u \in \mathbb{Z}^r\}$ denote the lattice spanned by the columns of $A$. Note that $NA \subseteq ZA \cap \text{cone}(A)$ but in general the containment might be strict.

**Example 8.2.4** (Small Semigroup with Holes). Let $A = \begin{pmatrix} 0 & 1 & 3 \\ 1 & 1 & 1 \end{pmatrix}$. Then the vector $\begin{pmatrix} 2 \\ 1 \end{pmatrix}$ belongs to $ZA \cap \text{cone}(A)$ but is not in the semigroup $NA$. Figure 8.2.1 shows the semigroup and the facets of $\text{cone}(A)$.

**Definition 8.2.5.** The semigroup $NA$ is called *saturated* if $NA = ZA \cap \text{cone}(A)$. Alternate terminology is that the matrix $A$ is *normal*. Elements of $(ZA \cap \text{cone}(A)) \setminus NA$ are called *holes*.

In Example 8.2.4, the matrix $A$ is nonnormal, and $\begin{pmatrix} 2 \\ 1 \end{pmatrix}$ is a hole. In fact, the set of all holes for this semigroup are the points of the form $\begin{pmatrix} 3t - 1 \\ t \end{pmatrix}$ with $t = 1, 2, 3, \ldots$.

Here we illustrate these ideas on a family of examples, the binary graph models, or heterogeneous Ising models. Let $G = (V, E)$ be an undirected graph with vertex set $V$ and edge set $E$. We create a matrix $A_G$ with $\#V + \#E + 1$ rows and $2\#V$ columns as follows. Let $\#V = m$. The columns are indexed by elements of $\{0, 1\}^m$. For each such string $\epsilon = (\epsilon_1, \ldots, \epsilon_m)$ we form the vector

$$a_\epsilon = (1, \epsilon_1, \ldots, \epsilon_m) \oplus (\epsilon_i \epsilon_j)_{ij \in E} \in \mathbb{Z}_{\geq 0}^{1+m+\#E}.$$
and let $A_G$ be the matrix which these vectors as its columns.

The corresponding log-linear model gives distributions for $m$ binary random variables of the form

$$P_\theta(X_1 = x_1, \ldots, X_m = x_m) = \frac{1}{Z(\theta)} \prod_{i=1}^m \theta_i^{x_i} \prod_{ij \in E} \theta_{ij}^{x_i x_j}$$

where $\theta = (\theta_i)_{i \in [m]}, (\theta_{ij})_{ij \in E} \in \mathbb{R}^{m + \#E}$.

The cone of sufficient statistics for this model is $\text{cone}(A_G)$ a $1 + m + \#E$ dimensional polyhedral cone with $2^m$ extreme rays $a_e$. The coordinates on $\mathbb{R}^{1 + m + \#E}$ are indexed by $\{\emptyset\} \cup V \cup E$, and we denote them by $p_0, p_i$ and $p_{ij}$, respectively.

**Example 8.2.6** (Cone of a Short Path). Let $G$ be the graph with 3 vertices and two edges $E = \{12, 23\}$. The matrix $A_G$ is a $6 \times 8$ matrix:

$$A_G = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}.$$  

For example, the 7th column is $a \{1,1,0\}$. The cone $A_G$ is described by 8 facet-defining inequalities:

$$p_{12} \geq 0, \quad p_{23} \geq 0, \quad p_1 \geq p_{12}, \quad p_2 \geq p_{12}, \quad p_2 \geq p_{23}, \quad p_3 \geq p_{23}, \quad p_{12} - p_1 - p_2 + p_0 \geq 0, \quad p_{23} - p_2 - p_3 + p_0 \geq 0.$$  

The cone $\text{cone}(A_G)$ is closely related to the polytope $\text{conv}(A_G)$, which is a well-studied object in optimization. It is known as the correlation polytope or moment polytope. The correlation polytope is also affinely isomorphic, via an appropriate change of coordinates, to the cut polytope. See [DL97] for extensive background on correlation and cut polytopes. Cut polytopes and the connection to correlation polytopes are discussed in Section 19.1. The interpretation and the name correlation polytope come from the following proposition.

**Proposition 8.2.7.** Let $P \in \Delta_{2^m - 1}$ be a probability distribution for $m$ binary random variables $X_1, \ldots, X_m \in \{0, 1\}$. Let $G = ([m], E)$ be an undirected graph. Let $p = A_G P$, whose coordinates are indexed by the set $\{\emptyset\} \cup V \cup E$. Then $p_0 = 1$, $p_i = \mathbb{E}[X_i]$ and $p_{ij} = \mathbb{E}[X_i X_j]$. 
Proof. We calculate the entries of $p$. For $P \in \Delta_{2^m-1}$ and $\epsilon \in \{0,1\}^m$, let $P_\epsilon$ denote the probability $P_\epsilon = P(X_1 = \epsilon_1, \ldots, X_m = \epsilon_m)$. Then

$$p_0 = \sum_{\epsilon \in \{0,1\}^m} a_{0,\epsilon} P_\epsilon = \sum_{\epsilon \in \{0,1\}^m} 1 \cdot P_\epsilon = 1$$

$$p_i = \sum_{\epsilon \in \{0,1\}^m} a_{i,\epsilon} P_\epsilon = \sum_{\epsilon \in \{0,1\}^m} \epsilon_i P_\epsilon = \mathbb{E}[X_i]$$

$$p_{ij} = \sum_{\epsilon \in \{0,1\}^m} a_{ij,\epsilon} P_\epsilon = \sum_{\epsilon \in \{0,1\}^m} \epsilon_i \epsilon_j P_\epsilon = \mathbb{E}[X_iX_j].$$

To this end, inequality constraints on the cone of sufficient statistics $\text{cone}(A_G)$ can be used to check consistency of a given set of lower-order moments of a distribution. Indeed, a potential vector of moments $p \in \mathbb{R}^{1+2m+\#E}$ is actually the vector of moments of some probability distribution if and only if $p_0 = 1$ and $p \in \text{cone}(A_G)$. Inequality constraints can also be used to produce bounds on moments given some collection of other moments, as will be illustrated in Chapter 10. To do either of these things in practice requires explicit knowledge of the combinatorial structure of the facet-defining inequalities of $\text{cone}(A_G)$. It seems unlikely that there is a simple, compact description of the facet-defining inequalities of $\text{cone}(A_G)$ for all graphs $G$. However, there are some classes of graphs for which the description is known. From the moment description, we always have the following valid inequalities on $\text{cone}(A_G)$:

\begin{align}
(8.2.1) \quad p_{ij} &\geq 0, \quad p_i \geq p_{ij}, \quad p_j \geq p_{ij}, \quad p_0 + p_{ij} \geq p_i + p_j \text{ for all } ij \in E.
\end{align}

Moreover, those inequalities always define facets of $\text{cone}(A_G)$. Cycles in $G$ impose further facet defining inequalities on $\text{cone}(A_G)$.

Definition 8.2.8. A sequence of vertices $v_0, v_1, \ldots, v_k$ in a graph $G$ is a cycle if $v_0 = v_k$, there are no other repeated vertices in the sequence, and $v_iv_{i+1}$ is an edge in the $G$.

- Given a subset $K \subseteq V$ of the vertices of $G = (V,E)$, the induced subgraph $G_K$ is the graph with vertex set $K$ and edge set $E_K = \{ij : i, j \in K \text{ and } ij \in E\}$.
- A cycle in a graph $G$ is induced if the induced subgraph on the vertex set of the cycle consists of exactly the edges in the cycle.

Proposition 8.2.9. Let $G$ be a graph, $C$ be any induced cycle of $G$, and $O$ a subset of the edges of $C$ of odd cardinality. Then

$$\sum_{jk \in O} p_{jk} - \sum_{jk \in C \setminus O} p_{jk} - \sum_{j \in V(O)} p_j + \sum_{j \in V(C \setminus O)} p_j + \frac{\#O - 1}{2} p_0 \geq 0$$

is a valid inequality for the cone $\text{cone}(A_G)$ and is facet-defining.
The proof of Proposition 8.2.9 will be postponed until Chapter 19 where it will be more easy to state in the language of cut polytopes.

**Theorem 8.2.10.** [BM86] The cone of sufficient statistics \( \text{cone}(A_G) \) is defined by the inequalities from Equation 8.2.1 if and only if \( G \) is a forest.

The cone of sufficient statistics \( \text{cone}(A_G) \) is defined by the inequalities from Equation 8.2.1 plus the cycle inequalities of Proposition 8.2.9 from every induced cycle if and only if \( G \) is free of \( K_4 \) minors.

The problem of studying the integral points in \( \text{cone}(A_G) \) can be made into a more interesting/relevant/practical problem if we allow ourselves to change the matrix to one that has the same row span. Changing the matrix but leaving the row span the same does not change the resulting log-linear model, nor does it change the geometry of the cone of sufficient statistics, but the new coordinates that arise might change our interpretation of the sufficient statistics. To this end, for each such string \( \epsilon = (\epsilon_1, \ldots, \epsilon_m) \in \{0,1\}^m \) we form the vector

\[
c_\epsilon = (\epsilon_i \epsilon_j, (1-\epsilon_i) \epsilon_j, \epsilon_i (1-\epsilon_j), (1-\epsilon_i) (1-\epsilon_j))_{ij \in E} \in \mathbb{Z}^{4#E},
\]

and let \( C_G \) be the \( 4#E \times 2^m \) matrix consisting of the \( c_\epsilon \) as column vectors. Note that for a vector of counts \( u \), the vector \( C_G u \) consists of all of the marginal counts for pairs of random variables \( (X_i, X_j) \) such that \( ij \) is an edge in \( G \). The matrices \( A_G \) and \( C_G \) have the same integral row span and hence they are either both normal or both not normal. Characterizing the semigroup \( \mathbb{N}A_G \) can be used to determine which collections of integer marginals could have come from a \( 2^m \) integer contingency table. If the semigroup \( \mathbb{N}A_G \) is normal then one only needs to check that it lies in the lattice \( \mathbb{Z}A_G \) and lies in the cone \( \text{cone}(A_G) \).

The characterization of the normal binary graph models is the content of the following theorem.

**Theorem 8.2.11.** [Ohs10, Sul10] The semigroup \( \mathbb{N}A_G \) is saturated if and only if \( G \) is free of \( K_4 \) minors.

### 8.3. Gaussian Exponential Families

For exponential families besides the discrete exponential families, the cone of sufficient statistics is still convex but usually not polyhedral. To study such cones requires more general tools from convex geometry. For Gaussian exponential families the cones under consideration have natural interpretations as cones related to the cone of positive semidefinite matrices.

Let \( PD_m \) denote the cone of \( m \times m \) symmetric positive definite matrices, and \( PSD_m \) the cone of \( m \times m \) symmetric positive semidefinite matrices. For any two real symmetric matrices \( A, B \), let \( A \succeq B \) mean that \( A - B \in PSD_m \).
and $A \succ B$ that $A - B \in PD_m$. So $A \succeq 0$ means that $A$ is a positive semidefinite matrix.

Let $L \subseteq \mathbb{R}^{m(m+1)/2}$ be a linear space such that $L \cap PD_m$ is nonempty, which determines a Gaussian exponential family of covariance matrices via the set $L^{-1} = \{K^{-1} : K \in L \cap PD_m\}$. Given data $X^{(1)}, \ldots, X^{(n)}$, the sufficient statistics $T(X)$ of this model take the sample covariance matrix $S$ and map it via orthogonal projection onto $L$. The sample covariance matrix is, in principle, any positive semidefinite matrix and $T$ is a linear map, so $T(PD_m)$ is the cone of sufficient statistics. Following [SU10], we denote this set by $C_L$.

The cone of sufficient statistics for a Gaussian exponential family is usually not a polyhedral cone. However, since $PD_m$ is a semialgebraic set and the orthogonal projection map $T$ is a linear function, the image $C_L$ is also a semialgebraic set, though it need not be a closed basic semialgebraic set.

We illustrate the basic ideas with coordinate subspaces. Let $G = (V, E)$ be a graph. Let

$$L(G) = \{K \in \mathbb{R}^{m(m+1)/2} : K_{ij} = 0, \text{ if } i \neq j \text{ and } ij \notin E\}.$$ 

The resulting model $\mathcal{M}_G$ consisting of Gaussian random variables whose covariance matrices belong to the inverse linear space $L(G)^{-1}$ is called a concentration graph model, or Gaussian graphical model for an undirected graph. The sufficient statistic map is the map

$$T : \mathbb{R}^{m(m+1)/2} \to \mathbb{R}^V \oplus \mathbb{R}^E, \quad T(S) = (s_{ii})_{i \in V} \oplus (s_{ij})_{ij \in E}$$

that extracts the diagonal elements of $S$ and the off-diagonal elements corresponding to edges in the graph $G$. To describe the image of $T$ amounts to solving the following problem:

**Problem 8.3.1.** Given a graph $G$ and a partially observed symmetric matrix $T(S)$, determine whether or not there exists a $S' \in PD_m$ such that $T(S') = T(S)$.

Hence, giving a description of the cone of sufficient statistics of of a Gaussian exponential family is closely related to the positive semidefinite matrix completion problem (see e.g. [Lau98] for more background).

There are some obvious necessary conditions that a completable matrix must satisfy.

**Proposition 8.3.2.** Let $G = (V, E)$ be a graph and let $K \subseteq V$ be a set of vertices such that $G_K$ is a complete subgraph. Then if there exists an $S' \in PD_m$ such that $T(S') = T(S)$, it must be that $S'_{K,K}$ is a positive semidefinite matrix.
A subset of vertices $K \subseteq V$ such that the induced subgraph $G_K$ is a complete graph is called a \textit{clique} of $G$. The proof of Proposition 8.3.2 is left as an exercise. Proposition 8.3.2 rarely gives a necessary and sufficient condition. This holds if and only if $G$ satisfies strong conditions, which we describe now.

**Definition 8.3.3.** A graph $G$ is \textit{chordal} if on every cycle $v_0, v_1, \ldots, v_k = v_0$ in $G$ of length $k \geq 4$, there is a pair of vertices $v_i, v_j$ with $i - j \not\equiv -1, 0, 1 \mod k$ such that $v_iv_j \in E(G)$.

An equivalent definition of chordal graph is that every induced cycle is a three cycle. Chordal graphs play an important role in the theory of graphical models and hence in algebraic statistics. Their study originates in graph theory and they play a major role in the theory of graph minors. There are a number of different ways to describe these graphs besides the definition above. A second description is in terms of graph decompositions.

**Definition 8.3.4.** Say that a graph $G = (V, E)$ has a reducible decomposition into induced subgraphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ if

1. $G_i = G_{V_i}$, $i = 1, 2$
2. $G_{V_1 \cap V_2}$ is a complete subgraph of $G$, and
3. $G = G_1 \cup G_2$.

A graph with a reducible decomposition is \textit{reducible}. A graph is \textit{decomposable} if it is either a complete graph or it is reducible and both $G_1$ and $G_2$ are decomposable.

Reducible graphs are obtained from two graphs by gluing along a common clique. A decomposable graph can be obtained by a succession of gluing operations, starting with a collection of complete graphs. Chordal graphs are related to decomposable graphs by the following:

**Theorem 8.3.5.** [Dir61] A graph is chordal if and only if it is decomposable.

Figure 8.3.1 shows an example of a chordal graph and its decomposition into maximal cliques.

Say that a symmetric matrix $S$ satisfies the \textit{clique condition} with respect to the graph $G$ if all the determinants $|S_{K,K}|$ are nonnegative for $K$ a clique of $G$. Proposition 8.3.2 implies that if there exists a $S' \in PSD_m$ such that $T(S) = T(S')$ then $S$ must satisfy the clique condition with respect to $G$. This can be turned around for chordal graphs.

**Theorem 8.3.6.** [GJdSW84] If $G$ is a chordal graph then a symmetric matrix $S$ satisfies the clique condition with respect to $G$ if and only if there exists an $S' \in PSD_m$ such that $T(S) = T(S')$. 
Proof. Let $G$ be a chordal graph. By Theorem 8.3.5 and induction, we can assume that the graph $G$ consists of two cliques $A \cup B$ and $B \cup C$ meeting at a common subclique $B$, and that $A \cup B \cup C = [m]$. To start with, we will assume that $T(S)$ is a partially positive definite matrix (i.e., all cliques have positive determinant). Our partial matrix has the following form

$$T(S) = \begin{pmatrix} S_{A,A} & S_{A,B} & X \\ S_{B,A} & S_{B,B} & S_{B,C} \\ X^T & S_{C,B} & S_{C,C} \end{pmatrix}$$

where $X$ is the unknown matrix that we are trying to fill in. We claim we can take $X = S_{A,B}S_{B,B}^{-1}S_{B,C}$ to get a positive definite matrix. Indeed, let $S'$ be the resulting matrix with this fill-in. Let $U$ be the upper triangular matrix

$$U = \begin{pmatrix} I & 0 & 0 \\ 0 & -S_{A,B}S_{B,B}^{-1} & 0 \\ 0 & 0 & I \end{pmatrix}.$$

Then

$$US'U^T = \begin{pmatrix} S_{A,A} - S_{A,B}S_{B,B}^{-1}S_{B,A} & 0 & 0 \\ 0 & S_{B,B} & S_{B,C} \\ 0 & S_{C,B} & S_{C,C} \end{pmatrix}.$$ 

Note that $S_{A,A} - S_{A,B}S_{B,B}^{-1}S_{B,A}$ is positive definite since it is the Schur complement of the positive definite matrix $S_{A:B;A:B}$. This means that $US'U^T$ is positive definite, so $S'$ must have been positive definite.

Now suppose that $T(S)$ satisfies that clique property. Because the clique property is determined by nonnegativity of finitely many matrix inequalities, the set of partial matrices satisfying the clique property is a closed set. The interior consists of all the partial positive definite matrices. Let $T(S_1), T(S_2), \ldots$ be a sequence of partial positive definite matrices converging to $T(S)$, with $S_1, S_2, \ldots$ the positive definite matrices guaranteed to exist by the first part of our proof. Since $T(S)$ extracts the diagonal entries of $S$, this implies that the diagonal entries of all the $S_1, S_2, \ldots$ are bounded by some constant. By positive definiteness, this implies that all the entries of the $S_1, S_2, \ldots$ are bounded. Hence, by the Bolzano-Weierstrass Theorem,
there exists a subsequence $S_{i_1}, S_{i_2}, \ldots$, converging to some positive semidefinite matrix $S'$. This matrix satisfied $T(S') = T(S)$. □

On the other hand, as soon as $G$ contains an induced cycle without a chord, there are other conditions that must be satisfied to guarantee that there is a positive semidefinite matrix completion.

**Example 8.3.7** (Four Cycle). Let $G$ be the four cycle $C_4$, with edges $\{12, 23, 34, 14\}$, and consider the partial matrix

$$T(S) = A = \begin{pmatrix}
1 & a & x & -a \\
a & 1 & a & y \\
x & a & 1 & a \\
-a & y & a & 1
\end{pmatrix}.$$ 

Here $a \in [-1, 1]$ is a given parameter value and $x$ and $y$ are the unknown values. The problem is to determine if there are values of $x$ and $y$ that make the matrix positive semidefinite. We consider what happens as $a$ varies.

Note that every $1 \times 1$ and $2 \times 2$ principal submatrix of $A$ using observed values is positive semidefinite for any choice of $a \in [-1, 1]$. On the other hand, when we require that

$$|A_{\{1,2,3\},\{1,2,3\}}| = -(x + 1 - 2a^2)(x - 1) \geq 0$$

we deduce that $x \in [2a^2 - 1, 1]$. On the other hand the condition

$$|A_{\{1,3,4\},\{1,3,4\}}| = -(x - 1 + 2a^2)(x + 1) \geq 0$$

forces that $x \in [-1, 1 - 2a^2]$. So if $a > \frac{1}{\sqrt{2}}$ there are no values of $x$ and $y$ that complete the matrix to a positive semidefinite matrix.

For cycles of arbitrary length, there is a general structural result characterizing the partial matrices that can be completed to positive semidefinite matrices.

**Theorem 8.3.8.** [BJT93, Buh93] Let $T(S)$ be a partial matrix associated to the $m$-cycle graph $C_m$ with edges $\{12, 23, \ldots, m1\}$. This partial matrix can be completed to a positive definite matrix if and only if

1. All $s_{ii} > 0$,
2. All $s_{ii}s_{i+1,i+1} - s_{i,i+1}^2 > 0$, and
3. Let $x_i = \arccos\left( \frac{s_{i,i+1}}{\sqrt{s_{ii}s_{i+1,i+1}}} \right)$ then for all $i$, $x_i < \sum_{j \neq i} x_j$.

Here we consider the indices modulo $m$ so $m + 1 = 1$.

The conditions of Theorem 8.3.8 involve the $\arccos$ of functions of entries of partially observed matrix $T(S)$ and so does not appear to be algebraic. In fact, this condition can actually be rephrased as an algebraic
condition. Indeed, the equation that makes up part of the algebraic boundary of the cone of sufficient statistics then comes from making one of the inequalities an equality. In this case, it consists of a single hypersurface obtained from the equation

\[(8.3.1) \quad \frac{s_{1,2}}{\sqrt{s_{11} s_{22}}} = \cos \left( \sum_{i=2}^{n} \arccos \left( \frac{s_{i,i+1}}{\sqrt{s_{ii} s_{i+1,i+1}}} \right) \right)\]

which can be converted to an algebraic expression using trigonometric identities. In the case that \(m = 3\), this is just the \(3 \times 3\) determinant. The degree of the resulting equation is calculated in [SU10] for small values of \(m\).

Further work studying when these cycle conditions plus the conditions that come from cliques appeared in [BJL96], and the graphs where these conditions suffice to characterize the existence of a positive semidefinite completion have been completely characterized. The boundary of the cone of sufficient statistics was studied in depth for other graphs on small numbers of vertices in [SU10].

As in the discrete case, a second motivation for studying the cone of sufficient statistics is to determine when maximum likelihood estimates exist for given datasets. Here we collect some sample data \(X^{(1)}, \ldots, X^{(n)}\), compute a sample mean and sample covariance matrix, and ask when these lie on the boundary of the cone of sufficient statistics, or not. The sample mean does not affect this question (since the mean is constrained to a linear space in such models) and so we will ignore the mean in this discussion. To be precise, we assume that we consider a model of the form \(\{0\} \times \Theta\) where \(\Theta \subseteq PD_m\) is an inverse linear space. In such a centered Gaussian exponential family, the sample covariance matrix given the data is \(S = \sum_{i=1}^{n} X^{(i)}(X^{(i)})^T\) which is generically a positive semidefinite matrix of rank \(\min(m, n)\). We want to know when \(T(S)\) lies on the boundary of the cone of sufficient statistics. Since our data was generated by a random process on \(\mathbb{R}^m\) we can assume that it is generic among positive semidefinite matrices of rank \(\min(m, n)\). To lie in the interior of the cone of sufficient statistics is equivalent to the existence of a positive definite matrix \(\Sigma\) such that \(T(S) = T(\Sigma)\). Clearly if \(n \geq m\), then any generic \(S\) is of full rank \(m\), and positive definite, so we can just take \(\Sigma = S\). So the question of existence of maximum likelihood estimates becomes interesting when \(n < m\), that is, there are fewer data points \(n\) than coordinates \(m\) of our random variables. The best case would be that the maximum likelihood estimates exists for generic data
8.3. Gaussian Exponential Families

sets with a certain number \( n \) of data points. This leads to the following definition.

**Definition 8.3.9.** Let \( L \subset PD_m \) be a linear space with associated inverse linear space \( L^{-1} \). Let \( T_L \) be the associated projection onto the cone of sufficient statistics. The maximum likelihood threshold of \( L \), denoted \( \text{mlt}(L) \) is the smallest \( n \) such that for all generic positive semidefinite matrices \( S \in PSD_m \) of rank \( n \), there exists a positive definite matrix \( \Sigma \in PD_m \) such that \( T_L(S) = T_L(\Sigma) \).

The special case of coordinate subspaces (corresponding to the graphical Gaussian models mentioned above) has been the most studied, with partial results appearing in [Buh93, GS18, Uhl12]. In this case, the notation \( \text{mlt}(G) \) is used to represent \( \text{mlt}(L_G) \) where \( L_G = \{ \Sigma \in PD_m : \sigma_{ij} = 0 \text{ if } i \neq j \text{ and } ij \notin E(G) \} \). Some sample results from this theory:

**Theorem 8.3.10.** [Buh93] Suppose that \( G \) is a chordal graph. Then \( \text{mlt}(G) \) is the size of the largest clique in \( G \). More generally, if \( G \) is a graph and \( H \) is a chordal graph such that \( E(G) \subseteq E(H) \) then \( \text{mlt}(G) \leq \) the size of the largest clique in \( H \).

**Theorem 8.3.11.** [GS18] If \( G \) is a planar graph then \( \text{mlt}(G) \leq 4 \).

The maximum likelihood threshold is related to another invariant of a graph related to matrix completion, the generic completion rank of the linear subspace.

**Definition 8.3.12.** Let \( L \subset PD_m \) be a linear space with associated inverse linear space \( L^{-1} \). Let \( T_L \) be the associated projection onto the cone of sufficient statistics. The generic completion rank of \( L \), denoted \( \text{gcr}(L) \) is the smallest \( r \) such that for all generic positive definite matrices \( \Sigma \in PD_m \), there exists a (not necessarily positive semidefinite) matrix \( S \) of rank \( r \) such that \( T_L(S) = T_L(\Sigma) \).

Uhler showed the following result relating the maximum likelihood threshold to the generic completion rank, which can be used to make some useful statements about the maximum likelihood threshold.

**Theorem 8.3.13.** [Uhl12] Let \( L \subset PD_m \) be a linear space with associated inverse linear space \( L^{-1} \). Then \( \text{mlt}(L) \leq \text{gcr}(L) \).

**Proof.** An alternate way to define the maximum likelihood threshold is that is it is the smallest \( n \) such that all generic positive semidefinite matrices \( S \) of rank \( n \) has \( T_L(S) \) in the interior of the cone of sufficient statistics \( T_L(PSD_m) \). If \( L \) has \( \text{gcr}(L) = r \), then every generic matrix in the cone of sufficient statistics \( T_L(PSD_m) \) is the image of some matrix of rank \( r \).
In particular, this implies that the image of the positive semidefinite matrices of rank $r$ could not be confined to a subvariety, in particular, a generic positive semidefinite matrix of rank $r$ must lie in the interior of $T_L(PSD_m)$. Hence $\text{mlt}(L) \leq \text{gcr}(L)$. \hfill $\square$

The generic completion rank and the maximum likelihood threshold have received the most attention in cases where the linear space $L$ is a coordinate subspace, in which case we can associate the maximum likelihood threshold and generic completion rank to associated graphs $G$. Two recent papers studying these quantities are [GS18, BS17b]. In particular, [BS17b] gives the first examples of graphs where $\text{mlt}(G) \neq \text{gcr}(G)$ and shows that these two quantities can in fact be very far from each other.

8.4. Exercises

Exercise 8.1. Does there exist a probability distribution on three binary random variables such that $E[X_1] = E[X_2] = E[X_3] = 1/2$ and $E[X_1X_2] = E[X_1X_3] = E[X_2X_3] = 1/8$?

Exercise 8.2. (1) Compute the facet defining inequalities of the cone of sufficient statistics for the binary graph model associated to the graph $K_4$. Which inequalities do not come from (8.2.1) or Proposition 8.2.9?

(2) The matrix of the binary graph model $A_{K_4}$ is not normal. Compute the holes of the semigroup.

Exercise 8.3. Let $G$ be the graph with three vertices and three edges $1-2, 2-3, 1-3$. Determine all minimal patterns of zeroes in a $2 \times 2 \times 2$ table $\nu$ such that the maximum likelihood estimate does not exist for the binary graph model associated to $G$.

Exercise 8.4. Prove that for a graph $G$ the two matrices $A_G$ and $C_G$ associated to a binary graph model have the same row span.

Exercise 8.5. Prove Proposition 8.2.3.

Exercise 8.6. Prove Proposition 8.3.2.

Exercise 8.7. Find an explicit algebraic representation of the boundary equation (8.3.1) in the case that $m = 4$. What is the degree of the resulting polynomial?

Exercise 8.8. Let $L \subseteq \mathbb{R}^{10}$ be the linear space of concentration matrices of the form:

$$
\begin{pmatrix}
a & c & 0 & d \\
c & b & d & 0 \\
0 & d & a & c \\
d & 0 & c & b
\end{pmatrix}.
$$
(1) What are the sufficient statistics of the associated Gaussian exponential family?

(2) Determine the boundary of the cone of sufficient statistics $C_L$.

(3) What is the maximum likelihood threshold of $L$?

**Exercise 8.9.** Determine the maximum likelihood threshold and generic completion rank of cycles.
Chapter 9

Fisher’s Exact Test

Diaconis and Sturmfels’ paper [DS98], widely considered to be the first paper in algebraic statistics, established a connection between toric ideals and an important sampling problem related to contingency tables and discrete data. The sampling problem amounts to generating random lattice points from polytopes (with respect to an appropriate distribution), or said more statistically, sampling from the set of all contingency tables with a given set of sufficient statistics, with respect to the hypergeometric distribution. These samples is used in a Monte Carlo estimation of \( p \)-values in Fisher’s exact test.

In this chapter we explain the connection between Fisher’s exact test and toric ideals. Because of their connection to the sampling problem for Fisher’s exact test using Markov chain Monte Carlo algorithms, generating sets of toric ideals are often known as Markov bases. We also introduce the hierarchical log-linear models which are an important general class of models, and are related to graphical models that will be studied in more detail in Chapter 13. Beyond Markov bases, we will also discuss Graver bases as another sampling tool when bounds on cell entries are imposed, as well as some other strategies for sampling when Markov bases are too difficult to compute.

9.1. Conditional Inference

Consider a discrete exponential family \( \mathcal{M}_{A,h} \subseteq \Delta_{r-1} \) and suppose that we collect data \( X^{(1)}, \ldots, X^{(n)} \in [r] \) which are independent and identically distributed according to a distribution \( p \in \text{int}(\Delta_{r-1}) \). We would like to test
the hypothesis

\[ H_0 : p \in \mathcal{M}_{A,h} \quad \text{versus} \quad H_1 : p \notin \mathcal{M}_{A,h}. \]

Suppose that the null hypothesis is true: \( p \in \mathcal{M}_{A,h} \). So there is a parameter \( \theta \in \mathbb{R}_{\geq 0}^k \) such that \( p_j = Z(\theta)^{-1}h_j \theta^\alpha \) for \( j = 1, 2, \ldots, r \). Let \( u \in \mathbb{N}^r \) be the vector of counts or summary contingency table: that is, \( u_j = \#\{i : X^{(i)} = j\} \) for \( j \in [r] \). Then the probability of observing the table \( u \) given a parameter value \( \theta \) is

\[
L(u \mid \theta) = \binom{n}{u} h^u \theta^A Z(\theta)^{-n}
\]

where

\[
\binom{n}{u} = \frac{n!}{u_1! \cdots u_r!}
\]

is the multinomial coefficient. Let \( t : \mathbb{N}^r \rightarrow \mathbb{R} \) be a “nice” statistic which is zero if and only if \( u/n \in \mathcal{M}_{A,h} \) and increases away from \( \mathcal{M}_{A,h} \). Then we could calculate the p-value

\[
E_p[1_{t(v) > t(u)}(V)]
\]

where the expected value is taken over all \( v \in \mathbb{N}^r \) such that \( \|v\| = n \), the sample size, sampled with respect to the distribution \( p \). Here \( 1_{t(v) > t(u)}(v) \) is the indicator function of the event that \( t \) is greater than the value of the test statistics on the data point \( u \). The random variable \( V \) is a multinomial random variable sampled with respect to the probability distribution \( p \).

If the p-value is small, then \( u/n \) is far from a distribution that lies in the model \( \mathcal{M}_{A,h} \) (or, more precisely, the value of the statistic \( t(u) \) was unlikely to be so large if the true distribution belonged to the model). In this case, we reject the null hypothesis.

Of course, it is not possible to actually perform such a test as indicated since it depends on knowing the underlying distribution \( p \), which is unknown. If we knew \( p \), we could just test directly whether \( p \in \mathcal{M}_{A,h} \), or not. There are two ways out of this dilemma: asymptotics and conditional inference.

In the asymptotic perspective, we choose a test statistic \( t \) such that as the sample size \( n \) tends to infinity, the distribution of the test statistic converges to a distribution that does not depend on the true underlying distribution \( p \in \mathcal{M}_{A,h} \). The p-value computation can then be approximated based on this asymptotic assumption.

The usual choice of a test statistic which is made for log-linear models is Pearson’s \( X^2 \) statistic, which is defined as

\[
X_n^2(u) = \sum_{j=1}^r \frac{(u_j - \hat{u}_j)^2}{\hat{u}_j}
\]
Table 9.1.1. A contingency table relating the length of hospital stay to how often relatives visited of 132 schizophrenic hospital patients.

<table>
<thead>
<tr>
<th>Length of Stay</th>
<th>Visited regularly</th>
<th>Visited rarely</th>
<th>Visited never</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 ≤ Y &lt; 10</td>
<td>43</td>
<td>6</td>
<td>9</td>
<td>58</td>
</tr>
<tr>
<td>10 ≤ Y &lt; 20</td>
<td>16</td>
<td>11</td>
<td>18</td>
<td>45</td>
</tr>
<tr>
<td>20 ≤ Y</td>
<td>3</td>
<td>10</td>
<td>16</td>
<td>29</td>
</tr>
<tr>
<td>Totals</td>
<td>62</td>
<td>27</td>
<td>43</td>
<td>132</td>
</tr>
</tbody>
</table>

where \( \hat{u} = n\hat{p} \) is the maximum likelihood estimate of counts under the model \( \mathcal{M}_{A,h} \). Note that \( \hat{p} \) can be computed using iterative proportional scaling when there are no exact formulas for it, as described in Section 7.3. A key result about the limiting distribution of Pearson’s \( X^2 \) statistics is the following proposition, whose proof can be found, for example, in [Agr13, Section 16.3.3].

**Proposition 9.1.1.** If the joint distribution of \( X \) is given by a distribution \( p \in \mathcal{M}_{A,h} \), then

\[
\lim_{n \to \infty} P(X_2^2(\hat{V}) \geq t) = P(\chi_{df}^2 \geq t) \quad \text{for all} \quad t \geq 0,
\]

where \( df = r - 1 - \dim \mathcal{M}_{A,h} \) is the codimension or number of degrees of freedom of the model. In other words, for a true distribution lying in the model, the Pearson \( X^2 \) statistic converges in distribution to a chi-square distribution with \( df \) degrees of freedom.

The following example illustrates the application of the asymptotic \( p \)-value calculation.

**Example 9.1.2 (Schizophrenic Hospital Patients).** Consider the following data from [EGG89] which categorizes 132 schizophrenic hospital patients based on the length of hospital stay (in year \( Y \)) and the frequency \( X \) with which they were visited by relatives:

The most basic hypothesis one could test for this data is whether or not \( X \) and \( Y \) are independent. The maximum likelihood estimate for the data under the independence model is

\[

\begin{pmatrix}
27.2424 & 21.1364 & 13.6212 \\
11.8636 & 9.20455 & 5.93182 \\
18.8939 & 14.6591 & 9.44697
\end{pmatrix}

\]

The value of Pearson’s \( X^2 \)-statistic at the data table \( u \) is \( X^2(u) = 32.3913 \). The model of independence \( \mathcal{M}_{X \perp \perp Y} \) is four dimensional inside the probability simplex \( \Delta_8 \), so this model has \( df = 8 - 4 = 4 \). The probability that a \( \chi_4^2 \) random variable is greater than \( X^2(u) = 35.17109 \) is approximately
4.2842 \times 10^{-7}$. This asymptotic $p$-value is very small, in which case the asymptotic test tells us we should reject the null hypothesis of independence. The following code performs the asymptotic test in R.

```r
tab <- as.table(matrix(c(43, 16, 3, 6, 11, 10, 9, 18, 16), nrow = 3, byrow = TRUE))
dimnames(tab) <- list(X = c("X1","X2","X3"), Y = c("Y1","Y2","Y3"))
MASS::loglm(~ X + Y, data = tab) # asymptotic test
```

A similar strategy is based on the likelihood ratio test, where we use the $G$-statistic, instead of the $X^2$-statistic. The $G$-statistic is defined by:

$$G_n(u) = 2 \sum_{j=1}^r u_j (\log u_j - \log \hat{u}_j).$$

As with the $X^2$-statistic, the $G$ statistic converges to a chi-square distribution with $df = r - 1 - \dim \mathcal{M}_{A,h}$ degrees of freedom as $n$ tends to infinity, which follows from results in Section 7.4 since $\mathcal{M}_{A,h}$ is smooth everywhere.

Appealing to an asymptotic distribution is natural when viewing a statistical problem through the lens of probability theory. On the other hand, in many cases the sample size is small, and in particular it is not reasonable to consider that the sample size is “tending to infinity”. The typical rule of thumb that has been used in these problems if “most” entries of $u$ are greater than 5, then the asymptotics are reasonably accurate. However, since many large contingency tables are sparse, and data can be rare and expensive to obtain, this suggests that other strategies besides the asymptotic tests might be appropriate. Thus we are posed the question: Is there any other way to eliminate the dependence of the non-asymptotic $p$-value on the unknown parameter values $\theta$? This leads to the framework of conditional inference.

**Definition 9.1.3.** Let $A \in \mathbb{Z}^{k \times r}$ and let $u \in \mathbb{N}^r$. The fiber of the table $u$ with respect to the the sufficient statistic $A$ is the set

$$\mathcal{F}(u) = \{ v \in \mathbb{N}^r : Av = Au \}$$

consisting of all other vectors $v \in \mathbb{N}^r$ with the same sufficient statistics as $u$.

The conditional inference framework is based on the following observation. The conditional likelihood function $L(v \mid v \in \mathcal{F}(u), \theta)$ does not depend on the unknown parameter $\theta$. This unknown parameter is often
9.1. Conditional Inference

called a nuisance parameter. Indeed, this likelihood function is

\[ L(v \mid v \in \mathcal{F}(u), \theta) = \frac{\binom{n}{v} h^v \theta^v Z(\theta)^{-n}}{\sum_{v \in \mathcal{F}(u)} \binom{n}{v} h^v \theta^v Z(\theta)^{-n}} \]

and since \( Av = Au \) for all \( v \in \mathcal{F}(u) \), all the terms involving \( \theta \) drop out and we are left with

\[ L(v \mid v \in \mathcal{F}(u), \theta) = L(u \mid v \in \mathcal{F}(u)) = \frac{\binom{n}{v} h^v}{\sum_{v \in \mathcal{F}(u)} \binom{n}{v} h^v}. \]

The resulting distribution on the fiber \( \mathcal{F}(u) \) where \( P(v) \propto \binom{n}{v} h^v \) is called the generalized hypergeometric distribution. Thus, conditional inference provides us with a means for computing a \( p \)-value that does not depend on the unknown parameter \( \theta \). However, we are left with computing or approximating the sum

\[ \frac{1}{\# \mathcal{F}(u)} \sum_{v \in \mathcal{F}(u)} 1_{l(v) \geq l(u)} L(u \mid v \in \mathcal{F}(u)). \]  

(9.1.1)

**Remark 9.1.4.** When considering exact tests, we are no longer constrained by the motivation to choose a test statistic whose asymptotic distribution we can find explicitly. One common choice for the exact test besides Pearson’s \( X^2 \)-statistic and the \( G \)-statistic, another choice of test statistic is \( -p(v) \). Since the hypergeometric probability \( p(v) \) is largest on the fiber precisely at the unique data point that lies in the model (if it exists), the function \( -p(v) \) also has the desirable properties that make it a good test statistic. With this choice, the \( p \)-value is equal to the probability that the given table \( u \) was more likely to occur than a random table in the fiber.

**Example 9.1.5 (Schizophrenic Hospital Patients).** We now apply Fisher’s exact test to the data from Example 9.1.2. In this case, the fiber consists of 258909 \( 3 \times 3 \) tables. For each of these tables, we determine if the \( X^2 \) statistic is greater than that of the data table, and sum the probabilities of all such tables. The resulting \( p \)-value is 3.18356 \( \times 10^{-7} \), which is on the same order of magnitude as the \( p \)-value for the asymptotic test. Both tests suggest that we should reject the null hypothesis of independence. Because the size of the fiber is relatively small, the Fisher’s exact test can be run directly in \( R \) using the following line of code:

```r
fisher.test(dat)
```

In the previous example, all the cell entries are greater than 5, so the asymptotics and Fisher’s exact test give similar results. Fisher’s exact test is mainly of interest in the situation where the table is sparse, with small cell entries. Even in this case, the fibers \( \mathcal{F}(u) \) can be enormous and impossible to enumerate. Furthermore, it is unlikely that there exists any analytic
expression for (9.1.1). So the usual strategy is to approximate the exact test $p$-value by randomly generating samples from $\mathcal{F}(u)$ with respect to the generalized hypergeometric distribution. This suggests the following problem, which will occupy us for the majority of this chapter.

**Problem 9.1.6.** Generate random samples from the fiber $\mathcal{F}(u)$ with respect to the generalized hypergeometric distribution $P(v) \propto \binom{n}{v} h^v$.

### 9.2. Markov Bases

As discussed in the previous section, we can usually not enumerate the entire fiber $\mathcal{F}(u)$ to perform Fisher’s exact test. Instead we must generate random samples from the fiber $\mathcal{F}(u)$ with respect to the generalized hypergeometric distribution, to get a Monte Carlo estimate of the $p$-value. Markov bases provide a tool for generating such random samples, which can be applied to any distribution on $\mathcal{F}(u)$, and we focus on that general version of the problem. The problem of generating random tables from fibers with respect to other distributions besides the generalized hypergeometric distribution has other applications in statistics. For example, the Diaconis-Efron test [DE85] requires samples from the uniform distribution on the fiber $\mathcal{F}(u)$. In Bayesian statistics as part of a larger posterior distribution calculation, we might be required to generate random samples from essentially arbitrary distributions on $\mathcal{F}(u)$.

**Definition 9.2.1.** Let $A \in \mathbb{Z}^{k \times r}$. Let $\ker \mathbb{Z} A = \{ v \in \mathbb{Z}^r : Av = 0 \}$ be the integer kernel of $A$. A finite set $\mathcal{B} \subset \ker \mathbb{Z} A$ is called a Markov basis for $A$ if for all $u \in \mathbb{N}^n$ and $u' \in \mathcal{F}(u)$, there is a sequence $v_1, \ldots, v_L \in \pm \mathcal{B}$ such that

$$u' = u + \sum_{k=1}^L v_k \quad \text{and} \quad u + \sum_{k=1}^l v_k \geq 0 \quad \text{for all } l = 1, \ldots, L.$$ 

The elements of the Markov basis are called *moves*.

An alternate description of a Markov basis is via a graph on the fiber. To the fiber $\mathcal{F}(u)$ and the set of moves $\mathcal{B} \subseteq \ker \mathbb{Z} A$ we make a graph $\mathcal{F}(u)_\mathcal{B}$ as follows. The vertices of $\mathcal{F}(u)_\mathcal{B}$ are the elements of $\mathcal{F}(u)$. Two elements $v, v' \in \mathcal{F}(u)$ are connected by an edge if $v - v' \in \mathcal{B}$ or $v' - v \in \mathcal{B}$. To say that $\mathcal{B}$ is a Markov basis for $A$ is equivalent to saying that the graph $\mathcal{F}(u)_\mathcal{B}$ is connected for all $u$.

We will discuss the existence and computation of Markov bases later in this section, and in Section 9.3 the construction of Markov bases in the important case of hierarchical models. The importance of Markov bases is that they can be used to generate random samples from a distribution on $\mathcal{F}(u)$ using the Metropolis-Hastings algorithm.
Algorithm 9.2.2 (Metropolis-Hastings).
Input: A contingency table $u \in \mathbb{N}^r$, a Markov basis $B$ for $A$, and a probability distribution $p$ on $\mathcal{F}(u)$ which gives nonzero probability to every $u' \in \mathcal{F}(u)$.
Output: A sequence of tables $u_1, u_2, \ldots \in \mathcal{F}(u)$.
Step 1: Initialize $u_1 = u$.
Step 2: For $t = 1, 2, \ldots$ repeat the following steps:
   (i) Select uniformly at random a move $v_t \in \pm B$.
   (ii) If any coordinate of $u_t + v_t$ is negative, then set $u_{t+1} = u_t$, else set
        $u_{t+1} = \begin{cases} u_t + v_t & \text{with probability } q \\ u_t & \text{with probability } 1 - q \end{cases}$,
        where
        \[ q = \min\left\{ 1, \frac{p(u_t + v_t)}{p(u_t)} \right\}. \]
Step 3: Output the sequence $u_1, u_2, \ldots, u_t, \ldots$.

The sequence of tables that is produced by the Metropolis-Hastings algorithm has the desired property of eventually converging to a random sample from the desired distribution $p$.

Theorem 9.2.3. The output sequence of tables $u_t$ from Algorithm 9.2.2 is an aperiodic, reversible, and irreducible Markov chain that has stationary distribution equal to the distribution $p$ on $\mathcal{F}(u)$.

A proof of this theorem can be found in [DS98, Lemma 2.1] or [RC99, Chapter 6]. The consequence of Theorem 9.2.3 for Monte Carlo algorithms is that if we let the Markov chain in the Metropolis-Hastings algorithm run long enough, then we can draw samples from the distribution $p$ on $\mathcal{F}(u)$, and hence use those samples to approximate the $p$-value from (9.1.1). It remains a major unsolved research problem to determine bounds on the time until a sample that is drawn is close to the desired distribution. This time is called the mixing time of the Markov chain (see [LPW09] for an in-depth treatment of mixing times of general Markov chains). Even in the case of sampling from the uniform distribution on the fiber $\mathcal{F}(u)$, the mixing time remains unknown (but see [Win16] for related results). In practice it is possible to increase the connectivity of the Markov chain by using integer multiples or linear combinations of the moves $v \in B$, but the consequences for the mixing time seem even more difficult to analyze.

One important feature of the Metropolis-Hastings algorithm is that it is not necessary to be able to compute exactly the probability of the particular table $p(u)$, one need only know this probability distribution up to the normalization constant. Indeed, the only place in the Metropolis-Hastings
algorithm where the probability distribution $p$ appears in the calculation of $q$, and the normalizing constant drops out of the ratio $p(u_t + v_t)/p(u_t)$. This makes the application of the algorithm for the generalized hypergeometric distribution feasible, since the nonnormalized version $p(u) \propto \binom{u}{n} k^u$ is tractable. Similarly for the uniform distribution, the normalizing constant is the cardinality of the fiber $F(u)$, but the Metropolis-Hastings algorithm does not require us to know this difficult-to-obtain quantity.

Example 9.2.4 (Metropolis-Hastings for the Uniform Distribution). Note that for the uniform distribution, the quantity $p(u_t + v_t)/p(u_t)$ is always equal to one. Hence, provided that $\min(u_t + v_t) \geq 0$, we always move to $u + v_t$.

Markov bases are fundamentally connected to computational algebra via toric ideals. Recall, that the toric ideal $I_A$ associated to the integer matrix $A$, is the vanishing ideal of the monomial parametrization $\phi^A$. By Proposition 6.2.4,

$$I_A = \langle p^u - p^{u'} : u, u' \in \mathbb{N}^r \text{ and } A u = A u' \rangle$$

so the toric ideal $I_A$ gives an encoding of the integer lattice $\ker \mathbb{Z} A$. In particular, if we are given a vector $v \in \ker \mathbb{Z} A$, we can write $v = v^+ - v^-$ where the coordinates of $v^+$ and $v^-$ are given by $v^+_j = \max(v_j, 0)$ and $v^-_j = -\max(-v_j, 0)$. To the vector $v \in \ker \mathbb{Z}$, we associate the binomial $p^v^+ - p^v^- \in I_A$. Conversely, if $p^u - p^{u'} \in I_A$, then $u - u' \in \ker \mathbb{Z} A$. The following theorem of Diaconis and Sturmfels [DS98] connects the generating sets of the toric ideal $I_A$ to Markov bases.

**Theorem 9.2.5 (Fundamental Theorem of Markov Bases).** A finite subset $B \subseteq \ker \mathbb{Z} A$ is a Markov basis for $A$ if and only if the corresponding set of binomials $\{p^{v^+} - p^{v^-} : v \in B\}$ is a generating set for the toric ideal $I_A$.

**Proof.** ($\Rightarrow$) Let $B$ be a Markov basis for $A$. We will show that the set of binomials $B = \{p^{v^+} - p^{v^-} : v \in B\}$ is a generating set of $I_A$. According to Proposition 6.2.4, it suffices to show that any binomial $p^{u} - p^{u'} \in I_A$ is in the ideal generated by the binomials in $B$. Since $A u = A u'$ and $B$ is a Markov basis, there is a sequence of moves $v_1, \ldots, v_L \in B$ such that

$$u' = u + \sum_{k=1}^{L} v_k \quad \text{and} \quad u + \sum_{k=1}^{L} v_k \geq 0 \text{ for all } l = 1, \ldots, L.$$ 

This implies that

$$p^{u'} = p^u + \sum_{k=1}^{L} p^{u+v_1+\cdots+v_{k-1}-v_k} (p^{v^+_k} - p^{v^-_k})$$
and each of the exponent vectors

\[ u + v_1 + \cdots + v_{k-1} - v_k^- \]

is nonnegative. Thus, \( p^u - p^{u'} \) is in the ideal generated by \( B \).

(\( \Leftarrow \)) Conversely, suppose that \( B = \{ p^{v^+} - p^{v^-} : v \in B \} \) is a binomial generating set for the ideal \( I_A \). Then any \( p^u - p^{u'} \in I_A \) has an expression as

\[ p^u - p^{u'} = \sum_{k=1}^{L} p^{m_k}(p^{v^+_k} - p^{v^-_k}). \]

This follows because algebraic containment of a pure difference binomial in an ideal generated by pure difference binomials can be decided by looking at the containment problem over the prime subfield (i.e. either \( \mathbb{Z}_p \) or \( \mathbb{Q} \)). In the case of \( \mathbb{Q} \) coefficients we can clear denominators so we get an expression only involving integers. This sum can be arranged in such a way that \( p^{m_k} p^{v^+_k} = p^u \) and each successive pair of terms \( p^{m_k} p^{v^-_k} = p^{m_{k+1}} p^{v^+_k} \). This forces \( p^{m_k} p^{v^-_k} = p^{u'} \), and implies that

\[ u = u' + \sum_{k=1}^{L} v_k \quad \text{and} \quad u' + \sum_{k=1}^{l} v_k \geq 0 \text{ for all } l = 1, \ldots, L. \]

Thus, \( B \) is a Markov basis for \( A \).

Because of the Fundamental Theorem of Markov Bases, a binomial generating set of a toric ideal is often called a Markov basis. One immediate consequence of Theorem 9.2.5 is that Markov bases exist; that is, for any \( A \), there exists a finite set of moves \( B \subset \ker \mathbb{Z} A \) which connects all the fibers \( F(u) \). This follows from the Fundamental Theorem of Markov Bases plus the Hilbert Basis Theorem, since the ideal \( I_A \) must have a finite generating set consisting of binomials. A second consequence of Theorem 9.2.5 is a method for computing Markov bases: since \( I_A \) is the homogeneous vanishing ideal of the discrete exponential family \( M_A \), we can compute a binomial generating set using implicitization via Gröbner basis calculation. In fact, the special combinatorial nature of toric ideals yields sophisticated lattice-based algorithms for computing generating sets and Gröbner bases of toric ideals. These algorithms are implemented in the software 4ti2 [HH03], 4ti2 can be run as a stand-alone program, as part of the lattice point package LattE [BBDL+15], or run from inside Macaulay2 as the following example illustrates.
Example 9.2.6 (Markov Basis Computation for the Independence Model).
Consider the matrix
\[
A = \begin{pmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1
\end{pmatrix}
\]

Letting \( u = (u_{11}, u_{12}, u_{13}, u_{21}, u_{22}, u_{23}, u_{31}, u_{32}, u_{33})^T \), we see that \( Au = (u_{1+}, u_{2+}, u_{3+}, u_{+1}, u_{+2}, u_{+3})^T \) is the vector of sufficient statistics of the independence model for two discrete random variables with 3 states. We compute the Markov basis in Macaulay2 using the following code:

```plaintext
loadPackage "FourTiTwo"
A = matrix{
  {1,1,0,0,0,0,0,0,0},
  {0,0,1,1,0,0,0,0,0},
  {0,0,0,0,0,0,1,1,1},
  {1,0,0,1,0,0,1,0,0},
  {0,1,0,0,1,0,0,1,0},
  {0,0,1,0,0,1,0,0,1}};
toricMarkov A
```

which produces the output:

\[
\begin{pmatrix}
0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 \\
0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & -1 & -1 & 0 & 1 \\
0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 \\
0 & 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 \\
1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\
1 & 0 & -1 & -1 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1
\end{pmatrix}
\]

\( oo4 : \text{Matrix} \ ZZ \leftarrow ZZ \)

The rows of this output matrix are the elements of the Markov basis. Written as \( 3 \times 3 \) contingency tables, these are:

\[
\begin{pmatrix}
0 & 0 & 0 \\
0 & 1 & -1 \\
0 & -1 & 1
\end{pmatrix}, \quad
\begin{pmatrix}
0 & 0 & 0 \\
1 & -1 & 0 \\
-1 & 1 & 0
\end{pmatrix}, \ldots, \quad
\begin{pmatrix}
1 & 0 & -1 \\
0 & 0 & 0 \\
-1 & 0 & 1
\end{pmatrix}.
\]
According to Theorem 9.2.5, these tables convert to binomials which are a generating set for $I_A$:

$$P_{22}P_{33} - P_{23}P_{32}, \quad P_{21}P_{33} - P_{22}P_{31}, \quad \ldots, \quad P_{11}P_{33} - P_{13}P_{31}.$$ 

Another consequence of Theorem 9.2.5 is that it gives us a useful combinatorial procedure to construct a generating set of a toric ideal. That is, to prove that a given set of binomials is a generating set of a toric ideal $I_A$ we can show that every fiber $F(u)$ is connected by the corresponding set of moves.

**Theorem 9.2.7** (Markov Basis of the Independence Model). Let $e_{11}, \ldots, e_{r1}$ and $e'_{11}, \ldots, e'_{r2}$ denote the standard unit vectors in $\mathbb{R}^{r1}$ and $\mathbb{R}^{r2}$, respectively. Let $A$ be the matrix whose columns are the $r1 \times r2$ vectors $e_{i1} \otimes e'_{i2}$ with $i1 \in [r1]$ and $i2 \in [r2]$. So $A : \mathbb{Z}^{r1 \times r2} \to \mathbb{Z}^{r1} \otimes \mathbb{Z}^{r2}$ represents the linear transformation that computes the row and column sums of and $r1 \times r2$ tables. Let

$$B_{r1r2} := \{ e_{i1i2} + e_{j1j2} - e_{i1j2} - e_{j1i2} : i1, j1 \in [r1], i2, j2 \in [r2] \}$$

where $e_{11}, \ldots, e_{r1r2}$ denote the standard unit vectors in $\mathbb{R}^{r1 \times r2}$. Then $B_{r1r2}$ is a Markov basis for $A$.

**Proof.** Let $u \neq v \in \mathbb{N}^{r1 \times r2}$ be two non-negative integral tables that have the same row and column sums. It suffices to show that there is an element $b \in B_{r1r2}$, such that $u + b \geq 0$ and $\|u - v\|_1 > \|u + b - v\|_1$, because this implies that we can use elements of $B_{r1r2}$ to bring points in the same fiber closer to one another. Since $u$ and $v$ are not equal and $Au = Av$, there is at least one positive entry in $u - v$. Without loss of generality, we may suppose $u_{11} - v_{11} > 0$. Since $u - v \in \ker A$, there is an entry in the first row of $u - v$ that is negative, say $u_{12} - v_{12} < 0$. By a similar argument $u_{22} - v_{22} > 0$. But this implies that we can take $b = e_{12} + e_{21} - e_{11} - e_{22}$ which attains $\|u - v\|_1 > \|u + b - v\|_1$ and $u + b \geq 0$ as desired. \(\Box\)

Note that the Markov basis in Theorem 9.2.7 is a minimal Markov basis, that is, no move can be omitted to obtain a set that satisfies the Markov basis property. Indeed, if, for example, the move $e_{11} + e_{22} - e_{12} - e_{21}$ is omitted, then the fiber

$$F(e_{11} + e_{22}) = \{ e_{11} + e_{22}, e_{12} + e_{21} \}$$

cannot be connected by any of the remaining moves of $B_{r1r2}$.

Theorem 9.2.7 on Markov bases of two-way contingency tables allows us to fill in a missing detail from our discussions on conditional independence from Chapter 4.
Table 9.2.1. Birth and Death Months of 82 Descendants of Queen Victoria

<table>
<thead>
<tr>
<th>Month of Birth</th>
<th>J</th>
<th>F</th>
<th>M</th>
<th>A</th>
<th>M</th>
<th>J</th>
<th>J</th>
<th>A</th>
<th>S</th>
<th>O</th>
<th>N</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>February</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>March</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>April</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>May</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>June</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>July</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>August</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>September</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>October</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>November</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>December</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Corollary 9.2.8. The homogeneous vanishing ideal of the independence model on two discrete random variables is the prime toric ideal

\[ I_{X \perp Y} = \langle p_{i_1 i_2} p_{j_1 j_2} : i_1, j_1 \in [r_1], i_2, j_2 \in [r_2] \rangle. \]

Proof. Discrete random variables \( X \) and \( Y \) are independent if and only if \( P(X = i, Y = j) = P(X = i)P(Y = j) \) for all \( i \) and \( j \). Treating the marginal probability distributions as parameters we can express this as

\[ P(X = i, Y = j) = \frac{1}{Z} \alpha_i \beta_j \]

where \( Z \) is a normalizing constant. Hence, the model of independence is a discrete exponential family. The sufficient statistics of that family are the row and column sums of a table \( u \in \mathbb{N}^{r_1 \times r_2} \). Since the coefficient vector \( h = 1 \), the vector of ones, we see that the homogeneous vanishing ideal of \( M_{X \perp Y} \) is the toric ideal associated to the matrix \( A \) from Theorem 9.2.7. Each of the moves in the Markov basis \( B_{r_1 r_2} \)

\[ e_{i_1 i_2} + e_{j_1 j_2} - e_{i_1 j_2} - e_{j_1 i_2} \]

translates into a binomial

\[ p_{i_1 i_2} p_{j_1 j_2} - p_{i_1 j_2} p_{j_1 i_2}. \]

The fundamental theorem of Markov bases says that these binomials generate the toric ideal. \( \square \)

Example 9.2.9 (Birthday/Deathday). Consider Table 9.2.1 which records the birth month and death month of 82 descendants of Queen Victoria by 1977, excluding a few deaths by nonnatural causes [AH85, p. 429]. This
dataset can be used to test the independence of birthday and deathday. Note that this table is extremely sparse and all cell entries are small, and hence we should be reluctant to use results based on asymptotics for analyzing the data.

This dataset can be analyzed using Markov bases via the \texttt{algstat} package in R \cite{KGPY17a}. (Related R packages are \texttt{mpoly} \cite{Kah13} for working with multivariate polynomials in R, \texttt{latter} \cite{KGPY17b} which uses the software \texttt{LattE} from R, and \texttt{m2r} \cite{KOS17} which allows users to access Macaulay2 from R.) To perform Fisher's exact test using the \texttt{algstat} package requires having \texttt{LattE} \cite{BBDL15} installed on the computer and then properly linked to R. Once that is done, the following sequence of commands will perform the exact tests, with respect to various test statistics.

\begin{verbatim}
if(!requireNamespace("algstat")) install.packages("algstat")
library("algstat"); data("birthdeath")
loglinear(~ Birth + Death, data = birthdeath,
iter = 1e5, thin = 500)
\end{verbatim}

This series of commands generates 10000 samples using the random walk based on a Markov basis. The resulting $p$-value for the exact test with Pearson’s $X^2$ test statistic is 0.6706, suggesting that we cannot reject the null hypothesis of independence between birth month and death month. The $p$-value obtained from the asymptotic approximation is 0.6225, which shows that the asymptotic approximation is not very good at this sample size. Figure 9.2.1 shows a histogram of the $X^2$ values of the 10000 samples, in comparison with the chi-squared distribution with $df = (12 - 1)(12 - 1)$ which the asymptotic approximation uses. The vertical line denotes the value of $X^2(u)$ for the table $u$. Note the poor disagreement between the true distribution and the approximation.

As a second example of a Markov basis, we consider the model of Hardy-Weinberg equilibrium, which represents a well-known use of the exact test and Monte Carlo computations from the paper of Guo and Thompson \cite{GT92}. The Hardy-Weinberg model is a model for a discrete random variable $X$ whose states are unordered pairs $ij$ with $i, j \in [r]$. The probability distribution of $X$ is given by

$$P(X = ij) = \begin{cases} 2\theta_i\theta_j & i \neq j \\ \theta_i^2 & i = j. \end{cases}$$

The Hardy-Weinberg distribution is the result of taking two independent draws from a multinomial distribution with parameter $\theta = (\theta_1, \ldots, \theta_r)$. This special instance of a multinomial model has a special name because of its applications in biology. Suppose that a gene has $r$ alleles (i.e. different versions). In a diploid species, each organism has two copies of each
chromosome and hence two copies of each gene. The genotype of each organism is, hence, an unordered pair of elements from \([r]\). If the population is well-mixed and there is no positive or negative selection against any combinations of alleles, the population frequencies will be in Hardy-Weinberg equilibrium, since the probability of a genotype will just be the product of the probabilities of each allele times a multinomial coefficient. Hardy-Weinberg equilibrium was postulated in [Har08]. To test whether a given data set is consistent with the model of Hardy-Weinberg equilibrium, we need to perform an exact test.

In particular, consider the following data set from Guo and Thompson [GT92] of genotype frequencies in Rhesus data, which we display as an upper triangular matrix:

\[
\begin{pmatrix}
1236 & 120 & 18 & 982 & 32 & 2582 & 6 & 2 & 115 \\
3 & 0 & 55 & 1 & 132 & 0 & 0 & 5 \\
0 & 7 & 0 & 20 & 0 & 0 & 2 \\
249 & 12 & 1162 & 4 & 0 & 53 \\
0 & 29 & 0 & 0 & 1 \\
1312 & 4 & 0 & 149 \\
0 & 0 & 0 \\
0 & 0 \\
4
\end{pmatrix}
\]

(9.2.1)

In spite of the large sample size in this example, asymptotic approximations appear unwise since there are so many small entries in the table, and it is also relatively sparse.
In general, data for the model of Hardy-Weinberg equilibrium is an upper triangular data table \( u = (u_{ij}) \) such that \( 1 \leq i \leq j \leq r \). The sufficient statistics for this model are the sums for each \( i \)
\[
\sum_{j=1}^{i} u_{ji} + \sum_{j=i}^{r} u_{ij}
\]
which gives the number of times that the allele \( i \) appears in the population. Note that we count the diagonal entry \( u_{ii} \) twice. Writing this out in matrix form for \( r = 4 \) yields the matrix:
\[
\begin{pmatrix}
2 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 2 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 2 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 2 \\
\end{pmatrix}
\]

**Theorem 9.2.10.** The Markov basis for the model of Hardy-Weinberg equilibrium consists of the following family of moves
\[
B_r = \{ e_{ij} + e_{kl} - e_{ik} - e_{jl} : i, j, k, l \in [r] \}
\]
In the theorem, we use the convention that \( e_{ij} = e_{ji} \). Note that this set of moves includes moves like \( e_{ii} + e_{jj} - 2e_{ij} \) so that this is an example of a Markov basis that does not consist only of moves whose coefficients are \( 1, -1, 0 \). Applying the fundamental theorem of Markov bases, we can also state Theorem 9.2.10 in the language of toric ideals. That is, translating the moves in \( B_r \) into binomials, we also have the following equivalent representation.

**Corollary 9.2.11.** The homogeneous vanishing ideal of the model of Hardy-Weinberg equilibrium is the ideal generated by \( 2 \times 2 \) minors of the symmetric matrix:
\[
\begin{pmatrix}
2p_{11} & p_{12} & p_{13} & \cdots & p_{1r} \\
p_{12} & 2p_{22} & p_{23} & \cdots & p_{2r} \\
p_{13} & p_{23} & 2p_{33} & \cdots & p_{3r} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
p_{1r} & p_{2r} & p_{3r} & \cdots & 2p_{rr}
\end{pmatrix}
\]

The proofs of Theorem 9.2.10 and Corollary 9.2.11 are similar to those of Theorem 9.2.7 and Corollary 9.2.8 are left as exercises.

### 9.3. Markov Bases for Hierarchical Models

Among the most important discrete exponential families are the hierarchical log-linear models. These statistical models are for collections of random variables. In a hierarchical model a simplicial complex is used to encode
interaction factors among the random variables. The sufficient statistics of a hierarchical model are also easily interpretable as lower-dimensional margins of a contingency table. Hierarchical models are, thus, a broad generalization of the independence model whose Markov basis we have seen in Theorem 9.2.7. However, it is usually not straightforward to describe their Markov bases and significant effort has been put towards developing theoretical tools for describing Markov bases for these models.

**Definition 9.3.1.** For a set $S$, let $2^S$ denote its power set, that is the set of all of its subsets. A simplicial complex with ground set $S$ is a set $\Gamma \subseteq 2^S$ such that if $F \in \Gamma$ and $F' \subseteq F$ then $F' \in \Gamma$. The elements of $\Gamma$ are called the faces of $\Gamma$ and the inclusion maximal faces are the facets of $\Gamma$.

Note Definition 9.3.1 is more properly called an abstract simplicial complex. This abstract simplicial complex has a geometric realization where a face $F \subseteq S$ with $i+1$ elements yields an $i$-dimensional simplex in the geometric realization. To describe a simplicial complex we need only list its facets.

Hierarchical models are log-linear models, so they can be described as $\mathcal{M}_A$ for a suitable matrix $A$ associated to the simplicial complex $\Gamma$. However, it is usually easiest to describe this model first by giving their monomial parametrization, and then recovering the associated matrix. Both descriptions are heavy on notation.

Hierarchical models are used for describing interactions between collections of random variables. We have $m$ discrete random variables $X_1, \ldots, X_m$. Random variable $X_i$ has state space $[r_i]$ for some integer $r_i$. We denote the joint state space of the random vector $X = (X_1, \ldots, X_m)$ by $\mathcal{R} = \prod_{i=1}^{m} [r_i]$.

We use the following convention for writing subindices. If $i = (i_1, \ldots, i_m) \in \mathcal{R}$ and $F = \{f_1, f_2, \ldots, \} \subseteq [m]$ then $i_F = (i_{f_1}, i_{f_2}, \ldots)$. For each subset $F \subseteq [m]$, the random vector $X_F = (X_f)_{f \in F}$ has the state space $\mathcal{R}_F = \prod_{f \in F} [r_f]$.

**Definition 9.3.2.** Let $\Gamma \subseteq 2^{[m]}$ be a simplicial complex and let $r_1, \ldots, r_m \in \mathbb{N}$. For each facet $F \in \Gamma$, we introduce a set of $\# \mathcal{R}_F$ positive parameters $\theta_{i_F}^{(F)}$. The hierarchical log-linear model associated with $\Gamma$ is the set of all probability distributions

\[
\mathcal{M}_\Gamma = \left\{ p \in \Delta_{\mathcal{R}^{m-1}} : p_i = \frac{1}{Z(\theta)} \prod_{F \in \text{facet}(\Gamma)} \theta_{i_F}^{(F)} \text{ for all } i \in \mathcal{R} \right\},
\]
where $Z(\theta)$ is the normalizing constant (or partition function)

$$Z(\theta) = \sum_{i \in \mathcal{R}} \prod_{F \in \mathrm{facet}(\Gamma)} \theta_{i_F}^{(F)}.$$ 

The parameters $\theta_{i_F}^{(F)}$ are sometimes called potential functions, especially when hierarchical models are used in their relationship to graphical models.

**Example 9.3.3 (Independence).** Let $\Gamma = [1][2]$. Then the hierarchical model consists of all positive probability matrices $(p_{i_1i_2})$

$$p_{i_1i_2} = \frac{1}{Z(\theta)} \theta_{i_1}^{(1)} \theta_{i_2}^{(2)}$$

where $\theta^{(j)} \in (0, \infty)^{r_j}$, $j = 1, 2$. That is, the model consists of all positive rank one matrices. It is the positive part of the model of independence $\mathcal{M}_{X \perp \perp Y}$, or in algebraic geometric language, the positive part of the Segre variety. The normalizing constant is

$$Z(\theta) = \sum_{i_1=1}^{r_1} \sum_{i_2=1}^{r_2} \theta_{i_1} \theta_{i_2}.$$ 

In this case, the normalizing constant factorizes as

$$Z(\theta) = \left( \sum_{i_1=1}^{r_1} \theta_{i_1} \right) \left( \sum_{i_2=1}^{r_2} \theta_{i_2} \right).$$

Complete factorization of the normalizing constant as in this example is a rare phenomenon.

**Example 9.3.4 (No 3-way interaction).** Let $\Gamma = [12][13][23]$ be the boundary of a triangle. The hierarchical model $\mathcal{M}_{\Gamma}$ consists of all $r_1 \times r_2 \times r_3$ tables $(p_{i_1i_2i_3})$ with

$$p_{i_1i_2i_3} = \frac{1}{Z(\theta)} \theta_{i_1i_2}^{(12)} \theta_{i_1i_3}^{(13)} \theta_{i_2i_3}^{(23)}$$

for some positive real tables $\theta^{(12)} \in (0, \infty)^{r_1 \times r_2}$, $\theta^{(13)} \in (0, \infty)^{r_1 \times r_3}$, and $\theta^{(23)} \in (0, \infty)^{r_2 \times r_3}$. Unlike the case of the model of independence, this important statistical model does not have a correspondence with any classically studied algebraic variety. In the case of binary random variables, its implicit representation is given by the equation

$$P_{111}P_{122}P_{212}P_{221} = P_{112}P_{121}P_{211}P_{222}.$$ 

That is, the log-linear model consists of all positive probability distributions that satisfy this quartic equation.
Example 9.3.5 (Something more general). Let $\Gamma = [12|23|345]$. The hierarchical model $\mathcal{M}_\Gamma$ consists of all $r_1 \times r_2 \times r_3 \times r_4 \times r_5$ probability tensors $(p_{i_1i_2i_3i_4i_5})$ with

$$p_{i_1i_2i_3i_4i_5} = \frac{1}{Z(\theta)} \theta^{(12)}_{i_1i_2} \theta^{(23)}_{i_2i_3} \theta^{(345)}_{i_3i_4i_5},$$

for some positive real tables $\theta^{(12)} \in (0, \infty)^{r_1 \times r_2}$, $\theta^{(23)} \in (0, \infty)^{r_2 \times r_3}$, and $\theta^{(345)} \in (0, \infty)^{r_3 \times r_4 \times r_5}$. These tables of parameters represent the potential functions.

To begin to understand the Markov bases of hierarchical models, we must come to terms with the 0/1 matrices $A$ that realize these models in the form $M_A$. In particular, we must determine what linear transformation the matrix $A$ represents. Let $u \in \mathbb{N}^{R}$ be an $r_1 \times \cdots \times r_m$ contingency table. For any subset $F = \{f_1, f_2, \ldots\} \subseteq [m]$, let $u|_F$ be the $r_{f_1} \times r_{f_2} \times \cdots$ marginal table such that $(u|_F)_{i_F} = \sum_{j \in R_{[m] \setminus F}} u_{i_F, j}$. The table $u|_F$ is called the $F$-marginal of $u$.

Example 9.3.6 (Margins of 3-way tables). Let $u$ be the $2 \times 2 \times 3$ table

$$
\begin{array}{ccc}
  u_{111} & u_{112} & u_{113} \\
  u_{121} & u_{122} & u_{123} \\
  u_{211} & u_{212} & u_{213} \\
  u_{221} & u_{222} & u_{223} \\
\end{array}
= \begin{pmatrix}
  4 & 6 & 0 \\
  3 & 0 & 12 \\
  5 & 4 & 1 \\
  17 & 0 & 3 \\
\end{pmatrix}
$$

The marginal $u|_{23}$ is the $2 \times 3$ table

$$
\begin{array}{ccc}
  (u|_{23})_{11} & (u|_{23})_{12} & (u|_{23})_{13} \\
  (u|_{23})_{21} & (u|_{23})_{22} & (u|_{23})_{23} \\
\end{array}
= \begin{pmatrix}
  9 & 10 & 1 \\
  20 & 0 & 15 \\
\end{pmatrix}
$$

and the table $u|_1$ is the vector $\begin{pmatrix} 25 \\ 30 \end{pmatrix}$.

Proposition 9.3.7. Let $\Gamma = [F_1|F_2] \cdots$. The matrix $A_\Gamma$ represents the linear transformation

$$u \mapsto (u|_{F_1}, u|_{F_2}, \ldots),$$

and the $\Gamma$-marginals are minimal sufficient statistics of the hierarchical model $\mathcal{M}_\Gamma$.

Proof. We can read off the matrix $A_\Gamma$ from the parametrization. In the parametrization, the rows of $A_\Gamma$ correspond to parameters, and the columns correspond to states of the random variables. The rows come in blocks that correspond to the facets $F$ of $\Gamma$. Each block has cardinality $\#R_F$. Hence, the rows of $A_\Gamma$ are indexed by pairs $(F, i_F)$ where $F$ is a facet of $\Gamma$ and $i_F \in R_F$. The columns of $A_\Gamma$ are indexed by all elements of $R$. The entry in $A_\Gamma$ for row index $(F, i_F)$ and column index $j \in R$ equals 1 if $jF = i_F$ and equals zero otherwise. This description follows by reading the
parametrization from (9.3.1) down the column of $A_{\Gamma}$ that corresponds to $p_j$. The description of minimal sufficient statistics as marginals comes from reading this description across the rows of $A_{\Gamma}$, where the block corresponding to $F$, yields the $F$-marginal $u_{|F}$.

Example 9.3.8 (Sufficient Statistics of Hierarchical Models). Returning to our examples above, for $\Gamma = [1][2]$ corresponding to the model of independence, the minimal sufficient statistics are the row and column sums of $u \in \mathbb{N}^{r_1 \times r_2}$. That is

$$A_{[1][2]}u = (u_{|1}, u_{|2}).$$

Above, we abbreviated these row and column sums by $u_+$ and $u_+$, respectively.

For the model of no 3-way interaction, with $\Gamma = [12][13][23]$, the minimal sufficient statistics consist of all 2-way margins of the three way table $u$. That is

$$A_{[12][13][23]}u = (u_{|12}, u_{|13}, u_{|23})$$

and $A_{[12][13][23]}$ is a matrix with $r_1 r_2 + r_1 r_3 + r_2 r_3$ rows and $r_1 r_2 r_3$ columns.

As far as explicitly writing down the matrix $A_{\Gamma}$, this can be accomplished in a uniform way by assuming that the rows and columns are ordered lexicographically.

Example 9.3.9 (Marginals of a 4-way table). Let $\Gamma = [12][14][23]$ and $r_1 = r_2 = r_3 = r_4 = 2$. Then $A_{\Gamma}$ is the matrix

$$
\begin{pmatrix}
11 & 1111 & 1112 & 1121 & 1122 & 1211 & 1212 & 1221 & 1222 & 2111 & 2112 & 2121 & 2122 & 2211 & 2212 & 2221 & 2222 \\
12 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
21 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
22 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
11 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
12 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
22 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
11 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
12 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
21 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
22 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\end{pmatrix}
$$

where the rows correspond to ordering the facets of $\Gamma$ in the order listed above and using the lexicographic ordering $11 > 12 > 21 > 22$ within each facet. For example, the tenth row of $A_{\Gamma}$, whose label is $1 \cdot 2$, tells us we are looking at the 12 entry of the 14 margin.
To perform the asymptotic tests with the hierarchical models, we need to know the dimension of these models, from which we can compute the degrees of freedom (i.e. the codimension). For a proof of this formula see \[HS02\], Cor 2.7 or Exercise 9.8.

**Proposition 9.3.10.** Let \(\Gamma\) be a simplicial complex on \([m]\), and \(r_1, \ldots, r_m \in \mathbb{N}\). The rank of the matrix \(A_{\Gamma}\) associated to these parameters is

\[
\sum_{F \in \Gamma} \prod_{j \in F} (r_j - 1)
\]

where the sum runs over all faces of \(\Gamma\). The dimension of the associated hierarchical model \(\mathcal{M}_{\Gamma}\) is one less than the rank of \(A_{\Gamma}\).

Note that the rank of the matrix \(A_{\Gamma}\) is sufficiently smaller than the number rows in \(A_{\Gamma}\) in the standard representation of the model. All that matters for defining a log-linear model, however, is the row space of the corresponding matrix. Hence, for the hierarchical models, it makes sense to try to find an alternate matrix representation for hierarchical models where the number of rows of the resulting matrix \(B_{\Gamma}\) is equal to the dimension \((9.3.2)\). This can be done as follows. The rows of \(B_{\Gamma}\) are indexed by pairs of a face \(F \in \Gamma\) and a tuple \(i_F = (i_{f_1}, \ldots, i_{f_k})\) such that \(F = \{f_1, \ldots, f_k\}\) and each \(i_j \in [r_j - 1]\). The columns of \(B_{\Gamma}\) are still indexed by index vectors \(i \in \mathcal{R}\). An entry of \(B_{\Gamma}\) indexed with row indices \((F, j_F)\) and column index \(i\) will be a one if \(i_F = j_F\), and will be zero otherwise.

One of the big challenges in the study of Markov bases of hierarchical models is to find descriptions of the Markov bases as the simplicial complex \(\Gamma\) and the numbers of states of the random variables vary. When it is not possible to give an explicit description of the Markov basis (that is, a list of all types of moves needed in the Markov basis), we might still hope to provide structural or asymptotic information about the types of moves that could arise. In the remainder of this section, we describe some results of this type.

For a simplicial complex \(\Gamma\), let \(|\Gamma| = \bigcup_{F \in \Gamma} F\) denote the ground set of \(\Gamma\).

**Definition 9.3.11.** A simplicial complex \(\Gamma\) is reducible, with reducible decomposition \((\Gamma_1, S, \Gamma_2)\) and separator \(S \subset |\Gamma|\), if it satisfies \(\Gamma = \Gamma_1 \cup \Gamma_2 \) and \(\Gamma_1 \cap \Gamma_2 = 2^S\). Furthermore, we here assume that neither \(\Gamma_1\) nor \(\Gamma_2\) is \(2^S\). A simplicial complex is decomposable if it is reducible and \(\Gamma_1\) and \(\Gamma_2\) are decomposable or simplices (that is, of the form \(2^R\) for some \(R \subseteq [m]\)).

Of the examples we have seen so far, the simplicial complexes \([1][2]\) and \([12][23][345]\) are decomposable, whereas the simplicial complex \([12][13][23]\)
is not reducible. On the other hand, the complex \( \Gamma = [12][13][23][345] \) is reducible but not decomposable, with decomposition \(((12)(13)(23), \{3\}, [345])\). Also, any complex with only two facets is decomposable.

**Remark 9.3.12.** Comparing Definition 9.3.11 for decomposable complex and Definition 8.3.4 for a decomposable graph, we see that these notations are not the same. That is a decomposable graph, when considered as a simplicial complex is probably not decomposable. To make the two definitions related, a graph is a decomposable graph if and only if the complex of cliques of the graph is a decomposable simplicial complex.

For reducible complexes, there is a relatively straightforward procedure to build the Markov basis for \( \Gamma \) from the Markov bases for \( \Gamma_1 \) and \( \Gamma_2 \) which we describe now. To do this, first we introduce **tableau notation**, which is a useful unary representation of elements in the lattice \( \ker \mathbb{Z} A_\Gamma \), or, of binomials in the ring \( \mathbb{K} [p] \).

To construct the tableau notation for a table \( u \in \mathbb{N}^R \) first represent \( u = e_{i_1} + e_{i_2} + \cdots + e_{i_r} \), with each \( i_j \in R \). The tableau for \( u \) is the array

\[
\begin{bmatrix}
  i_1 \\
  i_2 \\
  \vdots \\
  i_r
\end{bmatrix}
\]

Two tableau represent the same table \( u \) if and only if they differ by permuting rows. Tableau notation also gives a representation for a monomial \( p^u \), and can be used to represent binomials. For an element \( u \in \mathbb{Z}^R \), we write \( u = u^+ - u^- \), construct the tableau for each of the nonnegative integer vectors \( u^+ \) and \( u^- \), and write the formal difference of the resulting tableau. For instance, the move

\[ e_{i_1 i_2} + e_{j_1 j_2} - e_{i_1 j_2} - e_{i_2 j_1} \]

for the independence model translates into the tableau notation

\[
\begin{bmatrix}
  i_1 & i_2 \\
  j_1 & j_2
\end{bmatrix}
- \begin{bmatrix}
  i_1 & j_2 \\
  j_1 & i_2
\end{bmatrix}.
\]

In tableau notation, it is easy to compute the marginals of a given table.

**Proposition 9.3.13.** Let \( u \in \mathbb{N}^R \) be a table and \( F \subseteq [m] \), and let \( T \) be the tableau associated to \( u \). Then the tableau for the \( F \)-marginal \( u|_F \) is \( T|_F \), obtained from \( T \) by taking the only columns of \( T \) indexed by \( F \).
Proof. It suffices to see the statement for a tableau consisting of a single row. This corresponds to a table that is a standard unit vector \( e_i \). Its \( F \)-marginal is precisely the table \( e_i|_F = e_iF \). The index \( iF = (ij)_{j \in F} \) which is obtained by taking only the parts of \( i \) indexed by \( F \).

Proposition 9.3.13 allows for us to easily check whether or not two tables \( u \) have the same \( \Gamma \) marginals, equivalently, when an element is in the kernel of \( A_\Gamma \), equivalently, when the corresponding binomial belongs to the toric ideal \( I_{A_\Gamma} \). This is achieved by checking when each of the restricted tableau are the same up to permuting rows.

Example 9.3.14 (Marginals of Tableaux). Consider the complex \( \Gamma = [12][13][234] \). Then the following move

\[
\begin{bmatrix}
1 & 1 & 1 & 2 \\
1 & 2 & 2 & 1 \\
2 & 1 & 2 & 2 \\
2 & 2 & 1 & 2
\end{bmatrix}
\begin{bmatrix}
1 & 1 & 2 & 2 \\
1 & 2 & 1 & 2 \\
2 & 1 & 1 & 2 \\
2 & 2 & 1 & 1
\end{bmatrix}
\]

represented in tableau notation, belongs to \( \ker Z A_\Gamma \), since for each facet \( F \) of \( \Gamma \), the restriction of the tableaux to the columns indexed by \( F \) represent the same table.

Let \( \Gamma \) be a reducible complex with decomposition \((\Gamma_1, S, \Gamma_2)\). We let \( V_1 = |\Gamma_1| \setminus S \) and \( V_2 = |\Gamma_2| \setminus S \), where \(|\Gamma_i| = \bigcup_{F \in \Gamma_i} F \) is the ground set of \( \Gamma_i \). After renaming the elements of the ground sets, we can assume that \( V_1 = \{1, 2, \ldots, m_1\} \), \( S = \{m_1 + 1, \ldots, m_2\} \), and \( V_2 = \{m_2 + 1, \ldots, m\} \). With respect to the resulting tripartition \( V_1|S|V_2 \), we can represent the tableau as having three groups of columns, as

\[
\begin{bmatrix}
i_1 & j_1 & k_1 \\
\vdots & \vdots & \vdots \\
i_d & j_d & k_d
\end{bmatrix}
\]

so that, \( i_1 \in R_{V_1}, j_1 \in R_S, \ldots, k_d \in R_{V_2} \). Similarly, the tableau involved in representing a move for \( A_{\Gamma_1} \) or \( A_{\Gamma_2} \) will look like

\[
\begin{bmatrix}
i_1 & j_1 \\
\vdots & \vdots \\
i_d & j_d
\end{bmatrix}
\quad \text{or} \quad
\begin{bmatrix}
j_1 & k_1 \\
\vdots & \vdots \\
j_d & k_d
\end{bmatrix}
\]

respectively, where the \( j_1, \ldots, j_d \) are in \( R_S \).

For a reducible complex \( \Gamma \) with decomposition \((\Gamma_1, S, \Gamma_2)\) we first introduce an operation to take moves for \( \Gamma_1 \) and \( \Gamma_2 \) to produce moves for \( \Gamma \). This
is the lifting operation, which we explain now. Let
\[
\begin{bmatrix}
i_1 & j_1 \\
i_d & j_d \\
\end{bmatrix}
- 
\begin{bmatrix}
i'_1 & j'_1 \\
i'_d & j'_d \\
\end{bmatrix}
\]
be a move for \( \Gamma_1 \). Since \( S \) is a face of \( \Gamma_1 \), we must have that the multisets of \( \{j_1, \ldots, j_d\} \) and \( \{j'_1, \ldots, j'_d\} \) are the same, by Proposition 9.3.13. Hence, after reordering rows and renaming indices, we can assume this move has the form
\[
\begin{bmatrix}
i_1 & j_1 \\
i_d & j_d \\
\end{bmatrix}
- 
\begin{bmatrix}
i'_1 & j'_1 \\
i'_d & j'_d \\
\end{bmatrix}
\]
(9.3.3)
Now let \( k_1, \ldots, k_d \in \mathcal{R}_{V_2} \) and form the move
\[
\begin{bmatrix}
i_1 & j_1 & k_1 \\
i_d & j_d & k_d \\
\end{bmatrix}
- 
\begin{bmatrix}
i'_1 & j'_1 & k_1 \\
i'_d & j'_d & k_d \\
\end{bmatrix}
\]
(9.3.4)

**Proposition 9.3.15.** Let \( \Gamma \) be reducible with decomposition \((\Gamma_1, S, \Gamma_2)\). Consider the move (9.3.3) which belongs to \( I_{\Gamma_1} \), and form the move (9.3.4) by applying the lifting procedure above. Then the resulting move (9.3.4) belongs to \( I_{\Gamma} \).

**Proof.** Let \( T \) and \( T' \) be the resulting tableau in (9.3.4). We must show that for any face \( F \in \Gamma \), \( T|_F = T'|_F \). Since \( \Gamma \) is reducible, either \( F \in \Gamma_1 \) or \( F \in \Gamma_2 \). In the first case, forming the restriction of \( T|_{\Gamma_1} - T'|_{\Gamma_1} \) yields the starting move. Since that move belonged to \( I_{\Gamma_1} \) we must have \( T|_F = T'|_F \). On the other hand if \( F \in \Gamma_2 \) we already have that \( T|_{\Gamma_2} = T'|_{\Gamma_2} \) which implies that \( T|_F = T'|_F \). \( \square \)

If \( \mathcal{F}_1 \subseteq I_{\Gamma_1} \) is a set of moves, then we denote by \( \text{Lift}(\mathcal{F}_1) \) the set of all the moves in \( I_{\Gamma} \) by applying this procedure, using all possible elements \( k_1, \ldots, k_d \in \mathcal{R}_{V_2} \). Similarly, we can run a similar lifting procedure to moves \( \mathcal{F}_2 \subseteq I_{\Gamma_2} \) to obtain the lifted set of moves \( \text{Lift}(\mathcal{F}_2) \).

The second set of moves that we use are quadratic moves associated to the tripartition \( V_1, S, V_2 \).

**Proposition 9.3.16.** Let \( \Gamma \) be a decomposable simplicial complex with two facets \( F_1, F_2 \) with \( S = F_1 \cap F_2 \), \( V_1 = F_1 \setminus S \), and \( V_2 = F_2 \setminus S \). The set of quadratic moves \( \text{Quad}(V_1, S, V_2) \) of the form
\[
\begin{bmatrix}
i_1 & j & k_1 \\
i_2 & j & k_2 \\
\end{bmatrix}
- 
\begin{bmatrix}
i_1 & j & k_2 \\
i_2 & j & k_1 \\
\end{bmatrix}
\]
(9.3.5)
form a Markov basis for \( I_{\Gamma} \), where \( i_1, i_2 \in \mathcal{R}_{V_1} \), \( j \in \mathcal{R}_S \), and \( k_1, k_2 \in \mathcal{R}_{V_2} \).
Proof. This Markov basis problem reduces to the case of 2-way tables with fixed row and column sums. The details are left to the reader. □

The set of quadratic moves $\text{Quad}(V_1, S, V_2)$ associated to a tripartition of the ground set $[m]$, will play an important role in the next Theorem, which gives a description of the Markov bases of reducible hierarchical models in terms of the Markov bases of each of the pieces.

**Theorem 9.3.17** (Markov bases of reducible models \([\text{DS04, HS02]}\)). Let $\Gamma$ be a reducible simplicial complex with decomposition $(\Gamma_1, S, \Gamma_2)$, and suppose that $\mathcal{F}_1$ and $\mathcal{F}_2$ are Markov bases for $I_{\Gamma_1}$ and $I_{\Gamma_2}$ respectively. Then

\begin{equation}
\mathcal{G} = \text{Lift}(\mathcal{F}_1) \cup \text{Lift}(\mathcal{F}_2) \cup \text{Quad}(\lvert \Gamma_1 \setminus S, S, \lvert \Gamma_2 \setminus S)
\end{equation}

is a Markov basis for $I_{\Gamma}$.

**Proof.** Let $u$ and $u'$ be two tables belonging to the same fiber with respect to $\Gamma$, and $T$ and $T'$ the corresponding tableaux. We must show that $T$ and $T'$ can be connected using the moves from $\mathcal{G}$.

Restricting to the set $\lvert \Gamma_1 \rvert$, we know that $T\lvert_{\Gamma_1}$ and $T'\lvert_{\Gamma_1}$ can be connected using the moves in $\mathcal{F}_1$, since $\mathcal{F}_1$ is a Markov basis for $\Gamma_1$. Suppose that the move

\[
f = \begin{bmatrix}
i_1 & j_1 \\
\vdots & \vdots \\
i_d & j_d
\end{bmatrix} - \begin{bmatrix}
i'_1 & j_1 \\
\vdots & \vdots \\
i'_d & j_d
\end{bmatrix}
\]

is to be applied first to $T\lvert_{\Gamma_1}$. For it to be possible to apply this move to $T\lvert_{\Gamma_1}$, it must be the case that, up to permuting the rows of $T$, the first $d$ rows of $T$ are

\[
\begin{bmatrix}
i_1 & j_1 & k_1 \\
\vdots & \vdots & \vdots \\
i_d & j_d & k_d
\end{bmatrix}
\]

for some $k_1, \ldots, k_d$. Hence, we can lift $f$ to the move

\[
\begin{bmatrix}
i_1 & j_1 & k_1 \\
\vdots & \vdots & \vdots \\
i_d & j_d & k_d
\end{bmatrix} - \begin{bmatrix}
i'_1 & j_1 & k_1 \\
\vdots & \vdots & \vdots \\
i'_d & j_d & k_d
\end{bmatrix}
\]

in $\text{Lift}(\mathcal{F}_1)$ which can be applied to $T$. This same reasoning can be applied to all the moves that are used in sequence to connect $T\lvert_{\Gamma_1}$ and $T'\lvert_{\Gamma_1}$. These moves connect $T$ with a new tableau $T''$ which has the property that $T''\lvert_{\Gamma_1} = T'\lvert_{\Gamma_1}$.

Applying the same reasoning with respect to $\Gamma_2$, we can find a set of moves in $\text{Lift}(\mathcal{F}_2)$ which connect $T''$ to a new tableau $T'''$ and such that $T'''\lvert_{\Gamma_2} = T'\lvert_{\Gamma_2}$. Furthermore, the moves in $\text{Lift}(\mathcal{F}_2)$ have the property that...
they do not change the marginal with respect to $|\Gamma_1|$. Hence $T''|_{|\Gamma_1|} = T'|_{|\Gamma_1|}$, as well. So $T''$ and $T'$ have the same marginals with respect to the simplicial complex whose only facets are $|\Gamma_1|$ and $|\Gamma_2|$. By Proposition 9.3.16, $T''$ and $T'$ can be connected using the moves in $\text{Quad}(|\Gamma_1| \setminus S, S, |\Gamma_2| \setminus S)$. Thus, $\mathcal{G}$ is a Markov basis for $\Gamma$. □

Using the fact that the Markov basis for the model when $\Gamma$ is a simplex is just the empty set, and applying induction, one gets a complete description of the Markov bases of decomposable models.

**Corollary 9.3.18** ([Dob03, Tak99]).

If $\Gamma$ is a decomposable simplicial complex, then the set of moves

$$
\mathcal{B} = \bigcup_{(\Gamma_1, S, \Gamma_2)} \text{Quad}(|\Gamma_1| \setminus S, S, |\Gamma_2| \setminus S),
$$

with the union over all reducible decompositions of $\Gamma$, is a Markov basis for $A_{\Gamma}$.

**Example 9.3.19** (Markov Bases of the 4-chain).

Consider the four-chain $\Gamma = [12][23][34]$. This graph has two distinct reducible decompositions with minimal separators, namely $([12], \{2\}, [23][34])$ and $([12][23], \{3\}, [34])$. Therefore, the Markov basis consists of moves of the set of moves

$$
\text{Quad}(\{1\}, \{2\}, \{3, 4\}) \cup \text{Quad}(\{1, 2\}, \{3\}, \{4\}).
$$

In tableau notation these moves look like:

$$
\begin{bmatrix}
  i_1 & j & i_3 & i_4 \\
  i'_1 & j & i'_3 & i'_4
\end{bmatrix}
\quad \text{and} \quad
\begin{bmatrix}
  i_1 & i_2 & j & i_3 \\
  i'_1 & i'_2 & j & i'_4
\end{bmatrix}
\quad \text{and} \quad
\begin{bmatrix}
  i_1 & i_2 & j & i_4 \\
  i'_1 & i'_2 & j & i'_4
\end{bmatrix}.
$$

Note that the decomposition $([12][23], \{2, 3\}, [23][34])$ is also a valid reducible decomposition of $\Gamma$, but it does not produce any new Markov basis elements.

Theorem 9.3.17 can also be generalized beyond reducible models in some instances using the theory of toric fiber products [Sul07, RS16, EKS14]. Theorem 9.3.17 shows that to determine Markov bases of all hierarchical models it suffices to focus on attention on hierarchical models that are not reducible. While there a number of positive results in the literature about the structure of Markov bases in the nonreducible case, a general description is probably impossible, and it is already difficult to describe Markov bases completely in even the simplest nonreducible models.

**Example 9.3.20** (Markov Basis of Model of No-3-way Interaction).

The simplest nonreducible hierarchical model has simplicial complex $\Gamma = [12][13][23]$. Suppose that $r_1 = r_2 = 3$ and $r_3 = 5$. The Markov basis for this model consists of 2670 moves. Among these are moves that generalize the ones for
2-way tables:

\[
\begin{array}{ccc}
1 & 1 & 1 \\
1 & 2 & 2 \\
2 & 1 & 2 \\
2 & 2 & 1
\end{array}
\quad \quad \quad
\begin{array}{ccc}
1 & 1 & 2 \\
1 & 2 & 1 \\
2 & 1 & 1 \\
2 & 2 & 2
\end{array}
\]

but also unfamiliar moves where it is hard to imagine how they generalize to larger tables, for example:

\[
\begin{array}{ccc}
1 & 1 & 1 \\
1 & 2 & 2 \\
1 & 2 & 5 \\
1 & 3 & 4 \\
2 & 1 & 3 \\
2 & 2 & 1 \\
2 & 3 & 5 \\
3 & 1 & 2 \\
3 & 2 & 4 \\
3 & 3 & 3
\end{array}
\quad \quad \quad
\begin{array}{ccc}
1 & 1 & 2 \\
1 & 2 & 1 \\
1 & 2 & 4 \\
1 & 3 & 5 \\
2 & 1 & 1 \\
2 & 2 & 5 \\
2 & 3 & 3 \\
3 & 1 & 3 \\
3 & 2 & 2 \\
3 & 3 & 4
\end{array}
\]

The no-3-way interaction model with \( r_1 = r_2 = r_4 = 4 \) has 148968 moves in its minimal Markov basis.

One of the remarkable consequences of Corollary 9.3.18 is that the structure of the Markov basis of a decomposable hierarchical log-linear model does not depend on the number of states of the underlying random variables. In particular, regardless of the sizes of \( r_1, r_2, \ldots, r_m \), the Markov basis for a decomposable model always consists of moves with one-norm equal to four, with a precise and global combinatorial description. The following theorem of De Loera and Onn [DLO06] says that this nice behavior fails, in the worst possible way, already for the simplest non-decomposable model. We fix \( \Gamma = [12][13][23] \) and consider \( 3 \times r_2 \times r_3 \) tables, where \( r_2, r_3 \) can be arbitrary. De Loera and Onn refer to these as slim tables.

**Theorem 9.3.21** (Slim tables). Let \( \Gamma = [12][13][23] \) be the 3-cycle and let \( v \in \mathbb{Z}^k \) be any integer vector. Then there exist \( r_2, r_3 \in \mathbb{N} \) and a coordinate projection \( \pi : \mathbb{Z}^{3 \times r_2 \times r_3} \to \mathbb{Z}^k \) such that every minimal Markov basis for \( \Gamma \) on \( 3 \times r_2 \times r_3 \) tables contains a vector \( u \) such that \( \pi(u) = v \).

In particular, Theorem 9.3.21 shows that there is no hope for a general bound on the one-norms of Markov basis elements for non-decomposable models, even for a fixed simplicial complex \( \Gamma \). On the other hand, if only one of the table dimensions is allowed to vary, then there is a bounded finite structure to the Markov bases. This theorem was first proven in [HS07b] and generalizes a result in [SS03a].
9.4. Graver Bases and Applications

Theorem 9.3.22 (Long tables). Let $\Gamma$ be a simplicial complex and fix $r_2, \ldots, r_m$. There exists a number $b(\Gamma, r_2, \ldots, r_m) < \infty$ such that the one-norms of the elements of any minimal Markov basis for $\Gamma$ on $s \times r_2 \times \cdots \times r_m$ tables are less than or equal to $b(\Gamma, r_2, \ldots, r_m)$. This bound is independent of $s$, which can grow large.

From Corollary 9.3.18, we saw that if $\Gamma$ is decomposable and not a simplex, then $b(\Gamma, r_2, \ldots, r_m) = 4$. One of the first discovered results in the non-decomposable case was $b([12][13][23], 3, 3) = 20$, a result obtained by Aoki and Takemura [AT03]. The most complicated move that appears for $3 \times 3 \times r_3$ tables is the one shown in Example 9.3.20.

In general, it seems a difficult problem to actually compute the values $b(\Gamma, r_2, \ldots, r_m)$, although some recent progress was reported by Hemmecke and Nairn [HN09]. The proof of Theorem 9.3.22 only gives a theoretical upper bound on this quantity, involving other numbers that are also difficult to compute. Further generalizations of these results appear in [HS12].

The Markov basis database [KR14] collects known results about specific Markov bases, including Markov bases of hierarchical models. Further theoretical results on Markov bases for many models, including hierarchical models, can be found in [AHT12].

9.4. Graver Bases and Applications

In some contingency table analysis problems, the types of tables that can occur are limited by structural conditions. A typical example is that cell entries are constrained to be in $\{0, 1\}$, or that certain entries are forced to be zero, but other types of cell entry constraints are also common. In this case, the notion of Markov basis might need to be changed to take into account these new constraints. One tool that is helpful in these contexts is the Graver basis, which can solve Markov basis problems for all bounded cell entry problems simultaneously.

We start this section with a very general definition of a Markov basis.

**Definition 9.4.1.** Let $\mathcal{F}$ be a collection of finite subsets of $\mathbb{N}^r$. A finite set $\mathcal{B} \subseteq \mathbb{Z}^r$ is called a Markov basis Markov basis for $\mathcal{F}$ if for all $\mathcal{F} \in \mathcal{F}$ and for all $u, u' \in \mathcal{F}$, there is a sequence $v_1, \ldots, v_L \in \mathcal{B}$ such that

$$u' = u + \sum_{k=1}^{L} v_k \quad \text{and} \quad u + \sum_{k=1}^{l} v_k \in \mathcal{F} \text{ for all } l = 1, \ldots, L.$$  

The elements of the Markov basis are called moves.

This definition contains our previous definition of Markov basis as a subcase with $\mathcal{F} = \{\mathcal{F}(u) : u \in \mathbb{N}^r\}$. In this section, we consider more
general sets of fibers. Let $L \in \mathbb{N}^r$ and $U \in \mathbb{N}^r$. For a given matrix $A \in \mathbb{Z}^{d \times r}$, and $u \in \mathbb{N}^r$, let the fiber $F(u, L, U)$ be the set

$$F(u, L, U) = \{v \in \mathbb{N}^r : Av = Au \text{ and } L_i \leq v_i \leq U_i \text{ for all } i \in [r]\}.$$  

Hence $L$ and $U$ are vectors of lower and upper bound constraints on cell entries. Let $F_A^{\text{bound}}$ be the set of all the fibers $F(u, L, U)$ with $u, L, U \in \mathbb{N}^r$. We will describe the Markov basis for this set of fibers.

**Definition 9.4.2.** Let $u \in \mathbb{Z}^r$. A conformal decomposition of $u$ is an expression $u = v + w$ where $v, w \in \mathbb{Z}^r \setminus \{0\}$ and $|u_i| = |v_i| + |w_i|$ for all $i \in [r]$. Given a matrix $A \in \mathbb{Z}^{d \times r}$, the Graver basis of $A$, denoted $Gr_A$, is the set of all nonzero $u \in \ker A$ such that $u$ does not have a conformal decomposition with $v$ and $w$ in $\ker A$.

Conformal decomposition of a vector is also known as a sign consistent decomposition. A simple example showing the structure of a Graver basis concerns the independence model for 2-way tables.

**Proposition 9.4.3.** Let $\Gamma = [1][2]$ and $r = (r_1, r_2)$ and let $A_\Gamma$ be the resulting matrix representing the model of independence of two random variables. Let $i = i_1, \ldots, i_k \in [r_1]$ be a sequence of disjoint elements, and $j = j_1, \ldots, j_k \in [r_2]$ be a sequence of disjoint elements. The move

$$f_{i,j} = e_{i_1j_1} - e_{i_1j_2} + e_{i_2j_2} - e_{i_2j_3} + \cdots + e_{i_kj_k} - e_{i_kj_1},$$

belongs to the Graver basis of $A_\Gamma$, and the Graver basis of $A_\Gamma$ consists of all such moves for $k = 2, \ldots, \min(r_1, r_2)$.

**Proof.** First, suppose that $f_{i,j}$ was not in the Graver basis. Then there are $v, w \in \ker_A A\Gamma$ such that $f_{i,j} = v + w$ and this is a conformal decomposition. Since $f_{i,j}$ is a vector with all entries in $\{-1, 0, 1\}$, each of the unit vectors contributing to its definition ends up in either $v$ or $w$. Say that $e_{i_1j_1}$ appears in $v$. Since $v$ has all row and column sums equal to zero, this forces $-e_{i_1j_2}$ to appear in $v$ as well. By reasoning through the cyclic structure, we must have $v = f_{i,j}$ and $w = 0$ contradicting that $f_{i,j} = v + w$ is a conformal decomposition. This shows that $f_{i,j}$ is in the Graver basis.

To see that this set of moves is a Graver basis, let $u \in \ker_A A\Gamma$ assume it is not an $f_{i,j}$ move. After permuting rows and columns, we can assume that $u_{11} > 0$. Since the row and columns sums of $u$ are zero, then there is some entry in row one that is negative, say $u_{12} < 0$. Since the row and column sums of $u$ are zero, then there is some entry in column two that is positive, say $u_{22} > 0$. Now one of two things could happen. If $u_{21} < 0$, then $u = f_{12} + w$ will be a conformal decomposition. If $u_{21} \geq 0$, then there is some entry say $u_{23}$ in the first row that is negative. Repeating this procedure eventually must close off a cycle to obtain $u = f_{i,j} + w$ as
9.4. Graver Bases and Applications

a conformal decomposition. This shows that the set of \( f_{i,j} \) is a Graver basis.

In general, the Graver basis of a matrix \( A \) is much larger and more complex than a minimal Markov basis of that matrix. The Graver basis of a matrix \( A \) can be computed using tools from computational algebra, and direct algorithms for computing it are available in 4ti2 [HH03], which can be accessed through Macaulay2.

**Example 9.4.4** (Graver Basis Computation). Here we illustrate how to compute the Graver basis of the hierarchical model with \( \Gamma = [12][13][14][234] \) and \( r = (2, 2, 2, 2) \) in Macaulay2.

```plaintext
loadPackage "FourTiTwo"
S = ZZ[x1,x2,x3,x4]
listofmonomials = matrix{{1,x1,x2,x3,x4,x1*x2,x1*x3, x1*x4, x2*x3,x2*x4,x3*x4,x2*x3*x4}};
A = transpose matrix toList apply((0,0,0,0)..(1,1,1,1),
    i -> flatten entries sub(listofmonomials,
    {x1 => i_0, x2 => i_1, x3=> i_2, x4 => i_3}))
L = toricGraver(A)
```

The output is a 20 \( \times \) 16 matrix whose rows are the elements of the Graver basis up to sign. For example, the first row is the vector:

\[
2 \ -1 \ -1 \ 0 \ 1 \ 0 \ 0 \ -1
\]

which can be represented in tableau notation as

\[
\begin{bmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 2 & 2 & 2 \\
2 & 1 & 1 & 2 \\
2 & 1 & 2 & 1 \\
2 & 2 & 1 & 1 \\
\end{bmatrix}
\begin{bmatrix}
1 & 1 & 1 & 2 \\
1 & 1 & 2 & 1 \\
1 & 2 & 1 & 1 \\
2 & 1 & 1 & 1 \\
2 & 1 & 1 & 1 \\
2 & 2 & 2 & 2 \\
\end{bmatrix}
\]

**Theorem 9.4.5.** Let \( A \in \mathbb{Z}^{d\times r} \). Then the Graver basis of \( A \) is a Markov basis for the set of fibers \( F_A \).

**Proof.** Let \( u, u' \in \mathcal{F}(u, L, U) \). We must show that \( u \) and \( u' \) can be connected by elements of \( Gr_A \) without leaving \( \mathcal{F}(u, L, U) \). The vector \( u - u' \in \ker_2 A \). It is either a Graver basis element, in which case \( u \) and \( u' \) are connected by a Graver basis element, or there is a Graver basis element \( v \) such that \( u - u' = v + w \) is a conformal decomposition, where \( w \in \ker_2 \). But then \( u - v \in \mathcal{F}(u, L, U) \) since

\[
L_i \leq \min(u_i, u'_i) \leq (u - v)_i \leq \max(u_i, u'_i) \leq U_i
\]
for all \( i \in [r] \). Furthermore, \( \|u - u'\|_1 > \|(u - v) - u'\|_1 \) so \((u - v)\) and \(u'\) are closer than \(u\) and \(u'\). So by induction, it is possible to connect \(u\) and \(u'\) by a sequence of moves from \(Gr_A\).

\[ \square \]

In some problems of statistical interest, it is especially interesting to focus only on the case where all lower bounds are zero and all upper bounds are 1, that is, restricting to the case where all elements of the fiber have 0/1 entries. In this case, it is only necessary to determine the Graver basis elements that have 0, ±1 entries. Even so, this set of Graver basis elements might be larger than necessary to connect the resulting fibers. We use \(0\) and \(1\) to denote the vector or table of all 0’s or 1’s, respectively.

**Theorem 9.4.6.** Let \( \Gamma = [1][2] \) and \( r = (r_1, r_2) \) and let \( A_{r'} \) be the resulting matrix representing the model of independence of two random variables. Then the set of moves

\[
(9.4.1) \quad \{ e_{i_1,j_1} + e_{i_2,j_2} - e_{i_1,j_2} - e_{i_2,j_1} : i_1, i_2 \in [r_1], j_1, j_2 \in [r_2] \}
\]

is a Markov basis that connects the fibers \( \mathcal{F}(u, 0, 1) \) for all \( u \).

**Proof.** We show that these standard moves for the independence model can be used to bring any two elements in the same fiber closer to each other. Let \( u, v \in \mathcal{F}(u, 0, 1) \). First of all, we can assume that \( u \) and \( v \) have no rows in common, otherwise we could delete this row and reduce to a table of a smaller size. We will show that we can apply moves from (9.4.1) to make the first rows of \( u \) and \( v \) the same. After simultaneously permuting columns of \( u \) and \( v \), we can assume that the first rows of each \((u_1 \text{ and } v_1, \text{ respectively})\) have the following form:

\[
\begin{align*}
  u_1 &= (1 \ 0 \ 0 \ 1) \quad v_1 = (1 \ 0 \ 1 \ 0). 
\end{align*}
\]

The corresponding blocks of each component are assumed to have the same size. In other words, we have arranged things so that we have first a region where \(u_1\) and \(v_1\) completely agree, and then a region where they completely disagree. It suffices to show that we can decrease the length of the region where they completely disagree.

If it were not possible to decrease the region where they disagree then, in particular, we could not apply a single move from (9.4.1) that involves the first row entries corresponding to the last two blocks. We focus on what the rest of the matrices \(u\) and \(v\) look like below those blocks. In particular, \(u\) will have the following form after permuting rows:

\[
\begin{pmatrix}
  1 & 0 & 0 & 1 \\
  A_1 & A_2 & A_3 & 1 \\
  B_1 & B_2 & 0 & B_4
\end{pmatrix}
\]
The salient feature here is that if a 1 appeared in any position in the third block of columns, then in the corresponding row in the fourth block must be all 1’s, otherwise it would be possible to use a move to decrease the amount of discrepancy in the first row. Similarly, if there were a 0 in the fourth block of columns there must be all 0’s in the corresponding row in the third block. A similar argument shows that there is a (possibly different) permutation of rows so that \( v \) has the following form:

\[
\begin{pmatrix}
1 & 0 & 1 & 0 \\
C_1 & C_2 & C_3 & 0 \\
D_1 & D_2 & 1 & D_3
\end{pmatrix}.
\]

Let \( a \) and \( b \) be the number rows in the corresponding blocks of rows in \( u \) and \( c \) and \( d \) the number of rows in the corresponding blocks of rows in \( v \). Since \( u \) and \( v \) have the same column sums, we deduce the following inequalities

\[
1 + d \leq a \quad \text{and} \quad 1 + a \leq d
\]

which is clearly a contradiction. Hence, there must be a move that could be applied that would decrease the discrepancy in the first row. \( \square \)

A common setting in analysis of contingency table data is that some entries are constrained to be zero. For instance, in the genetic data of the Hardy-Weinberg model (Theorem 9.2.10) it might be the case that certain pairs of alleles cannot appear together in an organism because they cause greatly reduced fitness. These forced zero entries in the contingency table are called \textit{structural zeroes} and the Graver basis can be useful in their analysis. Indeed, in the context of Markov bases for fibers with bounds, this is equivalent to setting the upper bounds to 0 for certain entries. Hence, having the Graver basis for a certain matrix \( A \) can be useful for any structural zero problem associated to the corresponding log-linear model [Rap06]. See also Exercise 9.3.

9.5. Lattice Walks and Primary Decompositions

A Markov basis for a matrix \( A \) has the property that it connects every fiber \( \mathcal{F}(u) \). In practical applications, we are usually interested in the connectivity of a specific fiber \( \mathcal{F}(u) \). Furthermore, it might not be possible to obtain a Markov basis because it is too big to compute, and there might not exist theoretical results for producing the Markov basis. In this setting, a natural strategy is to find and use a set of easy to obtain moves \( \mathcal{B} \subset \ker \mathbb{Z} A \) to perform a Markov chain over the fiber \( \mathcal{F}(u) \), which a priori may not be connected. Then we hope to understand exactly how much of the fiber \( \mathcal{F}(u) \) is connected by \( \mathcal{B} \), and, in particular, conditions on \( u \) that guarantee that the fiber \( \mathcal{F}(u) \) is connected. One strategy for addressing this problem
Fisher’s Exact Test is based on primary decomposition of binomial ideals, as introduced in the paper [DES98]. Because the set of moves \( B \) does not connect every fiber, these sets are sometimes called Markov subbases.

The strategy for using primary decomposition is based on the following simple generalization of Theorem 9.2.5, whose proof is left as an exercise.

**Theorem 9.5.1.** Let \( B \subseteq \ker A \) be a subset of the lattice \( \ker A \) and let \( J_B \) be the ideal \( J_B = \langle p^{v'} - p^v : v \in B \rangle \). Then \( u, u' \in \mathcal{F}(u) \) are in the same connected component of the graph \( \mathcal{F}(u)_B \) if and only if \( p^u - p^{u'} \in J_B \).

Note that Theorem 9.5.1 immediately implies Theorem 9.2.5 since we require every \( u, u' \in \mathcal{F}(u) \) to be in the same connected component to have a Markov basis.

Of course, deciding whether or not \( p^u - p^{u'} \in J_B \) is a difficult ideal membership problem in general. The method suggested in [DES98] is to compute a decomposition of the ideal \( J_B \), as \( J_B = \cap_i I_i \) (for example, a primary decomposition). Then deciding membership in \( J_B \) is equivalent to deciding membership in each of the ideals \( I_i \). The hope is that it is easier to decide membership in each of the ideals \( I_i \) and combine this information to determine membership in \( J_B \). We illustrate the point with an example.

**Example 9.5.2** (2 \( \times \) 3 table). Consider the problem of connecting 2 \( \times \) 3 tables with fixed row and column sums. We know that the set

\[
B' = \left\{ \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & -1 \\ -1 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix} \right\}
\]

forms a Markov basis for the corresponding \( A \) via Theorem 9.2.7. But let us consider the connectivity of the random walk that arises from using the first and last moves in this set. We denote the corresponding set of moves \( B \). The resulting ideal is the ideal generated by two binomials:

\[
J_B = \langle p_{11}p_{22} - p_{12}p_{21}, p_{12}p_{23} - p_{13}p_{22} \rangle.
\]

This ideal is an example of an ideal of adjacent minors [HS04]. The primary decomposition of \( J_B \) is

\[
J_B = \langle p_{11}p_{22} - p_{12}p_{21}, p_{12}p_{23} - p_{13}p_{22}, p_{11}p_{23} - p_{13}p_{21} \rangle \cap \langle p_{12}, p_{22} \rangle.
\]

We want to use this decomposition to decide when two 2 \( \times \) 3 tables are connected by the moves of \( B \). The first component is the ideal of 2 \( \times \) 2 minors of a generic 2 \( \times \) 3 matrix, which is the toric ideal of the independence model, i.e. \( J_{B'} \). A binomial \( p^u - p^v \) belongs to \( J_B \) if and only if \( u \) and \( v \) have the same row and column sums. For a binomial \( p^u - p^v \) to belong to the second component, we must have each of \( p^u \) and \( p^v \) divisible by either \( p_{12} \) or \( p_{22} \). Thus we deduce that \( u \) and \( v \) are connected by the set of moves \( B \) if and only if
(1) \( u \) and \( v \) have the same row and column sums and
(2) the sum \( u_{12} + u_{22} = v_{12} + v_{22} > 0 \).

Attempts to generalize this example to arbitrary sets of adjacent minors appear in [DES98, HS04].

For hierarchical models, there are many natural families of moves which could be used as a starting point to try to employ the strategy for analyzing connectivity based on primary decomposition. In the case where the underlying complex \( \Gamma \) is the set of cliques in a graph \( G \), then the natural choice is the set of degree 2 moves that come from the conditional independence statements implied by the graph. (The connection between conditional independence structures and graphs will be explained in Chapter 13.) A computational study along these lines was undertaken in [KRS14].

### 9.6. Other Sampling Strategies

In the previous sections we saw Markov Chain Monte Carlo strategies for generating tables from the fibers \( \mathcal{F}(u) \) based on fixed marginal totals. These were based on computing Markov bases or Graver bases or using appropriate Markov subbases which connect many fibers. In this section we describe some alternative strategies for sampling from the fibers \( \mathcal{F}(u) \), which can be useful in certain situations.

The urn scheme is a method for obtaining exact samples from the hypergeometric distribution on 2-way contingency tables with fixed row and column sums, with a natural generalization to decomposable hierarchical models. It can also be applied to speed up the generation of random samples from the hypergeometric distribution on fibers associated to reducible models. Note, however, that if a different distribution on the fiber \( \mathcal{F}(u) \) is required, it does not seem possible to modify the urn scheme to generate samples. For example, even introducing a single structural zero can make it impossible to use the urn scheme to generate samples from the hypergeometric distribution. The basic urn scheme for two-way tables is described as follows.

**Algorithm 9.6.1** (Urn Scheme for 2-way tables). *Input:* A table \( u \in \mathbb{N}^{r_1 \times r_2} \).

*Output:* A sample \( v \in \mathbb{N}^{r_1 \times r_2} \) from the hypergeometric distribution on \( \mathcal{F}(u) \).

(1) Let \( b_i = \sum_j u_{ij}, \ c_j = \sum_i u_{ij}, \) and \( n = \sum_{i,j} u_{ij} \)
(2) Let \( P \) be a uniformly random permutation matrix of size \( n \times n \).
(3) For each $i,j$ define

$$v_{ij} = \sum_{k=b_1+\cdots+b_{i-1}+1}^{b_1+\cdots+b_i} \sum_{l=c_1+\cdots+c_{j-1}+1}^{c_1+\cdots+c_j} P_{kl}$$

(4) Output $v$.

In other words, to generate a random table from the hypergeometric distribution on $F(u)$, we just need to generate a uniformly random permutation and then bin the rows and columns into groups whose sizes are the row and column sums of $u$. This is doable, because it is easy to generate uniformly random permutation matrices, by sampling without replacement.

**Proof of correctness of Algorithm 9.6.1.** We just need to show that the probability of sampling a table $v$ is proportional to $\frac{1}{Q_{ij} v_{ij}!}$. To do this, we calculate the number of permutations that lead to a particular table $v$. This is easily seen to be

$$\prod_{i=1}^{r_1} \binom{b_i}{v_{i1}, v_{i2}, \ldots, v_{ir_i}} \prod_{j=1}^{r_2} \binom{c_j}{v_{1j}, v_{2j}, \ldots, v_{r_1j}} \prod_{i=1}^{r_1} \prod_{j=1}^{r_2} v_{ij}!.$$

The first product shows the numbers of ways to choose columns for the 1’s in the permutation to live, compatible with the $v_{ij}$, the second product gives the number of ways to choose columns for the permutation, compatible with the $v_{ij}$, and the $v_{ij}!$ give the number of permutations on the resulting $v_{ij} \times v_{ij}$ submatrix. Expanding the multinomial coefficients yields the equivalent expression:

$$\prod_{i=1}^{r_1} b_i! \prod_{j=1}^{r_2} c_j! \prod_{i,j} v_{ij}!.$$

Since the vectors $b$ and $c$ are constants, this shows that the probability of the table $v$ is given by the hypergeometric distribution.

For the special case of reducible models, there is a general recursive strategy using the urn scheme that takes hypergeometric samples from the pieces of the decomposition and constructs a hypergeometric sample from the entire model.

Let $\Gamma$ be a reducible simplicial complex with reducible decomposition $(\Gamma_1, S, \Gamma_2)$, and let $u$ be a table. Let $V_i = [\Gamma_i \setminus S, i = 1, 2$. Let $u_i = \pi_{\Gamma_i \cup S}(u)$ be the projection onto the random variables appearing in $\Gamma_i$. Let $v_i$ be a sample from the hypergeometric distribution on the fiber $F_{\Gamma_i}(u_i)$. To generate a hypergeometric sample on $F(u)$, do the following.

For each index $i_S \in R_S$, use the urn scheme on $\#R_{\Gamma_1} \times \#R_{\Gamma_2}$ tables to generate a random hypergeometric table with row indices and column
indices equal to

\[(v_1)_{i_1,i_S} \in R_{i_1} \quad \text{and} \quad (v_2)_{i_2,i_S} \in R_{i_2}\]

respectively. For each \(i_S\) this produces a table which comes from the hypergeometric distribution and the combination of all of these tables will be a table from the hypergeometric distribution.

**Example 9.6.2** (Urn Scheme in a Decomposable Model). Consider the decomposable simplicial complex \(\Gamma = [12][23][345]\) and suppose that \(d = (2,3,4,5,6)\). To generate a random sample from the hypergeometric distribution on the fiber \(F_{\Gamma}(u)\) for this decomposable model, we proceed as follows. First we restrict to the subcomplex \(\Gamma_1 = [12][23]\), and we need to generate hypergeometric samples on the fiber \(F_{\Gamma_1}(u_{123})\). To do this, we must generate samples from the hypergeometric distribution on \(d_2 = 3 \times 4\) tables with fixed row and column sums. This leads to a random hypergeometric distributed sample \(v\) from \(F_{\Gamma_1}(u_{123})\). Now we proceed to the full model \(\Gamma\). To do this, we need to generate \(d_3 = 4\) random hypergeometric tables of size \(6 \times 30\), using the margins that come from \(v\) and \(u|_{345}\).

Direct sampling from the underlying distribution of interest on \(F(u)\) is usually not possible. This is the reason for the Markov chain approaches outlined in previous sections. An alternate Markov chain strategy, which was described in [HAT12], is based on simply having a spanning set for the lattice \(\ker ZA\) and using that to construct arbitrarily complicated moves and perform a random walk. Let \(B \subseteq \ker ZA\) be a spanning set of \(\ker Z\). Suppose that \(B = \{b_1, \ldots, b_d\}\) has \(d\) elements. Let \(P\) be a probability distribution on \(Z^d\) that is symmetric about the origin, \(P(m) = P(-m)\), and such that every vector in \(Z^d\) has positive probability. To get a random element in \(\ker ZA\), take a random draw \(m \in Z^d\) with respect to \(P\) and form the vector \(\sum m_i b_i\). This produces a random element of \(\ker ZA\), which can then be used to run the Markov chain. The fact that the distribution \(P\) is symmetric about the origin guarantees that the resulting Markov chain is irreducible, and then the Metropolis-Hastings algorithm can be employed to generate samples converging to the true distribution.

An alternate strategy for computing approximate samples that does not use Markov chains is sequential importance sampling, which works by building a random table entry by entry. This method does not produce a random table according to the hypergeometric distribution, but then importance weights are used to correct for the fact that we sample from the incorrect distribution. To generate a random table entry by entry requires to know upper and lower bounds on cell entries given table margins. We discuss this method further in Chapter 10.
Table 9.7.1. A contingency table for the study of the relation between trophic level and vegetable composition in rivers.

<table>
<thead>
<tr>
<th></th>
<th>(r, p, e)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0, 0, 0) (1, 0, 0) (0, 1, 0) (0, 0, 1) (1, 1, 0) (0, 1, 1)</td>
</tr>
<tr>
<td>Oligotrophic</td>
<td>0 0 3 0 3 2</td>
</tr>
<tr>
<td>Mesotrophic</td>
<td>2 1 0 2 1 0</td>
</tr>
<tr>
<td>Eutrophic</td>
<td>2 0 3 1 1 0</td>
</tr>
</tbody>
</table>

9.7. Exercises

Exercise 9.1. For the generalized hypergeometric distribution, write as simple an expression as possible for the fraction $p(u + v_t)/p(u_t)$ that is computed in the Metropolis-Hastings algorithm (i.e. give an expression that involves as few multiplications as possible).

Exercise 9.2. Prove Theorem 9.2.10 and Corollary 9.2.11.

Exercise 9.3. Let $A \in \mathbb{Z}^{k \times r}$ be an integer matrix with $I_A$ the corresponding toric ideal. Suppose that $A'$ is a matrix obtained as the subset of columns of $A$, indexed by the set of columns $S$.

1. Show that $I_{A'} = I_A \cap \mathbb{K}[p_i : i \in S]$.
2. Explain how the previous result can be used for Markov basis problems for models with structural zeros.
3. Compute a Markov basis for Hardy-Weinberg equilibrium on seven alleles with structural zeros in all the $(i, i + 1)$ positions.


Exercise 9.5. Consider the dataset displayed in Table 9.7.1. This table appears in [Fin10] which is extracted from the original ecological study in [TCHN07]. The dataset was used to search for a link between trophic levels and vegetable composition in rivers in northeast France. A river can be either oligotrophic, mesotrophic, or eutrophic if nutrient levels are poor, intermediate, or high, respectively. To each river, a triple of binary characters was associated $(r, p, e)$ indicating the presence or absence $(1/0)$ of rare $(r)$, exotic $(e)$, or pollution-tolerant $(p)$ species of vegetables. Two empty columns have been removed to leave a $3 \times 6$ table. Is there sufficient data to reject the null hypothesis that trophic level and vegetable composition of the rivers are independent?

Exercise 9.6. Let $\Gamma = [1][2][3]$ and $r = (2, 2, 2)$ be the model of complete independence for 3 binary random variables, and let $u$ be the $2 \times 2 \times 2$ table

\[
\begin{pmatrix}
16 & 4 \\
2 & 3
\end{pmatrix}
\quad \quad
\begin{pmatrix}
3 & 22 \\
11 & 6
\end{pmatrix}
\]
What is the expected value of the (1, 1, 1) coordinate of \( v \in F(u) \) selected with respect to the generalized hypergeometric distribution and the uniform distribution?

Exercise 9.7. Perform Fisher’s exact test on the \( 2 \times 2 \times 2 \) table
\[
\begin{pmatrix}
1 & 5 \\
3 & 5
\end{pmatrix}
\begin{pmatrix}
4 & 4 \\
9 & 6
\end{pmatrix}
\]
with respect to the model of complete independence, i.e. where \( \Gamma = \{1\}|2|\{3\} \).

Exercise 9.8. Show that the matrices \( A_\Gamma \) and \( B_\Gamma \) (defined after Proposition 9.3.10) have the same row span. Use this to show that the rank of \( A_\Gamma \) is given by (9.3.2).

Exercise 9.9. Let \( \Gamma \) be the simplicial complex \( \Gamma = \{1\}|2|\{3\} \) and \( r = (2, 2, 2, 4) \). Use the fact that \( \Gamma \) is reducible to give the Markov basis for \( A_{\Gamma, r} \).

Exercise 9.10. Let \( A \in \mathbb{Z}^{k \times r} \) and let \( \Lambda(A) \) denote the \( (r + k) \times 2r \) matrix
\[
\Lambda(A) = \begin{pmatrix}
A & 0 \\
I & I
\end{pmatrix}
\]
where \( I \) denotes the \( r \times r \) identity matrix. The matrix \( \Lambda(A) \) is called the Lawrence lifting of \( A \). Note that \( u \in \ker_\mathbb{Z}(A) \) if and only if \( (u, -u) \in \ker_\mathbb{Z}(\Lambda(A)) \). Show that a set of moves \( B \in \ker_\mathbb{Z}(A) \) is the Graver basis for \( A \) if and only if the set of moves \( \{(u, -u) : u \in B\} \) is a minimal Markov basis for \( \Lambda(A) \).

Exercise 9.11. One setting where Graver bases and Lawrence liftings naturally arise is in the context of logistic regressions. We have a binary response variable \( X \in \{0, 1\} \), and another random variable \( Y \in [r] \) for some \( r \). In the logistic regression model, we assume that
\[
\frac{P(X = 1 \mid Y = i)}{P(X = 0 \mid Y = i)} = \exp(\alpha + \beta i)
\]
for some parameters \( \alpha \) and \( \beta \).

(1) Show that with no assumptions on the marginal distribution on \( Y \), the set of joint distributions that is compatible with the logistic regression assumption is an exponential family whose design matrix can be chosen to have the form \( \Lambda(A) \) for some matrix \( A \) (see Exercise 9.10).

(2) Compute the Graver basis that arises in the case that \( r = 6 \).

(3) Use the logistic regression model to analyze the snoring and heart disease data in Table 9.7.2, which is extracted from a data set in [ND85]. Perform a hypothesis test to decide if the logistic regression fits this data. Note that in the original analysis of this data,
Table 9.7.2. A contingency table for the study of the relation between snoring and heart disease.

<table>
<thead>
<tr>
<th>Heart Disease</th>
<th>Never</th>
<th>Occasionally</th>
<th>Nearly every night</th>
<th>Every night</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>24</td>
<td>35</td>
<td>21</td>
<td>30</td>
</tr>
<tr>
<td>No</td>
<td>1355</td>
<td>603</td>
<td>192</td>
<td>224</td>
</tr>
</tbody>
</table>

the mapping of the states Never, Occasionally, Almost every night, and Every night to integers is Never = 1, Occasionally = 3, Almost every night = 5, and Every night = 6.


Exercise 9.13. Let $\Gamma = [12][23][34][45][15]$ be the five cycle and suppose that $r = (2, 2, 2, 2, 2)$.

(1) Describe all the Markov basis elements of $A_\Gamma$ which have 1-norm 4.

(2) Compute the primary decomposition of the ideal generated by the corresponding degree 2 binomials described in the previous part.

(3) Use the primary decomposition from the previous part to deduce that if $u$ satisfies $A_\Gamma u > 0$, then the fiber $F(u)$ is connected by the moves of 1-norm 4.

Exercise 9.14. Describe an urn scheme for generating hypergeometric samples for the model of Hardy-Weinberg equilibrium. (See [HCD+06] for an approach to calculate samples even more quickly.) Use both the Markov basis and the urn scheme to perform a hypothesis test for the Rhesus data (9.2.1).
Chapter 10

Bounds on Cell Entries

In this chapter we consider the problem of computing upper and lower bounds on cell entries in contingency tables given lower dimensional marginal totals. The calculation of lower and upper bounds given constraints is an example of an integer program. There is an established history of developing closed formulas for upper and lower bounds on cell entries given marginal totals \([DF00, DF01]\), but these typically only apply in special cases. We explain how computational algebra can be used to solve the resulting integer programs. While computational algebra is not the only framework for solving integer programs, this algebraic perspective also reveals certain results on toric ideals and their Gröbner bases that can be used to analyze the effectiveness of linear programming relaxations to approximate solutions to the corresponding integer programs. We begin the chapter by motivating the problem of computing bounds with two applications.

10.1. Motivating Applications

In this section, we consider motivating applications for computing upper and lower bounds on cell entries given marginal totals. The first application is to sequential importance sampling, a procedure for generating samples from fibers \(F(u)\) that avoids the difficulty of sampling from the hypergeometric distribution. A second application is to disclosure limitation, which concerns whether or not privacy is preserved when making certain data releases of sensitive data.

10.1.1. Sequential Importance Sampling. In Chapter 9 we saw strategies for sampling from the fiber \(F(u)\) to approximate the expected value \(\mathbb{E}_p[f(u)]\) where \(p\) is the hypergeometric distribution or uniform distribution
and $f$ is an indicator function pertaining to the test statistic of the particular hypothesis test we are using. Since generating direct samples from either the hypergeometric distribution or uniform distribution are both difficult, we might content ourselves with sampling from an easy-to-sample-from distribution $q$ on $\mathcal{F}(u)$, which has the property that the probability of any $v \in \mathcal{F}(u)$ is positive. Then we have

$$E_q[f(v)] = \sum_{v \in \mathcal{F}(u)} f(v) \frac{p(v)}{q(v)} \cdot q(v) = \sum_{v \in \mathcal{F}(u)} f(v) p(v) = E_p[f(v)].$$

So to approximate the expected value we are interested in computing random samples $v^1, \ldots, v^n$ from the distribution $q$ and calculate

$$\frac{1}{n} \sum_{i=1}^n f(v^i) \frac{p(v^i)}{q(v^i)}.$$ as an estimator of this expectation. This procedure is called importance sampling and the ratios $\frac{p(v^i)}{q(v^i)}$ are the importance weights.

A further difficulty in applying importance sampling for Fisher’s exact test, is that our target distribution $p$ on $\mathcal{F}(u)$ is the hypergeometric distribution on $\mathcal{F}(u)$ where we cannot even compute the value $p(v)$ for $v \in \mathcal{F}(u)$. This is because we usually only have the formula for $p$ up to the normalizing constant. Indeed, if $p'$ is the unnormalized version of $p$, so that $p(v) = p'(v) / \left( \sum_{w \in \mathcal{F}(u)} p'(w) \right)$, then for $n$ samples $v^1, \ldots, v^n$ according to the distribution $q$ the expression

$$\frac{1}{n} \sum_{i=1}^n \frac{p'(v^i)}{q(v^i)}.$$ converges to the normalizing constant, hence

$$\frac{\left( \sum_{i=1}^n f(v^i) \frac{p'(v^i)}{q(v^i)} \right)}{\left( \sum_{i=1}^n \frac{p'(v^i)}{q(v^i)} \right)}$$ converges to the mean of the distribution $p$.

The question remains: how should we choose the distribution $q$? The sequential part of sequential importance sampling relies on the fact that the distribution $p$ is usually over some high dimensional sample space. The strategy is then to compute a distribution $q$ by sequentially filling in each entry of the vector we intend to sample. In the case of sampling from a fiber $\mathcal{F}(u)$, we want to fill in the entries of the vector $v \in \mathcal{F}(u)$ one at a time. To do this, we need to know what the set of all possible first coordinates of all $v \in \mathcal{F}(u)$ are. This can be approximated by computing upper and lower bounds on cell entries.
Example 10.1.1 (Sequential importance sampling in a $3 \times 3$ table). Consider generating a random $3 \times 3$ nonnegative integer array with fixed row and column sums as indicated in the following table.

<table>
<thead>
<tr>
<th>$v_{11}$</th>
<th>$v_{12}$</th>
<th>$v_{13}$</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_{21}$</td>
<td>$v_{22}$</td>
<td>$v_{23}$</td>
<td>3</td>
</tr>
<tr>
<td>$v_{31}$</td>
<td>$v_{32}$</td>
<td>$v_{33}$</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

We compute upper and lower bounds on entry $v_{11}$ and see that $v_{11} \in \{0, 1, 2, 3\}$. Suppose we choose $v_{11} = 2$, and we want to fill in values for $v_{12}$. We compute upper and lower bounds on $v_{12}$ given that $v_{11} = 2$ and see that $v_{12} \in \{1, 2, 3, 4\}$, and we might choose $v_{12} = 3$. At this point, $v_{13}$ is forced to be 1 by the row sum constraints. We can proceed to the second row by treating this as a random table generation problem for a $2 \times 3$ table. Continuing in this way, we might produce the table

<table>
<thead>
<tr>
<th>2</th>
<th>3</th>
<th>1</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Assuming that when we have a range of values to choose from at each step, we pick them uniformly at random, the probability of selecting the table $v$ above is $\frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{2} \cdot \frac{1}{3}$.

A key step in the sequential importance sampling algorithm for contingency tables is the calculation of the upper and lower bounds on cell entries. That is, we need to calculate, for example,

$$ \max / \min v_{11} \quad \text{subject to} $$

$$ \sum_{j=1}^{r_2} v_{ij} = u_{i+} \text{ for } i \in [r_1], \quad \sum_{i=1}^{r_1} v_{ij} = u_{+j} \text{ for } j \in [r_2] \text{ and } v \in \mathbb{N}^{r_1 \times r_2}. $$

For 2-way tables, such bounds are easy to calculate given the marginal totals.

Proposition 10.1.2. Let $u \in \mathbb{N}^{r_1 \times r_2}$ be a nonnegative integral table. Then for any $v \in \mathcal{F}(u)$ we have

$$ \max(0, u_{i+} + u_{+j} - u_{++}) \leq v_{ij} \leq \min(u_{i+}, u_{+j}) $$

and these bounds are tight. Furthermore, for every integer $z$ between then upper and lower bounds, there is a table $v \in \mathcal{F}(u)$ such that $v_{ij} = z$.

Proof. Clearly the upper bound is valid, since all entries of the table are nonnegative. To produce a table that achieves the upper bound, we use a
procedure that is sometimes referred to as "stepping-stones" or the "north-west corner rule". Without loss of generality, we can focus on $v_{11}$. We prove by induction on $r_1$ and $r_2$, that there is a table whose 11 entry is $\min(u_{1+}, u_{+1})$. Suppose that the minimum is attained by the row sum $u_{1+}$. Set $v_{11} = u_{1+}$ and $v_{1j} = 0$ for all $j = 2, \ldots, r_2$. Then we can restrict attention to the bottom $r_1 - 1$ rows of the matrix. Updating the column sums by replacing $u_{+j} = u_{1+} - u_{1j}$, we arrive at a smaller version of the same problem. Since the row and columns sums of the resulting $(r_1 - 1) \times r_2$ matrix are consistent, induction shows that there exists a way to complete $v$ to a nonnegative integral table satisfying the original row and columns and with $v_{11}$ equal to the upper bound. The procedure is called stepping-stones because if you follow it through from the 11 entry down to the bottom produces a sequence of nonnegative entries in the resulting table $v$.

For the lower bound, we note an alternate interpretation of this formula. Suppose that in a given row $i$ we were to put the upper bound value $u_{+k}$ in each entry $u_{jk}$ for $k \neq j$. Then this forces

$$u_{ij} = u_{i+} - \sum_{k \neq j} u_{+k} = u_{i+} + u_{+j} - u_{++}$$

which is a lower bound since we made the other entries in the $i$th row as large as possible. This lower bound is attainable (and nonnegative) precisely when all the $u_{+k}$ for $k \neq j$ are smaller than $u_{i+}$, which again follows from the stepping-stones procedure. Otherwise the lower bound of 0 is attained by putting some $u_{ik} = u_{i+}$.

To see that every integer value between the bounds is attained by some table $v$ in the fiber, note that the Markov basis for 2-way tables with fixed row and columns sums consists only of moves with 0, ±1 entries. Hence in a connected path from a table that attains the upper bound to one that attains the lower bound we eventually see each possible entry between the upper and lower bounds.

Example 10.1.3 (Schizophrenic Hospital Patients). We apply the basic strategy described above to the 2-way table from Example 9.1.2. Generating 1000 samples from the proposal distribution generates an approximation to the Fisher’s exact test $p$-value of $1.51 \times 10^{-7}$.

Note however that this basic sequential importance sampling method can be problematic in more complicated problems, because it does not put enough probability mass on the small proportion of tables that have high probability under the hypergeometric distribution. For this reason, sequential importance sampling has been primarily used to approximate samples
from the uniform distribution where that is appropriate. One example application is to use sequential importance sampling to approximate the cardinality of the fiber $\mathcal{F}(u)$. Using the arguments above, and taking $p$ to be the uniform distribution on $\mathcal{F}(u)$, it is not difficult to see that the estimator

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{q(v^{(i)})}$$

converges to $|\mathcal{F}(u)|$ as the number of samples $n$ tends to infinity. Using the same 1000 samples above, we estimate $|\mathcal{F}(u)| \approx 2.6 \times 10^5$.

Further details about applications of sequential importance sampling to the analysis of tables appear in [CDHL05] where the focus is on 2-way tables, with special emphasis on 0/1 tables. The paper [CDS06] considers the extensions to higher-order tables under fixed margins, using some of the tools discussed in this chapter.

Computing bounds on cell entries in 2-way tables is especially easy. In general, it is difficult to find general formulas for the bounds on cell entries given marginal totals. We discuss this in more detail in Section 10.5. It remains a problem to compute bounds on tables of larger sizes with respect to more complex marginals, or for more general log-linear models besides hierarchical models. General easy to compute formulas are lacking and usually algorithmic procedures are needed, which we will see in subsequent sections.

10.1.2. Disclosure Limitation. Computing bounds on entries of contingency tables can also be useful in statistical disclosure limitation or other areas of data privacy. Census bureaus and social science researchers often collect data via surveys with a range of questions including various demographic questions, questions about income and job status, health care information, and personal opinions. Often there are sensitive questions mixed in among a great many other innocuous questions.

To assure that the survey respondents give truthful answers, such surveys typically offer guarantees to the participants that individual personal information will not be revealed, that respondent files will not be linked with names, and that the data will only be released in aggregate. In spite of such guarantees from the survey collectors, the summary of data in a cross-classified contingency table together with the knowledge that a particular person participated in the survey can often be enough to identify sensitive details about individuals, especially when those people belong to minority groups.
Table 10.1.1. A hypothetical table of survey responses by an academic department of 31 tenured/tenure-track faculty.

<table>
<thead>
<tr>
<th></th>
<th>Excellent</th>
<th>Neutral</th>
<th>Terrible</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L M H</td>
<td>L M H</td>
<td>L M H</td>
</tr>
<tr>
<td>Male</td>
<td>Asst</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Assoc</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Full</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Male</td>
<td>Assoc</td>
<td>1 1/2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Full</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Female</td>
<td>Asst</td>
<td>1</td>
<td>1 1/2 1/2</td>
</tr>
<tr>
<td></td>
<td>Assoc</td>
<td>1</td>
<td>1 1/2 1/2</td>
</tr>
<tr>
<td></td>
<td>Full</td>
<td>1</td>
<td>1 1/2 1/2</td>
</tr>
</tbody>
</table>

To illustrate the problem, consider the following hypothetical contingency table cross-classifying a group of mathematics professors at a particular university according to rank (R), gender (G), sexual orientation (SO), income (I), and opinion about university administration (UA). Possible answers for these survey questions are

- Rank: Assistant Professor, Associate Professor, Full Professor,
- Gender: Male, Female
- Sexual Orientation: Heterosexual, Gay/Lesbian/Bisexual
- Income: Low, Medium, High
- Opinion: Excellent job, Neutral, Terrible job

The table shows the results of the hypothetical survey of a department of 31 faculty. For example the 1 entry denotes the one female, lesbian or bisexual, associate professor, in the middle income range, and neutral about the university administration. As should be clear from this example, even with only 5 survey questions, the resulting tables are very sparse, and with many singleton entries. Table observers with access to public information (e.g. a departmental website, social media, etc.) can combine this information to deduce sensitive information, for example, sexual orientation or the faculty member’s opinion on the administration. Research in disclosure limitation is designed to analyze what data can be released in anonymized form without revealing information about the participants.
If we focus only on the information contained in the table, we should be concerned about any entry that is a 1 or a 2, as these are considered the most sensitive. A 1 means a person is uniquely identified, and a 2 means that two participants can identify each other. One easy to implement strategy to preserve the knowledge of the exact table values in a large table is to only release lower dimensional marginals of the tables with no ones or twos. However, an agency or researcher releasing such information must be sure that this release does not further expose cell entries in the table. “How much information about \( u \) can be recovered from the released margins \( A_{\Gamma} u \), for some particular choice of margins \( \Gamma \)”, is a typical question in this area. The most naive way to assess this is to compute bounds on individual cell entries given the marginal \( A_{\Gamma} u \). If the bounds are too narrow and the lower bounds are positive, then they probably reveal too much about the cell entry.

For example, in Table 10.1.1, the release of all 3-way marginals of the table does not mask the table details at all: in fact, it is possible to recover all table entries given all 3-way margins in this case! If we restrict to just 2-way marginals, then by computing linear programming upper and lower bounds, we are uniquely able to recover one of the table entries namely the position marked by 1. This example shows that even releasing quite low dimensional marginals on a 5-way table that is sparse can still reveal table entries that are sensitive.

The issue of privacy from databases is a very active research area, and the problem of computing bounds on entries represents only a tiny part. Perhaps the most widely studied framework is called differential privacy [Dwo06]. A survey of statistical aspects of the disclosure limitation problem is [FS08].

10.2. Integer Programming and Gröbner Bases

To discuss computing bounds in more detail requires some introduction to the theories of integer and linear programming. This is explained in the present section which also makes the connection to Gröbner bases of toric ideals. More background on linear and integer programming can be found in [BT97], and the book [DLHK13] has substantial material on applications of algebra to optimization.

Let \( A \in \mathbb{Z}^{k \times r}, b \in \mathbb{Z}^k \) and \( c \in \mathbb{Q}^r \). The standard form integer program is the optimization problem

\[
\text{min } c^T x \text{ subject to } Ax = b \text{ and } x \in \mathbb{N}^r.
\]

The optimal value of the integer program will be denoted \( IP(A, b, c) \). The linear programming relaxation of the integer program is the optimization problem

\[
\text{min } c^T x \text{ subject to } Ax = b, x \geq 0 \text{ and } x \in \mathbb{R}^r.
\]
Its optimal value will be denoted $LP(A, b, c)$. Since the optimization in the linear programming relaxation is over a larger set than in the integer program, $LP(A, b, c) \leq IP(A, b, c)$.

For general $A$, $b$, and $c$, integer programming is an NP-hard problem [Kar72] while linear programs can be solved in polynomial time in the input complexity (e.g. via interior point methods) [Hac79, Kar84]. Since integer programs are difficult to solve in general, there are a range of methods that can be used. Here we discuss methods based on Gröbner bases. In the optimization literature these are closely related to optimization procedures based on test sets (see [DLHK13]), and also related to the theory of Markov bases from Chapter 9.

Recall that $F(u) = \{v \in \mathbb{N}^r : Au = Av\}$ is the fiber in which $u$ lies. This is the set of feasible points of the integer program (10.2.1) when $b = Au$. Let $B \subseteq \ker Z A$ be a set of moves and construct the directed graph $F(u)_B$ whose vertex set is $F(u)$ and with an edge $v \to w$ if $c^T v \geq c^T w$ and $v - w \in \pm B$.

To use the directed fiber graph $F(u)_B$ to solve integer programs, we need to first identify some key properties that these graphs can have. A directed graph is strongly connected if there is a directed path between any pair of vertices. A strongly connected component of a directed graph is a maximal induced subgraph that is strongly connected. A sink component of a strongly connected graph is a strongly connected component that has no outgoing edges.

**Definition 10.2.1.** Let $A \in \mathbb{Z}^{k \times r}$ and $c \in \mathbb{Q}^r$. Assume that $\ker A \cap \mathbb{R}_{\geq 0}^r = \{0\}$ (so that $F(u)$ is finite for every $u$). A subset $B \subseteq \ker Z A$ is called a Gröbner basis for $A$ with respect to the cost vector $c$ if for each nonempty fiber $F(u)$, the directed fiber graph $F(u)_B$ has a unique sink component.

**Example 10.2.2 (A Connected Fiber Graph but Not a Gröbner Basis).**

Let $A = \begin{pmatrix} 2 & 3 \\ 3 & 5 \end{pmatrix}$, $c = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, and $B = \left\{ \begin{pmatrix} 3 \\ -2 \\ 0 \end{pmatrix}, \begin{pmatrix} 5 \\ 0 \\ -2 \end{pmatrix}, \begin{pmatrix} 0 \\ 5 \\ -3 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \right\}$

The fiber $F(u)$ consists of seven elements. There is a unique sink component consisting of the two elements $\{(0, 5, 0)^T, (0, 0, 3)^T\}$. However, the set $B$ is not a Gröbner basis for $A$ with respect to the cost vector $c$. For example, the fiber $F((1, 0, 2)^T)$ consists of the two elements $\{(1, 0, 2)^T, (0, 4, 0)^T\}$ which is disconnected, in particular it has two sink components.

The existence of a single sink component guarantees that the greedy search algorithm terminates at an optimum of the integer program (10.2.1). We use the term Gröbner basis here, which we have already used in the
10.2. Integer Programming and Gröbner Bases

...context of computing in polynomial rings previously. There is a close connection.

**Definition 10.2.3.** Let $c \in \mathbb{Q}^r$. The weight order $\prec_c$ on $\mathbb{K}[p_1, \ldots, p_r]$ is defined by $p^u \prec_c p^v$ if $c^T u < c^T v$. The weight of a monomial $p^u$ is $c^T u$.

Note that under the weight order, it can happen that two or more monomials all have the same weight. For a polynomial $f = \sum_{u \in \mathbb{N}^r} c_u p^u$ let $\text{in}_c(f)$ be the sum of all terms in $f$ of the highest weight. As usual, the initial ideal in $c(I)$ is the ideal generated by all initial polynomials of all polynomials in $I$, $\text{in}_c(I) = \langle \text{in}_c(f) : f \in I \rangle$. A weight order $\prec_c$ is said to be generic for an ideal $I$ if $\text{in}_c(I)$ is a monomial ideal.

**Theorem 10.2.4.** Let $A \in \mathbb{Z}^{k \times r}$, $\ker A \cap \mathbb{R}_{\geq 0} = \{0\}$, $c \in \mathbb{Q}^r$, and suppose that $c$ is generic for $I_A$. Then $B \subseteq \ker_2 A$ is a Gröbner basis for $A$ with respect to $c$ if and only if the set of binomials $\{p^{b^+} - p^{b^-} : b \in B\}$ is a Gröbner basis for $I_A$ with respect to the weight order $\prec_c$.

**Proof.** Let $p^u$ be a monomial and suppose we apply one single step of polynomial long division with respect to a binomial $p^{b^+} - p^{b^-}$ in the collection $\{p^{b^+} - p^{b^-} : b \in B\}$ assuming that $p^{b^+}$ is the leading term. A single step of long division produces the remainder $p^{u-b}$ which must be closer to optimal than $p^u$. In the fiber graph $F(u)_B$ we should connect $u$ and $u - b$ with a directed edge. Since $\{p^{b^+} - p^{b^-} : b \in B\}$ is a Gröbner basis we end this process with a unique end monomial $p^v$. The point $v$ must be the unique sink in $F(u)_B$, else there is a binomial $p^{b^+} - p^{b^-}$ where neither term is divisible by any leading term in $\{p^{b^+} - p^{b^-} : b \in B\}$, contradicting the Gröbner basis property.

In the case that $c$ is not generic, it is still possible to use Gröbner bases to solve integer programs. To do this, we choose an arbitrary term order $\prec'$, and use this as a tie breaker in the event that $c^T u = c^T v$. In particular, we define the term order $\prec_c'$ by the rule

$$p^u \prec_c' p^v \quad \text{if} \quad \begin{cases} c^T u < c^T v, \text{ or} \\ c^T u = c^T v \text{ and } p^u \prec' p^v. \end{cases}$$

**Theorem 10.2.5.** Let $A \in \mathbb{Z}^{k \times r}$, $\ker A \cap \mathbb{R}_{\geq 0} = \{0\}$, $c \in \mathbb{Q}^r$, and $\prec'$ an arbitrary term order. Let $B \subseteq \ker_2 A$. If $\{p^{b^+} - p^{b^-} : b \in B\}$ is a Gröbner basis for $I_A$ with respect to the weight order $\prec_c'$, then $B$ is a Gröbner basis for $A$ with respect to $c$.

The proof of Theorem 10.2.5 follows that of Theorem 10.2.4 and is left as an exercise. Note, in particular, that Theorems 10.2.4 and 10.2.5 tell us...
Bounds on Cell Entries

exactly how to solve an integer program using Gröbner bases. The normal form of \( p^u \) modulo a Gröbner basis of \( I_A \) with respect to the weight order \( \prec \) will be a monomial \( p^v \) that is minimal in the monomial order. This result and the Gröbner basis perspective on integer programming originated in [CT91, Tho95].

Note that for the optimization problems we consider here associated with computing upper and lower bounds on cell entries, the cost vector \( c \) is a standard unit vector or its negative. These cost vectors are usually not generic with respect to a toric ideal \( I_A \). However, there are term orders that can be used that will do the same job from the Gröbner basis standpoint.

**Proposition 10.2.6.** Let \( A \in \mathbb{Z}^{k \times r} \) with \( \text{ker} \ A \cap \mathbb{R}_{>0} = \{0\} \). Let \( \prec_{\text{lex}} \) be the lexicographic order with \( p_1 \succ_{\text{lex}} p_2 \succ_{\text{lex}} \cdots \) and let \( \prec_{\text{revlex}} \) be the reverse lexicographic order with \( p_1 \prec_{\text{revlex}} p_2 \prec_{\text{revlex}} \cdots \).

1. The normal form of \( p^u \) with respect to the toric ideal \( I_A \) and the term order \( \prec_{\text{lex}} \) is a monomial \( p^v \) where \( v_1 \) is minimal among all \( v \in F(u) \).

2. The normal form of \( p^u \) with respect to the toric ideal \( I_A \) and the term order \( \prec_{\text{revlex}} \) is a monomial \( p^v \) where \( v_1 \) is maximal among all \( v \in F(u) \).

**Proof.** We only prove the result for the lexicographic order, since the proof for the reverse lexicographic order is similar. Suppose that \( p^u \) is the normal form of \( p^w \), but \( w \in F(u) \) has smaller first coordinate then \( v \). Then \( p^u - p^w \) is in \( I_A \) and the leading term with respect to the lex order is \( p^v \) since \( v_1 > w_1 \). But this contradicts the fact the \( p^u \) is the normal form, and hence not divisible by any leading term of any element of the Gröbner basis.

**Example 10.2.7.** Continuing Example 10.2.2, note that since we were using the cost vector \( c = (1, 0, 0) \), we can compute a Gröbner basis for \( I_A \) with respect to the lexicographic order to solve the resulting integer program. The Gröbner basis consists of the following five binomials

\[
p_2^2 - p_3^2, p_1 p_2^2 - p_2^3, p_1 p_2 - p_3, p_2^2 p_3 - p_3^2, p_3^2 - p_2^2
\]

from which we see that the following list of vectors is a Gröbner basis for solving the associated integer programs

\[
\left\{ \begin{pmatrix} 0 \\ 5 \\ -3 \end{pmatrix}, \begin{pmatrix} 1 \\ -4 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ -2 \\ 0 \end{pmatrix} \right\}.
\]
10.3. Quotient Rings and Gröbner Bases

A key feature of Gröbner bases is that they give a well-defined computational strategy for working with quotient rings of polynomial rings modulo ideals. Performing division with respect to a Gröbner basis chooses a distinct coset representative for each coset, namely the normal form under polynomial long-division. In the case of toric ideals, the unique coset representative of a coset of the form \( p^n + IA \) also gives the solution to the corresponding integer programming problem. Structural information about the set of standard monomials can also give information about how well linear programming relaxations perform, which we discuss in the next section. Computing with quotient rings also plays an important role in the design of experiments, covered in detail in Chapter 12.

Let \( R \) be a ring, \( I \) an ideal of \( R \), and \( f \in R \). The coset of \( f \) is the set \( f + I := \{ f + g : g \in I \} \). The set of cosets of the ideal \( I \) form a ring with addition and multiplication defined by:

\[
(f + I) + (g + I) = (f + g) + I \quad \text{and} \quad (f + I)(g + I) = fg + I.
\]

The property of being an ideal guarantees that these operations do not depend on which element of the coset is used in the calculation. The set of all cosets under these operations is denoted by \( R/I \) and is called the quotient ring. One way to think about quotients is that the set of elements \( I \) are all being “set to zero”. Or, we can think about the elements in \( I \) as being new relations that we are adding to the defining properties of our ring. The following is a key example of a quotient ring arising from algebraic geometry.

**Proposition 10.3.1.** Let \( V \subseteq \mathbb{K}^n \) be an affine variety. The quotient ring \( \mathbb{K}[p]/I(V) \) is called the coordinate ring of \( V \) and consists of all polynomial functions defined on \( V \).

**Proof.** Two polynomials \( f, g \in \mathbb{K}[p] \) define the same function on \( V \) if and only if \( f - g = 0 \) on \( V \), that is, \( f - g \in I(V) \). But \( f - g \in I(V) \) if and only if \( f + I(V) = g + I(V) \).

At first glance, it might not be obvious how to perform computations using quotient rings. For instance, how can we check whether or not \( f + I = g + I \)? One solution is provided by Gröbner bases.

**Proposition 10.3.2.** Let \( I \subseteq \mathbb{K}[p] \) be an ideal, and \( \mathcal{G} \) a Gröbner basis for \( I \) with respect to a term order \( \prec \) on \( \mathbb{K}[p] \). Two polynomials \( f, g \in \mathbb{K}[p] \) belong to the same coset of \( I \) if and only if \( NF_\mathcal{G}(f) = NF_\mathcal{G}(g) \).
234

10. Bounds on Cell Entries

Proof. If $NF_G(f) = NF_G(g)$ then $f - g \in I$ (this is true, whether or not $G$ is a Gröbner basis). On the other hand, if $NF_G(f) \neq NF_G(g)$ then

$$NF_G(f - g) = NF_G(NF_G(f) - NF_G(g)) = NF_G(f) - NF_G(g) \neq 0.$$ 

By Theorem 3.3.9, $f - g \notin I$, so $f$ and $g$ are not in the same coset. □

Definition 10.3.3. Let $M$ be a monomial ideal. The standard monomials of $M$ are the monomials $p^u \in \mathbb{K}[p]$ such that $p^u / 2 M$.

Proposition 10.3.2 guarantees that the normal form can be used as a unique representative as elements of a coset. Note that the only monomials that can appear in the normal form $NF_G(f)$ are the standard monomials of $in_{\omega}(I)$. Clearly, any polynomial whose support is in only the standard monomials of $in_{\omega}(I)$ will be a normal form. This proves:

Proposition 10.3.4. For any ideal $I \subseteq \mathbb{K}[p]$, and any monomial order $\prec$, the set of standard monomials of $in_{\omega}(I)$ form a $\mathbb{K}$-vector space basis for the quotient ring $\mathbb{K}[p]/I$.

The structure of the set of standard monomials of an initial ideal $in_{\omega}(I)$ is particularly interesting in the case of toric ideals, as it relates to integer programming.

Proposition 10.3.5. Let $A \in \mathbb{Z}^{k \times r}$ be an integer matrix and $c \in \mathbb{Q}^r$ be a generic cost vector with respect to the toric ideal $I_A \subseteq \mathbb{K}[p]$. The standard monomials $p^u$ of the initial ideal $in_c(I_A)$ are the optimal points of all integer programs (10.2.1) for $b \in NA$.

Proof. If $c$ is generic then the integer program (10.2.1) has a unique solution for each $b \in NA$. For a fixed $b$ call this optimum $u$. Computing the normal form of any monomial $p^v$ where $v \in F(u)$ with respect to the Gröbner basis of $I_A$ under the term order $\prec_c$ produces the monomial $p^u$ which is a standard monomial. This shows that every monomial $p^v$ is either in $in_c(I_A)$ or is optimal. □

Note that for generic $c$, Proposition 10.3.5 gives a bijection between the elements of the semigroup $NA$ and the standard monomials of $in_c(I_A)$. In the nongeneric case, each fiber might have more than one optimum. A tie-breaking term order $\prec'_c$ can be used to choose a special optimum element in each fiber.

10.4. Linear Programming Relaxations

Integer programming is generally an NP-hard problem so it is unlikely that there is a polynomial time algorithm that solves general integer programs. In the applications in Sections 10.1.1 and 10.1.2, we might be only interested
in some bounds we can compute quickly without worrying about whether or not they are tight all the time. In Section 10.5, we explore the possibility of developing general closed formulas for bounds on cell entries. In this section, we focus on a general polynomial time algorithm to get rough bounds, namely the computation of linear programming bounds as an approximation to integer programming bounds.

For given $A$, $b$ and $c$ recall that $IP(A, b, c)$ is the optimal value of the integer program (10.2.1) and $LP(A, b, c)$ is the optimal value of the linear program (10.2.2). Linear programs can be solved in polynomial time using interior point methods, so we might hope that perhaps for the bounds problem that arises in the analysis of contingency tables, that $IP(A, b, c) = LP(A, b, c)$ for all $b$ for the particular $A$ and $c$ that arise in contingency table problems. There is a precise algebraic way to characterize when this equality happens for all $b$ in the case of generic cost vectors $c$.

**Theorem 10.4.1.** Let $A \in \mathbb{Z}^{k \times r}$ and $c \in \mathbb{Q}^r$ generic with respect to the toric ideal $I_A$. Then $LP(A, b, c) = IP(A, b, c)$ for all $b \in \mathbb{N}A$ if and only if $in_c(I_A)$ is a squarefree monomial ideal.

**Proof.** Suppose that $in_c(I_A)$ is not a squarefree monomial ideal and let $p^u$ be a minimal generator of $I_A$ that has some exponent greater than one. We may suppose $u_1 > 1$. Since $p^u$ is a minimal generator of $in_c(I_A)$, $p^{u-e_1}$ is a standard monomial of $in_c(I_A)$. In particular, the vector $u - e_1$ is optimal for the integer program in its fiber. Now let $v \in \mathbb{N}^r$ so that $p^u - p^v$ is in the Gröbner basis of $I_A$ with respect to $\prec_c$ with $p^u$ as leading term. Since $c$ is generic, $c^T u < c^T v$ which means that $(u - e_1) - \frac{w_{e_1}}{w_{e_1}}(u - v)$ is a nonnegative vector rational vector which has a better cost than $u - e_1$. Hence, $LP(A, b, c)$ does not equal $IP(A, b, c)$ for the corresponding vector $b = A(u - e_1)$.

Conversely, suppose that $in_c(I_A)$ is a squarefree monomial ideal such that $v$ is LP optimal, $u$ is IP optimal, and $c^T v < c^T u$. For this to happen $v$ must have rational coordinates. Let $D$ be the least common multiple of the denominators of the nonzero coordinates of $v$. Consider the linear and integer programs where $b$ has been replaced by $Db$. Clearly $Dv$ is the optimum for the LP since it continues to be a vertex of the corresponding polytope. Since we cleared denominators, it is also an integer vector so is IP optimal. This means that $Du$ is not IP optimal. So there exists a binomial $p^w - p^{w'}$ in the Gröbner basis such that $Du - w + w' \geq 0$. Since the cost vector is generic, $Du - w + w'$ is IP feasible and must have improved cost. However, $w$ is a zero/one vector. This forces that $u \geq w$ coordinatewise. This implies that $u - w + w' \geq 0$ as well and since $c$ is generic, $u - w + w'$ also has improved cost. This contradicts our assumption that $u$ was IP optimal. $$\square$$
For the minimization and maximization of a cell entry given marginals the cost vectors $c$ are not generic. There is an analogous statement to Theorem 10.4.1 for the lexicographic and reverse lexicographic term orders.

**Proposition 10.4.2.** Let $A \in \mathbb{Z}^{k \times r}$ with $1 \in \text{rowspan}(A)$. Let $\prec_{\text{lex}}$ be the lexicographic order with $p_1 \succ_{\text{lex}} p_2 \succ_{\text{lex}} \cdots$ and let $\prec_{\text{revlex}}$ be the reverse lexicographic order with $p_1 \prec_{\text{revlex}} p_2 \prec_{\text{revlex}} \cdots$.

1. If $\text{in}_{\text{lex}}(I_A)$ is a squarefree monomial ideal then $LP(A, b, e_1) = IP(A, b, e_1)$ for all $b \in N A$.

2. If $\text{in}_{\text{revlex}}(I_A)$ is a squarefree monomial ideal then $LP(A, b, -e_1) = IP(A, b, -e_1)$ for all $b \in N A$.

The proof combines the arguments of the proofs of Theorem 10.4.1 and Proposition 10.2.6 and is left as an exercise. Note that whether or not $\text{in}_{\text{lex}}(I_A)$ or $\text{in}_{\text{revlex}}(I_A)$ are squarefree monomial ideals might depend on the precise ordering of the variables that are taken.

**Example 10.4.3** (LP Relaxation for 3-way Tables). Consider the matrix $A_{\Gamma}$ for $\Gamma = [1][2][3]$ and $r_1 = r_2 = r_3 = 2$. Let $\prec_{\text{lex}1}$ be the lexicographic order with

$$p_{111} \succ_{\text{lex}1} p_{112} \succ_{\text{lex}1} p_{121} \succ_{\text{lex}1} p_{122} \succ_{\text{lex}1} p_{211} \succ_{\text{lex}1} p_{212} \succ_{\text{lex}1} p_{221} \succ_{\text{lex}1} p_{222}. $$

Then $\text{in}_{\text{lex}1}(I_{A_{\Gamma}})$ is a squarefree monomial ideal:

$$\text{in}_{\text{lex}1}(I_{A_{\Gamma}}) = \langle p_{211}p_{222}, p_{121}p_{222}, p_{121}p_{212}, p_{112}p_{222}, p_{112}p_{212}, p_{112}p_{122}, p_{111}p_{222}, p_{111}p_{212}, p_{111}p_{122} \rangle. $$

On the other hand, with respect to the lexicographic order with

$$p_{111} \succ_{\text{lex}2} p_{112} \succ_{\text{lex}2} p_{121} \succ_{\text{lex}2} p_{122} \succ_{\text{lex}2} p_{211} \succ_{\text{lex}2} p_{212} \succ_{\text{lex}2} p_{221} \succ_{\text{lex}2} p_{222} $$

the initial ideal $\text{in}_{\text{lex}2}(I_{A_{\Gamma}})$ is not squarefree, since

$$\text{in}_{\text{lex}2}(I_{A_{\Gamma}}) = \langle p_{221}^2 p_{112}, p_{212}^2 p_{121}, p_{212} p_{221}, p_{121} p_{222}, p_{121} p_{122}, p_{111} p_{212}, p_{111} p_{222}, p_{111} p_{122}, p_{111} p_{112} \rangle. $$

Although the second Gröbner basis does not reveal whether or not the linear programming relaxation solves the integer programs (since we consider a nongeneric cost vector), the first Gröbner basis shows that in fact linear programming does solve the integer program in this case.

More generally, for a nongeneric cost vector $c$ and an ideal $I_A$, we can define a refinement of $\prec_c$ to be any term order $\prec$ such that $\text{in}_{\prec}(p^u - p^v) = \text{in}_{\prec_c}(p^u - p^v)$ for all $p^u - p^v \in I_A$ such that $\text{in}_{\prec}(p^u - p^v)$ is a monomial. (For example, the tie breaking term orders $\prec'_c$ are refinements of $\prec_c$.) Proposition
10.4.2 can be generalized to show that if $\prec$ is a refinement of the weight order $\prec_c$ then if $\text{in}_{\prec_c}(I_A)$ is a squarefree monomial ideal then $LP(A, b, c) = IP(A, b, c)$ for all $b \in \mathbb{N}A$.

Example 10.4.3 shows that we might need to investigate many different lexicographic term orders before we find one that gives a squarefree initial ideal. On the other hand, a result of [Sul06] shows that toric ideals that are invariant under a transitive symmetry group on the variables either have every reverse lexicographic initial ideal squarefree or none are, and that this is equivalent to linear programming solving the integer programs for the upper bounds on coordinates problem. Since this symmetry property is satisfied for the toric ideals $I_{A_r}$ of hierarchical models, one can check for the existence of a squarefree reverse lexicographic initial ideal for these models in a single ordering of the variables.

In general, the problem of classifying which $A$ and $r = (r_1, \ldots, r_m)$ have $I_{A_r}$ possessing squarefree lexicographic and/or reverse lexicographic initial ideal is an interesting open problem. One partial result is the following from [Sul06].

**Theorem 10.4.4.** Let $\Gamma$ be a graph and suppose that $r = 2$, so we consider a binary graph model. Then $I_{A_r}$ has a squarefree reverse lexicographic initial ideal if and only if $\Gamma$ is free of $K_4$ minors and every induced cycle in $\Gamma$ has length $\leq 4$.

Extending this result to arbitrary simplicial complexes and also to lexicographic term orders is an interesting open problem. Computational results appear in [BS17a].

In the other direction, one can ask how large the difference between $IP(A, b, c)$ and $LP(A, b, c)$ can be for these cell entry bound problems, in cases where the linear programming relaxations do not always solve the integer program. A number of negative results are known producing instances with exponential size gaps between the LP relaxations and the true IP bounds for lower bounds [DLO06, Sul05]. These examples for gaps on the lower bounds can often be based on the following observation.

**Proposition 10.4.5.** Let $A \in \mathbb{Z}^{k \times r}$ and suppose that $F(u) = \{u, v\}$ is a two element fiber with $\text{supp}(u) \cap \text{supp}(v) = \emptyset$. Then $p^u - p^v$ belongs to every minimal binomial generating of $I_A$ and every reduced Gröbner basis of $I_A$. Furthermore, if $u_1 > 1$ then

$$IP(A, A(u - e_1), e_1) - LP(A, A(u - e_1), e_1) = u_1 - 1.$$ 

**Proof.** The fact that $p^u - p^v$ is in any binomial generating set follows from the Markov basis characterization of generating sets: the move $u - v$ is
the only one that can connect this fiber. A similar argument holds for the Gröbner basis statement.

For the statement on gaps, note the \( \mathcal{F}(u - e_1) = \{u - e_1\} \) consists of a single element, so the minimal value for the first coordinate in the fiber is \( u_1 - 1 \). However, the nonnegative real table \( u - e_1 - \frac{u_1 - 1}{u_1}(u - v) \) has first coordinate zero.

Example 10.4.6 (2-way Margins of a 5-way Table). Let \( \Gamma = K_5 \) be the simplicial complex on \([5]\) whose facets are all pairs of vertices, and let \( r = (2, 2, 2, 2, 2) \). Among the generators of the toric ideal \( I_{A_{K_5}} \) is the degree 8 binomial

\[
p_00000p_00111p_00101p_01001p_10001p_{11110} - p_00001p_00110p_00100p_01000p_10000p_{11111}.
\]

The corresponding vectors \( u, v \) are the only elements of their fiber. Proposition 10.4.5 shows that there is a \( b \) such that

\[
IP(A_{K_5}, b, c_{00000}) - LP(A_{K_5}, b, c_{00000}) = 2.
\]

Hosten and Sturmfels [HS07a] showed that for any fixed \( A \) and \( c \), there is always a maximum discrepancy between the integer programming optimum and the linear programming relaxation. That is:

**Theorem 10.4.7.** Fix \( A \in \mathbb{Z}^{k \times r} \) and \( c \in \mathbb{Q}^r \). Then

\[
\max_{b \in \mathbb{N}^A} (IP(A, b, c) - LP(A, b, c)) < \infty.
\]

The number \( \text{gap}(A, c) = \max_{b \in \mathbb{N}^A} (IP(A, b, c) - LP(A, b, c)) \) is called the **integer programming gap**, and exhibits the worst case behavior of the linear programming relaxation. It can be computed via examining the irreducible decomposition of the monomial ideal \( \text{in}_c(I_A) \) or a related monomial ideal in the case that \( c \) is not generic with respect to \( I_A \) [HS07a], as follows.

For an arbitrary \( c \in \mathbb{Q}^r \), let \( M(A, c) \) be the monomial ideal generated by all monomials \( p^u \) such that \( u \) is not optimal for the integer program (10.2.1) for \( b = Au \). Note that every monomial \( p^v \in M(A, c) \) has \( v \) nonoptimal since if \( u \) is nonoptimal in its fiber and \( w \) is optimal in this same fiber, and \( p^u|p^v \), then \( v \) could not be optimal since \( v + (w - u) \) is in the same fiber as \( v \) and has a lower weight.

Recall that an ideal \( I \) is irreducible if it cannot be written nontrivially as \( I = J \cap K \) for two larger ideals \( J \) and \( K \). Each irreducible monomial ideal in \( \mathbb{K}[p] \) is generated by powers of the variables, specifically a monomial ideal is irreducible if and only if it has the form

\[
I_{u, \tau} = \langle p_i^{\tau_i + 1} : i \in \tau \rangle,
\]
where \( u \in \mathbb{N}^r \) and \( \tau \subseteq [r] \), and \( u_i = 0 \) if \( i \notin \tau \). Every monomial ideal \( M \) can be written in a unique minimal way as the intersection of irreducible monomial ideals. This is called the \textit{irreducible decomposition} of \( M \).

For each such irreducible monomial ideal \( I_{u,\tau} \) in the irreducible decomposition of \( M(A,c) \) we calculate the gap value associated to \( c \) and \( A \), which is the optimal value of the following linear program:

\[
\text{(10.4.1)} \quad \max c^T u - c^T v \quad \text{such that} \quad Av = Au, \quad \text{and} \quad v_i \geq 0 \quad \text{for} \quad i \in \tau.
\]

**Theorem 10.4.8.** [HS07a] The integer programming gap \( \text{gap}(A,c) \) equals the maximum gap value (10.4.1) over all irreducible components \( I_{u,\tau} \) of the monomial ideal \( M(A,c) \).

**Example 10.4.9** (IP Gap for a 4-way Table). Consider the hierarchical model with \( \Gamma = [12][13][14][234] \) and \( r = (2,2,2,2) \). Consider the nongeneric optimization problem of minimizing the first cell entry in a \( 2 \times 2 \times 2 \times 2 \) contingency table given the \( \Gamma \) margins. Hence \( c = e_{1111} \). The integer programming gap \( \text{gap}(A,e_{1111}) \) measures the discrepancy between integer programming and linear programming in this case. Note that \( A_\Gamma \) is not normal in this case, so no initial ideal of \( I_{A_\Gamma} \) is squarefree (see Exercise 10.6) and we must do further analysis to determine the value of \( \text{gap}(A,e_{1111}) \).

Since \( e_{1111} \) is not generic with respect to the matrix \( A_\Gamma \), \( M(A_\Gamma,e_{1111}) \) is not an initial ideal of \( I_{A_\Gamma} \). We can compute the ideal \( M(A_\Gamma,e_{1111}) \) as follows. First we compute the non-monomial initial ideal \( \text{in}_{e_{1111}}(I_{A_\Gamma}) \). The ideal \( M(A_\Gamma,e_{1111}) \) is the ideal generated by all the monomials contained in \( \text{in}_{e_{1111}}(I_{A_\Gamma}) \). This ideal can be computed using Algorithm 4.4.2 in [SST00], which has been implemented in Macaulay2. The following Macaulay2 code computes the ideal \( M(A_\Gamma,e_{1111}) \) and its irreducible decomposition:

```plaintext
loadPackage "FourTiTwo"
S = ZZ[x1,x2,x3,x4]
listofmonomials = matrix{{1,x1,x2,x3,x4,x1*x2,x1*x3, x1*x4, x2*x3,x2*x4,x3*x4,x2*x3*x4}};
A = transpose matrix toList apply((0,0,0,0)..(1,1,1,1),
   i -> flatten entries sub(listofmonomials,
   {x1 => i_0, x2 => i_1, x3 => i_2, x4 => i_3}))
L = toricGraver(A);
R = QQ[p_(1,1,1,1)..p_(2,2,2,2), MonomialOrder => {Weights =>{
   1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0}}]
I = ideal gens toBinomial(L, R)
J = ideal leadTerm(1, I)
M = monomialSubideal(J)
transpose gens K
L = irreducibleDecomposition M
```
Among the 16 monomial irreducible ideals in the irreducible decomposition of $M(A, e_{1111})$ is the ideal
\[ \langle p_{1222}, p_{2112}, p_{2121}, p_{2211}^2 \rangle. \]

To compute its gap value amounts to solving the linear program:
\[
\max -v_{1111} \text{ subject to } A \mathbf{v} = (A_G)_{2211}, \text{ and } v_{1222}, v_{2112}, v_{2121}, v_{2211} \geq 0.
\]
The solution to this linear program shows that the gap value for this irreducible component is 1/2. Looking at all 16 irreducible components in this case produced one component with gap value 1, which was the largest value. Hence, $\text{gap}(A_G, e_{1111}) = 1$.

## 10.5. Formulas for Bounds on Cell Entries

The formulas in Proposition 10.1.2 for upper and lower bounds on cell entries of a two-way contingency table given the row and column sums are extremely appealing and it is natural to wonder if there are generalizations of these formulas for larger tables and more complex marginals. For example, in the case of two-way contingency tables none of the Gröbner basis machinery is needed to compute bounds on cell entries and use sequential importance sampling, so one might wonder if the Gröbner basis procedure is always needed. In principle there is always a closed formula for the linear programming upper and lower bounds, which can be revealed by the polyhedral geometry of the cone $\mathbb{R}_{\geq 0} A_G$. As we have seen in previous sections, in some cases those LP optima are equal to IP optima, and this can be useful. In most cases, the resulting formula that arises becomes increasingly complicated as the size of the contingency table increases.

We first consider the general integer program:
\[
(10.5.1) \quad \max x_1 \text{ subject to } Ax = b \text{ and } x \in \mathbb{R}_{\geq 0}^n
\]

associated with maximizing a coordinate. This optimization problem can be reformulated as follows in terms of $\text{cone}(A)$.

**Proposition 10.5.1.** The linear program 10.5.1 is equivalent to the optimization problem
\[
(10.5.2) \quad \max z \text{ subject to } b - zA_1 \in \text{cone}(A)
\]
where $A_1$ denotes the first column of $A$.

The proof of Proposition 10.5.1 is straightforward and left to the reader. The application of Proposition 10.5.1 comes from knowing the facet defining inequalities of the cone $\mathbb{R}_{\geq 0} A$. Indeed, suppose that $c^T y \geq 0$ is a valid inequality for $\text{cone}(A)$. Then we must have $c^T (b - zA_1) \geq 0$. If $c^T A_1 > 0$ then we deduce that $z \leq \frac{c^T b}{c^T A_1}$. Taking the minimum over all such formulas...
gives the solution to the linear program (10.5.1). For example, this can be used to give a simple proof of the upper bound in the case of 2-way tables.

Example 10.5.2 (Upper Bounds on 2-way Tables). Let $\Gamma = [1|2]$ and $r = (r_1, r_2)$, and consider the matrix $A_\Gamma$ and the corresponding optimization problem of maximizing the first coordinate of $x_{11}$ subject to the constraints $A_\Gamma x = b$, and $x \geq 0$. The cone $cone(A_\Gamma) \subseteq \mathbb{R}^{r_1+r_2}$ is given by a simple explicit description

$$cone(A_\Gamma) = \{ (y^{(1)}, y^{(2)}) \in \mathbb{R}^{r_1} \oplus \mathbb{R}^{r_2} : y_i^{(1)} \geq 0 \text{ for all } i \in [r_1], y_i^{(2)} \geq 0 \text{ for all } i \in [r_2], \sum_{i=1}^{r_1} y_i^{(1)} = \sum_{i=1}^{r_2} y_i^{(2)} \}.$$ 

The first column vector of $A_\Gamma$ is the vector $e_1^{(1)} \oplus e_1^{(2)}$ where $e_j^{(i)}$ denotes the first standard unit vector in $\mathbb{R}^{r_j}$. Of the $r_1 + r_2$ facet defining inequalities $c^T y \geq 0$, of $cone(A_\Gamma)$, only two of them satisfy that $c^T A_1 > 0$. These correspond to the inequalities $y_1^{(1)} \geq 0$ and $y_1^{(2)} \geq 0$. Now we apply the formula $z \leq \frac{c^T b}{c^T A_1}$. Note that in this context $b = (u_1, \ldots, u_{r_1}, u_{r_1+1}, \ldots, u_{r_1+r_2})$. So the inequality $y_1^{(1)} \geq 0$ implies that $x_{11} \geq u_{1+}$ and the inequality $y_1^{(2)} \geq 0$ implies that $x_{11} \leq u_{1}$. In summary, $x_{11} \leq \min(u_{1}, u_{1+})$.

Example 10.5.3 (Bounds on 3-way Table). For a more complex example consider the case of the binary graph model from Chapter 8 with $G = K_3$, the complete graph on 3 vertices. This is equivalent to the hierarchical model with $\Gamma = [12][13][23]$ and $r = (2, 2, 2)$, but the binary graph representation has the advantage that the matrix $A_G$ has full row rank, so there is no ambiguity in the choice of the facet defining inequalities.

The facet defining inequalities of the polyhedral cone $cone(A_G)$ were given for general $K_4$ minor free graphs in Theorem 8.2.10. For the case of the graph $K_3$, there are sixteen inequalities in total. These are the twelve basic inequalities:

$$p_{ij} \geq 0, p_i - p_{ij} \geq 0, p_{ij} + p_{i+} - p_i - p_j \geq 0$$

for $i, j \in \{1, 2, 3\}$ and four triangle inequalities:

$$p_{12} - p_{13} - p_{23} + p_3 \geq 0$$
$$p_{13} - p_{12} - p_{23} + p_2 \geq 0$$
$$p_{23} - p_{12} - p_{13} + p_1 \geq 0$$
$$p_{i+} + p_{i2} + p_{13} - p_1 - p_2 - p_{i3} \geq 0$$

The column vector corresponding to the coordinate $x_{111}$ is $(1, 0, 0, 0, 0, 0, 0)^T$. So the bounds we seek are only formulas that come from the inequalities
where $p_0$ appears with a positive coefficient. From this, we deduce the following upper bound on the coordinate $x_{111}$:

$$x_{111} \leq \min(p_0 - p_1 - p_2 + p_{12}, p_0 - p_1 - p_2 + p_{12}, p_0 - p_2 - p_3 + p_{23}, p_0 - p_1 - p_2 - p_3 + p_{13} + p_{23}).$$

We can also express these formulas in terms of entries in the 2-way marginals alone, those these formulas are not unique. In particular, there are many choices for the final formula:

$$x_{111} \leq \min(u_{11} + u_{1+1}, u_{11}, u_{11+}, u_{21} + u_{2+1} + u_{22} + u_{+12} - u_{+1} - u_{2+1}).$$

For lower bounds we can use a similar idea to find a formula, but the relevant cone changes from cone($A$) to a different cone. Indeed, we need to perform the following minimization:

**Proposition 10.5.4.** The linear program $\min x_1 : Ax = b, x \geq 0$, is equivalent to the optimization problem

$$\min z \text{ subject to } b - zA_1 \in \text{cone}(A'), z \geq 0$$

where $A_1$ denotes the first column of $A$, and $A'$ is the matrix obtained from $A$ by deleting the first column.

From Proposition 10.5.4 we deduce that $z \geq \frac{c^T b}{c^T A_1}$ for any valid inequality $c^T y \geq 0$ for cone($A'$) for which $c^T A_1 < 0$. The proof of Proposition 10.5.4 follows because we want to choose the smallest nonnegative value of $z$ such that $b - zA_1$ can be written as a nonnegative combination of the other columns. Note that this is also true for the maximization problem, but we do not need to consider cone($A'$) because we automatically end up in cone($A'$) when performing the maximization (whereas, this is not forced in the minimization, i.e. we can always choose $z = 0$, if we worked with cone($A$) in the minimization problem).

The polyhedron cone($A'$) is less well-studied for contingency table problems, but can be worked out in simple examples, and deserves further study.

**Example 10.5.5 (Lower Bounds on 2-way Tables).** Let $\Gamma = [1][2]$ and $r = (r_1, r_2)$, and consider the matrix $A_{\Gamma}$ and the corresponding optimization problem of minimizing the first coordinate of $x_{11}$ subject to the constraints $A_{\Gamma} x = b$, and $x \geq 0$. The cone cone($A_{\Gamma}$) $\subseteq \mathbb{R}^{r_1 + r_2}$ is given by a simple explicit description:

$$\text{cone}(A_{\Gamma}) = \{(y^{(1)}, y^{(2)}) \in \mathbb{R}^{r_1} \oplus \mathbb{R}^{r_2} : y^{(1)}_i \geq 0 \text{ for all } i \in [r_1], y^{(2)}_i \geq 0 \text{ for all } i \in [r_2], \sum_{i=2}^{r_1} y^{(1)}_i - y^{(2)}_i \geq 0\}$$
10.5. Formulas for Bounds on Cell Entries

The only inequality $c^Ty \geq 0$ such that $c^TA_1 < 0$ is the new inequality,

$$\sum_{i=2}^{r_2} y_i \geq 0$$

from which we deduce that

$$x_{11} \geq u_{+1} - \sum_{i=2}^{r_1} u_{i+}.$$ 

Note that because cone($A^T_1$) is not full-dimensional, there might be many different ways to represent this expression in terms of the marginals. For instance $x_{11} \geq u_{+1} + u_{1+} - u_{++}$ is an equivalent inequality known as the Bonferroni bound. Always we will need to take the max of whatever expressions we find and 0.

As we have seen, the decomposable hierarchical models possess similar properties to 2-way contingency tables. This similarity extends to the existence of simple formulas for the upper and lower bounds for the cell entries given the decomposable marginals [DF00].

**Theorem 10.5.6.** Let $\Gamma$ be a decomposable simplicial complex with facets $F_1, F_2, \ldots, F_k$. Suppose further that these facets are ordered so that for each $j = 2, \ldots, k$, the complex $\Gamma_j$ whose facets are $\{F_1, \ldots, F_{j-1}\}$ is also decomposable and $F_j \cap \Gamma_j$ is a simplex $2^S_j$. Let $u \in \mathbb{N}^R$ be a contingency table. Then for each $i \in \mathcal{R}$ and $v \in \mathcal{F}(u)$,

$$\min \left( (u|_{F_1})_{i_{F_1}}, \ldots, (u|_{F_2})_{i_{F_2}} \right) \geq v_i \geq \max \left( \sum_{j=1}^{k} (u|_{F_j})_{i_{F_j}} - \sum_{j=2}^{k} (u|_{S_j})_{i_{S_j}}, 0 \right).$$

Furthermore, the upper and lower bounds are tight, in the sense that for each coordinate there exist two tables $v^u, v^l \in \mathcal{F}(u)$ achieving the upper and lower bounds, respectively.

The ordering of the facets of $\Gamma$ in Theorem 10.5.6 is called a perfect elimination ordering.

**Example 10.5.7 (A Large Decomposable Model).** Let $\Gamma = [12][23][24][567]$. A perfect elimination ordering of the facets is given by the ordering that the facets are listed in the complex. The resulting sequence of separators is $S_2 = \{2\}, S_3 = \{2\}, S_4 = \emptyset$. Theorem 10.5.6 gives the following tight upper and lower bounds on the cell entry $v_{i_{12}i_{23}i_{24}i_{56}i_{7}}$ given the marginals $A_{11}u$:

$$\min \left( (u|_{12})_{i_{12}}, (u|_{23})_{i_{23}}, (u|_{24})_{i_{24}}, (u|_{567})_{i_{56}i_{7}}, 1 \right) \geq v_{i_{12}i_{23}i_{24}i_{56}i_{7}} \geq \max \left( (u|_{12})_{i_{12}} + (u|_{23})_{i_{23}} + (u|_{24})_{i_{24}} + (u|_{567})_{i_{56}i_{7}} - 2(u|_{2})_{i_{2}} - u_{i_{2}}, 0 \right).$$
10.6. Exercises

Exercise 10.1. Show that the estimator
\[ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{q(v_i)} \]
converges to the number of elements in the fiber \( F(u) \) (see Example 10.1.3).

Exercise 10.2. Let \( u \) be the table from Table 10.1.1. Prove the claims made about this table at the end of Section 10.1.2. In particular:

1. Show that if \( \Gamma \) is the simplicial complex whose facets are all 3-element subsets of [5], then the fiber \( F(u) \) consists of just \( u \).
2. Show that if \( \Gamma \) is the simplicial complex whose facets are all 2-element subsets of [5], then linear programming upper and lower bounds can be used to determine some nonzero cell entries in \( u \).

Exercise 10.3. Prove Theorem 10.2.5.

Exercise 10.4. Let \( A \in \mathbb{N}^{k \times r} \) be an integer matrix and \( b \in \mathbb{R}^k \). Suppose \( L^t, U^t \in \mathbb{R}^r \) are vectors such that \( L^t_i \leq x_i \leq U^t_i \) for all \( x \) in the polytope \( P(A, b) = \{ x \in \mathbb{R}^r : Ax = b, x \geq 0 \} \) and for all \( i = 1, \ldots, r \). That is, the vectors \( L^t \) and \( U^t \) are valid lower and upper bounds for each coordinate.

Define \( L^{t+1} \) and \( U^{t+1} \) as follows:
\[
L^{t+1}_i = \max(0, L^t_i, \max_{j : A_{ji} \neq 0} (\frac{1}{A_{ji}} \sum_{k \neq i} A_{jk} U_k))
\]
\[
U^{t+1}_i = \min(U^t_i, \min_{j : A_{ji} \neq 0} (\frac{1}{A_{ji}} \sum_{k \neq i} A_{jk} L_k)).
\]
Show that \( L^{t+1} \) and \( U^{t+1} \) are also valid lower and upper bounds on coordinates of \( x \) in the polytope \( P(A, b) \), that improve \( L^t \) and \( U^t \).

Exercise 10.4 is the basis for the shuttle algorithm of Dobra and Fienberg [DF01, Dob02] for computing bounds on cell entries in contingency tables.

Exercise 10.5. Prove Proposition 10.4.2.

Exercise 10.6. Show that if there is a term order such that \( \text{in}_\omega(I_A) \) is a squarefree monomial ideal then \( NA \) is a normal semigroup.

Exercise 10.7. Show that if \( NA \) is normal and if \( p^u - p^v \in I_A \) is a binomial that appears in every minimal binomial generating set of \( I_A \), then either \( p^u \) or \( p^v \) is a squarefree monomial.

Exercise 10.8. Let \( A \in \mathbb{Z}^{k \times r} \). Let \( \pi : \mathbb{N}^r \rightarrow \mathbb{N} \) be the projection map onto the first coordinate: \( \pi(u) = u_1 \). In sequential importance sampling, a useful
property is that every fiber satisfies $\pi(F(u)) = [a, b] \cap \mathbb{Z}$ for some integers $a$ and $b$ [CDS06]. In this case we say that $A$ has the interval of integers property. Show that $A$ has the interval of integers property if and only if in the reduced lexicographic Gröbner basis for $I_A$, each binomial $p^u - p^v$ has $\pi(u) \in \{0, 1\}$.

**Exercise 10.9.** Let $\Gamma = K_4$ and take $r = (2, 2, 2, 2)$. Show that there is a $b$ such that $IP(A_\Gamma, b, e_{1111}) - LP(A_\Gamma, b, e_{1111}) = 1$.

**Exercise 10.10.** Let $\Gamma = K_4$ and take $r = (2, 2, 2, 2)$. Show that $\text{gap}(A_\Gamma, -e_{1111}) = 5/3$.

**Exercise 10.11.** Let $\Gamma = K_4$ and take $r = (2, 2, 2, 2)$. Find an explicit formula for the real upper and lower bounds on the cell entry $x_{1111}$ given the marginal totals akin to the formula we derived in Example 10.5.3.

**Exercise 10.12.** Let $\Gamma = [1][2][3]$. Find formulas for the upper and lower bounds on $u_{ijk}$ given the $A_\Gamma u = b$ and $u \in \mathbb{N}^{r_1 \times r_2 \times r_3}$. Show that your bounds are tight and that every value in between the upper and lower bounds occurs for some table with the given fixed margins. (Prove your results directly, do not use Theorem 10.5.6.)
Chapter 11

Exponential Random Graph Models

In this chapter we study the exponential random graph models, which are statistical models that are frequently used in the study of social networks. While exponential random graph models are a special case of discrete exponential family models, their study presents some subtleties because data sets do not consist of i.i.d. samples. For a typical exponential family model, we have i.i.d. samples from an underlying distribution in the model. However, for an exponential random graph model, we typically only have a sample size of one, that is, we observe a single graph and would like to make inferences about the process that generated it. Another way to think about this small sample size is that we observe not one sample but many samples of edges between pairs of vertices in the graph, but that the presence or absence of edges are not independent events. This issue of small sample size/non-i.i.d. data presents problems about the existence of maximum likelihood estimates, and also means that we cannot usually use hypothesis tests based on asymptotics. These issues can be addressed via analyzing the polyhedral structure of the set of sufficient statistics, in the first case, and developing tools for performing Fisher’s exact test in the second.

11.1. Basic Setup

Models for random graphs are widely used and are of interest in network science, including the study of social networks. Generally, we create a statistical model for a random graph where the probability of a particular graph
depends on some features of the graph. Exponential random graphs models provide a precise way to set up a very general family of such models.

**Definition 11.1.1.** Let $G_m$ denote the set of all undirected simple graphs on $m$ vertices. A graph statistic is a function $t : G_m \rightarrow \mathbb{R}$.

Examples of graph statistics that are commonly used in this area include: the number of edges in the graph, the number of isomorphic copies of a certain small graph, the degree of a particular vertex, etc.

**Definition 11.1.2.** Let $T = (t_1, \ldots, t_k)$ be graph statistics. The exponential random graph model associated to $T$ is the exponential family where

$$P(X = G) = \frac{1}{Z(\theta)} \exp(t_1(G)\theta_1 + \cdots + t_k(G)\theta_k)$$

where $\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$ is a real parameter vector and $Z(\theta)$ is the normalizing constant

$$Z(\theta) = \sum_{G \in G_m} \exp(t_1(G)\theta_1 + \cdots + t_k(G)\theta_k).$$

**Example 11.1.3 (Erdős-Rényi random graphs).** The prototypical example of an exponential random graph model is the Erdős-Rényi random graph model. In the language of exponential random graph models, the Erdős-Rényi model has a single parameter $\theta$, and the corresponding graph statistic $t(G)$ is the number of edges in $G$. Setting $p = \frac{\exp(\theta)}{1 + \exp(\theta)}$ we see that the probability of observing a particular graph $G$ is

$$P(X = G) = p^{|E(G)|}(1 - p)^{m^2 - |E(G)|}.$$

An interpretation of this probability is that each edge in the graph is present independently with probability $p$, which is the usual presentation for the Erdős-Rényi random graph model. Note that the small sample size issue we warned about in the introduction does not really appear in this example, because the presence/absence of edges can be treated as independent trials, by the definition of this model.

**Example 11.1.4 (Stochastic Block Model).** A natural extension of the Erdős-Rényi model is the stochastic block model. In this exponential random graph model, each edge is either present or absent independently of other edges in the graph, however the probabilities associated to each edge varies. In particular, suppose that $B_1, \ldots, B_k$ is a partition of the vertex set $V$. Let $z : V \rightarrow [k]$ be the function that assigns an element of $V$ to the index of its corresponding block. For each pair $i, j \in [k]$ let $p_{ij} \in (0, 1)$ be a probability that represents the probability of the type of an edge between a vertex in block $i$ and a vertex in block $j$. Let $g$ denote a graph which is represented as a 0/1 array where $g_{ij} = 1$ if there is an edge between vertices
Then, given the assignment of blocks $B_1, \ldots, B_k$, the probability of observing the graph $g$ is

$$P(G = g \mid B_1, \ldots, B_k) = \prod_{1 \leq i < j \leq |V|} p_{z(i)z(j)}^{z(i)z(j)} (1 - p_{z(i)z(j)})^{(1 - p_{z(i)z(j)})^2}.$$ 

This model can be seen to be an exponential random graph model with $k + 1$ parameters.

One nice property of the Erdős-Rényi random graph model is that the normalizing constant $Z(\theta)$ ends up having a simple factorization. Indeed, $Z(\theta) = (1 + \exp(\theta))^{m \choose 2}$. This makes this exponential random graph model especially easy to work with, since we do not need to actually sum over all graphs to find the probability of a graph. There usually is not such a simple formula for the normalizing constant in most exponential random graph models.

**Example 11.1.5 (Edge-Triangle Model).** Consider the exponential random graph model with graph statistics $T = (t_1, t_2)$ where $t_1(G)$ is the number of edges in $G$ and $t_2(G)$ is the number of triangles in $G$ (that is, the number of triples of vertices $i, j, k$ such that all of $ij, ik, jk$ are edges in $G$).

Although there are $2^{n(m-1)/2}$ different graphs on $m$ vertices, there are significantly fewer possible pairs $(t_1(G), t_2(G))$. For example, for $m = 7$, there are 2097152 graphs but only 110 different pairs of values of the number of edges and number of triangles. These 110 points and their convex hull is illustrated in Figure 11.1.1. Note that the fact that the line segment connecting the points $(0, 0)$ and $(21, 35)$ does not contain any other points of the form $(t_1(G), t_2(G))$ implies that the normalizing constant $Z(G)$ is an irreducible polynomial in the quantities $\exp(\theta_1)$ and $\exp(\theta_2)$.

One great difference between using exponential random graph models and the majority of other exponential families we have discussed thus far is that the datasets that are observed do not typically consist of i.i.d. samples from an underlying distribution. Rather, the dataset is simply a single graph (or perhaps, a time series of graphs on the same node set). In this setting, large sample theory does not apply, so hypothesis tests based on asymptotic approximations to test statistics are not appropriate. In particular, hypothesis testing should be performed using Fisher’s exact test.

To perform Fisher’s exact test for an exponential random graph model, we need to sample from the set of all graphs with given sufficient statistics. As we only have one sample, there is no multinomial coefficient that appears in the likelihood function, so we sample from the uniform distribution on a fiber.
Example 11.1.6 (Sampling in the Stochastic Block Model). In the stochastic block model, the sufficient statistics are the numbers of edges between each pair of blocks. To sample a random graph with these sufficient statistics fixed, we simply pick a random subset of the edges between the two blocks among all possible edge sets of the appropriate size. Sampling in the stochastic block model becomes more challenging when the block structure is unknown, see [KPP+16].

For even quite simple models, it is not at all clear how to construct a Markov basis for the sampling problem.

Example 11.1.7 (Sampling in the Edge-Triangle Model). For the model in Example 11.1.5, we would need to sample from the set of graphs with a fixed number of vertices, edges, and triangles. It is unclear what a general strategy for such sampling might be. For instance, suppose we want to sample from the set of graphs with $2m$ vertices, $m^2$ edges, and no triangles. The only such graphs are the complete bipartite graphs with bipartition of vertices into parts of size $m$ and $n$. A move that connects two such graphs will have to involve swapping two vertices between the different parts of the bipartition. This has the effect of changing $2m - 2$ edges. In fact, for this model there is no set of moves that connects fibers that only involves a bounded number of edges.

For most exponential random graph models, sampling from fibers becomes a purely combinatorial undertaking. However, for certain specific exponential random graphs models, more algebraic tools can be employed to design sampling schemes. We will see examples of this in Section 11.2.
One reason that the sampling problem in Example 11.1.6 appears to be easier than the sampling problem in Example 11.1.7 is that the sufficient statistics in the first case are linear functions of the adjacency matrix $g$, whereas in the second case, they are not linear functions of the adjacency matrix. Indeed, in the second case, the number of triangles can be seen as a cubic function of the adjacency matrix, namely

$$t_2(G) = \sum_{i<j<k} g_{ij} g_{ik} g_{jk}.$$ 

One key issue in using the exponential random graph models is estimating their parameters from data. As we discussed in Chapter 8, the maximum likelihood estimates will exist if and only if the observed sufficient statistics lie in the interior of the convex hull of the set of all sufficient statistics. For many exponential graph models, a large proportion of graphs have their sufficient statistics on or near to the boundary of the convex hull, and this presents a problem for inference with these models.

**Example 11.1.8** (Estimation in the Edge-Triangle Model). Consider the exponential random graph model from Example 11.1.5, and suppose that we observe a graph $G$ with seven vertices and sufficient statistics $(13, 8)$, that is 13 edges and 8 triangles. Letting $y_1 = \exp(\theta_1)$ and $y_2 = \exp(\theta_2)$, we need to optimize the function $y_1^{13} y_2^8 Z(y_1, y_2)^{-1}$ where $Z(y_1, y_2)$ is the normalizing constant, a polynomial of degree 56 in $y_1$ and $y_2$ with 110 terms. Since this is an exponential family, there is a unique maximum in the positive orthant. This is obtained for $y_1 \approx 2.0396$ and $y_2 \approx 0.8880$. On the other hand, if we observe a graph with 7 vertices 13 edges and only 4 triangles, while this data point is not on the boundary of the convex hull, it is quite close to the boundary, and the computation of maximum likelihood estimates is quite slow to converge, suggesting a limiting value with $y_1$ large and $y_2$ close to zero. These issues about the existence and difficulty of computing maximum likelihood estimates in ERGMs are studied in detail in [RPF13]. See [YRF16] for more details on the geometry of the polytope of the Edge-Triangle model.

### 11.2. The Beta Model and Variants

In this section, we discuss some widely used exponential random graph models that are based on degrees of vertices. Unlike the simple model that was discussed in Example 11.1.5 in Section 11.1, these models have a number of parameters that increases with the graph size. This makes it especially difficult to talk about asymptotic properties of these models. However, the beta model and its variants also possess certain nice properties that make them easier to analyze and where connections to toric ideals can be made.
Definition 11.2.1. Let $G$ be a graph with vertex set $[m]$. For $i \in [m]$ let $\deg_i(G)$ denote the number of edges of $G$ incident to vertex $i$. This number is called the degree of vertex $i$. The beta model is the exponential random graph model for an $m$ vertex graph whose sufficient statistics are $(\deg_1(G), \ldots, \deg_m(G))$.

Among the nice properties of the beta model are that its normalizing constant is simple to compute.

Proposition 11.2.2. The normalizing constant for the beta model is

$$Z(\theta) = \prod_{1 \leq i < j \leq m} (1 + \exp(\theta_i + \theta_j)).$$

Proof. For each pair $i, j$, the corresponding edge is either in the graph, or not in the graph. If it is in the graph, it contributes $+1$ to $\deg_i(G)$ and $\deg_j(G)$. If it is not in the graph, the degree remains unchanged. This explains the two terms in the product for each pair $1 \leq i < j \leq m$. □

An integer vector $d = (d_1, \ldots, d_m)$ is called a degree sequence if it is the sequence of vertex degrees of some graph on $m$ vertices. The polytope of all sufficient statistics of the beta model consists of the convex hull of all degree sequence vectors. These possible degree sequences are characterized by the Erdős-Gallai theorem [EG60].

Theorem 11.2.3. A sequence of nonnegative integers $(d_1, \ldots, d_m)$ with $d_1 \geq d_2 \geq \cdots \geq d_m$ is the degree sequence of a simple graph if and only if $d_1 + \cdots + d_m$ is even and for all $k = 1, \ldots, m$,

$$\sum_{i=1}^{k} d_i \leq k(k-1) + \sum_{i=k+1}^{m} \min(d_i, k).$$

A closely related model to the beta model is the Rasch model, which can be seen in various guises. The closest relation to the beta model follows from taking the beta model restricted to bipartite graphs with a fixed bipartition of the vertices.

Definition 11.2.4. Let $V_1, V_2$ be two sets of vertices. The Rasch model is the beta model where we only allow edges between $V_1$ and $V_2$ (no edges connecting vertices within $V_1$ or within $V_2$). The sufficient statistic of a graph $G$ is the vector of all vertex degrees.

There is also a simple description of which vectors are actually possible as sufficient statistics vectors in this model, provided by the Gale-Ryser theorem [Gal57, Rys57].
Theorem 11.2.5. A pair of nonnegative integer sequences \((d_1, \ldots, d_m), (e_1, \ldots, e_{m'})\) with \(d_1 \geq d_2 \geq \cdots \geq d_m\) is a pair of degree sequences of a bipartite graph if and only if
\[
\sum_{i=1}^{k} d_i = \sum_{i=1}^{m'} e_i \quad \text{and for } k = 1, \ldots, m,
\]
\[
\sum_{i=1}^{k} d_i \leq \sum_{i=1}^{m'} \min(e_i, k).
\]

Each simple bipartite graph with bipartition of the vertex set with \(m\) and \(m'\) vertices respectively, naturally corresponds to an \(m \times m'\) table of zeros and ones. Fixing the vertex degrees is equivalent to fixing the row and columns sums of that matrix. Hence, the fiber of bipartite graphs with fixed vertex degrees can be connected using moves that involve changing only two edges at a time, if we apply Theorem 9.4.6.

It is also possible to give a simple description of a Markov basis that will allow for a random walk on the fiber \(F(G)\) consisting of the set of arbitrary graphs which have the same degree sequence as the given graph \(G\), using a similar argument as Theorem 9.4.6.

Proposition 11.2.6. Consider the beta model and let \(G\) be a graph with degree sequence \((d_1, \ldots, d_m)\). The fiber \(F(G)\) consisting of all graphs with degree sequence \((d_1, \ldots, d_m)\) can be connected by local moves which replace the pair of edges in the graph \(ij, kl\) with the pair of edges \(ik, jl\) provided that neither of these edges already appeared in the graph.

The proof of Proposition 11.2.6 is left as an exercise. It follows the basic outline of the proof of Theorem 9.4.6.

The polyhedral descriptions of the set of degree sequences for simple graphs and bipartite graphs obtained in the Erdős-Gallai theorem and the Gale-Ryser theorem allow to easily check whether a given degree sequence lies on the boundary of the polytope of all degree sequences. In particular, a degree sequence lies on the boundary of the degree sequence polytope if one of the \(d_i = 0\), or if one of the inequalities in Theorem 11.2.3 or 11.2.5 is satisfied with equality. Another use for these theorems is that they can be used in the development of sequential importance sampling machinery for approximately generating samples from the corresponding fibers. These strategies are described in [BD10] and [CDHL05] for graphs and bipartite graphs, respectively. We describe the basic idea of this procedure here for bipartite graphs, phrased in the language of sampling \(0/1\) matrices.

The basic sequential importance sampling scheme described in Section 10.1.1, involves filling entries in a contingency table one at a time, and computing bounds at each step. For sampling \(0/1\) matrices with fixed row and column sums, [CDHL05] advocates sampling each column one at a time (that is, filling a whole column in one step). Let the row sums and
columns sums be \((d_1, \ldots, d_m)\) and \((e_1, \ldots, e_m')\) respectively. To choose a possible column with column sum \(e_1\), amounts to choosing a 0/1 vector in \(\mathbb{N}^m\). There are \(\binom{m}{e_1}\) possible 0/1 vectors to choose. The most naive algorithm would be to simply pick such a vector uniformly at random to fill in the column, then update the row sums accordingly. One can check for rejection after each step using the Gale-Ryser theorem (i.e. check if the updated row and column sums are actually the row and columns sums of a 0/1 matrix). According to [CDHL05] this is not a very effective strategy because, intuitively, if a row sum is large the corresponding entry in that row should be 1 with higher probability than if the row sum were small. To remedy this problem, [CDHL05] proposes using the conditional Poisson distribution to sample columns. For \(i = 1, \ldots, m\) let \(p_i = \frac{d_i}{d_1 + \cdots + d_m}\) let \(Z_i\) be a Bernoulli random variable with parameter \(p_i\). The conditional Poisson distribution is the distribution of \((Z_1, \ldots, Z_m)\) conditional on \(Z_1 + \cdots + Z_m = e_m'\). It is not difficult to check that

\[
P(Z = z \mid Z_1 + \cdots + Z_m = e_m') \propto \prod_{i=1}^{m} \left( \frac{p_i}{1 - p_i} \right)^{z_i}
\]

which ensures that rows with large row sums are more likely to have a 1 put in the corresponding position. [CDHL05] reports significantly improved performance under this sampling strategy. Further performance improvements can be obtained by making better use of the Gale-Ryser theorem beyond forcing rejections. For instance, analysis with the Gale-Ryser theorem can identify positions that must be 0 or must be 1, or regions that must contain a large number of 1s. These observations can lead to adjustments in how the conditional Poisson distribution is modified and further improvements. Similar strategies can also be employed for sampling graphs with fixed degree sequences as explained in [BD10].

The beta model has a number of generalizations that involve either expanding the class of combinatorial structures considered or allowing more complicated graph interactions. For example, [SSR+14] generalizes the beta model to allow not just pairwise connections but connections of arbitrary size. This results in a hypergraph version of the beta model.

A widely used generalization of the beta model in social network analysis is the \(p_1\) model [HL81]. This model essentially allows for different types of edge interactions between pairs of vertices. Between a pair of vertices \(i, j\) there might be no edge, a directed edge from \(i\) to \(j\), a directed edge from \(j\) to \(i\), or a bidirected edge between \(i\) and \(j\). In the social network setting, this might be seen as having one user follow another on a social networking platform, so also both of the pair could follow each other. The full \(p_1\) model has \(2\binom{m}{2} + 2m - 1\) parameters which are for each pair \(\lambda_{ij}\), and
$\rho_{ij}$, a parameter $\theta$, and $\alpha_i$ and $\beta_j$. It is typically represented in log-linear form by looking at the log of the probability of the different types of edges appearing, via the following equations:

\[
\begin{align*}
\log p_{ij}(0, 0) &= \lambda_{ij} \\
\log p_{ij}(1, 0) &= \lambda_{ij} + \alpha_i + \beta_j + \theta \\
\log p_{ij}(0, 1) &= \lambda_{ij} + \alpha_j + \beta_i + \theta \\
\log p_{ij}(1, 1) &= \lambda_{ij} + \alpha_i + \beta_j + \alpha_j + \beta_i + 2\theta + \rho_{ij}
\end{align*}
\]

These quantities are interpreted as follows: $p_{ij}(0, 0)$ is the probability of no edge between $i$ and $j$, $p_{ij}(1, 0)$ is the probability of a directed edge from $i$ to $j$, $p_{ij}(0, 1)$ is the probability of a directed edge from $j$ to $i$, and $p_{ij}(1, 1)$ is the probability of a bidirected edge between $i$ and $j$. The parameter $\alpha_i$ is an effect local to $i$ measuring its strength of sending outgoing edges and similarly $\beta_j$ is an effect local to $j$ measuring its strength of receiving incoming edges. The parameter $\theta$ is an intensity of the appearance of any edge (so that the bidirected edge could be considered a double edge). The parameter $\rho_{ij}$ is an intensity factor for reciprocation between $i$ and $j$. The parameters $\lambda_{ij}$ are essentially normalizing constants that are present to ensure that

\[p_{ij}(0, 0) + p_{ij}(1, 0) + p_{ij}(0, 1) + p_{ij}(1, 1) = 1\]

for each $i, j$ pair. Note that to avoid over-parametrization of the model, we can assume that $\sum \alpha_i = 0$ and $\sum \beta_j = 0$.

Submodels of the $p_{1}$ model are studied based on reducing the degrees of freedom in the $\rho_{ij}$ parameters. These possible submodels include:

- $\rho_{ij} = 0$, i.e. no reciprocation effect.
- $\rho_{ij} = \rho$, i.e. constant reciprocation effect.
- $\rho_{ij} = \rho + \rho_i + \rho_j$, vertex dependent reciprocation effect.

It is worth thinking through the sufficient statistics of these various models, to try to understand what the fibers of the models look like and how this affects the problem of developing a scheme for sampling from the corresponding fibers. Common to all the models is the requirement that each pair of vertices has exactly one type of interaction. The $\alpha_i$ and $\beta_j$ terms say that we want to look at graphs that have the same outdegree and indegree at each vertex (where a bidirected edge contributes both out and in at each vertex). If we were to set both $\theta$ and $\rho_{ij}$ to zero this would correspond to a directed analogue of the beta model. The $\theta$ parameter says that the total number of edges appearing in the graph should be constant, where the bidirected edge is counted as two edges. Adding the constant reciprocation effect $\rho$ says that the total number of bidirected edges in the graph should remain constant. Lastly, adding the vertex dependent reciprocation terms
Describing methods to sample from these fibers of the $p_1$ model seems a difficult problem in general. An initial study of the Markov bases and geometry of the marginal polytopes for these models appears in [FPR11], and a more detailed study appears in [PRF10]. For the model without reciprocation, it is possible to give a straightforward description of a Markov basis, closely related to the problem of sampling 0/1 matrices with fixed row and column sums that we have discussed previously. Without reciprocation, each bidirected edge $i \leftrightarrow j$ can be thought of as the pair of directed edges $i \rightarrow j$ and $j \rightarrow i$. Then to distinguish in and out degrees, we convert this to a bipartite graph on twice as many vertices, $\{u_1, \ldots, u_m\}$ and $\{v_1, \ldots, v_m\}$. We include an edge $u_i - v_j$ in this new undirected graph if there is a directed edge $i \rightarrow j$ in the original graph. Since we do not allow loops $i \rightarrow i$, we must exclude the possibility of edges $u_i - v_i$ for all $i$. In the usual way, we can translate a bipartite graph into a 0/1 incidence matrix. Our condition on not allowing edges $u_i - v_i$ for all $i$ forces that this matrix has all zeroes on the diagonal. The condition on fixing indegrees and outdegrees in the original graph translates to fixing vertex degrees in the bipartite graph, which amounts to fixing row and column sums in the 0/1 matrix. In summary:

**Proposition 11.2.7.** Sampling from the fiber of graphs in the $p_1$ model with no reciprocation effect is equivalent to sampling 0/1 $m \times m$ matrices with zeroes on the diagonal and fixed row and column sums.

The connection between the $p_1$ model with no reciprocation and 0/1 tables, means that we can use some of the standard machinery on Markov bases and Graver bases to compute random walks on the appropriate fibers. Proposition 11.2.7 together with Theorem 9.4.5 shows that the Graver basis for the 2-way independence model can be used to connect fibers. This will come from moves that come from even cycles in the bipartite graph (see Proposition 9.4.3). In fact, it is possible to connect these fibers using a smaller set of moves, as explained in Exercise 11.4.

It seems a difficult problem to give a general description of the Markov basis or Graver basis for the $p_1$ model with other conditions on the reciprocation parameters $p_{ij}$. On the level of our 0/1 matrices, the corresponding sufficient statistics do not correspond to linear constraints on the matrix. One option is to try to generate Graver basis moves on the fly (see [GPS17, OHT13]). Two options for dealing with this problem are to directly use the corresponding linear constraints as implied by the exponential family representation, or use nonlinear constraints (which is potentially complicated for setting up sampling schemes).
To illustrate the ideas, consider the constant reciprocation effect, with \( \rho_{ij} = \rho \). The corresponding sufficient statistic is the number of bidirected edges in the graph. In terms of our \( m \times m \) 0/1 matrix \( X = (x_{ij})_{i,j \in [m]} \), the number of reciprocal pairs is

\[
\sum_{1 \leq i < j \leq m} x_{ij}x_{ji}.
\]

This quadratic constraint can be converted into linear constraints by lifting into higher dimensions using the geometry of the cut polytope. See [DL97] for the general approach and [XS14] for an application in algebraic statistics. However, it is not clear in general how to sample from the set of graphs which have this complicated mixture of constraints.

11.3. Models from Subgraphs Statistics

In this section, we discuss some generalizations of Example 11.1.5, where all of the statistics being considered are subgraph statistics. The geometry of the resulting polytopes of all sufficient statistics have interesting connections to problems in real algebraic geometry.

**Definition 11.3.1.** Let \( H \) and \( G \) be graphs. The subgraph count statistic \( t_H(G) \) counts the number of isomorphic copies of the graph \( H \) in \( G \). Suppose that \( G \) has \( m \) vertices. The normalized subgraph count statistic or subgraph density is \( \bar{t}_H(G) = \frac{t_H(G)}{t_H(K_m)} \).

Note that in the subgraph count statistic we are counting the number of isomorphic copies of a graph in \( H \) in \( G \). There might be many such copies on the same set of vertices. For instance, the graph \( K_4 \setminus e \) is a subgraph 6 times in the graph \( K_4 \).

**Definition 11.3.2.** Let \( \mathcal{H} = \{H_1, \ldots, H_k\} \) be a collection of graphs. The exponential random graph model associated to \( \mathcal{H} \) is the exponential family on the set of graphs of a fixed size where the vector of sufficient statistics is \( (t_{H_1}(G), \ldots, t_{H_k}(G)) \).

One advantage of working with the normalized graph count statistics is that the normalization guarantees that \( t_H(G) \in [0, 1] \) regardless of the size of the graph \( G \). This makes it easier to compare the density of the two graphs of different sizes.

**Definition 11.3.3.** Let \( \mathcal{H} = \{H_1, \ldots, H_k\} \) be a collection of graphs. The polytope of subgraph statistics defined by \( \mathcal{H} \), denoted \( P_{\mathcal{H},n} \), is the convex hull of all the vectors \( (t_{H_1}(G), \ldots, t_{H_k}(G)) \) as \( G \) ranges over all graphs with \( n \) vertices.
Since the normalized graph count statistics put all these polytopes into the cube $[0,1]^k$, it is natural to compare these polytopes. One result is:

**Proposition 11.3.4.** [EN11, Proposition 2.7] Suppose that each graph in $\mathcal{H}$ has at most $n$ vertices, and let $n \leq n' \leq n''$. Then $P_{\mathcal{H},n''} \subseteq P_{\mathcal{H},n'}$.

**Proof.** Let $G$ be any graph, and let $U$ be a subset of the vertices of $G$ and denote by $G[U]$ the induced subgraph of $G$ on the vertex set $U$. A basic fact about the subgraph density is that it can be averaged over subsets, which gives the following formula

$$\bar{t}_H(G) = \frac{1}{\binom{|G|}{n'}} \sum_{U \subseteq V(G) : |U| = n'} \bar{t}_H(G[U]),$$

where $|G|$ denotes the number of vertices in the graph $G$. Indeed, each subgraph isomorphic to $H$ contributes the same amount to each of the subgraph statistics $\bar{t}_H(G[U])$ for any $U$ that contains that subgraph. For any subgraph isomorphic to $H$, the number of $U$ that contain it is the same. This equation implies that the vector $(\bar{t}_H(G[U]), \ldots, \bar{t}_H(G[U])) \in P_{\mathcal{H},n'}$ is in the convex hull of the vectors $(\bar{t}_H(G[U]), \ldots, \bar{t}_H(G[U])) \in P_{\mathcal{H},n''}$. \hfill $\Box$

Since the polytopes $P_{\mathcal{H},n}$ form a decreasing sequences as $n \to \infty$, there is a well-defined limiting object $\cap_{n \geq n'} P_{\mathcal{H},n}$ where $n$ is the number of vertices in the largest subgraph of $\mathcal{H}$. A natural goal in this area is to try to understand the geometry of these polytopes or at the least the limiting object. One useful fact is that all the polytopes contain certain semialgebraic sets formed by considering random graph models.

**Definition 11.3.5.** Let $\mathcal{H} = \{H_1, \ldots, H_k\}$ be a collection of graphs and suppose that graph $H_i$ has $e_i$ edges. The spine of the polytope of subgraph statistics consists of the set of points $\{(p^{e_1}, p^{e_2}, \ldots, p^{e_k}) : p \in [0,1]\}$.

The spine is a coordinate project of the familiar moment curve from combinatorial geometry.

**Proposition 11.3.6.** Let $\mathcal{H} = \{H_1, \ldots, H_k\}$ be a collection of graphs, suppose that $H_1 = K_2$, suppose that graph $H_i$ has $e_i$ edges, and let $n$ be the largest number of vertices among the $H_i$. Then the spine is contained in $P_{\mathcal{H},n'}$ for all $n' \geq n$.

**Proof.** Consider an Erdős-Rényi random graph $G(n',p)$, with $n'$ vertices and probability $p$ of an edge occurring. Note that this probability distribution on graphs with $n'$ vertices is a distribution in the exponential random graph model associated to $\mathcal{H}$, and can be obtained by setting all the exponential parameters associated to graphs $H_2, \ldots, H_k$ equal 0.
The probability that a particular subgraph $H$ occurs as a subgraph on a specific set of $|H|$ vertices is $p^e$, where $e$ is the number of edges in $H$. Hence, the expected density of $E[\ell_H(G)] = p^e$ as well. This implies that the point on the spine $(p^{e_1}, p^{e_2}, \ldots, p^{e_k})$, arising as the expected value of a probability distribution on random graphs, belongs to the polytope $P_{\mathcal{H}, n'}$.

Example 11.3.7 (Spine of the Edge-Triangle Model). Consider the graph statistics from Example 11.1.5, which consist of a $K_2$ and a $K_3$. The spine in this case consists of the points $\{(p, p^3) : p \in [0, 1]\}$. This is illustrated in Figure 11.3.1 which shows the spine contained in the polytope $P_{\{K_2, K_3\}, n}$.

It remains a difficult open problem to try to give a complete description of the polytopes $P_{\mathcal{H}, n}$, or the limiting objects and $n \to \infty$. Engström and Norén [EN11] have some further work on inner approximations to these polytopes as well as conjectures related to the limiting objects. The theory of flag algebras [Raz07] and the theory of graph limits [Lov12] are closely related and also shed light on these structures.

11.4. Exercises

Exercise 11.1. Prove that the conditions of Theorem 11.2.3 are necessary for a sequence to be a degree sequence of a graph.

Exercise 11.2. Determine the formula for the normalizing constant in the Rasch model, analogous to the formula in Proposition 11.2.2 for the beta model.

Exercise 11.3. Prove Proposition 11.2.6.
Exercise 11.4. Show that the degree 2 and 3 moves suffice to provide a Markov basis for the $p_1$ model with no reciprocation. (See [KTV99]. Fast mixing of the Markov chain is even known for various classes of degree sequences, see e.g. [EMT16].)

Exercise 11.5. Determine the size of the fiber under the Edge-Triangle model for a graph with 8 vertices, 16 edges and 8 triangles.

Exercise 11.6. Construct the subgraph statistics polytope associated to the set of graphs $\{K_2, K_3, K_4 \setminus e\}$ on 7 vertices, and draw the associated spine.
Statistics is concerned with answering scientific questions using data. There might be many differing interacting covariates that can explain a response variable and one is interested in determining what these interactions might be. In a controlled experiment, the researcher tries to measure the magnitudes of different effects by setting up each of a number of possible configurations of the covariates to measure the effects of each of the random variables on the response variables. This “design of experiments” originated in agricultural and industrial experiments. The researchers were interested in knowing how various factors affected the yield of various crops. In these settings it can be prohibitively expensive to run experiments in all possible combinations of the covariates, so the researchers are interested in designing experiments that only involve testing a small subset of all possible combinations that still captures the main interaction effects of the random variables measured. The experimental design is thus a choice of the particular combinations to use to actually run the experiments on. Typically this subset is chosen with particular assumptions about what effects might matter, which effects are assumed to be zero or inconsequential, and the experimental design is chosen in such a way that all of the desired effects can be measured. Further combinatorial considerations might also be made, for instance, for every pair of covariates every combination of those pairs should be measured.

The study of experimental designs, and the combinatorial aspects which are part of combinatorial design theory, have a long tradition using tools from both algebra and combinatorics. Pistone, Riccomagno, and Wynn realized that methods from computational algebraic geometry could be used to study the design of experiments, which was the subject of the first book...
that used the title “Algebraic Statistics” [PRW01]. We discuss some of these aspects in this chapter.

12.1. Designs

From a mathematical standpoint, a design $D$ is a set of points in a finite dimensional vector space. Computational methods from algebraic geometry and commutative algebra can be used to analyze properties of designs and decide which models can be estimated using a given design.

Let $K$ be a field, and let $D \subseteq K^m$ be a finite subset of a finite dimensional vector space over $K$. Such a set is called a design.

Example 12.1.1 (Full Factorial Design). Let $D \subseteq K$ be a finite set of numbers and let $D = D^m$. Designs of this type are often called full factorial designs. The most important special cases include $D = \{0, 1\}$, $D = \{-1, 0, 1\}$, or $D = \{0, 1, \ldots, p\}$.

Example 12.1.2 (Fractional Factorial Design). If $D$ is a full factorial design, a subset $D' \subseteq D$ is sometimes called a fractional factorial design. Sometimes this expression is used to denote subsets with specific properties, though there does not appear to be a standard definition in the literature on what properties a set $D'$ must satisfy to be a fractional factorial design.

For example, if $D = \{0, 1\}^m$ and $D' \subseteq D$ consists of all $(x_1, \ldots, x_m) \in D$ such that $\sum_{i=1}^m x_i \equiv 0 \mod 2$, this is called the standard $2^{m-1}$ fractional factorial design.

A key tool in the algebraic approach to the design of experiments is the vanishing ideal $I(D) \subseteq K[x]$ of the design $D$. Note that we use the indeterminates $x_1, \ldots, x_m$ in our polynomial ring when discussing the design of experiments (as opposed to working in the polynomial ring $R[p]$ or $R[\Sigma]$ as we have done in previous chapters). This is because for a design we are actually looking at the states of random variables we are considering (rather than coordinates on the probability simplex, or in the space of symmetric matrices, etc.). Also we need not require that $K = \mathbb{R}$, and there are many situations where finite fields are relevant in the design of experiments.

A key property for vanishing ideals of finite sets is the following proposition regarding the quotient ring $K[x]/I(D)$.

**Proposition 12.1.3.** Let $D \subseteq K^m$ be a finite set. Then the quotient ring $K[x]/I(D)$ is the vector space of all functions from $D$ to $K$.

**Proof.** We have already seen in Proposition 10.3.1 that for any variety $V$, $K[x]/I(V)$ consists of all polynomial functions from $V$ to $K$. The claim in the special case of finite sets is that every function from $D$ to $K$ can be
represented as a polynomial function. To prove this, it suffices to show that for each \( y \in D \), there is a polynomial \( f \in \mathbb{K}[x] \) such that \( f(y) \neq 0 \), and \( f(z) = 0 \) for all \( z \in D \setminus \{y\} \). The resulting function \( f(z)/f(y) \) is a standard unit vectors in the space of functions from \( D \) to \( \mathbb{K} \), and the collection of such functions is then a basis for all functions from \( D \) to \( \mathbb{K} \). An example of such a polynomial that satisfies these conditions is

\[
\prod_{i=1}^{m} \prod_{z \in D, z \neq y} (x_i - z_i).
\]

The existence of such a polynomial completes the proof.

One basic goal in the design of experiments is to determine which polynomial regression models are estimable given a particular design. In this setting, we have a polynomial with generic coefficients:

\[
f = \sum_{a \in A} c_a x^a
\]

where the \( c_a \) are the coefficients and \( A \) is a specific set of exponent vectors characterizing the model. We are typically interested in modeling a response variable \( Y \) as the polynomial function \( f(X) \) of covariates \( X \) plus a noise term \( \epsilon \). The goal in such an analysis is to determine which terms in the regression are important (e.g. large values of \( c_a \)) and which terms can be ignored.

To say that a particular polynomial regression model is *estimable* given the design \( D \), means that the coefficients of the polynomial \( c_a \) can be determined from data. This is equivalent to saying that the set of functions \( \{x^a : a \in A\} \) are linearly independent functions in \( \mathbb{K}[x]/I(D) \). This is clearly the same concept as identifiability studied in detail in Chapter 16. However, for the polynomial regression models studied in this section, the methods are more straightforward.

In statistics, the estimability of the design is expressed in terms of the design matrix: \( X(A, D) \) whose \((y, a)\) entry is the evaluation of \( x^a \) at \( y \).

**Definition 12.1.4.** The polynomial regression model with support \( A \) is *estimable* from the design \( D \) if the matrix \( X(A, D) \) has full column rank.

**Example 12.1.5 (An Estimable Regression Model).** Consider the design \( D = \{(0, 0), (2, 0), (1, 3), (4, 5)\} \) and the set \( A \) consisting of exponent vectors \( A = \{(0, 0), (1, 0), (2, 0), (0, 1)\} \). The set \( A \) corresponds to the set of monomials \( \{1, x_1, x_1^2, x_2\} \). The design matrix in this case is

\[
\begin{pmatrix}
1 & 0 & 0 & 0 \\
1 & 2 & 4 & 0 \\
1 & 1 & 1 & 3 \\
1 & 4 & 16 & 5
\end{pmatrix}
\]
which has rank 4. So the polynomial regression model
\[ f(x_1, x_2) = c_{00} + c_{10} x_1 + c_{20} x_1^2 + c_{01} x_2 \]
is estimable from the design \( D \).

A basic problem in design of experiments is to determine which polynomial regressions models are estimable from a given design \( D \). Usually arbitrary polynomial regression models are not of interest. Instead, we are interested in models that satisfy certain properties that make them reasonable from a practical standpoint. A typical assumption is that the set of monomials that appears is \textit{hierarchical}. This means that if a monomial \( x^u \) appears in the expansion, then any monomial \( x^v \) such that \( x^v | x^u \) also appears in the expansion. In other words:

\textbf{Definition 12.1.6.} A set of monomials \( M \) is \textit{hierarchical} if and only if it is the set of standard monomials of some monomial ideal.

So the main problem in the algebraic theory of design of experiments is to determine which monomial ideals correspond to identifiable models with respect to a given design \( D \). A large class is obtained by looking at initial ideals.

\textbf{Proposition 12.1.7.} Let \( D \subseteq \mathbb{K}^m \) be a design, and \( \prec \) a term order. Then the standard monomials of \( \text{in}_\prec(I(D)) \) are the support of an estimable model with respect to the design \( D \).

\textbf{Proof.} By Proposition 10.3.4, the standard monomials of an initial ideal of \( I(D) \) form a \( \mathbb{K} \)-vector space basis for \( \mathbb{K}[x]/I(D) \), so these standard monomials are linearly independent modulo \( I(D) \).

\textbf{Example 12.1.8} (5 Point Design). Let \[ D = \{(0,0), (1,1), (-1,1), (1,-1), (-1,-1)\} \subseteq \mathbb{R}^2. \] The vanishing ideal of the design is \[ I(D) = (x_1^2 - x_2^2, x_1^3 - x_1, x_2^3 - x_2). \] A reduced Gröbner basis of \( I(D) \) with respect to lexicographic order with \( x_1 \prec x_2 \) is:

\[ \{x_2^2 - x_1^2, x_1^2 x_2 - x_2, x_1^3 - x_1\} \]
where the initial terms have been underlined. This means that the set of monomials \[ \{1, x_1, x_1^2, x_2, x_1 x_2\} \]
provides the support for an estimable model with respect to the design \( D \).
12.1. Designs

Computing Gröbner bases and analyzing initial ideals of the ideal of points $I(\mathcal{D})$ is a useful tool for finding identifiable polynomial regression models. Understanding which ideals can arise in this way is an interesting question. Recall that a condition is said to be generic if it holds off of a set of measure zero. In algebraic geometry a condition is generic if it holds on the complement of the union of countably many hypersurfaces (here we need to assume that our field is uncountable, to be certain that there are any points satisfying the condition). For example, we have the following proposition about generic designs.

**Proposition 12.1.9.** Let $\mathcal{D} \subseteq \mathbb{K}^m$ be a generic design where $\mathbb{K}$ is an uncountable field. Then every polynomial regression model with $\leq |\mathcal{D}|$ terms is estimable given $\mathcal{D}$.

**Proof.** Fix a number $d$ representing the size of the design. A set $\mathcal{D}$ of $d$ points in $\mathbb{K}^m$ can be thought of as the columns of an $m \times d$ matrix with entries in $\mathbb{K}$. We just need to show that there is a set of countably many hypersurfaces in $\mathbb{K}^{m \times d}$, such that off of those hypersurfaces any polynomial regression will be identifiable. For each set of $d$ exponent vectors we get the design matrix $X(A, \mathcal{D})$, which we can consider as a matrix whose entries are monomials in the coordinates on the space $\mathbb{K}^{m \times d}$. The determinant of $X(A, \mathcal{D})$ is zero if and only the model is not estimable given the design. Note that the determinant of $X(A, \mathcal{D})$ is not identically zero because each monomial in the expansion has a unique occurrence. Thus, each set $A$ gives a condition on the set of points under consideration. There are countably many sets $A$ of exponent vectors with $|\mathcal{D}|$ elements.

On the other hand, not every hierarchical set of monomials with $\mathcal{D}$ elements can arise as the set of standard monomials of an initial ideal of $I(\mathcal{D})$. In fact, for generic sets of points $\mathcal{D}$ there is a precise characterization of which initial ideals are possible.

**Definition 12.1.10.** A monomial ideal $M \subseteq \mathbb{K}[x]$ is called a corner cut ideal if there exists a vector $a \in \mathbb{R}^m_{>0}$ and $b > 0$ such that

$$M = \langle x^u : a^T u > b \rangle.$$ 

**Theorem 12.1.11.** [OS99] Let $\mathbb{K}$ be an infinite field and $\mathcal{D} \subseteq \mathbb{K}^m$ a generic set of $d$ points. A monomial ideal $M \subseteq \mathbb{K}[x]$ is an initial ideal of $I(\mathcal{D})$ if and only if $M$ is a corner cut ideal with exactly $d$ standard monomials.

Of course, most designs that arise in practice have specific combinatorial structure, arising as fractions of a full factorial design. So they are rarely generic and will have initial ideals that are not corner cut ideals. Conversely,
there will be corner cut ideals that are not estimable from the design. Furthermore, there can be more hierarchical polynomial regression models that are estimable that do not arise as an initial ideal.

**Example 12.1.12** (5 Point Design). Consider the design from Example 12.1.8. The only estimable hierarchical polynomial models for this design have monomial support

\[ \{1, x_1, x_1^2, x_2, x_1x_2\} \quad \text{and} \quad \{1, x_1, x_2^2, x_1x_2\}. \]

Both of these sets are standard monomials of corner cut ideals. However, other corner cut ideals do not yield estimable models with respect to this design. For example, the set of monomials

\[ \{1, x_1, x_2^2, x_3^2\} \]

is the set of standard monomials of a corner cut ideal but is not estimable with respect to the design \( \mathcal{D} \).

Every hierarchical set of monomials can be realized as the set of standard monomials of a design ideal \( I(\mathcal{D}) \). This can be seen by employing a construction called the distraction.

**Definition 12.1.13.** Let \( M \) be a monomial ideal such that \( \mathbb{K}[x]/M \) is a finite dimensional \( \mathbb{K} \)-vector space. For each monomial \( x^a \) we associate the product of linear polynomials

\[ x^{(a)} := \prod_{i=1}^m x_i(x_i - 1) \cdots (x_i - a_i + 1) \]

The distraction of \( M \) is the ideal

\[ \text{Dist}(M) := \langle x^{(a)} : x^a \in M \rangle. \]

**Proposition 12.1.14.** For any monomial ideal \( M \in \mathbb{K}[x] \) such that \( \mathbb{K}[x]/M \) is a finite dimensional \( \mathbb{K} \)-vector space, \( V(\text{Dist}(M)) \) is equal to the set of exponent vectors of the standard monomials of \( M \). For any term order \( \preceq \), \( \text{in}_{\preceq}(\text{Dist}(M)) = M \).

**Proof.** Since \( \mathbb{K}[x]/M \) is a finite dimensional \( \mathbb{K} \)-vector space, for each \( i \), there is \( r_i \) such that \( x_i^{(r_i)} \in M \). The polynomial

\[ x_i^{(r_i)} = x_i(x_i - 1) \cdots (x_i - r_i + 1) \]

belongs to \( \text{Dist}(M) \), so \( V(\text{Dist}(M)) \) consists of only points with nonnegative integer coordinates. Let \( x^a \) be a standard monomial of \( M \) and let \( x^a \in M \). Then there in \( i \) such that \( u_i < a_i \). This implies that \( u \in V(\{x^{(a)}\}) \), and thus
12.2. Computations with the Ideal of Points

Let $u \in V(\text{Dist}(M))$. Conversely, $a \not\in V(\{x^{(a)}\})$ so this implies that $V(\text{Dist}(M))$ is the set of exponent vectors of standard monomials of $M$.

Now let $\preceq$ be an arbitrary term order. Since each polynomial $x^{(a)}$ is a product of linear terms, each of whose leading terms is $x_i$, this forces $\text{in}_{\preceq}(x^{(a)}) = x^a$. Hence $M \subseteq \text{in}_{\preceq}(\text{Dist}(M))$. However this initial ideal $\text{in}_{\preceq}(\text{Dist}(M))$ could not be bigger then $M$, since otherwise $\text{Dist}(M)$ would have fewer standard monomials that the number of points in $V(\text{Dist}(M))$, contradicting Proposition 12.1.3.

Example 12.1.15 (A Non-Corner Cut Ideal). The set of monomials

$$\{1, x_1, x_1^2, x_2, x_2^2\}$$

is not the set of standard monomials of a corner cut. Indeed, a set of standard monomials which is a corner cut ideal that does not include $x_1x_2$ either does not include $x_1^2$ or $x_2^2$. The monomial ideal with this set of standard monomials is $M = \langle x_1^1, x_1x_2, x_2^3 \rangle$. The distraction is the ideal

$$\text{Dist}(M) = \langle x_1(x_1 - 1)(x_1 - 2), x_1x_2, x_2(x_2 - 1)(x_2 - 2) \rangle$$

and the resulting design consists of the points

$$\{(0, 0), (0, 1), (0, 2), (1, 0), (2, 0)\}.$$  

12.2. Computations with the Ideal of Points

A key role in the algebraic approach to the design of experiments is played by $I(D)$, the vanishing ideal of the finite set of points $D \subseteq \mathbb{K}^m$. To make this useful requires methods for developing quick computations with the ideal of points, in particular, computing generating sets, Gröbner bases, and interpolating polynomials. Besides design of experiments, these methods also have applications in other contexts, including to the analysis of discrete dynamical systems [DJLS07, LS04]. We describe methods for computing with the ideal of points in this section.

Let $D = \{y_1, \ldots, y_r\} \subseteq \mathbb{K}^m$ be a design. The simplest theoretical approach to computing the vanishing ideal $I(D)$ is to use the relations between unions of varieties and intersections of ideals to compute the vanishing ideal of $D$. Indeed, for the point $y_i = (y_{i1}, \ldots, y_{im}) \in D$, the vanishing ideal is

$$I(\{y_i\}) = \langle x_1 - y_{i1}, \ldots, x_m - y_{im} \rangle.$$ 

Hence, by Proposition 3.2.13, we have

$$I(D) = \cap_{i=1}^r \langle x_1 - y_{i1}, \ldots, x_m - y_{im} \rangle.$$ 

The intersection of two ideals can be computed via elimination using the following:
Proposition 12.2.1. Let $I, J \in \mathbb{K}[x_1, \ldots, x_m]$ and let $t$ be a new indeterminate. Then

$$I \cap J = (t \cdot I + (1 - t) \cdot J) \cap \mathbb{K}[x_1, \ldots, x_m].$$

The ideal $(t \cdot I + (1 - t) \cdot J)$ is considered inside the ring $\mathbb{K}[x_1, \ldots, x_m, t]$. The proof of Proposition 12.2.1 is left as an exercise. The resulting procedure for computing intersections of ideals based on elimination is implemented in most computer algebra packages. In practice this is not the most efficient way to compute the vanishing ideal of a finite set of points. A more efficient algorithm is the Buchberger-Möller algorithm [MB82] which we describe here in broad strokes. We use $\text{Mono}(\mathbb{K}[x_1, \ldots, x_m])$ to denote the set of all monomials in $\mathbb{K}[x_1, \ldots, x_m]$.

Algorithm 12.2.2. Input: A set of points $D = \{y_1, \ldots, y_r\} \subseteq \mathbb{K}^m$ and a term order. Output: A reduced Gröbner basis $G$ for $I(D)$ with respect to $\preceq$, a set $S$ of exponent vectors of standard monomials of $\mathbb{K}[x_1, \ldots, x_m]/I(D)$, and the initial ideal $M = \text{in}_{\preceq}(I(D))$.

Initialize: $S = \emptyset$, $G = \emptyset$, $M = \langle 0 \rangle$.

While: $\text{Mono}(\mathbb{K}[x_1, \ldots, x_m]) \setminus (S \cup M) \neq \emptyset$

- Let $x^a$ be the smallest monomial in $\text{Mono}(\mathbb{K}[x_1, \ldots, x_m]) \setminus (S \cup M)$ with respect to $\preceq$.
- If $\text{rank}(X(S \cup \{a\}, D)) > \text{rank}(X(S, D))$
  - Then, $S := S \cup \{a\}$
  - Else, there is a unique polynomial $g = x^a + \sum_{b \in S} c_b x^b \in I(D)$.
  Set $G = G \cup \{g\}$ and $M := M + \langle x^a \rangle$.


The algorithm terminates in finitely many steps because one never needs to look at any monomial of degree larger than $|D|$. A more detailed description of the algorithm shows how to directly compute the polynomial $x^a + \sum_{b \in S} c_b x^b$ while checking the rank condition $\text{rank}(X(S \cup \{a\}, D)) > \text{rank}(X(S, D))$ (and technically avoids any rank computations at all). See [MB82] for the details.

Once we have a method to compute the ideal of points, we can use our Gröbner basis for the ideal of points to answer various questions about the underlying designs. A first topic of consideration is the notion of aliasing of two polynomial regression models.

Definition 12.2.3. Given two finite sets of exponent vectors $A, B \subseteq \mathbb{N}^m$ defining polynomial regression models, we say they are aliased with respect to the design $D$ if the two vector spaces of functions $\{\sum_{a \in A} c_a x^a : c_a \in \mathbb{K}\}$
and \( \{ \sum_{b \in B} c_b x^b : c_b \in \mathbb{K} \} \) are the same vector space in the quotient ring \( \mathbb{K}[x]/I(\mathcal{D}) \).

Among the simplest types of designs where it is easy to analyze aliasing are when the design has a group structure. For example, consider the full factorial design \( \{\pm 1\}^m \). This design is an abelian group under coordinate-wise multiplication. Any subgroup is obtained by adding binomial constraints of the form \( x^a - 1 = 0 \), which induces a subdesign \( \mathcal{D} \subseteq \{\pm 1\}^m \). Since all equations in the resulting defining ideal are binomial, this will force aliasing relations to show that certain monomials in a polynomial regression model over the design will be equal to each other as functions.

**Example 12.2.4** (Subgroup Design). Let \( \mathcal{D} \) be the subgroup design consisting of the following 8 points in \( \mathbb{R}^5 \):

\[
\begin{align*}
(1,1,1,1,1), (1,1,1,-1,-1), (1,-1,-1,1,-1), (1,-1,-1,-1,1), \\
(-1,1,-1,1,-1), (-1,-1,-1,-1,1), (-1,-1,1,1,1), (-1,-1,1,-1,-1)
\end{align*}
\]

The vanishing ideal of the design is

\[
I(\mathcal{D}) = (x_1^2 - 1, x_2^2 - 1, x_3^2 - 1, x_4^2 - 1, x_5^2 - 1, x_1 x_2 x_3 - 1, x_3 x_4 x_5 - 1).
\]

The Gröbner basis for \( I(\mathcal{D}) \) with respect to the lexicographic term order where \( x_1 > x_2 > x_3 > x_4 > x_5 \) is

\[
\mathcal{G} = \{ x_2^2 - 1, x_4^2 - 1, x_3 - x_4 x_5, x_2^2 - 1, x_1 - x_2 x_4 x_5 \}.
\]

From the Gröbner basis we already see some aliasing relations between monomial terms in polynomial regression with respect to \( \mathcal{D} \). For example, since \( x_3 - x_4 x_5 \) is in the Gröbner basis, the two functions \( x_3 \) and \( x_4 x_5 \) are equal as functions on \( \mathcal{D} \). Similarly \( x_1 = x_2 x_4 x_5 \) as functions on \( \mathcal{D} \). But, there are many other relations besides these. Two monomials \( x^a \) and \( x^b \) represent that same function on \( \mathcal{D} \) if and only if \( x^a - x^b \) is the zero function on \( I(\mathcal{D}) \), if and only if \( x^a - x^b \in I(\mathcal{D}) \). This can be checked by determining if \( NF_{\mathcal{G}}(x^a - x^b) = 0 \) or not.

More generally, aliasing not just between individual monomials but entire polynomial regression models can be detected by performing normal form calculations. To do this, one considers the two polynomial regression models \( \{ \sum_{a \in A} c_a x^a : c_a \in \mathbb{K} \} \) and \( \{ \sum_{b \in B} c_b x^b : c_b \in \mathbb{K} \} \), computes the normal form of each monomial appearing to get equivalent representations as the models \( \text{span}(NF_{\mathcal{G}}(x^a) : a \in A) \) and \( \text{span}(NF_{\mathcal{G}}(x^b) : b \in B) \) and checks whether or not these two vector spaces are equal.
12. Design of Experiments

12.3. The Gröbner Fan and Applications

As we have seen in Section 12.1, variation of term order allows for the choice of different estimable models with respect to a given design. The results of all possible term orders are encoded in a polyhedral object called the Gröbner fan, whose structure can be useful for choosing estimable models that satisfy certain features. We describe the Gröbner fan here and discuss some of its applications. More details on the Gröbner fan and the related state polytope can be found in [Stu95, Chapter 2].

Recall the definition of a weight order discussed in Chapter 10. A vector $c$ defines a weight order $\prec_c$ on the set of monomials in $\mathbb{K}[x]$ by $x^u \prec_c x^v$ if and only if $c^T u < c^T v$. The weight of a monomial $x^u$ with respect to $c$ is $c^T u$. Given a polynomial $f \in \mathbb{K}[x]$, its initial form $\text{in}_c(f)$ with respect to the weight order $c$ is the sum of all terms of $f$ that have the highest weight. Note that the weight order $\prec_c$ might not be a term order because there can be monomials that have the same weight, and the monomial 1 might not be the smallest weight monomial with respect to the weight order. Given an ideal $I \subseteq \mathbb{K}[x]$ we can also define the initial ideal

$$\text{in}_c(I) = \langle \text{in}_c(f) : f \in I \rangle.$$ 

Not every term order can be realized by a weight order. For example, for the lexicographic term order, there is no weight vector $c \in \mathbb{R}^m$ such that $\text{in}_c(f) = \text{in}_{\text{lex}}(f)$ for all $f \in \mathbb{K}[x]$. However, once we fix a particular ideal and focus on computing the initial ideal, it is always possible to find a weight vector that realizes a particular term order.

**Proposition 12.3.1.** Let $I \subseteq \mathbb{K}[x]$ be an ideal and $\prec$ be a term order on $\mathbb{K}[x]$. Then there exists a weight vector $c \in \mathbb{R}^m$ such that $\text{in}_c(I) = \text{in}_\prec(I)$.

**Proof.** Let $\mathcal{G} = \{g_1, \ldots, g_k\}$ be a reduced Gröbner basis for $I$ with respect to the term order $\prec$. Each polynomial $g_i \in \mathcal{G}$ can be written as

$$g_i = x^{u_i} + \sum_{j=1}^{l_i} c_{ij} x^{v_{ij}}$$

where each of the monomials $x^{u_i}$ is larger in the term order than any $x^{v_{ij}}$ for $j = 1, \ldots, l_i$. The set of all possible cost vectors that could produce the initial ideal $\text{in}_\prec(I)$ is the open polyhedral cone

$$\{c \in \mathbb{R}^m : c^T u_i > c^T v_{ij} \text{ for } i \in [k], j \in [l_i] \}.$$ 

The proof of the proposition amounts to showing that this cone is nonempty. This can be shown via an application of the Farkas Lemma of convex geometry: if the system of inequalities were inconsistent, the ordering $\prec$ would not be a term order. $\square$
We say that a weight vector \( c \in \mathbb{R}^m \) is in the \textit{Gröbner region} of the ideal \( I \), if it is in the closure of the set of all cost vectors for which there exists a term order \( \prec \) on \( \mathbb{K}[x] \) such that \( \text{in}_c(I) = \text{in}_{\prec}(I) \). If \( c \in \mathbb{R}^m \) with all positive entries that are linearly independent over the rational numbers, then \( c \) defines a term order on \( \mathbb{K}[x] \) (regardless of the ideal). So this shows that the Gröbner region of an ideal contains the positive orthant.

The proof of Proposition 12.3.1 shows that the set of weight vectors that realize a specific monomial initial ideal of an ideal \( I \in \mathbb{K}[x] \) is an open polyhedral cone. For any term order \( \prec \), let \( C(\prec) \) denote the closure of this cone, the \textit{Gröbner cone} of the term order.

**Example 12.3.2 (A Small Gröbner Cone).** Consider the ideal \( I = \langle x_1^2 - x_2, x_2^2 - x_1 \rangle \). Then two polynomial generators are a Gröbner basis of \( I \) with respect to any graded term order (e.g. graded reverse lexicographic). The set of possible weight vectors that realize the same initial ideal is the cone

\[
\{(c_1, c_2) \in \mathbb{R}^2 : 2c_1 > c_2, 2c_2 > c_1\},
\]

which is the Gröbner cone \( C(\prec_{(1,1)}) \).

The Gröbner region is a union of polyhedral cones, one for each distinct initial ideal of \( I \). In fact, there are only finitely many Gröbner cones in the Gröbner region.

**Theorem 12.3.3.** For any ideal \( I \in \mathbb{K}[x] \) there are only finitely many distinct monomial initial ideals of \( I \). Hence the Gröbner region of \( I \) is the union of finitely many polyhedral cones.

**Proof.** The set of all monomial ideals in a polynomial ring is partially ordered by inclusion. Since the set of standard monomials of an initial ideal of an ideal are a basis for the quotient ring (Proposition 10.3.4), two different initial ideals of the same monomial ideal must be incomparable in this containment order. However, there are no infinite sets of monomial ideals that are pairwise incomparable [Mac01]. So the set of distinct initial monomial ideals of \( I \) is finite. \( \square \)

The cones of the Gröbner region are arranged with respect to each other in a nice way.

**Definition 12.3.4.** Let \( C \) be a set of polyhedra in \( \mathbb{R}^m \). This set is called a \textit{polyhedral cell complex} if it satisfies the following conditions:

1. \( \emptyset \in C \).
2. If \( F \in C \) and \( F' \) is a face of \( F \) then \( F' \in C \).
3. If \( F, F' \in C \) then \( F \cap F' \) is a face of both \( F \) and \( F' \).
A polyhedral complex $C$ is called a fan if all the polyhedra in $C$ are cones with apex at the origin.

**Proposition 12.3.5.** Let $C_1$ and $C_2$ be two Gröbner cones of the ideal $I \subseteq K[x]$. Then $C_1 \cap C_2$ is a common face of both cones. In particular, the set of all the Gröbner cones and their faces forms a fan called the Gröbner fan of the ideal $I$, denoted $GF(I)$.

**Proof.** See [Stu95, Chapter 2].

A fan is called complete if the union of all the cones in the fan is $\mathbb{R}^m$. Note that the Gröbner fan need not be complete in general, but they are in special cases.

**Proposition 12.3.6.** Let $I \subseteq K[x]$ be a homogeneous ideal. Then the Gröbner fan of $I$ is a complete fan.

**Proof.** Let $I$ be homogeneous. Let $c \in \mathbb{R}^m$, let $\alpha \in \mathbb{R}$, and let $1 \in \mathbb{R}^m$ denote the vector of all ones. Then for any homogeneous polynomial $f \in I$, $\text{in}_c(f) = \text{in}_{c+\alpha 1}(f)$. By choosing $\alpha$ very large, we see that for any weight vector $c$, there is a vector of all positive entries that realizes the same initial ideal. Since vectors in the positive orthant are in the Gröbner region for any ideal, this implies that every vector is in the Gröbner region.

More generally, the Gröbner fan will be a complete fan if the ideal is homogeneous with respect to a grading that gives every variable a positive weight. The simplest situation to think about the Gröbner fan of an ideal is in the case of principal ideals. This requires a few definitions.

**Definition 12.3.7.** Let $f = \sum_{u \in \mathbb{N}^m} c_u x^u \in K[x]$ be a polynomial. The Newton polytope of $f$ is the convex hull of all exponent vectors of monomials appearing in $f$; that is,

$$\text{Newt}(f) = \text{conv}(u : c_u \neq 0).$$

Let $P$ be a polytope. To each face $F \subseteq P$, let $C(F)$ be the set of cost vectors such that $F$ is the optimal set of points in $P$:

$$C(F) = \{ c \in \mathbb{R}^m : c^T x \geq c^T y \text{ for all } x \in F \text{ and } y \in P \}.$$  

The set $C(F)$ is a polyhedral cone, since it can be defined by finitely many linear inequalities (e.g. by taking all the inequalities $c^T x \geq c^T y$ where $x \in \text{vert}(F)$ and $y \in \text{vert}(P)$).

**Proposition 12.3.8.** Let $P$ be a polytope. The set of cones $N_P = \{ C(F) : F \text{ if a face of } P \}$ is a complete fan called the normal fan of the polytope $P$. 

12.3. The Gröbner Fan and Applications

Figure 12.3.1. The Newton polytope of the polynomial of Example 12.3.10 and its normal fan.

Proof. See [Zie95, Example 7.3].

The normal fan of a polytope can be thought of as the parametric solution to all possible linear programs over that polytope. As the cost vector varies, the optimal solution moves from one vertex of $P$ to another.

Proposition 12.3.9. Let $f \in \mathbb{K}[x]$. Then $\text{GF}(\langle f \rangle)$ consists of all cones in the normal fan of the Newton polytope of $f$ that intersect the positive orthant. If $f$ is homogeneous then $\text{GF}(\langle f \rangle) = \mathcal{N}_{\text{Newt}(f)}$.

Proof. Every monomial initial ideal of $\langle f \rangle$ arises from taking a vertex of the Newton polytope. The corresponding Gröbner cone associated to a given weight vector is the cone of the normal fan that realizes that vertex as the optimum. Vectors from the normal fan will actually yield term orders if and only if the corresponding normal cone intersects the positive orthant.

Example 12.3.10 (Pentagon). Consider the polynomial $f = 3 + 7x_1 + 2x_1x_2 - 4x_1^2x_2 + 6x_1x_2^3 + x_1^3x_2^3$. The Newton polytope $\text{Newt}(f)$ is a pentagon, and the normal fan $\mathcal{N}_{\text{Newt}(f)}$ has five 2-dimensional cones. The Gröbner fan consists of the two of those cones that intersect the positive orthant. The Newton polytope and its normal fan are illustrated in Figure 12.3.1.

Example 12.3.11 (Computation of a Gröbner Fan). The following Macaulay2 code uses the package \texttt{gfan} [Jen] to compute the Gröbner fan and all monomial initial ideals of the ideal $I = \langle x_1^2 - x_2^2, x_1^3 - x_1, x_2^3 - x_2 \rangle$ from Example 12.1.8.

```markdown
loadPackage "gfanInterface"
R = QQ[x1,x2];
I = ideal(x1^2-x2^2, x1^3-x1, x2^3-x2);
gfan I
groebnerFan I
```
274  12. Design of Experiments

The first command \texttt{gf}an \texttt{I} computes all the distinct initial ideals of \texttt{I} and also shows the corresponding reduced Gröbner bases. The second command \texttt{groebnerFan} \texttt{I} computes a representation of the Gröbner fan as the union of polyhedral cones. In this case, there are two distinct initial ideals:

\[
\langle x_1^2, x_1 x_2, x_2^3 \rangle \quad \text{and} \quad \langle x_1^1, x_1^1 x_2, x_2^2 \rangle
\]

corresponding to the two estimable sets of monomials seen in Example 12.1.12. The Gröbner fan consists of two cones:

\[
\{(c_1, c_2) \in \mathbb{R}^2 : c_1 \geq c_2 \geq 0\} \quad \text{and} \quad \{(c_1, c_2) \in \mathbb{R}^2 : c_2 \geq c_1 \geq 0\}
\]
corresponding to the first and second initial ideal, respectively. The Gröbner region is the positive orthant.

We saw in Proposition 12.3.9 that the Gröbner fan of a principal ideal \(\langle f \rangle\) can be realized as (part of) the normal fan of a polytope, namely, the Newton polytope of the \(f\). One might wonder if the Gröbner fan always has such a nice polyhedral realization. A polytope \(P\) such that \(NF(P)\) is the Gröbner fan of an ideal \(I\) is called a state polytope for \(I\). Every homogeneous ideal has a state polytope [MR88] but they do not need to exist for nonhomogeneous ideals, since the Gröbner fan need not be complete.

Since the positive orthant usually has the most interesting weight vectors for computing Gröbner bases, we could restrict to the intersection of the Gröbner fan with the positive orthant. For this restricted Gröbner fan, we could ask if there is a state polyhedron whose normal fan is the resulting restricted Gröbner fan. Denote by \(GF_r(I)\) the restricted Gröbner fan of \(I\). State polyhedra for the restricted Gröbner fan need not exist in general [Jen07]. However, state polyhedra do always exist for the ideals \(I(D)\) where \(D\) is a finite set of any affine space (over any field), and there is a straightforward construction. This construction can be phrased in terms of estimable staircase models associated to the design \(D\).

**Definition 12.3.12.** Let \(D \subseteq \mathbb{K}^m\) be a design. An estimable polynomial regression model for \(D\) is saturated if it has the same number of terms as \(|D|\). Let \(\text{Est}(D)\) denote the set of estimable saturated hierarchical polynomial regression models for the design \(D\).

**Theorem 12.3.13.** [OS99, Thm 5.2] Let \(D \subseteq \mathbb{K}^m\) be a finite set. To each \(S \in \text{Est}(D)\) associate the vector \(v_S = \sum_{s \in S} s\). Then the polyhedron

\[
\text{conv}\{\{-v_S : S \in \text{Est}(D)\}\} + \mathbb{R}_{\geq 0}^m,
\]

is a state polyhedron for the restricted Gröbner fan of \(I(D)\).

**Example 12.3.14** (A Generic Design). Suppose that \(D \subseteq \mathbb{R}^2\) is a generic design with four elements. There are five estimable saturated hierarchical
12.3. The Gröbner Fan and Applications

Figure 12.3.2. The state polyhedron of the ideal of four generic points of Example 12.3.14 and the corresponding restricted Gröbner fan.

designs, whose monomials are
\{1, x_1, x_1^2, x_1^3\}, \{1, x_1, x_1^2, x_2\}, \{1, x_1, x_2, x_1x_2\}, \{1, x_1, x_2, x_2^2\}, \{1, x_2, x_2^2, x_3^2\}.

Their corresponding vectors vs are
(6, 0), (3, 1), (2, 2), (1, 3), (0, 6)
respectively. The state polyhedron and restricted Gröbner fan are illustrated in Figure 12.3.2. Note that the point (-2, -2) does not appear as a vertex of the state polyhedron, because the monomial ideal with standard monomial set \{1, x_1, x_2, x_1x_2\} is not a corner cut ideal.

The state polyhedron/Gröbner fan of a design is a useful object associated to the design because it parametrizes all the estimable models for the given design which have an algebraic representation. One application of this machinery is to regression models of minimal aberration.

Definition 12.3.15. Consider a linear model consisting of exponent vector set \(B \subseteq \mathbb{N}^m\) with \(d\) elements. Let \(w \in \mathbb{R}_{\geq 0}^m\) be a vector of nonnegative weights such that \(\sum_{i=1}^m w_i = 1\). The weighted linear aberration of the model is

\[ A(w, B) = \frac{1}{d} \sum_{b \in B} w^T b. \]

Theorem 12.3.16. [BMAO+10] Given a design \(D \subset \mathbb{R}^m\) with \(r\) points and a weight vector \(w \in \mathbb{R}_{\geq 0}^m\), there is at least one estimable hierarchical polynomial regression model \(B\) that minimizes \(A(w, B)\). Such a model can be chosen to be the set of standard monomials of an initial ideal of \(I(D)\). In particular, such a model can be obtained by computing a Gröbner basis with respect to the weight vector \(w\), or, equivalently, maximizing the linear functional \(w\) over the state polyhedron of \(w\).

Proof. First of all, note that if a set of exponent vectors \(B\) is not hierarchical, then there is an exponent vector \(b \in B\) and an exponent vector \(b' \notin B\)
such that $b' \leq b$ coordinate-wise. Since $w \in \mathbb{R}_{>0}^m$, we have

$$A(w, B) \geq A(w, B \setminus \{b\} \cup \{b'\}).$$

Furthermore, if $B$ was estimable, we can choose $b'$ in such a way that $B \setminus \{b\} \cup \{b'\}$ is also estimable. Hence, if we are looking for a model of minimal weighted linear aberration, we can restrict to hierarchical sets of exponent vectors.

Minimizing $A(w, B)$ over all estimable hierarchical polynomial regression models is equivalent to maximizing $-w^Tv_B$ where $v_B$ are the vector associated to a monomial ideal defined in Theorem 12.3.13. By Theorem 12.3.13, this amounts to maximizing over the vertices of the state polyhedron of $I(D)$. Since an optimum can always be attained at a vertex, this means that a minimum aberration model can always be taken to correspond to one coming from the term order induced by $w$.

Note that if we also want to find a design and a polynomial regression model that has as small an aberration as possible, among all designs with a fixed number of points $r$, this is also provided by a combination of Theorems 12.3.16 and 12.1.11. Indeed, on a fixed number of points $r$, the largest state polytope possible is obtained by taking a design $D$ whose initial ideals are precisely all the corner cut ideals with $r$ standard monomials. Hence, simply taking a generic design will achieve the desired optimality.

12.4. 2-level Designs and System Reliability

In this section we focus on certain two level designs, their connections to system reliability, and how tools from commutative algebra can be used to gain insight into them. Key tools are the study of squarefree monomial ideals and their Hilbert series.

Consider a system consisting of $m$ components, $[m]$, and suppose that for the system to function properly, certain subsets of the components must not fail simultaneously. We suppose that the set of failure modes is closed upwards, so if $S \subseteq [m]$ is a failure mode, and $S \subseteq T \subseteq [m]$ then $T$ is also a failure mode. Thus, the set $\Gamma$ of all subsets of $[m]$ whose failure does not cause the system to fail is a simplicial complex, as defined in Section 9.3.

**Example 12.4.1** (Network Connectivity). Consider the graph with 5 vertices and 7 edges as shown in Figure 12.4.1. The vertices $s$ and $t$ are the source and receiver of a transmission which will be sent over the network. The various edges in the network work or fail each with their own independent probability. This system will function (i.e. be able to successfully transmit a message from $s$ to $t$) if the set of failed edges do not disconnect $s$ and $t$. The set of failure modes which still allow the system to function will
form a simplicial complex, since failure states that break the system will still break the system if further elements are added. The list of minimal failure states that break the system are minimal sets of edges that disconnect $s$ and $t$. These are

$12, 67, 234, 457, 1357, 2356$

By associating to each subset $S \subseteq [m]$ the 0/1 indicator vector $u_S \in \mathbb{R}^m$, the reliability system naturally corresponds to a design. As in previous sections we could associate the vanishing ideal $I(\mathcal{D})$ to the design. However, it becomes substantially more useful to associate an alternate ideal, the Stanley-Reisner ideal, to the resulting simplicial complex.

**Definition 12.4.2.** Let $\Gamma \subseteq 2^{|m|}$ be a simplicial complex on $m$ vertices. The **Stanley-Reisner ideal** associated to $\Gamma$ is the ideal

$$\text{SR}_{\Gamma} := \langle x^{|u_S|} : S \notin \Gamma \rangle.$$  

That is, $\text{SR}_{\Gamma}$ is generated by monomials corresponding to the subsets of $[m]$ that are not faces of the complex $\Gamma$.

Note that we only need to consider the minimal nonfaces of $\Gamma$ when writing down the generating sets of $\text{SR}_{\Gamma}$. The Stanley-Reisner ideal is a squarefree monomial ideal, in particular, it is a radical ideal, and its variety $V(\text{SR}_{\Gamma})$ is a union of coordinate subspaces. It is easy to describe the primary decomposition of this ideal.

**Proposition 12.4.3.** Let $\Gamma$ be a simplicial complex on $m$ vertices. Then

$$\text{SR}_{\Gamma} = \bigcap_{S \in \text{facet}(\Gamma)} \langle x_i : i \notin S \rangle.$$  

In other words, the variety $V(\text{SR}_{\Gamma})$ consists of the union of those coordinate subspaces, one for each facet of $S$, and such that nonzero coordinates appear precisely at the coordinates corresponding to the elements of the facet.
Note that the semialgebraic set $V(SR_\Gamma) \cap \Delta_{m-1}$ gives a geometric realization of the simplicial complex $\Gamma$.

**Proof.** Since $SR_\Gamma$ is a monomial ideal generated by squarefree monomials it is a radical ideal. Hence, the primary decomposition of $SR_\Gamma$ is automatically a decomposition into prime monomial ideals, and it suffices to understand the geometric decomposition of the variety $V(SR_\Gamma)$. Let $a \in \mathbb{R}^m$, and suppose that the coordinates of $a$ where $a_i = 0$ are indexed by the set $S \subseteq [m]$. Note that $a \in V(SR_\Gamma)$, if and only if $S \cap T \neq \emptyset$ for all $T$ that are nonfaces of $\Gamma$. In particular, this shows that $S$ must be the complement of a face of $\Gamma$, since if $S$ is the complement of a nonface $T$ then $S \cap T = \emptyset$. In other words, the support of any $a \in V(SR_\Gamma)$ is a face of $\Gamma$. The maximal such support sets correspond to facets, which yield the ideals arising in the primary decomposition. □

**Example 12.4.4** (Network Connectivity). Consider the network connectivity problem from Example 12.4.1 with associated simplicial complex of failure modes that allow the system to transmit messages $\Gamma$. The Stanley-Reisner ideal $SR_\Gamma$ is generated by all of the monomials corresponding to minimal system failures in Example 12.4.1, so

$$
RS_\Gamma = (x_1x_2, x_6x_7, x_2x_3x_4, x_4x_5x_7, x_1x_3x_5x_7, x_2x_3x_5x_6).
$$

This Stanley-Reisner ideal has primary decomposition

$$
RS_\Gamma = \langle x_2, x_7 \rangle \cap \langle x_1, x_3, x_7 \rangle \cap \langle x_1, x_4, x_6 \rangle \cap \langle x_2, x_5, x_6 \rangle \cap \\
\langle x_1, x_3, x_5, x_6 \rangle \cap \langle x_2, x_3, x_4, x_6 \rangle \cap \langle x_1, x_4, x_5, x_7 \rangle.
$$

Note that the variety of each of the minimal primes of $RS_\Gamma$ corresponds to a maximal set of edges that could fail but so that the network will continue to transmit messages. Conversely, the generators of the minimal primes correspond to minimal sets of edges which, when all function, will allow the network to transmit messages. In this particular problem, these are precisely the simple paths that connected $s$ with $t$.

A key problem in reliability theory is to determine what is the probability that the system actually functions. An idealized setting is to imagine that the components function or do not function independently of each other, and component $i$ has probability $p_i$ of functioning. The probability of the system functioning is then given by the following formula:

$$(12.4.1) \quad \sum_{S \subseteq \Gamma} \left( \prod_{i \in S} p_i \prod_{j \in S} (1 - p_j) \right).$$

This formula is related to the Stanley-Reisner ideal through its Hilbert series.
12.4. 2-level Designs and System Reliability

Definition 12.4.5. Let $M \subseteq \mathbb{K}[x_1, \ldots, x_m]$ be a monomial ideal. The multigraded Hilbert series of the quotient ring $\mathbb{K}[x_1, \ldots, x_m]/M$ is the multivariable power series:

$$H(\mathbb{K}[x_1, \ldots, x_m]/M; x) = \sum_{x^n \in M} x^n.$$

If the Hilbert series of $H(\mathbb{K}[x_1, \ldots, x_m]/M; x)$ is expressed as a rational function in the form

$$H(\mathbb{K}[x_1, \ldots, x_m]/M; x) = \frac{K(\mathbb{K}[x_1, \ldots, x_m]/M; x)}{(1 - x_1) \cdots (1 - x_m)}$$

then $K(\mathbb{K}[x_1, \ldots, x_m]/M; x)$ is the K-polynomial of $\mathbb{K}[x_1, \ldots, x_m]/M$.

We can also define the multigraded Hilbert series for any multigraded module, though we will not need this here (see [MS04]). For monomial ideals $M$, the multigraded Hilbert series is defined by

$$H(M; x) = \sum_{x^n \in M} x^n.$$

Hilbert series and reliability are connected via the following theorem and its corollary.

Theorem 12.4.6. The K-polynomial of the quotient ring $\mathbb{K}[x_1, \ldots, x_m]/RS_\Gamma$ is

$$(12.4.2) \quad K(\mathbb{K}[x_1, \ldots, x_m]/RS_\Gamma; x) = \sum_{S \in \Gamma} \left( \prod_{i \in S} x_i \prod_{j \notin S} (1 - x_j) \right).$$

Proof. The Hilbert series of $\mathbb{K}[x_1, \ldots, x_m]/RS_\Gamma$ is the sum over all monomials that are not in $RS_\Gamma$. Each such monomial $x^n$ has a vector $a$ whose support $\text{supp}(a)$ is a face of $\Gamma$ (where $\text{supp}(a) = \{ i \in [m] : a_i \neq 0 \}$). Hence we can write

$$H(\mathbb{K}[x_1, \ldots, x_m]/RS_\Gamma; x) = \sum \{ x^n : a \in \mathbb{N}^n \text{ and } \text{supp}(a) \in \Gamma \}$$

$$= \sum_{S \in \Gamma} \sum \{ x^n : a \in \mathbb{N}^n \text{ and } \text{supp}(a) = S \}$$

$$= \sum_{S \in \Gamma} \prod_{i \in S} x_i \prod_{j \notin S} (1 - x_j).$$

The formula for the K-polynomial is obtained by putting all terms over the common denominator $\prod_{i=1}^m (1 - x_i)$. \hfill $\square$

Combining Theorem 12.4.6 with Equation 12.4.1 we immediately see the following.
Corollary 12.4.7. Let $\Gamma$ be a simplicial complex of functioning failure modes of a reliability system. Suppose that each component functions independently with probability $p_i$. Then the probability that the system functions is

$$K(\mathbb{K}[x_1, \ldots, x_m]/SR_\Gamma; p).$$

While Corollary 12.4.7 demonstrates the connection of the K-polynomial to the systems reliability problem, the formula of Equation 12.4.1 is typically an inefficient method for actually calculating the K-polynomial, since it involves determining and summing over every element of $\Gamma$. For applications to problems of system reliability, one needs a more compact format for the K-polynomial. A straightforward approach that can save time in the computation is to use the principle of inclusion and exclusion to compute the multigraded Hilbert series of the quotient $\mathbb{K}[x_1, \ldots, x_m]/M$ for a monomial ideal $M$.

Example 12.4.8 (K-Polynomial via Inclusion/Exclusion). Let

$$M = \langle x_1^2 x_2, x_2^3 x_3^2 \rangle.$$ 

We want to enumerate the monomials not in $M$. Clearly, this is the set of all monomials in $\mathbb{K}[x_1, x_2, x_3]$ minus the monomials in $M$, so that

$$H(\mathbb{K}[x_1, x_2, x_3]/M; x) = H(\mathbb{K}[x_1, x_2, x_3]; x) - H(M; x).$$

The Hilbert series of $\mathbb{K}[x_1, x_2, x_3]$ is

$$H(\mathbb{K}[x_1, x_2, x_3]; x) = \frac{1}{(1 - x_1)(1 - x_2)(1 - x_3)}.$$

Similarly, if we have a principal monomial ideal $\langle x^a \rangle$ then

$$H(\langle x^a \rangle; x) = \frac{x^a}{(1 - x_1)(1 - x_2)(1 - x_3)}.$$

Since our ideal is generated by two monomials, we can reduce to the principal case using the principle of inclusion and exclusion. Monomials in $M$ are either divisible by $x_1^2 x_2$ or $x_2^3 x_3^2$ or both. If they are divisible by both, they are divisible by the least common multiple. Putting these all together yields the following as the Hilbert series of $\mathbb{K}[x_1, x_2, x_3]/M$:

$$H(\mathbb{K}[x_1, x_2, x_3]/M; x) = \frac{1 - x_1^2 x_2 - x_2^3 x_3^2 + x_1^2 x_2^3 x_3^2}{(1 - x_1)(1 - x_2)(1 - x_3)}.$$

More generally, we have the following general form for the multigraded Hilbert series of a monomial ideal using the principle of inclusion and exclusion. The proof is left as an exercise.
Theorem 12.4.9. Let $M = (x^{a_1}, \ldots, x^{a_k})$. Then

$$K(\mathbb{K}[x_1, \ldots, x_m]/M; x) = \sum_{S \subseteq [m]} (-1)^{|S|} LCM(\{x^{a_i} : i \in S\}).$$

Theorem 12.4.9 can be useful for computing the $K$ polynomial in small examples, however it is clearly exponential in the number of generators (of course, it might not be possible to avoid exponential complexity, however). For instance, in Example 12.4.4, the monomial ideal $SR_T$ has six generators so the inclusion exclusion formula will involve summing over 64 terms, whereas expanding this expression and cancelling out, and collecting like terms yields 27 terms in total in the $K$-polynomial.

Developing methods for computing the $K$-polynomials of monomial ideals is an active and complex area that we will not cover here. A key tool for computing the $K$-polynomial is the free resolution of the resulting quotient ring. This is an important topic in combinatorial commutative algebra [MS04]. Applications to reliability theory include in tree percolation [MSdCW16] and developing formulas for failure probabilities for communication problems on networks [Moh16].

12.5. Exercises

Exercise 12.1. Suppose that $D = D^m \subseteq \mathbb{K}^m$ is a full factorial design. Show that the only polynomial regression model that is estimable is the model with support set:

$$\{x^u : u_i < |D| \text{ for all } i\}.$$

Exercise 12.2. How many corner cut ideals $M \subseteq \mathbb{K}[x_1, x_2]$ are there such that $M$ has exactly 7 standard monomials?

Exercise 12.3. Prove Proposition 12.2.1. Design an algorithm that takes as input the generating sets of two ideals $I$ and $J$ and outputs a generating set for $I \setminus J$.

Exercise 12.4. Consider the design $D = \{(0, 0, 0), (\pm 2, 0, 0), (0, \pm 2, 0), (0, 0, \pm 2), (\pm 1, \pm 1, 0), (\pm 1, 0, \pm 1), (0, 0, 0)\} \subseteq \mathbb{R}^3$ consisting of 19 points in $\mathbb{R}^3$. Compute the vanishing ideal of the design, determine the number of distinct initial ideals, and calculate the Gröbner fan.

Exercise 12.5. Let $D = D^m \subseteq \mathbb{K}^m$ be a full factorial design, $D' \subseteq D$ a fraction, and $D'' = D \setminus D'$ be the complementary fraction. Let $\prec$ be a term order. Explain how in $\prec(I(D'))$ and in $\prec(I(D''))$ are related [MASdCW13].
Exercise 12.6. Consider the graph from Example 12.4.1 but now consider a system failure to be any failure of edges that makes it impossible to transmit messages between any pair of vertices. If each edge $i$ has probability $p_i$ of functioning, calculate the probability that the system does not fail.

Exercise 12.7. Prove Theorem 12.4.9.
Chapter 13

Graphical Models

In this chapter we describe graphical models, an important family of statistical models that are used to build complicated interaction structures between many random variables by specifying local dependencies between pairs or small subsets of the random variables. This family of statistical models has been widely used in statistics [HEL12, Whi90], computer science [KF09], computational biology [SM14] and many other areas. A standard reference for a mathematical treatment of graphical models is [Lau96]. Graphical models will play an important role throughout the remainder of the book as they are an important example about which we can ask very general questions. They are also useful from a computational standpoint. Special cases of graphical models include the phylogenetic models from Chapter 15 and the hidden markov models which are widely used in biological applications and discussed in Chapter 18.

13.1. Conditional Independence Description of Graphical Models

Let \( X = (X_v \mid v \in V) \) be a random vector and \( G = (V, E) \) a graph with vertex set \( V \) and some set of edges \( E \). In the graphical model associated to the graph \( G \) each edge \((u, v) \in E\) of the graph denotes some sort of dependence between the variables \( X_u \) and \( X_v \). The type of graph that is used (for example, undirected graph, directed graph, chain graph [AMP01, Drt09a], mixed graph [STD10], ancestral graph [RS02], etc.) will determine what is meant by “dependence” between the variables, and often this can be expressed as a parametrization of the model, which will be explained in more detail in Section 13.2. Since the edges correspond to dependence between
the random variables, the non-edges will typically correspond to independence or conditional independence between collections of random variables. It is this description which we begin with in the present section. We focus in this chapter on the two simplest versions of graphical models, those whose underlying graph is an undirected graph, and those whose underlying graph is a directed acyclic graph.

Let $G = (V, E)$ be an undirected graph. Formally this means that $E$ is a set of ordered pairs $(u, v)$ with $u, v \in V$ and such that if $(u, v) \in E$ then so is $(v, u)$. The pairs $(u, v) \in E$ are called the edges of the graph and the set $V$ is the set of vertices of the graph. We will assume that the graph has no loops, that is, for no $v \in V$ is the pair $(v, v) \in E$.

Associated to the undirected graph are various Markov properties which are lists of conditional independence statements that must be satisfied by all random vectors $X$ consistent with the graph $G$. To describe these Markov properties, we need some terminology from graph theory. A path between vertices $u$ and $w$ in an undirected graph $G = (V, E)$ is a sequence of vertices $u = v_1, v_2, \ldots, v_k = w$ such that each $(v_{i-1}, v_i) \in E$. A pair of vertices $a, b \in V$ is separated by a set of vertices $C \subseteq V \setminus \{a, b\}$ if every path from $a$ to $b$ contains a vertex in $C$. If $A, B, C$ are disjoint subsets of $V$, we say that $C$ separates $A$ and $B$ if $a$ and $b$ are separated by $C$ for all $a \in A$ and $b \in B$. The set of neighbors of a vertex $v$ is the set $N(v) = \{u : (u, v) \in E\}$.

**Definition 13.1.1.** Let $G = (V, E)$ be an undirected graph.

(i) The **pairwise Markov property** associated to $G$ consists of all conditional independence statements $X_u \indep X_v | X_{V \setminus \{u, v\}}$ where $(u, v)$ is not an edge of $G$.

(ii) The **local Markov property** associated to $G$ consists of all conditional independence statements $X_v \indep X_{V \setminus (N(v) \cup \{v\})} | X_{N(v)}$ for all $v \in V$.

(iii) The **global Markov property** associated to $G$ consists of all conditional independence statements $X_A \indep X_B | X_C$ for all disjoint sets $A, B, C$ such that $C$ separates $A$ and $B$ in $G$.

Let pairwise($G$), local($G$), and global($G$) denote the set of pairwise, local, and global Markov statements associated to the graph $G$.

Note that the Markov properties are arranged in order of increasing strength. If a distribution satisfies the global Markov property associated to the graph $G$, it necessarily satisfies the local Markov property on the graph $G$. And if a distribution satisfies the local Markov property on the graph $G$, it necessarily satisfies the pairwise Markov property. The reverse implication is not true in general.
Example 13.1.2 (Comparison of Pairwise and Local Markov Properties). Let $G$ be the graph with vertex set $V = \{3\}$ and the single edge $2 - 3$. The pairwise Markov property for this graph yields the two conditional independence statements $X_1 \perp X_2 | X_3$ and $X_1 \perp X_3 | X_2$. On the other hand, both the local and global Markov property for this graph consists of the single statement $X_1 \perp (X_2, X_3)$. For discrete random variables, there are distributions that satisfy the pairwise statements and not the local statements, in particular, the distributions that fail to satisfy the intersection axiom (see Proposition 4.1.5 and Theorem 4.3.3).

Example 13.1.3 (Path of Length 5). Consider the path of length five $P_5$ with vertex set $\{5\}$ and edges $\{1,2\}, \{2,3\}, \{3,4\}, \{4,5\}$. In this example the pairwise, local, and global Markov properties associated to $P_5$ are all different. For example, the local Markov statement associated to the vertex 3, which is $X_3 \perp (X_1, X_5) | (X_2, X_4)$ is not implied by the pairwise Markov statements. The global Markov statement $(X_1, X_2) \perp (X_4, X_5) | X_3$ is not implied by the local Markov statements.

Characterizations of precisely which graphs satisfy the property that the pairwise Markov property equals the local Markov property, and the local Markov property equals the global Markov property can be found in [Lau96]. Remarkably, the failure of the intersection axiom is the only obstruction to the probabilistic equality of the three Markov properties.

Theorem 13.1.4. If the random vector $X$ has a joint distribution $P$ that satisfies the intersection axiom, then $P$ obeys the pairwise Markov property for the undirected graph $G$ if and only if it obeys the global Markov property for the graph $G$. In particular, if $P(x) > 0$ for all $x$ then the pairwise, local, and global Markov properties are all equivalent.

Proof. ($\Leftarrow\Rightarrow$): If $u$ and $v$ are vertices in $G$ that are not adjacent, then $V \setminus \{u, v\}$ separates $\{u\}$ and $\{v\}$ in $G$. Hence the pairwise conditional independence statement $X_u \perp X_v | X_{V \setminus \{u,v\}}$ appears among the global conditional independence statements.

($\Rightarrow\Leftarrow$): Suppose that $C$ separates nonempty subsets $A$ and $B$. First of all, we can assume that $A \cup B \cup C = V$. If not, there are supersets $A'$ and $B'$ of $A$ and $B$, respectively, such that $A' \cup B' \cup C = V$ and $C$ separates $A'$ and $B'$. Then if we have proven that $X_{A'} \perp X_{B'} | X_C$ is implied by the local Markov statements, this will also imply that $X_A \perp X_B | X_C$ follows from the local Markov statements by applying the decomposition axiom (Proposition 4.1.4 (ii)).

So suppose that $A \cup B \cup C = V$. We will proceed by induction on $\#(A \cup B)$. If $\#(A \cup B) = 2$ then both $A$ and $B$ are singletons, $A = \{a\}$ and $B = \{b\}$, there is no edge between $a$ and $b$, and so $X_A \perp X_B | X_C$ is a local
Markov statement. So suppose \( \#(A \cup B) > 2 \). Without loss of generality we may assume that \( \#B \geq 2 \). Let \( B_1 \cup B_2 = B \) be any bipartition of \( B \) into nonempty parts. Then \( C \cup B_1 \) separates \( A \) and \( B_2 \) and \( C \cup B_2 \) separates \( A \) and \( B_1 \). Also, \( \#(A \cup B_1) < \#(A \cup B) \) and \( \#(A \cup B_2) < \#(A \cup B) \) so we can apply the induction hypothesis to see that

\[
X_A \independent X_{B_1}|(X_{B_2}, X_C) \quad \text{and} \quad X_A \independent X_{B_2}|(X_{B_1}, X_C)
\]

are implied by the local Markov statements for \( G \), plus the intersection axiom. But then application of the intersection axiom to the two statements in Equation 13.1.1 shows that \( X_A \independent X_B|X_C \) is implied by the local Markov statements and the intersection axiom. The fact that the intersection axiom holds for positive densities is the content of Proposition 4.1.5.

For multivariate Gaussian random variables with nonsingular covariance matrix, the density function is strictly positive. So all the Markov properties are equivalent in this class of densities. Note that for multivariate Gaussian random variables, the interpretation of the local Markov statement is especially simple. A conditional independence statement of the form \( X_u \independent X_v|X_{V \setminus \{u, v\}} \) holds if and only if \( \det \Sigma_{V \setminus \{v\}, V \setminus \{u\}} = 0 \). By the cofactor expansion formula for the inverse matrix, this is equivalent to requiring that \( \Sigma_{uv}^{-1} = 0 \) (see Proposition 6.3.2). Hence:

**Proposition 13.1.5.** The set of covariance matrices of nonsingular multivariate Gaussian distributions compatible with the Markov properties for the graph \( G \) is precisely the set

\[
\mathcal{M}_G = \{ \Sigma \in PD_{|V|} : \Sigma_{uv}^{-1} = 0 \text{ if } u \neq v \text{ and } (u, v) \notin E \}.
\]

In particular, the Gaussian graphical model is an example of a Gaussian exponential family, where the subspace of concentration matrices that is used in the definition is a coordinate subspace.

The global Markov property for undirected graphs satisfies an important completeness property.

**Proposition 13.1.6.** Let \( G \) be an undirected graph and suppose that \( A, B, \) and \( C \) are disjoint subsets of \( V \) such that \( C \) does not separate \( A \) and \( B \) in \( G \). Then there is a probability distribution satisfying all the global Markov statements of \( G \) and not satisfying \( X_A \independent X_B|X_C \).

**Proof.** We show this with the example of a multivariate normal random vector. Consider a path \( \pi \) of vertices from \( a \in A \) to \( b \in B \) that does not pass through any vertex in \( C \), and construct a concentration matrix \( \Sigma^{-1} = K \) which has ones along the diagonal, a small nonzero \( \epsilon \) in entries corresponding to each edge in the path \( \pi \), and zeroes elsewhere. The corresponding covariance matrix belongs to the Gaussian graphical model associated to \( G \) since
13.1. Conditional Independence Description of Graphical Models

\( K \) only has nonzero entries on the diagonal and in positions corresponding to edges of \( G \), and by the fact that the pairwise property implies the global property for multivariate normal random vector (whose density is positive).

After reordering vertices, we can assume that the concentration matrix is a tridiagonal matrix of the following form

\[
K = \begin{pmatrix}
1 & \epsilon & & \\
\epsilon & 1 & \ddots & \\
& \ddots & \ddots & \epsilon \\
& & \ddots & 1 \\
& & & \epsilon & 1 \\
\end{pmatrix} I_{V \setminus \pi}
\]

where \( I_{V \setminus \pi} \) denotes an identity matrix on the vertices not involved in \( \pi \).

This implies that \( \Sigma = K^{-1} \) is a block matrix, with nonzero off-diagonal entries only appearing precisely in positions indexed by two elements of \( \pi \). Note that \( K \) is positive definite for \(|\epsilon|\) sufficiently close to zero. The conditional independence statement \( X_A \perp X_B | X_C \) holds if and only if \( \Sigma_{A,B} - \Sigma_{A,C} \Sigma_{C,B}^{-1} \Sigma_{C,B} = 0 \). However, \( \Sigma_{A,C} \) and \( \Sigma_{C,B} \) are both zero since \( C \cap \pi = \emptyset \). But \( \Sigma_{A,B} \neq 0 \) since both \( a, b \in \pi \). So the conditional independence statement \( X_A \perp X_B | X_C \) does not hold for this distribution.

A second family of graphical models in frequent use are graphical models associated to directed acyclic graphs. Let \( G = (V,E) \) be a directed graph. A directed cycle in \( G \) is a sequence of vertices \( u_0, u_1, \ldots, u_k \) with \( u_0 = u_k \) and \( u_i \to u_{i+1} \in E \) for all \( i = 0, \ldots, k - 1 \). The directed graph \( G \) is acyclic if there does not exist a directed cycle in \( G \). In this setting, the directions of the arrows are sometimes given causal interpretations (i.e. some events happen before and affect other events), though we will primarily focus on only the probabilistic interpretation of these models. See \[Pea09\] for further information on using directed acyclic graphs for causal inference. Directed acyclic graph is often abbreviated as DAG.

In a directed graph \( G \) we have directed paths, which are sequences of vertices \( u_0, \ldots, u_k \) such that for each \( i, u_i \to u_{i+1} \in E \). On the other hand, an undirected path is a sequence of vertices with \( u_0, \ldots, u_k \) such that for each \( i \) either \( u_i \to u_{i+1} \) is an edge or \( u_{i+1} \to u_i \) is an edge. In an undirected path, a collider is a vertex \( u_i \) such that \( u_{i-1} \to u_i \) and \( u_{i+1} \to u_i \) are both edges of \( G \).

Let \( v \in V \). The parents of \( v \), denoted \( \text{pa}(v) \) consists of all \( w \in V \) such that \( w \to v \) is an edge of \( G \). The descendants of \( v \), denoted \( \text{de}(v) \) consists of all \( w \in V \) such that there is a directed path from \( v \) to \( w \). The nondescendants
of \( v \) are denoted \( \text{nd}(v) = V \setminus \{v\} \cup \text{de}(v) \). The *ancestors* of \( v \), denoted \( \text{an}(v) \) consists of all \( w \in V \) such that there is a directed path from \( w \) to \( v \).

**Definition 13.1.7.** Two nodes \( v \) and \( w \) in a directed acyclic graph \( G \) are \( d \)-**connected** given a set \( C \subseteq V \setminus \{v, w\} \) if there is an undirected path \( \pi \) from \( v \) to \( w \) such that

1. all colliders on \( \pi \) are in \( C \cup \text{an}(C) \) and
2. no non-collider on \( \pi \) is in \( C \).

If \( A, B, C \subseteq V \) are pairwise disjoint with \( A \) and \( B \) nonempty, then \( C \) \( d \)-**separates** \( A \) and \( B \) if no pair of nodes \( a \in A \) and \( b \in B \) are \( d \)-connected given \( C \).

The definition of \( d \)-separation can also be reformulated in terms of the usual separation criterion in an associated undirected graph, the *moralization*. For a directed acyclic graph \( G \), let the moralization \( G^m \) denote the undirected graph which has an undirected edge \((u, v)\) for each directed edge in \( u \to v \) in \( G \), plus we add the undirected edge \((u, v)\) if \( u \to w \), \( v \to w \) are edges in \( G \). Thus the moralization “marries the parents” of each vertex. For a set of vertices \( S \subseteq V \), let \( G_S \) denote the induced subgraph on \( S \).

**Proposition 13.1.8.** Let \( G \) be a directed acyclic graph. Then \( C \) \( d \)-separates \( A \) and \( B \) in \( G \) if and only if \( C \) separates \( A \) and \( B \) in the moralization \((G_{\text{an}(A \cup B \cup C)})^m\).

See, for example, [Lau96, Prop. 3.25] for a proof.

As for undirected graphs, we can define pairwise, local, and global Markov properties associated to any directed acyclic graph \( G \).

**Definition 13.1.9.** Let \( G = (V, E) \) be a directed acyclic graph.

(i) The *directed pairwise Markov property* associated to \( G \) consists of all conditional independence statements \( X_u \perp \!\!\!\!\perp X_v | \text{nd}(u) \setminus \{v\} \) where \((u, v)\) is not an edge of \( G \).

(ii) The *directed local Markov property* associated to \( G \) consists of all conditional independence statements \( X_v \perp \!\!\!\!\perp X_{\text{nd}(v) \setminus \text{pa}(v)} | X_{\text{pa}(v)} \) for all \( v \in V \).

(iii) The *directed global Markov property* associated to \( G \) consists of all conditional independence statements \( X_A \perp \!\!\!\!\perp X_B | X_C \) for all disjoint sets \( A, B, \) and \( C \) such that \( C \) \( d \)-separates \( A \) and \( B \) in \( G \).

Let \( \text{pairwise}(G), \text{local}(G), \) and \( \text{global}(G) \) denote the set of directed pairwise, local, and global Markov statements associated to the DAG \( G \).
Example 13.1.10 (Comparison of Markov Properties in DAGs). Consider the DAG with 6 vertices illustrated in Figure 13.1.1. In this DAG, we can see that the conditional independence statement $X_1 \perp X_2$ holds, since $\emptyset$ d-separates 1 and 2 in the graph. This is an example of a pairwise Markov statement. On the other hand, the conditional independence statement $(X_1, X_2) \perp (X_5, X_6) | (X_3, X_4)$ holds in this graph, and is an example of a directed global Markov statement that is not a directed local Markov statement.

Although the list of conditional independence statements can be strictly larger as we go from directed pairwise to local to global Markov statements, the direct global and local Markov properties are, in fact, equivalent on the level of probability distributions.

Theorem 13.1.11. If the random vector $X$ has a joint distribution $P$ that obeys the directed local Markov property for the directed acyclic graph $G$ then $P$ obeys the directed global Markov property for the graph $G$.

Theorem 13.1.11 is intimately connected with the parametrizations of graphical models which we will explore in the next section, where we will also give its proof. The directed global Markov property is also complete with respect to a given directed acyclic graph $G$.

Proposition 13.1.12. Let $G$ be a directed acyclic graph and suppose that $A, B, C \subseteq V$ are disjoint subsets such that $C$ does not d-separate $A$ and $B$ in $G$. Then there is a probability distribution satisfying all the directed global Markov statements of $G$ and not satisfying $X_A \perp X_B | X_C$.

Proof Sketch. As in the proof of Proposition 13.1.6, this can be shown by constructing an explicit example of a distribution that satisfies the direct global Markov statements but not satisfying the statement $X_A \perp X_B | X_C$. By the moralization argument, this reduces to the case of undirected graphs, assuming that we have a reasonable method to parametrize distributions.
satisfying the directed global Markov property. This will be dealt with in the next section.

For undirected graphical models, each graph gives a unique set of global Markov statements, and hence each graph yields a different family of probability distributions. The same is not true for directed graphical models, since different graphs can yield precisely the same conditional independence statements, because two different directed graphs can have the same moralization on all induced subgraphs. For example, the graphs $G_1 = ([3], \{1 \to 2, 2 \to 3\})$ and $G_2 = ([3], \{3 \to 2, 2 \to 1\})$ both have the directed global Markov property consisting of the single independence statement $X_1 \perp X_3 | X_2$.

Two directed acyclic graphs are called Markov equivalent if they yield the same set of directed global Markov statements. Markov equivalence is completely characterized by the following result.

**Theorem 13.1.13.** Two directed acyclic graphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ are Markov equivalent if and only if the following two conditions are satisfied:

1. $G_1$ and $G_2$ have the same underlying undirected graph,
2. $G_1$ and $G_2$ have the same unshielded colliders, which are triples of vertices $u, v, w$ which induce a subgraph $u \to v \leftarrow w$.

A proof of this is found in [AMP97, Thm 2.1].

**Example 13.1.14** (Comparison of Undirected and Directed Graphical Models). Note that DAG models and undirected graphical models give fundamentally different families of probability distributions. For a given DAG, $G$ there is usually not an undirected graph $H$ whose undirected global Markov property is the same the directed global Markov property of $G$. For instance, consider the directed acyclic graph $G = ([3], \{1 \to 3, 2 \to 3\})$. Its only directed global Markov statement is the marginal independence statement $X_1 \perp X_2$. This could not be the only global Markov statement in an undirected graph on 3 vertices.

### 13.2. Parametrizations of Graphical Models

A useful counterpoint to the definition of graphical model via conditional independence statements from Section 13.1 is the parametric definition of graphical models which we describe here. In many applied contexts, the parametric description is the natural description for considering an underlying process that generates the data. As usual with algebraic statistical models, we have a balance between the model’s implicit description (via conditional independence constraints) and its parametric description. For both
undirected and directed acyclic graphical models there are theorems which relate the description via conditional independence statements to parametric description of the model (the Hammersley-Clifford Theorem (Theorem 13.2.3) and the recursive factorization theorem (Theorem 13.2.10), respectively). We will see two important points emerge from the comparison of the parametric and implicit model descriptions: 1) for discrete undirected graphical models, there are distributions which satisfy all the global Markov statements but are not in the closure of the parametrization, 2) even when the parametric description of the model gives exactly the distributions that satisfy the global Markov statements of the graph, the prime ideal of all functions vanishing on the model might not equal the global conditional independence ideal of the graph.

First we consider parametrizations of graphical models for undirected graphs. Let $G$ be an undirected graph with vertex set $V$ and edge set $E$. A clique $C \subseteq V$ in $G$ is a set of vertices such that $(i, j) \in E$ for all $i, j \in C$. The set of all maximal cliques in $G$ is denoted $\mathcal{C}(G)$. For each $C \in \mathcal{C}(G)$, we introduce a continuous potential function $\phi_C(x_C) \geq 0$ which is a function on $\mathcal{X}_C$, the state space of the random vector $X_C$.

**Definition 13.2.1.** The *parametrized undirected graphical model* consists of all probability density functions on $\mathcal{X}$ of the form

$$f(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}(G)} \phi_C(x_C)$$

for some potential functions $\phi_C(x_C)$, where

$$Z = \int_{\mathcal{X}} \prod_{C \in \mathcal{C}(G)} \phi_C(x_C) d\mu(x)$$

is the normalizing constant (also called the partition function). The parameter space for this model consists of all tuples of potential functions such that the normalizing constant is finite and nonzero. A probability density is said to factorize according to the graph $G$ if it can be written in the product form (13.2.1) for some choice of potential functions.

**Example 13.2.2 (Triangle-Square).** Let $G$ be the graph with vertex set $\{5\}$ and edge set $\{12, 13, 23, 25, 34, 45\}$, pictured in Figure 13.2.1. The set of maximal cliques of $G$ is $\{123, 25, 34, 45\}$. The parametric description from (13.2.1) asks for probability densities with the following factorization:

$$f(x_1, x_2, x_3, x_4, x_5) = \frac{1}{Z} \phi_{123}(x_1, x_2, x_3) \phi_{25}(x_2, x_5) \phi_{34}(x_3, x_4) \phi_{45}(x_4, x_5).$$

The Hammersley-Clifford theorem gives the important result that the parametric graphical model described in Definition 13.2.1 yields the same
family of distributions as the pairwise Markov property when restricted to strictly positive distributions. The original proof of this theorem has an interesting story, see [Cli90].

**Theorem 13.2.3** (Hammersley–Clifford). A continuous positive probability density $f$ on $X$ satisfies the pairwise Markov property on the graph $G$ if and only if it factorizes according to $G$.

**Proof.** ($\Leftarrow$) Suppose that $f$ factorizes according to $G$ and that $i$ and $j$ are not connected by an edge in $G$. Then no potential function in the factorization of $f$ depends on both $x_i$ and $x_j$. Letting $C = V \setminus \{i,j\}$, we can multiply potential functions together that depend on $x_i$ as $\psi_i(x_i, x_C)$ and those together that depend on $x_j$ as $\psi_j(x_i, x_C)$, and the remaining functions that depend on neither has $\psi_C(x_C)$, to arrive at the factorization:

$$f(x_i, x_j, x_C) = \frac{1}{Z} \psi_i(x_i, x_C)\psi_i(x_i, x_C)\psi_C(x_C).$$

This function plainly satisfies the relation

$$f(x_i, x_j, x_C)f(x'_i, x'_j, x_C) = f(x_i, x'_j, x_C)f(x'_i, x_j, x_C)$$

for all $x_i, x'_i, x_j, x'_j$. By integrating out the $x'_i$ and $x'_j$ variables on both sides of this equation, we deduce that $X_i \perp X_j | X_C$ holds. Hence, $f$ satisfies the local Markov property.

($\Rightarrow$) Suppose that $f$ satisfies the local Markov property on $G$. Let $y \in X$ be any point. For any subset $C \subseteq V$ define a potential function

$$\phi_C(x_C) = \prod_{S \subseteq C} f(x_S, y_{V \setminus S})^{(-1)^{|S-C|-|S|}}.$$

Note that by the principle of Möbius inversion,

$$f(x) = \prod_{C \subseteq V} \phi_C(x_C).$$
We claim that if \( C \) is not a clique of \( G \), then \( f(x) = \prod_{C \subseteq C^*(G)} \phi_C(x_C) \).

where \( C^*(G) \) is the set of all cliques in \( G \). This proves that \( f \) factorizes according to the model since we can merge the potential functions associated to nonmaximal cliques into those associated to one of the maximal cliques containing it.

To prove that claim, suppose that \( C \) is not a clique of \( G \). Then there is a pair \( i,j \subseteq C \) such that \( (i,j) \notin E \). Let \( D = V \setminus \{i,j\} \). We can write \( \phi_C(x_C) \) as

\[
\phi_C(x_C) = \prod_{S \subseteq C \cap D} \left( \frac{f(x_i, x_j, x_S, y_D, S)}{f(x_i, y_j, x_S, y_D, S)} \right)^{(1 - \#C_\cap S)}.
\]

Since the conditional independence statement \( X_i \perp \!\!\!\!\perp X_j | X_D \) holds, each of the terms inside the product is equal to one, hence the product equals one. Note that there can be probability distributions that satisfy the global Markov property of an undirected graph \( G \), but do not factorize according to that graph. By the Hammersley-Clifford theorem, such distributions must have some nonempty events with zero probability. Examples and the connection to primary decomposition of conditional independence ideals will be discussed in Section 13.3.

While the Hammersley–Clifford theorem only compares the local Markov statements to the parametrization, it also implies that any distribution in the image of the parametrization associated to the graph \( G \) satisfies the global Markov statements.

**Corollary 13.2.4.** Let \( P \) be a distribution that factors according to the graph \( G \). Then \( P \) satisfies the global Markov property on the graph \( G \).

**Proof.** The Hammersley–Clifford theorem shows that if \( P \) is positive, then \( P \) satisfies the local Markov property. All positive \( P \) satisfy the intersection axiom, and hence by Theorem 13.1.4, such positive \( P \) must satisfy the global Markov property. However, the global Markov property is a closed condition, so all the distributions that factorize and have some zero probabilities, which must be limits of positive distributions that factorize, also satisfy the global Markov property.

Focusing on discrete random variables \( X_1, \ldots, X_m \), let \( X_j \) have state space \([r_j]\). The joint state space of the random vector \( X = (X_1, \ldots, X_m) \) is \( \mathcal{R} = \prod_{j=1}^m [r_j] \). In the discrete case, the condition that the potential functions \( \psi_C(x_C) \) are continuous on \( \mathcal{R}_C \) is vacuous, and hence we are allowed any
nonnegative function as a potential function. Using an index vector \(i_C = (i_j)_{j \in C} \in \mathcal{R}_C\), to represent the states we can write the potential function \(\psi_C(x_C)\) as \(\theta_{i_C}^{(C)}\), so that the parametrized expression for the factorization in the Equation 13.2.1 becomes a monomial parametrization. Hence, the discrete undirected graphical model is a hierarchical model. For discrete graphical models and a graph (either directed, undirected, or otherwise) let \(I_G\) denote the homogeneous vanishing ideal of the graphical model associated to \(G\). In summary we have the following proposition whose proof is left to the reader.

**Proposition 13.2.5.** The parametrized discrete undirected graphical model associated to \(G\) consists of all probability distributions in \(\Delta_R\) given by the probability expression

\[
p_{i_1 i_2 \cdots i_m} = \frac{1}{Z(\theta)} \prod_{C \in \mathcal{C}(G)} \theta_{i_C}^{(C)}
\]

where \(\theta = (\theta^{(C)})_{C \in \mathcal{C}(G)}\) is the vector of parameters. In particular, the positive part of the parametrized graphical model is the hierarchical log-linear model associated to the simplicial complex of cliques in the graph \(G\).

Proposition 13.2.5 implies that the homogeneous vanishing ideal of a discrete undirected graphical model \(I_G\) is a toric ideal.

**Example 13.2.6** (Triangle-Square). Consider the graph from Example 13.2.2, and let \(r_1 = \cdots = r_5 = 2\), so that all \(X_i\) are binary random variables. The homogenized version of the parametrization of the undirected graphical model has the form

\[
p_{i_1 i_2 i_3 i_4 i_5} = \theta_{i_1 i_2 i_3}^{(123)} \theta_{i_2 i_5}^{(25)} \theta_{i_3 i_4}^{(34)} \theta_{i_4 i_5}^{(45)}.
\]

The toric vanishing ideal of the graphical model can be computed using techniques described in Chapter 6. In this case the vanishing ideal \(I_G\) is generated by homogeneous binomials of degrees 2 and 4. This can be computed directed in Macaulay 2, or using results from Section 9.3 on the toric ideal of hierarchical models for reducible complexes.

Turning now to Gaussian undirected graphical models, the structure is determined in a very simple way from the graph. The density of the multivariate normal distribution \(\mathcal{N}(\mu, \Sigma)\) can be written as

\[
f(x) = \frac{1}{Z} \prod_{i=1}^{m} \exp \left( -\frac{1}{2} (x_i - \mu_i)^2 k_{ii} \right) \prod_{1 \leq i < j \leq m} \exp (- (x_i - \mu_i)(x_j - \mu_j) k_{ij})
\]

where \(K = (k_{ij}) = \Sigma^{-1}\) is the concentration matrix, and \(Z\) is the normalizing constant. In particular, we see that the density always factorizes into pairwise potential according to the graph \(G\) if and only if \(k_{ij} = 0\) for all
13.2. Parametrizations of Graphical Models

(i, j) ∉ E. In summary, we have the following proposition whose proof is left to the reader.

**Proposition 13.2.7.** The parametrized Gaussian undirected graphical model corresponds to the set of pairs \((\mu, \Sigma) \in \mathbb{R}^m \times PD_m\) with \((\Sigma^{-1})_{ij} = 0\) for all \((i, j) \notin E\).

There are a number of approaches that can be used to compute the vanishing ideal of the Gaussian graphical model associated to \(G\). One strategy is to use the classical adjoint formula for the matrix inverse to get a parametrization for the ideal system. Alternately, we can perform an elimination on the system \(Id - \Sigma K = 0\), where \(Id\) denotes an \(m \times m\) identity matrix. It is not clear which of these is best in general, but we illustrate the second method with an example. For a given graph \(G\), whether undirected or directed, we use \(J_G\) to denote the vanishing ideal of the associated Gaussian graphical model.

**Example 13.2.8 (Four Cycle).** Let \(G\) be the four cycle graph with edges set \(E = \{12, 23, 34, 14\}\). Since the Gaussian graphical model makes no restrictions on the mean vector of our Gaussian random variables, we can focus on the covariance matrix. We want to find the vanishing ideal of the set of covariance matrices \(\Sigma\) such that \((\Sigma^{-1})_{13} = (\Sigma^{-1})_{24} = 0\). The following Singular code computes the vanishing ideal in this case.

```plaintext
ring R = 0, (s11,s12,s13,s14,s22,s23,s24,s33,s34,s44, k11,k12,k14,k22,k23,k33,k34,k44),dp;
matrix K[4][4] = k11,k12,0,k14,k12,k22,k23,0, 0,k23,k33,k44,k14,0,k34,k44;
matrix S[4][4] = s11,s12,s13,s14,s12,s22,s23,s24, s13,s23,s33,s34,s14,s24,s34,s44;
matrix eye[4][4] = 1,0,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,1,0,0,0,0,0,1,0,0,0,1;ideal JG = K*S-eye;
LIB "elim.lib";
elim(I,11..18);
```

In this case, we see that the vanishing ideal of the parametrization is equal to the conditional independence ideal associated to the global Markov statements of this model. It is the ideal

\[ J_G = J_{\text{global}}(G) = \langle |\Sigma_{124,234}|, |\Sigma_{123,134}| \rangle. \]

Now we consider the parametric description of graphical models for directed acyclic graphs. Let \(G = (V, E)\) be such a DAG. For each node \(j \in V\) we have a conditional distribution of variable \(X_j\) conditional on the parents of node \(j\) in the graph \(G\), \(f_j(x_j | x_{pa(j)})\) and consider probability densities
of the form

\[ f(x) = \prod_{j \in V} f_j (x_j \mid x_{\text{pa}(j)}). \]

**Definition 13.2.9.** The parametric directed graphical model associated to the directed acyclic graph \( G \) consists of all probability densities that factorize as the product of conditionals (13.2.2).

The recursive factorization structure that arises in graphical models associated to directed acyclic graphs can be quite useful for Bayesian data analysis. Indeed, some more ancestral vertices in the graph can represent prior distributions, which hierarchically determine the distributions of other variables given the values of those priors. For this reason, DAG models are often called Bayesian networks.

The situation in the directed case is much better than in the undirected case: factorizations are equivalent to satisfying the directed local Markov property.

**Theorem 13.2.10** (Recursive Factorization). A probability density satisfies the recursive factorization property (13.2.2) associated to the directed acyclic graph \( G \) if and only if it satisfies the directed local Markov property associated to \( G \).

**Proof.** Since \( G \) is a directed acyclic graph, we can assume the vertex set is \( V = [m] \) and that after renaming vertices, we have that \( i \to j \) implies that \( i < j \). This is called a topological ordering of the vertices in the graphical models literature.

\( (\Rightarrow) \) We will show the stronger statement that a distribution that factorizes satisfies the global Markov property. Let \( A, B, C \) disjoint subsets of \( V \) such that \( C \) d-separates \( A \) and \( B \), and let \( f \) be a distribution with factorization (13.2.2). We can assume that \( V = \text{an}(A \cup B \cup C) \), otherwise we marginalize over \( V \setminus \text{an}(A \cup B \cup C) \) which does not change the factorization structure. To say that \( C \) d-separates \( A \) and \( B \) we have that \( C \) separates \( A \) and \( B \) in the moral graph of \( G \). Moralization makes each set \( \{j\} \cup \text{pa}(j) \) a clique in the graph \( G^m \). Hence, the factorization (13.2.2) gives a parametrization by potential functions in the graph \( G^m \). Hence, by Corollary 13.2.4 the conditional independence statement \( X_A \perp X_B \mid X_C \) is satisfied.

\( (\Leftarrow) \) Any density \( f \) whatsoever can be factorized as

\[ f(x) = \prod_{j=1}^{m} f_j (x_j \mid x_{[j-1]}), \]
13.2. Parametrizations of Graphical Models

where \( f_j(x_j \mid x_{[j-1]}) \) is the conditional density of \( X_j \) given the variables \( X_1, \ldots, X_{j-1} \). Since the vertices of \( G \) are topologically ordered, we always have \( \text{pa}(j) \subseteq [j-1] \), and similarly, \( [j-1] \subseteq \text{nd}(j) \). Let \( C = [j-1] \setminus \text{pa}(j) \subseteq \text{nd}(j) \setminus \text{pa}(j) \). Thus, we can split this representation as \( f_j(x_j \mid x_{\text{pa}(j)}, x_C) \).

But by the directed local Markov property \( X_j \perp \!\!\!\perp \text{nd}(j) \cap \text{pa}(j) \mid \text{pa}(j) \), and the decomposition axiom implies that \( X_j \perp \!\!\!\perp X_C \mid \text{pa}(j) \). Thus,

\[
    f_j(x_j \mid x_{[j-1]}) = f_j(x_j \mid x_{\text{pa}(j)}, x_C) = f_j(x_j \mid x_{\text{pa}(j)}),
\]

so the density \( f \) factorizes according to \( G \).

Theorem 13.2.10 allows us to explicitly parametrize both discrete or Gaussian directed graphical models and use those parametrizations to compute the vanishing ideal of the model. As we have mentioned previously, knowing the vanishing ideal can be useful for other statistical applications (i.e. it can be easier to compute the ML-degree), and we can try to understand the discrepancy between the parametrized model and the implicit model.

The recursive factorization theorem also allows us to prove that the directed local Markov property implies the directed global Markov property for DAGs.

**Proof of Theorem 13.1.11.** This is a consequence of the proof of Theorem 13.2.10. If \( f \) satisfies the directed local Markov property with respect to \( G \) then \( f \) has a factorization of the form

\[
    f(x) = \prod_{j \in V} f_j(x_j \mid x_{\text{pa}(j)}).
\]

But if \( f \) has such a recursive factorization, then it automatically satisfies the global Markov property associated to the DAG \( G \). This can be seen by applying the moralization version of the d-separation criterion and using the recursive factorization formula as an expression with potential functions to relate it to the undirected separation criterion.

In the discrete case, the parametric representation of the directed graphical model \( G \) takes the form

\[
    P_{i_1 i_2 \ldots i_m} = \prod_{j=1}^{m} \theta^{(j)}(i_j \mid i_{\text{pa}(j)})
\]

where the parameters \( \theta^{(j)}(i_j \mid i_{\text{pa}(j)}) \) represent the conditional probabilities in the recursive factorization. Although this looks like a monomial parametrization, and hence the model appears to be a log-linear model, the fact that
these are conditional probabilities means that the parameters satisfy linear constraints:

\[ \sum_{k=1}^{r_j} \theta^{(j)}(k|i_{\text{pa}(j)}) = 1 \]

for all \( i_{\text{pa}(j)} \in \mathcal{R}_{\text{pa}(j)}. \)

**Example 13.2.11 (A Nontoric DAG).** Let \( G \) be the graph with vertex set \([3]\) and edges \( 1 \rightarrow 3 \) and \( 2 \rightarrow 3 \). The parametric description of this model will be:

\[ p_{i_1i_2i_3} = \theta^{(1)}(i_1)\theta^{(2)}(i_2)\theta^{(3)}(i_3|i_1i_2) \]

where

\[ \sum_{i_1 \in [r_1]} \theta^{(1)}(i_1) = 1, \quad \sum_{i_2 \in [r_2]} \theta^{(2)}(i_2) = 1, \quad \text{and} \quad \sum_{i_3 \in [r_3]} \theta^{(3)}(i_3|i_1i_2) = 1. \]

Letting \( r_1 = r_2 = r_3 = 2 \), we can compute the vanishing ideal of the model in Macaulay 2 using the following code:

```plaintext
S = QQ[t,a,b,c11,c12,c21,c22];
R = QQ[p111,p112,p121,p122,p211,p212,p221,p222];
f = map(S,R, { t*a*b*c11, t*a*b*(1-c11), t*a*(1-b)*c12,
            t*a*(1-b)*(1-c12), t*(1-a)*b*c21, t*(1-a)*b*(1-c21),
            t*(1-a)*(1-b)*c22, t*(1-a)*(1-b)*(1-c22)});
I = kernel f
```

Note that we use the following substitutions \( a = \theta^{(1)}(1) \) (so \( 1 - a = \theta^{(1)}(2) \)), \( b = \theta^{(2)}(1) \), \( c_{i_2i_3} = \theta^{(3)}(1|i_1i_2) \). The homogenization parameter \( t \) is introduced so that we eliminate the trivial equation \( p_{111} + \cdots + p_{222} = 1 \). The vanishing ideal in this case is equal to the conditional independence ideal of all directed global Markov statements of the graph, namely it is equal to \( I_{1\rightarrow 2} \). In particular, the vanishing ideal is not toric because it is not generated by binomials.

We now turn attention to the case of Gaussian Bayesian networks. The recursive factorization of the joint density translates into a sequence of recursive regressions of random variables lower in the graph in terms of random variables further up the graph. Indeed, in the jointly normal case, the conditional density of \( X_j|X_{\text{pa}(j)} \) is also normal and is specified parametrically by a formula like

\[ (13.2.3) \quad X_j = \sum_{k \in \text{pa}(j)} \lambda_{jk}X_k + \epsilon_j, \]

where \( \epsilon_j \sim \mathcal{N}(\nu_j, \omega_j) \) is a univariate normal distribution, and \( \lambda_{jk} \) is a real parameter, sometimes called the regression coefficient.
Let $\Lambda$ be the $m \times m$ upper triangular matrix such that
\[
\Lambda_{kj} = \begin{cases} 
\lambda_{kj} & \text{if } k \rightarrow j \in E \\
0 & \text{otherwise.}
\end{cases}
\]
We use the notations $\mathbb{R}^E$ to denote the set of all such matrices $\Lambda$. We can write (13.2.3) in matrix format as
\[
X = \Lambda^TX + \epsilon
\]
where $\epsilon = (\epsilon_1, \ldots, \epsilon_m)$ is the jointly normal random vector with mean $\nu = (\nu_1, \ldots, \nu_m)$ and diagonal covariance matrix $\Omega = \text{diag}(\omega_1, \ldots, \omega_m)$. Solving for $X$ yields
\[
X = (\text{Id} - \Lambda)^{-1}\epsilon
\]
where $\text{Id}$ is an $m \times m$ identity matrix and $(\text{Id} - \Lambda)^{-1}$ denotes the transpose of the inverse of the matrix $\text{Id} - \Lambda$. By Lemma 2.4.3, $X$ is a multivariate normal with covariance matrix $\Sigma = (\text{Id} - \Lambda)^{-T}\Omega(\text{Id} - \Lambda)^{-1}$. In summary, the Gaussian DAG model is parameterized as follows. The proof is left to the reader.

**Proposition 13.2.12.** The parametrized Gaussian graphical model associated to the directed acyclic graph $G$ consists of all pairs $(\mu, \Sigma) \in \mathbb{R}^m \times \text{PD}_m$ such that $\Sigma = (\text{Id} - \Lambda)^{-T}\Omega(\text{Id} - \Lambda)^{-1}$ for some $\Omega$ diagonal with positive entries and upper triangular $\Lambda \in \mathbb{R}^E$.

**Example 13.2.13 (Directed Four Cycle).** Let $G = ([4], \{1 \rightarrow 2, 1 \rightarrow 3, 2 \rightarrow 4, 3 \rightarrow 4\})$. The matrices $\Lambda$ and $(\text{Id} - \Lambda)^{-1}$ are
\[
\begin{pmatrix}
0 & \lambda_{12} & \lambda_{13} & 0 \\
0 & 0 & \lambda_{24} & 0 \\
0 & 0 & 0 & \lambda_{34} \\
0 & 0 & 0 & 0
\end{pmatrix},
\begin{pmatrix}
1 & \lambda_{12} & \lambda_{13} \lambda_{24} + \lambda_{13}\lambda_{34} \\
0 & 1 & 0 & \lambda_{24} \\
0 & 0 & 1 & \lambda_{34} \\
0 & 0 & 0 & 1
\end{pmatrix}.
\]
From the formula $\Sigma = (\text{Id} - \Lambda)^{-T}\Omega(\text{Id} - \Lambda)^{-1}$ we see, for example, that $\sigma_{24} = \omega_1\lambda_{12}\lambda_{13}\lambda_{34} + \omega_2\lambda_{24}$.

We can compute the vanishing ideal of the parametrization in the ring $\mathbb{R}[\Sigma]$ of polynomial functions of the covariance matrix. This vanishing ideal is
\[
\langle [\Sigma_{12,13}], [\Sigma_{123,234}] \rangle,
\]
which equals the conditional independence ideal of the directed global Markov statements, $J_{\text{global}(G)} = J_{2 \perp 3|1} + J_{1 \perp 4|\{2,3\}}$.

For a DAG $G$, let $J_G$ denote the vanishing ideal of the Gaussian graphical model associated to $G$. At the time of this writing there is no known general characterization of which graphs the vanishing ideal of the model $J_G$ equals the global Markov ideal for a Gaussian Bayesian network $J_{\text{global}(G)}$. It is known to hold for trees, and all graphs with $\leq 4$ vertices [Sul08]. Already
300 13. Graphical Models

on 5 vertices there are graphs for which the global Markov ideal does not equal the vanishing ideal of the model. Somewhat surprisingly, there are graphs where the vanishing ideal contains vanishing subdeterminants of $\Sigma$ that do not belong to the global Markov ideal.

Example 13.2.14 (An Extra Generator). Let $G$ be the directed acyclic graph illustrated in Figure 13.2.2. The vanishing ideal $J_G$ contains the minor $|\Sigma_{12,45}|$ which is not contained in the global Markov ideal $J_{\text{global}}(G)$. On the other hand, $J_G = J_{\text{global}}(G) + \langle |\Sigma_{12,45}| \rangle$, in particular, $J_G$ is generated by all the determinants of $\Sigma$ that vanish on the model.

The characterization of which subdeterminants of $\Sigma$ belong to the ideal $J_G$ is through a combinatorial conditional called t-separation.

Definition 13.2.15. Let $G = (V, D)$ be a directed acyclic graph. A trek between two vertices $a$ and $b$ is a pair $(P_a, P_b)$ of directed paths $P_a$ and $P_b$ that have the same starting vertex and such that $P_a$ ends at $a$ and $P_b$ ends at $b$.

Definition 13.2.16. Let $A$, $B$, $C_A$, and $C_B$ be four subsets of $V$, not necessarily disjoint. The pair of sets $(C_A, C_B)$ t-separates $A$ and $B$ if for all $a \in A$ and $b \in B$ and any trek $(P_a, P_b)$ from $a$ to $b$, either $P_a$ has a vertex in $C_A$, or $P_b$ has a vertex in $C_B$, or both.

Theorem 13.2.17. [STD10] Let $G = (V, D)$ be a directed acyclic graph, and $A$ and $B$ two subsets of $V$ with $\#A = \#B = k$. Then $|\Sigma_{A,B}| \in J_G$ if and only if there exist a pair of sets $(C_A, C_B)$ that t-separate $A$ and $B$ such that $\#C_A + \#C_B < k$.

The proof of Theorem 13.2.17 is too complicated for this text. It depends on combinatorial tools related to determinants of matrices associated to graphs, in particular the Gessel-Viennot-Lindström Lemma.

Example 13.2.18 (t-separation Determinant). Considering the graph $G$ from Example 13.2.14, the pair of sets $(\emptyset, \{4\})$ t-separates $\{1, 2\}$ and $\{4, 5\}$. This is because any trek $(P_L, P_R)$ from $\{1, 2\}$ to $\{4, 5\}$ has $P_L = \emptyset$, and must

![Figure 13.2.2. The directed acyclic graph from Example 13.2.14](image-url)
13.3. Failure of the Hammersley-Clifford Theorem

contain 4 as a vertex in $P_R$. Thus, $|\Sigma_{12,45}|$ belongs to the vanishing ideal of the Gaussian graphical model $J_G$.

Although we have tended to build our conditional independence and graphical model examples from scratch in this book, there is a package in Macaulay 2 specifically for working with the ideals and parametrizations of graphical models and conditional independence ideals (see \[GPPS13\] for more details). For example, the following code verifies that the vanishing ideal of the graph in Example 13.2.14 does not equal the global Markov ideal.

\begin{verbatim}
loadPackage "GraphicalModels"
G = digraph{{1,{4}},{2,{4}},{3,{4,5}},{4,{5}}}
R = GaussianRing G
I = conditionalIndependenceIdeal(R,globalMarkov(G))
J = GaussianVanishingIdeal(R)
I == J
\end{verbatim}

13.3. Failure of the Hammersley-Clifford Theorem

The comparison of the implicit description of a graphical model by conditional independence statements to the parametric description of the model by factorization is especially interesting in the case of undirected graphical models on discrete variables. Of the four primary cases we consider in this chapter (undirected discrete, directed discrete, undirected Gaussian, directed Gaussian) this is the only case where it can happen that not every distribution that satisfies the conditional independence constraints need factor according to the model. In this case, primary decomposition can be helpful to determine which extra probability distributions there are that satisfy the conditional independence constraints but do not factor.

The first main theorem in this area classifies which graphs give equality of the ideals $I_{\text{global}}(G)$ and $I_G$.

**Theorem 13.3.1** (\[GMS06\]). Let $G$ be an undirected graph. Then the vanishing ideal of the discrete graphical model, $I_G$, is equal to the ideal of global Markov statements, $I_{\text{global}}(G)$, if and only if $G$ is a chordal graph.

Recall that a graph is chordal if and only if there are no induced cycles of length $\geq 4$, or equivalently a graph is chordal if and only if its complex of cliques is a decomposable simplicial complex (Theorem 8.3.5). We provide a sketch of the proof, see \[GMS06\] for complete details.

**Proof Sketch.** If $G$ is chordal, $I_G$ is the toric ideal of a decomposable hierarchical model, hence has a Markov basis/ generating set consisting of
quadratic binomials by Corollary 9.3.18. It can be shown that those quadrics can be chosen to come from conditional independence statements.

Conversely, if $G$ is not chordal, then it has an induced subgraph that is a cycle of length $\geq 4$. The toric ideal $I_H$ associated to an induced subgraph $H$ naturally corresponds to a facial subset of the vector configuration associated to $G$. Hence, the generators of $I_H$ appear as generators of $I_G$. But the toric ideal of a cycle contains generators of degree $\geq 4$, so it could not equal the conditional independence ideal, which is necessarily generated by quadrics.

Beyond chordal graphs, the set of distributions which satisfy the global Markov statements associated to the graph $G$ will decompose into a number of irreducible components. Only in a small number of examples is the explicit decomposition of the binomial ideal $I_{\text{global}}(G)$ known. Here is one example explicitly computed in [KRS14].

**Theorem 13.3.2.** Let $C_m$ be the $m$-cycle graph with edges $(i, i+1)$ for $i = 1, \ldots, m-1$ and edge $(1, m)$. Suppose that all random variables $X_i$ are binary, i.e. $r_i = 2$ for all $i \in [m]$. The global independence ideal $I_{\text{global}}(C_m)$ is radical and has $1 + \binom{m}{3} \times 2^{m-3}$ minimal primes. The minimal primes of $I_{\text{global}}(C_m)$ are of two forms:

- The toric vanishing ideal $I_{C_m}$ of the graphical model.
- $\binom{m}{3} \times 2^{m-3}$ monomial prime ideals, each of dimension 8.

To explicitly describe these minimal primes, we need some notation. If $i = (i_1, \ldots, i_m) \in [2]^m$ is an index string, let $\overline{i}$ denote the binary complement $\overline{i} = (3-i_1, \ldots, 3-i_m)$. Now, let $A|B|C$ be any cyclic partition of the numbers $[m]$, that is, each of the sets $A, B, C$ is an interval of integers, considered modulo $m$. Let $(i_A, i_B, i_C)$ be any binary index string, with $i_A \in \mathcal{R}_A$, etc. Associate to this pair of information the monomial ideal

$$P = \langle p_{j_A,j_B,j_C} | (j_A, j_B, j_C) \notin \{ (i_A, i_B, i_C), (i_A, i_B, \overline{i_C}), \ldots, (\overline{i_A}, \overline{i_B}, \overline{i_C}) \} \rangle.$$  

There are $\binom{m}{3} \times 2^{m-3}$ such monomial ideals, and these are precisely the non-toric minimal prime ideals of the ideal $I_{\text{global}}(C_m)$.

**Example 13.3.3 (Four Cycle).** Letting $m = 4$, the binomial ideal $I_{\text{global}}(C_4)$ has nine minimal prime ideals. There is the toric vanishing ideal $I_{C_4}$ and eight other monomial ideals. Associated to the cyclic tripartition $1|23|4$ we have two monomial ideals:

$$\langle p_{1111}, p_{1112}, p_{1121}, p_{1122}, p_{1211}, p_{2111}, p_{2211}, p_{2221}, p_{2222} \rangle$$

$$\langle p_{1121}, p_{1122}, p_{1211}, p_{1212}, p_{2121}, p_{2122}, p_{2211}, p_{2212}, p_{2221}, p_{2222} \rangle$$
For example, from the first of these two components, we see that any probability distribution with
\[ p_{1111} = p_{1112} = p_{1221} = p_{2111} = p_{2112} = p_{2221} = p_{2222} = 0 \]
satisfies the conditional independence statements \( 1 \perp 3 \mid \{2, 4\} \) and \( 2 \perp 4 \mid \{1, 3\} \)
which are the global Markov statements of the four-cycle graph. However, most of these distributions will not factor. Indeed, the vanishing ideal of the distributions that factor contains the binomial
\[ p_{1121}p_{2122}p_{1212}p_{2211} - p_{1122}p_{2121}p_{1211}p_{2212} \]
which a generic distribution in the first component will not satisfy.

The structure of this primary decomposition can also be used to determine the connectivity of random walks associated to the quadratic binomials in the ideal. See Section 9.5 for the background and [KRS14] for more details.

Note that in Theorem 13.3.2 the conditional independence ideal \( I_{\text{global}}(G) \) was a radical ideal. The global Markov ideal for the complete bipartite graph \( K_{2,m} \) with all binary random variables was also shown to be radical in [KRS14]. Global Markov ideals are not, however, radical in general: for the complete bipartite graph \( K_{3,3} \), \( I_{\text{global}}(K_{3,3}) \) is not radical already for binary random variables [KRS14]. In general, it is an open problem to give a characterization of those graphs and dimension vectors \( r \) for which \( I_{\text{global}}(G) \) is a radical ideal. Here is a specific conjecture:

**Conjecture 13.3.4.** Let \( G \) be a graph with no \( K_4 \) minors, and suppose that all random variables are binary. Then \( I_{\text{global}}(G) \) is radical.

For each of the other combination of type of graph and discrete or Gaussian random variables, it is an open problem to determine for which family of graphs \( I_G = I_{\text{global}}(G) \) or \( J_G = J_{\text{global}}(G) \). Further conjectures of characterizations for when the global Markov ideal are radical are also open at this point, and an interesting area for further study.

### 13.4. Examples of Graphical Models from Applications

Graphical models are widely used in the applied sciences. We will give some examples of particular situations where they are used. More examples will follow in subsequent chapters. One feature of the way they are used in applications is that various types of restrictions might be made on some of the parameters, for example, some parameter might be required to be equal to (functions of) other parameters in the model.

**Example 13.4.1 (Markov Chain).** Let \( X_1, X_2, \ldots, X_m \) be a sequence of discrete random variables, each with state space \([r]\). Consider the directed
Figure 13.4.1. The caterpillar tree, the underlying graph of a hidden Markov model.

path graph $P_m$, with edges $i \to i + 1$ for $i \in [m - 1]$. According to the recursive factorization theorem, the joint probability distribution for the Bayesian network associated to this DAG has the form:

$$P(X_1 = x_1, \ldots, X_m = x_m) = P_1(X_1 = x_1) \prod_{i=2}^{m} P_i(X_i = x_i | X_{i-1} = x_{i-1}).$$

Note that the conditional independence structure in this model will have the form $X_{i+1} \perp\!\!\!\perp X_{[i-1]} | X_i$ for $i = 2, \ldots, m - 1$, so that the sequence of random variables forms a Markov chain, as discussed in Chapter 1.

If, in addition, we require that each of the conditional distributions $P_i$ are equal:

$$P_i(X_i = x | X_{i-1} = y) = P_2(X_2 = x | X_1 = y)$$

for all $i, x, y$ then we have a homogeneous Markov chain, which is the way that discrete Markov chains are often discussed, for example in [Nor98]. Letting $A = (a_{ij})_{i,j \in [r]}$ denote the matrix representing the conditional distribution, that is $a_{ij} = P(X_{i+1} = j | X_i = i)$, then $A$ is the transition matrix of the homogeneous Markov chain. Thus, the example of a Markov chain from Chapter 1 is an example of a Bayesian network.

As we saw in Example 13.4.1, models that can be expressed by a sequence of simple Markov processes connecting random variables can usually be thought of as graphical models on a directed acyclic graph, with the underlying DAG being a directed tree. Another important example is the hidden Markov model.

**Example 13.4.2 (Hidden Markov Model).** Let $Y_1, \ldots, Y_m$ be a sequence of discrete random variables each with state space $[r]$, and $X_1, \ldots, X_m$ a sequence of discrete random variables, each with state space $[s]$. Consider the directed graph with edges $Y_i \to Y_{i+1}$ for $i \in [m - 1]$ and $Y_i \to X_i$ for $i \in [m]$. This DAG is illustrated in Figure 13.4.1, a type of tree called a caterpillar.
The recursive factorization property for this graph yields the joint distribution
\[
P(X_1 = x_1, \ldots, X_m = x_m, Y_1 = y_1, \ldots, Y_m = y_m) =
\]
\[
P_1(Y_1 = Y_1) \prod_{i=2}^{m} P_i(Y_i = y_i | Y_{i-1} = y_{i-1}) \prod_{i=1}^{m} Q_i(X_i = x_i | Y_i = y_i).
\]
In this model, we typically require that each of the conditional distributions of $Y_i$ given $Y_{i-1}$ are the same, and that each of the conditional distributions of $X_i$ given $Y_i$ are the same. Furthermore, we assume that the random variables $Y_1, \ldots, Y_m$ are all hidden random variables, that is, they are unobserved (we discuss this point in more detail in later chapters).

In applications, one should think of the hidden Markov model in the following way. The sequence of random variables $Y_1, \ldots, Y_m$ evolves according to a homogeneous Markov chain model. However, we are not able to observe $Y$, we only observe a corrupted or noisy version of it or some information $X$ computed from $Y$ in a possibly random way. Usually we want to recover $Y$ from $X$. The hidden Markov model is commonly used in computational biology where it is used to align DNA sequences and annotate DNA for genes (see e.g. [DEKM98]). We will discuss it in more detail in Chapter 18.

The Ising model is an important interpretation of undirected graphical models. Although we usually associate the Ising model with statistical physics, this model and variations on it are also using in spatial statistical and agriculture.

**Example 13.4.3 (Ising Model).** Let $G = (V, E)$ be an undirected graph and let $X$ be a discrete random vector where all random variables are binary with state space $\{-1, 1\}$. Introduce a probability distribution
\[
P(x_1, \ldots, x_n) = \frac{1}{Z} \exp \left( \sum_{(i,j) \in E} \alpha_{i,j} x_i x_j + \sum_{i \in V} \beta_i x_i \right).
\]
The number $Z$ is the normalizing constant, in this context often called a partition function. Note that, as written, we have expressed this as an exponential family, and the resulting distribution gives a distribution in the hierarchical model associated whose underlying simplicial complex is the graph $G$. With appropriately chosen potential functions this can be realized by the graphical model associated to $G$. If the graph $G$ has no triangles, then the Ising model is equal to the positive part of the graphical model associated to $G$. The most commonly used graph for the Ising model is the grid graph.
In some applications, e.g. spatial statistics, we restrict to a submodel where all the $\alpha_{ij}$ parameters are equal and all the $\beta_i$ parameters are equal. See [Bes74] for more on applications of the Ising model in statistical settings.

### 13.5. Exercises

**Exercise 13.1.** List all the pairwise, local, and global conditional independence statements associated to the $G$ with $V = \{1, 2, 3, 4, 5, 6\}$ and undirected edge set $E = \{12, 13, 15, 23, 26, 34, 46, 56\}$.

**Exercise 13.2.** Let $G$ be the undirected path graph on 5 vertices. Give an example of distribution $P$ that satisfies all the local Markov statements associated to $G$ but not all the global Markov statements associated to $G$.

**Exercise 13.3.**

1. Let $G$ be a chordal graph. Show that there is a directed graph $H$ which realizes exactly the same global Markov statements as $G$.

2. Conversely, if $G$ is an undirected four-cycle, show that there is no directed graph which yields exactly the same global Markov statements as $G$.

**Exercise 13.4.** Give a proof of Proposition 13.1.6 by giving an example of discrete random vector satisfying the local Markov statements of a graph but that does not satisfying the conditional independence statement $X_A \perp X_B | X_C$ when $C$ does not separate $A$ and $B$.

**Exercise 13.5.** Prove Proposition 13.1.8.

**Exercise 13.6.** Give a direct proof of Corollary 13.2.4 analogous to the proof of the only if direction of the proof of the Hammersley-Clifford theorem.

**Exercise 13.7.** Among the undirected Gaussian graphical models on 5 vertices, which one has the largest maximum likelihood degree?

**Exercise 13.8.** Complete one direction of the proof of Theorem 13.3.1 as follows: Let $G$ be an undirected graph, and $I_G$ the toric ideal vanishing on the parametrized graphical models. Show that the degree 2 part of the toric ideal is spanned by the quadratic binomials associated to the global Markov property associated to $G$.

**Exercise 13.9.** Show that the maximum likelihood degree of any Bayesian network is 1. In particular, explain how to determine rational formulas for the maximum likelihood estimates of parameters in graphical models with directed acyclic graphs.
A random variable in a model is hidden or latent if it is not measured. This chapter introduces the detailed study of hidden variable models, whose various aspects will occupy us for the remainder of the book. Hidden variables tend to make models more complicated to analyze and lead to interesting geometric challenges.

Hidden random variables in a model might be present for any number of reasons. Hidden variables might represent quantities that we would directly like to measure but cannot or for which it is too expensive to measure directly. For example, in a mixture model, observed variables might represent symptom status of a subject (which are measured) and the hidden mixture classes might correspond to whether or not a subject has a particular disease. Hidden variables might represent quantities that are theoretically hypothesized to exist but are not directly measurable (in particular, the observed variables might be proxies for the unmeasured hidden variables). For example, in a factor analysis model, the hidden factors typically represent unmeasurable quantities that are supposed to give a low-dimensional explanation for the correlation between a collection of observed random variables. Or, the hidden variables might simply be covariates that were not collected in a data analysis, for whatever reason, but that are still affecting the interactions between the observed random variables.

Hidden variables in a model mean that the probability densities on the observed random variables are obtained by computing marginals of the joint distribution of a fully observed model on both observed and hidden random variables. While the fully observed model might be a familiar statistical model with nice properties (e.g. exponential family, graphical model, etc.), the simple operation of computing marginals usually produces a model that
is significantly more complex than the fully observed model. For example, hidden variable models typically have singularities, have nontrivial algebraic boundaries inside the probability simplex or cone of positive definite matrices, can fail to be identifiable in non-obvious ways, and can have multiple local maxima in their likelihood functions. The richness of the expressive power of hidden variable models is balanced by the increased difficulty of using them in practical settings. We begin to address the study of such hidden variable models in this and subsequent chapters.

14.1. Mixture Models

Mixture models are among the simplest hidden variable statistical models to describe, though they already lead to a complex mathematical structure. Mixture models can be used in many different settings. We focus in this chapter on the setting of discrete random variables. Mixture models with discrete random variables are closely related to secant and join varieties in algebraic geometry. We already introduced mixture models in Example 3.2.10 in the case of binomial random variables, but here we study them in more detail.

Consider an underlying statistical model $\mathcal{M} \subseteq \Delta_{r-1}$. Here we consider an $r$-element state space but in practice this might be naturally considered as a product space (if there are many random variables involved in the model). We let $X$ denote the random variable being modeled by $\mathcal{M}$. Let $H$ be a (hidden) random variable with state space $[k]$. In a mixture model, we suppose that for some $\pi = (\pi_1, \ldots, \pi_k) \in \Delta_{k-1}$, and $p^1, \ldots, p^k \in \mathcal{M}$ we have that

$$P(H = i) = \pi_i \quad \text{and} \quad P(X = j \mid H = i) = p^i_j,$$

where $p^i = (p^i_1, \ldots, p^i_r)$. Then we have $P(X = j) = \sum_{i=1}^k \pi_i p^i_j$ so that the probability distribution vector of $X$ is the convex combination

$$\pi_1 p^1 + \cdots + \pi_k p^k$$

of the distributions $p^1, \ldots, p^k$.

**Definition 14.1.1.** The $k$-th mixture model of the model $\mathcal{M} \subseteq \Delta_{r-1}$ is the family of probability distributions

$$\text{Mix}^k(\mathcal{M}) = \left\{ \pi_1 p^1 + \cdots + \pi_k p^k : \pi \in \Delta_{k-1}, p^1, \ldots, p^k \in \mathcal{M} \right\}.$$

A natural interpretation for the mixture model is as follows: we have a population that is divided into $k$ subpopulations. The $i$-th subpopulation follows a distribution $p^i \in \mathcal{M}$ for some observed covariate $X$. The weight $\pi_i$ is the relative proportion of the total population that belongs to the $i$-th
14.1. Mixture Models

subpopulation. Finally, it is assumed that we do not observe which subpopulation an individual belongs to. Introducing a random variable $H$ with state space $[k]$, we have that $P(H = i) = \pi_i$ and the conditional distributions of $X$ given $H$ satisfy $P(X = j \mid H = i) = p_{ij}^i$. The distributions in the $k$-th mixture model then consist of the resulting marginal distributions of the variable $X$. Each of these marginal distributions is the convex combination of $k$ distributions in $\mathcal{M}$.

A natural situation when a mixture model can be used is when $H$ represents the (binary) infection status of individuals with respect to a disease and the discrete states of $X$ records various symptoms (e.g. cough, fever, rash, etc.) that a patient might have. If it is assumed that the distribution of $X \mid H = \text{Healthy}$ and $X \mid H = \text{Infected}$ each follow a distribution in some model $\mathcal{M}$, then the distribution of just $X$ will belong to $\text{Mixt}^2(\mathcal{M})$.

**Example 14.1.2** (Mixture of Independence Model). Let $\mathcal{M}_{X_1 \perp \perp X_2}$ be the independence model on two discrete random variables. The model consists of probability matrices $P \in \Delta_R$ that are rank one as matrices. The $k$-th mixture model $\text{Mixt}^k(\mathcal{M}_{X_1 \perp \perp X_2})$ consists of all convex combinations of $k$ elements of $\mathcal{M}_{X_1 \perp \perp X_2}$. Said another way, these are probability matrices that can be written as the sum of $k$ nonnegative matrices of rank 1.

Here is a “classic” example illustrating the mixture model of the independence model [MSvS03]. Let $X_1$ be a random variable describing how much a person watches soccer, and $X_2$ a random variable describing how much hair a person has. Suppose that $X_1$ has three states: Never, Sometimes, Frequently, and $X_2$ has 4 states: Bald, Short, Medium, and Long. Although soccer watching habits and hair length seem like they are unrelated, they are not independent from each other: people with short hair tend to watch soccer more frequently. However, it might be reasonable to assume that $X_1 \perp \perp X_2 \mid H$ where $H$ is a binary hidden variable, namely, the gender. Hence, if we only observe the joint distribution of $X_1$ and $X_2$, we would observe a distribution in $\text{Mixt}^2(\mathcal{M}_{X_1 \perp \perp X_2})$.

**Example 14.1.3** (Mixture of Complete Independence Model). Consider the complete independence model $\mathcal{M}_{X_1 \perp \perp X_2 \perp \perp X_3}$ on three discrete random variables. This complete independence model consists of probability tensors $P \in \Delta_R$ that are rank one tensors. Indeed, in the model of complete independence, we have

$$P(X_1 = i_1, X_2 = i_2, X_3 = i_3) = P(X_1 = i_1)P(X_2 = i_2)P(X_3 = i_3)$$

for all $i_1 \in [r_1], i_2 \in [r_2], i_3 \in [r_3]$.

Replacing the marginal distributions of $X_1$, $X_2$, and $X_3$ with parameter vectors $\alpha \in \Delta_{r_1 - 1}$, $\beta \in \Delta_{r_2 - 1}$, and $\gamma \in \Delta_{r_3 - 1}$, we have

$$p_{i_1i_2i_3} = \alpha_{i_1}\beta_{i_2}\gamma_{i_3}$$
which represents a nonnegative $r_1 \times r_2 \times r_3$ tensor of rank 1. The $k$-th mixture model $\text{Mixt}^k(M_{X_1 \perp X_2 \perp X_3})$ consists of convex combinations of $k$ of these probability distributions. These are $r_1 \times r_2 \times r_3$ probability tensors that have nonnegative rank $\leq k$. Parametrically we can represent this model as distributions of the form

$$p_{i_1,i_2,i_3} = \sum_{h=1}^{k} \pi_h \alpha_{h,i_1} \beta_{h,i_2} \gamma_{h,i_3}.$$ 

See [Lan12] for more background on tensor rank.

It is also natural to consider more general mixture models, where we do not require that all the conditional distributions $p^i$ belong to the same underlying statistical model.

**Definition 14.1.4.** Let $M_1, \ldots, M_k \subseteq \Delta_r$ be $k$ statistical models. The mixture model $M_1 \ast \cdots \ast M_k$ consists of all distributions

$$M_1 \ast \cdots \ast M_k = \left\{ \pi_1 p^1 + \cdots + \pi_k p^k : \pi \in \Delta_{k-1}, p^i \in M_i, i = 1, \ldots, k \right\}.$$ 

Upon taking the Zariski closure, mixtures models in statistics correspond to join varieties and secant varieties in algebraic geometry.

**Definition 14.1.5.** Let $V_1, V_2, \ldots, V_k \subseteq \mathbb{K}^r$ be algebraic varieties. The join variety is the variety $V_1 \ast V_2 \ast \cdots \ast V_k = \{ \pi_1 p^1 + \cdots + \pi_k p^k : \sum_i \pi_i = 1 \text{ and } p^i \in V_i \text{ for all } i \}$ where the bar denotes the Zariski closure. The $k$-th secant variety of $V$ is the variety

$$\text{Sec}^k(V) = \text{copies of } V \ast V \ast \cdots \ast V.$$ 

The second secant variety $\text{Sec}^2(V)$ is often called the secant variety of the variety $V$ in the algebraic geometry literature.

We use the symbol $\ast$ in both the mixture model and the join. These two operations are rather different if we apply them to arbitrary sets, since the join includes a Zariski closure while the mixture operation does not. Furthermore, the mixture model has a restriction that all $\pi_i$ are nonnegative while there is no such restriction in the join. From the context, it should be clear which operation is meant: the mixture operation should only be applied to statistical models for discrete random variables, while the join operation should only be applied to algebraic varieties.
14.1. Mixture Models

Note that the Zariski closure of the mixture model equals the join of the Zariski closures, that is,
\[ \overline{\mathcal{M}_1 \cdots \mathcal{M}_k} = \overline{\mathcal{M}_1} \ast \cdots \ast \overline{\mathcal{M}_k}. \]
The definition of the join variety explicitly includes a Zariski closure. This is because there can be points on the join variety that are the limits of points on secant planes but are not themselves on secant planes. For example, this is usually the case with points that lie on the tangent planes to a variety.

For any model \( \mathcal{M} \subseteq \Delta_{r-1} \), we have the containment
\[ \text{Mixt}^k(\mathcal{M}) \subseteq \left( \text{Sec}^k(\mathcal{M}) \cap \Delta_R \right). \]
Examples 14.1.6 and 14.1.7 show that these two sets need not be equal, in general.

**Example 14.1.6** (Mixture of Independence Model). Consider the mixture model of the independence model of discrete random variables \( X_1 \) and \( X_2 \). As we discussed in Example 14.1.2, the mixture model \( \text{Mixt}^k(\mathcal{M}_{X_1 \perp X_2}) \) consists of all matrices that are probability distributions and that can be written as a sum of \( k \) nonnegative matrices of rank 1. For a matrix \( A \in \mathbb{R}_{\geq 0}^{r_1 \times r_2} \) the nonnegative rank of \( A \), denoted \( \text{rank}_+(A) \), is the smallest \( k \) such that \( A \) can be written as the sum of \( k \) nonnegative matrices of rank 1. So
\[ \text{Mixt}^k(\mathcal{M}_{X_1 \perp X_2}) = \{ P \in \Delta_R : \text{rank}_+(P) \leq k \}. \]
On the other hand,
\[ \text{Sec}^k(\mathcal{M}_{X_1 \perp X_2}) \cap \Delta_R = \{ P \in \Delta_R : \text{rank}(P) \leq k \}. \]
For an arbitrary matrix \( A \), \( \text{rank}(A) \leq \text{rank}_+(A) \). However, there are abundant examples of matrices where \( \text{rank}(A) < \text{rank}_+(A) \). For example, for the probability distribution:
\[
\begin{pmatrix}
1 & 1 & 0 & 0 \\
0 & 1 & 1 & 0 \\
0 & 0 & 1 & 1 \\
1 & 0 & 0 & 1
\end{pmatrix}
\]

it is easy to see that \( \text{rank}(P) = 3 \) but \( \text{rank}_+(P) = 4 \). This shows, in particular, that
\[ \text{Mixt}^3(\mathcal{M}_{X_1 \perp X_2}) \neq \text{Sec}^3(\mathcal{M}_{X_1 \perp X_2}) \cap \Delta_R \]
where \( X_1 \) and \( X_2 \) both have 4 or more states. In particular, the mixture model \( \text{Mixt}^3(\mathcal{M}_{X_1 \perp X_2}) \) has a nontrivial boundary. A complete description of this semialgebraic set appears in [EHtK16].
Example 14.1.7 (Mixture of Complete Independence Model). Consider
the complete independence model $\mathcal{M}_{X_1 \perp X_2 \perp X_3}$ for binary random vari-
ables $X_1, X_2, X_3$ and its mixture model $\text{Mixt}^2(\mathcal{M}_{X_1 \perp X_2 \perp X_3}) \subseteq \Delta_7$. Note
that on the algebraic geometric side, the complete independence model
corresponds to the Segre variety $\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \subseteq \mathbb{P}^7$. The secant variety $\text{Sec}^2(\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1)$ has dimension seven (this can be proven using Kruskal’s
Theorem (Theorem 16.3.2)) and so fills all of $\mathbb{P}^7$. The mixture model, how-
ever, $\text{Mixt}^2(\mathcal{M}_{X_1 \perp X_2 \perp X_3})$ is very far from filling all of $\Delta_7$. As shown in
[ARSZ15], a probability distribution lies in $\text{Mixt}^2(\mathcal{M}_{X_1 \perp X_2 \perp X_3})$ if and
only if it is $\pi$-supermodular for some $\pi \in S_2 \times S_2 \times S_2$. As a semialgebraic
set, the mixture model decomposes into 4 components, one of which is the
set of distributions satisfying the following system of inequalites:

\[
p_{111}p_{222} \geq p_{112}p_{212}, \quad p_{111}p_{222} \geq p_{121}p_{212}, \quad p_{111}p_{222} \geq p_{211}p_{122},
\]
\[
p_{112}p_{222} \geq p_{122}p_{212}, \quad p_{112}p_{222} \geq p_{212}p_{122}, \quad p_{211}p_{122} \geq p_{212}p_{211}, \quad p_{211}p_{122} \geq p_{221}p_{121}, \quad p_{211}p_{122} \geq p_{221}p_{211}.
\]

The other 3-components are obtained from this one by applying the sym-
metries in $S_2 \times S_2 \times S_2$. Simulations suggest that only about 8% of the
probability simplex $\Delta_7$ lies in $\text{Mixt}^2(\mathcal{M}_{X_1 \perp X_2 \perp X_3})$.

Even the mixture model $\text{Mixt}^3(\mathcal{M}_{X_1 \perp X_2 \perp X_3})$ does not fill the proba-
bility simplex. As shown in [MM15], a distribution in the mixture model
$\text{Mixt}^k(\mathcal{M}_{X_1 \perp X_2 \perp \ldots \perp X_m})$ can have at most $k$ strong maxima. A strong
maximum is an index string $(i_1, \ldots, i_m) \in [r_1] \times \cdots \times [r_m]$ such that

\[
p_{i_1 \ldots i_m} > \sum_{(j_1 \ldots j_m) : d_H((i_1 \ldots i_m), (j_1 \ldots j_m)) = 1} p_{j_1 \ldots j_m}
\]

where $d_H(i, j)$ denotes the Hamming distance between strings $i$ and $j$. The
distribution $p \in \Delta_7$ with

\[
p_{111} = p_{122} = p_{212} = p_{221} = \frac{1}{4} \quad \text{and} \quad p_{112} = p_{121} = p_{211} = p_{222} = 0
\]

has four strong maxima, so does not belong to $\text{Mixt}^3(\mathcal{M}_{X_1 \perp X_2 \perp X_3})$. Fur-
thermore, all probability distributions in $\Delta_7$ that are in a small neigh-
borhood of $p$ also have four strong maxima and so are also excluded from
the third mixture model. Seigal and Montúfar [SG17] gave a precise semi-
algebraic description of the three class mixture model $\text{Mixt}^3(\mathcal{M}_{X_1 \perp X_2 \perp X_3})$
and showed via simulations that it fills approximately 75.3% of the proba-
bility simplex.

On the other hand, it can be shown that $\text{Mixt}^4(\mathcal{M}_{X_1 \perp X_2 \perp X_3}) = \Delta_7$.
As this example illustrates, there can be a significant discrepancy between
the mixture model and the intersection of its Zariski closure with the proba-
bility simplex.
A common feature of hidden variable models is that they have singularities. This makes traditional asymptotic results about statistical inference procedures based on hidden variable models more complicated. For example, the asymptotics of the likelihood ratio test statistic will depend on the structure of the tangent cone at singular points, as studied in Section 7.4. Singularities will also play an important role in the study of marginal likelihood integrals in Chapter 17. Here is the most basic result on singularities of secant varieties.

**Proposition 14.1.8.** Let $V \subseteq \mathbb{P}^{r-1}$ and suppose that $\text{Sec}^k(V)$ is not a linear space. Then $\text{Sec}^{k-1}(V)$ is in the singular locus of $\text{Sec}^k(V)$.

**Proof.** The singular locus of $\text{Sec}^k(V)$ is determined by the drop in rank of the Jacobian evaluated at a point. However, if $\text{Sec}^k(V)$ is not a linear space, the main result of [SS09] shows that the Jacobian matrix of the equations of $\text{Sec}^k(V)$ evaluated at a point of $\text{Sec}^{k-1}(V)$ is the zero matrix. Hence all such points are singular. 

**Example 14.1.9 (Singular Locus of Secant Varieties of Segre Varieties).** Consider the Segre variety $V = \mathbb{P}^{r_1-1} \times \mathbb{P}^{r_2-1} \subseteq \mathbb{P}^{r_1+r_2-1}$, the projectivization of the set of rank one matrices. The secant variety $\text{Sec}^k(V)$ is the projectivization of the set of matrices of rank $\leq k$, and its vanishing ideal $I(\text{Sec}^k(V))$ is generated by the $k+1$ minors of a generic $r_1 \times r_2$ matrix, provided $k < \min(r_1,r_2)$. The partial derivative of a generic $k+1$ minor with respect to one of its variables is a generic $k$-minor. Hence we see that the Jacobian matrix of the ideal $I(\text{Sec}^k(V))$ vanishes at all points on $\text{Sec}^{k-1}(V)$ and, in particular, is rank deficient. For statistical models, this example implies that the mixture model $\text{Mixt}^k(M_{X_1 \sqcup X_2})$ is singular along the submodel $\text{Mixt}^{k-1}(M_{X_1 \sqcup X_2})$, provided that $k < \min(r_1,r_2)$.

One situation where mixture models arise unexpectedly is in the study of exchangeable random variables. De Finetti’s Theorem shows that an infinite sequence of exchangeable random variables can be realized as a unique mixture of i.i.d. random variables. Interesting convex algebraic geometry questions arise when we only consider finite sequences of exchangeable random variables.

**Definition 14.1.10.** A finite sequence $X_1, X_2, \ldots, X_n$ of random variables on a countable state space $\mathcal{X}$ is exchangeable if for all $\sigma \in S_n$, and $x_1, \ldots, x_n \in \mathcal{X}$,

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_{\sigma(1)}, \ldots, X_n = x_{\sigma(n)}).$$

An infinite sequence $X_1, X_2, \ldots$, of random variables on a countable state space $\mathcal{X}$ is exchangeable if each of its finite subsequences is exchangeable.
De Finetti’s Theorem relates infinite sequences of exchangeable random variables to mixture models. We state this result in only its simplest form, for binary random variables.

**Theorem 14.1.11 (De Finetti’s Theorem).** Let $X_1, X_2, \ldots$ be an infinite sequence of exchangeable random variables with state space $\{0, 1\}$. Then there exists a unique probability measure $\mu$ on $[0, 1]$ such that for all $n$, and $x_1, \ldots, x_n \in \{0, 1\}$

$$P(X_1 = x_1, \ldots, X_n = x_n) = \int_0^1 \theta^k (1 - \theta)^{n-k} d\mu(\theta)$$

where $k = \sum_{i=1}^n x_i$.

Another way to say De Finetti’s theorem is that an infinite exchangeable sequence of binary random variables is a mixture of i.i.d. Bernoulli random variables. Because we work with infinite sequences we potentially require a mixture of an infinite and even uncountable number of i.i.d. Bernoulli random variables, and this is captured in the mixing distribution $\mu$.

De Finetti’s theorem is useful because it allows us to relate the weak Bayesian notion of exchangeability back to more familiar strong notion of independent identically distributed samples. It is also a somewhat surprising result since finite exchangeable sequences need not be mixtures of i.i.d. Bernoulli random variables. Indeed, consider the exchangeable sequence of random variables $X_1, X_2$ where

$$p_{00} = p_{11} = 0 \quad \text{and} \quad p_{01} = p_{10} = \frac{1}{2}.$$  

This distribution could not be a mixture of i.i.d. Bernoulli random variables, because any such mixture must have either $p_{00} > 0$ or $p_{11} > 0$ or both.

One approach to proving Theorem 14.1.11 is to develop a better geometric understanding of the failure of a finite exchangeable distribution to be mixtures. This is the approach taken in [Dia77, DF80]. Consider the set of $EX_n \subseteq \Delta_{2^n-1}$ of exchangeable distributions for $n$ binary random variables. Since the exchangeable condition is a linear condition on the elementary probabilities, this set $EX_n$ is a polytope inside $\Delta_{2^n-1}$. This polytope is itself a simplex of dimension $n+1$, and its vertices are the points $e_0, \ldots, e_n$, where

$$(e_i)_{j_1, \ldots, j_n} = \begin{cases} \frac{1}{\binom{n}{i}} & \text{if } j_1 + \cdots + j_n = i \\ 0 & \text{otherwise.} \end{cases}$$

The set of mixture distributions of i.i.d. Bernoulli random variables, is the convex hull of the curve

$$C_n = \{ (\theta^{j_1 + \cdots + j_n} (1 - \theta)^{n-j_1 - \cdots - j_n})_{j_1, \ldots, j_n \in \{0, 1\}} : \theta \in [0, 1] \}.$$
Example 14.1.12 (Exchangeable Polytope). If \( n = 2 \), the polytope \( EX_2 \) has vertices \((1, 0, 0, 0), (0, 1, 0, 0), \) and \((0, 0, 0, 1)\). The curve \( C_2 \) is the set
\[
C_2 = \{ ((1 - \theta)^2, \theta(1 - \theta), \theta(1 - \theta), \theta^2) : \theta \in [0, 1] \}.
\]
In this example, the basic outline of the polytope \( EX_2 \) and the curve \( C_2 \) is the same as in Figure 3.1.1.

For \( k < n \) we can consider the linear projection map \( \pi_{n,k} : EX_n \rightarrow EX_k \) that is obtained by computing a marginal over \( n - k \) of the variables. The image is the subset of \( EX_k \) consisting of exchangeable distributions on \( k \) random variables that lift to an exchangeable distribution on \( n \) random variables. In other words, these correspond to exchangeable sequences \( X_1, \ldots, X_k \) which can be completed to an exchangeable sequence \( X_1, \ldots, X_n \). Let \( EX_{n,k} = \pi_{n,k}(EX_n) \). Since it is more restrictive to be exchangeable on more variables, we have that \( EX_{n+1,k} \subseteq EX_{n,k} \). Diaconis’ proof of Theorem 14.1.11 amounts to proving, in a suitable sense, that
\[
\lim_{n \to \infty} EX_{n,k} = \text{Mix}^k(C_k)
\]
for all \( k \).

14.2. Hidden Variable Graphical Models

One of the simplest ways to construct hidden variable statistical models is using a graphical model where some variables are not observed. Graphical models where all variables are observed are simple models in numerous respects: They are exponential families (in the undirected case), they have closed form expressions for their maximum likelihood estimates (in the directed acyclic case), they have distributions largely characterized by conditional independence constraints, and the underlying models are smooth. Graphical models with hidden variables tend to lack all of these nice features. As such, these models present new challenges for their mathematical study.

Example 14.2.1 (Claw Tree). As a simple example in the discrete case, consider the claw tree \( G \) with four vertices and with edges \( 1 \rightarrow 2, 1 \rightarrow 3, 1 \rightarrow 4 \), and let vertex 1 be the hidden variable, illustrated in Figure 14.2.1.

This graphical model is parametrized by the recursive factorization theorem:
\[
p_{i_2i_3i_4} = \phi(\pi, \alpha, \beta, \gamma) = \sum_{i_1=1}^{\tau_1} \pi_{i_1} \alpha_{i_1i_2} \beta_{i_1i_3} \gamma_{i_1i_4},
\]
where
\[
\pi_{i_1} = P(X_1 = i_1), \quad \alpha_{i_1i_2} = P(X_2 = i_2 | X_1 = i_1), \quad \beta_{i_1i_3} = P(X_3 = i_3 | X_1 = i_1), \quad \text{and} \quad \gamma_{i_1i_4} = P(X_4 = i_4 | X_1 = i_1).\]
This parametrization is the same as the one we saw for the mixture model \( \text{Mixt}^1(\mathcal{M}_{X_2 \perp X_3 \perp X_4}) \) in the previous section. Since the complete independence model \( \mathcal{M}_{X_2 \perp X_3 \perp X_4} \) is just the graphical model associated to the graph with 3 vertices and no edges, the hidden variable graphical model associated to \( G \) corresponds to a mixture model of a graphical model.

Note that the fully observed model, as a graphical model with a directed acyclic graph, has its probability distributions completely characterized by conditional independence constraints. The conditional independence constraints for the fully observed model are \( X_2 \perp (X_3, X_4) | X_1 \), \( X_3 \perp (X_2, X_4) | X_1 \), and \( X_4 \perp (X_2, X_3) | X_1 \). Since there are no conditional independence statements in this list, or implied using the conditional independence axioms, that do not involve variable \( X_1 \), we see that we cannot characterize the distributions of this hidden variable graphical model by conditional independence constraints among only the observed variables.

Graphical models with hidden variables where the underlying graph is a tree play an important role in phylogenetics. These models will be studied in detail in Chapter 15. The model associated to the claw tree is an important building block for the algebraic study of phylogenetic models and plays a key role in the Allman-Rhodes-Draisma-Kuttler Theorem (Section 15.5).

Now we move to the study of hidden variable models under a multivariate normal distribution. Let \( X \sim \mathcal{N}(\mu, \Sigma) \) be a normal random vector, and \( A \subseteq [m] \). Recall from Theorem 2.4.2 that \( X_A \sim \mathcal{N}(\mu_A, \Sigma_{A,A}) \). This immediately implies that when we pass to the vanishing ideal of a hidden variable model we can compute it by computing an elimination ideal.

**Proposition 14.2.2.** Let \( \mathcal{M} \subseteq \mathbb{R}^m \times PD_m \) be an algebraic exponential family with vanishing ideal \( I = I(\mathcal{M}) \subseteq \mathbb{R}[\mu, \Sigma] \). Let \( H \sqcup O = [m] \) be a partition of the index labeling into hidden variables \( H \) and observed variables \( O \). The hidden variable model consists of all marginal distributions on the variables \( X_O \) for a distribution with parameters in \( \mathcal{M} \). The vanishing ideal of the hidden variable model is the elimination ideal

\[
I \cap \mathbb{R}[\mu_O, \Sigma_{O,O}].
\]
Example 14.2.3 (Gaussian Claw Tree). Consider the graphical model associated to the directed claw graph from Example 14.2.1 but now under a multivariate normal distribution. The vanishing ideal of the fully observed model is generated by the degree 2 determinantal constraints:

\[ I(\mathcal{M}_G) = \langle \sigma_{11}\sigma_{23} - \sigma_{13}\sigma_{14}, \sigma_{12}\sigma_{14} - \sigma_{14}\sigma_{23}, \sigma_{11}\sigma_{34} - \sigma_{13}\sigma_{14}, \sigma_{12}\sigma_{34} - \sigma_{14}\sigma_{23} \rangle. \]

The hidden variable graphical model where \( X_1 \) is hidden is obtained by computing the elimination ideal

\[ I(\mathcal{M}_G) \cap \mathbb{R}[\sigma_{22}, \sigma_{23}, \sigma_{24}, \sigma_{33}, \sigma_{34}, \sigma_{44}]. \]

In this case we get the zero ideal. Note that the resulting model is the same as the 1-factor model \( F_{3,1} \) which was studied in Example 6.5.4. In particular, the resulting hidden variable model, in spite of having no nontrivial equality constraints vanishing on it, is subject to complicated semialgebraic constraints.

Our usual intuition about hidden variable models is that they are much more difficult to describe than the analogous fully observed models, and that these difficulties should appear in all contexts (e.g. computing vanishing ideals, describing the semialgebraic structure, solving the likelihood equations, etc.). Surprisingly, for Gaussian graphical models, computing the vanishing ideal is just as easy for hidden variable models as for the fully observed models, at least in one important case.

Definition 14.2.4. Let \( G = (V,D) \) be a directed acyclic graph, and let \( H \uplus O = [m] \) be a partition of the index labeling into hidden variables \( H \) and observed variables \( O \). The hidden variables are said to be upstream of the observed variables if there is no edge \( o \rightarrow h \), where \( o \in O \) and \( h \in H \).

Suppose that \( H \uplus O \) is a partition of the variables in a DAG, where the hidden variables are upstream. We introduce the following two dimensional grading on \( \mathbb{R}[\Sigma] \), associated to this partition of the variables:

\[
\deg \sigma_{ij} = \left( \frac{1}{\#(\{i\} \cap O) + \#(\{j\} \cap O)} \right).
\]

Proposition 14.2.5. Let \( G = (V,D) \) be a directed acyclic graph and let \( H \uplus O = [m] \) be a partition of the index labeling into hidden variables \( H \) and observed variables \( O \), where the \( H \) variables are upstream. Then the ideal \( J_G \subseteq \mathbb{R}[\Sigma] \) is homogeneous with respect to the upstream grading in Equation 14.2.1. In particular, any homogeneous generating set of \( J_G \) in this grading contains as a subset a generating set of the vanishing ideal of the hidden variable model \( J_G \cap \mathbb{R}[\Sigma_{O,O}] \).
Proof. The fact that the ideal $J_G$ is homogeneous in this case is proven in [Sul08]. The second statement follows by noting that the subring $\mathbb{R}[\Sigma_{O,O}]$ consists of all elements of $\mathbb{R}[\Sigma_{O,O}]$ whose degree lies on the face of the grading semigroup generated by the ray $(1)$. □

For instance, in the case of the factor analysis model from Example 14.2.3 all the generators of the ideal $J_G$ that are listed are homogeneous with respect to the upstream grading. Since none of these generating polynomials involve only the indices $\{2, 3, 4\}$, we can see that the vanishing ideal of the hidden variable model must be the zero ideal. We can also look at this the other way around, namely the hardness of studying vanishing ideals of hidden variable models leads to difficulties in computing vanishing ideals of Gaussian graphical models without hidden variables.

Example 14.2.6 (Factor Analysis). The factor analysis model $F_{m,s}$ is the hidden variable model associated to the graph $K_{s,m}$ the directed complete bipartite graph with vertex set $H$ and $O$ of sizes $s$ and $m$ respectively with every edge $h \to o$ for $h \in H$ and $o \in O$, and no other edges. The vanishing ideal of the factor analysis model $F_{5,2}$ consists of a single polynomial of degree 5, the pentad (see e.g. [DSS07]), which is the polynomial

$$f = \sigma_{12}\sigma_{13}\sigma_{24}\sigma_{35}\sigma_{45} - \sigma_{12}\sigma_{13}\sigma_{25}\sigma_{34}\sigma_{45} - \sigma_{12}\sigma_{14}\sigma_{23}\sigma_{35}\sigma_{45}$$

$$+\sigma_{12}\sigma_{14}\sigma_{25}\sigma_{34}\sigma_{35} + \sigma_{12}\sigma_{15}\sigma_{23}\sigma_{34}\sigma_{45} - \sigma_{12}\sigma_{15}\sigma_{24}\sigma_{34}\sigma_{35}$$

$$+\sigma_{13}\sigma_{14}\sigma_{23}\sigma_{25}\sigma_{45} - \sigma_{13}\sigma_{14}\sigma_{24}\sigma_{25}\sigma_{35} - \sigma_{13}\sigma_{15}\sigma_{23}\sigma_{24}\sigma_{45}$$

$$+\sigma_{13}\sigma_{15}\sigma_{24}\sigma_{25}\sigma_{34} - \sigma_{14}\sigma_{15}\sigma_{23}\sigma_{25}\sigma_{45} + \sigma_{14}\sigma_{15}\sigma_{23}\sigma_{24}\sigma_{35}.$$  

Proposition 14.2.5 implies that any minimal generating set of the fully observed model associated to the graph $K_{2,5}$ must contain a polynomial of degree 5.

Graphical models with hidden variables are most useful in practice if we know explicitly all the hidden structures that we want to model. For example, in the phylogenetics models studied in Chapter 15, the hidden variables correspond to ancestral species, which, even if we do not know exactly what these species are, we know they must exist by evolutionary theory. On the other hand, in some situations it is not clear what the hidden variables are, nor how many hidden variables should be included. In some instances with Gaussian graphical models, there are ways around this difficulty.

Example 14.2.7 (Instrumental Variables). Consider the graph on four nodes in Figure 14.2.2. Here $X_1, X_2, X_3$ are the observed variables and $H$ is a hidden variable. Such a model is often called an instrumental variable model. Here is the way that it is used. The goal of an investigator is to measure the direct effect of variable $X_2$ on $X_3$. However, there is an unobserved
14.2. Hidden Variable Graphical Models

hidden variable $H$ which affects both $X_2$ and $X_3$, confounding our attempts to understand the effect of $X_2$ on $X_3$. To defeat this confounding factor we measure a new variable $X_1$ which we know has a direct effect on $X_2$ but is otherwise independent of $X_3$. The variable $X_1$ is called the instrumental variable. (We will discuss issue of whether or not it is possible to recover the direct effect of $X_2$ on $X_3$ in such a model in Chapter 16 on identifiability.)

To make this example concrete, suppose that $X_2$ represents how much a mother smokes and $X_3$ represents the birth weight of that mother’s child. One wants to test the magnitude of the causal effect of smoking on the birth weight of the child. Preventing us from making this measurement directly is the variable $H$, which might correspond to some hypothesized genetic factor affecting how much the mother smokes and the child’s birth weight (R.A. Fisher famously hypothesized there might be a genetic factor that makes both a person want to smoke more and be more likely to have lung cancer [Sto91]), or a socioeconomic indicator (e.g. how rich/poor the mother is), or some combination of these. The instrumental variable $X_1$ would be some variable that directly affects how much a person smokes, but not the child development, e.g. the size of taxes on cigarette purchases.

The complication for modeling that arises in this scenario is that there might be many variables that play a direct or indirect role confounding $X_2$ and $X_3$. It is not clear a priori which variables to include in the analysis. Situations like this can present a problem for hidden variable modeling.

To get around the problem illustrated in Example 14.2.7 of not knowing which hidden variables to include in the model, in the case of Gaussian graphical models there is a certain work-around that can be used. The resulting models are called linear structural equation models. Let $G = (V, B, D)$ be a mixed graph with vertex set $V$ and two types of edges: bidirected edges $i \leftrightarrow j \in B$ and directed edges $i \rightarrow j \in D$. Technically, the bidirected edges are really no different from undirected edges, but the double arrowheads are useful for a number of a reasons. It is allowed that both

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{instrumental_variables_graph.png}
\caption{Graph for an instrumental variables model with a hidden variable}
\end{figure}
Hidden Variables

$i \rightarrow j$ and $i \leftrightarrow j$ might both be edges in the same mixed graph. Let

$$PD(B) = \{\Omega \in PD_m : \omega_{ij} = 0 \text{ if } i \neq j \text{ and } i \leftrightarrow j \notin B\}$$

be the set of positive definite matrices with zeroes in off-diagonal positions corresponding to $ij$ pairs where $i \leftrightarrow j$ is not a bidirected edge. To each directed edge $i \rightarrow j$ we have a real parameter $\lambda_{ij} \in \mathbb{R}$. Let

$$\mathbb{R}^D = \{\Lambda \in \mathbb{R}^{m \times m} : \lambda_{ij} = \lambda_{ji} \text{ if } i \rightarrow j \in D, \text{ and } \lambda_{ij} = 0 \text{ otherwise}\}.$$ 

Let $\epsilon \sim \mathcal{N}(0, \Omega)$ where $\Omega \in PD(B)$. Define the random variables $X_j$ for $k \in V$ by

$$X_j = \sum_{k \in pa(j)} \lambda_{kj} X_k + \epsilon_j.$$ 

Note that this is the same linear relations we saw for directed Gaussian graphical models (Equation 13.2.3) except that now we allow correlated error terms $\epsilon \sim \mathcal{N}(0, \Omega)$. Following the proof of Proposition 13.2.12, a similar factorization for the covariance matrix of $X$ can be computed.

**Proposition 14.2.8.** Let $G = (V, B, D)$ be a mixed graph. Let $\Omega \in PD(B)$, and $\epsilon \sim \mathcal{N}(0, \Omega)$ and $\Lambda \in \mathbb{R}^D$. Then the random vector $X$ is a multivariate normal random variable with covariance matrix

$$\Sigma = (Id - \Lambda)^{-T} \Omega (Id - \Lambda)^{-1}.$$ 

The set of all such covariance matrices that arises is the linear structural equation model.

**Definition 14.2.9.** Let $G = (V, B, D)$ be a mixed graph. The linear structural equation model $\mathcal{M}_G \subseteq PD_m$ consists of all covariance matrices

$$\mathcal{M}_G = \{(Id - \Lambda)^{-T} \Omega (Id - \Lambda)^{-1} : \Omega \in PD(B), \Lambda \in \mathbb{R}^D\}.$$ 

In the linear structural equation model, instead of having hidden variables we have bidirected edges that represent correlations between the error terms. This allows us to avoid specifically modeling what types of hidden variables might be causing interactions between the variables. The bidirected edges in these models allow us to explicitly relate error terms that we think should be correlated because of unknown confounding variables.

The connection between graphical models with hidden variables and the linear structural equation models is obtained by imagining the bidirected edge $i \leftrightarrow j$ as a subdivided pair of edges $i \leftarrow k \rightarrow j$.

**Definition 14.2.10.** Let $G = (V, B, D)$ be a mixed graph. Let $G^{sub}$ be the directed graph obtained from $G$ whose vertex set is $V \cup B$, and with edge set

$$D \cup \{b \rightarrow i : i \leftrightarrow j \in B\}.$$
The resulting graph $G_{\text{sub}}$ is the bidirected subdivision of $G$.

For example, the graph in Figure 14.2.2 is the bidirected subdivision of the graph in Figure 14.2.3.

**Proposition 14.2.11.** Let $G = (V, B, D)$ be a mixed graph with $m$ vertices and $G_{\text{sub}}$ the bidirected subdivision. Let $\mathcal{M}_G \subseteq PD_m$ be the linear structural equation model associated to $G$ and $\mathcal{M}'_{G_{\text{sub}}}$ the Gaussian graphical model associated to the directed graph $G_{\text{sub}}$ where all variables $B$ are hidden variables. Then $\mathcal{M}'_{G_{\text{sub}}} \subseteq \mathcal{M}_G$, and $I(\mathcal{M}'_{G_{\text{sub}}}) = I(\mathcal{M}_G)$.

One point of Proposition 14.2.11 is to apply it in reverse: take a model with hidden variables and replace the hidden variables and outgoing edges with bidirected edges instead, which should, in principle, be easier to work with. To prove Proposition 14.2.11 we will need to pay attention to a comparison between the parametrization of the model $\mathcal{M}_G$ and the model $\mathcal{M}'_{G_{\text{sub}}}$. The form of the parametrization of a linear structural equation model is described by the trek rule.

**Definition 14.2.12.** Let $G = (V, B, D)$ be a mixed graph. A trek from $i$ to $j$ in $G$ consists of either

1. a directed path $P_L$ ending in $i$, and a directed path $P_R$ ending in $j$ which have the same source, or
2. a directed path $P_L$ ending in $i$, and a directed path $P_R$ ending in $j$ such that the source of $P_L$ and $P_R$ are connected by a bidirected edge.

Let $\mathcal{T}(i, j)$ denote the set of treks in $G$ connecting $i$ and $j$.

To each trek $T = (P_L, P_R)$ we associate the trek monomial $m_T$ which is the product with multiplicities of all $\lambda_{st}$ over all edges appearing in $T$ times $\omega_{st}$ where $s$ and $t$ are the sources of $P_L$ and $P_R$. Note that this generalizes the definition of trek that we saw in Definition 13.2.15 to include bidirected edges. The proof of Proposition 14.2.11 will be based on the fact that the bidirected edges in a trek act exactly like the corresponding pair of edges in a bidirected subdivision.

**Proposition 14.2.13 (Trek Rule).** Let $G = (V, B, D)$ be a mixed graph. Let $\Omega \in PD(B)$, $\Lambda \in \mathbb{R}^D$, and $\Sigma = (Id - \Lambda)^{-T} \Omega (Id - \Lambda)^{-1}$. Then

$$
\sigma_{ij} = \sum_{T \in \mathcal{T}(i,j)} m_T.
$$

**Proof.** Note that in the matrix

$$(Id - \Lambda)^{-1} = Id + \Lambda + \Lambda^2 + \cdots$$
The \((i, j)\) entries is the sum over all directed paths from \(i\) to \(j\) in the graph \(G\) of a monomial which consists of the product of all edges appearing on that path. The transpose yields paths going in the reverse direction. Hence the proposition follows from multiplying the three matrices together. \(\square\)

**Example 14.2.14** (Instrumental Variables Mixed Graph). Consider the mixed graph in Figure 14.2.3. We have

\[
\Omega = \begin{pmatrix}
\omega_{11} & 0 & 0 \\
0 & \omega_{22} & \omega_{23} \\
0 & \omega_{23} & \omega_{33}
\end{pmatrix}, \quad \Lambda = \begin{pmatrix}
0 & \lambda_{12} & 0 \\
0 & 0 & \lambda_{23} \\
0 & 0 & 0
\end{pmatrix}.
\]

So in \(\Sigma = (Id - \Lambda)^{-T}\Omega(Id - \Lambda)^{-1}\) we have

\[
\sigma_{23} = \omega_{11}\lambda_{12}^2\lambda_{23} + \omega_{22}\lambda_{23} + \omega_{23}
\]

which is realized as the sum over the three distinct treks from 2 to 3 in the graph.

**Proof of Proposition 14.2.11.** We need to show how we can choose parameters for the graph \(G\) to get any covariance matrix associated to the hidden variable model associated to the graph \(G^{sub}\). Use the same \(\lambda_{ij}\) parameters for each \(i \rightarrow j\) edge in \(D\). Let \(\Omega'\) be the diagonal matrix of parameters associated to the bidirected subdivision \(G^{sub}\). For each bidirected edge \(b = i \leftrightarrow j\) let

\[
\omega_{ij} = \omega'_{bb}\lambda_{bi}\lambda_{bj} \quad \text{and} \quad \omega_{ii} = \omega'_{ii} + \sum_{b=i \rightarrow j \in B} \omega'_{bb}\lambda_{bi}^2.
\]

Note that the resulting matrix \(\Omega\) is a positive definite matrix in \(PD(B)\) since it can be expressed as a sum of a set of PSD rank one matrices (one for each edge in \(B\)) plus a positive diagonal matrix. Furthermore, the trek rule implies that this choice of \(\Omega\) gives the same matrix as from the hidden variable graphical model associated to the graph \(G^{sub}\).
To see that the vanishing ideals are the same, it suffices to show that we can take generic parameter values in the model $\mathcal{M}'_{G_{srh}}$ and produce parameter values for $\mathcal{M}_G$ that yield the same matrix. However, this inversion process will not always yield a positive definite matrix for $\Omega'$, which is why we do not yield equality for the two models. To do this, we invert the construction in the previous paragraph. In particular, take the same $\lambda_{ij}$ parameters in both graphs. For each bidirected edge $b = i \leftrightarrow j$ let

$$
\omega'_{bb} = \omega_{ij} \lambda_{bi}^{-1} \lambda_{bj}^{-1} \quad \text{and} \quad \omega'_{ii} = \omega_{ii} - \sum_{b = i \to j \in B} \omega_{bb} \lambda_{bi}^2.
$$

\[\square\]

**Example 14.2.15** (Instrumental Variables Mixed Graph). For the instrumental variable model from Example 14.2.7, the associated graph with bidirected edges is illustrated in Figure 14.2.3. Note that now we do not need to make any particular assumptions about what might be the common cause for correlations between smoking and babies birth weight. These are all subsumed in the condition that there is nontrivial covariance $\omega_{23}$ between $\varepsilon_2$ and $\varepsilon_3$.

While Proposition 14.2.11 shows that the models $\mathcal{M}_G$ and $\mathcal{M}'_{G_{srh}}$ are related by containment, they need not be equal for general graphs $G$. The argument in the proof of Proposition 14.2.11 implies that they have the same vanishing ideal, but the semialgebraic structure might be different. The paper \cite{DY10} has a detailed analysis in the case of graphs with only bidirected edges. See also Exercise 14.7.

### 14.3. The EM Algorithm

Even if one begins with a statistical model with concave likelihood function, or a model with closed form expressions for maximum likelihood estimates, in associated hidden variable models the likelihood function might have multiple local maxima and large maximum likelihood degree. When the underlying fully observed model is an exponential family, the EM algorithm \cite{DLR77} is a general purpose algorithm for attempting to compute the maximum likelihood estimate in the hidden variable model that is guaranteed to not decrease the likelihood function at each step, though it might not converge to the global maximum of the likelihood function. Because the EM algorithm is so widely used for calculating maximum likelihood estimates, it deserves further mathematical study.

The EM algorithm consists of two steps, the Expectation Step and the Maximization Step, which are iterated until some termination criterion is reached. In the Expectation Step we compute the expected values of the sufficient statistics in the fully observed model given the current parameter estimates. In the Maximization Step these estimated sufficient statistics are
used to compute the maximum likelihood estimates in the fully observed model to get new estimates of the parameters. The fact that this procedure leads to a sequence of estimates with increasing likelihood function values appears in [DLR77], but we will give a simplified argument shortly in the case of discrete random variables.

Here is the structure of the algorithm with discrete random variables in a hidden variable model. Assume for simplicity that we work with only two random variables, an observed random variable $X$ with state space $\{r_1\}$ and a hidden random variable $Y$ with state space $\{r_2\}$. Let $p_{ij} = P(X = i, Y = j)$ denote a coordinate of the joint distribution, and let $p_i = \sum_{j=1}^{r_2} p_{ij}$ be a coordinate of the marginal distribution of $X$. Let $\mathcal{M} \subseteq \Delta_{r_1 r_2 - 1}$ be a fully observed statistical model and $\mathcal{M}' \subseteq \Delta_{r_1 - 1}$ be the hidden variable model obtained by computing marginal distributions.

Let $p(t) \in \mathcal{M} \subseteq \Delta_{r_1 r_2 - 1}$ denote the $t$-th estimate for the maximum likelihood estimate of the fully observed model.

**Algorithm 14.3.1. (Discrete EM algorithm)**

**Input:** Data $u \in \mathbb{N}^r$, an initial estimate $p^{(0)} \in \mathcal{M}$.

**Output:** A distribution $p \in \mathcal{M}'$ (candidate for the MLE)

**While:** Some termination condition is not yet satisfied do

- **(E-Step)** Compute estimates for fully observed data

  \[ v_{ij} = u_i \cdot \frac{p_{ij}^{(t)}}{p_i^{(t)}} \]

- **(M-Step)** Compute $p^{(t+1)}$ the maximum likelihood estimate in $\mathcal{M}$ given the “data” $v \in \mathbb{R}^{r_1 \times r_2}$.

  **Return:** The marginal distribution of the final $p^{(t)}$.

We wish to show that the likelihood function of the hidden variable model does not decrease under application of a step of the EM-algorithm. To do this let

\[ \ell'(p) = \sum_{i=1}^{r_1} u_i \log p_i \]

be the log-likelihood function of the hidden variable model $\mathcal{M}'$, and let

\[ \ell(p) = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} v_{ij} \log p_{ij} \]

be the log-likelihood of the fully observed model $\mathcal{M}$.

**Theorem 14.3.2.** For each step of the EM algorithm $\ell'(p^{(t+1)}) \geq \ell'(p^{(t)})$. 
Proof. Let $p = p^{(t)}$ and $q = p^{(t+1)}$. Consider the difference $\ell'(q) - \ell'(p)$ which we want to show is nonnegative. We compute:

\[
\ell'(q) - \ell'(p) = \sum_{i=1}^{r_1} u_i (\log p_i - \log q_i)
\]

\[
= \sum_{i=1}^{r_1} u_i \log \frac{q_i}{p_i}
\]

\[
= \sum_{i=1}^{r_1} u_i \log \frac{q_i}{p_i} + \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} v_{ij} \log \frac{q_{ij}}{p_{ij}} - \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} v_{ij} \log \frac{q_{ij}}{p_{ij}}
\]

\[
= \sum_{i=1}^{r_1} u_i \left( \log \frac{q_i}{p_i} - \sum_{j=1}^{r_2} v_{ij} \log \frac{q_{ij}}{p_{ij}} \right) + \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} v_{ij} \log \frac{q_{ij}}{p_{ij}}
\]

Looking at this final expression, we see that the second term

\[
\sum_{i=1}^{r_1} \sum_{j=1}^{r_2} v_{ij} \log \frac{q_{ij}}{p_{ij}}
\]

is equal to $\ell(q) - \ell(p)$. Since $q$ was chosen to be the maximum likelihood estimate in the model $\mathcal{M}$ with data vector $v$, this expression is nonnegative. So to complete the proof we must show nonnegativity of the remaining expression. To this end, it suffices to show nonnegativity of

\[
\log \frac{q_i}{p_i} - \sum_{j=1}^{r_2} v_{ij} \log \frac{q_{ij}}{p_{ij}}
\]

for each $i$. Since $\sum_{j=1}^{r_2} v_{ij} = u_i$ we get that this expression equals

\[
\sum_{j=1}^{r_2} v_{ij} \log \frac{p_{ij}}{q_{ij}} = \sum_{j=1}^{r_2} v_{ij} \log \left( \frac{p_{ij}}{p_i} . \frac{q_{ij}}{q_i} \right)
\]

By the definition of $v_{ij}$ we have that $\frac{v_{ij}}{u_i} = \frac{p_{ij}}{p_i}$ to arrive at the expression

\[
\sum_{j=1}^{r_2} p_{ij} \log \frac{p_{ij}}{p_i} = \sum_{j=1}^{r_2} p_{ij} \left( -\log \frac{p_{ij}q_{ij}}{q_ip_i} \right)
\]

\[
\geq \sum_{j=1}^{r_2} p_{ij} \left( 1 - \frac{p_{ij}q_{ij}}{q_ip_i} \right)
\]

\[
= \sum_{j=1}^{r_2} \left( \frac{p_{ij}}{p_i} - \frac{q_{ij}}{q_i} \right)
\]

\[
= 0
\]
where we used the inequality $-\log x \geq 1 - x$. This shows that the log-likelihood function can only increase with iteration of the EM algorithm.

The behavior of the EM algorithm is not understood well, even in remarkably simple statistical models. It is known that if the maximum likelihood estimate lies in the relative interior of the model $\mathcal{M}$, then it is a fixed point of the EM algorithm. However, fixed points of the EM algorithm need not be critical points when those points lie on the boundary of the model, and the maximum likelihood estimate itself need not be a critical point of the likelihood function in the boundary case. Kubjas, Robeva, and Sturmfels \cite{KRS15} studied in detail the case of mixture of independence model $\text{Mixt}(\mathcal{M}_{\mathbb{X} \perp \mathbb{Y}})$, and also gave a description of general issues that might arise with the EM algorithm. Even in this simple mixture model, the EM algorithm might have fixed points that are not critical points of the likelihood function. There might be an infinite number of fixed points of the EM algorithm. The two steps of the EM-algorithm are quite simple for the model $\text{Mixt}(\mathcal{M}_{\mathbb{X} \perp \mathbb{Y}})$, so we write it out here explicitly to help focus ideas on this algorithm.

**Algorithm 14.3.3. (EM algorithm for $\text{Mixt}(\mathcal{M}_{\mathbb{X} \perp \mathbb{Y}})$)**

**Input:** Data $u \in \mathbb{N}^{r_1 \times r_2}$, an initial parameter estimates $\pi \in \Delta_{r-1}$, $A \in \mathbb{R}^{r \times r_1}$, $B \in \mathbb{R}^{r \times r_2}$

**Output:** A distribution $p \in \text{Mixt}(\mathcal{M}_{\mathbb{X} \perp \mathbb{Y}})'$ (candidate for the MLE)

**While:** Some termination condition is not yet satisfied do

- **(E-Step)** Compute estimates for fully observed data

$$v_{ijk} = u_{ijk} \cdot \frac{\pi^{(t)}_k a^{(t)}_{ki} b^{(t)}_{kj}}{\sum_k \pi^{(t)}_k a^{(t)}_{ki} b^{(t)}_{kj}}$$

- **(M-Step)**

$$\pi^{(t+1)}_k = \frac{\sum_{i,j} v_{ijk}}{\sum_{i,j,k} v_{ijk}}$$

$$a^{(t+1)}_{ki} = \frac{\sum_j v_{ijk}}{\sum_{j,k} v_{ijk}}$$

$$b^{(t+1)}_{kj} = \frac{\sum_i v_{ijk}}{\sum_{i,k} v_{ijk}}$$

**Return:** The final $\pi^{(t)}$, $A^{(t)}$, and $B^{(t)}$.

When working with models with continuous random variables, it is important to remember that in the E-step we do not compute the expected values of the missing data but the expected values of the missing sufficient statistics of the data. For continuous random variables this can involve doing complicated integrals.

**Example 14.3.4 (Censored Gaussians).** For $i = 1, 2, 3$, let $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, and let $T \sim \mathcal{N}(0, 1)$, and suppose that $X_1 \perp X_2 \perp X_3 \perp T$. For $i = 1, 2, 3$, let
$Y_i \in \{0, 1\}$ be defined by

$$Y_i = \begin{cases} 1 & \text{if } X_i < T \\ 0 & \text{if } X_i > T \end{cases}$$

Suppose that we only observe the discrete random variable $Y \in \{0, 1\}^3$. If we want to compute maximum likelihood estimates of the parameters $\mu_i, \sigma_i^2$ we can use the EM algorithm. In this case the M-step is trivial (sample mean and variance of a distribution), but the E-step requires a nontrivial computation. In particular, we need to compute the expected sufficient statistics of the Gaussian model namely

$$E_{\mu, \sigma}[X_i \mid Y] \quad \text{and} \quad E_{\mu, \sigma}[X_i^2 \mid Y].$$

Let

$$f(x_i \mid \mu_i, \sigma_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2}(x_i - \mu_i)^2\right)$$

be the density function of the univariate Gaussian distribution. Computing $E_{\mu, \sigma}[X_i \mid Y]$ and $E_{\mu, \sigma}[X_i^2 \mid Y]$ amounts to computing the following integrals:

$$E_{\mu, \sigma}[X_i \mid Y] = \frac{1}{P(Y)} \iiint_V x_i \prod_{j=1}^3 f(x_j \mid \mu_j, \sigma_j) \times f(t \mid 0, 1)dxdt$$

$$E_{\mu, \sigma}[X_i^2 \mid Y] = \frac{1}{P(Y)} \iiint_V x_i^2 \prod_{j=1}^3 f(x_j \mid \mu_j, \sigma_j) \times f(t \mid 0, 1)dxdt$$

where

$$P(Y) = \iiint_V \prod_{j=1}^3 f(x_j \mid \mu_j, \sigma_j) \times f(t \mid 0, 1)dxdt$$

and $V$ is the region

$$V = \{(x_1, x_2, x_3, t) \in \mathbb{R}^4 : x_i < t \text{ if } Y_i = 1, \text{ and } x_i > t \text{ if } Y_i = 0\}.$$

A detailed analysis of generalizations of these models appears in [Boc15], including their application to the analysis of certain mutation data. That work uses MonteCarlo methods to approximate the integrals. The integrals that appear here are integrals of Gaussian density functions over polyhedral regions and can also be computed using the holonomic gradient method [Koy16, KT15].

14.4. Exercises

**Exercise 14.1.** Show that $\text{Mix}^2(\mathcal{M}_{X_1, \perp X_2})$ is equal to the set of joint distributions $P \in \Delta_{r_1 r_2 - 1}$ for two discrete random variables such that rank $P \leq 2$. 
Exercise 14.2. Let $X_1, X_2, X_3$ be binary random variables. Prove that
\[ \text{Mixt}^4(\mathcal{M}_{X_1 \perp X_2 \perp X_3}) = \Delta_7. \]

Exercise 14.3. Show that a distribution in
\[ \text{Mixt}^k(\mathcal{M}_{X_1 \perp X_2 \perp \cdots \perp X_m}) \]
can have at most $k$ strong maxima.

Exercise 14.4. Estimate the proportion of distributions that are in $EX_{4,2}$ but are not mixtures of i.i.d. Bernoulli random variables.

Exercise 14.5. Consider the DAG $G$ on five vertices with edges $1 \rightarrow 2$, $1 \rightarrow 4$, $1 \rightarrow 5$, $2 \rightarrow 3$, $3 \rightarrow 4$, and $3 \rightarrow 5$. Let $X_1, \ldots, X_5$ be binary random variables. Compute the vanishing ideal of the hidden variable graphical model where $X_1$ is hidden. Is the corresponding variety the secant variety of the variety associated to some graphical model?

Exercise 14.6. Show that the factor analysis model $\mathcal{F}_{m,s}$ as introduced in Example 6.5.4 gives the same family of covariance matrices as the hidden variable Gaussian graphical model associated to the complete directed bipartite graph $K_{s,m}$ as described in Example 14.2.6.

Exercise 14.7. Consider the complete bidirected graph on 3 vertices, $K_3$. The linear structural equation model for this graph consists of all positive definite matrices $\mathcal{M}_{K_3} = PD_3$. Compare this to the set of covariance matrices that arise when looking at the hidden variable model on the bidirected subdivision. In particular, give an example of a positive definite matrix that could not arise from the hidden variable model on the bidirected subdivision.

Exercise 14.8. Show that if $p$ is in the relative interior of $\mathcal{M}'$ and $\ell'(p^{(t)}) = \ell'(p^{(t+1)})$ then $p^{(t)}$ is a critical point of the likelihood function. This shows that the termination criterion in the EM algorithm can be taken to run the algorithm until $\ell(p^{(t+1)}) - \ell(p^{(t)}) < \epsilon$ for some predetermined tolerance $\epsilon$.

Exercise 14.9. Implement the EM-algorithm for the mixture model over two independent discrete random variables $\text{Mixt'}(\mathcal{M}_{X_1 \perp X_2})$ and run it from 1000 random starting points with $r = 3$ on the data set
\[ u = \begin{pmatrix} 10 & 1 & 1 & 3 \\ 1 & 10 & 1 & 1 \\ 1 & 1 & 10 & 1 \\ 1 & 1 & 1 & 10 \end{pmatrix}. \]
Report on your findings.
Phylogenetic Models

Phylogenetics is the area of mathematical biology concerned with reconstructing the evolutionary history of collections of species. It has an old history, originating in the works of Linnaeus on classification of species and Darwin in the development of the theory of evolution.

The phylogeny of a collection of species is typically represented by a rooted tree. The tips (leaves) of the tree correspond to extant species (i.e. those that are current alive), and internal nodes in the tree represent most recent common ancestors of the extant species which lie below them. The major goal of phylogenetics is to reconstruct this evolutionary tree from data about species at the leaves.

Before the advent of modern sequencing technology, the primary methods for reconstructing phylogenetic trees were to look for morphological similarities between species and group species together that had similar characteristics. This, of course, also goes back to the Linnaean classification system: all species with a backbone are assumed to have a common ancestor, bird species should be grouped based on how many toes they have, etc. Unfortunately, a focus only on morphological features might group together organisms that have developed similar characteristics through different pathways (convergent evolution) or put apart species that are very closely related evolutionarily simply because they have a single disparate adaptation. Furthermore, collections of characteristics of different species might not all compatibly arise from a single tree, so combinatorial optimization style methods have been developed to reconstruct trees from such character data. Systematic development of such methods eventually led to the maximum parsimony approach to reconstruct phylogeny, which attempts
to find the tree on which the fewest mutations are necessary to observe the species with the given characteristics.

Since the 1970’s sequencing technology has been developed that can take a gene and reconstruct the sequence of DNA bases that make up that gene. We can also construct an amino acid sequence for the corresponding protein that the gene encodes. Modern methods for reconstructing phylogenetic trees are based on comparing these sequences for genes that appear in all species under consideration, using a phylogenetic model to model how these sequences might have descended through mutations from a common ancestor, and applying statistical procedures to recover the tree that best fits the data.

Phylogenetics is a vast field, with a wealth of papers, results, and algorithms. The methods of phylogenetics are applied throughout the biological sciences. We could not hope to cover all this material in the present book but refer the interested reader to the standard texts [Fel03, SS03b, Ste16]. Our goal here is to cover the basics of phylogenetic models and how these models can be analyzed using methods from algebraic statistics. We will return to discuss phylogenetics from a more polyhedral and combinatorial perspective in Chapter 19.

The algebraic study of phylogenetic models was initiated by computational biologists [CF87, Lak87] under the name phylogenetic invariants. In the algebraic language, the phylogenetic invariants of a model are just the generators of the vanishing ideal of that model. The determination of phylogenetic invariants for many of the most commonly used models is understood, but for more complex models, it remains open to determine them. We will discuss tools that make it easier to determine phylogenetic invariants in this chapter. In Chapter 16 we discuss applications of phylogenetic invariants to identifiability problems for phylogenetic models.

15.1. Trees and Splits

In this section we summarize some basic facts about trees that we will need throughout our discussion of phylogenetics. We typically treat the phylogenetic models on trees that we will study as directed graphical models with all arrows directed away from a given root vertex, but the unrooted perspective plays an important role because of identifiability issues.

A tree $T = (V, E)$ is a graph that is connected and does not have any cycles. These two conditions imply that the number of edges in a tree with $m$ vertices is $m - 1$. A leaf is a vertex of degree 1 in $T$. Because phylogenetics is usually concerned with constructing trees based on data associated only with extant species, which are associated with leaves of the tree, it is useful to discuss combinatorial features of trees with only the leaves labeled.
15.1. Trees and Splits

Definition 15.1.1. Let $X$ be a set of labels and let $\mathcal{T} = (V, E)$ be a tree. Let $\phi : X \rightarrow V$ be a map from the label set to the vertices of $\mathcal{T}$.

- The pair $T = (\mathcal{T}, \phi)$ is called an $X$-tree if each vertex of degree one and two in $\mathcal{T}$ is in the image of $\phi$.
- The pair $T = (\mathcal{T}, \phi)$ is called a phylogenetic $X$-tree if the image of $\phi$ is exactly the set of leaves of $\mathcal{T}$ and $\phi$ is injective.
- A binary phylogenetic $X$-tree is a phylogenetic $X$-tree where each non-leaf vertex has degree 3.

We will usually just write that $T$ is a phylogenetic $X$-tree (or just phylogenetic tree if $X$ is clear) without referring to the map $\phi$. In a rooted tree we designate a prescribed vertex as the root, which is also required to have a special label $\phi$. In a binary rooted phylogenetic tree, the root vertex has degree 2 and all other non-leaf vertices have degree 3. Although our rooted trees could be considered as undirected trees, we will typically draw them as directed with all edges directed away from the root node. This will be useful when we consider phylogenetic models on these trees, as they are naturally related to DAG models from Chapter 13. An example of a $[6]$-tree and a rooted binary phylogenetic $[6]$-tree appear in Figure 15.1.1.

In addition to the standard representation of a tree as a vertex set and list of edges, it can be useful to also encode the tree by a list of splits, which are bipartitions of the label set $X$ induced by edges.

Definition 15.1.2. A split $A|B$ of $X$ is a partition of $X$ into two disjoint nonempty sets. The split $A|B$ is valid for the $X$-tree $T$ if it is obtained from $T$ by removing some edge $e$ of $\mathcal{T}$ and taking $A|B$ to be the two sets of labels of the two connected components of $\mathcal{T} \setminus e$. The set of all valid splits for the $X$-tree $T$ is $\Sigma(T)$. 

Figure 15.1.1. A $[6]$-tree (left), and a rooted binary phylogenetic $[6]$-tree (right)
Example 15.1.3 (Splits in an $X$-tree). Let $X = [6]$ and consider the $X$-tree $T$ on the left in Figure 15.1.1. The associated set of valid splits $\Sigma(T)$ is

$$\Sigma(T) = \{1|23456, 14|2356, 26|1345, 3|12456, 5|12346\}.$$  

For the rooted binary phylogenetic $[6]$-tree on the right in Figure 15.1.1, the associated set of valid splits (including the root) are

$$\Sigma(T) = \{i|\{1, 2, 3, 4, 5, 6, \rho\} \setminus \{i\} : i = 1, \ldots, 6\} \cup \{16|2345\rho, 23|145\rho, 1236|45\rho, 12346|5\rho\}.$$  

Not every set of splits could come from an $X$-tree. For example, on four leaves, there is no tree that has both of the splits $12|34$ and $13|24$. This notion is encapsulated in the definition of compatible splits.

Definition 15.1.4. A pair of splits $A_1|B_1$ and $A_2|B_2$ is called compatible if at least one of the sets

$$A_1 \cap A_2, \quad A_1 \cap B_2, \quad B_1 \cap A_2, \quad B_1 \cap B_2$$

is empty. A set $\Sigma$ of splits is pairwise compatible if every pair of splits $A_1|B_1, A_2|B_2 \in \Sigma$ is compatible.

Proposition 15.1.5. If $T$ is an $X$-tree then $\Sigma(T)$ is a pairwise compatible set of splits.

Proof. The fact that removing an edge of $T$ disconnects it, guarantees that at least one pair of intersections in Equation 15.1.2 is empty.

The splits equivalence theorem says that the converse to Proposition 15.1.5 is also true:

Theorem 15.1.6 (Splits Equivalence Theorem). Let $\Sigma$ be a set of $X$-splits that is pairwise compatible. Then there exists a unique $X$-tree $T$ such that $\Sigma = \Sigma(T)$.

Proof Sketch. The idea of the proof is to “grow” the tree by adding one split at a time, using induction on the number of splits in a pairwise compatible set. Suppose that $T'$ is the unique tree we have constructed from a set of pairwise compatible splits, and we want to add one more split $A|B$. The pairwise compatibility of $A|B$ with the splits in $T'$ identifies a unique vertex in $T'$ that can be grown out to produce an edge in a tree $T$ that has all the given splits. This process is illustrated in Figure 15.1.2 with the set of splits in (15.1.1), adding the splits in order from left to right.

The splits equivalence theorem allows us to uniquely associate a tree with its set of edges (the splits). This fact can be useful in terms of descriptions of phylogenetic models associated to the trees. For instance, it will
Figure 15.1.2. The tree growing procedure in the proof of the Splits Equivalence Theorem

be used in the parameterizations of group-based models, in the descriptions of vanishing ideals of phylogenetic models, and in proofs of identifiability of phylogenetic models.

15.2. Types of Phylogenetic Models

We describe in this section the general setup of phylogenetic models broadly speaking. We will explain how these models are related to the latent variable DAG models of previous chapters. However, in most applications, the phylogenetic models are not equal to a graphical model, but is merely a submodel of a graphical model, because of certain structures that are imposed on the conditional distributions. We will illustrate what types of assumptions are made in these models, to illustrate the imposed structures. Subsequent sections will focus on algebraic aspects of these models.

Phylogenetic models are used to describe the evolutionary history of sequence data over time. Let $T$ be a rooted tree, which we assume is a directed graph with all edges pointing away from the root vertex. To each vertex $i$ in the tree associate a sequence $S_i$ on some alphabet. For simplicity the alphabet is assumed to be the set $\{\alpha\}$ for some integer $\alpha$. For biological applications, the most important value of $\alpha$ is $\alpha = 4$, where the alphabet consists of the four DNA bases $\{A, C, G, T\}$ and then the sequence $S_i$ is a fragment of DNA, usually corresponding to a gene, and the node corresponds to a species that has that gene. The value $\alpha = 20$ is also important, where the alphabet consists of the 20 amino acids, and so the sequences $S_i$ are protein sequences. Other important values include $\alpha = 61$, corresponding to the 61 codons (DNA triplets that code for amino acids, excluding three stop codons) and $\alpha = 2$, still corresponding to DNA sequences but where purines $(A, G)$ and pyrimidines $(C, T)$ are grouped together as one element.

The sequences $S_i$ are assumed to be related to each other by common descent, i.e. they all code for a gene with the same underlying function in the species corresponding to each node. So the sequence corresponding to the root node is the most ancestral sequence, and all the other sequences
in the tree have descended from it via various changes in the sequence. We typically do not have access to the sequence data of ancestral species (since those species are long extinct), and so the sequence data associated to internal nodes in the tree are treated as hidden variables.

In the most basic phylogenetic models that are used, the only changes that happen to the sequences are point mutations. That is, at random points in time, a change to a single position in a sequence occurs. This means that all the sequences will have the same length. Furthermore, it is assumed that the point mutations happen independently at each position in the sequence. So our observed data under these assumptions looks something like this:

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>ACCGTGCAACGTGACGA</td>
</tr>
<tr>
<td>Chimp</td>
<td>ACCCTTGGAAAGTAAACGA</td>
</tr>
<tr>
<td>Gorilla</td>
<td>ACCGTGCAACGTAAACTA</td>
</tr>
</tbody>
</table>

These are the DNA sequences for sequences at the leaves of some tree explaining the evolutionary history of these three genes. Such a diagram is called an alignment of the DNA sequences, since at some point a decision needs to be made about what positions in one sequence correspond to positions that are related by common descent in another sequence. Note that with the modeling assumptions we take in this chapter, the determination of an alignment is trivial, but the more general problem of aligning sequences (which could have gaps of unknown positions and lengths, or could be rearranged in some way) is a challenging one (see [DEKM98]). The goal of phylogenetics is to figure out which (rooted) tree produced this aligned sequence data. Each tree yields a different family of probability distributions, and we would like to find the tree that best explains the observed data.

The assumption of site independence means that each column of the alignment has the same underlying probability distribution. For example, the probability of observing the triple AAA is the same at any position in the alignment. Hence, with the assumptions of only point mutations and site independence, to describe a phylogenetic model, we need to use the tree and model parameters to determine the probability of observing each column in the alignment. For example, with 3 observed sequences, and \( \kappa = 4 \), we need to write down formulas for each of the 64 probabilities

\[ p_{AAA}, p_{AAC}, p_{AAG}, \ldots, p_{TTT}. \]

For general \( \kappa \) and number of taxa at the leaves, \( m \), we need a probability distribution for each of the \( \kappa^m \) different possible columns in the alignment.

Consider the hidden variable graphical model associated to the rooted phylogenetic tree \( T \). We have a random variable associated to each vertex that has \( \kappa \) states and all random variables associated to internal vertices (that is, vertices that are not tips in the tree) are hidden random variables.
According to our discussion above, this gives a phylogenetic model because it assigns probabilities to each of the $\kappa^m$ possible columns in the alignment of the $m$ sequences. The resulting phylogenetic model that arises is called the \textit{general Markov model}. It has interesting connections to mathematics and has received wide study, though it is not commonly used in phylogenetics practice because its large number of parameters make it computationally unwieldy. We will study this model in detail in Section 15.4.

\textbf{Example 15.2.1 (Three Leaf Tree).} Consider the three leaf tree in Figure 15.2.1. Here root is the node labeled $Y_1$. The variables $Y_1$ and $Y_2$ are hidden random variables, whereas the random variables associated to the leaf nodes $X_1, X_2, X_3$ are observed random variables. In the DAG model associated to this tree, the parameters are the marginal probability distribution of $Y_1$, and the conditional probability distributions of $Y_2$ given $Y_1$, $X_1$ given $Y_1$, $X_2$ given $Y_2$, and $X_3$ given $Y_3$. The probability of observing $X_1 = x_1, X_2 = x_2, X_3 = x_3$ at the leaves has the form

\[
P(x_1, x_2, x_3) = \sum_{y_1 \in \mathcal{Y}} \sum_{y_2 \in \mathcal{Y}} P(Y_1 = y_1) P(Y_2 | Y_1 = y_1) P(X_1 | Y_1 = y_1) (x_1 | y_1) P(X_2 | Y_2 = y_2) P(X_3 | Y_2 = y_2)
\]

where we used shorthand like

\[
P(Y_2 | Y_1 = y_1) = P(Y_2 = y_2 | Y_1 = y_1).
\]

The structure of these conditional distribution matrices associated to each edge in the tree are important for the analysis of the phylogenetic models. Point substitutions are happening instantaneously across the sequence and across the tree. The conditional distribution like $P(Y_2 = y_2 | Y_1 = y_1)$ along a single edge summarizes (or integrates out) this continuous time process.

Since $T$ is a rooted tree and all edges are directed away from the root, each vertex in $T$ except the root has exactly one parent. This means that the conditional probability distributions that appear in the directed graphical model $P(X_1 = x_i | X_{pa(i)} = x_{pa(i)})$ are square matrices, and can be thought
of as transition matrices of Markov chains. (This explains the name “general Markov model”, above). Most phylogenetic models that are used in practice are obtained as submodels of the general Markov model by placing restrictions on what types of transition matrices are used.

The starting point for discussing the general structure of the Markov models that are used in practice is to start with some basic theory of Markov chains. Let $Q \in \mathbb{R}^{n \times n}$ be a matrix such that $q_{ij} \geq 0$ for all $i \neq j$ and such that for each $i$, $\sum_{j=1}^{n} q_{ij} = 0$. The resulting matrix describes a continuous time Markov chain. Roughly speaking the probability that our random variable jumps from state $i$ to state $j$ in infinitesimal time $dt$ is $q_{ij}dt$. The matrix $Q$ is called the rate matrix of the continuous time Markov chain.

To calculate the probability that after $t$ time units our random variable ends up in state $j$ given that it started at $t = 0$ in state $i$ we can solve a system of linear differential equations to get the conditional probability matrix $P(t) = \exp(Qt)$, where

$$\exp(Qt) = I + Q t + \frac{Q^2 t^2}{2!} + \frac{Q^3 t^3}{3!} + \cdots$$

denotes the matrix exponential. See a book on Markov chains for more details (e.g. [Nor98]).

The various phylogenetic models arise by requiring that all the transition matrices are of the form $\exp(Qt)$ for some rate matrix $Q$, and then further specifying what types of rate matrices are allowed. The parameter $t$ is called the branch length and can be thought of as the amount of time that has passed between major speciation events (denoted by vertices in the tree). This time parameter is an important one from the biological standpoint because it tells how far back in time the most recent common ancestors occurred. Each edge $e$ in the tree typically has its own branch length parameter $t_e$. Note that even if we allowed each edge $e$ of the tree to have a different, and completely arbitrary, rate matrix $Q_e$, the resulting model that arises is smaller than the general Markov model, since a conditional probability distribution of the form $\exp(Qt)$ has a restricted form.

**Example 15.2.2** (A Non-Markovian Transition Matrix). The transition matrix

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

could not arise as $P = \exp(Qt)$ for any rate matrix $Q$ or any time $t$. Indeed, from the identity

$$\det \exp(A) = \exp(\text{tr} A)$$
we see that \( \det \exp(Qt) \) is positive for any rate matrix \( Q \) and any \( t \in [0, \infty) \). On the other hand, \( \det P = -1 \), in the preceding example.

Phylogenetic models are usually specified by describing the structure of the rate matrix \( Q \). Here are some common examples.

The simplest model for 2 state Markov chain is the Cavender-Farris-Neyman model (CFN) \cite{Cav78, Far73, Ney71} which has rate matrices of the form:

\[
Q_{\text{CFN}} = \begin{pmatrix}
-\alpha & \alpha \\
\alpha & -\alpha
\end{pmatrix}
\]

for some positive parameter \( \alpha \). This model is also called the binary symmetric model since the transition matrix

\[
P(t) = \exp(Q_{\text{CFN}}t) = \begin{pmatrix}
\frac{1+\exp(-2\alpha t)}{2} & \frac{1-\exp(-2\alpha t)}{2} \\
\frac{1-\exp(-2\alpha t)}{2} & \frac{1+\exp(-2\alpha t)}{2}
\end{pmatrix}
\]

is symmetric. The CFN model is the only significant model for \( \kappa = 2 \) besides the general Markov model.

From here on we focus on models for the mutation of DNA-bases. We order the rows and columns of the transition matrix \( A, C, G, T \). The simplest model for DNA bases is the Jukes-Cantor model (JC69) \cite{JC69}.

\[
Q_{\text{JC}} = \begin{pmatrix}
-3\alpha & \alpha & \alpha & \alpha \\
\alpha & -3\alpha & \alpha & \alpha \\
\alpha & \alpha & -3\alpha & \alpha \\
\alpha & \alpha & \alpha & -3\alpha
\end{pmatrix}
\]

Note that JC69 model, while widely used, is not considered biologically reasonable. This is because it gives equal probability to each of the mutations away from a given DNA base. This is considered unrealistic in practice because the mutations are often more likely to occur within the purine group \( \{A, G\} \) and within the pyrimidine group \( \{C, T\} \), than from purine to pyrimidine or vice versa. In phylogenetics, a within group mutation is called a transition and mutation across groups is called a transversion. Kimura \cite{Kim80} tried to correct this weakness of the Jukes-Cantor model by explicitly adding separate transversion rates. These are summarized in the Kimura 2-parameter and Kimura 3-parameter models:

\[
Q_{\text{K}2P} = \begin{pmatrix}
-\alpha - 2\beta & \beta & \alpha & \beta \\
\beta & -\alpha - 2\beta & \beta & \alpha \\
\alpha & \beta & -\alpha - 2\beta & \beta \\
\beta & \alpha & \beta & -\alpha - 2\beta
\end{pmatrix}
\]
338

15. Phylogenetic Models

\[ Q^{K3P} = \begin{pmatrix}
-\alpha - \beta - \gamma & \beta & \alpha & \gamma \\
\beta & -\alpha - \beta - \gamma & \gamma & \alpha \\
\alpha & \gamma & -\alpha - \beta - \gamma & \beta \\
\gamma & \alpha & \beta & -\alpha - \beta - \gamma \\
\end{pmatrix}. \]

The CFN, JC69, K2P, and K3P models all have a beautiful mathematical structure which we will analyze in detail in Section 15.3.

Before we go on to introduce more complex transition structures, we mention an important feature of the model which is important to consider as one develop more complex transition structures. This key feature of a Markov chain is its stationary distribution, which is defined below. Let \( P \) be the transition matrix of a Markov chain. Form a graph \( G(P) \) with vertex set \([\mathbb{V}]\), and with an edge \( i \to j \) if \( p_{ij} > 0 \) (including loops \( i \to i \) if \( p_{ii} > 0 \)). The Markov chain is said to be irreducible if between any ordered pair of vertices \((i, j)\) there is directed path from \( i \) to \( j \) in \( G(P)\). The Markov chain is aperiodic the greatest common divisor of cycle lengths in the graph is 1.

**Theorem 15.2.3** (Perron-Frobenius Theorem). Let \( Q \) be a rate matrix such that \( P(t) = \exp(Qt) \) describes an irreducible, aperiodic Markov chain. Then there is a unique probability distribution \( \pi \in \Delta_{\mathbb{V} - 1} \) such that \( \pi Q = 0 \), or equivalently \( \pi \exp(Qt) = \pi \). This distribution is called the stationary distribution of the Markov chain.

For each of the CFN, JC69, K2P, and K3P models described above, the stationary distribution is the uniform distribution, that is \( \pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}) \) for the CFN model and \( \pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}) \) for the other models. For aperiodic irreducible Markov chains, the stationary distribution also has the property that if \( \mu \) is any probability distribution on \([\mathbb{V}]\) then \( \mu P^t \) converges to \( \pi \) as \( t \to \infty \). This is part of the reason \( \pi \) is called the stationary distribution.

Considering the Markov chain associated to a rate matrix \( Q \) arising from a phylogenetic model, if one believes that the Juke-Cantor, or Kimura models truly describe the long-term behavior of mutational processes in DNA sequences, one would expect to see a relatively uniform distribution of bases in DNA, with one fourth of bases being, e.g. \( A \). However, this is generally not what is observed in practice as \( C \)'s and \( G \)'s are generally rarer than \( A \)'s and \( T \)'s. Various models have been constructed which attempt to allow for other stationary distributions to arise, besides the uniform distributions that are inherent in the CFN, JC69, K2P, and K3P models.

The Felsenstein model (F81) \cite{Fel81} and the Hasegawa-Kishino-Yano (HKY85) \cite{HKaY85} model are two simple models that take into account these structures. Their rate matrices have the following forms where the \( * \) in each row denotes the appropriate number so that the row sums are all
15.2. Types of Phylogenetic Models

zero:

\[
Q^{FS} = \begin{pmatrix}
\pi_C & \pi_G & \pi_T \\
\pi_A & \pi_G & \pi_T \\
\pi_A & \pi_C & \pi_T \\
\pi_A & \pi_C & \pi_G
\end{pmatrix}
\]

\[
Q^{HKY} = \begin{pmatrix}
\pi_C & \alpha \pi_G & \pi_T \\
\pi_A & \alpha \pi_C & \pi_T \\
\alpha \pi_A & \pi_C & \pi_T \\
\pi_A & \alpha \pi_C & \pi_G
\end{pmatrix}
\]

Note that the addition of the parameter \( \alpha \) to the HKY model allows to increase the probability for transitions versus transversions (so \( \alpha \geq 1 \)).

A further desirable property of the Markov chains that arise in phylogenetics is that they are often time reversible. Intuitively, this means that random point substitutions looking forward in time should look no different from random point substitutions looking backwards in time. This is a reasonable assumption on shorter time-scales for which we could discuss comparing two DNA sequences to say they evolved from a common source. Time reversibility is a mathematically convenient modeling assumption that is present in many of the commonly used mutation models. Time reversibility amounts to the following condition on the rate matrix

\[ \pi_i Q_{ij} = \pi_j Q_{ji} \]

for all pairs \( i, j \). All the models we have encountered so far, except the general Markov model, are time reversible. The general time reversible model (GTR) is the model of all rate matrices that satisfy the time reversibility assumption. These rate matrices have the following general form:

\[
Q^{GTR} = \begin{pmatrix}
* & \pi_C \alpha & \pi_G \beta & \pi_T \gamma \\
\pi_A \alpha & * & \pi_G \delta & \pi_T \epsilon \\
\pi_A \beta & \pi_C \delta & * & \pi_T \zeta \\
\pi_A \gamma & \pi_C \epsilon & \pi_G \zeta & *
\end{pmatrix}
\]

where the * in each row is chosen so that the row sums are zero.

These Markov mutation models are just a fraction of the many varied models that might be used in practice. The most widely used models are the ones that are implemented in phylogenetic software, probably the most commonly used model is the GTR model.

Up to this point we have added conditions to our models based on increasing assumptions of biological realism. However, there are also mathematical considerations which can be made which might be useful from a theoretical standpoint, or lead us to rethink their usefulness as biological tools. We conclude this section by discussing one such practical mathematical consideration now.

It is often assumed that the rate matrix \( Q \) is fixed throughout the entire evolutionary tree. This is mostly for practical reasons since this assumption greatly reduces the number of parameters in the model. It might be a
reasonable assumption for very closely related species but is probably unreasonable across large evolutionary distances. So let us assume throughout the remainder of this section that we use different rate matrices along each edge of the tree.

Note that if we construct a phylogenetic tree that has a vertex of degree two we do not really get any information about this vertex as it is a hidden node in the tree. If \( \exp(Q_1 t_1) \) is the transition matrix associated with the first edge and \( \exp(Q_2 t_2) \) is the transition matrix associated to the second edge the new transition matrix we get by suppressing the vertex of degree two will be the matrix product \( \exp(Q_1 t_1) \exp(Q_2 t_2) \). The mathematical property we would like this matrix to satisfy for a model to be reasonable is that this resulting matrix also has the form \( \exp(Q_3 (t_1 + t_2)) \) for some \( Q_3 \) in our model class. It seems difficult to characterize which sets of rate matrices are closed in this way. One useful way to guarantee this property is handled by the theory of Lie Algebras.

**Definition 15.2.4.** A set \( L \subseteq \mathbb{K}^{n \times n} \) is a matrix Lie algebra if \( L \) is a \( \mathbb{K} \)-vector space and for all \( A, B \in L \)

\[
[A, B] := AB - BA \in L.
\]

We say a set of rate matrices \( L \) is a Lie Markov model if it is a Lie algebra.

The quantity \( [A, B] = AB - BA \) is called the Lie bracket or commutator of the matrices \( A \) and \( B \).

**Theorem 15.2.5. [SFSJ12a]** Let \( L \subseteq \mathbb{R}^{n \times n} \) be a collection of rate matrices. If \( L \) is a Lie Markov model then for any \( Q_1, Q_2 \in L \) and \( t_1, t_2 \), there is a \( Q_3 \in L \) such that \( \exp(Q_1 t_1) \exp(Q_2 t_2) = \exp(Q_3 (t_1 + t_2)) \).

**Proof.** The Baker-Campbell-Hausdorff [Cam97] identity for the product of two matrix exponentials is the following:

\[
\exp(X) \exp(Y) = \exp\left(X + Y + \frac{1}{2}[X, Y] + \frac{1}{12}[X, [X, Y]] + \cdots\right).
\]

The expression on the left involves infinitely many terms but is a convergent series. In particular, this show that if \( X, Y \in L \), then \( \exp(X) \exp(Y) = \exp(Z) \) where \( Z \in L \). \( \square \)

The group-based phylogenetic models (CFN, JC69, K2P, K3P) all have the property that the matrices in the models are simultaneously diagonalizable (this will be discussed in more detail in Section 15.3). This in particular, implies that any pair of rate matrices \( Q_1, Q_2 \) in one of these models commute with each other, and hence, the commutator \( [Q_1, Q_2] = 0 \). Thus, the group-based models (CFN, JC69, K2P, K3P) give abelian Lie algebras, and so are Lie Markov models. The general Markov model includes transition
matrices that are not of the form \( \exp(Qt) \) so is not really a Lie Markov model. However, if we restrict the general Markov model to only matrices of the form \( \exp(Qt) \), then this general rate matrix model is also a Lie Markov model. The F81 model is also a Lie Markov model, though it is not abelian.

On the other hand, the HKY, and GTR models are not Lie Markov models, as can be verified by directly calculating the commutators between a pair of elements. If one works with models with heterogeneous rate matrices across the tree, this might present difficulties for using these models in practice \([\text{SFSJ}+12b]\).

### 15.3. Group-based Phylogenetic Models

In this section, we study the group-based phylogenetic models. The mostly commonly used group-based models in phylogenetics are the CFN, JC, K2P, and K3P models, though, in theory, the idea of a group-based model holds much more generally. The rate matrices of these models (and hence the transition matrices) all have the remarkable property that they can be simultaneously diagonalized (without regard to what particular rate matrix is used on a particular edge in a tree). In the resulting new coordinate system, the models correspond to toric varieties and it is considerably easier to describe their vanishing ideals.

To explain the construction of explicit generating sets of the vanishing ideals of group-based models we need some language from group theory. Let \( G \) be a finite abelian group, which we write with an additive group operation. In all the applications we consider either \( G = \mathbb{Z}_2 \) or \( G = \mathbb{Z}_2 \times \mathbb{Z}_2 \), though any finite abelian group would work for describing a group-based model. We replace the state space \([\kappa]\) with the finite abelian group \( G \) (whose cardinality equals \( \kappa \)). Let \( M = \exp(Qt) \) be the transition matrix along an edge of the tree \( T \).

**Definition 15.3.1.** A phylogenetic model is called a group-based model if for all transition matrices \( M = \exp(Qt) \) that arise from the model, there exists a function \( f : G \to \mathbb{R} \) such that \( M(g, h) = f(g - h) \).

Here \( M(g, h) \) denotes the \((g, h)\) entry of \( M \), the probability of going to \( h \) given that we are in state \( g \) at the previous position. Explicit calculation of the matrix exponentials shows that the CFN, JC, K2P, K3P models are all group-based. For example, in the CFN model:

\[
Q = \begin{pmatrix}
-\alpha & \alpha \\
\alpha & -\alpha
\end{pmatrix}
\quad
\exp(Qt) = \begin{pmatrix}
\frac{1}{2}(1 + e^{-2\alpha t}) & \frac{1}{2}(1 - e^{-2\alpha t}) \\
\frac{1}{2}(1 - e^{-2\alpha t}) & \frac{1}{2}(1 + e^{-2\alpha t})
\end{pmatrix}.
\]
So the function \( f : \mathbb{Z}_2 \to \mathbb{R} \) has \( f(0) = \frac{1}{2}(1 + e^{-2\alpha t}) \) and \( f(1) = \frac{1}{2}(1 - e^{-2\alpha t}) \).

In the K3P model we have

\[
Q = \begin{pmatrix}
-\alpha - \beta - \gamma & \beta & \alpha & \gamma \\
\beta & -\alpha - \beta - \gamma & \gamma & \alpha \\
\alpha & \gamma & -\alpha - \beta - \gamma & \beta \\
\gamma & \alpha & \beta & -\alpha - \beta - \gamma
\end{pmatrix}
\]

and

\[
\exp(Qt) = \begin{pmatrix}
a & b & c & d \\
b & a & d & c \\
c & d & a & b \\
d & c & b & a
\end{pmatrix}
\]

where

\[
\begin{aligned}
a &= \frac{1}{4} \left(1 + e^{-2(\alpha + \beta)t} + e^{-2(\alpha + \gamma)t} + e^{-2(\beta + \gamma)t}\right) \\
b &= \frac{1}{4} \left(1 - e^{-2(\alpha + \beta)t} - e^{-2(\alpha + \gamma)t} + e^{-2(\beta + \gamma)t}\right) \\
c &= \frac{1}{4} \left(1 - e^{-2(\alpha + \beta)t} + e^{-2(\alpha + \gamma)t} - e^{-2(\beta + \gamma)t}\right) \\
d &= \frac{1}{4} \left(1 + e^{-2(\alpha + \beta)t} - e^{-2(\alpha + \gamma)t} - e^{-2(\beta + \gamma)t}\right).
\end{aligned}
\]

So if we make the identification of the nucleotides \( \{A, C, G, T\} \) with elements of the group \( \mathbb{Z}_2 \times \mathbb{Z}_2 = \{(0,0), (0,1), (1,0), (1,1)\} \) then we see that the K3P model is a group-based model, and the corresponding function \( f \) has \( f(0,0) = a, f(0,1) = b, f(1,0) = c, \) and \( f(1,1) = d \). The JC and K2P models can be seen to be group-based models with respect to \( \mathbb{Z}_2 \times \mathbb{Z}_2 \) as they arise as submodels of K3P.

The presence of expressions like \( f(g - h) \) in the formulas for the conditional probabilities suggest that the joint distributions in these models are convolutions of a number of functions. This suggests that applying the discrete Fourier transform might simplify the representation of these models. We now explain the general setup of the discrete Fourier transform which we then use to analyze the group-based models.

**Definition 15.3.2.** A one dimensional character of a group \( G \) is a group homomorphism \( \chi : G \to \mathbb{C}^* \).

**Proposition 15.3.3.** The set of one dimensional characters of a group \( G \) is a group under the coordinate-wise product: \((\chi_1 \cdot \chi_2)(g) = \chi_1(g) \cdot \chi_2(g)\). It is called the dual group of \( G \) and is denoted \( \hat{G} \). If \( G \) is a finite abelian group then \( \hat{G} \cong \hat{G} \).
15.3. Group-based Phylogenetic Models

Proof. The conditions to check that \( \hat{G} \) is a group are straightforward. For example, since \( C^* \) is abelian, we have that

\[
(\chi_1 \cdot \chi_2)(g_1 g_2) = \chi_1(g_1 g_2) \cdot \chi_2(g_1) = \chi_1(g_1) \cdot \chi_2(g_1 g_2) = \chi_1(g_1) \cdot (\chi_2(g_1) \cdot \chi_2(g_2))
\]

which shows that \((\chi_1 \cdot \chi_2)\) is a group homomorphism, so the product operation in \( \hat{G} \) is well-defined.

It is also straightforward to see that \( \hat{G} \times H \rightarrow \hat{G} \times \hat{H} \). If \( G \) is a cyclic group \( \mathbb{Z}_n \), then each group homomorphism sends the generator of \( \mathbb{Z}_n \) to a unique \( n \)-th root of unity. The group of \( n \)-th roots of unity is isomorphic to \( \mathbb{Z}_n \).

The most important group for group-based models arising in phylogenetics is \( G = \mathbb{Z}_2 \times \mathbb{Z}_2 \).

Example 15.3.4 (Character Table of the Klein 4 Group). Consider the group \( \mathbb{Z}_2 \times \mathbb{Z}_2 \). The four characters of \( \mathbb{Z}_2 \times \mathbb{Z}_2 \) are the rows in the following character table:

<table>
<thead>
<tr>
<th>(\chi_1)</th>
<th>(\chi_2)</th>
<th>(\chi_3)</th>
<th>(\chi_4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( (0,0) )</td>
<td>( (0,1) )</td>
<td>( (1,0) )</td>
<td>( (1,1) )</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>

Note that there is no canonical isomorphism between \( G \) and \( \hat{G} \) in this case. Some choice must be made in establishing the isomorphism.

Definition 15.3.5. Let \( G \) be a finite abelian group, and \( f : G \rightarrow \mathbb{C} \) a function. The discrete Fourier transform of \( f \) is the function \( \hat{f} : \hat{G} \rightarrow \mathbb{C} \) defined by

\[
\hat{f}(\chi) = \sum_{g \in G} f(g) \chi(g).
\]

The discrete Fourier transform is a linear change of coordinates. If we think about \( f \) as a vector in \( \mathbb{C}^G \) then \( \hat{f} \) is a vector in \( \mathbb{C}^{\hat{G}} \) and two vectors are related by multiplication by the matrix whose rows are the characters of the group \( G \). This linear transformation can be easily inverted via the inverse Fourier transform stated as the following theorem.

Theorem 15.3.6. Let \( G \) be a finite abelian group and \( f : G \rightarrow \mathbb{C} \) a function. Then

\[
f(g) = \frac{1}{|G|} \sum_{\chi \in G} \hat{f}(\chi) \overline{\chi(g)}
\]

where the bar denotes complex conjugation.
Proof. The rows of the character table are (complex) orthogonal to each other and each has norm $|G|$.

A key feature of the discrete Fourier transform is that, in spite of being a matrix-vector multiplication, which seems like it might require $O(|G|^2)$ operations to calculate, the group symmetries involved imply that it can be calculated rapidly, in $O(|G| \log |G|)$ time, when $|G| = 2^k$ [CT65]. This is the fast Fourier transform (though we will not need this here). A final piece we will need is the convolution of two functions and its relation to the discrete Fourier transform.

**Definition 15.3.7.** Let $f_1, f_2 : G \to \mathbb{C}$ be two functions on a finite abelian group. Their **convolution** is the function

$$(f_1 \ast f_2)(g) = \sum_{h \in G} f_1(h) f_2(g - h).$$

**Proposition 15.3.8.** Let $f_1, f_2 : G \to \mathbb{C}$ be two functions on a finite abelian group. Then

$$\hat{f_1 \ast f_2} = \hat{f}_1 \cdot \hat{f}_2.$$ i.e. the Fourier transform of a convolution is the product of the Fourier transforms.

**Proof.** This follows by a directed calculation:

$$(\hat{f_1 \ast f_2})(\chi) = \sum_{g \in G} \sum_{h \in G} f_1(h) f_2(g - h) \chi(g)$$

$$= \sum_{g' \in G} \sum_{h \in G} f_1(h) f_2(g') \chi(g' + h)$$

$$= \sum_{g' \in G} \sum_{h \in G} f_1(h) f_2(g') \chi(g') \chi(h)$$

$$= \left( \sum_{h \in G} f_1(h) \chi(h) \right) \cdot \left( \sum_{g' \in G} f_2(g') \chi(g') \right)$$

$$= \hat{f}_1(\chi) \cdot \hat{f}_2(\chi).$$

Note that in the second line we made the replacement $g' = g - h$ to simplify the calculation.

With these preliminaries out of the way, we are ready to apply the discrete Fourier transform to our group-based model. We first explain the structure in an example on a small tree, then explain how this applies for general group-based models on trees.
Example 15.3.9 (Fourier Transform on Two Leaf Tree). Consider a general group-based model on a tree with only two leaves. This model has a root distribution \( \pi \in \Delta_{k-1} \), and transition matrices \( M_1 \) and \( M_2 \) pointing from the root to leaf 1 and 2 respectively. The probability of observing the pair \((g_1, g_2)\) at the leaves is

\[
p(g_1, g_2) = \sum_{h \in G} \pi(h)M_1(h, g_1)M_2(h, g_2)
\]

\[
= \sum_{h \in G} \pi(h)f_1(g_1 - h)f_2(g_2 - h)
\]

where \( f_i(g - h) = M_i(g, h) \) which is guaranteed to exist since we are working with a group-based model. Note that it is useful to think about the root distribution \( \pi \) as a function \( \pi : G \rightarrow \mathbb{R} \) in these calculations. Now we compute the discrete Fourier transform over the group \( G^2 \).

\[
\hat{p}(\chi_1, \chi_2) = \sum_{(g_1, g_2) \in G^2} \sum_{h \in G} \pi(h)f_1(g_1 - h)f_2(g_2 - h)\chi_1(g_1)\chi_2(g_2)
\]

\[
= \sum_{g_1' \in G} \sum_{g_2' \in G} \sum_{h \in G} \pi(h)f_1(g_1')f_2(g_2')\chi_1(g_1' + h)\chi_2(g_2' + h)
\]

\[
= \left( \sum_{g_1' \in G} f_1(g_1')\chi_1(g_1') \right) \left( \sum_{g_2' \in G} f_2(g_2')\chi_2(g_2') \right) \times
\]

\[
\left( \sum_{h \in G} \pi(h)\chi_1(h)\chi_2(h) \right)
\]

\[
= \hat{f}_1(\chi_1)\hat{f}_2(\chi_2)\hat{\pi}(\chi_1 \cdot \chi_2).
\]

In particular, this shows that the probability of observing group elements \((g_1, g_2)\) can be interpreted as a convolution over the group \( G^2 \) and hence in the Fourier coordinate space (i.e. phase space) the transformed version factorizes as a product of transformed versions of the components. A special feature to notice is the appearance of the \( \chi_1 \cdot \chi_2 \) term.

Definition 15.3.10. Call a group-based model a **generic group-based model** if the transition matrices are completely unrestricted, except for the condition that \( M(g, h) = f(g - h) \) for some function \( f \), and there is no restriction on the root distributions that can arise.

For example, the CFN model is the generic group-based model with respect to the group \( \mathbb{Z}_2 \) and the K3P model is the generic group-based model with respect to the group \( \mathbb{Z}_2 \times \mathbb{Z}_2 \). For a generic group-based model on a two leaf tree we can easily read off the representation of the model that is suggested in the Fourier coordinates.
Proposition 15.3.11. Consider the map $\phi_G : \mathbb{C}^G \times \mathbb{C}^G \times \mathbb{C}^G \to \mathbb{C}^G \times \mathbb{C}^G$, such that

$$(a_g)_{g \in G} \times (b_g)_{g \in G} \times (c_g)_{g \in G} \mapsto (a_{g_1}b_{g_2}c_{g_1+g_2})_{g_1,g_2 \in G}.$$ 

The Zariski closure of the image of $\phi_G$ equals the Zariski closure of the cone over the image of the generic group-based model on a two-leaf tree. In particular, since $\phi_G$ is a monomial map, the Zariski closure of the image is a toric variety.

Proof. From Example 15.3.9, we see that

$$\hat{p}(\chi_1, \chi_2) = \hat{f}_1(\chi_1)\hat{f}_2(\chi_2)\hat{\pi}(\chi_1 \cdot \chi_2).$$

Since we have assumed a generic group-based model, each of the functions $f_1$, $f_2$, and $\pi$ were arbitrary functions, subject to $\sum_{g \in G} f_i(g) = 1$, and $\sum_{g \in G} \pi(g) = 1$. Then the Fourier transforms are arbitrary functions subject to the constraints $\hat{f}_i(id) = 1$, $\hat{\pi}(id) = 1$, where $id$ denotes the identity element of $G$. Taking the cone and the Zariski closure yields the statement of the proposition. \[\square\]

In Proposition 15.3.11 note that we have implicitly made an identification between $G$ and $\hat{G}$. Again, this is not canonically determined and some choice must be made.

Example 15.3.12 (General Group-Based Model of Klein Group). To illustrate Proposition 15.3.11 consider the group $\mathbb{Z}_2 \times \mathbb{Z}_2$. We identify this group with the set of nucleotide bases (as in the K3P model) and declare that $A = (0,0)$, $C = (0,1)$, $G = (1,0)$, and $T = (1,1)$. The map $\phi_G$ in this example is

$$(a_A, a_C, a_G, a_T, b_A, b_C, b_G, b_T, c_A, c_C, c_G, c_T) \mapsto$$

$$\begin{pmatrix}
  a_A b_A c_A & a_A b_C c_C & a_A b_G c_G & a_B b_T c_T \\
  a_C b_A c_A & a_C b_C c_C & a_C b_G c_G & a_C b_T c_T \\
  a_G b_A c_A & a_G b_C c_C & a_G b_G c_G & a_G b_T c_T \\
  a_T b_A c_A & a_T b_C c_C & a_T b_G c_G & a_T b_T c_T
\end{pmatrix}.$$ 

We can compute the vanishing ideal in the ring $\mathbb{C}[q_{AA}, \ldots, q_{TT}]$, which is a toric ideal generated by 16 binomials of degree 3 and 18 binomials of degree 4. Here we are using $q_{g_1g_2}$ to be the indeterminate corresponding to the Fourier transformed value $\hat{p}(g_1, g_2)$. These Fourier coordinates are the natural coordinate system after applying the Fourier transform. The following code in Macaulay2 can be used to compute those binomial generators:

```macaulay2
S = QQ[aA,aC,aG,aT,bA,bC,bG,bT,cA,cC,cG,cT];
R = QQ[qAA,qAC,qAG,qAT,qCA,qCC,qCG,qCT,
       qGA,qGC,qGG,qGT,qTA,qTC,qTG,qTT];
```
$f = \text{map}(S, R, \{aA*bA*cA, aA*bC*cC, aA*bG*cG, aA*bT*cT, aC*bA*cC, aC*bC*cA, aC*bG*cT, aC*bT*cG, aG*bA*cG, aG*bC*cT, aG*bG*cA, aG*bT*cC, aT*bA*cT, aT*bC*cG, aT*bG*cT, aT*bT*cA\})$

$I = \text{kernel } f$

Among the generators of the vanishing ideal of this toric variety is the binomial

$q_{AA}q_{CG}q_{TC}q_{TA}q_{CC}q_{AG}$

in the Fourier coordinates. To express this equation in terms of the original probability coordinates we would need to apply the Fourier transform, which expresses each of the $q_{g_1}$ as a linear combination of the probability coordinates. For example, if we make the identification that $A = (0, 0), C = (0, 1), G = (1, 0), T = (1, 1)$ in both $\mathbb{Z}_2 \times \mathbb{Z}_2$ and its dual group, then

$q_{CG} = \sum_{(i,j),(k,l) \in \mathbb{Z}_2 \times \mathbb{Z}_2} (-1)^{i+k} p_{(i,j),(k,l)}$

$= \sum_{e \in T} M^e \cdot (g_T(e), g_{H(e)})^t$

Here we are using the following forms of the characters of $\mathbb{Z}_2 \times \mathbb{Z}_2$: $\chi_A(i,j) = 1, \chi_C(i,j) = (-1)^i, \chi_G(i,j) = (-1)^j, \text{ and } \chi_T(i,j) = (-1)^{i+j}$.

Now we state the general theorem about the Fourier transform of an arbitrary group-based model. To do this we need some further notation. Again, suppose that $T$ is a rooted tree with $m$ leaves. To each edge in the tree we have the transition matrix $M^e$, which satisfies $M^e(g, h) = f^e(g - h)$ for some particular function $f^e : G \to \mathbb{R}$. Also we have the root distribution $\pi$ which we write in function notation as $\pi : G \to \mathbb{R}$. Each edge $e$ in the graph has a head vertex and a tail vertex. Let $H(e)$ denote the head vertex and $T(e)$ denote the tail vertex. For each edge $e$, let $\chi(e)$ be the set of leaves which lie below the edge $e$. Let $Int$ be the set of interior vertices of $T$.

We can write the parametrization of the probability of observing group elements $(g_1, g_2, \ldots, g_m)$ at the leaves of $T$ as:

$(15.3.1) \quad p(g_1, \ldots, g_m) = \sum_{g \in G^{int}} \pi(g_r) \prod_{e \in T} M^e(g_T(e), g_{H(e)})$

$(15.3.2) \quad = \sum_{g \in G^{int}} \pi(g_r) \prod_{e \in T} f^e(g_{H(e)} - g_T(e)).$

**Theorem 15.3.13.** Let $p : G^m \to \mathbb{R}$ denote the probability distribution of a group-based model on an $m$-leaf tree. Then the Fourier transform of $p$
satisfies:

\[
\hat{p}(\chi_1, \ldots, \chi_m) = \hat{\pi} \left( \prod_{i=1}^{m} \chi_i \right) \cdot \prod_{e \in T} \hat{f}^e \left( \prod_{i \in \lambda(e)} \chi_i \right).
\]

The use of the Fourier transform to simplify the parametrization of a group-based model was discovered independently by Hendy and Penny [HP96] and Evans and Speed [ES93]. Hendy and Penny refer to this technique as Hadamard conjugation (technically, a more involved application of the method is Hadamard conjugation) because the matrix that represents the Fourier transform for \( \mathbb{Z}_2^n \) is an example of a Hadamard matrix from combinatorial design theory. See Example 15.3.4 for an example of the Hadamard matrix structure.

**Proof of Theorem 15.3.13.** We directly apply the Fourier transform to equation 15.3.1.

\[
\hat{p}(\chi_1, \ldots, \chi_m) = \sum_{(g_1, \ldots, g_m) \in G^n} \sum_{g \in G^{\text{int}}} \pi(g) \prod_{e \in T} f^e(gH(e) - gT(e)) \prod_{i=1}^{m} \chi_i(g).
\]

As in Example 15.3.9 we change variables by introducing a new group element \( g_e \) associated to each edge in the tree

\[
g_e = gH(e) - gT(e).
\]

With this substitution, the group elements associated to the leaves are rewritten as \( g_i = g_r + \sum_{e \in P(i)} g_e \) where \( P(i) \) is the set of edges on the unique path from the root of \( T \) to leaf \( i \). In summary we can rewrite our expression for \( \hat{p} \) as

\[
\hat{p}(\chi_1, \ldots, \chi_m) = \sum_{g_r \in G} \sum_{g_e \in G^{E(T)}} \pi(g_r) \prod_{e \in T} f^e(g_e) \prod_{i=1}^{m} \chi_i \left( g_r + \sum_{e \in P(i)} g_e \right).
\]

Using the fact that \( \chi(g + h) = \chi(g) \chi(h) \) for any one-dimensional character, we can regroup this product with all terms that depend on \( g_e \) together to produce:

\[
\hat{p}(\chi_1, \ldots, \chi_m) = \left( \sum_{g_r \in G} \pi(g_r) \prod_{i=1}^{m} \chi_i(g_r) \right) \times \prod_{e \in E(T)} \left( \sum_{g_e \in G} f^e(g_e) \prod_{i \in \lambda(e)} \chi_i(g_e) \right).
\]
15.3. Group-based Phylogenetic Models

Figure 15.3.1. A three leaf tree

\[
= \hat{\pi} \left( \prod_{i=1}^{m} \chi_i \right) \cdot \prod_{e \in T} \hat{f}_e \left( \prod_{t \in \lambda(e)} \chi_t \right)
\]
as claimed.

Sturmfels and the author [SS05] realized that in certain circumstances (the Zariski closure of) a group-based model can be identified as a toric variety in the Fourier coordinates and so combinatorial techniques can be used to find the phylogenetic invariants in these models. The key observation is that the Equation 15.3.3 expresses the Fourier transform of the probability distribution as a product of Fourier transforms. If the functions \( f_e(g) \) depended in a complicated way on parameters, this might not be a monomial parametrization. However, if those transition functions were suitably generic, then each entry of the Fourier transform of the transition function can be treated as an independent “Fourier parameter”. The simplest case is the general group-based model.

**Corollary 15.3.14.** Let \( G \) be a finite abelian group and \( T \) a rooted tree with \( m \) leaves. In the Fourier coordinates, the Zariski closure of the cone over the general group-based model is equal to the Zariski closure of the image of the map \( \phi^{G,T} \)

\[
\phi^{G,T} : \mathbb{C}^{(|G| \times (|E(T)| + 1)}} \rightarrow \mathbb{C}^{(|G|)}
\]

\[
\phi^{G,T}((a_e^g)_{e \in E(T), g \in G}, (a_g^r)_{g \in G}) = a_r^{\Sigma_{i \in [m]} g_i} \cdot \prod_{e \in E(T)} a_e^{\Sigma_{t \in \lambda(e)} g_t}
\]

**Proof.** The proof merely consists of translating the result of Theorem 15.3.13 into a more algebraic notation. Since the model is general, the function values of the Fourier transform are all algebraically independent. Hence the \( a_e^g \) parameters play the role of the Fourier transform of the root distribution and the \( a_r^g \) play the role of the Fourier transform of the transition functions. Note that we have used the (nonunique) isomorphism \( G \cong \hat{G} \) to write our indices using the \( g \in G \) rather than with \( \chi \in \hat{G} \).
Example 15.3.15 (General Group-Based Model on Three Leaf Tree). Consider a general group-based model for the three leaf tree illustrated in Figure 15.3.1. There are $|G|$ parameters associated to each of the edges in the tree, and another $|G|$ parameters associated to the root. The map $\phi^{G,T}$ has coordinate functions

$$
\phi_{g_1,g_2,g_3}^{G,T}(a) = a_{g_1+g_2+g_3}a_{g_2+g_3}a_{g_1}a_{g_2}a_{g_3},
$$

where the $a_i$ indeterminates (for $i = 1, 2, 3$) correspond to the edges incident to leaves, the $a_r$ indeterminates correspond to the root, and the $a_e$ indeterminates correspond to the single internal edge.

When a group-based model is not a general group-based model, some care is needed to see whether or not it is a toric variety, and, if so, how this affects the parametrization. First of all, we deal with the root distribution. Up to this point, in the group-based models we have treated the root distribution as a set of free parameters. This can be a reasonable assumption. However, often in phylogenetic models one takes the root distribution to be the (common) stationary distribution of all the transition matrices, should such a distribution exist. For all the group-based models it is not difficult to see that the uniform distribution is the common stationary distribution. Fixing this throughout our model, we can take the Fourier transform of this function.

Proposition 15.3.16. Let $G$ be a finite abelian group and let $f : G \to \mathbb{C}$ be the constant function $f(g) = \frac{1}{|G|}$. Then $\hat{f}$ is the function

$$
\hat{f}(\chi) = \begin{cases} 
1 & \text{if } \chi = \text{id} \\
0 & \text{otherwise}.
\end{cases}
$$

Proof. The rows of the character table are orthogonal to each other, and the row corresponding to the trivial character is the constant function. Hence, in the Fourier coordinates, the trivial character appears with nonzero weight and all others with weight 0. Since

$$
\hat{f}(\text{id}) = \sum_{g \in G} f(g)
$$

we see that $\hat{f}(\text{id}) = 1$. 

A straightforward combination of Corollary 15.3.14 and Proposition 15.3.16 produces the following result for group-based models. We leave the details to the reader.

Corollary 15.3.17. Let $G$ be a finite abelian group and $T$ a rooted tree with $m$ leaves. In the Fourier coordinates, the Zariski closure of the cone
over the general group-based model with uniform root distribution is equal to the Zariski closure of the image of the map \( \phi^{G,T} \):

\[
\phi^{G,T} : \mathbb{C}^{|G|} \times (|E(T)|) \rightarrow \mathbb{C}^{|G|^m}
\]

\[
\phi^{G,T}_{g_1, \ldots, g_m}((a_g^e)_{e \in E(T), g \in G}) = \left\{ \begin{array}{ll}
0 & \text{if } \sum_{i \in |m|} g_i = \text{id}\\
\prod_{e \in E(T)} a_{\sum_{i \in |m|} g_i} & \text{otherwise.}
\end{array} \right.
\]

From here we begin to focus on the other cases where \( f^e(g) \) is not allowed to be an arbitrary function. The main cases of interest are the Jukes-Cantor model and the Kimura 2-parameter model, but the questions make sense in more general mathematical setting.

For the Jukes-Cantor model we have

\[
 f^e(A) = a, \quad f^e(C) = f^e(G) = f^e(T) = b
\]

where \( a \) and \( b \) are specific functions of \( a \) and \( t \). Indeed, for the rate matrix \( Q^{JC} \) we have

\[
\exp(Q^{JC}t) = \begin{pmatrix}
1 + 3e^{-4t} & 1 - e^{-4t} & 1 - e^{-4t} & 1 - e^{-4t} \\
1 - e^{-4t} & 1 + 3e^{-4t} & 1 - e^{-4t} & 1 - e^{-4t} \\
1 - e^{-4t} & 1 - e^{-4t} & 1 + 3e^{-4t} & 1 - e^{-4t} \\
1 - e^{-4t} & 1 - e^{-4t} & 1 - e^{-4t} & 1 + 3e^{-4t}
\end{pmatrix}.
\]

Computing the Fourier transform of \( f^e(g) \) produces the function

\[
\hat{f}^e(A) = a + 3b, \quad \hat{f}^e(C) = \hat{f}^e(G) = \hat{f}^e(T) = a - b.
\]

As \( a + 3b \) and \( a - b \) are algebraically independent, we can treat these quantities as free parameters, so that the resulting model is still parametrized by a monomial map, akin to that in Corollary 15.3.17 but with the added feature that for each \( e \in E(T) \) \( a^e_G = a^e_C = a^e_T \). This has the consequence that there will exists pairs of tuples of group elements \( (g_1, \ldots, g_m) \neq (h_1, \ldots, h_m) \) such that \( \hat{p}(g_1, \ldots, g_m) = \hat{p}(h_1, \ldots, h_m) \). The resulting linear polynomial constraints on the phylogenetic model are called linear invariants.

**Example 15.3.18 (Linear Invariants of Jukes-Cantor Model).** Consider the tree from Figure 15.3.1, under the Jukes-Cantor model with the uniform root distribution. We use the indeterminates \( q_{AAA}, \ldots, q_{TTT} \) to denote coordinates on the image space. Then, for example

\[
q_{CGT} = a_C^1 a_C^2 a_C^3 a^e_C = a_C^1 a_C^2 a_C^3 a^e_C
\]

and

\[
q_{GTC} = a_C^1 a_C^2 a_C^3 a^e_C = a_C^1 a_C^2 a_C^3 a^e_C
\]

so \( q_{CGT} - q_{GTC} \) is a linear invariant for the Jukes-Cantor model on this three leaf tree.
A similar analysis can be run for the Kimura 2-parameter model to show that it is also a toric variety [SS05]. Analysis of the conditions on the functions $f'(g)$ that yield toric varieties in the Fourier coordinates appears in [Mic11].

To conclude this section, we note one further point about the parametrizations in the case where $G$ is a group with the property that every element is its own inverse (e.g. $\mathbb{Z}_2$ and $\mathbb{Z}_2 \times \mathbb{Z}_2$). If $G$ satisfies this property and we are working with the model with uniform root distribution, we can pass to an unrooted tree and the parametrization can be written as:

$$
\phi^G_{T \mid g_1, \ldots, g_m}((a_B^{[A]})_{A \mid B \in \Sigma(T), g \in G}) = \begin{cases} 
\prod_{A \mid B \in \Sigma(T)} a_B^{[A]} & \text{if } \sum_{i \in [m]} g_i = id \\
0 & \text{otherwise.}
\end{cases}
$$

The reason we can ignore the root in this case is because if $\sum_{i \in [m]} g_i = id$ then $\sum_{a \in A} g_i = \sum_{i \in B} g_i$ for groups with the property that every element is its own inverse. This can be a useful alternate representation of the parametrization which allows us to work with nonrooted trees.

15.4. The General Markov Model

As discussed previously, the general Markov model is the phylogenetic model that arises as the hidden variable graphical model on a rooted directed tree, where all vertices have random variables with the same number of states, $\kappa$, and all of the non-leaf vertices are hidden variables. While this model is not so widely used in phylogenetic practice, it does contain all the other models discussed in Sections 15.2 and 15.3 as submodels and so statements that can be made about this model often carry over to the other models. Furthermore, the general Markov model has nice mathematical properties that make it worthwhile to study. We highlight some of the mathematical aspects here.

A first step in analyzing the general Markov model on a tree is that we can eliminate any vertex of degree two from our tree. Furthermore, we can relocate the root to any vertex in the tree that we wish, without changing the family of probability distributions that arises. Often, then, the combinatorics can be considered only in terms of the underlying unrooted tree alone.

**Proposition 15.4.1.** Let $T_1$ and $T_2$ be rooted trees with the same underlying undirected tree. Then under the general Markov model $M_{GMM}^{T_1} = M_{GMM}^{T_2}$.

**Proof.** To see why this is so, we just need to recall results about graphical models from Chapter 13. First to see that we can move the root to any vertex of the tree without changing the underlying distribution, we use Theorem 13.1.13, which characterizes the directed graphical models with all
15.4. The General Markov Model

variables observed that yield the same family of probability distributions. Once we fix the underlying undirected graph to be a tree, choosing any node as the root yields a new directed tree with the same unshielded colliders (specifically, there are no unshielded colliders). Since they give the same probability distributions when all vertices are observed, they will give the same probability distributions after marginalizing the internal nodes. □

**Proposition 15.4.2.** Let $T$ and $T'$ be two rooted trees where $T'$ is obtained from $T$ by suppressing a vertex of degree 2. Then $\mathcal{M}_{T}^{GMM} = \mathcal{M}_{T'}^{GMM}$.

**Proof.** From Proposition 15.4.1, we can assume that the vertex of degree 2 is not the root. Let $v$ be this vertex. We consider what happens locally when we marginalize just the random variable $X_v$. Note that because the underlying graph is a tree, it will appear in the distribution a sequence of random variables $Y \rightarrow X_v \rightarrow Z$ and since the vertex has degree 2, these are the only edges incident to $X_v$. So, in the joint distribution for all the random variables in the tree will appear the product of conditional distribution $P(X_v = x | Y = y)P(Z = z | X_v = x)$. Since we consider the general Markov model, each of these conditional distributions is an arbitrary stochastic matrix. Summing out over $x$, produces an arbitrary stochastic matrix which represents the conditional probability $P(Z = z | Y = y)$. Thus, marginalizing this one variable produces the same distribution family as the fully observed model on $T$. This proves the proposition. □

Perhaps the best known results about the algebraic structure of the general Markov model concerns the case when $\kappa = 2$. The vanishing ideal is completely characterized in this case, combining results of Allman and Rhodes [AR08] and Raicu [Rai12]. To explain these results we introduce some results about flattenings of tensors.

Let $P \in \mathbb{C}^{r_1} \otimes \mathbb{C}^{r_2} \otimes \cdots \otimes \mathbb{C}^{r_m}$ be an $r_1 \times r_2 \times \cdots \times r_m$ tensor. Note that for our purposes a tensor in $\mathbb{C}^{r_1} \otimes \mathbb{C}^{r_2} \otimes \cdots \otimes \mathbb{C}^{r_m}$ is the same as an $r_1 \times r_2 \times \cdots \times r_m$ array with entries in $\mathbb{C}$. The language of tensors is used more commonly in phylogenetics. Let $A \mid B$ be a split of $[m]$. We can naturally regroup the tensor product space as

$$\left( \bigotimes_{a \in A} \mathbb{C}^{r_a} \right) \otimes \left( \bigotimes_{b \in B} \mathbb{C}^{r_b} \right)$$

and consider $P$ inside of this space. This amounts to considering $P$ as a $\prod_{a \in A} r_a \times \prod_{b \in B} r_b$ matrix. The resulting matrix is called the flattening of $P$ and is denoted $\text{Flat}_{A \mid B}(P)$. 
Example 15.4.3 (Flattening of a 4-way Tensor). Suppose that $\kappa = r_1 = r_2 = r_3 = r_4 = 2$, and let $A|B = \{1,3\}\{2,4\}$. Then

$$\text{Flat}_{13|24}(P) = \begin{pmatrix} P_{1111} & P_{1112} & P_{1121} & P_{1122} \\ P_{1211} & P_{1212} & P_{1221} & P_{1222} \\ P_{2111} & P_{2112} & P_{2211} & P_{2221} \\ P_{2121} & P_{2122} & P_{2221} & P_{2222} \end{pmatrix}.$$  

In Section 15.1, we introduced the set of splits $\Sigma(T)$ of a tree $T$. For the purposes of studying phylogenetic invariants of the general Markov model, it is also useful to extend this set slightly for nonbinary trees. Define $\Sigma'(T)$ to be the set of all splits of the leaf set of a tree that could be obtained by resolving vertices of degree $>3$ into vertices of lower degree. For example, if $T$ is a star tree with one internal vertex and $m$ leaves, then $\Sigma'(T)$ will consist of all splits of $[m]$.

Proposition 15.4.4. Let $T$ be a tree and let $\mathcal{M}_T$ be the phylogenetic model under the general Markov model with random variables of $\kappa$ states. Then for any $A|B \in \Sigma'(T)$, and any $P \in \mathcal{M}_T$

$$\text{rank}(\text{Flat}_{A|B}(P)) \leq \kappa.$$  

Thus all $(\kappa + 1) \times (\kappa + 1)$ minors of $\text{Flat}_{A|B}(P)$ belong to the vanishing ideal of $\mathcal{M}_T$. Furthermore, if $A|B \notin \Sigma'(T)$ then for generic $P \in \mathcal{M}_T$

$$\text{rank}(\text{Flat}_{A|B}(P)) \geq \kappa^2.$$  

Proof. Recall that we can subdivide an edge in the tree and make the new internal node of degree 2 the root vertex (directed all other edges away from this edge). Since the underlying graph of our model is a tree, the joint distribution of states at the leaves can be expressed as

$$(15.4.1) \quad P_{i_1...i_m} = \sum_{j=1}^{\kappa} \pi_j F(j, i_A) G(j, i_B)$$  

where $F(j, i_A)$ is the function obtained by summing out over all the hidden variables between the root node and the set of leaves $A$, and $G(j, i_B)$ is the function obtained by summing out over all the hidden variables between the root node and the set of leaves $B$. Let $F$ denote the resulting $\kappa^{[A]} \times \kappa$ matrix that is the flattening of this tensor and similarly let $G$ denote the resulting $\kappa \times \kappa^{[B]}$ matrix. Then Equation 15.4.1 can be expressed in matrix form as:

$$\text{Flat}_{A|B}(P) = F \text{diag}(\pi) G,$$

which shows that $\text{rank}(\text{Flat}_{A|B}(P)) \leq \kappa$.

For the second part of the theorem, note that it suffices to find a single $P \in \mathcal{M}_T$ such that $\text{rank}(\text{Flat}_{A|B}(P)) \geq \kappa^2$. This is because the set of matrices with $\text{rank}(\text{Flat}_{A|B}(P)) \leq r$ is a variety, and the Zariski closure
of the model $\mathcal{M}_T$ is an irreducible variety. So suppose that $A|B \notin \Sigma'(T)$. This means that there exists another split $C|D \in \Sigma(T)$ such that $A|B$ and $C|D$ are incompatible. Note that since $C|D \in \Sigma(T)$ that split actually corresponds to an edge in $T$ (whereas splits in $\Sigma'(T)$ need not correspond to edges in $T$). We construct the distribution $P$ as follows. Set all parameters on all edges besides the one corresponding to the edge $C|D$ to the identity matrix. Set the transition matrix corresponding to the edge $C|D$ to be a generic matrix $\alpha^{2 \times 2}$. The resulting joint distribution $P$ has the following form:

$$P_{i_1 \cdots i_m} = \begin{cases} 
\alpha_{jk} & \text{if } i_c = j \text{ and } i_d = k \text{ for all } c, d \in D \\
0 & \text{otherwise}
\end{cases}$$

Since $A|B$ and $C|D$ are not compatible this means that none of $A \cap C$, $A \cap D$, $B \cap C$, nor $B \cap D$ are empty. Now form the submatrix of $M$ of Flat $A|B(P)$ whose row indices consists of all tuples $i_A$ such that $i_{A|C}$ and $i_{A|D}$ are constant, and whose column indices consist of all tuples $i_B$ such that $i_{B|C}$ and $i_{B|D}$ are constant. The matrix $M$ will be a $\kappa^2 \times \kappa^2$ matrix which has one nonzero entry in each row and column. Such a matrix has rank $\kappa^2$, which implies that Flat $A|B(P) \succeq \kappa^2$.

A strengthening of the second part of Proposition 15.4.4 due to Eriksson [Eri05] shows how the rank of Flat $A|B(P)$ could be even larger if the split $A|B$ is incompatible with many splits in $T$.

**Theorem 15.4.5.** [AR08, Rai12] Consider the general Markov model on a tree $T$ and suppose that $\kappa = 2$. Then the homogeneous vanishing ideal of $\mathcal{M}_T$ is generated by all the $3 \times 3$ minors of all flattenings Flat $A|B(P)$ for $A|B \in \Sigma'(T)$:

$$I^h(\mathcal{M}_T) = \sum_{A|B \in \Sigma'(T)} \langle 3 \times 3 \text{ minors of Flat } A|B(P) \rangle \subseteq \mathbb{R}[P].$$

Allman and Rhodes [AR08] proved Theorem 15.4.5 in the case of binary trees and Raicu [Rai12] proved the theorem in the case of star trees (in the language of secant varieties of Segre varieties). The Allman-Rhodes-Draisma-Kuttler theorem, to be explained in the next section, shows that it suffices to prove the full version of Theorem 15.4.5 in the case of star trees to get the result for general (nonbinary) trees.

**Example 15.4.6.** [General Markov Model on Five Leaf Tree] Consider the unrooted tree with 5 leaves in Figure 15.4.1. This tree has two nontrivial splits 14/235 and 25/134. According to Theorem 15.4.5 the homogeneous vanishing ideal $I^h(\mathcal{M}_T) \subseteq \mathbb{R}[P_{i_1i_2i_3i_4i_5}]$ is generated by the $3 \times 3$ minors of the following matrices:
It remains an open problem to describe the vanishing ideal of the general Markov model for larger $\kappa$. For the three leaf star tree and $\kappa = 4$, a specific set of polynomials of degree 5, 6, and 9 are conjectured to generate the vanishing ideal. While some progress has been made using numerical techniques to prove that these equations described the variety set-theoretically \cite{BO11, FG12} or even ideal theoretically \cite{DH16} it remains an open problem to give a symbolic or theoretical proof of this fact.

Besides the vanishing ideal of the general Markov model, a lot is also known about the semialgebraic structure of these models. For example, the paper \cite{ART14} gives a semialgebraic description of the general Markov model for all $\kappa$ (assuming the vanishing ideal is known) for the set of all distributions that can arise assuming that the root distribution does not have any zeroes in it and all transition matrices are non-singular matrices. Both of these assumptions are completely reasonable from a biological standpoint.

The paper \cite{ZS12} gave a complete semialgebraic description of the general Markov model in the case that $\kappa = 2$. While the techniques of that paper do not seem to extend to handle larger values of $\kappa$, the methods introduced there extend to other models on binary random variables and are worthwhile to study for these reasons. In some sense, the idea is similar to what has been done for group-based models above, namely make a coordinate change that simplifies the parametrization. In the case of the general

![Figure 15.4.1. The unrooted 5-leaf tree from Example 15.4.6](image-url)
Markov model, the coordinate change is a nonlinear change of coordinates that changes to cumulant coordinates (or what Zwiernik and Smith call tree cumulants). Here we focus on ordinary cumulants and their algebraic geometric interpretation (as studied in [SZ13]).

For binary random variables (with state space \{0,1\}) we encode the distribution by its probability generating function

\[ P(x_1, \ldots, x_m) = \sum_{I \subseteq [m]} p_I \prod_{i \in I} x_i. \]

Here \( p_I \) denotes the probability that \( X_i = 1 \) for \( i \in I \) and \( X_i = 0 \) for \( i \not\in I \). The moment generating function of this distribution is

\[ M(x_1, \ldots, x_m) = P(1 + x_1, \ldots, 1 + x_m) = \sum_{I \subseteq [m]} \mu_I \prod_{i \in I} x_i. \]

In the case of binary random variables, the moment \( \mu_I \) is the probability that \( X_i = 1 \) for \( i \in I \). The logarithm of the moment generating function gives the cumulants:

\[ K(x_1, \ldots, x_m) = \sum_{I \subseteq [m]} k_I \prod_{i \in I} x_i = \log M(x_1, \ldots, x_m). \]

This power series expansion is understood to be modulo the ideal \((x_1^2, \ldots, x_m^2)\) so that we only keep the squarefree terms. The coefficients \( k_I \) for \( I \subseteq [m] \) are called the cumulants of the binary random vector \( X \).

It is known how to write an explicit formula for the cumulants in terms of the moments:

\[ k_I = \sum_{\pi \in \Pi(I)} (-1)^{|\pi|-1} (|\pi| - 1)! \prod_{B \in \pi} \mu_B \]

where \( \Pi(I) \) denotes the lattice of set partitions of the set \( I \), \(|\pi|\) denotes the number of blocks of the partition \( I \), and \( B \in \pi \) denotes the blocks of \( \pi \). For example

\[ k_i = \mu_i, \quad k_{ij} = \mu_{ij} - \mu_i \mu_j, \quad k_{ijk} = \mu_{ijk} - \mu_i \mu_{jk} - \mu_j \mu_{ik} - \mu_k \mu_{ij} + 2\mu_i \mu_j \mu_k. \]

We use the cumulant transformation to analyze the three leaf claw tree under the general Markov model with \( \kappa = 2 \). This is, of course, the same model as the mixture model \( \text{Mixt}^2(\mathcal{M}_{X_1 \perp X_2 \perp X_3}) \). Since the passage from the probability generating function to the moment generating function is linear, we can compute this first for the model \( \mathcal{M}_{X_1 \perp X_2 \perp X_3} \) and then pass to the mixture model. However, it is easy to see directly that the moment generating function for the complete independence model is

\[ (1 + \mu_1 x_1)(1 + \mu_2 x_2)(1 + \mu_3 x_3). \]
Passing to the secant variety, we have the moment generating function of \( \text{Mixt}^2(\mathcal{M}_{X_1} \sqcup X_2 \sqcup X_3) \) as

\[
1 + (\pi_0 \mu_{01} + \pi_1 \mu_{11})x_1 + (\pi_0 \mu_{02} + \pi_1 \mu_{12})x_2 + (\pi_0 \mu_{03} + \pi_1 \mu_{13})x_3 \\
+ (\pi_0 \mu_{01} \mu_{02} + \pi_1 \mu_{11} \mu_{12})x_1 x_2 + (\pi_0 \mu_{01} \mu_{03} + \pi_1 \mu_{11} \mu_{13})x_1 x_3 \\
+ (\pi_0 \mu_{02} \mu_{03} + \pi_1 \mu_{12} \mu_{13})x_2 x_3 + (\pi_0 \mu_{01} \mu_{02} \mu_{03} + \pi_1 \mu_{11} \mu_{12} \mu_{13})x_1 x_2 x_3.
\]

Then we pass to the cumulants. We have

\[
k_i = \mu_i = (\pi_0 \mu_{0i} + \pi_1 \mu_{1i})
\]

\[
k_{ij} = -\pi_0 \pi_1 (\mu_{0i} - \pi_1 \mu_{0j} - \mu_{0j} + \pi_1 \mu_{1j})
\]

\[
k_{123} = -\pi_0 \pi_1 (\mu_{01} - \pi_1 \mu_{02} - \mu_{02} + \pi_1 \mu_{12}) (\mu_{03} - \pi_1 \mu_{13}).
\]

From this we see that

\[
k_{123}^2 + 4k_{12}k_{13}k_{23} = \pi_0^2 \pi_1^2 (\mu_{01} - \mu_{11})^2 (\mu_{02} - \mu_{12})^2 (\mu_{03} - \mu_{13})^2
\]

and in particular is nonnegative. In summary, we have the following result.

**Proposition 15.4.7.** Let \( p \) be a distribution coming from the three leaf trivalent tree under the general Markov model with \( \kappa = 2 \). Then the cumulants satisfy the inequality:

\[
k_{123}^2 + 4k_{12}k_{13}k_{23} \geq 0.
\]

This inequality together with other more obvious inequalities on the model turn out to characterize the distributions that can arise for a three leaf tree under the general Markov model. Similar arguments extend to arbitrary trees where it can be shown that a suitable analogue of the cumulants factorizes into a product. Hence, in cumulant coordinates these models are also toric varieties. This is explored in detail in [ZS12]. See also the book [Zwi16] for further applications of cumulants to tree models.

### 15.5. The Allman-Rhodes-Draisma-Kuttler Theorem

The most general tool for constructing generators of the vanishing ideals of phylogenetic models is a theorem of Allman and Rhodes [AR08] and Draisma and Kuttler [DK09] which allows one to reduce the determination of the ideal generators to the case of star trees. This result holds quite generally in the algebraic setting, for models that are called equivariant phylogenetic models. These models include the group-based models, the general Markov model, and the strand symmetric model. The full version of the result was proven by Draisma and Kuttler [DK09] based on significant partial results that appeared in earlier papers for group-based models [SS05], the general Markov model [AR08] and the strand symmetric model [CS05]. In particular, the main setup and many of the main ideas already appeared in [AR08] and for this reason we also credit Allman and Rhodes.
15.5. The Allman-Rhodes-Draisma-Kuttler Theorem

with this theorem. We outline this main result and the idea of the proof in
this section.

**Theorem 15.5.1.** Let $T$ be a tree and $\mathcal{M}_T$ the model of an equivariant
phylogenetic model on $T$. Then the homogeneous vanishing ideal of $\mathcal{M}_T$ is
a sum

$$I(\mathcal{M}_T) = \sum_{e \in E(T)} I_e + \sum_{v \in \text{Int}(T)} I_v$$

where $I_e$ and $I_v$ are certain ideals associated to flattenings along edges and
vertices of $T$ respectively.

In the case of the general Markov model, the $I_e$ ideals are generated by
the $(\kappa + 1) \times (\kappa + 1)$ minors of the flattening matrix $\text{Flat}_{A|B}(P)$
associated to the edge $e$. In the case of binary trees, the ideals $I_v$ correspond to
flattening to 3-way tensors associated to the three-way partition from a
vertex $v$ and the requirement that that 3-way tensor have rank $\leq \kappa$. For the
other equivariant models, similar types of statements hold though we refer
the reader to the references above for the precise details.

The main tool to prove Theorem 15.5.1 is a gluing lemma, which we
describe in some detail for the case of the general Markov model. Let $\mathcal{M}_T$
be a phylogenetic model and let $V_T$ be the corresponding projective variety
(obtained by looking at the cone over $\mathcal{M}_T$). We assume that $T$ has $m$
leaves. This variety sits naturally inside of the projective space over the
tensor product space $\mathbb{P}(\mathbb{C}^\kappa \otimes \cdots \otimes \mathbb{C}^\kappa)$. The product of general linear groups
$\text{Gl}_\kappa^m$ naturally acts on the vector space $\mathbb{C}^\kappa \otimes \cdots \otimes \mathbb{C}^\kappa$ with each copy of $\text{Gl}_\kappa$
acting on one coordinate. Explicitly on tensors of rank one this action has
the following form:

$$(M_1, \ldots, M_m) \cdot a_1 \otimes \cdots \otimes a_m = M_1 a_1 \otimes \cdots \otimes M_m a_m.$$ 

**Proposition 15.5.2.** The variety $V_T \subseteq \mathbb{P}(\mathbb{C}^\kappa \otimes \cdots \otimes \mathbb{C}^\kappa)$ is fixed under the
action of $\text{Gl}_\kappa^m$.

**Proof.** On the level of the tree distribution, multiplying by the matrix $M_i$
at the $i$th leaf is equivalent to adding a new edge $i \to i'$ with $i$ now being a
hidden variable and $i'$ the new leaf, and $M_i$ the matrix on this edge. Since we
can contract vertices of degree two without changing the model (Proposition
15.4.2), this gives us the same family of probability distributions, and hence
the same variety $V_T$. \qed

Now let $T_1$ and $T_2$ be two trees, and let $T_1 * T_2$ denote the new tree
obtained by gluing these two trees together at a leaf vertex in both trees.
A specific vertex must be chosen on each tree to do this, but we leave the
notation ambiguous. This is illustrated in Figure 15.5.1. Suppose that $T_i$
Figure 15.5.1. Glueing two trees together at a vertex.

has \( m_i \) leaves for \( i = 1, 2 \). As in the proof of Proposition 15.4.4, we note that any distribution on the tree \( T_1 \ast T_2 \) could be realized by taking a distribution \( P_1 \) for the tree \( T_1 \), and \( P_2 \) for the tree \( T_2 \), flattening the first distribution down to a \( \kappa^{m_1-1} \times \kappa \) matrix, flattening the second distribution down to a \( \kappa^{m_2-1} \) matrix, and forming the matrix product \( P_1 P_2 \). The resulting \( \kappa^{m_1-1} \times \kappa^{m_2-1} \) matrix is an element of \( V_{T_1 \ast T_2} \). Clearly that matrix has rank \( \leq \kappa \) by Proposition 15.4.4, which tells us some equations in the vanishing ideal \( I(V_{T_1 \ast T_2}) \). Lemma 15.5.4 shows that all generators for \( I(V_{T_1 \ast T_2}) \) can be recovered from these minors plus equations that come from \( I(V_{T_1}) \) and \( I(V_{T_2}) \). We state this result in some generality.

**Definition 15.5.3.** Let \( V \subseteq \mathbb{P}(\mathbb{C}^{r_1} \otimes \mathbb{C}^\kappa) \) and \( W \subseteq \mathbb{P}(\mathbb{C}^\kappa \otimes \mathbb{C}^{r_2}) \) be projective varieties. The **matrix product variety** is the variety of \( V \) and \( W \) is

\[
V \ast W = \{ v \cdot w : v \in V, w \in W \}
\]

where \( v \cdot w \) denotes the product of the \( r_1 \times \kappa \) matrix, \( v \), and \( \kappa \times r_2 \) matrix, \( w \).

**Lemma 15.5.4.** Let \( V \subseteq \mathbb{P}(\mathbb{C}^{r_1} \otimes \mathbb{C}^\kappa) \) and \( W \subseteq \mathbb{P}(\mathbb{C}^\kappa \otimes \mathbb{C}^{r_2}) \) be projective varieties. Suppose that both \( V \) and \( W \) are invariant under the action of \( \text{Gl}_\kappa \) (acting as right matrix multiplication on \( V \) and left matrix multiplication on \( W \)). Then

\[
I(V \ast W) = \langle (\kappa + 1) \times (\kappa + 1) \text{ minors of } P \rangle + \text{Lift}(I(V)) + \text{Lift}(I(W)).
\]

The ideal \( I(V \ast W) \) belongs to the polynomial ring \( \mathbb{C}[P] \) where \( P \) is an \( r_1 \times r_2 \) matrix of indeterminates. The lifting operation \( \text{Lift}(I(V)) \) is defined as follows. Introduce a matrix of indeterminates \( Z \) of size \( r_2 \times \kappa \). The resulting matrix \( PZ \) has size \( r_1 \times \kappa \). Take any generator \( f \in I(V) \) and form the polynomial \( f(PZ) \) and extract all coefficients of all monomials in \( Z \) appearing in that polynomial. The resulting set of polynomials that arises is \( \text{Lift}(I(V)) \) as we range over all generators for \( I(V) \). An analogous construction defines the elements of \( \text{Lift}(I(W)) \). Note that the proof of Lemma 15.5.4 depends on some basic constructions of classical invariant theory (see [DK02]).
Proof of Lemma 15.5.4. We clearly have the containment $\subseteq$, so we must show the reverse containment. Let $f \in I(V \star W)$. Since $(\kappa+1) \times (\kappa+1)$ minors of $P$ $\subseteq I(V \star W)$, we can work modulo this ideal. This has the effect of plugging in $P = YZ$ in the polynomial $f$ where $Y$ is a $r_1 \times \kappa$ matrix of indeterminates and $Z$ is a $\kappa \times r_2$ matrix of indeterminates. The resulting polynomial $f(YZ)$ satisfies
\[ f(YZ) \in I(V) + I(W) \subseteq \mathbb{C}[Y, Z] \]
by construction. We can explicitly write
\[ f(YZ) = h_1(Y, Z) + h_2(Y, Z) \]
where $h_1 \in I(V)$ and $h_2 \in I(W)$.

Now both the varieties $V$ and $W$ are invariant under the action of $Gl_\kappa$ on the left and right, respectively. We let $Gl_\kappa$ act on $V \times W$ by acting by multiplication by $g$ on $V$ and by $g^{-1}$ on $W$, with the corresponding action on the polynomial ring $\mathbb{C}[Y, Z]$. The First Fundamental Theorem of Classical Invariant Theory says that when $Gl_\kappa$ acts this way on the polynomial ring $\mathbb{C}[Y, Z]$, the invariant ring is generated by the entries of $YZ$.

Apply the Reynolds operator $\rho$ with respect to the action of $Gl_\kappa$ to the polynomial $f(YZ)$. Note that $f(YZ)$ is in the invariant ring by the First Fundamental Theorem, so $\rho(f)(YZ) = f(YZ)$. This produces
\[ \rho(f)(YZ) = \rho(h_1)(Y, Z) + \rho(h_2)(Y, Z) \]
\[ f(YZ) = \tilde{h}_1(YZ) + \tilde{h}_2(YZ). \]
The Reynolds operator produces $\tilde{h}_1(YZ)$ and $\tilde{h}_2(YZ)$ which both necessarily belong to the invariant ring $\mathbb{C}[Y, Z]$. The polynomials $\tilde{h}_1(YZ)$ and $\tilde{h}_2(YZ)$ are now in the invariant ring, so this means we can lift these polynomials to polynomials $h_1'\,(P)$ which is in the ideal generated by $\text{Lift}(I(V))$ and $h_2'(P)$ which is in the ideal generated by $\text{Lift}(I(W))$. But then $f - h_1' - h_2'$ is a polynomial which evaluates to zero when we plug in $P = YZ$. This implies that $f - h_1' - h_2'$ is in the ideal generated by the $(\kappa+1) \times (\kappa+1)$ minors of $P$, which completes the proof.

To apply Lemma 15.5.4 to deduce Theorem 15.5.1 one successively breaks the trees into smaller and smaller pieces, gluing together a large tree out of a collection of star trees. One only needs to verify that the lift operation turns determinants into determinants, and preserves the structure of the ideal associated to internal vertices of the trees. See [DK09] for details of the proof, as well as further details on how variations on Lemma 15.5.4 can be applied to the group-based models and the strand symmetric model.
15.6. Exercises

**Exercise 15.1.** Check that the stationary distributions of the CFN, JC69, K2P, and K3P models are uniform, and that the stationary distributions of the F81 and HKY models are \((\pi_A, \pi_C, \pi_G, \pi_T)\).

**Exercise 15.2.** Show that the F81 model is a Lie Markov model, but that the HKY model is not a Lie Markov model.

**Exercise 15.3.** Show that any GTR rate matrix is diagonalizable, and all eigenvalues are real.

**Exercise 15.4.** Prove the analogue of Proposition 15.4.1 for group-based models.

**Exercise 15.5.** Compute phylogenetic invariants for the CFN model with uniform root distribution on a five leaf binary tree in both Fourier coordinates and probability coordinates.

**Exercise 15.6.** Consider the Jukes-Cantor model on an \(m\)-leaf binary tree \(T\) with uniform root distribution. Show that the number of linearly independent linear invariants for this model is \(4^m - F_{2m}\) where \(F_k\) denotes the \(k\)th Fibonacci number \(F_0 = 1, F_1 = 1, F_{k+1} = F_k + F_{k-1}\). That is, the number of distinct nonzero monomials that arise in the parametrization is \(F_{2m}\). (Hint: Show that the distinct nonzero monomials that arise are in bijection with the subforests of \(T\). See [SF95] for more details.)

**Exercise 15.7.** Let \(G\) be an undirected graph, \(H \uplus O\) a partition of the vertices into hidden and observed variables, and \(A \upharpoonright B\) a split of the observed variables. Let \(P \in M_G\) be a distribution from the graphical model on the observed variables. What can be said about the rank of \(\text{Flat}_{A \upharpoonright B}(P)\) in terms of the combinatorial structure of the graph and the positioning of the vertices of \(A\) and \(B\)?

**Exercise 15.8.** Generalize Proposition 15.4.4 to mixture models over the general Markov model. In particular, if \(P \in \text{Mixture}^k(M_T)\), and \(A \upharpoonright B \in \Sigma(T)\) show that

\[
\text{rank}(\text{Flat}_{A \upharpoonright B}(P)) \leq k\kappa.
\]

**Exercise 15.9.** Consider the general Markov model with \(\kappa = 2\) on an \(m\) leaf star tree. How what does the parametrization of this model look like in the cumulant coordinates? What sort of equalities and inequalities arise for this model?
Identifiability

A model is identifiable if all of the model parameters can be recovered from the observables of the model. Recall that a statistical model is a map \( p : \Theta \to P_\Theta \), \( \theta \mapsto p_\theta \). Thus, a statistical model is identifiable if \( \theta \) can be recovered from \( p_\theta \); that is, the mapping is one-to-one. Identifiability is an important property for a statistical model to have if the model parameters have special meaning for the data analysis. For example, in phylogenetic models the underlying rate matrices and branch lengths of the model tell how quickly or slowly evolution happened along a particular edge of the tree and the goal of a phylogenetic analysis is to recover those mutation rates/branch lengths. If a model is not identifiable, and we care about recovering information about the specific values of the parameters from data, there is no point in doing data analysis with such a model. Identifiability problems are often most difficult, and of most practical interest, in cases where there are hidden variables in the models.

Since the mappings that arise for many of these statistical models can be seen as polynomial or rational mappings, identifiability of a statistical model often is equivalent to the injectivity of a polynomial or rational map. Hence, techniques from computational algebra and algebraic geometry can be used to decide the identifiability of these models. Another common feature is that models might have both continuous and discrete parameters. For example, in a phylogenetic model, the continuous parameters are the mutation rates/branch lengths, and the discrete parameter is the underlying tree. Identifiability of parameters of both of these types can be addressed using algebraic tools.

Identifiability analysis has been one of the successful areas where tools from algebra have been applied to the study of statistical models. In this
chapter we give an overview of these methods, as well as a number of applications highlighting the methods’ applicability. We also discuss identifiability of dynamical systems models (i.e. ordinary differential equations) where similar algebraic techniques are also useful.

16.1. Tools for Testing Identifiability

A statistical model is identifiable if the parameters of the model can be recovered from the probability distributions. There are many variations on this definition and sometimes these distinctions are not so clear in research articles from different sources and in different research areas. We highlight these distinctions here. We also describe some of the basic computational algebra approaches that can be used to test identifiability.

In this chapter we assume throughout that we are working with an algebraic exponential family (see Section 6.5). So the map \( \phi : \Theta \to N \) maps the parameter space \( \Theta \) into (a possibly transformed version of) the natural parameter space \( N \) of an exponential family. We assume that \( \phi \) is a rational map, defined everywhere on \( \Theta \), and that \( \Theta \subseteq \mathbb{R}^d \) is a semialgebraic set of dimension \( d \). The model \( \mathcal{M} \) is the image of \( \phi \).

Definition 16.1.1. Let \( \phi : \Theta \to N \) be as above, with model \( \mathcal{M} = \text{im} \phi \). The parameter vector \( \theta \) is:

- globally identifiable if \( \phi \) is a one-to-one map on \( \Theta \);
- generically identifiable if \( \phi^{-1}(\phi(\theta)) = \{ \theta \} \) for almost all \( \theta \in \Theta \);
- rationally identifiable if there is a dense open subset of \( U \subseteq \Theta \) and a rational function \( \psi : N \to \Theta \) such that \( \psi \circ \phi(\theta) = \theta \) on \( U \);
- locally identifiable if \( |\phi^{-1}(\phi(\theta))| < \infty \) for almost all \( \theta \in \Theta \);
- non-identifiable if \( |\phi^{-1}(\phi(\theta))| > 1 \) for some \( \theta \in \Theta \); and
- generically non-identifiable if \( |\phi^{-1}(\phi(\theta))| = \infty \) for almost all \( \theta \in \Theta \).

Note that some of these definitions also have varying names in the literature. Generic identifiability is also called almost everywhere identifiability. Local identifiability is sometimes called finite identifiability or algebraic identifiability. Note that the name local identifiability arises from the following alternate interpretation: if a model is locally identifiable then around a generic point \( \theta \) of parameter space, there exists an open neighborhood \( U_\theta \) such that \( \phi : U_\theta \to \mathcal{M} \) is identifiable. In some instances of local identifiability, it might be reasonable to call this generic identifiability if the finite-to-one behavior is due to an obvious symmetry among states of a hidden variable. This point will be elaborated upon in Section 16.3.
Example 16.1.2 (Independence Model). Consider the independence model \( M_{X_1 \perp \perp X_2} \). This model is defined parametrically via the map
\[
\phi : \Delta_{r_1-1} \times \Delta_{r_2-1} \to \Delta_{r_1 r_2-1}, \quad (\alpha, \beta) \mapsto \alpha \beta^T.
\]
The independence model is globally identifiable because it is possible to recover \( \alpha \) and \( \beta \) from the probability distribution \( p = \alpha \beta^T \). Indeed
\[
\alpha_i = \sum_j p_{ij} \quad \text{and} \quad \beta_j = \sum_i p_{ij}
\]
for \( p \in M_{X_1 \perp \perp X_2} \). Also, these formulas to recover the parameters are rational functions, so the parameters are rationally identifiable as well.

Example 16.1.3 (Mixture of Independence Model). Consider the mixture model \( \text{Mixt}_k(M_{X_1 \perp \perp X_2}) \). The number of parameters of this model is
\[
k(r_1 + r_2 - 2) + k - 1 = k(r_1 + r_2 - 1) - 1.
\]
This count arises by picking \( k \) points on the model \( M_{X_1 \perp \perp X_2} \) which has dimension \( r_1 + r_2 - 2 \), and then a distribution \( \pi \in \Delta_{k-1} \) for the mixing weights.

On the other hand, the model \( \text{Mixt}_k(M_{X_1 \perp \perp X_2}) \) is contained inside the set of probability distributions that are rank \( \leq k \) matrices. The dimension of this set can be calculated as
\[
k(r_1 + r_2 - k) - 1.
\]
Since the number of parameters of \( \text{Mixt}_k(M_{X_1 \perp \perp X_2}) \) is greater than the dimension of the model (when \( k > 1 \)) we see that this model is generically non-identifiable.

Example 16.1.4 (1-factor Model). Consider the 1-factor model \( F_{m,1} \) introduced in Example 6.5.4. This consists of all covariance matrices that are diagonal plus a rank one matrix:
\[
F_{m,1} = \{ \Omega + \lambda \lambda^T \in PD_m : \Omega \succ 0 \text{ is diagonal, and } \lambda \in \mathbb{R}^m \}.
\]
Here \( \Omega \succ 0 \) means \( \Omega \) is positive definite, i.e. it has all positive entries along the diagonal. Our parametrization map is
\[
\phi : \mathbb{R}_{>0}^m \times \mathbb{R}^m \to PD_m \quad \phi(\omega, \lambda) = \left\{ \begin{array}{ll}
\lambda_i \lambda_j & \text{if } i \neq j \\
\omega_i + \lambda_i^2 & \text{if } i = j.
\end{array} \right.
\]
Clearly \( \phi \) is not a one-to-one map since \( \phi(\omega, \lambda) = \phi(\omega, -\lambda) \) but perhaps this is the only obstruction to identifiability.

Let \( \Sigma \in F_{m,1} \). Suppose that \( i \neq j \neq k \) and \( \sigma_{jk} \neq 0 \). Then
\[
\frac{\sigma_{ij} \sigma_{ik}}{\sigma_{jk}} = \frac{\lambda_i \lambda_j \lambda_j \lambda_k}{\lambda_j \lambda_k} = \lambda_i^2
\]
identifiability giving two choices for $\lambda_i$. If this value is not zero, once we make a choice of $\pm \lambda_i$ the entry $\sigma_{ij} = \lambda_i \lambda_j$ can be used to solve for $\lambda_j$. Furthermore, we can solve for $\omega_i$ via

$$\omega_i = \sigma_{ii} - \frac{\sigma_{ik}\sigma_{jk}}{\sigma_{jk}}.$$ 

This implies that the map $\phi$ parametrizing the 1-factor model is generically 2-to-1, so the 1-factor model is locally identifiable.

Note however, that model parameters are non-identifiable in this case, even if we restrict to require $\lambda \geq 0$ (in which case the map becomes generically 1-to-1). This is because there are some parameter choices $\theta$ where the preimage jumps in dimension. Specifically, when $\lambda = 0$ and $\phi(\omega,0)$ is a diagonal matrix there are many choices of $\lambda$ and changes to $\omega$ that could yield the same diagonal covariance matrix. For example, if we take $N' = (\lambda_1, 0, 0, \ldots, 0)$, and set $\omega' = (\omega_1 - \lambda_1^2, \omega_2, \ldots, \omega_m)$ then $\phi(\omega, 0) = \phi(\omega', \lambda')$. This gives an (at least) one dimensional set that has the same image as the pair $(\omega, 0)$.

Often we might have a model that is (generically) non-identifiable but we might be interested in whether some individual parameters of the model are identifiable. For example, in the 1-factor model from Example 16.1.4, the $\omega_i$ parameters are generically identifiable while the $\lambda_i$ parameters are locally identifiable, and the value $\lambda_i^2$ is generically identifiable. We formalize this in a definition.

**Definition 16.1.5.** Let $\phi : \Theta \to N$ be as above, with model $\mathcal{M} = \text{im} \phi$. Let $s : \Theta \to \mathbb{R}$ be another function, called a **parameter**. We say the parameter $s$ is

- **identifiable** from $\phi$ if for all $\theta, \theta'$ such that $\phi(\theta) = \phi(\theta')$, $s(\theta) = s(\theta')$;
- **generically identifiable** from $\phi$ if there is a dense open subset $U \subseteq \Theta$ such that $s$ is identifiable from $\phi$ on $U$;
- **rationally identifiable** from $\phi$ if there is a rational function $\psi : N \to \mathbb{R}$ such that $\psi \circ \phi(\theta) = s(\theta)$ almost everywhere;
- **locally identifiable** from $\phi$ if for almost all $\theta \in \Theta$ the set $\{s(\theta') : \phi(\theta) = \phi(\theta')\}$ is finite;
- **non-identifiable** from $\phi$ if there exists a $\theta, \theta' \in \Theta$ such that $\phi(\theta) = \phi(\theta')$ but $s(\theta) \neq s(\theta')$; and
- **generically non-identifiable** from $\phi$ if for almost all $\theta \in \Theta$, $|\{s(\theta') : \phi(\theta) = \phi(\theta')\}| = \infty$.

**Example 16.1.6 (Instrumental Variables).** Consider the instrumental variables model from Example 14.2.15 on the graph with a bidirected edge. The parametrization for this model has the form:
16.1. Tools for Testing Identifiability

\begin{equation}
\begin{pmatrix}
\sigma_{11} & \sigma_{12} & \sigma_{13} \\
\sigma_{12} & \sigma_{22} & \sigma_{23} \\
\sigma_{13} & \sigma_{23} & \sigma_{33}
\end{pmatrix} = \\
\begin{pmatrix}
1 & 0 & 0 \\
\lambda_{12} & 1 & 0 \\
\lambda_{12}\lambda_{23} & \lambda_{23} & 1
\end{pmatrix}
\begin{pmatrix}
\omega_{11} & 0 & 0 \\
0 & \omega_{22} & \omega_{23} \\
0 & \omega_{23} & \omega_{33}
\end{pmatrix}
\begin{pmatrix}
1 & \lambda_{12} & \lambda_{12}\lambda_{23} \\
0 & 1 & \lambda_{23} \\
0 & 0 & 1
\end{pmatrix}
\end{equation}

Note that we only display the upper entries to save space; since \(\Sigma\) is symmetric this is no loss of information. Recall that in the main use of this model, the desire is to uncover the effect of variable \(X_2\) on variable \(X_3\), that is, recover the parameter \(\lambda_{23}\). It is easy to see that

\[
\lambda_{23} = \frac{\sigma_{13}}{\sigma_{12}} = \frac{\omega_{11}\lambda_{12}\lambda_{23}}{\omega_{11}\lambda_{12}}
\]

in which case we see that \(\lambda_{23}\) is generically identifiable (since \(\sigma_{12}\) could be zero, this formula is not defined everywhere). On the other hand \(\lambda_{12}\) is globally identifiable by the formula \(\lambda_{12} = \frac{\sigma_{12}}{\sigma_{11}}\) which is defined everywhere since \(\sigma_{11} > 0\). Note that in the case that \(\lambda_{23}\) is not identified, namely when \(\lambda_{12} = 0\), the model fails to work in the desired manner. Namely, the instrumental variable \(X_1\) is supposed to be chosen in a manner so that it does directly affect \(X_2\) (i.e. \(\lambda_{12} \neq 0\)) so we can use this effect to uncover the effect of \(X_2\) on \(X_3\).

There are a number of methods for directly checking identifiability of a statistical model, or for checking identifiability of individual parameters. Perhaps the simplest first approach is to check the dimension of the image of \(\Theta\). If \(\dim \text{im} \phi = \dim \Theta\) then, since the map \(\phi\) is a rational map, the number of preimages of a generic point has a fixed finite value (over the complex numbers). Checking the dimension is easy to do by calculating the rank of the Jacobian of \(\phi\).

**Proposition 16.1.7.** Let \(\Theta \subseteq \mathbb{R}^d\) with \(\dim \Theta = d\) and suppose that \(\phi\) is a rational map. Then \(\dim \text{im} \phi\) is equal to the rank of the Jacobian matrix evaluated at a generic point:

\[
J(\phi) = \begin{pmatrix}
\frac{\partial \phi_1}{\partial \theta_1} & \cdots & \frac{\partial \phi_1}{\partial \theta_d} \\
\vdots & \ddots & \vdots \\
\frac{\partial \phi_s}{\partial \theta_1} & \cdots & \frac{\partial \phi_s}{\partial \theta_d}
\end{pmatrix}.
\]

In particular, the parameter vector \(\theta\) is locally identifiable if \(\text{rank} J(\phi) = d\) and the parameter vector is generically non-identifiable if \(\text{rank} J(\phi) < d\).
16. Identifiability

Proof Sketch. The fact that the dimension of the image is equal to the rank of the Jacobian at a generic point can be seen in an algebraic context, for instance, in [Har77, Thm I.5.1] (applied to the graph of the morphism $\phi$). The fact that the fibers of a map with full rank Jacobian are finite follows from an application of Noether normalization.

Proposition 16.1.7 can only tell us about local identifiability or generic non-identifiability. In practice, we would like to prove results about global identifiability or generic identifiability, as these are the most desirable identifiability properties for a model to possess. Algebraic geometry can be used to test for these conditions. A prototypical result of this type is Proposition 16.1.8. For a given map $\phi$ and parameter $s$, let $\tilde{\phi}$ be the augmented map $\tilde{\phi} : \Theta \to \mathbb{R}^{m+1}$, $\theta \mapsto (s(\theta), \phi(\theta))$. We denote the coordinates on $\mathbb{R}^{m+1}$ by $q, p_1, \ldots, p_m$.

Proposition 16.1.8. Let $\tilde{\phi} : \Theta \to \mathbb{R}^{m+1}$ be a rational map. Suppose that $g(q, p) \in I(\tilde{\phi}(\Theta)) \subseteq \mathbb{R}[q, p]$ is a polynomial such that $q$ appears in this polynomial, $g(q, p) = \sum_{i=0}^{d} g_i(p) q^i$ and $g_d(p)$ does not belong to $I(\tilde{\phi}(\Theta))$.

1. If $g$ is linear in $q$, $g = g_1(p)q - g_0(p)$ then $s$ is generically identifiable by the rational formula $s = \frac{g_0(p)}{g_1(p)}$. If, in addition, $g_1(p) \neq 0$ for $p \in \phi(\Theta)$ then $s$ is globally identifiable.

2. If $g$ has higher degree $d$ in $q$, then $s$ may or may not be generically identifiable. Generically, there are at most $d$ possible choices for the parameter $s(\theta)$ given $\phi(\theta)$.

3. If no such polynomial $g$ exists then the parameter $s$ is not generically identifiable.

Proof. If $g(q, p) \in I(\tilde{\phi}(\Theta))$ then $g(q, p)$ evaluates to zero for every choice of $\theta \in \Theta$. If there is a linear polynomial in $I(\tilde{\phi}(\Theta))$, this means that $g_1(\phi(\theta))s(\theta) - g_0(\phi(\theta)) = 0$. Since $g_1 \notin I(\phi(\Theta))$, we can solve for $s(\theta)$ as indicated.

If $g(q, p)$ has degree $d$ in $q$, we can solve an equation of degree $d$ to find $s(\theta)$. This may mean the parameter is locally identifiable, but it might also mean the parameter is globally identifiable. For instance, if $g(q, p) = q^d - g(p)$ and we know that $s(\theta)$ must be a positive real, then the existence of this polynomial would imply global identifiability.

Finally, if there is no such polynomial then generically there is no constraint on $s(\theta)$ given $\phi(\theta)$ so the parameter is not generically identifiable.

Note that Gröbner basis computations can be used to verify the existence or nonexistence of polynomials in $I(\tilde{\phi}(\Theta))$ of the indicated type.
16.1. Tools for Testing Identifiability

Proposition 16.1.9. Let $G$ be a reduced Gröbner basis for $I(\hat{\phi}(\Theta)) \subseteq \mathbb{R}[q,p]$ with respect to an elimination ordering such that $q > p_i$ for all $i$. Suppose that there is a polynomial $g(q,p) = \sum_{i=0}^{d} g_i(p)q^i \in I(\hat{\phi}(\Theta))$ with $g_d(p) \notin I(\phi(\Theta))$ that has nonzero degree in $q$. Then a polynomial of lowest nonzero degree in $q$ of this form appears in $G$. If no such polynomials exist in $I(\phi(\Theta))$, then $G$ does not contain any polynomials involving the indeterminate $q$.

Proof. Assume that $g(q,p) \in I(\hat{\phi}(\Theta))$ is a polynomial of lowest nonzero degree in $q$ of the form $g(q,p) = \sum_{i=0}^{d} g_i(p)q^i \in I(\hat{\phi}(\Theta))$ with $g_d(p) \notin I(\phi(\Theta))$. However, suppose there is no polynomial of that degree in $q$ in the reduced Gröbner basis $G$. Since $g_d(p) \notin I(\phi(\Theta))$, we can reduce $g(q,p)$ with respect to $G \cap \mathbb{R}[p]$ (which by the elimination property is a Gröbner basis for $I(\phi(\Theta))$). The resulting polynomial $\tilde{g}(q,p)$ will have the property that its leading term is not divisible by the leading term of any of polynomial in $G \cap \mathbb{R}[p]$. But then the leading term of $\tilde{g}(q,p)$ is divisible by $q_d$ by the elimination property. Since $G$ is a Gröbner basis, this leading term is divisible by some leading term of a polynomial in $G$. The term that divides this must be a polynomial of the form of Proposition 16.1.8, and must have degree in $q$ less than or equal to $d$. Since $d$ is the smallest possible degree for such a polynomial, we see that $G$ must contain a polynomial of the desired type. □

Similar reasoning can give qualitative information about the number of possible preimages over the complex numbers when we analyze the entire map at once.

Proposition 16.1.10. Consider a polynomial map $\phi : \mathbb{C}^d \to \mathbb{C}^r$. Let $I = \langle p_1 - \phi_1(\theta), \ldots, p_r - \phi_r(\theta) \rangle$ be the vanishing ideal of the graph of $\phi$. Let $G$ be a Gröbner basis for $I$ with respect to an elimination ordering such that $\theta_i > p_j$ for all $i$ and $j$. Let $M \subseteq \mathbb{C}[\theta]$ be the monomial ideal generated by all monomials $\theta^u$ such that $v \neq 0$ and $p^v \theta^u$ is the leading term of some polynomial in $G$. Then the number of complex preimages of a generic point in $\phi(\Theta)$ equals the number of standard monomials of $M$.

Recall that the standard monomials of a monomial ideal $M$ are the monomials that are not in $M$.

Proof. Let $J$ be the elimination ideal $I \cap \mathbb{C}[p]$, which is the vanishing ideal of the image of $\phi$. Let $K = K(\mathbb{C}[p]/I)$ be the fraction field of the coordinate ring of the image of $\phi$. If we consider the ideal

$$\tilde{I} = \langle p_1 - \phi_1(\theta), \ldots, p_r - \phi_r(\theta) \rangle \subseteq K[\theta]$$

then the number of standard monomials of $\tilde{I}$ with respect to any term order on the $\theta$ indeterminates equals the generic number of preimages for a point $p$ in the model.
The point of the proposition is that this initial ideal can be calculated by computing a Gröbner basis in the ring \( \mathbb{C}[p, \theta]\) rather than in the ring \( \mathbb{K}[\theta]\). Indeed, any polynomial \( f \in \mathcal{G} \subseteq \mathbb{C}[p, \theta]\) also yields a polynomial \( \tilde{f} \in \mathbb{K}[\theta]\). Since we have chosen an elimination term order on \( \mathbb{C}[p, \theta]\), the leading term of the highest coefficient appearing in \( f \), is not in the elimination ideal \( J \). This means that the leading monomial \( p^v \theta^u \) of \( f \) yields the leading monomial \( \theta^u \) of \( \tilde{I} \). Conversely if \( \tilde{f} \) is a nonzero polynomial of \( \tilde{I} \) with leading monomial \( \theta^u \), then by clearing denominators we get a nonzero polynomial of \( I \) with leading monomial \( p^v \theta^u \) for some \( v \). This shows that the monomial ideal \( M \) is the leading term ideal of \( \tilde{I} \).

**Example 16.1.11** (Identifiability for 1-factor Model). Consider the 1-factor model with \( m = 3 \). The following Macaulay 2 code computes the initial ideal with respect to the indicated lexicographic term order:

```plaintext
S = QQ[w1,w2,w3,l1,l2,l3,s11,s12,s13,s22,s23,s33,
       MonomialOrder => Lex];
I = ideal(s11 - w1 - l1^2, s12 - l1*l2, s13 - l1*l3,
             s22 - w2 - l2^2, s23 - l2*l3, s33 - w3 - l3^2)
leadTerm I
```

and the resulting initial ideal is

\[
\langle \lambda_3^2 \sigma_{12}, \lambda_2 \sigma_{13}, \lambda_2 \lambda_3, \lambda_1 \lambda_3, \lambda_1 \lambda_2, \omega_1, \omega_2, \omega_3 \rangle
\]

and the monomial ideal \( M \) is

\[
\langle \lambda_3^2, \lambda_2, \lambda_1, \omega_1, \omega_2, \omega_3 \rangle
\]

from which we deduce that a generic point in the image has 2 complex preimages. This confirms our calculations from Example 16.1.4.

Beyond identifiability of numerical parameters of models, as we have seen up to this point, it is also of interest in applications to decide the identifiability of certain discrete parameters of models. For example, we would like to know that two different phylogenetic trees \( T_1 \neq T_2 \) cannot produce the same probability distributions. So global identifiability of tree parameters would mean that \( M_{T_1} \cap M_{T_2} = \emptyset \). Generic identifiability of tree parameters means that

\[
\text{dim}(M_{T_1} \cap M_{T_2}) < \min(\text{dim} M_{T_1}, \text{dim} M_{T_2}).
\]

For graphical models we might desire that two different graphs \( G_1 \) and \( G_2 \) should yield different models, i.e. \( M_{G_1} \neq M_{G_2} \). This is also called model distinguishability. Both the problem of discrete parameter identifiability and model distinguishability can be addressed by using algebraic geometry techniques. A prototypical algebraic tool for this problem is the following:
Proposition 16.1.12. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two algebraic exponential families which sit inside the same natural parameter space of an exponential family. If there exists a polynomial $f \in I(\mathcal{M}_1)$ such that $f \notin I(\mathcal{M}_2)$ then $\mathcal{M}_1 \neq \mathcal{M}_2$. If $\mathcal{M}_1$ and $\mathcal{M}_2$ are both irreducible (e.g. they are parametrized models) and there exists polynomials $f_1$ and $f_2$ such that

$$f_1 \in I(\mathcal{M}_1) \setminus I(\mathcal{M}_2) \quad \text{and} \quad f_2 \in I(\mathcal{M}_2) \setminus I(\mathcal{M}_1)$$

then $\dim(\mathcal{M}_1 \cap \mathcal{M}_2) < \min(\dim \mathcal{M}_1, \dim \mathcal{M}_2)$.

Proof. Since $\mathcal{M}_1$ and $\mathcal{M}_2$ are both irreducible, the ideals $I(\mathcal{M}_1)$ and $I(\mathcal{M}_2)$ are prime ideals. Then if $\langle f_1 \rangle + I(\mathcal{M}_2)$ yields a variety strictly contained in $V(I(\mathcal{M}_2))$, in particular, it must have smaller dimension than $\mathcal{M}_2$. Since $\mathcal{M}_1 \cap \mathcal{M}_2 \subseteq V(\langle f_1 \rangle + I(\mathcal{M}_2))$ this shows that $\dim(\mathcal{M}_1 \cap \mathcal{M}_2) < \dim \mathcal{M}_2$. A similar argument shows the necessary inequality for $\mathcal{M}_1$ to complete the proof. □

Proposition 16.1.12 is a standard tool in proving identifiability results of discrete parameters, for instance, in proofs of identifiability of tree parameters in phylogenetic models and mixture models [APRS11, AR06, LS15, RS12]. It also arises in the well-known result in the theory of graphical models on when two directed acyclic graphs give the same graphical model (Theorem 13.1.13). Here is a simple example of how this method might be applied in phylogenetics.

Theorem 16.1.13. The unrooted tree parameter in the phylogenetic model under the general Markov model for binary trees is generically identifiable.

Proof. It suffices to show that for any pair of unrooted binary trees $T_1$ and $T_2$ on the same number of leaves, there exists a polynomials $f_1$ and $f_2$ such that

$$f_1 \in I(\mathcal{M}_{T_1}) \setminus I(\mathcal{M}_{T_2}) \quad \text{and} \quad f_2 \in I(\mathcal{M}_{T_2}) \setminus I(\mathcal{M}_{T_1}).$$

Since $T_1$ and $T_2$ are binary trees, there exists splits $A_1|B_1$ and $A_2|B_2$ such that $A_1|B_1 \in \Sigma(T_1) \setminus \Sigma(T_2)$ and $A_2|B_2 \in \Sigma(T_2) \setminus \Sigma(T_1)$. Then Proposition 15.4.4 implies that we can choose $f_i$ among the $(k+1) \times (k+1)$ minors of Flat$_{A_i|B_i}(P)$. □

16.2. Linear Structural Equation Models

Among the most well-studied identifiability problems in statistics is the identifiability of parameters in linear structural equation models. This problem originates in the work of Sewell Wright [Wri34] where SEMs were first introduced, and has been much studied in the econometrics literature [Fis66]. Still, a complete classification of when linear SEMs models are generically
identifiable, or necessary and sufficient conditions for when individual parameters are identifiable, remains elusive. In this section we highlight some of the main results in this area.

Let \( G = (V, B, D) \) be a mixed graph with \( m = |V| \) vertices. Recall that the linear structural equation model associated to the mixed graph \( G \) consists of all positive definite matrices of the form

\[
\Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}
\]

where \( \Omega \in PD(B) \) and \( \Lambda \in \mathbb{R}^D \) (see Section 14.2). Identifiability analysis of the linear structural equation model concerns whether or not the map \( G : PD(B) \times \mathbb{R}^D \rightarrow PD_m \) is one-to-one, or fails to be one-to-one under the various settings of Definition 16.1.1. One case where this is completely understood is the case of global identifiability which we explain now.

**Theorem 16.2.1.** [DFS11] Let \( G = (V, B, D) \) be a mixed graph. The parameters of the linear structural equation model associated to \( G \) are globally identifiable if and only if

1. The graph \( G \) has no directed cycle and
2. There is no mixed subgraph \( H = (V', B', D') \) with \( |V'| > 1 \) such that the subgraph of \( H \) on the bidirected edges is connected and the subgraph of \( H \) on the directed edges forms a converging arborescence.

A **converging arborescence** is a directed graph whose underlying undirected graph is a tree and has a unique sink vertex.

**Example 16.2.2** (Globally Identifiable and Nonidentifiable SEMs). Consider the two mixed graphs in Figure 16.2.1. The first mixed graph yields

\[\text{Figure 16.2.1.} \quad \text{The graph on the top yields a globally identifiable linear structural equation model while the graph on the bottom does not.}\]
16.2. Linear Structural Equation Models

a globally identifiable SEM. This follows from Theorem 16.2.1 since there are no directed cycles and the only possible set of vertices \( V' = \{2, 4\} \) that yields \( (V', B') \) connected, has no directed edges, so is not a converging arborescence.

On the other hand, the second mixed graph does not yield a globally identifiable SEM. The set of vertices \( V' = V \) already yields a connected graph on the bidirected edges and the set of directed edges forms a converging arborescence converging to node 6.

We do not give a complete proof of Theorem 16.2.1 here, but just show that condition (1) is necessary. Note that since we are concerned with global identifiability we always have the following useful proposition.

**Proposition 16.2.3.** Let \( G \) be a mixed graph and \( H \) a subgraph. If the linear structural equation model associated to \( G \) is globally identifiable, so is the linear structural equation model associated to \( H \).

**Proof.** Passing to a subgraph amounts to setting all parameters corresponding to deleted edges equal to zero. Since global identifiability means identifiability holds for all parameter choices, it must hold when we set some parameters to zero. \( \Box \)

Hence to show that the two conditions of Theorem 16.2.1 are necessary, it suffices to show that directed cycles are not globally identifiable, and graphs that fail to satisfy the second condition are not globally identifiable.

**Proposition 16.2.4.** Suppose that \( G = (V, D) \) is a directed cycle of length \( \geq 3 \). Then the map \( \phi_G : \mathbb{R}^D \to PD_m \) is generically 2-to-1.

**Proof.** Without loss of generality, we can take \( G \) to be the directed cycle with edges 1 \( \to \) 2, 2 \( \to \) 3, \( \ldots \), \( m \to 1 \). Note that the determinant of \( (I - \Lambda) \) in this case is \( 1 + (-1)^{m-1} \prod_{i=1}^{m} \lambda_{i+1} \). We assume that this quantity is not equal to zero, and that all \( \lambda_{i+1} \neq 0 \). Since we are considering generic behavior, we will use the useful trick that we can parametrize the concentration matrix, rather than the covariance matrix. This has the parametrization:

\[
K = (I - \Lambda)^T \Delta (I - \Lambda)
\]

where \( \Delta = \Omega^{-1} \) is the diagonal matrix with diagonal entries \( \delta_{ii} = \omega_{ii}^{-1} \). Expanding this parametrization explicitly gives

\[
K = \begin{pmatrix}
\delta_{11} + \delta_{22} \lambda_{12}^2 & -\delta_{22} \lambda_{12} & 0 & \cdots \\
-\delta_{22} \lambda_{12} & \delta_{22} + \delta_{33} \lambda_{23}^2 & -\delta_{33} \lambda_{23} & \cdots \\
0 & -\delta_{33} \lambda_{23} & \delta_{33} + \delta_{44} \lambda_{34}^2 & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{pmatrix}
\]

where \( \delta_{ij} = \omega_{ij}^{-1} \).
This tridiagonal matrix has
\[ K_{ii} = \delta_{ii} + \delta_{i+1,i+1} \lambda_{i,i+1}^2, \quad K_{i,i+1} = -\delta_{i+1,i+1} \lambda_{i,i+1}. \]
We can solve for \( \lambda_{i,i+1} \) in terms of \( K_{i,i+1} \) and \( \delta_{i+1,i+1} \), which produces the equation system
\[ K_{ii} = \delta_{ii} + \frac{K_{i,i+1}^2}{\delta_{i+1,i+1}}. \]
Linearly eliminating \( \delta_{ii} \) and plugging into the next equation for this system eventually produces a quadratic equation in \( \delta_{mm} \). Since we know that one of the roots for this equation is real the other root must be as well. This produces all real roots for all the coefficients with two preimages over the reals, generically. We must show that the second pair \( (\Delta', \Lambda') \) also corresponds to model parameters, i.e. \( \Delta' \) is positive definite. To see this note that the second matrix \( \Lambda' \) will also yield \( (I - \Lambda')^{-1} \) invertible generically. Since \( K \) is positive definite, this means that \( (I - \Lambda')^{-T}K(I - \Lambda')^{-1} = \Delta' \) is also positive definite. \( \square \)

To show that the second condition of Theorem 16.2.1 is necessary requires a different argument. For graphs that have a subgraph connected in \( B' \) and a converging arborescence in \( D' \) it can be the case that the map \( \phi_G \) is generically injective, so that the model is generically identifiable. So we need to find specific parameter values \( \Omega, \Lambda \) such that \( \phi_G(\Omega, \Lambda) \) has multiple preimages. In fact, it always happens in these cases that there are fibers that are not zero dimensional. See [DFS11] for the details.

The simplest example of the bad substructure in the second condition of Theorem 16.2.1 is a pair of vertices \( i, j \) such that we have both edges \( i \rightarrow j \) and \( i \leftrightarrow j \). By passing to the subgraph with only those two edges, we see that this model could not be globally identifiable. Remarkably, if we just rule out this one subgraph and assume we have no directed cycles, we have generic identifiability of the associated linear SEM. Graphs with no pair of vertices with both \( i \rightarrow j \in D \) and \( i \leftrightarrow j \in B \) are called simple mixed graphs or bow-free graphs.

**Theorem 16.2.5.** [BP02] Let \( G = (V, B, D) \) be a simple directed acyclic mixed graph. Then \( \phi_G \) is generically injective.

**Proof.** We assume that the ordering of vertices matches the acyclic structure of the graph (so, \( i \rightarrow j \in D \) implies that \( i < j \)). We proceed by induction on the number of vertices. The upper triangular nature of the matrix \( (I - \Lambda)^{-1} \) implies that the submatrix \( \Sigma_{[m-1],[m-1]} \) has the following factorization:
\[ \Sigma_{[m-1],[m-1]} = (I - \Lambda_{[m-1],[m-1]})^{-T}\Omega_{[m-1],[m-1]}(I - \Lambda_{[m-1],[m-1]})^{-1}. \]
By induction on the number of vertices, we can assume that all parameters associated to the model on vertices in \([m-1]\) are identifiable.

So consider the vertices \(N(m)\) which are the neighbors of \(m\) (either a directed or bidirected edge from \(i \in N(m)\) to \(m\)). Every trek from a vertex in \(N(m)\) to \(m\) has a single edge involving one of the yet underdetermined parameters which connects to vertex \(m\). Let \(P\) be the length \(|N(m)|\) vector consisting of all parameters corresponding to either the directed or bidirected edges connecting any \(i \in N(m)\) to \(m\). Since the graph is simple we can write a square system

\[
\Sigma_{N(m),m} = A(\Omega_{[m-1],[m-1]}, \Lambda_{[m-1],[m-1]}))P
\]

where \(A(\Omega_{[m-1],[m-1]}, \Lambda_{[m-1],[m-1]}))\) is a square matrix whose entries are polynomial functions of the parameters whose values we know by induction. If we can show that \(A(\Omega_{[m-1],[m-1]}, \Lambda_{[m-1],[m-1]}))\) is generically invertible, we will identify all parameters corresponding to edges pointing to \(m\). However, if we choose \(\Omega_{[m-1],[m-1]} = I\) and \(\Lambda_{[m-1],[m-1]} = 0\), \(A(\Omega_{[m-1],[m-1]}, \Lambda_{[m-1],[m-1]}))\) will be the identity matrix, which is clearly invertible. Requiring a matrix to be invertible is an open condition, so \(A(\Omega_{[m-1],[m-1]}, \Lambda_{[m-1],[m-1]}))\) will be invertible for generic choices of the parameters.

Finally, the entry \(\omega_{m,m}\) is identifiable as it appears in only the entry

\[
\sigma_{m,m} = \omega_{m,m} + f(\Omega, \Lambda)
\]

where \(f\) is a polynomial only in identifiable entries. \(\square\)

**Example 16.2.6** (Bow-Free Graph). The second graph in Figure 16.2.1 yields a generically identifiable, but not globally identifiable SEM model, since the graph is bow-free but does not satisfy the condition of Theorem 16.2.1.

In contrast to this result, the basic instrumental variables model is generically identifiable (Example 16.1.6), though the graph is not simple. It remains a major open problem in the study of identifiability to find a complete characterization of those mixed graphs for which the parametrization \(\phi_G\) is generically injective. Most partial results of this type attempt to follow the strategy of the proof of Theorem 16.2.5, by recursively solving for parameters in the model by writing linear equation systems whose coefficients involve either parameters that have already been solved for or entries of the matrix \(\Sigma\). One strong theorem of this type is the half-trek criterion [FDD12]. That paper also contains a similar necessary condition involving half-treks that a graph must satisfy to be identifiable. Further recent results extend the half-trek criterion by considering certain graph decompositions of the mixed graph \(G\) [DW16]. See [CTP14] for related work. To state the results on the half-trek criterion requires a number of definitions.
Definition 16.2.7. Let $G = (V, B, D)$ be a mixed graph. Let $\text{sib}(v) = \{i \in V : i \leftrightarrow v \in V\}$ be the siblings of node $v \in V$. A half trek is a trek with left side of cardinality one. Let $\text{htr}(v)$ be the subset of $V \setminus (\{v\} \cup \text{sib}(v))$ of vertices reachable from $v$ via a half-trek.

Definition 16.2.8. A set of nodes $Y \subseteq V$ satisfies the half-trek criterion with respect to node $v \in V$ if

1. $|Y| = |\text{pa}(v)|$,
2. $Y \cap (\{v\} \cup \text{sib}(v)) = \emptyset$, and
3. there is a system of half-treks with no sided intersection from $Y$ to $\text{pa}(v)$.

Note that if $\text{pa}(v) = \emptyset$ then $Y = \emptyset$ will satisfy the half-trek criterion with respect to $v$.

Theorem 16.2.9. [FDD12] Let $(Y_v : v \in V)$ be a family of subsets of the vertex set $V$ of the mixed graph $G$. If, for each node $v$, the set $Y_v$ satisfies the half-trek criterion with respect to $v$, and there is a total ordering $\prec$ of the vertex set $V$ such that $w \prec v$ whenever $w \in Y_v \cap \text{htr}(v)$, then $\phi_G$ is generically injective and has a rational inverse.

We do not attempt to prove Theorem 16.2.9 here but illustrate it with an example from [FDD12].

Example 16.2.10 (Identifiability Via Half-Treks). Consider the graph in Figure 16.2.2. This graph is generically identifiable via the half trek criterion. Take

$$Y_1 = \emptyset, \quad Y_2 = \{5\}, \quad Y_3 = \{2\}, \quad Y_4 = \{2\}, \quad Y_5 = \{3\}.$$ 

Each $Y_v$ satisfies the half-trek criterion with respect to $v$. For example, for $v = 3, 2 \rightarrow 3$ is a half-trek from a nonsibling to 3. Taking the total order on the vertices to be the standard order $1 \prec 2 \prec 3 \prec 4 \prec 5$ we see that the conditions on the sets $\text{htr}(v)$ are also satisfied. For example, for $v = 3$, $\text{htr}(3) = \{4, 5\}$ so $Y_3 \cap \text{htr}(3) = \emptyset$. 

Figure 16.2.2. A graph whose linear SEM is proven identifiable by the half trek criterion.
16.3. Tensor Methods

For hidden variable statistical models with discrete random variables, a powerful tool for proving identifiability results comes from methods related to the tensor rank of three way tensors. The main theorem in this area is Kruskal’s theorem \([\text{Kru76, Kru77}]\). The methods for extending the technique to complicated hidden variable models appears in \([\text{AMR09}]\). In this section, we will explore the applications of Kruskal’s theorem to proving identifiability for hidden variable graphical models and mixture models.

The starting point for Kruskal’s theorem is the comparison of two mixture models, \(\text{Mixt}_k(M_{X_1 \parallel X_2})\) and \(\text{Mixt}_k(M_{X_1 \parallel X_2 \parallel X_3})\). As we saw in Example 16.1.3, the first model is always generically nonidentifiable for any \(k > 1\). Perhaps the first natural guess is that \(\text{Mixt}_k(M_{X_1 \parallel X_2 \parallel X_3})\) should be generically nonidentifiable as well since it contains the first model after marginalizing. However this intuition turns out to be wrong, and this three-factor mixture turns out to be generically identifiable if \(k\) satisfies a simple bound in terms of the number of states of \(X_1, X_2, X_3\). Kruskal’s theorem can be stated purely in statistical language but it can also be useful to state it in the language of tensors.

Let \(a_1, \ldots, a_k, b_1, \ldots, b_k, c_1, \ldots, c_k \in \mathbb{K}^{r_1}\), \(a_1, \ldots, a_k, b_1, \ldots, b_k, c_1, \ldots, c_k \in \mathbb{K}^{r_2}\), and \(a_1, \ldots, a_k, b_1, \ldots, b_k, c_1, \ldots, c_k \in \mathbb{K}^{r_3}\). Arrange these vectors into matrices \(A \in \mathbb{K}^{r_1 \times k}\), \(B \in \mathbb{K}^{r_2 \times k}\), and \(C \in \mathbb{K}^{r_3 \times k}\) whose columns are \(a_1, \ldots, a_k, b_1, \ldots, b_k, c_1, \ldots, c_k\), respectively. The \((i, j)\) entry of \(A\) is \(a_{ij}\) the \(i\)th entry of \(a_j\), and similarly for \(B\) and \(C\).

**Definition 16.3.1.** Let \(A \in \mathbb{K}^{r \times k}\) be a matrix. The **Kruskal rank** of \(A\), denoted \(\text{rank}_k(A)\) is the largest number \(l\) such that any \(l\) columns of \(A\) are linearly independent.

We can form the tensor

\[
M = \sum_{j=1}^{k} a_j \otimes b_j \otimes c_j \in \mathbb{K}^{r_1} \otimes \mathbb{K}^{r_2} \otimes \mathbb{K}^{r_3}.
\]

The tensor product space \(\mathbb{K}^{r_1} \otimes \mathbb{K}^{r_2} \otimes \mathbb{K}^{r_3} \cong \mathbb{K}^{r_1 \times r_2 \times r_3}\) is the vector space of \(r_1 \times r_2 \times r_3\) arrays with entries in \(\mathbb{K}\). Each tensor of the form \(a_j \otimes b_j \otimes c_j\) is called a **rank one tensor** or **decomposable tensor**. A tensor \(M \in \mathbb{K}^{r_1} \otimes \mathbb{K}^{r_2} \otimes \mathbb{K}^{r_3}\) has rank \(k\) if it can be written as in Equation 16.3.1 for some \(a_1, \ldots, a_k, b_1, \ldots, b_k, c_1, \ldots, c_k \in \mathbb{K}^{r_1}\), \(b_1, \ldots, b_k, c_1, \ldots, c_k \in \mathbb{K}^{r_2}\), and \(c_1, \ldots, c_k, c_1, \ldots, c_k \in \mathbb{K}^{r_3}\) but cannot be written as a sum of fewer rank one tensors. Kruskal’s theorem concerns the uniqueness of the decomposition of a tensor into rank one tensors.

**Theorem 16.3.2** (Kruskal). Let \(M \in \mathbb{K}^{r_1} \otimes \mathbb{K}^{r_2} \otimes \mathbb{K}^{r_3}\) be a tensor and

\[
M = \sum_{j=1}^{k} a_j \otimes b_j \otimes c_j \in \mathbb{K}^{r_1} \otimes \mathbb{K}^{r_2} \otimes \mathbb{K}^{r_3},
\]
Let $M = \sum_{j=1}^{k} a'_j \otimes b'_j \otimes c'_j \in K^{r_1} \otimes K^{r_2} \otimes K^{r_3}$ be two decompositions of $M$ into rank one tensors. Let $A, B, C$ be the matrices whose columns are $a_1, \ldots, a_k$, $b_1, \ldots, b_k$, and $c_1, \ldots, c_k$, respectively, and similarly for $A', B'$, and $C'$. If
\[ \text{rank}_K(A) + \text{rank}_K(B) + \text{rank}_K(C) \geq 2k + 2 \]
then there exists a permutation $\sigma \in S_k$, and nonzero $\lambda_1, \ldots, \lambda_k, \gamma_1, \ldots, \gamma_k \in K$ such that
\[ a_{\sigma(j)} = \lambda_j a'_j, \quad b_{\sigma(j)} = \gamma_j b'_j, \quad \text{and} \quad c_{\sigma(j)} = \lambda_j^{-1} \gamma_j^{-1} c'_j \quad \text{for all} \quad j \in [k]. \]

In particular, $M$ has rank $k$.

A useful compact notation for expressing Equation 16.3.1 is the triple product notation. We write $M = [A, B, C]$ to denote the rank one decomposition from (16.3.1). Theorem 16.3.2 can also be rephrased as saying that if $\text{rank}_K(A) + \text{rank}_K(B) + \text{rank}_K(C) \geq 2k + 2$ and $[A, B, C] = [A', B', C']$ there is a permutation matrix $P$ and invertible diagonal matrices $D_1, D_2, D_3$ with $D_1 D_2 D_3 = I$ such that
\[ D_1 PA = A', \quad D_2 PB = B', \quad \text{and} \quad D_3 PC = C'. \]

So Kruskal’s theorem guarantees that a given decomposition of tensors is as unique as possible provided that the Kruskal ranks of the given tensors are large enough, relative to the number of factors. This theorem was originally proven in [Kru76, Kru77] and simplified versions of the proof can be found in [Lan12, Rho10]. In this section, we explore applications of Kruskal’s theorem to the identifiability of hidden variable discrete models. A first application concerns the mixture model $\text{Mixt}^k(M_{X_1 \perp X_2 \perp X_3})$.

**Corollary 16.3.3.** Consider the mixture model $\text{Mixt}^k(M_{X_1 \perp X_2 \perp X_3})$ where $X_i$ has $r_i$ states, for $i = 1, 2, 3$. This model is generically identifiable up to relabeling of the states of the hidden variable if
\[ \min(r_1, k) + \min(r_2, k) + \min(r_3, k) \geq 2k + 2. \]

Note the phrase “up to relabeling of the states of the hidden variable” which appears in Corollary 16.3.3. This is a common feature in the identifiability of hidden variable models. Another interpretation of this result is that the model is locally identifiable and each point in the model has $k!$ preimages.
Proof of Corollary 16.3.3. Recall that \( \text{Mixt}^k(\mathcal{M}_{X_1 \perp \cdots \perp X_3}) \) can be parametrized as the set of all probability distributions in \( p \in \Delta_{r_1 r_2 r_3 - 1} \) with
\[
p_{11211} = \sum_{j=1}^{k} \pi_j \alpha_{i1j} \beta_{i2j} \gamma_{i3j}
\]
for some \( \pi \in \Delta_{k-1} \) and \( \alpha, \beta, \) and \( \gamma \) representing conditional distributions of \( X_1|Y, X_2|Y \) and \( X_3|Y \), respectively. In comparison to the parametrization of rank \( k \) tensors, we have extra restrictions on the matrices \( \alpha, \beta, \gamma \) (their columns sums are all equal to one) and we have the appearance of the extra \( \pi \). Let \( \Pi \) denote the diagonal matrix \( \text{diag}(\pi_1, \ldots, \pi_k) \). Let \( A = \alpha \Pi, B = \beta, C = \gamma \). The triple product \([A, B, C]\) gives the indicated probability distribution in the mixture model. For generic choices of the conditional distributions and the root distributions we have
\[
\text{rank}_K(A) + \text{rank}_K(B) + \text{rank}_K(C) = \min(r_1, k) + \min(r_2, k) + \min(r_3, k)
\]
so we are in the position to apply Kruskal’s theorem. Suppose that
\[
[A', B', C'] = [A, B, C].
\]
By Kruskal’s theorem, it must be the case that there is a permutation matrix \( P \) and invertible diagonal matrix \( D_1, D_2, D_3 \) with \( D_1 D_2 D_3 = I \) such that
\[
D_1 PA = A', \quad D_2 PB = B', \quad \text{and} \quad D_3 PC = C'.
\]
However, if we take \( B' \) and \( C' \) to give conditional probability distributions, this forces \( D_2 = I \) and \( D_3 = I \), which in turn forces \( D_1 = I \). So the matrices \( A, B, C, \) and \( A', B', C' \) only differ by simultaneous permutations of the columns. So up to this permutation, we can assume that \( A = A', B = B', \) and \( C = C' \). This means we recover \( \beta \) and \( \gamma \) up to permutation of states of the hidden variable. Lastly \( \pi \) and \( \alpha \) can be recovered from \( A \) since \( \pi \) is the vector of columns sums of \( A \), and then \( \alpha = \Pi \pi^{-1} \).

There are direct generalizations of Kruskal’s Theorem to higher order tensors. They prove essential uniqueness of the decomposition of tensors as sums of rank one tensors provided that the Kruskal ranks of certain matrices satisfy a linear condition in terms of the rank of the tensor. However, part of the power of Kruskal’s theorem for 3-way tensors is that it can be used to prove identifiability to much higher number of states of hidden variables if we are willing to drop the explicit check on identifiability that comes with calculating the Kruskal rank. For example:

Corollary 16.3.4. Let \( X_1, \ldots, X_m \) be discrete random variables where \( X_i \) has \( r_i \) states. Let \( \mathcal{M}_{X_1 \perp \cdots \perp X_m} \) denote the complete independence model. Let \( A|B|C \) be a tripartition of \([m]\) with no empty parts. Then the mixture
model $\text{Mixt}^k(M_{X_1, \ldots, X_m})$ is generically identifiable up to label swapping if
\[
\min(\prod_{a \in A} r_a, k) + \min(\prod_{b \in B} r_b, k) + \min(\prod_{c \in C} r_c, k) \geq 2k + 2.
\]

**Proof.** We would like to reduce to Corollary 16.3.3. To do this we “clump” the random variables into three blocks to produce three super random variables $Y_A$ with $\prod_{a \in A} r_a$ states, $Y_B$ with $\prod_{b \in B} r_b$ states, and $Y_C$ with $\prod_{c \in C} r_c$ states. Here we are flattening our $m$-way array into a 3-way array (akin to the flattening to matrices we saw in Chapter 15).

Clearly the Corollary is true for these three random variables and the resulting mixture model $\text{Mixt}^k(M_{Y_A, Y_B, Y_C})$. However, the resulting conditional distributions that are produced in this clumping procedure are not generic distributions, they themselves consist of rank 1 tensors. For example, the resulting conditional distribution of $Y_A|H = h$ is an $\prod_{a \in A} r_a$ rank 1 tensor. This means that we cannot apply Corollary 16.3.3 unless we can prove that generic parameters in the model $\text{Mixt}^k(M_{X_1, \ldots, X_m})$ map to the identifiable locus in $\text{Mixt}^k(M_{Y_A, Y_B, Y_C})$. At this point, however, we can appeal to the proof of Corollary 16.3.3. The resulting matrices whose columns are the conditional distributions obtained by the flattening will generically have full rank (hence, Kruskal rank will be $\min(\prod_{a \in A} r_a, k)$ for the first matrix, etc.). This follows because there are no linear relationships that hold on the set of rank 1 tensors, so we can always choose a generic set that are linearly independent. Then the proof of Corollary 16.3.3 implies we can recover all parameters in the model, up to renaming the hidden variables.

The “clumping” argument used in the Proof of Corollary 16.3.4 has been used to prove identifiability of a number of different models. See [AMR09] for numerous examples and [RS12] for an application to identifiability of phylogenetic mixture models. We close with another example from [ARSV15] illustrating how the method can be used in graphical models with a more complex graphical structure.

**Example 16.3.5 (Hidden Variable Graphical Model).** Consider the discrete directed graphical model in Figure 16.3.1 where all random variables are binary and variable $Y$ is assumed to be a hidden variable. We will show that this model is identifiable up to label swapping the two states of the hidden variable $Y$. The parametrization for this model takes the following form:
\[
\begin{align*}
p_{ij1234} &= \sum_{j=1}^{2} \pi_j a_{ij1} b_{ij2} c_{ij3} d_{ij4}
\end{align*}
\]
where $P(Y = j) = \pi_j$, $P(X_1 = i_1 \mid Y = j) = a_{i_1|j}$, etc., are the conditional distributions that go into the parametrization of the graphical model.

Fixing a value of $X_1 = i_1$ and regrouping terms yields the expression:

$$p_{i_1i_2i_3i_4} = 2 \sum_{j=1}^{2} (\pi_j a_{i_1|j} b_{i_2|j} c_{i_3|j} d_{i_4|j}).$$

For each fixed value of $i_1$, the resulting $2 \times 2 \times 2$ array that is being parameterized in this way is a rank $\leq 2$ tensor. Each of the matrices

$B^{i_1} = \begin{pmatrix} b_{1|1i_1} & b_{1|2i_1} \\ b_{2|1i_1} & b_{2|2i_1} \end{pmatrix}$, $C^{i_1} = \begin{pmatrix} c_{1|1i_1} & c_{1|2i_1} \\ c_{2|1i_1} & c_{2|2i_1} \end{pmatrix}$, $D = \begin{pmatrix} d_{1|1} & c_{1|2} \\ d_{2|1} & c_{2|2} \end{pmatrix}$

generically has full rank, and hence has Kruskal rank $2$. From the argument in the proof of Corollary 16.3.3, we can recover all of the entries in these matrices (up to label swapping) plus the quantities

$$\pi_1 a_{1|1}, \ \pi_2 a_{1|2}, \ \pi_1 a_{2|1}, \ \pi_2 a_{2|2}.$$ 

Since

$$\pi_1 a_{1|1} + \pi_1 a_{2|1} = \pi_1 \quad \text{and} \quad \pi_2 a_{1|2} + \pi_2 a_{2|2} = \pi_2$$

this allows us to recover all entries in the probability distributions that parametrize this model.

Corollaries to Kruskal’s theorem can also be restated directly in the language of algebraic geometry.

**Corollary 16.3.6.** Suppose that $\min(r_1, k) + \min(r_2, k) + \min(r_3, k) \geq 2k + 2$. Then a generic point on the secant variety $\text{Sec}_k(\text{Seg}(\mathbb{P}^{r_1-1} \times \mathbb{P}^{r_2-1} \times \mathbb{P}^{r_3-1}))$ lies on a unique secant $\mathbb{P}^{k-1}$ to $\text{Seg}(\mathbb{P}^{r_1-1} \times \mathbb{P}^{r_2-1} \times \mathbb{P}^{r_3-1})$.

While Corollary 16.3.6 gives a cleaner looking statement, the full version of Theorem 16.3.2 is significantly more powerful because it gives an explicitly checkable condition (the Kruskal rank) to test whether a given decomposition is unique. This is the tool that often allows the method to be applied to more complex statistical models. In fact, tools from algebraic geometry provide stronger statements than Corollary 16.3.6.
Theorem 16.3.7. [CO12] Let \( r_1 \geq r_2 \geq r_3 \). A generic point on the secant variety \( \text{Sec}^k(\text{Seg}(\mathbb{P}^{r_1-1} \times \mathbb{P}^{r_2-1} \times \mathbb{P}^{r_3-1})) \) lies on a unique secant \( \mathbb{P}^{k-1} \) to \( \text{Seg}(\mathbb{P}^{r_1-1} \times \mathbb{P}^{r_2-1} \times \mathbb{P}^{r_3-1}) \) provided that \( (r_2+1)(r_3+1)/16 \geq k \).

Note that Kruskal’s theorem gives a linear bound on the tensor rank that guarantees the uniqueness of tensor decomposition while Theorem 16.3.7 gives a quadratic bound on uniqueness of decomposition where \( r_1, r_2, r_3 \) have the same order of magnitude. However, Theorem 16.3.7 lacks any explicit condition that guarantees that a specific tensor lies in the identifiable locus, so at present it does not seem possible to apply this result to complex identifiability problems in algebraic statistics. It remains an interesting open problem in the theory of tensor rank to give any explicit checkable condition that will verify a tensor has a specific rank when that rank is large relative to the tensor dimensions.

16.4. State Space Models

Models based on differential equations are widely used throughout applied mathematics and nearly every field of science. For example, the governing equations for physical laws are usually systems of partial differential equations, and the time evolution of a biological system is usually represented by systems of ordinary differential equations. These systems often depend on unknown parameters that cannot be directly measured, and a large part of applied mathematics concerns developing methods for recovering these parameters from the observed data. The corresponding research areas and the mathematical techniques that are used are variously known as the theory of inverse problems, parameter estimation, or uncertainty quantification. A fundamental property that needs to be satisfied for a model to be useful is that the data observations need to be rich enough that it is actually possible to estimate the parameters from the data that is given. That is, the model parameters need to be identifiable given the observation. While the study of differential equation systems is far removed from the bulk of the statistical problems studied in this book, the techniques for testing identifiability for such systems often relies on the same algebraic geometric ideas that we have seen in previous sections of this chapter. For this reason, we include the material here. Our focus in this section is state space models, a special class of dynamical systems, that come from systems of ordinary differential equations.

Consider the state space model

\[
\dot{x}(t) = f(t, x(t), u(t), \theta), \quad y(t) = g(x(t), \theta).
\]

In this equation system, \( t \in \mathbb{R} \) denotes time, \( x(t) \in \mathbb{R}^n \) are the state variables, \( u(t) \in \mathbb{R}^n \) are the control variables or system inputs, \( y(t) \in \mathbb{R}^m \) are
the output variables or observations, and $\theta \in \mathbb{R}^d$ is the vector of model parameters. The goal of identifiability analysis for differential equations is to determine when the model parameters $\theta$ can be recovered from the trajectory of the input and output variables.

In the identifiability analysis of state space models, one often makes the distinction between structural identifiability and practical identifiability. Structural identifiability is a theoretical property of the model itself and concerns whether the model parameters can be determined knowing the trajectory of the input $u(t)$ and the output $y(t)$ exactly. Practical identifiability is a data specific property and concerns whether or not, in a model that is structurally identifiable, it is possible to estimate the model parameters with high certainty given a particular data set. Our focus here is on structural identifiability, the situation where algebraic techniques are most useful.

We address structural identifiability problems for state space models using the differential algebra approach [BSAD07, LG94]. In this setting, we assume that the functions $f$ and $g$ in the definition of the dynamical system (16.4.1) are polynomial or rational functions. Under certain conditions there will exist a vector-valued polynomial function $F$ such that $y, u$ and their derivatives satisfy:

\[(16.4.2)\quad F(y, \dot{y}, \ddot{y}, \ldots, u, \dot{u}, \ddot{u}, \ldots, \theta) = 0.\]

This differential equation system just in the inputs and outputs is called the input/output equation. When there is only one output variable, a basic result in differential algebra says that with a general choice of the input function $u(t)$ the coefficients of the differential monomials in the input/output equation can be determined from the trajectory of $u(t)$ and the output trajectory $y(t)$. These coefficients are themselves rational functions of the parameters $\theta$. Global identifiability for dynamical systems models reduces to the question of whether or not the coefficient mapping $c : \mathbb{R}^d \to \mathbb{R}^k$, $\theta \mapsto c(\theta)$ is a one-to-one function. The situation when there are more than one output function is more complicated, see [HOPY17].

**Example 16.4.1 (Two Compartment Model).** Consider the compartment model represented by the following system of linear ODEs:

\[
\begin{align*}
\dot{x}_1 &= -a_{21}x_1 + a_{12}x_2 + u_1 \\
\dot{x}_2 &= a_{21}x_1 + (-a_{12} - a_{02})x_2
\end{align*}
\]

with 1-dimensional output variable $y(t) = Vx_1(t)$. The state variable $x(t) = (x_1(t), x_2(t))^T \in \mathbb{R}^2$ represents the concentrations of some quantity in two compartments. Here parameters $a_{21}$ and $a_{12}$ represent the flow of material from compartment 1 to 2 and 2 to 1 respectively, $a_{02}$ represents the rate of flow out of the system through compartment 2, and $V$ is a parameter.
that shows that we only observe $x_1$ up to an unknown scale. The input $u(t) = (u_1(t), 0)$ is only added to compartment 1.

A natural setting where this simple model might be applied is as follows: Suppose we administer a drug to a person, and we want to know how much drug might be present in the body over time. Let $x_1$ represent the concentration of a drug in the blood and $x_2$ represents concentration of the same drug in the liver. The drug leaves the system through decomposition in the liver, and otherwise does not leave the system. The drug enters the system through the blood (e.g., by injection) represented by $u_1$ and its concentration can be measured through measurements from the blood $y$ which only give the concentration to an unknown constant.

To find the input/output equation in this case is simple linear algebra. We solve for $x_2$ in the first equation and plug the result in to the second, and then substitute $x_1 = y/V$. This yields the input/output equation:

$$\ddot{y} + (a_{21} + a_{12} + a_{02})\dot{y} + a_{21}a_{02}y = V\dot{u}_1 + V(a_{12} + a_{02})u_1.$$  

With generic input function $u(t)$ and initial conditions, this differential equation can be recovered from the input and output trajectories. Identifiability reduces to the question of whether the coefficient map $c : \mathbb{R}^4 \to \mathbb{R}^5$, $(a_{21}, a_{12}, a_{02}, V) \mapsto (1, a_{21} + a_{12} + a_{02}, a_{21}a_{02}, V, V(a_{12} + a_{02}))$ is one-to-one. In this particular example, it is not difficult to see that the map is rationally identifiable. Indeed, in terms of the coefficients $c_1, c_2, c_3, c_4$, we have that

$$V = c_3, \quad a_{21} = c_1 - \frac{c_4}{c_3}, \quad a_{02} = \frac{c_2c_3}{c_1c_3 - c_4}, \quad a_{12} = \frac{c_4}{c_3} - \frac{c_2c_3}{c_1c_3 - c_4}.$$  

In fact, the model is globally identifiable when we restrict to the biologically relevant parameter space where all parameters are positive. $\square$

Note that when we recover the input/output equation from data, we only know this equation up to scaling. To eliminate this ambiguity, we usually prefer to make a representation for the input/output equation that has one coefficient set to 1 by dividing through by one of the coefficient functions if necessary. Alternately, instead of viewing the coefficient map $c : \mathbb{R}^d \to \mathbb{R}^k$ we can view this as a map to projective space $c : \mathbb{R}^d \to \mathbb{P}^{k-1}$.

Compared with the statistical models whose identifiability we studied in previous sections, dynamical systems models have a two layer source of complexity in the study of their identifiability. First we must determine the input/output equation of the model, then we must analyze the injectivity of the coefficient map. The first problem is handled using techniques from differential algebra: the input/output equation can be computed using triangular set methods or differential Gröbner bases. These computations can
be extremely intensive and it can be difficult to determine the input/output equations except in relatively small models. It seems an interesting theoretical problem to determine classes of models where it is possible to explicitly calculate the input/output equation of the system. Having such classes would facilitate the analysis of the identifiability of such models and could possibly make it easier to estimate the parameters.

One general family of models where the input/output equations are easy to write down explicitly is the family of linear compartment models, generalizing Example 16.4.1. Let $G$ be a directed graph with vertex set $V$ and set of directed edges $E$. Each vertex $i \in V$ corresponds to a compartment in the model and an edge $j \to i$ represents a direct flow of material from compartment $j$ to $i$. Let $In, Out, Leak \subseteq V$ be three sets of compartments corresponding to the compartments with input, with output, and with a leak, respectively. Note that the sets $In, Out, Leak$ do not need to be disjoint. To each edge $j \to i$ we associate an independent parameter $a_{ij}$, the rate of flow from compartment $j$ to compartment $i$. To each leak node $i \in Leak$ we associate an independent parameter $a_{0i}$, the rate of flow from compartment $i$ out of the system.

We associate a matrix $A(G)$ to the graph and the set $Leak$ as follows:

$$A(G)_{ij} = \begin{cases} 
-a_{0i} - \sum_{k:j \to k \in E} a_{ki} & \text{if } i = j \text{ and } i \in Leak \\
-a_{0i} - \sum_{k:i \to k \in E} a_{ki} & \text{if } i = j \text{ and } i \notin Leak \\
 a_{ij} & \text{if } j \to i \in E \\
 0 & \text{otherwise.}
\end{cases}$$

We construct a system of linear ODEs with inputs and outputs associated to the quadruple $(G, In, Out, Leak)$ as follows:

$$\dot{x}(t) = A(G)(t) + u(t) \quad y_i(t) = x_i(t) \text{ for } i \in Out$$

where $u_i(t) \equiv 0$ for $i \notin In$. The functions $x_i(t)$ are the state variables, the functions $y_i(t)$ are the output variables, and the nonzero functions $u_i(t)$ are the inputs. The resulting model is called a linear compartment model. The following conventions are often used for presenting diagrams of linear compartment models: numbered vertices represent compartments, outgoing edges from compartments represent leaks, an edge with a circle coming out of a compartment represent an output measurement, and an arrowhead pointing into a compartment represents an input.

**Example 16.4.2** (Three Compartment Model). Consider the graph representing a linear compartment model from Figure 16.4.1. The resulting ODE system takes the form:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{pmatrix} = \begin{pmatrix} -a_{01} - a_{21} & a_{12} & a_{13} \\ a_{21} & -a_{02} - a_{12} & a_{23} \\ 0 & a_{32} & -a_{13} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ 0 \end{pmatrix}.$$
Proposition 16.4.3. Consider the linear compartment model defined by the quadruple \((G, In, Out, Leak)\). Let \(\partial\) denote the differential operator \(d/dt\) and let \((\partial I - A(G))^{ji}\) denote the matrix obtained by deleting the \(j\)th row and \(i\)th column of \(\partial I - A(G)\). Then there are input/output equations of this model having the form

\[
\frac{\det(\partial I - A(G))}{g_i} y_i = \sum_{j \in In} (-1)^{i+j} \left( \frac{\det(\partial I - A(G))^{ji}}{g_i} \right) u_j
\]

where \(g_i\) is the greatest common divisor of \(\det(\partial I - A(G))\) and the set of polynomials \(\det(\partial I - A(G))^{ji}\) such that \(j \in In\) for a given \(i \in Out\).

Proof. We perform formal linear algebra manipulations of the system (16.4.3). This differential equation system can be rewritten in the form:

\[(\partial I - A(G))x = u.\]

Using Cramer’s rule, for \(i \in Out\) we can “solve” for \(x_i = y_i\) by the formal expression

\[y_i = \frac{\det(B_i)}{\det(\partial I - A(G))}\]

where \(B_i\) is the matrix obtained from \(\partial I - A(G)\) by replacing the \(i\)th column by \(u\). Clearing denominators and using the Laplace expansion down the \(i\)th column of \(B_i\) yields the formula

\[
\det(\partial I - A(G))y_i = \sum_{j \in In} (-1)^{i+j} \left( \frac{\det(\partial I - A(G))^{ji}}{g_i} \right) u_j.
\]

However, this expression need not be minimal. This is the reason for the division by the common factor \(g_i\) in the proposition. \(\square\)

With various reasonable assumptions on the structure of the model, it is often possible to show that for generic parameter choices the greatest common divisor polynomials that appears in Proposition 16.4.3 is 1. Recall...
that a graph is strongly connected if for any pair of vertices \( i, j \in V \) there is a directed path from \( i \) to \( j \).

**Proposition 16.4.4.** Let \( G \) be a strongly connected graph. If \( \text{Leak} \neq \emptyset \) then the polynomial \( \det(\partial I - A(G)) \) is irreducible as a polynomial in \( \mathbb{K}(a)[\partial] \).

See [MSE15, Theorem 2.6] for a proof.

**Example 16.4.5 (Three Compartment Model).** Continuing Example 16.4.2 we see that the graph \( G \) is strongly connected, so the input/output equation for this system is

\[
\begin{align*}
y^{(3)} + (a_{01} + a_{21} + a_{13} + a_{02} + a_{32})y^{(2)} + (a_{01}a_{12} + a_{01}a_{13} + a_{21}a_{13} + a_{12}a_{13} + a_{01}a_{02} + a_{21}a_{02} + a_{13}a_{02} + a_{21}a_{32} + a_{13}a_{32})y^{(1)} \\
+ (a_{01}a_{12}a_{13} + a_{01}a_{13}a_{02} + a_{21}a_{13}a_{02} + a_{01}a_{13}a_{32})y &= u^{(2)}_1 + (a_{12} + a_{13} + a_{02} + a_{32})u^{(1)}_1 + (a_{12}a_{13} + a_{13}a_{02} + a_{13}a_{32})u_1 \\
+ a_{12}u^{(1)}_2 + (a_{12}a_{13} + a_{13}a_{32})u_2.
\end{align*}
\]

One difference between the identifiability analysis of state space models compared to the statistical models studied in the majority of this chapter is that state space models are often not generically identifiable. In spite of the non-identifiability, the models are still used and analysis is needed to determine which functions of the parameters can be identified from the given input/output data, and how the model might be reparametrized or non-dimensionalized so that meaningful inferences can be made.

**Example 16.4.6 (Modified Two Compartment Model).** Consider a modification of the two compartment model from Example 16.4.1 where now there are leaks from both compartments one and two:

\[
\begin{align*}
\dot{x}_1 &= -(a_{21} - a_{01})x_1 + a_{12}x_2 + u_1 \\
\dot{x}_2 &= a_{21}x_1 + (-a_{12} - a_{02})x_2
\end{align*}
\]

with 1-dimensional output variable \( y(t) = Vx_1(t) \). The resulting input-output equation is:

\[
\begin{align*}
\dot{y} + (a_{21} + a_{12} + a_{02} + a_{01})\dot{y} + (a_{01}a_{12} + a_{01}a_{02} + a_{21}a_{02})y &= V\dot{u}_1 + V(a_{12} + a_{02})u_1.
\end{align*}
\]

We see that the model has five parameters but there are only 4 nonmonic coefficients in the input/output equation. Hence the model is unidentifiable. However, it is not difficult to see that following quantities are identifiable from the input/output equation:

\[V, a_{12} + a_{02}, a_{21} + a_{01}, a_{12}a_{21}.\]

In general, the problem of finding a set of identifiable functions could be phrased in algebraic language as follows.
Problem 16.4.7. Let \( c : \mathbb{R}^d \to \mathbb{R}^k \) be a rational map.

- (Rational identifiable functions) Find a “nice” transcendence basis of the field extension \( \mathbb{R}(c_1, \ldots, c_k)/\mathbb{R} \).
- (Algebraic identifiable functions) Find a “nice” transcendence basis of the field extension \( \overline{\mathbb{R}}(c_1, \ldots, c_k)/\mathbb{R} \).

The word “nice” is used and appears in quotations because this meaning might depend on the context and does not seem to have a natural mathematical formulation. Nice might mean: sparse, low degree, with each element involving a small number of variables, or each element having a simple interpretation in terms of the underlying application. One strategy to attempt to find identifiable functions in the case when a model is non-identifiable is to use eliminations based on the following result:

Proposition 16.4.8. Let \( c : \mathbb{R}^d \to \mathbb{R}^k \) be a polynomial map and the \( I_c \) be the ideal

\[
I_c = \langle c_1(t) - c_1(\theta), \ldots, c_k(t) - c_k(\theta) \rangle \subseteq \mathbb{R}(\theta)[t].
\]

Let \( f \in \mathbb{R}[t] \). Then \( f \) is rationally identifiable if and only if \( f(t) - f(\theta) \in I_c \).

Proof. First we consider the variety \( V(I_c) \subseteq \mathbb{R}(\theta)^d \). The interpretation is that for a generic parameter choice \( \theta \), we look at the variety of preimages \( c^{-1}(c(\theta)) \). To say that \( f(t) - f(\theta) \in \sqrt{I_c} \) means that \( f(t) - f(\theta) \) vanishes on the variety \( V(I_c) \) by the Nullstellensatz. In turn, this means that every \( t \in V(I_c) \) has \( f(t) = f(\theta) \), which, since we work over an algebraically closed field implies that \( f(t) \) is rationally identifiable. Finally, Bertini’s theorem (see [Har77, Thm II.8.18]) implies that generic fibers of a rational map are smooth and reduced, which implies that \( I_c \) is radical, i.e. \( \sqrt{I_c} = I_c \).

Proposition 16.4.8 is used as a heuristic to find identifiable functions that depend on a small number of variables as described in [MED09]. The user inputs the ideal \( I_c \), computes Gröbner bases with respect to random lexicographic orders, and inspects the elements at the end of the Gröbner basis (which only depend on a small number of indeterminates with lowest weight in the lexicographic ordering) for polynomials of the form \( f(t) - f(\theta) \).

16.5. Exercises

Exercise 16.1. Prove that the dimension of \( \text{Mix}^k(M_{X_1 \perp X_2}) \) is \( k(r_1 + r_2 - k) - 1 \) when \( k \leq r_1, r_2 \).

Exercise 16.2. Explain how to modify Proposition 16.1.7 to check local identifiability or generic non-identifiability of individual parameters \( s(\theta) \).
Exercise 16.3. Consider the linear structural equation model on four random variables whose graph is indicated in Figure 16.5.1. Which parameters of this model are globally identifiable, locally identifiable, algebraically identifiable, etc.?

Exercise 16.4. Consider the factor analysis model $F_{m,s}$ consisting of all covariance matrices

$$
F_{m,s} = \{ \Omega + \Lambda \Lambda^T \in PD_m : \Omega > 0 \text{ diagonal}, \Lambda \in \mathbb{R}^{m \times s} \}.
$$

Perform an identifiability analysis of this model. In particular, consider the map $\phi : \mathbb{R}_{>0}^m \times \mathbb{R}^{m \times s} \to F_{m,s}$.

1. If $(\Omega^*, \Lambda^*)$ is generic, what are the dimension and geometry of the fiber $\phi^{-1}(\phi(\Omega^*, \Lambda^*))$?

2. If $\Omega^*$ is generic, what is the dimension of the fiber $\phi^{-1}(\phi(\Omega^*, 0))$?

Exercise 16.5. Complete Example 16.2.10 by showing that all the conditions of Theorem 16.2.9 are satisfied.

Exercise 16.6. Use Theorem 16.2.9 to give a direct proof of Theorem 16.2.5.

Exercise 16.7. Prove that every mixed graph $G$ is the subgraph of some graph $H$ whose linear structural equation model is generically identifiable. (Hint: For each node in $G$ add an instrumental variable that points to it.)

Exercise 16.8. Consider the graph shown in Figure 16.5.2. Suppose that all random variables are binary and that $Y$ is a hidden variable. Show that this model is identifiable up to label swapping on the hidden node $Y$. 
Exercise 16.9. Consider the following SIR model for the transmission of infectious diseases. Here \( S(t) \) is the number of susceptible individuals, \( I(t) \) is the number of infected individuals, \( R(t) \) is the number of recovered individuals, and \( \beta \) and \( \gamma \) are unknown parameters that represent the infection rate and the recovery rate, respectively:

\[
\frac{dS}{dt} = -\beta IS \\
\frac{dI}{dt} = \beta IS - \gamma I \\
\frac{dR}{dt} = \gamma I
\]

(1) Find the nonlinear second order differential equation in \( I \) for the dynamics of \( I \). (That is, perform differential elimination on this system to find a differential equation in terms of only \( I \)).

(2) Are the parameters \( \beta \) and \( \gamma \) generically identifiable from the observation of the trajectory of \( I \) alone?

Exercise 16.10. Verify that the linear two compartment model of Example 16.4.1 is identifiable when all parameters are real and positive.

Exercise 16.11. Explain how the input/output equation of Proposition 16.4.3 should be modified when each output measurement is of the form \( y_i(t) = V_i x_i(t) \) where \( V_i \) is an unknown parameter.

Exercise 16.12. Decide whether or not the model from Example 16.4.2 is identifiable.

Exercise 16.13. Consider the linear compartment model whose graph is a chain \( 1 \rightarrow 2 \rightarrow \cdots \rightarrow n \), with input in compartment 1, output in compartment \( n \) and a leak in compartment \( n \). Is the model identifiable? What happens if there are more leak compartments besides just compartment \( n \)?

Exercise 16.14. Explain how to modify Proposition 16.4.8 to search for rational functions that are generically rationally identifiable.
Model Selection and Bayesian Integrals

In previous chapters, especially Chapter 9, we have primarily concerned ourselves with hypothesis tests. These tests either reject a hypothesis (usually that an underlying parameter satisfies certain restrictions) or conclude that the test was inconclusive. Model selection concerns the situation where there is a collection of models that are under consideration and we would like to choose the best model for the data, rather than decide if we have evidence against a specific model. In this model selection problem, a typical situation is that the models under consideration are nested in various ways. For instance, in graphical modeling, we might want to choose the best directed acyclic graph for our given data set. Or we might consider a sequence of mixture models

\[ \mathcal{M} \subseteq \operatorname{Mixt}^2(\mathcal{M}) \subseteq \operatorname{Mixt}^3(\mathcal{M}) \subseteq \cdots \]

and try decide the number of mixture components. When two models are nested as \( \mathcal{M}_1 \subseteq \mathcal{M}_2 \), the maximum likelihood of \( \mathcal{M}_2 \) will necessarily be greater than or equal to that of \( \mathcal{M}_1 \), so the maximum likelihood alone cannot be used to decide between the two models. The development of a practical model selection criterion must include some penalty for the fact that the larger model is larger, and so it is easier for that model to fit better, when measured from the standpoint of maximum likelihood.

Two well-known model selection criteria are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These information criteria are determined by the penalty function they place on the likelihood score. The BIC can be shown to consistently choose the correct
model under relatively mild conditions on the underlying models (see Theorem 17.1.3). The derivation of the BIC, and the reason for the adjective “Bayesian”, are through its connection to the marginal likelihood. This derivation depends on the underlying model being smooth. Algebraic geometry enters into the development of model selection strategies when the statistical models are not smooth, and the derivation of the BIC no longer holds. The singularities of the model can negatively affect the accuracy of approximations to the marginal likelihood and the resulting performance of the standard information criteria. To deal with this problem, corrections to the information criteria are needed that take into account the model singularities. This idea was developed in the work of Watanabe, summarized in the book [Wat09]. A primary technique for dealing with the singularities involves analyzing the asymptotics of deterministic Laplace integrals. We illustrate these ideas in this chapter.

17.1. Information Criteria

Fix a statistical model \( \mathcal{P}_\Theta = \{ P_\theta : \theta \in \Theta \} \) with parameter space \( \Theta \subset \mathbb{R}^k \). Let

\[
X^{(1)}, \ldots, X^{(n)} \sim P_{\theta_0}
\]

be i.i.d. samples drawn from an unknown true distribution \( P_{\theta_0} \) where \( \theta_0 \in \Theta \). A submodel \( \mathcal{P}_{\Theta_0} \subset \mathcal{P}_\Theta \) given by \( \Theta_0 \subset \Theta \) is a true model if \( \theta_0 \in \Theta_0 \).

In this chapter, we consider the model selection problem, which attempts to use the information from the sample \( X^{(1)}, \ldots, X^{(n)} \) to find the “simplest” true model from a family of competing models associated with the sets

\[ \Theta_1, \ldots, \Theta_M \subset \Theta. \]

We will assume that \( \mathcal{P}_\Theta \) is an algebraic exponential family, and that each \( \Theta_i \) is a semialgebraic subset of \( \Theta \), so that the \( \mathcal{P}_{\Theta_i} \) is also an algebraic exponential family. Furthermore, we suppose that each distribution \( P_\theta \) has a density function \( p_\theta(x) \). Let

\[
\ell_n(\theta \mid X^{(1)}, \ldots, X^{(n)}) = \sum_{i=1}^n \log p_\theta(X^{(i)})
\]

be the log-likelihood function. Note that we include a specific dependence on \( n \) here because we will be interested in what happens as \( n \) tends to infinity.

A naive approach to deciding which of the models \( \mathcal{P}_{\Theta_1}, \ldots, \mathcal{P}_{\Theta_M} \) is a true model is to determine which of the models achieves the maximum likelihood. Letting

\[
\hat{\ell}_n(i) = \sup_{\theta \in \Theta_i} \ell_n(\theta \mid X^{(1)}, \ldots, X^{(n)})
\]
17.1. Information Criteria
denote the maximum value of the likelihood function when restricted to model \( \mathcal{P}_{\Theta_i} \), this approach to model selection would select the model such that \( \hat{\ell}_n(i) \) is largest. Such a method is not satisfactory in practice because it will tend to pick models that are more complex (e.g. high dimensional) over models that are simple (e.g. low dimensional). Indeed, if \( \Theta_i \subseteq \Theta_j \) then a model selection procedure based solely on maximum likelihood would always select the model \( \mathcal{P}_{\Theta_j} \) over \( \mathcal{P}_{\Theta_i} \). To correct for this problem, we introduce penalties for the model complexity.

**Definition 17.1.1.** The information criterion associated to a family of penalty functions \( \pi_n : [M] \to \mathbb{R} \) assigns the score \( \tau_n(i) = \hat{\ell}_n(i) - \pi_n(i) \) to the \( i \)-th model, for \( i = 1, \ldots, M \).

A number of general purpose information criteria have been proposed in the literature.

**Definition 17.1.2.** The Akaike information criterion (AIC) \([Aka74]\) uses the penalty function \( \pi_n(i) = \dim(\Theta_i) \). The Bayesian information criterion (BIC) \([Sch78]\) (sometimes called the Schwarz information criterion) has penalty function \( \pi_n(i) = \frac{1}{2} \dim(\Theta_i) \log(n) \).

The score-based approach to model selection now selects a model for which \( \tau_n(i) \) is maximized. Consistency results for model selection procedures can be formulated for very general classes of models in this context.

**Theorem 17.1.3** (Consistency of Model Selection Procedures). Consider a regular exponential family \( \{ \mathcal{P}_\theta : \theta \in \Theta \} \). Let \( \Theta_1, \Theta_2 \subseteq \Theta \) be arbitrary sets. Denote the standard topological closure of \( \Theta_1 \) by \( \overline{\Theta}_1 \).

1. Suppose that \( \theta_0 \in \Theta_2 \setminus \overline{\Theta}_1 \). If the penalty functions are chosen such that the sequence \( |\pi_n(2) - \pi_n(1)|/n \) converges to zero as \( n \to \infty \), then
   \[
   P_{\theta_0}(\tau_n(1) < \tau_n(2)) \to 1, \quad \text{as } n \to \infty.
   \]

2. Suppose that \( \theta_0 \in \Theta_1 \cap \Theta_2 \). If the sequence of differences \( \pi_n(1) - \pi_n(2) \) diverges to infinity as \( n \to \infty \) then
   \[
   P_{\theta_0}(\tau_n(1) < \tau_n(2)) \to 1, \quad \text{as } n \to \infty.
   \]

Theorem 17.1.3 appears in \([Hau88]\). Roughly speaking, it says that for reasonable conditions on the penalty functions \( \pi_n \), the model selection criterion will tend to pick the simplest true model as the sample size \( n \) tends to infinity. Note that the AIC satisfies the first part of the theorem and not the second, whereas the BIC satisfies both of these conditions. The takeaway message comparing these two information criteria in the context
Table 17.1.1. Table of opinions on same-sex marriage, schooling, and religion

<table>
<thead>
<tr>
<th>Education</th>
<th>Religion</th>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>High school or less</td>
<td>Fundamentalist</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Liberal</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>At least some college</td>
<td>Fundamentalist</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Liberal</td>
<td>22</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 17.1.2. Information criteria for the same-sex marriage data

<table>
<thead>
<tr>
<th>Model</th>
<th>Likelihood Value</th>
<th>Dimension</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independence</td>
<td>-145.3</td>
<td>3</td>
<td>-148.3</td>
<td>-151.7</td>
</tr>
<tr>
<td>No 3-Way</td>
<td>-134.7</td>
<td>6</td>
<td>-140.7</td>
<td>-147.4</td>
</tr>
</tbody>
</table>

of the consistency theorem is that the BIC will tend to pick the smaller dimensional of two true models, provided that there is enough data.

Example 17.1.4 (Opinion on Same-Sex Marriage). Consider the $2 \times 2 \times 2$ table of counts in Table 17.1.1. These data were extracted from Table 8.9 in [Agr13]. This table classifies 71 respondents, age 18-25, on three questions: their opinion on whether homosexuals should be allowed to marry, their education level, and whether their religious affiliation is fundamentalist or liberal. This table is only a small subset of the much larger 2008 General Social Survey which asked many questions with many degrees of gradation. We only extract this small subtable to illustrate the model selection idea.

Suppose that we want to decide from this data between the model of complete independence and the model of no 3-way interaction. The value of the likelihood functions, dimensions, AIC, and BIC scores are displayed in Table 17.1.2. Both the AIC and the BIC scores select the no-3-way interaction model.

Usually we are not interested in comparing just two models, but rather choosing amongst a whole family of models to find the one that best fits the data. An example of a model selection problem with a large family of models is the problem of finding the best phylogenetic tree model to explain observed molecular sequence data, as discussed in Chapter 15. Since the search space for finding a phylogenetic tree model is typically restricted to binary trees, and for a fixed underlying model structure the dimension of a phylogenetic model only depends on the number of leaves, the penalty terms will be the same for every binary tree under consideration, and so the likelihood score alone is sufficient to choose an optimal tree. On the other hand, choosing the number of mixture components, looking at models with
network structure, or choosing between underlying substitution models can create interesting challenges for model selection in phylogenetics.

A particularly large search space arises when we try to determine an appropriate graphical model, and search through all graphs to find the best one. For directed graphical models, this can be turned in a natural way into a convex optimization problem over a (high dimensional) polytope. Strategies for approximating these model selection polytopes can be helpful for this optimization problem.

To explain how this goes, we focus on the problem of choosing the best directed acyclic graph, using a penalized likelihood approach. We focus on the case of \( m \) discrete random variables, where random variable \( X_i \) has \( r_i \) states (a similar reasoning can also be used in other cases with continuous random variables). Recall that if \( G \) is a directed acyclic graph with vertex set \([m]\) and directed edge set \( D \), we have the parametric density factorization

\[
(17.1.1) \quad f(x) = \prod_{j \in V} f_j(x_j \mid x_{\text{pa}(j)}).
\]

As discussed in Section 13.2, this recursive factorization also gives a parametric description of the model. For each \( j \in [m] \) we have a parameter array \( \theta^{(j)} \). Equivalent to the recursive factorization (17.1.1) is the parametrization

\[
p_{i_1i_2\ldots i_m} = \prod_{j \in V} \theta^{(j)}(i_j \mid i_{\text{pa}(j)}).
\]

Each \( \theta^{(j)} \) is a conditional probability distribution. It is an \( r_j \times \prod_{k \in \text{pa}(j)} r_k \) array, and has \((r_j - 1) \prod_{k \in \text{pa}(j)} r_k\) free parameters in it.

Let \( \mathcal{R} = \prod_{j \in V} [r_j] \). Given a table of counts \( u \in \mathbb{N}^\mathcal{R} \) and a set \( F \subseteq [m] \), recall that \( u|_F \) denotes the marginal of \( u \) onto the indices given by \( F \). For the directed graphical model, the log-likelihood has the form

\[
\ell_n(\theta \mid u) = \sum_{i \in \mathcal{R}} u_i \log p_i = \sum_{i \in \mathcal{R}} u_i \log \prod_{j \in V} \theta^{(j)}(i_j \mid i_{\text{pa}(j)})
\]

\[
= \sum_{j \in V} \sum_{i_j, i_{\text{pa}(j)} \in \mathcal{R}_{j, i_{\text{pa}(j)}}} \log u_{i_j, i_{\text{pa}(j)}} \log \theta^{(j)}(i_j \mid i_{\text{pa}(j)}).
\]

The last equality is obtained by collecting terms. For any distinct \( j_1 \) and \( j_2 \), the parameters \( \theta^{(j_1)} \) and \( \theta^{(j_2)} \) are completely independent of each other. Hence the optimization to find the maximum likelihood estimate is obtained by simply matching the conditional distribution \( \theta^{(j)} \) to the empirical conditional distribution. This is summarized in the following:
Proposition 17.1.5. Let $G$ be a directed acyclic graph. The maximum likelihood estimates for the parameters $\theta^{(j)}$ are the empirical conditional distributions from the data vector $u$.

One consequence of this nice structure of the maximum likelihood estimates is that the likelihood function evaluated at the maximum likelihood estimate is the sum of independent terms, and the structure of those terms only depend locally on the graph: if node $j$ has parent set $\text{pa}(j)$, then the same corresponding expression appears in the likelihood of every DAG $G$ with that same local structure. This structure can further be incorporated into a model selection procedure, for either the AIC or BIC. We explain for the BIC and leave the analogous statements for the AIC to the reader.

For each pair of $j$ and set $F \subseteq [m] \setminus \{j\}$ we calculate the value

$$Q(j, F \mid u) = \sum_{i \in [r_j]} (u_{i,F})_{i,j_F} \log \left( \frac{(u_{j,F})_{i,j_F}}{\sum_{k=1}^{r_j} (u_{j,F})_{k,j_F}} \right) - \left( (r_j - 1) \prod_{k \in F} r_k \right) \log n.$$

Proposition 17.1.6. For a given directed acyclic graph $G$, the BIC score associated to the observed data $u$ is

$$\tau_n(G) = \sum_{j \in V} Q(j, \text{pa}(j) \mid u).$$

Proof. The score $Q(j, F \mid u)$ consists of two parts: the first part sums to give the likelihood function evaluated at the maximum likelihood estimate. The second part, $(r_j - 1) \prod_{k \in F} r_k \log n$, sums to give the penalty function consisting of the product of the logarithm of the number of sample points $n$ times the dimension of the model. \qed

The BIC for a graphical model breaks as the sum of pieces, each of which only depends on local structure of the graph, and each of which is independent to calculate. An approach for finding the best directed graph based on this decomposition appears in [SM06]. We explain a polyhedral approach that comes out of these ideas. Introduce an $m \times 2^{m-1}$ dimensional space $\mathbb{R}^{m \times 2^{m-1}}$ whose coordinates are indexed by pairs $j \in [m]$ and $F \subseteq [m] \setminus \{j\}$. Let $e_{j,F}$ denote the standard unit vector with a 1 in the position corresponding to the pair $(j, F)$ and zeros elsewhere. To each DAG $G$, we associate the vector

$$\delta_G = \sum_{j \in [m]} e_{j,\text{pa}(j)}.$$
17.2. Bayesian Integrals and Singularities

**Definition 17.1.7.** The *DAG model selection polytope* is the polytope:

\[ MS_m = \text{conv}(\delta_G : G = ([m], D) \text{ is a DAG on } [m]). \]

Finally, let \( q \in \mathbb{R}^{m \times 2^m - 1} \) be the vector whose component indexed by the pair \((j,F)\) is \( Q(j,F \mid u) \).

**Proposition 17.1.8.** Given a table of counts \( u \in \mathbb{N}^R \), the directed acyclic graph \( G \) with the maximal BIC score is the graph such that the vector \( \delta_G \) is the optimal solution of the linear programming problem:

\[
\max q^T x \quad \text{subject to} \quad x \in MS_m.
\]

**Proof.** One just needs to verify that for any DAG \( G \), \( q^T \delta_G = \tau_n(G) \). This follows from Proposition 17.1.6. \( \square \)

To optimize over the polytope \( MS_m \) in practice, one tries to find nice sets of inequalities that define a polytope whose only integral points are the points \( \delta_G \). Since this polytope has exponentially many vertices and is in an exponential dimensional space, there are significant issues with making these methodologies tractable. This includes changing to alternate coordinate systems in lower dimensions [HLS12], restricting to graphs that are relatively sparse or are structurally constrained [XY15], and determining methods for automatically generating inequalities [CHS17].

17.2. Bayesian Integrals and Singularities

Theorem 17.1.3 sheds light on the behavior of model selection criteria but it does not directly motivate a specific trade-off between model fit and model complexity, as seen in BIC or AIC. In the remainder of this section we focus on BIC and show how its specific penalty can be derived from properties of Bayesian integrals. Although the BIC is consistent whether or not the underlying models are smooth, these derivations are based on smoothness assumptions concerning the underlying models. In the case that the models are not smooth, we will see that further corrections can be made that take into account the nature of the singularities.

In the Bayesian approach we consider the distribution \( P_\theta \) to be the conditional distribution of our observations given the model being considered and the value of the parameter \( \theta \). The observations \( X^{(1)}, \ldots, X^{(n)} \) are assumed to be i.i.d. given \( \theta \). For each model we have a prior distribution \( Q_\theta \) on \( \Theta \). Finally, there is a prior probability \( P(\Theta) \) for each of the models. With these ingredients in place, we may choose the model that maximizes the posterior probability. Let \( L_n(\theta \mid X^{(1)}, \ldots, X^{(n)}) \) denote the likelihood
function. The posterior probability of the $i$-th model is

$$P(\Theta_i \mid X^{(1)}, \ldots, X^{(n)}) \propto \pi(\Theta_i) \int_{\Theta_i} L_n(\theta \mid X^{(1)}, \ldots, X^{(n)}) dQ_i(\theta).$$

The constant of proportionality in this expression is just a normalizing constant so that $\sum_i P(\Theta_i \mid X^{(1)}, \ldots, X^{(n)}) = 1$. The difficulty in computing the posterior probability comes from the difficulty in computing the integral. This integral is known as the marginal likelihood integral or just the marginal likelihood.

In most applications each of the submodels $\Theta_i$ is a parametric image of a map $g_i : \Gamma_i \rightarrow \mathbb{R}^k$, where $\Gamma_i$ is a full-dimensional subset of some $\mathbb{R}^{d_i}$, and the prior $Q_i$ can be specified via a Lebesgue integrable density $p_i(\gamma)$ on $\mathbb{R}^{d_i}$. Suppressing the index $i$ of the submodel, the marginal likelihood takes the form

$$\mu(X^{(1)}, \ldots, X^{(n)}) = \int_{\Gamma} L_n(g(\gamma) \mid X^{(1)}, \ldots, X^{(n)}) p(\gamma) d\gamma$$

which can be expressed in terms of the log-likelihood as

$$\mu(X^{(1)}, \ldots, X^{(n)}) = \int_{\Gamma} \exp(\ell_n(g(\gamma) \mid X^{(1)}, \ldots, X^{(n)})) p(\gamma) d\gamma.$$

We illustrate the structure of the marginal likelihood $\mu(X^{(1)}, \ldots, X^{(n)})$ in the case of Gaussian random vectors with covariance matrix equal to the identity matrix. While this example is quite special, it is revealing of the general structure that arises for (algebraic) exponential families.

**Example 17.2.1** (Multivariate Normal Model). Let $X^{(1)}, \ldots, X^{(n)}$ be independent $\mathcal{N}(\theta, Id_m)$ random vectors with $\theta \in \mathbb{R}^m$. From Equation 7.1.2, we see that the log-likelihood can be written as

$$\ell_n(\theta \mid X^{(1)}, \ldots, X^{(n)}) = -\frac{nm}{2} \log 2\pi - \frac{n}{2} \|X - \mu\|^2 - \frac{1}{2} \sum_{i=1}^{n} \|X^{(i)} - \bar{X}\|^2.$$

It follows that the marginal likelihood is

$$\mu(X^{(1)}, \ldots, X^{(n)}) = \left(\frac{1}{\sqrt{(2\pi)^m}}\right)^n \exp \left(-\frac{1}{2} \sum_{i=1}^{n} \|X^{(i)} - \bar{X}\|^2\right) \times \int_{\Gamma} \exp \left(-\frac{n}{2} \|X - g(\gamma)\|^2\right) p(\gamma) d\gamma.$$
The goal of the remainder of this chapter is to understand the asymptotics of this marginal likelihood as \( n \) goes to infinity. Symbolic algebra strategies can be used to compute the marginal likelihood exactly in small examples (small number of parameters and small sample size) \([LSX09]\), but this is not the focus of this chapter. A key observation in Example 17.2.1 is that as \( n \to \infty \), \( \bar{X} \) converges to the true value \( \theta \). If \( \theta \) is in the model \( g(\Gamma) \) then the logarithm of the marginal likelihood has behavior roughly like \( \ell_n - C_n \) where \( C_n \) is a correction term that comes from the remaining integral and the error in the approximation of \( \bar{X} \) to the true parameter value. This has the basic shape of the information criteria that were introduced in Section 17.1. As we will see, the marginal likelihood always has this form for reasonable statistical models. To make this precise, we need the following definition.

**Definition 17.2.2.** A sequence of random variables \( (R_n)_n \) is bounded in probability if for all \( \varepsilon > 0 \) there is a constant \( M_\varepsilon \) such that \( P(|R_n| > M_\varepsilon) < \varepsilon \) for all \( n \).

The notation \( O_p(1) \) can be used to denote a sequence of random variables that are bounded in probability when we do not want to explicitly name these random variables.

**Theorem 17.2.3** (Laplace Approximation). \([Hau88, \text{Thm 2.3}]\) Let \( \{P_\theta : \theta \in \Theta\} \) be a regular exponential family with \( \Theta \subseteq \mathbb{R}^k \). Let \( \Gamma \subseteq \mathbb{R}^d \) be an open set and \( g : \Gamma \to \Theta \) be a smooth injective map with a continuous inverse on \( g(\Gamma) \). Let \( \theta_0 = g(\gamma_0) \) be the true parameter. Assume that the Jacobian of \( g \) has full rank at \( \gamma_0 \) and that the prior \( \pi(\gamma) \) is a smooth function that is positive in a neighborhood of \( \gamma_0 \). Then

\[
\log \mu(X^{(1)}, \ldots, X^{(n)}) = \hat{\ell}_n - \frac{d}{2} \log(n) + O_p(1),
\]

where

\[
\hat{\ell}_n = \sup_{\gamma \in \Gamma} \ell_n(g(\gamma) \mid X^{(1)}, \ldots, X^{(n)}).
\]

Theorem 17.2.3 explains why BIC can be seen as an approximation to a Bayesian procedure based on selecting the model with highest posterior probability. Indeed, for large sample size, the logarithm of the marginal likelihood looks like the BIC plus an error term, assuming that the map \( g \) is smooth and invertible. Although the BIC is a consistent model selection criterion without smoothness assumptions (Theorem 17.1.3), Theorem 17.2.3 provides further justification for why the BIC gives reasonable behavior. On the other hand, if the conditions of Theorem 17.2.3 are not satisfied, then the conclusions can fail.
Example 17.2.4 (Monomial Curve). Continuing with Example 17.2.1, let $X^{(1)}, \ldots, X^{(n)}$ be i.i.d. random vectors with distribution $\mathcal{N}(\theta, \text{Id}_2)$, with $\theta \in \mathbb{R}^2$. We consider the singular model given as the image of the parametrization $g(\gamma) = (\gamma^3, \gamma^4)$. Let $\overline{X} = (\bar{X}_{n,1}, \bar{X}_{n,2})$ be the sample mean. From Example 17.2.1, we see that the evaluation of the marginal likelihood requires computation of the integral

$$
\bar{\mu}(\bar{X}) = \int_{-\infty}^{\infty} \exp \left( - \frac{n}{2} \| \bar{X} - g(\gamma) \|^2 \right) p(\gamma) d\gamma.
$$

$$
= \int_{-\infty}^{\infty} \exp \left( - \frac{n}{2} \left( (\sqrt{n} \gamma^3 - \sqrt{n} \bar{X}_{n,1})^2 + (\sqrt{n} \gamma^4 - \sqrt{n} \bar{X}_{n,2})^2 \right) \right) p(\gamma) d\gamma.
$$

If the true parameter $\theta_0 = g(\gamma_0)$ is non-zero, then Theorem 17.2.3 applies with $d = 1$. On the other hand, if $\theta_0 = g(0) = 0$, Theorem 17.2.3 does not apply because the Jacobian does not have full rank. In fact, in this case the marginal likelihood will have a different asymptotic behavior. Changing variables to $\tilde{\gamma} = n^{1/6} \gamma$ we obtain

$$
(17.2.3) \quad \bar{\mu}(\bar{X}) = n^{-1/6} \int_{-\infty}^{\infty} \exp \left( - \frac{1}{2} \left( \tilde{\gamma}^3 - \sqrt{n} \bar{X}_{n,1} \right)^2 + \left( \frac{\tilde{\gamma}^4}{n^{1/3}} - \sqrt{n} \bar{X}_{n,2} \right)^2 \right) p \left( \frac{\tilde{\gamma}}{n^{1/6}} \right) d\tilde{\gamma}.
$$

By the central limit theorem $\sqrt{n} \bar{X}_{n,1}$ and $\sqrt{n} \bar{X}_{n,2}$ converge to $\mathcal{N}(0, 1)$ random variables. Clearing denominators and letting $Z_1$ and $Z_2$ be independent $\mathcal{N}(0, 1)$ random variables, we see that

$$
n^{1/6} \bar{\mu}(\bar{X}) \to D \int_{-\infty}^{\infty} \exp \left( - \frac{1}{2} (\tilde{\gamma}^3 - Z_1)^2 + Z_2^2 \right) p(0) d\tilde{\gamma}.
$$

Since convergence in distribution implies boundedness in probability, we see that

$$
\log \bar{\mu}(\bar{X}) = - \frac{1}{6} \log n + O_p(1)
$$

from which it follows that

$$
\log \mu(X^{(1)}, \ldots, X^{(n)}) = \hat{\ell}_n - \frac{1}{6} \log n + O_p(1)
$$

when $\theta_0 = (0, 0)$. Recall that $\hat{\ell}_n$ denotes the maximized likelihood value in the model $\{(\gamma^3, \gamma^4) : \gamma \in \mathbb{R}\}$. This is correct because the difference between the log of the maximum likelihood value in (17.2.2) and the maximum log-likelihood of the submodel is $O_p(1)$ when the submodel is true.
The key detail to note in this example is that the asymptotics of the marginal likelihood changes when the true parameter value $\theta_0$ lies at a singularity: We do not see the asymptotic expansion
\[
\log \mu(X^{(1)}, \ldots, X^{(n)}) = \hat{\ell}_n - \frac{1}{2} \log n + O_p(1)
\]
that one would expect to see when the model is one dimensional and smooth.

While our analysis in Example 17.2.4 involved a probabilistic integral, it is, in fact, possible to reduce the study of the asymptotics to studying the asymptotics of a deterministic version of the Laplace integral. Indeed, let
\[
\ell_0(\theta) = E[\log p_{\theta}(X)]
\]
where $X$ is a random variable with density $p_{\theta}$, given by the true parameter value $\theta_0$. Then the deterministic Laplace integral is
\[
\mu(\theta_0) = \int_\Gamma \exp (n \ell_0(g(\gamma))) p(\gamma) \, d\gamma.
\]

Example 17.2.5 (Deterministic Laplace Integral of Multivariate Normal Model). Continuing our discussion of Gaussian random variables, suppose we have mean $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,m})$ and covariance matrix is constrained to always be the identity. Then
\[
\ell_0(\theta) = E[\log p_{\theta}(X)]
\]
\[
= E \left[ -\frac{1}{2} \sum_{i=1}^m (X_i - \theta_i)^2 \right] - \frac{m}{2} \log(2\pi)
\]
\[
= -\frac{1}{2} \sum_{i=1}^m E[X_i^2 - 2X_i\theta_i + \theta_i^2] - \frac{m}{2} \log(2\pi)
\]
\[
= -\frac{1}{2} \sum_{i=1}^m (1 + (\theta_{0,i} - \theta_i)^2) - \frac{m}{2} \log(2\pi).
\]
In the case of our parametrized model where $\theta_0 = g(\gamma_0)$ the deterministic Laplace integral has the form
\[
\mu(g(\gamma_0)) = \int_\Gamma \exp \left(-\frac{nm}{2} - \frac{nm}{2} \log(2\pi) - \frac{n}{2}||g(\gamma) - g(\gamma_0)||^2\right) p(\gamma) \, d\gamma.
\]
In the special case we considered in Example 17.2.4, the deterministic Laplace integral is the integral
\[
\mu(\theta) = \int_{-\infty}^{\infty} \exp \left(-n - n \log(2\pi) - \frac{n}{2}(\gamma^6 + \gamma^8)\right) p(\gamma) \, d\gamma.
\]

The following theorem relates the probabilistic Laplace integral to the deterministic Laplace integral.

**Theorem 17.2.6.** Let $\{P_{\theta} : \theta \in \Theta\}$ be a regular exponential family with $\Theta \subseteq \mathbb{R}^k$. Let $\Gamma \subseteq \mathbb{R}^d$ be an open set and $g : \Gamma \to \Theta$ a polynomial map. Let
\( \theta_0 = \gamma(\gamma_0) \) be the true parameter value. Assume that \( g^{-1}(\theta_0) \) is a compact set and that the prior density \( p(\gamma) \) is a smooth function on \( \Gamma \) that is positive on \( g^{-1}(\theta_0) \). Then
\[
\log \mu(X^{(1)}, \ldots, X^{(n)}) = \hat{\ell}_n - \frac{\lambda}{2} \log(n) + (m - 1) \log \log(n) + O_p(1)
\]
where the rational number \( \lambda \in (0, d] \) and the integer \( m \in [d] \) come from the deterministic Laplace integral via
\[
\log \mu(\theta_0) = n \ell_0(\theta_0) - \frac{\lambda}{2} \log(n) + (m - 1) \log \log(n) + O(1).
\]

The numbers \( \lambda \) and \( m \) are called the learning coefficient and its multiplicity, respectively. Note that they depend on the model, the prior, and the value \( \theta_0 \). Theorem 17.2.6 and generalizations are proven in [Wat09].

**Example 17.2.7** (Learning Coefficient of Monomial Curve). Concluding Example 17.2.4, recall that
\[
\mu(0) = \int_{-\infty}^{\infty} \exp(-n - n \log(2\pi) - \frac{n}{2}(\gamma^6 + \gamma^8) p(\gamma)) d\gamma.
\]
The asymptotics of \( \mu(0) \) are completely determined by the fact that in the integrand we have \( \gamma^6(1+\gamma^2) \) which is of the form a monomial times a strictly positive function. From this, we can easily see that
\[
\log(\mu(0)) = -n - n \log(2\pi) - \frac{1}{6} \log(n) + O(1).
\]
Thus \( \lambda = 1/6 \) and \( m = 1 \) in this example. The goal in subsequent sections will be to generalize this structure to more complicated examples.

Theorem 17.2.6 reduces the problem of determining the asymptotics of the marginal likelihood to the computation of the learning coefficient and multiplicity arising from the deterministic Laplace integral. Actually computing these numbers can be a significant challenge, and often depends on resolution of singularities of algebraic varieties, among other tools. We will discuss this in more detail in Section 17.3, and talk about applications in Section 17.4.

**17.3. The Real Log-Canonical Threshold**

The computation of the learning coefficient and its multiplicity remain a major challenge and there are but a few instances where these have been successfully worked out. One well-studied case is the case of reduced rank regression, where the learning coefficients associated to low rank matrices have been computed [AW05]. The case of monomial parametrizations was handled in [DLWZ17]. Our goal in this section is to expand the vocabulary enough that the statements of these results can be understood, and give a
sense of what goes into the proof. We also make the connection between the learning coefficients and the real log-canonical threshold.

Recall that the calculation of the learning coefficients amounts to understanding the asymptotics of the deterministic Laplace integral:

\[ \mu(\theta_0) = \int_\Gamma \exp (n \ell_0(g(\gamma))) \, p(\gamma) \, d\gamma. \]

The calculation of the learning coefficient can be generalized to arbitrary Laplace integrals of the form

\[ \int_\Gamma \exp(-nH(\gamma))\phi(\gamma) \, d\gamma, \]

where \( H \) is a nonnegative analytic function on \( \Gamma \). The asymptotic behavior of this generalized Laplace integral is related to the poles of another function associated to \( f \) and \( \phi \).

**Definition 17.3.1.** The \( \Gamma \subseteq \mathbb{R}^k \) be an open semianalytic set. Let \( f : \Gamma \to \mathbb{R}_{\geq 0} \) be a nonnegative analytic function whose zero set \( V(f) \cap \Gamma \) is compact and nonempty. The function

\[ \zeta(z) = \int_{\Theta} f(\theta)^{-z/2} \phi(\theta) \, d\theta, \quad \Re(z) \leq 0 \]

can be analytically continued to a meromorphic function on the complex plane. The poles of the analytic continuation are real and positive. Let \( \lambda \) be the smallest pole and \( m \) be its multiplicity. The pair \((\lambda, m)\) is called the real log-canonical threshold of the pair of functions \( f, \phi \). We use the notation

\[ \text{RLCT}(f; \phi) = (\lambda, m) \]

to denote this.

**Example 17.3.2 (RLCT of Monomial Function).** Consider a monomial function \( H(\gamma) = \gamma^{2u} = \gamma_1^{2u_1} \cdots \gamma_k^{2u_k} \), let \( \phi \equiv 1 \), and let \( \Gamma = [0, 1]^k \). The zeta function is

\[ \zeta(z) = \int_{[0,1]^k} \gamma_1^{-zu_1} \cdots \gamma_k^{-zu_k} \, d\gamma \]

whose analytic continuation is the meromorphic function

\[ \zeta(z) = \prod_{i=1}^k \frac{1}{1 - zu_i}. \]

The smallest pole of \( \zeta(z) \) is \( \min_{i \in [k]} 1/u_i \), and its multiplicity is the number of times the corresponding exponent appears in the vector \( u \).
The real log-canonical threshold of a function $H$ is determined by its zero set and the local behavior of $H$ as $\gamma$ moves away from the zero set. To use this to understand the asymptotics of the deterministic Laplace integral, we need to rewrite the deterministic Laplace integral

\begin{equation}
\int_{\Gamma} \exp \left(n \ell_0(g(\gamma))\right) p(\gamma) \, d\gamma.
\end{equation}

so that we have a function that achieves the lower bound of zero in the integral. To do this, we introduce the Kullback-Leibler divergence:

$$K_q(\theta) = \int \log \frac{q(x)}{p(x|\theta)} q(x) \, dx$$

and the entropy:

$$S(q) = \int q(x) \log q(x) \, dx.$$ 

Then we can rewrite (17.3.1) as

\begin{equation}
\int_{\Gamma} \exp \left(nS(q) - nK_q(g(\gamma))\right) p(\gamma) \, d\gamma.
\end{equation}

where $q$ denotes the data generating distribution. Since $S(q)$ is a constant, the interesting asymptotics are governed by the asymptotics of the integral

\begin{equation}
\int_{\Gamma} \exp \left(-nK_q(g(\gamma))\right) p(\gamma) \, d\gamma.
\end{equation}

In general, the deterministic Laplace integral is connected to the real log-canonical threshold by the following theorem.

**Theorem 17.3.3.** [Wat09, Thm 7.1] Let $H(\gamma)$ be a nonnegative analytic function on the open set $\Gamma$. Suppose that $V(f) \cap \Theta$ is compact and nonempty, and $\phi$ is positive on $\Gamma$. Then

$$\log \int_{\Gamma} \exp(-nH(\gamma))\phi(\gamma) \, d\gamma = -\frac{\lambda}{2} \log n + (m - 1) \log \log n + O(1)$$

where $(\lambda, m) = \text{RLCT}(H, \phi)$.

Theorem 17.3.3 reduces the calculation of the learning coefficients to the calculation of real log-canonical thresholds. When it comes to actually computing this value, there are a number of tricks that can simplify the calculations. A first observation is the following:

**Proposition 17.3.4.** Let $H : \Gamma \rightarrow \mathbb{R}_{\geq 0}$ and $K : \Gamma \rightarrow \mathbb{R}_{\geq 0}$ be two nonnegative functions such that there are positive constants $C_1, C_2$ with $C_1K(\gamma) \leq H(\gamma) \leq C_2K(\gamma)$ for all $\gamma$ in a neighborhood of the set where $H$ attains its minimum. Then

$$\text{RLCT}(H; \phi) = \text{RLCT}(K; \phi).$$
17.3. The Real Log-Canonical Threshold

Functions $H$ and $K$ from Proposition 17.3.4 are called equivalent, denoted $H \sim K$. Proposition 17.3.4 can allow us to replace a complicated function $H$ with an equivalent function $K$ that is easier to analyze directly.

**Proof.** Suppose that $H \sim K$. First suppose that the condition

$$C_1K(\gamma) \leq H(\gamma) \leq C_2K(\gamma)$$

holds for all $\gamma \in \Gamma$. Let

$$a(n) = \int_\Gamma \exp(-nH(\gamma)) \phi(\gamma) \, d\gamma \quad \text{and} \quad b(n) = \int_\Gamma \exp(-nK(\gamma)) \phi(\gamma) \, d\gamma$$

be the associated deterministic Laplace integrals. To say that $\text{RLCT}(H; \phi) = (\lambda, m)$ is equivalent to saying that

$$\lim_{n \to \infty} a(n) \frac{(\log n)^m}{n^\lambda}$$

exists and is a positive constant. Now suppose that $\text{RLCT}(K; \phi) = (\lambda, m)$. Then

$$\lim_{n \to \infty} a(n) \frac{(\log n)^{m-1}}{n^\lambda} = \lim_{n \to \infty} \frac{(\log n)^{m-1}}{n^\lambda} \int_\Gamma \exp(-nH(\gamma)) \phi(\gamma) \, d\gamma$$

$$\leq \lim_{n \to \infty} \frac{(\log n)^{m-1}}{n^\lambda} \int_\Gamma \exp(-nC_1K(\gamma)) \phi(\gamma) \, d\gamma$$

$$\leq \lim_{n \to \infty} b(C_1n) \frac{(\log n)^{m-1}}{n^\lambda}.$$

The last term can be seen to converge to a positive constant after a change of variables and using $\text{RLCT}(K; \phi) = (\lambda, m)$. Similarly,

$$\lim_{n \to \infty} a(n) \frac{(\log n)^{m-1}}{n^\lambda} = \lim_{n \to \infty} \frac{(\log n)^{m-1}}{n^\lambda} \int_\Gamma \exp(-nH(\gamma)) \phi(\gamma) \, d\gamma$$

$$\geq \lim_{n \to \infty} \frac{(\log n)^{m-1}}{n^\lambda} \int_\Gamma \exp(-nC_2K(\gamma)) \phi(\gamma) \, d\gamma$$

$$\geq \lim_{n \to \infty} b(C_1n) \frac{(\log n)^{m-1}}{n^\lambda}.$$

The last term also converges to a positive constant after change of variables.

This implies that $\lim_{n \to \infty} a(n) \frac{(\log n)^m}{n^\lambda}$ converges to a positive constant, and so $\text{RLCT}(H; \phi) = (\lambda, m)$. To prove the general form of the theorem, note that we can localize the integral in a neighborhood containing the zero set of $H(\gamma)$, since the integrand rapidly converges to zero outside this neighborhood as $n$ goes to infinity. $\square$
One consequence of Proposition 17.3.4 is that in many circumstances, we can replace the Kullback-Leibler divergence with a function of the form $\|g(\gamma) - g(\gamma_0)\|^2$ which is considerably easier to work with. The typical argument is as follows: Consider the case of a Gaussian model where the map $g(\gamma) \to P\mathbb{D}_m$ is parametrizing some algebraic subset of covariance matrices, and where we fix the mean vectors under consideration to always be zero. The Kullback-Leibler divergence in this case has the form

$$K(\gamma) = \frac{1}{2} \left( \text{tr}(g(\gamma)^{-1} \Sigma_0) - m - \log \left( \frac{\det \Sigma_0}{\det g(\gamma)} \right) \right)$$

where $\Sigma_0 = g(\gamma_0)$. Since we have assumed $\Sigma_0$ is positive definite, the function

$$(17.3.2) \quad \Sigma \mapsto \frac{1}{2} \left( \text{tr}(\Sigma^{-1} \Sigma_0) - m - \log \left( \frac{\det \Sigma_0}{\det \Sigma} \right) \right)$$

is zero when $\Sigma = \Sigma_0$, and positive elsewhere, and its Hessian has full rank and is positive definite. This means that in a neighborhood of $\Sigma_0$, the function (17.3.2) is equivalent to the function

$$\|\Sigma - \Sigma_0\|^2$$

and so the Kullback-Leibler divergence is equivalent to the

$$H(\gamma) = \|g(\gamma) - \Sigma_0\|^2.$$

Following [DLWZ17] we now work with the special case that $g(\gamma)$ is a monomial map, and consider a general function $H$ of the form

$$(17.3.3) \quad H(\gamma) = \sum_{i=1}^{k} (\gamma^{u_i} - c_i)^2, \quad \gamma \in \Gamma.$$ 

The numbers $c_i$ are constants where $c_i = \gamma_0^{u_i}$ for some particular true parameter value $\gamma_0$. The vector $u_i = (u_{i1}, \ldots, u_{id})$ is a vector of nonnegative integers.

Let $t$ be the number of summands that have $c_i \neq 0$. We can organize the sum so that $c_1, \ldots, c_t \neq 0$ and $c_{t+1} = \cdots = c_k = 0$. Furthermore, we can reorder the coordinates $\gamma_1, \ldots, \gamma_d$ so that $\gamma_1, \ldots, \gamma_s$ are the parameters that appear among that monomials $\gamma^{u_1}, \ldots, \gamma^{u_t}$, and that $\gamma_{s+1}, \ldots, \gamma_d$ do not appear among these first $t$ monomials.

Define the nonzero part $H^1$ of $H$ as

$$H^1(\gamma) = \sum_{i=1}^{t} (\gamma^{u_i} - c_i)^2$$
17.3. The Real Log-Canonical Threshold

and the zero part $H^0$ of $H$ by

$$H^0(\gamma) = \sum_{i=\ell+1}^{k} \prod_{j=s+1}^{d} \gamma_{ij}^{2u_{ij}}.$$  

**Definition 17.3.5.** The Newton polyhedron of the zero part is the unbounded polyhedron

$$\text{Newt}^+(H^0(\gamma)) = \text{Newt}(H^0(\gamma)) + \mathbb{R}^{d-s}_+. $$

In Definition 17.3.5, $\text{Newt}(H^0(\gamma)) \subseteq \mathbb{R}^{d-s}$ denotes the ordinary Newton polytope obtained by taking the convex hull of the exponent vectors. The + in the definition of the Newton polyhedron denotes the Minkowski sum of the Newton polytope with the positive orthant. Note that the Newton polyhedron naturally lives in $\mathbb{R}^{d-s}$ since all the terms in $H^0(\gamma)$ only depend on the parameters $\gamma_{s+1}, \ldots, \gamma_d$.

Let $1 \in \mathbb{R}^{d-s}$ denote the all one vector. The 1-distance of the Newton polyhedron is the smallest $\delta$ such that $\delta 1 \in \text{Newt}^+(H^0(\gamma))$. The associated multiplicity is the codimension of the smallest face of $\text{Newt}^+(H^0(\gamma))$ that contains $\delta 1$.

**Theorem 17.3.6.** [DLWZ17, Theorem 3.2] Suppose that $\Gamma = \prod_{i=1}^{d} [a_i, b_i]$ is a product of intervals, let $\Gamma_1$ and $\Gamma_0$ be the projections of $\Gamma$ onto the first $s$ and last $d-s$ coordinates, respectively. Consider the function $H(\gamma)$ of the form (17.3.3). Suppose that the zero set $\{\gamma \in \Gamma : H(\gamma) = 0\}$ is non-empty and compact. Let $\phi : \Gamma \to \mathbb{R}$ be a smooth positive bounded function on $\Gamma$. Then

$$\text{RLCT}_\Gamma(H; \phi) = (\lambda_0 + \lambda_1, m) $$

where $\lambda_1$ is the codimension of the set $\{\gamma \in \Gamma_1 : H^1(\gamma) = 0\}$ in $\mathbb{R}^s$, and $2/\lambda_0$ is the 1-distance of the Newton polyhedron $\text{Newt}^+(H^0(\gamma))$ with associated multiplicity $m$. We take $\lambda_0 = 0$ and $m = 1$ in the case that $H$ has no zero part.

The proof of Theorem 17.3.6 is contained in Appendix A of [DLWZ17]. The proof is involved and beyond the scope of this book. The main idea is to show that the real log-canonical threshold of $H$ breaks into the sum of the real log-canonical thresholds of $H^0$ and $H^1$. The proof that the real log-canonical threshold of $H^1$ is $\lambda_1$ appears in Appendix A of [DLWZ17]. The proof that the real log-canonical threshold of $H^0$ is $\lambda_0$ with multiplicity $m$ is older and appears in [AGZV88] and [Lin10].

**Example 17.3.7 (Positive Definite Quadratic Form).** Consider the case where $H(\gamma) = \gamma_1^2 + \cdots + \gamma_d^2$. The integral $\int_{\Gamma} \exp(-nH(\gamma)) d\gamma$ is a Gaussian integral. It can be seen that the asymptotics governing such an integral yield that $\text{RLCT}_\Gamma(H; 1) = (d, 1)$. This is matched by the application of Theorem
17.3.6. Indeed, the zero part of \( H \) is \( H \), so \( \lambda_1 = 1 \). The 1-distance to \( \text{Newt}^+(H^0(\gamma)) \) is \( 2/d \), and this meets in the center of a face of codimension 1, whereby \( \lambda_0 = d \) and \( m = 1 \).

**Example 17.3.8 (Monomial Function).** Consider the situation in Example 17.3.2 where \( H(\gamma) = \gamma^{2u} \) is a monomial. The zero part of \( H \) is just \( H \), and \( H^1 \equiv 0 \). The Newton polyhedron \( \text{Newt}^+(H^0(\gamma)) \) is just a shifted version of the positive orthant, with apex at the point \( 2u \). The smallest \( \delta \) for which \( \delta T \in \text{Newt}^+(H^0(\gamma)) \) is for \( \delta = 2 \max_i u_i \), and the codimension of the face in which this intersection occurs is the number of coordinates that achieve the maximum value. Hence the real log-canonical threshold calculated via Theorem 17.3.6 agrees with our previous calculation in Example 17.3.2.

The general computation of the real log-canonical threshold of a function \( H \) is quite challenging, and depends on the real resolution of singularities. In principle, this can be performed in some computer algebra systems, but in practice the large number of charts that arises can be overwhelming and difficult to digest, if the computations terminate at all. See [Wat09] and [Lin10] for more details about how to apply the resolution of singularities to compute the real log-canonical threshold. Usually we need specific theorems to say much about the learning coefficients.

In the remainder of this section, we apply Theorem 17.3.6 to the case of the 1-factor model which is a special case of the factor analysis model that we have seen in previous chapters. This follows the more general results of [DLWZ17] for Gaussian hidden tree models.

Recall that the factor analysis model \( F_{m,s} \) is the family of multivariate normal distributions \( N(\mu, \Sigma) \) where \( \Sigma \) is restricted to lie in the the set

\[
F_{m,s} = \{ \Omega + \Lambda \Lambda^T : \Omega \gg 0 \text{ diagonal and } \Lambda \in \mathbb{R}^{m \times s} \}.
\]

See Examples 6.5.4 and 14.2.3. We focus on the case that \( s = 1 \), the so-called 1-factor model. The model is described parametrically by the map

\[
\tilde{g} : \mathbb{R}_{>0}^m \times \mathbb{R}^m \rightarrow PD_m; \quad \tilde{g}_{ij}(\omega, \ell) = \begin{cases} 
\omega_i + \ell^2, & \text{if } i = j \\
\ell_i \ell_j & \text{if } i \neq j.
\end{cases}
\]

Note that in previous sections we used \( \lambda \) for the parameters. We choose to use \( \ell \) here to avoid confusion with the \( \lambda \) from the real log-canonical threshold. While this expression for \( \tilde{g} \) is not monomial in this example, we can replace it with one that is monomial and parametrizes the same algebraic variety. In particular, we can replace \( \tilde{g} \) with the map \( g \) where \( g_{ii}(\omega, \ell) = \omega_i \) otherwise \( g_{ij} = \tilde{g}_{ij} \) for \( i \neq j \).

For a Gaussian algebraic exponential family, parametrized by a function \( g \), computing the learning coefficients/ real log-canonical thresholds amounts
17.3. The Real Log-Canonical Threshold

To understanding the asymptotics of the deterministic Laplace integral where $H$ is the function

$$H(\gamma) = \|g(\gamma) - g(\gamma_0)\|_2^2$$

as we previously discussed. Since the function $g(\omega, \ell)$ is a monomial function for the 1-factor model, we can use Theorem 17.3.6 to determine the real log-canonical thresholds at all points. To do this, we need to understand what are the zero patterns possible in $g(\gamma_0)$. Since the entries of $\omega^*$ must be positive, the zero coordinates of $g(\gamma_0)$ are determined by which entries of $\ell^*$ are zero. Without loss of generality, we can suppose that the first $t$ coordinates of $\ell^*$ are nonzero and the last $m - t$ coordinates are zero. This means that $H_1$ has the form

$$H_1(\omega, \ell) = \sum_{i=1}^{m} (\omega_i - \omega^*_i)^2 + \sum_{1 \leq i < j \leq t} (\ell_i \ell_j - \ell^*_i \ell^*_j)^2$$

while the zero part, $H_0$ is

$$H_0(\ell) = \sum_{i=t+1}^{m} t \ell_i^2 + \sum_{t+1 \leq i < j \leq m} \ell_i^2 \ell_j^2$$

when $t \geq 2$ and

$$H_0(\ell) = \sum_{1 \leq i < j \leq m} \ell_i^2 \ell_j^2$$

where $t = 0$ or 1. (Note that if $t = 0$ or 1, then $\ell_i^2 \ell_j^2 = 0$ for all pairs $i \neq j$.)

We must compute the numbers $\lambda_0, \lambda_1$ and $m$ associated to $H_1$ and $H_0$ to determine the real log-canonical thresholds. To compute $\lambda_1$, we need to know the codimension of the fiber

$$\{(\omega, \ell) : \omega_i = \omega^*_i, i \in [m], \ell_i \ell_j = \ell^*_i \ell^*_j, 1 \leq i < j \leq t\} \subseteq \mathbb{R}^{m+t(t-1)/2}$$

where all the $\ell^*_i \neq 0$. There are a few cases to consider.

1. If $t > 2$ the fiber always consists of finitely many points (this follows from Example 16.1.4), so $\lambda_1 = m + t$ in those cases.
2. If $t = 2$, the fiber is one dimensional (since we can choose $\ell_2$ arbitrarily and set $\ell_1 = \ell^*_1 \ell^*_2 / \ell_2$) which implies that $\lambda_1 = m + 1$.
3. If $t = 1, 0$, it is clear that $\lambda_1 = m$.

For the calculation of $\lambda_0$, there are also two cases to consider.

1. If $t \geq 2$, then Newt+(H_0(\ell)) has vertices consisting of the $m - t$ vectors $2e_{t+1}, \ldots, 2e_m$. The 1-distance is $2/(m-t)$ and the codimension of the face of intersection is 1, so that $\lambda_0 = m - t$ and $m = 1$.
(2) If \( t = 0,1 \), then Newt\( ^+ (H^0 (\ell)) \) has vertices consisting of the \( \binom{m}{2} \) vectors \( 2e_i + 2e_j \) for \( 1 \leq i < j \leq m \). The 1-distance is \( 4/m \), and the codimension of the face of intersection is 1, so that \( \lambda_0 = m/2 \) and \( m = 1 \).

Putting these results together, we arrive at the following result about the learning coefficients of distributions in the 1-factor model.

**Theorem 17.3.9.** Consider the 1-factor model \( F_{m,1} \) with \( p(\omega, \ell) \) a smooth prior density positive on all of \( \mathbb{R}_+^m \times \mathbb{R}^m \). Let \( X^{(1)}, \ldots, X^{(n)} \) be i.i.d. observations that are Gaussian with mean zero and covariance matrix \( \Sigma_0 = g(\omega_0, \lambda_0) \) be the true parameter matrix. Then

\[
\log \mu(X^{(1)}, \ldots, X^{(n)}) = \hat{\ell}_n - \lambda \frac{1}{2} \log(n) + O_p(1)
\]

where

\[
\lambda = \begin{cases} 
\frac{3m}{2} & \text{if } \Sigma_0 \text{ is diagonal}, \\
2m - 1 & \text{if } \Sigma_0 \text{ has exactly one nonzero off diagonal entry}, \\
2m & \text{otherwise}.
\end{cases}
\]

**17.4. Information Criteria for Singular Models**

In previous sections, we have explored the general fact that the asymptotics of the marginal likelihood integral might change if the true parameter lies at a singular point of the model, or the Jacobian of the parametrization is not invertible. Since model selection criteria like the BIC are derived from the asymptotics of the marginal likelihood integral, one might wonder how to incorporate the non-standard asymptotics that the marginal likelihood can exhibit at singular points in an information criterion. One major challenge is that the asymptotics of the marginal likelihood depend on knowing the true parameter value. Since the true parameter value is never known, this requires some care to develop a meaningful correction. A solution to this problem was proposed by Drton and Plummer [DP17]. We describe the broad strokes of this correction, and illustrate the idea with the 1-factor model and some of its submodels.

To tell the story in full generality, we consider not just a pair of statistical models \( M_1 \subseteq M_2 \) but, instead, an entire family of models. In particular, suppose that we have an index set \( I \), and for each \( i \in I \) a model \( M_i \). Each of these models is naturally identified with its parameter space, so that \( M_i \subseteq \mathbb{R}^k \) where \( \mathbb{R}^k \) is a finite dimensional real parameter space in which all models live. We assume that the models are algebraic exponential families, and in particular each \( M_i \) is a semialgebraic set whose Zariski closure is irreducible. We have a prior distribution \( P(M_i) \) on models.
17.4. Information Criteria for Singular Models

Introduce a partial ordering on the index set $I$ so that $j \preceq i$ if and only if $\mathcal{M}_j \subseteq \mathcal{M}_i$. The following assumption appears to hold in all models where the real log-canonical thresholds are known.

**Assumption 17.4.1.** Let $\mathcal{M}_j \subseteq \mathcal{M}_i$, and suppose that the Zariski closures of both models are irreducible algebraic varieties. Suppose that the prior distribution $\phi$ is smooth and positive on $\mathcal{M}_j$. Then there exists a dense relatively open subset $U_{ij} \subseteq \mathcal{M}_j$ and numbers $\lambda_{ij}$ and $m_{ij}$ such for all $\theta_0 \in U_{ij}$

$$\log \mu(X^{(1)}, \ldots, X^{(n)}) = \hat{\theta}_n - \frac{\lambda_{ij}}{2} \log(n) + (m_{ij} - 1) \log \log(n) + O_p(1).$$

Following [DP17], we define

$$L'_{ij} = L(X^{(1)}, \ldots, X^{(n)} \mid \hat{\theta}_i, \mathcal{M}_i) \frac{\log(n)^{m_{ij} - 1}}{n^{\lambda_{ij}/2}},$$

where $\hat{\theta}_i$ is the maximum likelihood estimate of the model parameters in the model $\mathcal{M}_i$, and $L$ denotes the likelihood function evaluated at this optimizer. Note that this makes $L'_{ij}$ into an approximation to the marginal likelihood assuming that the true parameter lies on the submodel $U_{ij}$ of $\mathcal{M}_i$. Here we have ignored the term $O_p(1)$, of random variables of bounded probability.

Drton and Plummer define their score recursively by the equation system:

$$L'(\mathcal{M}_i) = \frac{1}{\sum_{j \leq i} L'(\mathcal{M}_j)P(\mathcal{M}_j)} \sum_{j \leq i} L'_{ij}L'(\mathcal{M}_j)P(\mathcal{M}_j), \quad i \in I. \quad (17.4.1)$$

This recursive equation gives a final score to model $\mathcal{M}_i$ that is a weighted combination of scores associated to all submodels, where the weights are determined by the prior probability of those models and the approximation to the marginal likelihood, $L'_{ij}$, at generic points in the submodel $\mathcal{M}_j$. This score is a principled way to use the prior distribution to take into account the fact that the true data generating distribution might belong to a submodel (even though, in a single model $\mathcal{M}_i$, the probability that the true distribution belongs to any proper submodel is typically zero).

Note that Equation 17.4.1 for the score $L'(\mathcal{M}_i)$ of model $\mathcal{M}_i$ includes $L'(\mathcal{M}_i)$, so after clearing denominators, and rearranging the equation system we arrive at the system of polynomial equations

$$\sum_{j \leq i}(L'(\mathcal{M}_i) - L'_{ij})L'(\mathcal{M}_j)P(\mathcal{M}_j) = 0, \quad i \in I. \quad (17.4.2)$$

A simple algebraic argument shows that this system of quadratic equations has a unique positive solution for the $L'(\mathcal{M}_i)$.

**Proposition 17.4.2.** The Drton-Plummer recursion (17.4.1) has a unique positive solution, provided that each submodel has positive prior probability.
Proof. We proceed by induction on the depth of a submodel $i$ in the partial order. If $i$ is a minimal element of the partial order then Equation 17.4.2 reduces to the single equation

$$\left( L'(M_i) - L'_{ii} \right) L'(M_i) P(M_i) = 0$$

which has the unique positive solution $L'(M_i) = L'_{ii}$.

Now consider a non-minimal index $i$. By the inductive hypothesis, we can assume that for all $j \prec i$ there is a unique value for $L'(M_j)$. Then the only unknown value in the $i$th copy of Equation 17.4.2 is $L'(M_i)$, and this equation is quadratic in $L'(M_i)$. Letting

$$b_i = -L'_{ii} + \sum_{j \prec i} L'(M_j) \frac{P(M_j)}{P(M_i)}$$

and

$$c_i = \sum_{j \prec i} L'_{ij} L'(M_j) \frac{P(M_j)}{P(M_i)}$$

we see that $L'(M_i)$ must be the solution to the quadratic equation

$$L'(M_i)^2 + b_i L'(M_i) - c_i = 0.$$ 

Since all the elements in $c_i$ are positive, we see that $c_i$ is also positive. Then the quadratic formula shows that there is a unique positive solution, which is

$$L'(M_i) = \frac{1}{2} \left( -b_i + \sqrt{b_i^2 + 4c_i} \right).$$

The quantities $L'(M_i)$ can be seen as an approximation to the marginal likelihood in the case of a singular model. For this reason, Drton-Plummer [DP17] refer to log $L'(M_i)$ as the singular Bayesian information criterion or sBIC. Note that unlike the ordinary BIC, the sBIC score of a model $M_i$ depends on all the other models that are being considered in the associated model selection problem. In particular, it might change if we redo an analysis by adding more models to consider.

Example 17.4.3 (0-factor and 1-factor Models). Suppose we are considering the model selection problem between two multivariate normal models, the one factor model $F_{m,1}$ and the model $F_{m,0}$ of complete independence. We only have two models in this case and the partial ordering on the indexes is a total order. The corrections involved in the diagonal terms $L'_{00}$ and $L'_{11}$ are just the exponentials of the ordinary BIC scores for these models. In particular

$$L'_{00} = L(X^{(1)}, \ldots, X^{(n)} | \hat{\theta}_0, F_{m,0}) \frac{1}{n^{m/2}}.$$
and

\[ L'_{11} = L(X^{(1)}, \ldots, X^{(n)} | \hat{\theta}_1, F_{m,1}) \frac{1}{\mu^m}. \]

On the other hand, to figure out the form for the \( L'_{10} \) term we need to use Theorem 17.3.9. In particular, note that the real log-canonical threshold at points of \( F_{m,0} \) as a submodel of \( F_{m,1} \) is \( 3m/2 \). Thus, we have

\[ L'_{11} = L(X^{(1)}, \ldots, X^{(n)} | \hat{\theta}_1, F_{m,1}) \frac{1}{\mu^{3m/4}}. \]

Supposing that we take \( P(F_{m,0}) = P(F_{m,1}) = 1/2 \), we get that

\[ L'(F_{m,0}) = L'_{00} \]

and

\[ L'(F_{m,1}) = \frac{1}{2} \left( L'_{11} - L'(F_{m,0}) + \sqrt{(L'_{11} - L'(F_{m,0}))^2 + 4L'_{10}L'(F_{m,0})} \right). \]

**Example 17.4.4 (1-factor Model with Zeros).** Consider a variation on the model selection problem concerning the 1-factor model from Example 17.4.3, where now we consider the model selection problem of determining which of the coordinates of the vector \( \beta \) are zero. That is, for each \( S \subseteq [m] \), we get a 1-factor type model

\[ F_{m,1,S} = \{ \Omega + \ell \ell^T \in PD_m : \Omega \succ 0 \text{ diagonal } \ell \in \mathbb{R}^m, \ell_i = 0 \text{ if } i \notin S \}. \]

Note that if \( S \) is a singleton, then \( F_{m,1,S} = F_{m,1,\emptyset} = F_{m,0} \). So we will ignore singleton sets. Also, we have the dimension formula of

\[ \dim F_{m,1,S} = \begin{cases} 
  m & \text{if } S = \emptyset \text{ or } \#S = 1, \\
  m + 1 & \text{if } \#S = 2, \\
  m + \#S & \text{if } \#S > 2 
\end{cases} \]

Hence, to use the singular BIC, we need to compute the appropriate \( \lambda_{S,T} \) and \( m_{S,T} \) coefficients for each \( T \subseteq S \) to use the recursive formula (17.4.1) to find the sBIC score. A straightforward generalization of Theorem 17.3.9 shows that for each \( T \subseteq S \), with \#S and \#T not 1, we have the following

\[ \lambda_{S,T} = \begin{cases} 
  m & \text{if } S = \emptyset \\
  m + 1 & \text{if } \#S = 2 \\
  m + \#S - \frac{T}{2} & \text{if } \#S > 2 \text{ and } T = \emptyset \\
  m + \#S - 1 & \text{if } \#S > 2 \text{ and } \#T = 2 \\
  m + \#S & \text{if } \#S > 2 \text{ and } \#T > 2. 
\end{cases} \]
17.5. Exercises

Exercise 17.1. Continue the analysis of the data in Example 17.1.4 for all hierarchical models, and for all directed acyclic graphical models on 3 random variables. Which models do the AIC and BIC select among these classes of models?

Exercise 17.2. Determine the dimensions, number of vertices, and number of facets of the model selection polytopes \( MS_3 \) and \( MS_4 \).

Exercise 17.3. Determine the general form of the deterministic Laplace integral in the case that the model is a parametrized model for discrete random variables.

Exercise 17.4. Extend Theorem 17.3.9 to general Gaussian hidden tree models. That is, consider a rooted tree \( T \) and the Gaussian hidden variable graphical model where all interior vertices are hidden. Determine all real log-canonical thresholds of all covariance matrices in the model, assuming that the prior distribution has a product form.
Chapter 18

MAP Estimation and Parametric Inference

In this chapter we consider the Bayesian procedure of maximum a posteriori estimation (MAP estimation). Specifically, we consider the estimation of the highest probability hidden states in a hidden variable model given fixed values of the observed states and either fixed values or a distribution on the model parameters. The estimation of hidden states given observed states and fixed parameter values is a linear optimization problem over potentially exponentially many possibilities for the hidden states. In graphical models based on trees, or other graphs with low complexity, it is possible to compute the MAP estimate rapidly using message passing algorithms. The most famous algorithm of this type is the Viterbi algorithm, which is specifically developed for hidden Markov models [Vit67]. More general message passing algorithms include the belief propagation algorithm [Pea82] and the generalized distributive law [AM00]. Computations for these message passing algorithms can be expressed in the context of tropical arithmetic, which we will introduce in Section 18.2. This connection appears to originate in the work of Pachter and Sturmfels [PS04b].

To perform true MAP estimation requires a prior distribution over the parameters and the calculation of the posterior distribution over all possible hidden state vectors. These posterior probability distributions are typically approximated using Monte Carlo algorithms. As we shall see, normal fans of polytopes can be an important tool in analyzing the sensitivity of the basic MAP estimate to changes in the underlying parameters. This variation on the problem is usually called parametric inference [FBSS02]. The computation of the appropriate normal fan for the parametric inference problem
can also be approached using a recursive message passing procedure akin to the Viterbi algorithm, but in an algebra where elements are the polytopes themselves (or, their normal fans). We illustrate these ideas in the present chapter focusing on the special case of the hidden Markov model.

18.1. MAP Estimation General Framework

We first consider a general setup for discussing the Bayesian problem of maximum a posteriori estimation in hidden variable models. We show how this leads to problems in optimization and polyhedral geometry in the discrete case, and how Newton polytopes can be useful for solving the MAP estimation problem and for sensitivity analysis.

Let $\Theta \subseteq \mathbb{R}^d$ be a parameter space and consider a statistical model $\mathcal{M} = \{p_\theta : \theta \in \Theta\}$, which is a model for a pair of random vectors $(X, Y)$ with $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. In particular, $p_\theta(x, y)$ denotes the probability of observing $X = x$ and $Y = y$ under the specified value $\theta$ of the parameter of the model. In the most basic setting in which MAP estimation is usually applied we will assume that $Y$ is a hidden variable and $X$ is an observed variable. Furthermore, we assume that $\pi(\theta)$ is a prior distribution on the parameter space $\Theta$. The most typical definition of the maximum a posteriori estimate (MAP) in the Bayesian literature is the MAP estimate for the parameter $\theta$ given the data $x$, which is defined as

$$\theta_{MAP} = \arg\max_{\theta \in \Theta} p_\theta(x) \pi(\theta).$$

Thus the maximum a posteriori estimate is the value of $\theta$ that maximizes the posterior distribution:

$$p_\theta(x) \pi(\theta) \int p_\theta(x) \pi(\theta) d\theta.$$

Since the MAP estimate is a maximizer of a certain function associated to the model and data it could be seen as a Bayesian analogue of the maximum likelihood estimate.

To make the connection to discrete optimization and algebraic statistics, we will be interested in a variation on the MAP estimate, where the hidden variables $Y$ are treated as unknown parameters.

**Definition 18.1.1.** The maximum a posteriori estimate (MAP estimate) of $y$ given $x$ and the prior distribution $\pi(\theta)$ is the value of $y \in \mathcal{Y}$ such that

$$\int_{\theta \in \Theta_y} p_\theta(x, y) \pi(\theta) d\theta$$

where

$$\Theta_y = \left\{ \theta \in \Theta : \arg\max_{z \in \mathcal{Y}} p_\theta(x, z) = y \right\}.$$
A special case that will help to illuminate the connection to algebraic statistics is the case that the prior distribution \( \pi(\theta) \) is a point mass concentrated on a single value \( \theta_0 \). In this case, the optimization problem in the definition of the MAP estimate becomes

\[
y_{MAP} = \arg \max_{y \in \mathcal{Y}} p_{\theta_0}(x, y).
\]

We further assume that the density function \( p_\theta(x, y) \) can be expressed in a product form in terms of the model parameters \( \theta \) (or simple functions of \( \theta \)). For example, if the joint distribution of \( X \) and \( Y \) is given by a discrete exponential family, then the density function \( p_\theta(x, y) \) has the form

\[
p_\theta(x, y) = \frac{1}{Z(\theta)} \theta^{a_{x,y}}
\]

where \( a_{x,y} \) is an integer vector. A similar product expression that is useful in this context is for directed graphical models where \( X = (X_1, \ldots, X_{m_1}) \) \( Y = (Y_1, \ldots, Y_{m_2}) \) and \( p_\theta(x, y) \) has the form

\[
p_\theta(x, y) = \prod_{i=1}^{m_1} P(x_i \mid (x, y)_{pa(x_i)}) \times \prod_{i=1}^{m_2} P(y_i \mid (x, y)_{pa(y_i)}).
\]

Either of these forms are useful for what we discuss next for MAP estimation.

For the remainder of this section we restrict to the first form where \( p_\theta(x, y) = \frac{1}{Z(\theta)} \theta^{a_{x,y}} \). Then, in the MAP estimation problem with \( \pi \) concentrated at \( \theta_0 \), we are computing

\[
\arg \max_{y \in \mathcal{Y}} p_{\theta_0}(x, y) = \arg \max_{y \in \mathcal{Y}} \frac{1}{Z(\theta_0)} \theta_0^{a_{x,y}}
\]

\[
= \arg \max_{y \in \mathcal{Y}} \theta_0^{a_{x,y}}
\]

\[
= \arg \max_{y \in \mathcal{Y}} (\log \theta_0)^T a_{x,y}.
\]

This follows because \( Z(\theta_0) \) is a constant and does not affect which \( y \in \mathcal{Y} \) is a maximizer, and because \( \log \) is an increasing function. Letting \( c = (\log \theta_0)^T \) and \( P_x = \text{conv}\{a_{x,y} : y \in \mathcal{Y}\} \) this optimization problem is equivalent to

\[
\arg \max_{y \in \mathcal{Y}} p_{\theta_0}(x, y) = \arg \max_{v \in P_x} c^Tv.
\]

This assumes that we can recover \( y \) from the exponent vector \( a_{x,y} \). With this assumption, MAP estimation is a linear optimization problem over the polytope \( P_x \).

The polytope \( P_x \) that arises is not an arbitrary polytope but one that is closely related to the algebraic structure of the underlying probabilistic
model. Recall that the marginal probability of $X$ has the density

$$p_\theta(x) = \sum_{y \in Y} p_\theta(x, y).$$

For our densities that have a product form, we can write this as

$$p_\theta(x) = \frac{1}{Z(\theta)} \sum_{y \in Y'} \theta^{a_{x,y}}.$$

The monomials appearing in this sum are exactly the monomials whose exponent vectors appear in the convex hull describing the polytope $P_x$. In summary:

**Proposition 18.1.2.** The polytope $P_x$ is the Newton polytope of the unnormalized probability $\sum_{y \in Y} \theta^{a_{x,y}}$.

Recall that the Newton polytope of a polynomial is the convex hull of the exponent vectors appearing in that polynomial (see Definition 12.3.7).

In the setup we have seen so far, the Newton polytope itself is not really needed. What is of main interest is to calculate the maximizer quickly. As we will see, this is closely related to the problem of being able to evaluate the polynomial $\sum_{y \in Y} \theta^{a_{x,y}}$ quickly. However, the polytope $P_x$ is of interest if we want to understand what happens to the MAP estimate as $\theta_0$ varies, or if we have a more general prior distribution $\pi$. These points will be illustrated in more detail in the subsequent sections.

### 18.2. Hidden Markov Models and the Viterbi Algorithm

The hidden Markov model is a special case of a graphical model with hidden variables, where the graph has a specific tree structure. The underlying tree is sometimes called the *caterpillar tree*, which is illustrated in Figure 18.2.1.

Let $Y_1, \ldots, Y_m$ be discrete random variables, all with the same state space $[r]$, and let $X_1, \ldots, X_m$ be discrete random variables all with the same
state space \([s]\). We assume that the sequence \(Y_1, \ldots, Y_m\) form a homogeneous Markov chain with transition matrix \(A = (a_{i_1i_2})_{i_1, i_2 \in [r]} \in \mathbb{R}^{r \times r}\), i.e.

\[
P(Y_k = i_2 \mid Y_{k-1} = i_1) = a_{i_1i_2} \text{ for } k = 2, \ldots, m.
\]

To start the Markov chain we also need a root distribution \(\pi \in \Delta_{r-1}\), where

\[
P(Y_1 = i) = \pi_i.
\]

There is also an emission matrix \(B = (b_{i_k j_k})_{i_k \in [r], j_k \in [s]} \in \mathbb{R}^{r \times s}\) which gives the probability of observing \(X_k\) given the values of \(Y_k\), i.e.

\[
P(X_k = j_k \mid Y_k = i_k) = b_{i_k j_k} \text{ for } k = 1, \ldots, m.
\]

In addition to the Markov chain assumption, the hidden Markov model makes the further assumption of conditional independence:

\[
X_k \perp (Y_1, \ldots, Y_{k-1}, X_1, \ldots, X_{k-1}) \mid Y_k
\]

for \(k = 2, \ldots, m\). With these assumptions on the random variables, the joint distribution of \(Y = i = (i_1, \ldots, i_m)\) and \(X = j = (j_1, \ldots, j_m)\) is given by the following formula

\[
p_{i,j} = \pi_{i_1} \prod_{k=2}^{m} a_{i_{k-1}i_k} \prod_{k=1}^{m} b_{i_k j_k}.
\]

In the Hidden Markov model, the variables \(Y_1, \ldots, Y_m\) are hidden, so only the variables \(X_1, \ldots, X_m\) are observed. Hence, the probability distribution that would actually be observed is given by the formula

\[
p_j = \sum_{i \in [r]^m} p_{i,j} = \sum_{i \in [r]^m} \pi_{i_1} \prod_{k=2}^{m} a_{i_{k-1}i_k} \prod_{k=1}^{m} b_{i_k j_k}.
\]

So in the hidden Markov model we assume that there is a homogeneous Markov chain \(Y = (Y_1, \ldots, Y_m)\) that we do not get to observe. Instead we only observe \(X = (X_1, \ldots, X_m)\) which could be considered a noisy version of \(Y\), or rather, some type of “probabilistic shadow” of \(Y\).

**Example 18.2.1 (Occasionally Dishonest Casino).** We illustrate the idea of the hidden Markov model with a simple running example that we learned in [DEKM98]. Consider a casino that uses a fair die and a loaded die in some of their rolls at the craps table. The fair die has a probability distribution \((\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})\) of getting 1, 2, \ldots, 6, respectively whereas the loaded die has probability distribution \((\frac{1}{20}, \frac{1}{20}, \frac{1}{20}, \frac{1}{20}, \frac{1}{20}, \frac{3}{4})\). If the die currently being used is a fair die, there is a .1 chance that the casino secretly swaps out the fair die with a loaded die. If the die currently being used is the loaded die there is a .05 chance that the casino secretly swaps out the loaded die for the fair die. Further suppose that the probability that the process starts with a fair
420  18. MAP Estimation and Parametric Inference

die is .8. Note that this setup gives us a hidden Markov model with the following parameters:

\[ \pi = (.8, .2), \quad A = \begin{pmatrix} .9 & .1 \\ .05 & .95 \end{pmatrix}, \quad B = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} \end{pmatrix} \]

After \( m \) rolls we have two sequences \( X \in [6]^m \) consisting of the sequence of die rolls, and \( Y \in [2]^m \) consisting of whether the die used was fair or loaded. As a visitor to the casino, we only observe \( X \) and not \( Y \). From a practical standpoint, we might be interested in estimating whether or not the casino was likely to have used the loaded die, given an observation \( X \) of rolls. Intuitively in this example, if we observe a long stretch with many sixes, we expect that the loaded die has been used. The Viterbi algorithm (Algorithm 18.2.4) can be used to determine the most likely sequence of hidden states given the observed die rolls, and can help a visitor to the casino decide if the loaded die was used.

Hidden Markov models are used in a wide range of contexts. Perhaps their most well-known uses are in text and speech recognition and more broadly in signal processing. In these applications, the hidden states might correspond to letters and phonemes, respectively. The observed states are noisy versions of these things (and typically in this context they would be modeled by continuous random variables). The underlying Markov chain would be a first order Markov chain that approximates the probability that one letter (or phoneme) follows another in the language being modeled. The HMM is used to try to infer what the writer wrote or the speaker said based on the noisy sequence of signals received. A number of texts are devoted to applications and practical aspects of using hidden Markov models [EAM95, ZM09].

HMMs in the discrete setting laid out in this chapter play an important role in bioinformatics. They are used to annotate DNA sequences to determine which parts of a chromosome code for proteins. More complex versions of the hidden Markov model (e.g. the pair hidden Markov model) can be used to give a probabilistic approach to DNA and protein sequence alignment. See [DEKM98] for more details about these biological applications.

A number of computational issues arise when trying to use the HMM on real data. One basic question is the computation of the probability \( p_j \) itself. In the HMM, it is typically assumed that \( m \) is large, so applying the formula (18.2.2) directly would involve calculating a sum over \( r^m \) terms, which is exponential in \( m \). The standard approach to simplify the computational complexity is to apply the distributive law to reduce the number of terms at each step, while adding temporary storage of size \( r \). In symbols, we can...
represent the situation like this:

\[
p_j = \sum_{i_1 \in [r]} \pi_{i_1} b_{i_1 j_1} \left( \sum_{i_2 \in [r]} a_{i_1 i_2} b_{i_2 j_2} \left( \sum_{i_3 \in [r]} a_{i_2 i_3} b_{i_3 j_3} \left( \cdots \right) \right) \right).
\]

Starting at the inside and working out, we see that at each step we are essentially performing a matrix-vector multiplication of an \( r \times r \) matrix with a vector of length \( r \). These matrix-vector multiplications need to be computed \( m \) times. This leads to an algorithm that has a running time that is \( O(r^2 m) \). Since \( r \) is usually not too big, and fixed, but \( m \) could grow large, this is a desirable form for the computational complexity.

**Algorithm 18.2.2** (Calculation of marginal probability in HMM).

- **Input:** \( A \in \mathbb{R}^{r \times r}, B \in \mathbb{R}^{r \times s}, \pi \in \Delta_{r-1}, j \in [s]^m \)
- **Output:** \( p_j \)

- **Initialize:** \( c = (c_1, \ldots, c_r) \) with \( c_\alpha = 1 \) for \( \alpha \in [r] \).
- **For** \( k = m, \ldots, 2 \) do
  - **Update** \( c_\alpha = \sum_{j=1}^r a_{\alpha j} b_{j_k} c_\beta \).
- **Output:** \( p_j = \sum_{\beta=1}^r \pi_{\beta j_k} c_\beta \).

Since at each step of the algorithm we perform \( 2r^2 \) multiplications and \( r(r-1) \) additions, this results in an \( O(r^2 m) \) algorithm, considerably faster than the naive \( O(r^m) \) algorithm. If \( r \) is very large, it can be necessary to consider the special structure built into \( A \) and \( B \), and use tools like the fast Fourier transform to try to speed up this computation further.

A variation on this computation can be performed to determine the conditional probability of an individual hidden state given the parameter values and the values of the observed variables, i.e.

\[
P(Y_k = i_k \mid X = j, \pi, A, B).
\]

This is explained in Exercise 18.1. Because the hidden Markov model is the marginalization of a discrete exponential family, the EM algorithm can be used to compute estimates of the model parameters \( \pi, A, B \), given many observations of sequences \( X \). Running the EM algorithm will require the frequent computation of the probabilities \( p_j \).

Having established basic properties and computations associated with the hidden Markov model, we now move on to performing MAP estimation for this model. The algorithm for performing MAP estimation in the HMM is the Viterbi algorithm. Given fixed values of the parameters \( \pi, A, B \), and a particular observation \( j \), the Viterbi algorithm finds the maximum a posteriori estimate of the hidden sequence \( i \). That is, we are interested in
computing the hidden sequence \( i \in [r]^m \) that maximizes the probability \( p_{i,j} \), i.e. solving the following problem

\[
(18.2.4) \quad \arg \max_{i \in [r]^m} p_{i,j}.
\]

The calculation of the most likely value of the hidden states involves potentially evaluating \( p_{i,j} \) for all the exponentially many different strings \( i \in [r]^m \).

The Viterbi algorithm implements the computation that finds the maximizing \( i \) with only \( O(r^2 n) \) arithmetic operations. A mathematically appealing way to explain how to run the Viterbi algorithm involves computations in the tropical semiring which we explain here.

**Definition 18.2.3.** The tropical semiring consists of the extended real numbers \( \mathbb{R} \cup \{-\infty\} \) together with the operations of tropical addition \( x \oplus y := \max(x, y) \) and tropical multiplication \( x \odot y := x + y \).

The triple \((\mathbb{R} \cup \{-\infty\}, \oplus, \odot)\) satisfies all the properties of a semiring, in particular the associative laws for both tropical addition and multiplication, and the distributive law of tropical multiplication over tropical addition. The “semi” part of the semiring is for the fact that elements do no have tropical additive inverses in the tropical semiring. (However, there are multiplicative inverses for all but the additively neutral element, so \((\mathbb{R} \cup \{-\infty\}, \oplus, \odot)\) is in fact, a semifield.)

We now explain how computations in the tropical semiring are useful for rapidly computing the optimal value \( \max_{i \in [r]^m} p_{i,j} \), which in turn leads to the Viterbi algorithm. The first observations is that

\[
\arg \max_{i \in [r]^m} \log p_{i,j}
\]

is a linear optimization problem. Indeed, substituting in the expression for \( p_{i,j} \), we have

\[
\max_{i \in [r]^m} p_{i,j} = \max_{i \in [r]^m} \pi_i \prod_{k=2}^{m} a_{i_{k-1}i_k} \prod_{k=1}^{m} b_{i_kj_k}.
\]

Let \( \tilde{a}_{ij} = \log a_{ij}, \tilde{b}_{ij} = \log b_{ij} \) and \( \tilde{\pi}_i = \log \pi_i \). Taking logarithms on both sizes yields

\[
\max_{i \in [r]^m} \log p_{i,j} = \max_{i \in [r]^m} \tilde{\pi}_i + \tilde{b}_{i_1j_1} + \sum_{k=2}^{m} \left( \tilde{a}_{i_{k-1}i_k} + \tilde{b}_{i_kj_k} \right).
\]

Now we convert this to an expression in terms of the tropical arithmetic operations:

\[
\bigoplus_{i_1 \in [r]} \cdots \bigoplus_{i_m \in [r]} \tilde{\pi}_{i_1} \odot \tilde{b}_{i_1j_1} \odot \bigodot_{k=2}^{m} \left( \tilde{a}_{i_{k-1}i_k} \odot \tilde{b}_{i_kj_k} \right).
\]
Note that this is the same form as the formula for computing $p_j$ with $a_{ij}, b_{ij}, \pi_i$ replaced by $\tilde{a}_{ij}, \tilde{b}_{ij}, \tilde{\pi}_i$ and with the operations of $+$ and $\times$ replaced by $\oplus$ and $\odot$ respectively. Since both $(\mathbb{R}, +)$ and $(\mathbb{R} \cup \{-\infty\}, \oplus, \odot)$ are semifields, we can apply the exact same form of the distributive law to simplify the calculation of $\max_{i \in [r]^m} \log p_{i,j}$. Indeed, the expression simplifies to

$$\bigoplus_{i_1 \in [r]} \tilde{\pi}_{i_1} \odot \tilde{b}_{i_1j_1} \odot \left( \bigoplus_{i_2 \in [r]} \tilde{a}_{i_1i_2} \odot \tilde{b}_{i_2j_2} \odot \left( \bigoplus_{i_3 \in [r]} \tilde{a}_{i_2i_3} \odot \tilde{b}_{i_3j_3} \odot \left( \cdots \right) \right) \right).$$

This formula/ rearrangement of the calculation yields an $O(r^2 n)$ algorithm to compute the maximum. Of course, what is desired in practice is not only the maximum value but the string $i$ that achieves the maximum. It is easy to modify the spirit of the formula to achieve this. Namely, one stores an $r \times m$ matrix of pointers which point from one entry to the next optimum value in the computation.

**Algorithm 18.2.4 (Viterbi Algorithm).**

Input: $A \in \mathbb{R}^{r \times r}, B \in \mathbb{R}^{r \times s}, \pi \in \Delta_{r-1}, j \in [s]^m$

Output: $i \in [r]^m$ such that $p_{i,j}$ is maximized, and $\log p_{i,j}$.

- Initialize: $c = (c_1, \ldots, c_r)$ with $c_\alpha = 0$ for $\alpha \in [r]$, $\tilde{a}_{ij} = \log a_{ij}, \tilde{b}_{ij} = \log b_{ij}, \tilde{\pi}_i = \log \pi_i$.
- For $k = m, \ldots, 2$ do
  - $d_{\alpha k} = \arg \max_{\beta \in [r]} (\tilde{a}_{\alpha \beta} \odot \tilde{b}_{\beta j_k} \odot c_\beta)$ for $\alpha \in [r]$.
  - Update $c_\alpha = \bigoplus_{\beta \in [r]} (\tilde{a}_{\alpha \beta} \odot \tilde{b}_{\beta j_k} \odot c_\beta)$ for $\alpha \in [r]$.
- Compute $d_1 = \arg \max_{\beta \in [r]} (\tilde{\pi}_\beta \odot \tilde{b}_{\beta j_1} \odot c_\beta)$

Define the sequence $i$ by $i_1 = d_1$ and $i_k = d_{i_{k-1}k}$ for $k = 2, \ldots, m$.

- $\log p_{i,j} = \bigoplus_{\beta \in [r]} (\tilde{\pi}_\beta \odot \tilde{b}_{\beta j_1} \odot c_\beta)$

Note that in Algorithm 18.2.4, it is the computation of the array $d$ which locally determines the maximizer at each step of the procedure. From the array $d$, the MAP estimate $i$ is computed which is the hidden state string that maximizes the probability of observing $j$.

**Example 18.2.5 (Occasionally Dishonest Casino).** Let us consider the occasionally dishonest casino from Example 18.2.1. Suppose that we observe the sequence of rolls:

$$1, 3, 6, 6, 1, 2, 6$$

and we want to estimate what was the most likely configuration of hidden states. At each step of the algorithm we compute the vector $c$, which shows
that current partial weight. These values are illustrated in the matrix:
\[
\begin{pmatrix}
-11.25 & -9.36 & -7.58 & -5.69 & -3.79 & -1.89 \\
-10.15 & -7.11 & -6.77 & -6.43 & -3.38 & -0.33 \\
\end{pmatrix}
\]
with a final optimal value of -13.27, obtained with the optimal argument of \( d_1 = 1 \). Here we are letting 1 denote the fair die, and 2 the loaded die. The resulting array of pointers is
\[
\begin{pmatrix}
1 & 2 & 1 & 1 & 1 \\
2 & 2 & 2 & 2 & 2 \\
\end{pmatrix}
\]
From this we see that the MAP estimate of hidden states is 1, 1, 2, 2, 2, 2.

18.3. Parametric Inference and Normal Fans

The calculation of the MAP estimate, for example, using the Viterbi algorithm for HMMs, assumes an absolute certainty about the values of the model parameters. This assumption is frequently made, but lacks justification in many setting. This naturally leads to the question: how does the MAP estimate vary as those “fixed” model parameters vary? This study of the sensitivity of the inference of the MAP estimate to the parameter choices is called parametric inference. For this specialized version of the MAP estimation problem to the more general MAP estimation problem, it becomes useful to understand the polyhedral geometry inherent in measuring how the variation of a cost vector in linear optimization affects the result of the optimization. From a geometric perspective, we need to compute the normal fan of the Newton polytope of the associated polynomial \( p_j = f(\pi, A, B) \).
The definitions of Newton polytope and normal fan appear in Section 12.3. We will illustrate these ideas in the case of the HMM, though the ideas also apply more widely. Perhaps the most well-known context where this approach is applied is in sequence alignment [FBSS02, GBN94, Vin09].

Now we explain how polyhedral geometry comes into play. To each pair \( i \in [r]^m \) and \( j \in [s]^m \), we have the monomial
\[
p_{i,j} = \pi_i b_{i_1,j_1} a_{i_1} \cdots a_{i_m-j_m} b_{i_m,j_m}
\]
Let \( \theta = (\pi, A, B) \) be the vector of model parameters. We can write
\[
p_{i,j} = \theta^{u_{i,j}}
\]
for some integer vector \( u_{i,j} \). For fixed values of \( \theta \) we want to compute
\[
\arg \max_{i \in [r]^m} \theta^{u_{i,j}} = \arg \max_{i \in [r]^m} (\log \theta)^T \cdot u_{i,j}
\]
Let \( c = \log \theta \), where the logarithm is applied coordinate-wse. Assuming that we can recover \( i \) from \( u_{i,j} \), we can express the problem instead as solving
the optimization problem
\[ \max_{c \in [r]^m} c^T u_{ij}. \]

Since \( c \) is a linear functional, we could instead solve the optimization problem
\[ \max_{u \in Q_j} c^T u \quad \text{where} \quad Q_j = \text{Newt}(p_j) = \text{conv}(u_{i,j} : i \in [r]^m). \]

The polytope \( Q_j = \text{Newt}(p_j) \) is the Newton polytope of the polynomial \( p_j \), as introduced in Proposition 18.1.2. This polytope reveals information about the solution to the MAP estimation problem. Indeed, if we want to know which parameter values \( \theta \) yield a given \( i \) as the optimum, this is obtained by determining the cone of all \( c \) that satisfy that
\[ \max_{u \in Q_j} c^T u = c^T u_{ij} \]

and intersecting that cone with the set of vectors of the form \( \log \theta \). Without the log-theta constraint, the set of all such \( c \) is the normal cone to \( Q_j \) at the point \( u_{ij} \). So the complete solution to the parametric optimization problem for all choices of optimum vector is the normal fan of the polytope \( Q_j \). This is summarized in the following proposition.

**Proposition 18.3.1.** The cones of the normal fan of the Newton polytope \( Q_j = \text{Newt}(p_j) \) give all parameter vectors that realize a specific optimum of the MAP estimation problem.

The normal fan of the Newton polytope is also useful in the more general Bayesian setting of MAP estimation. Here we would like to know the distribution over possible hidden sequences \( i \) given a distribution on the parameters and the observed sequence \( j \). The only sequences with non-zero probability will be the ones corresponding to vertices of the polytope \( \text{Newt}(p_j(A, B, \pi)) \). The posterior probability of such a sequence is obtained by integrating out the prior distribution on parameters over the normal cone at the corresponding vertex of \( \text{Newt}(p_j(A, B, \pi)) \). While these integrals can be rarely calculated in closed form, the geometry of the solution can help to calculate the posterior probabilities.

At this point, one might wonder if it is possible to actually compute the polytopes \( \text{Newt}(p_j(A, B, \pi)) \). At first glance, it seems hopeless to determine all the vertices or facets of these polytopes, as they are the convex hulls of the \( r^m \) points \( u_{i,j} \). Clearly there is an exponential number of points to consider as a function of the sequence length \( m \). However, these polytopes have only polynomially many vertices as a function of \( m \). There are also fast algorithms for computing these polytopes without enumerating all the \( r^m \) points first.

It is not difficult to see that there are at most a polynomial number of vertices of \( \text{Newt}(p_j(A, B, \pi)) \), as a function of \( m \). Indeed, there are a
total of \( r + r^2 + rs \) parameters in the \( \pi, A, B \), so each \( u_{i,j} \) is a vector of length \( r + r^2 + rs \). The sum of the coordinates of each \( u_{i,j} \) is 2, since this is the number of terms in the monomial representing \( p_{i,j} \). A basic fact from elementary combinatorics (see [Sta12]) is that the total number of distinct nonnegative integer vectors \( u \in \mathbb{N}^k \) such that \( \sum_{i=1}^{k} u_i = m \) is \( \binom{m+k-1}{k-1} \). When \( k \) is fixed, this is a polynomial of degree \( k - 1 \) as a function of \( m \). Hence, when computing the convex hull to get \( \text{Newt}(p_j(A, B, \pi)) \), there are at most \( \binom{2m+r^2+rs-1}{r^2+rs-1} \) distinct points, which gives a naive bound on the number of vertices of \( \text{Newt}(p_j(A, B, \pi)) \). As a function of \( m \), this is a polynomial of degree \( r + r^2 + rs - 1 \).

In fact, it is possible to derive a lower order polynomial bound on the number of vertices of \( Q_j \) than this, using a classic result of Andrews that bounds the number of vertices of a lattice polytope in terms of its volume.

**Theorem 18.3.2.** [And63] For every fixed integer \( d \) there is a constant \( C_d \) such that the number of vertices of any convex lattice polytope \( P \) in \( \mathbb{R}^d \) is bounded above by \( C_d \text{vol}(P)^{(d-1)/(d+1)} \).

Note that the polytope \( \text{Newt}(p_j(A, B, \pi)) \) is not \( r + r^2 + rs \) dimensional, and in fact, there will be linear relations that hold between entries in the \( u_{i,j} \) vectors. It is not hard to see that this polytope will be \( (r-1)(s+1) + r^2 \) dimensional for general \( j \) and large enough \( m \), and that for all \( j \) and \( m \), the dimension is bounded above by this number (see Exercise 18.3). We can project all the \( u_{i,j} \)’s onto a subset of the coordinates so that it will then be a full dimensional polytope in a suitable \( \mathbb{R}^{(r-1)(s+1)+r^2} \) and will still be a lattice polytope. Since all coordinates of the projections of \( u_{i,j} \) will be of size \( \leq m \), we see that the \( \text{vol}(\text{Newt}(p_j(A, B, \pi))) \) \( \leq n^{(r-1)(s+1)+r^2} \).

Combining this bound on the volume with Andrew’s theorem, we get the following bound:

**Corollary 18.3.3.** Let \( D = (r-1)(s+1) + r^2 \). The polytope \( Q_j \) has at most \( C_D m^{D(D-1)/(D+1)} \) vertices.

These types of bounds can be extended more generally to similar MAP estimation problems for directed graphical models with a fixed number of parameters. Results of this kind are developed in [EW07, Jos05]. Of course, we are not only interested in knowing bounds on the number of vertices, but we would like to have algorithms for actually computing the vertices and the corresponding normal cones. This is explained in the next section.

A second application of these bounds is that they can be used to bound the number of inference functions of graphical models. We explain this only for the HMM. For a given fixed \( \theta = (\pi, A, B) \), the associated inference
function is the induced map \( \phi : [s]^m \to [r]^m \), that takes an observed sequence \( j \) and maps it to the MAP estimate for the hidden sequence \( i \), given the parameter value \( \theta \). The total number of functions \( \phi : [s]^m \to [r]^m \) is \((r^m)^s^m = r^{ms^m}\) which is doubly exponential in \( m \). However, the Andrews bound can be used to show that the total number of inference functions must be much smaller than this. Indeed, for each fixed \( j \), the parameter space of \( \theta \) is divided into the normal cones of \( Q_j \), which parameterize all the different possible places that \( j \) can be sent in any inference function. The common refinement of these normal fans over all \( j \in [s]^m \) divides the parameter space of \( \theta \) into regions that correspond to the inference functions. The common refinement of normal fans corresponds to taking the Minkowski sums of the corresponding polytopes. This polytope is the Newton polytope of \( \prod_{j \in [s]^m} p_j \), which we denote \( Q \). The polytope \( Q \) has dimension \( r + r^2 + rs - 1 \), and it has degree \( 2ms^m \). The fact that the dimension is fixed allows us to again employ Theorem 18.3.2 to deduce.

**Corollary 18.3.4.** Let \( E = r + r^2 + rs - 1 \). The polytope \( Q \) has at most \( C^E (2ms^m)^E / (E^+1) \) vertices.

In particular, our analysis shows that the number of inference functions grows at most singly exponentially in \( m \). For this reason Corollary 18.3.4 is called the **few inference function theorem** [PS04a, PS04b].

### 18.4. Polytope Algebra and Polytope Propagation

In this section, we explain how the nice factorization and summation structure of the polynomials that appear for the HMM and other graphical models can be useful to more rapidly compute the Newton polytope and hence the corresponding normal fan. This allows to actually perform parametric inference rapidly in these models.

**Definition 18.4.1.** Let \( \mathcal{P}_k \) denote the set of all polytopes in \( \mathbb{R}^k \). Define two operations on \( \mathcal{P}_k \): for \( P, Q \in \mathcal{P}_k \) let

1. \( P \oplus Q := \text{conv}(P \cup Q) \)
2. \( P \odot Q := P + Q = \{ p + q : p \in P, q \in Q \} \).

The triple \( (\mathcal{P}_k, \oplus, \odot) \) is called the **polytope algebra**.

Note that \( P \odot Q \) is the Minkowski sum of \( P \) and \( Q \). The name polytope algebra comes from the fact that \( (\mathcal{P}_k, \oplus, \odot) \) is a semiring with the properties we discussed when first discussing tropical arithmetic. In particular, the operations \( \oplus \) and \( \odot \) are clearly associative, and they satisfy the distributive law

\[
P \odot (Q_1 \oplus Q_2) = (P \odot Q_1) \oplus (P \odot Q_2).
\]
Another useful semiring that is implicit in our calculations of Newton polytopes is the semiring \( R_+ = \mathbb{R}_+[\theta_1, \ldots, \theta_k] \) consisting of all polynomials in the indeterminates \( \theta_1, \ldots, \theta_k \) where the coefficients are restricted to nonnegative real numbers. This is a semiring, and not a ring, because we no longer have additive inverses of polynomials.

**Proposition 18.4.2.** The map \( \text{Newt} : R_+[\theta_1, \ldots, \theta_k] \to \mathcal{P}_k \) with \( f(\theta) \mapsto \text{Newt}(f(\theta)) \) is a semiring homomorphism.

**Proof.** We must show that
\[
\text{Newt}(f(\theta) + g(\theta)) = \text{Newt}(f(\theta)) \oplus \text{Newt}(g(\theta))
\]
and that
\[
\text{Newt}(f(\theta)g(\theta)) = \text{Newt}(f(\theta)) \odot \text{Newt}(g(\theta))
\]
For the first of these, note that our restriction to nonnegative exponents guarantees that there can be no cancellation when we add two polynomials together. Hence, the set of exponent vectors that appears is the same on both sides of the equation. For the second, note that for any weight vector \( c \) such that \( \text{in}_c(f(\theta)) \) and \( \text{in}_c(g(\theta)) \) is a single monomial, the corresponding initial term of \( f(\theta)g(\theta) \) is \( \text{in}_c(f(\theta))\text{in}_c(g(\theta)) \), and so this monomial yields a vertex of \( \text{Newt}(f(\theta)) + \text{Newt}(g(\theta)) \). This shows that
\[
\text{Newt}(f(\theta)) + \text{Newt}(g(\theta)) \subseteq \text{Newt}(f(\theta)g(\theta))
\]
since it shows that every vertex of \( \text{Newt}(f(\theta)) + \text{Newt}(g(\theta)) \) is contained in \( \text{Newt}(f(\theta)g(\theta)) \). But the reverse containment is trivial looking at what exponent vectors might appear in the product of two polynomials. \( \square \)

One consequence of Proposition 18.4.2 is that we immediately get an algorithm for computing the Newton polytope of \( f \), in situations where there is a fast procedure to evaluate or construct \( f \). As an example, consider the parametric inference problem for the hidden Markov model. For this model, we need to construct the Newton polytope of the polynomial \( p_j(\theta) \). The polynomial itself can be constructed using a variation on Algorithm 18.2.2, replacing values of the parameters \( A, B, \pi \) with indeterminates. This has the effect of replacing the sum and products of numbers in Algorithm 18.2.2 with sums and products of polynomials. Note that all coefficients in these polynomials are nonnegative so we are in a setting where we can apply Proposition 18.4.2. This means that we can use Algorithm 18.2.2 to compute \( \text{Newt}(p_j) \) where we make the following replacements in the algorithm: every polynomial that appears in the algorithm is replaced with a polytope, every \( + \) in the algorithm is replaced with \( \oplus \) (the convex hull of the union of
polytopes), and every \( \times \) in the algorithm is replaced with \( \odot \) (the Minkowski sum of two polytopes). Note that to initialize the algorithm, each indeterminate yields a polytope which is just a single point (the convex hull of its exponent vector). The algorithm that results from this procedure is known as \textit{polytope propagation} and its properties and applications were explored in [PS04a, PS04b].

In the case of the hidden Markov model, we have an expression using the distributive law for \( p_j \) which is shown in Equation 18.2.3. We apply the semiring homomorphism

\[
\text{Newt} : \mathbb{R}_+ [\pi, A, B] \to \mathcal{P}_{r+r^2+rs}
\]

which results in the following expansion in the polytope algebra:

\[
\text{Newt}(p_j) = \bigoplus_{i_1 \in [r]} \text{Newt}(\pi_{i_1} b_{i_1 j_1}) \odot \left( \bigoplus_{i_2 \in [r]} \text{Newt}(a_{i_1 i_2} b_{i_2 j_2}) \odot (\cdots) \right).
\]

Applications of the parametric inference paradigm, polytope propagation, and fast algorithms for computing the resulting Newton polytopes appear in [DHW+06, HH11].

18.5. Exercises

Exercise 18.1. Explain how to construct a variation on Algorithm 18.2.2 to compute the conditional probabilities \( P(Y_k = i_k \mid X = j, \pi, A, B) \). (Hint: First trying modifying Algorithm 18.2.2 to get a “forward” version. Then show how to combine the “forward” and “backward” algorithms to compute \( P(Y_k = i_k \mid X = j, \pi, A, B) \).)

Exercise 18.2. Implement the Viterbi algorithm. Use it to compute the MAP estimate of the hidden states in the occasionally dishonest casino where the parameters used are as in Example 18.2.1 and the observed sequence of rolls is:

1, 3, 6, 6, 3, 6, 6, 6, 6, 3, 6, 4, 4, 5, 6, 1, 4, 3, 2, 5, 6, 6, 2, 3, 2, 6, 6, 3, 6, 6, 1, 2, 3, 6

Exercise 18.3. Show that the dimension of the polytope Newt\((p_j(A, B, \pi))\) from the hidden Markov model is at most \((r - 1)(s + 1) + r^2\).

Exercise 18.4. Prove the distributive law (Equation 18.4.1) for the polytope algebra \( \mathcal{P}_k (\odot, \oplus) \).
Chapter 19

Finite Metric Spaces

This chapter is concerned with metric spaces consisting of finitely many points. Such metric spaces occur in various contexts in algebraic statistics. Particularly, finite metric spaces arise when looking at phylogenetic trees. Finite metric spaces also have some relationships to hierarchical models. A key role in this side of the theory is played by the cut polytope. We elucidate these connections in this chapter.

19.1. Metric Spaces and The Cut Polytope

In this section, we explore key facts about finite metric spaces and particularly results about embedding metrics within other metric spaces with specific properties. Here we are summarizing some of the key results that appear in the book of Deza and Laurent [DL97].

Definition 19.1.1. Let \( \mathcal{X} \) be a finite set. A dissimilarity map on \( \mathcal{X} \) is a function \( d : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \) such that \( d(x, y) = d(y, x) \) for all \( x, y \in \mathcal{X} \) and \( d(x, x) = 0 \) for all \( x \in \mathcal{X} \).

Note that we are explicitly allowing negative numbers as possible dissimilarities between elements of \( \mathcal{X} \).

Definition 19.1.2. Let \( \mathcal{X} \) be a finite set. A dissimilarity map on \( \mathcal{X} \) is a semimetric if it satisfies the triangle inequality

\[
d(x, y) \leq d(x, z) + d(y, z)
\]

for all \( x, y, z \in \mathcal{X} \). If \( d(x, y) > 0 \) for all \( x \neq y \in \mathcal{X} \) then \( d \) is called a metric. We denote the pair \( (\mathcal{X}, d) \) as a metric space when speaking of a semimetric or a metric.
Finite Metric Spaces

In the triangle inequality we allow that \( y = z \), in which case we deduce that \( d(x, y) \geq 0 \) for all \( x, y \in \mathcal{X} \) for any semimetric.

**Definition 19.1.3.** Let \((\mathcal{X}, d)\) and \((\mathcal{Y}, d')\) be metric spaces. We say that \((\mathcal{X}, d)\) embeds isometrically in \((\mathcal{Y}, d')\) if there is a map of sets \( \phi : \mathcal{X} \to \mathcal{Y} \) such that \( d'(\phi(x), \phi(y)) = d(x, y) \) for all \( x, y \in \mathcal{X} \).

Being able to embed \((\mathcal{X}, d)\) isometrically from one space into another is most interesting in the case where \( \mathcal{Y} \) is an infinite space and \( d' \) is some standard metric, for instance an \( L^p \) metric space, where \( \mathcal{Y} = \mathbb{R}^k \) and

\[
d'(x, y) = d_p(x, y) := \sqrt[p]{\sum_{i=1}^{k} |x_i - y_i|^p}.
\]

**Definition 19.1.4.** Let \((\mathcal{X}, d)\) be a finite metric space. We say that \((\mathcal{X}, d)\) is \( L^p \) embeddable if there exists a \( k \) such that \((\mathcal{X}, d)\) is isometrically embeddable in \((\mathbb{R}^k, d_p)\).

Let \( \mathcal{X} \) be a finite set with \( m \) elements. The set of all semimetrics on \( \mathcal{X} \) is a polyhedral cone in \( \mathbb{R}^{m(m-1)/2} \) defined by the \( 3\binom{m}{3} \) triangle inequalities together with the \( \binom{m}{2} \) nonnegativity inequalities. It is denoted \( \text{Met}_m \) and called the metric cone. The set of \( L^1 \) embeddable semimetrics forms a polyhedral subcone of the metric cone called the cut cone, which we describe now.

**Definition 19.1.5.** Let \( \mathcal{X} \) be a finite set with \( m \) elements and let \( A \parallel B \) be a split of \( \mathcal{X} \). The cut semimetric of \( A \parallel B \) is the semimetric \( \delta_{A \parallel B} \) with

\[
\delta_{A \parallel B}(x, y) = \begin{cases} 
1 & \text{if } \#(\{x, y\} \cap A) = 1 \\
0 & \text{otherwise}.
\end{cases}
\]

The cone generated by all \( 2^{m-1} \) cut semimetrics is the cut cone, denoted \( \text{Cut}_m \). The polytope obtained by taking the convex hull of all cut semimetrics is the cut polytope, denoted \( \text{Cut}_m^\ast \).

Note that in the cut polytope we are allowing the split \( \emptyset \parallel \mathcal{X} \) which gives the rather boring cut semimetric \( \delta_{\emptyset \parallel \mathcal{X}} = 0 \). This is necessary for reasons of symmetry in the cut polytope.

**Example 19.1.6** (Four Point Cut Polytope). Let \( m = 4 \). Then the cut polytope \( \text{Cut}_4^\ast \) is the polytope whose vertices are the eight cut semimetrics...
which are the columns of the following matrix
\[
\begin{pmatrix}
0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\
0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1
\end{pmatrix}.
\]

The cut cone characterizes $L^1$ embeddability.

**Proposition 19.1.7.** A finite metric space $([m], d)$ is $L^1$ embeddable if and only if $d \in \text{Cut}_m$.

**Proof.** First suppose that $d \in \text{Cut}_m$. This means
\[
d = \sum_{j=1}^k \lambda_{A_j|B_j} \delta_{A_j|B_j}
\]
for some $k$ and some positive constants $\lambda_{A_j|B_j}$. Let $u_1, \ldots, u_m \in \mathbb{R}^k$ be the vectors such that
\[
(u_i)_j = \begin{cases} 
\lambda_{A_j|B_j} & \text{if } i \in A_j \\ 
-\lambda_{A_j|B_j} & \text{if } i \in B_j 
\end{cases}
\]
Then $d_1(u_i, u_{i'}) = \sum_{j=1}^k |(u_i)_j - (u_{i'})_j| = d(i, i')$ since $|(u_i)_j - (u_{i'})_j| = \lambda_{A_j|B_j}$ if $\#(\{i, i'\} \cap A_j) = 1$ and $|(u_i)_j - (u_{i'})_j| = 0$ otherwise.

Conversely, suppose that there exist $k$ and $u_1, \ldots, u_m \in \mathbb{R}^k$ such that $d(i, i') = d_1(u_i, u_{i'})$. We must show that $d$ is in $\text{Cut}_m$. Being $L^1$-embeddable is additive on dimensions of the embedding, i.e. if $u_1, \ldots, u_m \in \mathbb{R}^k$ gives a realization of a metric $d$, and $u'_1, \ldots, u'_m \in \mathbb{R}^k$ gives a realization of a metric $d'$ then $(u_1, u'_1), \ldots, (u_m, u'_m)$, gives a realization of the metric $d + d'$ as an $L^1$ embedding. Hence, we may suppose that $k = 1$. Furthermore, suppose that $u_1 \leq u_2 \leq \cdots \leq u_m$. Then it is straightforward to check that
\[
d = \sum_{j=1}^{k-1} (u_{j+1} - u_j) \delta_{|[m]|j}.
\]

Although the cut cone and cut polytope have a simple description in terms of listing their extreme rays, any general description of their facet defining inequalities remains unknown. Part V of [DL97] contains descriptions of many types of facets of these polytopes, but the decision problem of determining whether or not a given dissimilarity map $d$ belongs to the cut cone is known to be NP-complete.

The cut cone and the cut polytope are connected to probability theory via their relation to joint probabilities and the correlation polytope. We
explain this material next by first relating the results to measure spaces. For two sets \( A, B \) let \( A \Delta B \) denote their symmetric difference.

**Proposition 19.1.8.** A dissimilarity map \( d : [m] \times [m] \to \mathbb{R} \) is in the cut cone \( \text{Cut}_m \) (respectively, cut polytope \( \text{Cut}_m^\square \)) if and only if there is a measure space (respectively, probability space) \( (\Omega, \mathcal{A}, \mu) \) and subsets \( S_1, \ldots, S_m \in \mathcal{A} \) such that \( d(i, j) = \mu(S_i \Delta S_j) \).

**Proof.** First suppose that \( d \in \text{Cut}_m \), so that

\[
d = \sum_{j=1}^{k} \lambda_{A_j|B_j} \delta_{A_j|B_j}
\]

with \( \lambda_{A_j|B_j} \geq 0 \). We will construct a measure space with the desired properties. Let \( \Omega = 2^m \), the power set of \([m]\), and let \( \mathcal{A} \) consist of all subsets of \( 2^m \). For a set \( S \subseteq \mathcal{A} \) define

\[
\mu(S) = \sum_{A_i \in S} \lambda_{A_i|B_i}
\]

(Here we will treat two splits \( A|B \) and \( B|A \) as different from each other, although they produce the same cut semimetrics \( \delta_{A|B} \).) Let \( S_i = \{ S \in \Omega : i \in S \} \). Then

\[
\mu(S_i \Delta S_j) = \mu(\{ A \in \Omega : \#(A \cap \{ i, j \}) = 1 \}) = \sum_{A \in \Omega, \#(A \cap \{ i, j \}) = 1} \lambda_{A|B} = d(i, j).
\]

Furthermore, if \( d \in \text{Cut}_m^\square \) then \( \sum \lambda_{A|B} = 1 \) which implies that \( \mu(\Omega) = 1 \), i.e. \( (\Omega, \mathcal{A}, \mu) \) is a probability space.

Similar reasoning can be used to invert this construction, i.e. given a measure space and subsets \( S_1, \ldots, S_m \), one can show that \( d(i, j) = \mu(S_i \Delta S_j) \) yields an element of the cut cone. \( \square \)

An analogous result concerns elements of the semigroup generated by the cut semimetrics, the proof of which is left as an exercise.

**Proposition 19.1.9.** A dissimilarity \( d : [m] \times [m] \to \mathbb{R} \) has the form \( d = \sum_{j=1}^{k} \lambda_{A_j|B_j} \delta_{A_j|B_j} \) where \( \lambda_{A_j|B_j} \) are nonnegative integers for \( j = 1, \ldots, k \) if and only if there are finite sets \( S_1, \ldots, S_m \) such that \( d(i, j) = \#(S_i \Delta S_j) \).

Proposition 19.1.8 suggests that the cut cone, or a cone over the cut polytope, should be related to the cone of sufficient statistics of a discrete exponential family, since, for many natural occurring models, those cones have the interpretation as cones that record moments of some particular functions of the random variables. This is made precise by the correlation
19.1. Metric Spaces and The Cut Polytope

cone, correlation polytope, and a mapping between this object and corresponding cut objects.

**Definition 19.1.10.** Let $\mathcal{X}$ be a finite set with $m$ elements. For each $S \subseteq \mathcal{X}$ let $\pi_S$ be the correlation vector in $\mathbb{R}^{(m+1)}$ whose coordinates are indexed by pairs $i, j \in \mathcal{X}$ of not necessarily distinct elements with

$$
\pi_S(i, j) = \begin{cases} 
1 & \text{if } \{i, j\} \subseteq S \\
0 & \text{otherwise}.
\end{cases}
$$

The correlation cone $\text{Corr}_m$ is the polyhedral cone generated by all $2^m$ correlation vectors $\pi_S$. The correlation polytope $\text{Corr}^{\square}_m$ is the convex hull of the $2^m$ correlation vectors.

**Example 19.1.11** (Correlation Polytope on Three Events). Let $m = 3$. Then the correlation polytope is the polytope whose vertices are the columns of the following matrix:

$$
\begin{pmatrix}
0 & 1 & 0 & 1 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}
$$

The covariance mapping turns the cut cone into the correlation cone, and the cut polytope into the correlation polytope. To describe this we need to denote the coordinates on the ambient spaces for these objects in the appropriate way. The natural coordinates on the space of dissimilarity maps is the set of $d(i, j)$ where $i, j$ is an unordered pair of distinct elements of $[m]$. The natural coordinates on the space of correlations are denoted $p_{ij}$ where $i, j$ is an unordered pair of not necessarily distinct elements of $[m]$.

**Definition 19.1.12.** Consider the mapping from $d(i, j)$ to $p_{ij}$ coordinates defined by the rule:

$$
p_{ii} = d(i, m + 1) \\
p_{ij} = \frac{1}{2}(d(i, m + 1) + d(j, m + 1) - d(i, j)) \\
$$

if $1 \leq i \leq m$ and $1 \leq i < j \leq m$. This map is called the covariance mapping and denoted $\xi$.

Note that the corresponding inverse mapping satisfies

$$
d(i, m + 1) = p_{ii} \\
d(i, j) = p_{ii} + p_{jj} - 2p_{ij} \\
$$

if $1 \leq i \leq m$ and $1 \leq i < j \leq m$.

**Proposition 19.1.13.** The covariance mapping sends the cut cone to the correlation cone and the cut polytope to the correlation polytope. That is

$$
\xi(\text{Cut}_{m+1}) = \text{Corr}_m \quad \text{and} \quad \xi(\text{Cut}^{\square}_{m+1}) = \text{Corr}^{\square}_m.
$$
Proof. Write a split of \([m + 1]\) in a unique way \(A|B\) so that \(m + 1 \in B\). It is straightforward to check that \(\xi(\delta_{A|B}) = \pi_A\).

The reader should try to verify Proposition 19.1.13 on Examples 19.1.6 and 19.1.11. Note that the covariance mapping allows us to easily transfer results about the polyhedral description from the cut side to the correlation side, and vice versa.

Applying the covariance mapping to Proposition 19.1.8, one can deduce the following Proposition, which explains the name correlation cone and correlation polytope.

**Proposition 19.1.14.** A correlation vector \(p\) belongs to the correlation cone (respectively correlation polytope) if and only if there exists events \(S_1, \ldots, S_m\) in a measure space \((\Omega, \mathcal{A}, \mu)\) (respectively, probability space) such that \(p_{ij} = \mu(S_i \cap S_j)\).

If \(X_1, \ldots, X_m\) are indicator random variables of the events \(S_1, \ldots, S_m\) respectively then clearly

\[
p_{ij} = \mu(S_i \cap S_j) = E[X_i X_j].
\]

This means that the cone over the correlation polytope (by lifting each correlation vector \(\pi_S\) to the vector \((1, \pi_S)\)) is the cone of sufficient statistics of a discrete exponential family associated to the the complete graph \(K_m\), that was studied in Section 8.2. In that section, these moment polytopes were discussed for arbitrary graphs \(G\) on \(m\) vertices, and it makes sense to extend both the cut and correlation side to this more generalized setting.

**Definition 19.1.15.** Let \(G = ([m], E)\) be a graph with vertex set \([m]\). The cut cone associated to the graph \(G\), denoted \(\text{Cut}(G)\) is the cone spanned by the projection of the \(2^m - 1\) cut semimetrics onto the coordinates indexed by edges \(E\). The cut polytope \(\text{Cut}^\square(G)\) is the convex hull of these projected cut semimetrics.

**Example 19.1.16 (Cut Polytope of a Five-Cycle).** Let \(m = 5\) and consider the graph \(G = C_5\), the 5-cycle with edges \(E = \{12, 23, 34, 45, 15\}\). Then the cut polytope \(\text{Cut}^\square(C_5)\) is the convex hull of the columns of the following matrix.

\[
\begin{pmatrix}
0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1
\end{pmatrix}
\]

The cut cone and polytope for general graphs have been studied by many authors, and many results are known about their geometry. For instance,
the switching operation can be used to take a complete set of facet defining inequalities for the cut cone and transform them to a complete set of facet defining inequalities for the cut polytope. Lists of facet-defining inequalities for the cut cone \( \text{Cut}(K_m) = \text{Cut}_m \) are known up to \( m = 8 \) (at least at the time of the writing of [DL97]), and given the difficulty of solving related optimization problems over these polytopes, it seems unlikely that a complete description is possible for all \( m \).

Certain combinatorial structures in the graphs force related facet defining inequalities. The simplest are the cycle inequalities.

**Proposition 19.1.17.** Let \( C \subseteq G \) be an induced cycle in the graph \( G \) consisting of edges \( \{i_1i_2, i_2i_3, \ldots, i_ti_1\} \). Then for each \( j \) the inequality

\[
d(i_j, i_{j+1}) \leq \sum_{k \neq j} d(i_k, i_{k+1})
\]

is a facet defining inequality of the cut cone \( \text{Cut}(G) \). (Where \( i_{t+1} := i_1 \).)

**Proof.** The fact that this inequality is valid for all metrics is a consequence of the triangle inequality. For simplicity, assume we are looking at the cycle \{12, 23, \ldots, t1\}. Suppose by way of induction that

\[
d(1, 2) \leq d(2, 3) + \cdots + d(t - 2, t - 1) + d(t - 1, 1).
\]

By the triangle inequality, we know that \( d(t - 1, 1) \leq d(t - 1, t) + d(t, 1) \) so combining these produces the inequality

\[
d(1, 2) \leq d(2, 3) + \cdots + d(t - 1, t) + d(t, 1)
\]

as desired. We will show that this inequality is facet-defining in the case that the underlying graph \( G \) is just the cycle \( C \) by itself. The corresponding cut cone \( \text{Cut}(C) \subseteq \mathbb{R}^t \), a \( t \) dimensional vector spaces, whose coordinates are \( d(1, 2), \ldots, d(t, 1) \). To see that this is facet-defining, note that the following points all lie in \( \text{Cut}(C) \):

\[
(1, 1, 0, 0, \ldots, 0), (1, 0, 1, 0, \ldots, 0), (1, 0, 0, 1, \ldots, 0), \ldots, (1, 0, 0, 0, \ldots, 1)
\]

and they also satisfy our inequality with equality. These \( t - 1 \) vectors are all the images of cut semimetrics, and they are linearly independent. This implies that the cycle inequality is facet-defining. To show that this inequality is facet defining in an arbitrary graph with an induced cycle, we must find further linearly independent vectors vectors that lie on the indicated facet. See [BM86] for details.

Note that when the cycle has length 3, the facet defining inequalities from Proposition 19.1.17 are just the usual triangle inequalities. The cone cut out by the cycle inequalities for the graph \( G \) plus nonnegativity constraints
Finite Metric Spaces

\(d(i, j) \geq 0\) is called the metric cone associated to the graph \(G\), denoted \(\text{Met}(G)\). Clearly \(\text{Cut}(G) \subseteq \text{Met}(G)\) for all graphs \(G\).

**Theorem 19.1.18.** [Sey81] The cut cone and metric cone of a graph \(G\) coincide if and only if the graph \(G\) is free of \(K_5\) minors.

Similarly, one can define the correlation cone and polytope for an arbitrary graph, and more generally for simplicial complexes.

**Definition 19.1.19.** Let \(\Gamma\) be a simplicial complex on \([m]\). For each \(S \subseteq [m]\) define the correlation vector \(\pi_S\) to be the vector in \(\mathbb{R}^F\) such that

\[
\pi_S(T) = \begin{cases} 
1 & \text{if } T \subseteq S \\
0 & \text{otherwise}
\end{cases}.
\]

The correlation cone of \(\Gamma\) is the cone generated by the \(2^m\) correlation vectors and is denoted \(\text{Corr}(\Gamma)\).

Note that this definition of correlation cone differs from our previous definition because we are including the face \(\emptyset\), which has the effect of lifting the correlation polytope up to one higher dimension and taking the cone over it. The correlation polytope of \(\Gamma\) arises in the study of hierarchical models.

**Proposition 19.1.20.** The correlation cone \(\text{Corr}(\Gamma)\) is isomorphic to the cone of sufficient statistics of the hierarchical model associated to \(\Gamma\) with all random variables having two states \(\text{cone}(A_1, 2)\).

The proof of Proposition 19.1.20 is left as an exercise. In the case of graphs, we can still apply the covariance mapping to map between the correlation polytope and the cut polytope (and back). We need to modify the graphs in the appropriate way. If \(G\) is a graph on \([m]\), let \(\hat{G}\) denote the graph obtained from \(G\) by adding the new vertex \(m+1\) and all edges \((i, m+1)\) for \(i \in [m]\). This graph is called the suspension of \(G\).

**Proposition 19.1.21.** Let \(G\) be a graph on \([m]\). Then the covariance mapping sends the cut polytope \(\text{Cut}(\hat{G})\) to the correlation polytope \(\text{Corr}(\Gamma)\).

**Proof.** For the definition of the covariance mapping to make sense for a graph \(\hat{G}\), the vertex \(m+1\) must be connected to all other vertices, hence \(G\) must be the suspension of a graph \(G\). Once this observation is made, it is straightforward to see that the projection from \(\text{Cut}(m+1)\) to \(\text{Cut}(\hat{G})\) and \(\text{Corr}^m(m)\) to \(\text{Corr}(\Gamma)\) commute with the covariance mapping \(\xi\).

Proposition 19.1.21 implies that the myriad results for the cut cone and cut polytope can be employed to deduce results on the facet structure of the cone of sufficient statistics of hierarchical models. For example, Proposition 19.1.21 together with Proposition 19.1.17 imply Proposition 8.2.9.
19.2. Tree Metrics

A metric is called a \textit{tree metric} if it can be realized by pairwise distances between leaves on a tree with edge lengths. These tree metrics play an important role in phylogenetics.

\textbf{Definition 19.2.1.} Let $\mathcal{T}$ be an $\mathcal{X}$-tree, with $\Sigma(\mathcal{T}) = \{A_1|B_1, \ldots, A_k|B_k\}$ the corresponding set of $\mathcal{X}$-splits. To each split $A_i|B_i \in \Sigma(\mathcal{T})$ let $\lambda_{A_i|B_i} > 0$ be a real edge weight. The \textit{tree metric} associated to these data is the \textit{semimetric}:

$$\sum_{i=1}^{k} \lambda_{A_i|B_i} \delta_{A_i|B_i}.$$  

A semimetric is called a tree metric if it can be expressed in this form for some tree and weights.

Note that an alternate description of the tree metric is to start with a tree $\mathcal{T}$, put the edge weight $\lambda_{A_i|B_i}$ on the edge corresponding to the split $A_i|B_i$ in the tree, and then define $d(x, y)$ to be the sum of the edge weights on the unique path between $x$ and $y$. This is equivalent to Definition 19.2.1, because the only splits that contribute to the distances between $x$ and $y$, are the ones where $\#(\{x, y\} \cap A_i) = 1$.

A key result about tree metrics is the four point condition of Buneman \cite{Bun74}, which shows that tree metrics can be detected in polynomial time. Furthermore, one only need to check locally that a metric is a tree metric.

\textbf{Theorem 19.2.2 (Four point condition).} A metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is a tree metric if and only if for all $x, y, z, w \in \mathcal{X}$

$$d(x, y) + d(z, w) \leq \max\{d(x, z) + d(y, w), d(x, w) + d(y, z)\}.$$  

We allow repetitions among $x, y, z, w$, which insures that $d$ satisfies non-negativity and the triangle inequalities. An equivalent description of the four point condition is that among the three quantities

$$d(x, y) + d(z, w), \quad d(x, z) + d(y, w), \quad d(x, w) + d(y, z)$$

the maximum value is achieved by two of them.

\textbf{Proof.} It is easy to see that a tree metric satisfies the four point condition. This follows by looking just at four leaf trees as illustrated by Figure 19.2.1.

To see that a metric that satisfies the four point condition comes from a tree, we proceed by induction on $\#\mathcal{X}$. If $\#\mathcal{X} = 3$ any metric is a tree metric, for example set

$$\lambda_{1|23} = \frac{d(1, 2) + d(1, 3) - d(2, 3)}{2} \quad \lambda_{2|13} = \frac{d(1, 2) + d(2, 3) - d(1, 3)}{2}$$


Figure 19.2.1. Graphical illustration of the four point condition. Note that the maximum of \(d(x,y) + d(z,w), d(x,z) + d(y,w), d(x,w) + d(y,z)\) is attained twice.

\[
\lambda_{3|12} = \frac{d(1,3) + d(2,3) - d(1,2)}{2}
\]

in which case

\[
d = \lambda_{1|23}\delta_{1|23} + \lambda_{2|13}\delta_{2|13} + \lambda_{3|12}\delta_{3|12}.
\]

So assume that \(\#X > 3\). Let \(x, y, z \in X\) such that

\[
d(x,y) + d(x,z) - d(y,z)
\]

is maximized. Note that for any \(w \in X\) this forces the pair of inequalities

\[
d(w,x) + d(y,z) \leq d(w,y) + d(x,z)\quad d(w,x) + d(y,x) \leq d(w,z) + d(x,y)
\]

so by the four point condition, we have

\[
d(w,x) + d(y,z) \leq \max(d(w,y) + d(x,z), d(w,z) + d(x,y))
\]

so that \(d(w,y) + d(x,z) = d(w,z) + d(x,y)\). This holds for any \(w \in X\).

Combining two such equations with \(w\) and \(w'\) yields

\[
(19.2.2)\quad d(w,y) + d(w',z) = d(w,z) + d(w',y) \geq d(w,w') + d(y,z).
\]

Define a new element \(t\) and set

\[
d(t,y) = \frac{d(x,y) + d(y,z) - d(x,z)}{2}, \quad \text{and}
\]

\[
d(t,u) = d(u,y) - d(t,y) \quad \text{for all } u \in X \setminus \{y,z\}.
\]

We claim that the set \(X \cup \{t\} \setminus \{y,z\}\) satisfies the four point condition with the dissimilarity map \(d\). This is clear for any \(a,b,c,u \in X \setminus \{y,z\}\), we must handle the case where one of these is \(t\). For example:

\[
d(a,b) + d(c,t) = d(a,b) + d(c,y) - d(t,y)
\]

\[
\leq \max(d(a,y) + d(c,b), d(a,c) + d(b,y)) - d(t,y)
\]

\[
= \max(d(a,y) + d(c,b) - d(t,y), d(a,c) + d(b,y) - d(t,y))
\]

\[
= \max(d(a,t) + d(c,b), d(a,c) + d(b,t)).
\]
From this we also deduce nonnegativity and that the extended $d$ forms a metric. By induction we have a tree representation for $d$ on $\mathcal{X} \cup \{t\} \setminus \{y, z\}$, and using the definitions of $d(t, y)$ and $d(t, z)$ we get a tree representation of $d$ on $\mathcal{X}$.

A key part of the proof of Theorem 19.2.2 is the identification of the pair $\{y, z\}$ which form a cherry in the resulting tree representation. Many algorithms for reconstructing phylogenetic trees from distance data work by trying to identify cherries and proceeding inductively. We will discuss this in more detail in the next section.

A key feature of tree metrics is that the underlying tree representation of a tree metric is uniquely determined by the metric. This can be useful in identifiability proofs for phylogenetic models (e.g. [ARS17]).

**Theorem 19.2.3.** Let $d$ be a tree metric. Then there is a unique tree $T$ with splits $\Sigma(T) = \{A_1|B_1, \ldots, A_k|B_k\}$ and positive constants $\lambda_{A_1|B_1}, \ldots, \lambda_{A_k|B_k}$ such that $d = \sum_{i=1}^{k} \lambda_{A_i|B_i} \delta_{A_i|B_i}$.

**Proof.** Suppose that $d$ is a tree metric. First we see that the splits are uniquely determined. From the four point condition we can determine, for each four element set, which of the four possible unrooted four leaf trees is compatible with the given tree metric. Indeed, we compute the three quantities

$$d(x, y) + d(z, w), \quad d(x, z) + d(y, w), \quad d(x, w) + d(y, z).$$

If one value is smaller than the other two, say $d(x, y) + d(z, w)$, this tells us our quartet tree has the split $xy|zw$. If all three values are the same, then the quartet tree is an unresolved four-leaf star tree.

From determining, for each quartet of leaves $x, y, z, w$, which quartet tree is present, we can then determine which splits must be present in the tree. This follows because a split $A|B$ will appear in the tree if and only if for each $x, y \in A$ and $z, w \in B$, $xy|zw$ is the induced quartet tree. Then by the Splits Equivalence Theorem (Theorem 15.1.6), the underlying tree is completely determined.

To see that the weights are uniquely determined, we use an analogue of the argument in the proof of Theorem 19.2.2 to show that we can recover each edge weight. Suppose that we have a leaf edge $e$ (corresponding to split $A|B$), which adjacent to the leaf $x$ and interior vertex $v$. Let $y$ and $z$ be leaves such that the path connecting $y$ and $z$ passes through $v$. Then

$$\lambda_{A|B} = \frac{1}{2}(d(x, y) + d(x, z) - d(y, z)).$$

Otherwise, let $e$ be an internal edge (corresponding to split $A|B$), with vertices $u$ and $v$. Let $x$ and $y$ be leaves such that the path from $x$ to $y$
passes through $u$ but not $v$. Let $w$ and $z$ be leaves such that the path from $w$ to $z$ passes through $v$ but not $u$. Then

$$\lambda_{A|B} = \frac{1}{2}(d(x, z) + d(y, w) - d(x, y) - d(w, z)).$$

Since we can recover the $\lambda_{A|B}$ parameters from the distances, this guarantees the uniqueness of the edge weights. $\square$

In addition to tree metrics, it is also natural to consider the tree metrics that arise naturally from rooted trees, as these also have both important mathematical connections and are connected to problems in phylogenetics.

**Definition 19.2.4.** A tree metric is **equidistant** if it arises from a root tree and the distance between a leaf and the root is the same for all leaves.

Note that in an equidistant tree metric the distance between any pair of leaves and their most recent common ancestor will be the same. Equidistant tree metrics are natural to consider in biological situations where there is a molecular clock assumption in the mutations, that is the rate of mutations are constant across the tree, and not dependent on the particular taxa. This might be reasonable to assume in if all taxa considered are closely related to one another and have similar generation times. On the other hand, distantly related species with widely different generation times are probably not well-modeled by an equidistant tree. In this sense the term branch length used in phylogenetics is not interchangeable with time.

**Definition 19.2.5.** A dissimilarity map $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called an **ultrametric** if for all $x, y, z \in \mathcal{X}$

$$d(x, y) \leq \max(d(x, z), d(y, z)).$$

Note that an ultrametric is automatically a semimetric. This follows since taking $x = y$ forces $d$ to be nonnegative, and $\max(a, b) \leq a + b$ for non-negative numbers implies that the triangle inequalities are satisfied. Sometimes it can be useful to only require the ultrametric condition on triples of distinct elements, in which case there can be negative entries in $d$.

In fact, equidistant tree metrics and ultrametrics are two different representations of the same structure, analogous to the four point condition for tree metrics.

**Proposition 19.2.6.** A semimetric is an ultrametric if and only if it has a representation as an equidistant tree metric.

**Proof.** Let $d$ be an equidistant tree metric. Since passing to a subset of $\mathcal{X}$ amounts to moving to a subtree, it suffices to prove that 3-leaf equidistant tree metrics are ultrametrics. This is easy to see by inspection, using the
19.2. Tree Metrics

fact that the ultrametric condition is equivalent to the condition that the maximum of \(d(x, y), d(x, z), d(y, z)\) is attained twice.

Now let \(d\) be an ultrametric. We will show that it is possible to build an equidistant tree representation of \(d\). We proceed by induction on \(\#\mathcal{X}\).

If \(\#\mathcal{X} = 2\), then clearly \(d\) can be represented by an equidistant tree metric, merely place the root at the halfway point on the unique edge of the tree.

Now suppose that \(\#\mathcal{X} > 2\). Let \(x, y \in \mathcal{X}\) be a pair such that \(d(x, y)\) is minimal. Let \(\mathcal{X}' = \mathcal{X} \setminus \{y\}\) and \(d'\) the associated ultrametric on this subset.

By induction there is an equidistant tree representation of \(d'\). Since the pair \(x, y\) had \(d(x, y)\) minimal, we can add \(y\) by making it a cherry with \(x\) in the tree, with the length of the resulting branch \(d(x, y)/2\). Since \(d\) is ultrametric and \(d(x, y)\) is the minimal distance, \(d(x, z) = d(y, z)\) for all \(z \in \mathcal{X} \setminus \{y, z\}\). Thus, we get an equidistant representation of \(d\).

We have seen in Theorem 19.2.3 how to give a unique representation of each tree metric as the nonnegative combination of unique cut semimetrics associated to splits in the tree. The description can also be interpreted as decomposing the set of tree metrics into simplicial polyhedral cones, one for each binary phylogenetic tree. Equidistant tree metrics can be described in a similar way, but this requires a different set of generating vectors for the polyhedral cones, which we describe now.

Each interior vertex \(v \in \text{Int}(T)\) in the tree induces a set of leaves \(A_v\) which are the leaves that lie below that vertex. The collection of sets \(\mathcal{H}(T) = \{A_v|v \in \text{Int}(T)\}\) is the hierarchy induced by the rooted tree \(T\).

To each set \(A \in \mathcal{H}(T)\) introduce a semimetric \(\delta_A\) with

\[
\delta_A(i, j) = \begin{cases} 
0 & \text{if } i, j \in A_v \\
1 & \text{otherwise}
\end{cases}.
\]

**Proposition 19.2.7.** The set of ultrametrics compatible with a given rooted tree structure \(T\) is the set of ultrametrics of the form

\[
\sum_{A \in \mathcal{H}(T)} \lambda_A \delta_A
\]

such that \(\lambda_A \geq 0\). For a given ultrametric \(d\), there is a unique rooted tree \(T\), and positive edge weights \(\lambda_A\) such that

\[
d = \sum_{A \in \mathcal{H}(T)} \lambda_A \delta_A.
\]

The proof of Proposition 19.2.7 is left as an exercise. Tree metrics are closely related to objects in tropical geometry. Though this does not play a major role in this work, we mention these connections to conclude our discussion.
The set of all tree metrics has the structure of a tropical variety, namely it is the tropical Grassmannian of tropical lines \[SS04\]. Recall that the Grassmannian of lines, \(G_{2,m}\), is the projective variety in \(P_2^{(m)^{-1}}\) defined by the vanishing of the Plücker relations:
\[
p_{ij}p_{kl} - p_{ik}p_{jl} + p_{il}p_{jk} = 0
\]
for all \(1 \leq i < j < k < l \leq m\). The tropical Grassmannian is obtained by tropicalizing these equations, that is, replacing each instance with + or − by tropical addition ⊕ and each instance of multiplication by tropical multiplication ⊙. The tropical Plücker relation has the form
\[
d_{ij} \odot d_{kl} \odot d_{ik} \odot d_{jl} \odot d_{il} \odot d_{jk} = 0
\]
where “\(= 0\)” is interpreted tropically to mean that the maximum is attained twice. That is, this tropical Plücker relation is equivalent to the statement that the maximum of the three quantities
\[
d_{ij} + d_{kl}, \quad d_{ik} + d_{jl}, \quad d_{il} + d_{jk}
\]
is attained twice. Hence, the tropical Plücker relations give the four point condition. Note however, that we only have tropical Plücker relations in the case that all of \(i, j, k, l\) are distinct, so that the tropical Grassmannian contains dissimilarity maps that are not tree metrics because negative entries are allowed on the leaf edges. This is summarized by the following theorem from \[SS04\].

**Theorem 19.2.8.** The tropical Grassmannian consists of the solutions to the tropical Plücker relations
\[
d_{ij} \odot d_{kl} \odot d_{ik} \odot d_{jl} \odot d_{il} \odot d_{jk} = 0 \quad \text{for all} \quad 1 \leq i < j < k < l \leq m.
\]
It consists of all dissimilarity maps \(d\) in \(\mathbb{R}_t^{(m)}\) such that there exists a tree \(T\) and edge weights \(\lambda_{A|B} \in \mathbb{R}\) with \(A|B \in \Sigma(T)\) with
\[
d = \sum_{A|B \in \Sigma(T)} \lambda_{A|B} \delta_{A|B}
\]
where \(\lambda_{A|B} \geq 0\) if both \(\#A > 1\) and \(\#B > 1\).

Similarly, the set of ultrametrics also has the structure of a tropical variety. It is the tropicalization of the linear space associated to the \(K_m\) matroid \[AK06\]. The corresponding tropical linear polynomials that define this linear space are given by
\[
d_{ij} \odot d_{ik} \odot d_{jk} = 0
\]
for all \(1 \leq i < j < k \leq m\). This is, of course, the ultrametric condition. Again, since we only have the case where all of \(i, j, k\) are distinct, this tropical variety has dissimilarity maps with negative entries.
19.3. Finding an Optimal Tree Metric

Even the property of a dissimilarity map being a metric has a tropical interpretation. Letting $D$ denote the matrix representation of the dissimilarity map, it will be a metric precisely when

$$D \geq (-D) \odot (-D)$$

where the $\odot$ in this formula denotes the tropical matrix multiplication.

19.3. Finding an Optimal Tree Metric

The main problem in phylogenetics is to reconstruct the evolutionary history of a collection of taxa from data on these taxa. In Chapter 15 we discussed phylogenetic models, their algebraic structure, and their use in reconstructing these evolutionary histories. The input for methods based on phylogenetic models is typically a collection of aligned DNA sequences, and statistical procedures like maximum likelihood estimation are used to infer the underlying model parameters and reconstruct the tree. An alternate approach is to summarize the alignment information by a dissimilarity map between the taxa and then use a method to try to construct a nearby tree metric from this “distance” data. We describe such distance-based methods in this section.

Let $T_X$ denote the set of all tree metrics and $U_X$ the set of ultrametrics on the leaf label set $X$. Note that if $\#X = m$, then $U_X \subset T_X \subset \mathbb{R}^{(m\choose 2)}$. Given a dissimilarity map $d \in \mathbb{R}^{(m\choose 2)}$ finding an optimal tree metric amounts to solving an optimization problem of the form

$$\min ||d - \hat{d}||_p \quad \text{subject to } \hat{d} \in T_X.$$ 

Here $|| \cdot ||_p$ denotes a $p$-metric on $\mathbb{R}^{(m\choose 2)}$, $1 \leq p \leq \infty$. One could also look at the analogous problem over the set $U_X$ of ultrametrics. In fact, the majority of these optimization problems are known to be NP-hard (see [KM86] for proof of this for $L^2$ optimization for $U_X$). One exception is the case of $p = \infty$ over the set of ultrametrics. There is an algorithm based for finding an $L^\infty$ optimal ultrametric given a dissimilarity map, based on finding the subdominant ultrametric and shifting it [CF00, Ard04]. However, there is almost never a unique closest ultrametric in the $L^\infty$ norm and there are full dimensional regions of the metric cone where multiple different tree topologies are possible for points that are closest in the $L^\infty$ norm [BL16].

Perhaps the most appealing optimization from the statistical standpoint is to optimize with respect to the standard Euclidean norm ($p = 2$). This problem is also known as the least-squares phylogeny problem. While least-squares phylogeny is also known to be NP-hard, there are greedy versions of least-squares phylogeny that are at least consistent and fast. We describe two such procedures here, the unweighted pair group method with arithmetic
finite metric spaces [SM58] and neighbor joining [SN87]. Both methods are agglomerative, in the sense that they use a criterion to decide which leaves in the tree form a cherry, then they join those two leaves together and continue the procedure.

UPGMA is a greedy algorithm that attempts to solve the NP-hard optimization problem of constructing the least-squares phylogeny among equidistant trees. UPGMA has a weighting scheme that remembers the sizes of groups of taxa that have been joined together. It can be seen as acting step-by-step to construct the rooted tree in the language of chains in the partition lattice.

**Definition 19.3.1.** The partition lattice \( \Pi_m \) consists of all unordered set partitions of the set \([m]\). The partitions are given a partial order \(<\) by declaring that the partition \( \pi = \pi^1 \cdot \pi^2 \cdot \cdots \cdot \pi^k < \tau = \tau^1 \cdot \tau^2 \cdot \cdots \cdot \tau^l \) if for each \( \pi^i \) there is a \( j \) such that \( \pi^i \subseteq \tau^j \).

The connection between the partition lattice and trees is that saturated chains in the partition lattice yield hierarchies which in turn correspond to rooted trees. Each covering relation in a saturated chain in the partition lattice corresponds to two sets in the set partition being joined together, and this yields a new set in the hierarchy.

**Algorithm 19.3.2 (UPGMA Algorithm).**

- **Input:** A dissimilarity map \( d \in \mathbb{R}^{m(m-1)/2} \) on \( X \).
- **Output:** A maximal chain \( C = \pi_m < \pi_{m-1} < \cdots < \pi_1 \) in the partition lattice \( \Pi_m \) and an equidistant tree metric \( \hat{d} \).
- **Initialize:** Set \( \pi_m = 1|2|\cdots|m \) and \( d^m = d \).
- For \( i = m-1, \ldots, 1 \) do
  - From partition \( \pi_{i+1} = \pi^1_{i+1} \cdot \pi^2_{i+1} \cdot \cdots \cdot \pi^k_{i+1} \) and dissimilarity map \( d^{i+1} \in \mathbb{R}^{(i+1)(i+2)/2} \) choose \( j, k \in [i+1] \) such that \( d^{i+1}(\pi^j_{i+1}, \pi^k_{i+1}) \) is minimized.
  - Set \( \pi_i \) to be the partition obtained from \( \pi_{i+1} \) by merging \( \pi^j_{i+1} \) and \( \pi^k_{i+1} \) and leaving all other parts the same. So \( \pi^1_i = \pi^1_{i+1} \cup \pi^k_{i+1} \).
  - Create new dissimilarity \( d^i \in \mathbb{R}^{(i-1)/2} \) by defining \( d^i(\tau, \tau') = d^{i+1}(\tau, \tau') \) if \( \tau, \tau' \) are both parts of \( \pi_{i+1} \) and

\[
d^i(\tau, \pi^j_i) = \frac{|\pi^j_{i+1}|}{|\pi^j_i|} d^{i+1}(\tau, \pi^j_{i+1}) + \frac{|\pi^k_{i+1}|}{|\pi^k_i|} d^{i+1}(\tau, \pi^k_{i+1})
\]

- For each \( x \in \pi^j_{i+1} \) and \( y \in \pi^k_{i+1} \) set \( \hat{d}(x, y) = d^{i+1}(\pi^j_{i+1}, \pi^k_{i+1}) \).
- Return the chain \( C = \pi_n < \cdots < \pi_1 \) and the equidistant metric \( \hat{d} \).
Note that Equation 19.3.1 is giving an efficient recursive procedure to calculate the dissimilarity

\[ d^*(\tau, \tau') = \frac{1}{|\tau||\tau'|} \sum_{x \in \tau, y \in \tau'} d(x, y). \]

The UPGMA algorithm is a consistent algorithm. In other words, if we input a dissimilarity map \( d \) that is actually an ultrametric, the algorithm will return the same ultrametric. This is because the minimum value that occurs at each step must be a cherry, and the averages that occur are always amongst values that are all the same. So one step of UPGMA takes an ultrametric and produces a new ultrametric on a smaller number of taxa. So to prove consistency of UPGMA one must adapt the proof of Proposition 19.2.6 that shows that ultrametrics are equidistant tree metrics.

The reader might be wondering why we say that UPGMA is a greedy procedure for solving the least-squares phylogeny problem. It is certainly greedy, always merging together pairs of leaves that are closest together, but what does this have to do with \( L^2 \) optimization. This is contained in the following proposition.

**Proposition 19.3.3. [FHKT08]** Let \( d \in \mathbb{R}^{m(m-1)/2} \) be a dissimilarity map on \( X \) and let \( C = \pi_m < \ldots < \pi_1 \) and \( \hat{d} \) be the chain and ultrametric returned by the UPGMA algorithm. Then \( \hat{d} \) is a linear projection of \( d \) onto one of the cones of \( \mathcal{U}_X \).

**Proof.** Let \( C = \pi_m < \ldots < \pi_1 \) be a chain of the \( \Pi_m \) and let \( L_C \) be the corresponding linear space. This space is obtained by setting \( d(x, y) = d(x', y') \) for all tuples \( x, x' \in \tau, y, y' \in \tau' \) where \( \tau \) and \( \tau' \) are a pair of parts of partition that get merged going from \( \pi_{i+1} \) to \( \pi_i \) for some \( i \). In other words, this linear space is a direct product of linear spaces of the form

\[ \left\{(t, t, \ldots, t) : t \in \mathbb{R} \right\} \subseteq \mathbb{R}^k \]

for some \( k \). The projection onto a direct product is the product of the projections so we just need to understand what the projection of a point \( (y_1, \ldots, y_k) \) onto such a linear space is. A straightforward calculation shows that this projection is obtained by the point \( (\bar{y}, \ldots, \bar{y}) \) where \( \bar{y} = \frac{1}{k} \sum_{i=1}^{k} y_i \). This shows that UPGMA is a linear projection onto the linear space spanned by one of the cones that make up the space of ultrametric trees. However, by the UPGMA procedure that always the smallest element of the current dissimilarity map is chosen, this forces that the inequalities that define the cone will also be satisfied, and hence UPGMA will map into a cone.

Although UPGMA does give a projection onto one of the cones of \( \mathcal{U}_X \), a property that the actual least squares phylogeny will satisfy, the result of
UPGMA need not be the actual least squares phylogeny. This is because the greediness of the algorithm might project onto the wrong cone in $U_X$. This is illustrated in the following example.

**Example 19.3.4** (UPGMA Projection). Consider the following dissimilarity map:

$$d = \begin{pmatrix}
    d(x, y) & d(x, z) & d(x, w) \\
    d(y, z) & d(y, w) & d(z, w)
\end{pmatrix} = \begin{pmatrix}
    10 & 40 & 90 \\
    60 & 90 & 48
\end{pmatrix}.$$

The first step of UPGMA makes $x$ and $y$ a cherry with distance 10, and produces the new dissimilarity map

$$\begin{pmatrix}
    50 & 90 \\
    48
\end{pmatrix}.$$

The next step of UPGMA makes $z$ and $w$ a cherry with distance 48. The remaining weight for other pairs is $(40 + 60 + 90 + 90)/4 = 70$. So the UPGMA reconstructed ultrametric is

$$\hat{d}_{UPGMA} = \begin{pmatrix}
    10 & 70 & 70 \\
    70 & 70 & 48
\end{pmatrix}$$

and $\|d - \hat{d}_{UPGMA}\|_2 = \sqrt{1800}$. On the other hand, the least-squares ultrametric is

$$\hat{d}_{LS} = \begin{pmatrix}
    10 & 50 & 76 \\
    50 & 76 & 76
\end{pmatrix}$$

with $\|d - \hat{d}_{LS}\|_2 = \sqrt{1376}$.

Studies of the polyhedral geometry of the UPGMA algorithm and comparisons with the least-squares phylogeny were carried out in [DS13, DS14, FHKT08].

Neighbor joining [SN87] is a similar method for constructing a tree from distance data, though it constructs a general tree metric, instead of an equidistant tree metric. It also is given by an agglomerative/recursive procedure, but uses the same formulas at each step, rather than relying on the structure of the partition of leaves that has been formed by previous agglomerations, as in the UPGMA algorithm. Associated to the given
dissimilarity map we compute the matrix $Q$ defined by

$$Q(x, y) = (m - 2)d(x, y) - \sum_{k=1}^{m} d(x, k) - \sum_{k=1}^{m} d(y, k).$$

Then we pick the pair $x, y$ such that $Q(x, y)$ is minimized. These will form a cherry in the resulting tree. A new node $z$ is created and estimates of the distance from $z$ to $x, y$, and the other nodes $w \in X$ are computed via the following formulas:

$$d(x, z) = \frac{1}{2}d(x, y) + \frac{1}{2(m - 2)} \left( \sum_{k=1}^{m} d(x, k) - \sum_{k=1}^{m} d(y, k) \right),$$

$$d(y, z) = d(x, y) - d(x, z),$$

and

$$d(w, z) = \frac{1}{2}(d(x, z) + d(y, z) - d(x, y)).$$

**Proposition 19.3.5.** Neighbor joining is a consistent algorithm for reconstructing tree metrics. That is, given a tree metric as input, the neighbor joining algorithm will return the same tree metric.

See [PS05, Corollary 2.40] for a proof. Recent work on the polyhedral geometry of neighbor joining include [EY08, HHY11, MLP09].

Methods based on $L^p$ optimization are not the only approaches to constructing phylogenetic trees from distance data, and there are a number of other methods that are proposed and used in practice. One other optimization criterion is the balanced minimum evolution criterion, which leads to problems in polyhedral geometry. The balanced minimum evolution problem is described as follows. Fixed an observed dissimilarity map $d : X \times X \rightarrow \mathbb{R}$. For each binary phylogenetic $X$-tree $T$, let $\hat{d}_T$ be the orthogonal projection onto the linear space spanned by $\{\delta_{A|B} : A|B \in \Sigma(T)\}$. Choose the tree $T$ such that the sum of the edge lengths in $\hat{d}_T$ is minimal.

Paulin [Pau00] showed that the balanced minimal evolution problem is equivalent to a linear optimization problem over a polytope. To each binary phylogenetic $X$-tree $T$, introduce a vector $w(T) \in \mathbb{R}^{\binom{m}{2}}$ by the rule:

$$w(T)_{ij} = 2^{-b_{ij}}$$

where $b_{ij}$ is the number of interior vertices of $T$ on the unique path from $i$ to $j$. Letting $P_m$ denote the convex hull of all $(2m - 5)!$ tree vectors $w(T)$, Paulin’s formula shows that the balanced minimum evolution tree is obtained as the vector that minimizes the linear function $d \cdot x$ over $x \in P_m$. Computing the balanced minimum evolution tree is NP-hard [Day87], so there is little hope for a simple description of $P_m$ via its facet defining inequalities. However, there has been recent work developing classes of facets, which might be useful for speeding up heuristic searches for the balanced minimum evolution tree [FKS16, FKS17, HHY11].
19.4. Toric Varieties Associated to Finite Metric Spaces

To conclude this chapter, we discuss the ways that the toric varieties associated to the polytopes from finite metric spaces arise in various contexts in algebraic statistics. In particular, for each graph $G$ the cut polytope $\text{Cut}^\ominus(G)$ defines a projective toric variety with a homogeneous vanishing ideal. This is obtained by lifting each cut semimetric $\delta_{A|B}$ into $\mathbb{R}^{(|G|+1)}$ by adding a single new coordinate and setting that equal to 1. These toric varieties and their corresponding ideals were introduced in [SS08]. We study the toric ideals in the ring $\mathbb{C}[q_{A|B}]$ where $A|B$ ranges over all $2^{m-1}$ splits and we denote the corresponding toric ideal by $I_G^\ominus$, which is called the cut ideal of the graph $G$.

**Example 19.4.1 (Cut Ideal of the Four Cycle).** Let $G = C_4$ be the four-cycle graph. The cut ideal

$$I_{C_4}^\ominus \subseteq \mathbb{C}[q_{0|234}, q_{1|234}, q_{2|134}, q_{12|34}, q_{123|24}, q_{23|14}, q_{123|4}]$$

is the toric ideal associated to the following matrix:

$$
\begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 1 & 1 & 1 & 0 & 0 & . \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1
\end{pmatrix}
$$

The order of the columns of this matrix matches the ordering of the indeterminates in the ring $\mathbb{C}[q_{A|B}]$. The toric ideal in this case is

$$I_{C_4}^\ominus = \langle q_{0|234}q_{13|24} - q_{1|234}q_{3|24} - q_{0|234}q_{12|34} - q_{0|123|24} - q_{23|14} - q_{123|4} \rangle.$$

The best general result characterizing generating sets of these cut ideals is due to Engström [Eng11]:

**Theorem 19.4.2.** The cut ideal $I_G^\ominus$ is generated by polynomials of degree $\leq 2$ if and only if $G$ is free of $K_4$-minors.

An outstanding open problem from [SS08] is whether or not similar excluded minor conditions suffice to characterize when the generating degree of $I_G^\ominus$ is bounded by any number. In particular, it is conjectured that $I_G^\ominus$ is generated by polynomials of degree $\leq 4$ if and only if $G$ is free of $K_5$ minors.

**Proposition 19.1.21** allows us to realize an equivalence between the cut ideal $I_G^\ominus$ of the suspension graph $\hat{G}$ and the toric ideal of the binary hierarchical model where the underlying simplicial complex is the graph $G$. This
allows a straightforward importation of results about these cut ideals for hierarchical models. For example, a direct application of Proposition 19.1.21 to Theorem 19.4.2 produces:

**Corollary 19.4.3.** The toric ideal of a binary graph model is generated by polynomials of degree $\leq 2$ if and only if the graph $G$ is a forest.

A surprising further connection to other book topics comes from the connection to the toric ideals that arise in the study of group-based phylogenetic models that were studied in Section 15.3, and some generalizations of those models. This relation was originally discovered in [SS08].

**Definition 19.4.4.** A cyclic split is a split $A|B$ where either $A$ or $B$ is of the form $\{i, i+1, \ldots, j-1, j\}$ for some $i$ and $j$. A cyclic split system is a set of cyclic splits on the same ground set $[m]$.

The Splits Equivalence Theorem (Theorem 15.1.6) says that each tree $T$ is equivalent to its set of displayed splits $\Sigma(T)$. Every isomorphism class of tree can be realized with a set of cyclic splits, after possibly relabeling the leaves. For instance, this can be obtained by drawing the tree in the plane with leaves all on a circle, and then cyclically labeling the leaves $1, 2, \ldots, m$ as the circle is traversed once. Hence, each phylogenetic tree corresponds to a cyclic split system. We will show that certain generalizations of the phylogenetic group-based models associated to cyclic split systems are in fact isomorphic to the cut ideal $I_G^e$.

Let $\Sigma$ be a cyclic split system on $[m]$. We associate the matrix whose rows are indexed by the elements of $\Sigma$, and whose columns are indexed by subsets of $[m]$ that have an even number of elements. To each such subset $S \subseteq [m]$ we associated the vector $\gamma_S$ where

$$\gamma_S(A|B) = \begin{cases} 1 & \text{if } |A \cap S| \text{ is odd} \\ 0 & \text{otherwise.} \end{cases}$$

The associated toric ideal is obtained from this set of vectors by adding a row of all ones to the matrix (as we did to associate the toric ideal to a cut polytope). Note that if the cyclic split system $\Sigma$ comes from a tree, then this toric ideal will be the same as the vanishing ideal of the Cavender-Farris-Neyman model introduced in Section 15.3, associated to the group $\mathbb{Z}_2$. This is true because firstly, subsets of $[m]$ with an even number of elements are in natural bijection with the subgroup of $\mathbb{Z}_2^m$ of elements $(g_1, \ldots, g_m)$ such that $g_1 + \cdots + g_m = 0$. And secondly, the sum of group elements associated to one side of an edge will contribute a nontrivial parameter precisely when the sum on one side of a split is nonzero in the group, i.e. when $|S \cap A|$ is odd.
To an edge \((i, j + 1)\) in the graph \(G\) we associate the split \(A_{ij} | B_{ij}\) with \(A_{ij} = \{i, i + 1, \ldots, j - 1, j\}\) where we consider indices cyclically modulo \(m\). This is clearly a bijection between unordered pairs in \([m]\) and cyclic splits. To an arbitrary split \(A | B\) we associate an even cardinality subset \(S(A | B)\) by the rule that \(k \in S\) if and only if \(|\{k, k + 1\} \cap A| = 1\). This is a bijection between splits of \([m]\) and even cardinality subsets of \([m]\). Lastly, by translating the two definitions, we see that

\[
\delta_{A | B}(i, j + 1) = \gamma_{S(A | B)}(A_{ij} | B_{ij}).
\]

To a graph \(G\) let \(\Sigma(G)\) denote the resulting set of cyclic splits that are induced by the map \(i, j + 1 \rightarrow A_{ij} | B_{ij}\). Then we have the following proposition.

**Proposition 19.4.5.** Let \(G\) be a graph with vertex set \([m]\). The cut ideal \(I_G^\Sigma\) of the graph \(G\) is equivalent to the toric ideal associated to the set of vectors \(\gamma_S\) associated to the set of cyclic splits \(\Sigma(G)\).

In the special case where the graph \(G\) is a triangulation of the \(n\) cycle graph, then the corresponding set of cyclic splits is \(\Sigma(G)\) is the set of cyclic splits of the dual tree. This is illustrated in the following example.

**Example 19.4.6 (Snowflake Tree).** Let \(G\) be the graph with vertex set \([m]\) and edges \(E = \{12, 23, 34, 45, 56, 16, 13, 35, 15\}\). The transformation from Proposition 19.4.5 gives the list of splits

\[
\Sigma(G) = \{1|23456, 2|13456, 3|12456, 4|12356, 5|12346, 6|12345, 12|3456, 34|1256, 56|1234\}
\]

which is the set of splits of a snowflake tree.

The motivation for studying the toric geometry of cut polytopes in [SS08], in fact, originates with Proposition 19.4.5. A growing trend in phylogenetics has been to studying increasingly more complex phylogenetic
models that allow for non-tree like structures. One proposal [Bry05] for such an extension was based on developing models on the tree-like structures of splits that are produced by the NeighborNet algorithm [BMS04]. The paper [SS08] explores the algebraic structure of such models. Each of the various alternate proposals for network and more general phylogenetic models requires its own project to uncover the underlying algebraic structure.

19.5. Exercises


Exercise 19.2. Prove Proposition 19.1.20. The representation of the hierarchical model given the by matrix $B_I$ in Section 9.3 might be helpful.

Exercise 19.3. Prove Proposition 19.2.7.

Exercise 19.4. Prove that the inequality (19.2.3) is equivalent to the triangle inequality.

Exercise 19.5. How many saturated chains in the partition lattice $\Pi_m$ are there? How does this compare to the number of rooted binary trees with $m$ leaf labels?

Exercise 19.6. Fill in the details of the argument that UPGMA is a consistent algorithm.

Exercise 19.7. Give a sequence of examples that shows that UPGMA is not a continuous function.

Exercise 19.8. Compute generators for the toric ideal $I_{K_5}^\triangle$.

Exercise 19.9. Let $C_n$ denote the caterpillar tree with $n$ leaves and $P_n$ denote the path graph with $n$ vertices. Let $I_n \subseteq \mathbb{C}[q]$ denote the vanishing ideal of the CFN model on $C_n$ in the Fourier coordinates. Let $J_n$ denote the vanishing ideal of the binary graph model on $P_n$. Show that after appropriately renaming variables $I_{n+1} = J_n$. 
Bibliography


[AW05] Miki Aoyagi and Sumio Watanabe, Stochastic complexities of reduced rank regression in Bayesian estimation, Neural Networks 18 (2005), no. 7, 924–933.


[BCR98] Jacek Bochnak, Michel Coste, and Marie-Françoise Roy, Real algebraic geometry, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], vol. 36, Springer-Verlag, Berlin, 1998, Translated from the 1987 French original, Revised by the authors. MR 1659509


<table>
<thead>
<tr>
<th>Reference</th>
<th>Details</th>
</tr>
</thead>
</table>
[GPS17] Elizabeth Gross, Sonja Petrović, and Despina Stasi, Goodness of fit for

[GPSS05] Luis D. García-Puente, Michael Stillman, and Bernd Sturmfels, Algebraic
geometry of Bayesian networks, J. Symbolic Comput. 39 (2005), no. 3-4,
331–355.

[Gre63] Ulf Grenander, Probabilities on algebraic structures, John Wiley &
Sons, Inc., New York-London; Almqvist & Wiksell, Stockholm-Göteborg-
Uppsala, 1963. MR 0206994 (34 #6810)

[GS] Daniel R. Grayson and Michael E. Stillman, Macaulay2, a software system
for research in algebraic geometry, Available at https://faculty.math.
ilinois.edu/Macaulay2/.

[GS18] Elizabeth Gross and Seth Sullivant, The maximum likelihood threshold of
a graph, Bernoulli 24 (2018), no. 1, 342–355. MR 3707672

[GT92] Sun Wei Guo and Elizabeth A. Thompson, Performing the exact test of
hardy-weinberg proportions for multiple alleles, Biometrics 48 (1992), 361–
372.

[Hac79] Leonid G. Hacijan, A polynomial algorithm in linear programming, Dokl.
Akad. Nauk SSSR 244 (1979), no. 5, 1093–1096. MR 522052

[Har08] G. H. Hardy, Mendelian proportions in a mixed population, Science
XXXVIII (1908), 49–50.

[Har77] Robin Hartshorne, Algebraic geometry, Springer-Verlag, New York-
Heidelberg, 1977, Graduate Texts in Mathematics, No. 52. MR 0463157

[Har95] Joe Harris, Algebraic geometry, Graduate Texts in Mathematics, vol. 133,
Springer-Verlag, New York, 1995, A first course, Corrected reprint of the
1992 original. MR 1416564

[Has07] Brendan Hassett, Introduction to algebraic geometry, Cambridge University

[HAT12] Hisayuki Hara, Satoshi Aoki, and Akimichi Takemura, Running Markov
chain without Markov basis, Harmony of Gröbner bases and the modern

[Hau88] Dominique M. A. Haughton, On the choice of a model to fit data from an

[HCD+06] Mark Huber, Yuguo Chen, Ian Dinwoodie, Adrian Dobra, and Mike
Nicholas, Monte Carlo algorithms for Hardy-Weinberg proportions, Bio-


[HH03] Raymond Hemmecke and Ralf Hemmecke, 4ti2 - Software for computa-
tion of Hilbert bases, Graver bases, toric Gröbner bases, and more, 2003,
Available at http://www.4ti2.de.

[HH11] Valerie Hower and Christine E. Heitsch, Parametric analysis of RNA

[HHY11] David C. Haws, Terrell L. Hodge, and Ruriko Yoshida, Optimality of
the neighbor joining algorithm and faces of the balanced minimum evolution


[KGPY17b] ______, *lattice: LattE and 4ti2 in R*, 2017, R package version 0.1.0.


[Kru77], Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics, Linear Algebra and Appl. 18 (1977), no. 2, 95–138. MR 0444690 (56 #3040)


David Mond, James Q. Smith, and Duco van Straaten, *Stochastic factorizations, sandwiched simplices and the topology of the space of explanations,*


Index

1-factor model, 131
   as graphical model, 317
identifiability, 365, 370
learning coefficients, 408, 410
sBIC, 412, 413
A
B|C, 70
A|B, 331
Ar, 202
Br, 204
EXn, 314
EXn,b, 315
Fm,s, 131
Gm, 288
G
ub, 321
H(K[x1,...,xm]/M; x), 279
I(W), 43
I(θ), 163
I : J∞, 137
I : J, 80, 137
I : J∞, 80
IP(A, b, c), 229
I/2, 430
I_{A\perp B|C}, 74
I_{A,b}, 121
Ia, 121
Ic, 75
I_{o,r}, 239
J_{A\perp B|C}, 76
Jc, 76
K(K[x1,...,xm]/M; x), 279
LP(A, b, c), 230
L1 embeddable, 433

Lp embeddable, 432
M(A, c), 238
NFp(f), 50
Op(1), 399
P(A), 12
P(A | B), 15
PD(B), 320
PD_m, 76
P_{m,n}, 257
Q^{CPN}, 337
Q^{JC}, 337
Q^{K2P}, 338
S-polynomial, 51
SR_C, 277
TC_{m,n}(Θ), 127
V(S), 40
V_1 * V_2 * ... * V_s, 310
V_{reg}, 145
V_{sing}, 145
X-tree, 331
   binary, 331
   phylogenetic, 331
X(A, D), 263
X^2(u), 110
X_A \perp X_B | X_C, 70
Corr[X, Y], 26
Cov[X, Y], 25
Δt, 42
E[X | Y], 27
E[X], 22
L_V, 147
Σ(T), 354
Σ(T), 332
Index

Var\{X\}, 24
\(\alpha\)-th moment, 103
\(\tilde{\ell}_N(G)\), 257
\(B_k\), 14
\(\mathcal{F}(U)\), 188
\(\mathcal{F}(u, L, U)\), 212
\(\mathcal{M}_1 \times \cdots \times \mathcal{M}_k\), 310
\(\mathcal{M}_G\), 320
\(\mathcal{M}_{\mathcal{A}(\mathcal{B})}\), 74
\(\mathcal{X}_m(p, \Sigma)\), 30
\(T_X\), 445
\(U_X\), 445
\(\mathcal{V}\)-polyhedron, 168
cone\((A)\), 169
cone\((V)\), 168
conv\((S)\), 166
\(-\rightarrow\), 43
\(\delta_A\), 443
\(\delta_{A \cap B}\), 432
exp\((Q)\), 336
\(\gamma_S\), 451
\(G\), 342
\(\mathbb{E}[p]\), 40
\(\Phi(\Omega)\), 12
Flat\(_{A \cap B}(P)\), 353
\(|\Gamma|\), 204
min\((f)\), 128
\(\mathcal{N}A\), 169
\(\pi\)-supermodular, 312
\(\pi^1 \mid \cdots \mid \pi^k\), 446
\(\pi_S\), 435
\(\prec\), 48
\(\preceq\), 48
rank\(_s(A)\), 311
AN\((v)\), 288
de\((v)\), 287
md\((v)\), 288
pa\((v)\), 287
\(\sigma\)-algebra, 14
\(\sqrt{\mathcal{T}}\), 44
d\((x, y)\), 431
p-value, 108
\(p_1\) model, 254
Markov basis, 260
\(tu(G)\), 257
Corr\(_0\), 435
Corr\(_m\), 435
Cut\((G)\), 436
Cut\(_0(G)\), 436
Cut\(_m\), 432
Cut\(_m\), 432
Dist\((M)\), 266
Jac\((f(V))\), 145
Met\((G)\), 438
Mixt\(_k(M)\), 308
Quad\((V_1, S, V_2)\), 207
RLCT\((f; \phi)\), 403
Sec\(_d(V)\), 310
deg\(_l(G)\), 252
gap\((A, c)\), 239
gcr\((L)\), 181
in\(_\prec(f)\), 49
in\(_\prec(J)\), 231
in\(_\prec(f)\), 231
mit\((L)\), 181
acyclic, 287
adjacent minors, 216
affine semigroup, 169
affinely isomorphic, 167
Akaike information criterion (AIC), 393
algebraic exponential family, 115, 130
algebraic tangent cone, 128
algebraic torus, 148
algebraically closed field, 44
aliasing, 268
alignment, 334, 424
Allman-Rhodes-Draisma-Kuttler theorem, 316, 358
almost sure convergence, 32
almost surely, 15
alternative hypothesis, 108
ancestors, 288
Andrews’ Theorem, 426
aperiodic, 338
argmax, 6
arrangement complement, 152
ascending chain, 52, 67
associated prime, 78, 79
balanced minimum evolution, 449
basic closed semialgebraic set, 126
basic semialgebraic set, 126
Bayes’ rule, 16
Bayesian information criterion (BIC), 393, 399
singular (sBIC), 412
Bayesian networks, 296
Bayesian statistics, 111
Bernoulli random variable, 20, 314
Bertini, 58
beta function, 113
beta model, 252
Index

Betti numbers, 61
bidirected edge, 319
bidirected subdivision, 321
binary graph model, 172, 182, 451
binomial ideal, 81, 121
primary decomposition, 81
binomial random variable, 18, 42, 44, 45, 98, 164
implicitization, 56
maximum likelihood estimate, 105
method of moments, 103
Bonferroni bound, 243
Borel $\sigma$-algebra, 14
bounded in probability, 399
bounds, 223
2-way table, 225
3-way table, 241
decomposable marginals, 243
bow-free graph, 374
branch length, 336, 442
Buchberger’s algorithm, 52
Buchberger’s criterion, 52
Buchburger-Möller algorithm, 268
canonical sufficient statistic, 116
caterpillar, 304, 453
Cavender-Farris-Neyman model, 337, 362, 451
cdf, 18
censored exponentials, 138, 146
censored Gaussians, 326
centered Gaussian model, 143
Central limit theorem, 35
CFN model, 337, 345
character
group theory, 342
characteristic class, 148
Chebyshev’s inequality, 24
Chern-Schwartz-Macpherson class, 148
cherry, 441
chi-square distribution, 32, 160
chordal graph, 177, 301
claw tree, 315
Gaussian, 317
clique, 177, 291
clique condition, 177
collider, 287
colon ideal, 80, 137
combinatorially equivalent, 167
compartment model
linear, 385
two, 383, 390
compatible splits, 332
pairwise, 332
complete fan, 272
complete independence, 17, 20, 394
and cdf, 21
and density, 21
complete intersection, 148
completeness of global Markov property, 286, 289
concave function, 150
concentration graph model, 176
concentration matrix, 118
conditional and marginal independence, 76
conditional density, 70
conditional distribution, 19
conditional expectation, 27
conditional independence, 70
conditional independence axioms, 71
conditional independence ideal, 74
conditional independence inference rules, 71
conditional independence model, 4
conditional inference, 185, 188
conditional Poisson distribution, 254
conditional probability, 15
cone generated by $V$, 168
cone of sufficient statistics, 168
and existence of MLE, 169
conformal decomposition, 212
conjugate prior, 113
consistent estimator, 102
constructible sets, 54
continuous random variable, 18
contraction axiom, 71, 84
control variables, 382
convergence in distribution, 34
convergence in probability, 33
converging arborescence, 372
convex function, 150
convex hull, 132, 166
convex set, 150, 165
convolution, 344
coordinate ring, 233
corner cut ideal, 265
correlation, 26
correlation cone, 435
correlation polytope, 173, 435
of simplicial complex, 438
coset, 233
covariance, 25
covariance mapping, 435
covariance matrix, 25
critical equations, 136
cumulants, 357
cumulative distribution function, 18
curved exponential families, 115
cuspidal cubic, 141, 161
cut cone, 432
  of a graph, 436
cut ideal, 450
cut polytope, 173, 432, 450
  of a graph, 436
cut semimetric, 432
cycle, 174
cyclic split, 451
cyclic split system, 451
d-connected, 288
d-separates, 288
DAG, 287
DAG model selection polytope, 397, 414
De Finetti’s theorem, 46, 101, 313
decomposable graph, 177
decomposable marginals, 243
decomposable model
markov basis, 209
urn scheme, 219
decomposable simplicial complex, 204, 301
decomposable tensor, 377
decomposition axiom, 71
defining ideal, 43
degree
  of a vertex, 252
degree sequence, 252
density function, 14
descendants, 287
design, 262
design matrix, 263
deterministic Laplace integral, 401–403
Diaconis-Efron test, 190
differential algebra, 383
differential privacy, 229
dimension, 53
directed acyclic graph, 287, 395
  maximum likelihood estimate, 396
  model selection, 397
  phylogenetic tree, 331
directed cycle, 287
Dirichlet distribution, 113
disclosure limitation, 227
discrete conditional independence
  model, 74
discrete linear model, 151
discrete random variable, 18
dissimilarity map, 431
distance-based methods, 445
distinguishability, 370
distraction, 266
distribution, 17
distributive law, 420, 429
division algorithm, 50
dual group, 342
dimensional, 167
degree, 249–251, 259, 260
elimination ideal, 55, 369
elimination order, 55, 61
EM algorithm, 323, 421
discrete, 324
embedded prime, 79
emission matrix, 419
empirical moment, 103
entropy, 404
equidistant tree metric, 442
equivalent functions, 405
equivariant phylogenetic model, 358
Erdős-Gallai theorem, 252
Erdős-Rényi random graphs, 248, 258
estimable, 263
estimable polynomial regression, 263
estimator, 102
Euclidean distance degree, 141
Euler characteristic, 148
event, 12
exact test, 108, 189
exchangeable, 101, 314
exchangeable polytope, 315
expectation, 22
expected value, 22
experimental design, 261
exponential distribution, 37
exponential family, 116, 165
  discrete random variables, 117, 119
  extended, 119
  Gaussian random variables, 117, 123
  likelihood function, 155
exponential random graph model, 248
exponential random variable, 114
extreme rays, 168
face, 166, 200
Index

facet, 167, 200
factor analysis, 131
identifiability, 389
pentad, 318
factorize, 291
fan, 272
Felsenstein model, 338, 362
few inference function theorem, 427
fiber, 188
Fisher information matrix, 163, 164
Fisher’s exact test, 7, 189, 224, 249
flattening, 353
four point condition, 439
Fourier transform
discrete, 342, 343
fast, 344, 421
fractional factorial design, 262
free resolution, 281
frequentist statistics, 111, 114
full factorial design, 262
fundamental theorem of Markov bases, 43
Gale-Ryser theorem, 253
gamma random variable, 114
Gaussian conditional independence
ideal, 76
Gaussian marginal independence
ML-degree, 143
Gaussian random variables
marginal likelihood, 398
maximum likelihood, 106, 140
Gaussian random vector, 30
Gaussoid axiom, 93
general Markov model, 335, 352
general time reversible model, 339, 341
generalized hypergeometric distribution, 189
generators of an ideal, 43, 51
generic, 137, 265
weight order, 231
generic completion rank, 181
generic group-based model, 345
Gröbner basis, 47, 49
integer programming, 230
to check identifiability, 369
Gröbner cone, 271
Gröbner fan, 270, 272
Gröbner region, 271
graph
of a polynomial map, 55
of a rational map, 57
graph statistic, 248
graphical model, 395
directed, 287
Gaussian, 176
undirected, 284
with hidden variables, 315, 380
Grassmannian, 444
Graver basis, 185, 212
ground set, 204
group, 341
group-based phylogenetic models, 340, 341, 451
half trek, 376
criterion, 376
half-space, 165
Hammersley-Clifford theorem, 291–293
Hardy-Weinberg equilibrium, 197
Markov basis, 199
Hasegawa-Kishino-Yano model, 338, 341, 362
hidden Markov model, 304, 418
pair, 420
hidden variable, 307
hierarchical model, 157, 200, 237
sufficient statistics, 203
hierarchical set of monomials, 264
hierarchy, 443, 446
Hilbert basis theorem, 43
Hilbert series, 279
holes, 172
holonomic gradient method, 327
homogeneous polynomial, 64
homogenization, 128
of parametrization, 66
Horn uniformization, 150
hypercube, 166
hypergeometric distribution, 189, 217
hypersurface, 43
hypothesis test, 107, 159
asymptotic, 111
i.i.d., 100
ideal, 43
ideal membership problem, 47, 50
ideal-variety correspondence, 46
identifiability
of phylogenetic models, 371, 441
practical, 383
structural, 383
identifiable, 98, 364
discrete parameters, 370
generically, 364
globally, 364
locally, 364
parameter, 366
rationally, 364
image of a rational map, 43
implicitization problem, 43, 54
importance sampling, 224
importance weights, 224
independence model, 74, 99
as exponential family, 122
as hierarchical model, 201
Euler characteristic, 149
Graver basis, 212
identifiability, 364
Markov basis, 193, 214
independent and identically distributed, 29, 100
independent events, 16
indeterminates, 40
indicator random variable, 20
induced subgraph, 174
inference functions, 426
information criterion, 393
initial ideal, 49
initial monomial, 49
initial term, 49
input/output equation, 383
instrumental variable, 318, 389
as mixed graph, 322, 323
identifiability, 366, 375
integer program, 229
integer programming gap, 238, 239
intersection axiom, 72, 85, 285
intersection of ideals, 267
interval of integers, 245
inverse linear space, 124
irreducible decomposition of a variety, 77
of an ideal, 78
of monomial ideal, 239
irreducible ideal, 78
irreducible Markov chain, 338
irreducible variety, 77
Ising model, 172, 305
isometric embedding, 432
iterative proportional fitting, 154
Jacobian, 145, 367
augmented, 146
JC69, 337
join variety, 310
joint distribution, 18
Jukes-Cantor model, 337, 362
linear invariants, 351, 362
K-polynomial, 279
K2P model, 338
K3P model, 338, 345
kernel, 62
Kimura models, 337, 362
Kruskal rank, 377
Kruskal’s Theorem, 377
Kullback-Leibler divergence, 404, 406
Lagrange multipliers, 144
Laplace approximation, 399
latent variable, 307
lattice basis ideal, 95
lattice ideal, 81, 121
law of total probability, 16
Lawrence lifting, 221
leading term, 49
learning coefficient, 402
least-squares phylogeny, 445
lexicographic order, 48, 232, 245
Lie Markov model, 340
likelihood correspondence, 147
likelihood function, 5, 104
likelihood geometry, 144, 147
likelihood ratio statistic, 159
likelihood ratio test
asymptotic, 161
linear invariants, 351
linear model, 151
linear program, 229
linear programming relaxation, 229, 234
log-affine model, 120
log-likelihood function, 105
log-linear model, 120
ML-degree, 164
log-odds ratios, 117
logistic regression, 221
long tables, 211
Macaulay2, 57
MAP, 112
marginal
as sufficient statistics, 203
marginal density, 19, 70
marginal independence, 71, 74, 88
marginal likelihood, 112, 398
Markov basis, 185, 190, 230, 253
and toric ideal, 192
minimal, 195
Markov chain, 2, 101, 303
homogeneous, 304, 419
Markov chain Monte Carlo, 191
Markov equivalent, 290
Markov property, 284
directed global, 288
directed local, 288
directed pairwise, 288
global, 284
local, 284
pairwise, 284
Markov subbasis, 216
matrix exponential, 336
matrix Lie algebra, 340
matrix product variety, 360
maximum a posteriori estimation, 112, 415, 416
maximum likelihood, 392
maximum likelihood degree, 138
maximum likelihood estimate, 5, 104, 135
existence, 169, 251
maximum likelihood threshold, 181
mean, 22
measurable sets, 14
method of moments, 103, 155
metric, 431
metric cone, 432
of graph, 438
metric space, 431
$L^p$, 432
Metropolis–Hastings algorithm, 190
for uniform distribution, 192
minimal Gröbner basis, 68
minimal prime, 79
minimal sufficient statistic, 100
Minkowski sum, 168, 427
Minkowski–Weyl theorem, 168
mixed graph, 319
simple, 374
mixing time, 191
mixture model, 45, 308, 310, 391
mixture of binomial random variable, 45, 164
mixture of complete independence, 309, 312, 316, 328, 357
identifiability, 378, 379
mixture of independence model, 309, 327
EM algorithm, 326, 328
identifiability, 365
nonnegative rank, 311
mixture of known distributions, 152
ML-degree, 135, 138
ML-degree 1, 149
MLE, 104
model selection, 391
consistency of, 393
molecular clock, 442
moment, 103
moment polytope, 173
monomial, 40
Monte Carlo algorithm, 191
moralization, 288
most recent common ancestor, 442
moves, 190, 211
multigraded Hilbert series, 279
multiplicity, 407
multivariate normal distribution, 29
mutual independence, 17, 19
natural parameter space, 116
neighbor joining, 446, 448
neighbors, 284
network connectivity, 276, 278, 282
Newton polyhedron, 272, 425
no 3-way interaction, 157, 201, 394
Markov basis, 209, 210
nodal cubic, 128, 142, 161
Noetherian ring, 67, 78
non-identifiable, 364
generically, 364
nondescendants, 288
nonparametric rank, 311
nonparametric statistical model, 98
nonsingular point, 145
normal distribution, 15
normal form, 272, 425
normal semigroup, 172, 244
normalized subgraph count statistic, 257
nuisance parameter, 189
null hypothesis, 108
Nullstellensatz, 44
occasionally dishonest casino, 419, 423, 429
output variables, 383
pairwise marginal independence, 88
parameter, 102, 366
parametric directed graphical model, 296
parametric inference, 415, 424
parametric sets, 41
parametric statistical model, 98
parametrized undirected graphical model, 291
parents, 287
partition lattice, 446
path, 284
Pearson’s $X^2$ statistic, 110
penalty function, 393
pentad, 318
perfect elimination ordering, 243
permutation test, 108
Perron-Frobenius theorem, 338
phylogenetic invariants, 330
phylogenetic model, 329
identifiability, 371
model selection, 394
phylogenetics and metrics, 439
planar graph, 181
plug-in estimator, 102
pointed cone, 168
pointwise convergence, 32
polyhedral cell complex, 271
polyhedral cone, 167
polyhedron, 126, 166
polynomial, 40
polynomial regression models, 263
polytope, 166
polytope algebra, 427
polytope of subgraph statistics, 257
polytope propagation, 429
positive definite, 76
positive semidefinite, 26
posterior distribution, 112
posterior probability, 398
potential function, 201, 291
power set, 12
precision matrix, 118
primary decomposition, 79
and graphical models, 302
application to random walks, 216
primary ideal, 78
prime ideal, 77
prior distribution, 112
probability measure, 12, 14
probability simplex, 42, 166
as semialgebraic set, 126
probability space, 14
projective space, 64
projective variety, 64
pure difference binomial ideal, 81
quartet, 441
quotient ideal, 80, 137
quotient ring, 233
R, 109
radical ideal, 44, 303
random variable, 17
random vector, 18
rank one tensor, 377
Rasch model, 252, 259
rate matrix, 336
rational map, 42
real algebraic variety, 126
real log-canonical threshold, 403
of monomial map, 407
real radical ideal, 67
recession cone, 168
recursive factorization theorem, 291, 296, 395
reduced Gr{"o}bner basis, 68
reducible decomposition, 177
reducible graph, 177
reducible model
Markov basis, 208
reducible simplicial complex, 204
reducible variety, 77, 145
regression coefficients, 30
regular exponential family, 116
regular point, 145
resolution of singularities, 408
reverse lexicographic order, 48, 232
ring, 40
ring of polynomials, 40
rooted tree, 331
sample space, 12
saturated
conditional independence statement, 87, 125
lattice, 82
polynomial regression, 274
semigroup, 172
saturation, 80, 137
schizophrenic hospital patients, 187, 189
score equations, 105, 136
secant variety, 310, 381, 382
Segre variety, 313, 381, 382
semialgebraic set, 126
semimetric, 431
seminiring, 422, 427
separated, 284
separator, 204
sequential importance sampling, 223, 253
shuttle algorithm, 244
siblings, 376
simplex, 166
simplicial complex, 200, 276
singular, 145
Singular (software), 57
singular locus, 145
sink component, 230
SIR model, 390
slim tables, 210
smooth, 127, 145, 161
spatial statistics, 101
spine, 258
split, 331
valid, 331
splits equivalence theorem, 332
squarefree monomial ideal, 235, 244
standard deviation, 24
standard monomial, 234, 369
standard normal distribution, 15
Stanley-Reisner ideal, 277
state polytope, 274
state space model, 382
state variables, 382
stationary distribution, 338
statistic, 99
statistical model, 98
stochastic block model, 248, 250
strand symmetric model, 358
strictly concave function, 151
Strong law of large numbers, 34
strong maximum, 312, 328
strongly connected, 230, 387
strongly connected component, 230
structural equation model, 319, 320
global identifiability, 372
identifiability, 371
structural zeroes, 215
subgraph count statistic, 257
subgraph density, 257
sufficient statistic, 7, 99
supermodular, 312
suspension, 438
symmetric difference, 434
system inputs, 382
system reliability, 276
t-t separates, 300
tableau notation, 205
tangent cone, 127, 160
tangent vector, 127
Tarski–Seidenberg theorem, 127
tensor, 377
tensor rank, 377
term order, 48
tie breaker, 231
time series, 101
topological ordering, 296
toric fiber product, 209
toric ideal, 82, 121
toric model, 120
toric variety, 120, 352
trace trick, 106
transition
  phylogenetics, 337
  transition matrix, 419
transversion, 337, 339
tree, 330
tree cumulants, 357
tree metric, 439
trek, 300, 321
trek rule, 321
triangle inequality, 431
tropical geometry, 443
tropical Grassmannian, 444
tropical semiring, 422
true model, 392
twisted cubic, 41, 122
ultrametric, 442
UPGMA, 446
upstream variables, 317
urn scheme, 217, 222
Vandermonde’s identity, 21
vanishing ideal, 43
variance, 24
variety, 40
vector of counts, 186
vertex, 166
very affine variety, 148
Viterbi algorithm, 420, 421, 423, 429
Weak law of large numbers, 33
weak union axiom, 71
weight order, 48, 231, 270
weighted linear aberration, 275
Zariski closed, 45
Zariski closure, 45
zero set, 40
zeta function, 403