An Invitation to Algebraic Statistics

Bernd Sturmfels
UC Berkeley

2008-09 SAMSI Program on
Algebraic Methods in Systems Biology and Statistics

Tutorial at the Opening Workshop
September 14, 2008

What is a Statistical Model ?

Wiki: *In mathematical terms, a statistical model is frequently thought of as a parameterized set of probability distributions of the form* $\{P_\theta \mid \theta \in \Theta\}$.

What is a Statistical Model ?

Wiki: *In mathematical terms, a statistical model is frequently thought of as a parameterized set of probability distributions of the form* $\{P_\theta \mid \theta \in \Theta\}$.

Planetmath.org: *A statistical model is usually parameterized by a function, called a parameterization*

$$\Theta \to \mathcal{P} \quad \textit{given by} \quad \theta \mapsto P_\theta \quad \textit{so that} \quad \mathcal{P} = \{P_\theta \mid \theta \in \Theta\}.$$

where Θ *is called a parameter space.* Θ *is usually a subset of* $\mathbb{R}^n$.

What is a Statistical Model ?

Wiki: *In mathematical terms, a statistical model is frequently thought of as a parameterized set of probability distributions of the form* $\{P_\theta \,|\, \theta \in \Theta\}$.

Planetmath.org: *A statistical model is usually parameterized by a function, called a parameterization*

$$\Theta \to \mathcal{P} \quad \textit{given by} \quad \theta \mapsto P_\theta \quad \textit{so that} \quad \mathcal{P} = \{P_\theta \,|\, \theta \in \Theta\}.$$

where Θ *is called a parameter space.* Θ *is usually a subset of* $\mathbb{R}^n$.

McCullagh, 2002: This should be defined using Category Theory.

What is a Statistical Model ?

Wiki: *In mathematical terms, a statistical model is frequently thought of as a parameterized set of probability distributions of the form* $\{P_\theta \,|\, \theta \in \Theta\}$.

Planetmath.org: *A statistical model is usually parameterized by a function, called a parameterization*

$$\Theta \to \mathcal{P} \quad \textit{given by} \quad \theta \mapsto P_\theta \quad \textit{so that} \quad \mathcal{P} = \{P_\theta \,|\, \theta \in \Theta\}.$$

where Θ *is called a parameter space.* Θ *is usually a subset of* $\mathbb{R}^n$.

McCullagh, 2002: This should be defined using Category Theory.

Today: Consider discrete data and suppose that the parameter space Θ and the function $\theta \mapsto P_\theta$ are described by polynomials.

Tomorrow: This makes sense also for Gaussian models.

Three-Way Contingency Tables

Let X, Y and Z be random variables that have a, b and c states respectively. A *probability distribution* P for these random variables is an $a \times b \times c$-table of non-negative real numbers that sum to one.

The entries of the table P are the probabilities

$$P_{ijk} \;=\; \mathrm{Prob}(X = i, Y = j, Z = k).$$

The set of all distributions is a simplex Δ of dimension $abc - 1$.

Three-Way Contingency Tables

Let X, Y and Z be random variables that have a, b and c states respectively. A *probability distribution* P for these random variables is an $a \times b \times c$-table of non-negative real numbers that sum to one.

The entries of the table P are the probabilities

$$P_{ijk} \;=\; \mathrm{Prob}(X = i, Y = j, Z = k).$$

The set of all distributions is a simplex Δ of dimension $abc - 1$.

A *statistical model* is a subset $\mathcal{M}$ of Δ which can be described by polynomial equations and inequalities in the coordinates P_{ijk}.

Typically, the model $\mathcal{M}$ is presented as the image of a polynomial map $P : \Theta \mapsto \Delta$ where Θ is a polynomially described subset of $\mathbb{R}^n$.

Independence

The distribution P is called *independent* if each probability is the product of the corresponding marginal probabilities:

$$P_{ijk} \ = \ P_{i++} \cdot P_{+j+} \cdot P_{++k}$$

Here, for instance,

$$P_{i++} \ = \mathrm{Prob}(X = i) \ = \ \sum_{j=1}^{b} \sum_{k=1}^{c} P_{ijk}$$

Independence

The distribution P is called *independent* if each probability is the product of the corresponding marginal probabilities:

$$P_{ijk} \; = \; P_{i++} \cdot P_{+j+} \cdot P_{++k}$$

Here, for instance,

$$P_{i++} \; = \mathrm{Prob}(X = i) \; = \; \sum_{j=1}^{b} \sum_{k=1}^{c} P_{ijk}$$

The independence model has the parametric representation

$$\begin{array}{rcl} \Theta \; = \; \Delta_{a-1} \times \Delta_{b-1} \times \Delta_{c-1} & \rightarrow & \Delta \; = \; \Delta_{abc-1} \\ (\alpha, \beta, \gamma) & \mapsto & (P_{ijk}) \; = \; (\alpha_i \beta_j \gamma_k) \end{array}$$

Independence

The distribution P is called *independent* if each probability is the product of the corresponding marginal probabilities:

$$P_{ijk} \;=\; P_{i++} \cdot P_{+j+} \cdot P_{++k}$$

Here, for instance,

$$P_{i++} \;= \mathrm{Prob}(X = i) \;=\; \sum_{j=1}^{b} \sum_{k=1}^{c} P_{ijk}$$

The independence model has the parametric representation

$$\begin{array}{rcl} \Theta \;=\; \Delta_{a-1} \times \Delta_{b-1} \times \Delta_{c-1} & \rightarrow & \Delta \;=\; \Delta_{abc-1} \\ (\alpha, \beta, \gamma) & \mapsto & (P_{ijk}) \;=\; (\alpha_i \beta_j \gamma_k) \end{array}$$

The image is known as the Segre variety in algebraic geometry. Its points are the $a \times b \times c$-tables of tensor rank one.

Three Binary Variables

If $a = b = c = 2$ then the independence model (Segre variety) is the threefold in Δ_7 (or in $\mathbb{P}^7$) which has the parametrization:

$$
\begin{aligned}
P_{000} &= \alpha\beta\gamma & P_{001} &= \alpha\beta(1-\gamma) \\
P_{010} &= \alpha(1-\beta)\gamma & P_{011} &= \alpha(1-\beta)(1-\gamma) \\
P_{100} &= (1-\alpha)\beta\gamma & P_{101} &= (1-\alpha)\beta(1-\gamma) \\
P_{110} &= (1-\alpha)(1-\beta)\gamma & P_{111} &= (1-\alpha)(1-\beta)(1-\gamma)
\end{aligned}
$$

Three Binary Variables

If $a = b = c = 2$ then the independence model (Segre variety) is the threefold in Δ_7 (or in $\mathbb{P}^7$) which has the parametrization:

$$\begin{aligned} P_{000} &= \alpha\beta\gamma & P_{001} &= \alpha\beta(1-\gamma) \\ P_{010} &= \alpha(1-\beta)\gamma & P_{011} &= \alpha(1-\beta)(1-\gamma) \\ P_{100} &= (1-\alpha)\beta\gamma & P_{101} &= (1-\alpha)\beta(1-\gamma) \\ P_{110} &= (1-\alpha)(1-\beta)\gamma & P_{111} &= (1-\alpha)(1-\beta)(1-\gamma) \end{aligned}$$

This threefold is cut out by the trivial constraint

$$P_{000} + P_{001} + P_{010} + P_{011} + P_{100} + P_{101} + P_{110} + P_{111} = 1$$

and the Markov basis which consists of nine *quadratic binomials*:

$$\begin{array}{lll} P_{100}P_{111} - P_{101}P_{110}, & P_{010}P_{111} - P_{011}P_{110}, & P_{010}P_{101} - P_{011}P_{100}, \\ P_{001}P_{111} - P_{011}P_{101}, & P_{001}P_{110} - P_{011}P_{100}, & P_{000}P_{111} - P_{011}P_{100}, \\ P_{000}P_{110} - P_{010}P_{100}, & P_{000}P_{101} - P_{001}P_{100}, & P_{000}P_{011} - P_{001}P_{010}. \end{array}$$

Markov bases

- make sense for every exponential family (log-linear model)
- are interesting for graphical models and hierarchical models
- minimally generate the corresponding toric ideal
- give Markov chains for sampling from conditional distributions
- can be computed in practise using the software `4ti2`

Markov bases

- make sense for every exponential family (log-linear model)
- are interesting for graphical models and hierarchical models
- minimally generate the corresponding toric ideal
- give Markov chains for sampling from conditional distributions
- can be computed in practise using the software `4ti2`

Theorem

The Markov basis for the independence model on three random variables consists of quadratic binomials $P_{\bullet\bullet\bullet}P_{\bullet\bullet\bullet} - P_{\bullet\bullet\bullet}P_{\bullet\bullet\bullet}$.
The number of binomials in this Markov basis equals

$$\frac{1}{8}abc\,(3abc - ab - ac - bc - a - b - c + 3).$$

Use representation theory to figure this out
and to compactly encode the Markov basis.

Mixtures

A distribution P is a mixture of independent distributions if

$$P = \lambda P' + (1 - \lambda) P''$$

where P' and P'' are independent and $0 \leq \lambda \leq 1$. The set of such mixtures is the *first mixture model* of the independence model.

Mixtures

A distribution P is a mixture of independent distributions if

$$P = \lambda P' + (1-\lambda)P''$$

where P' and P'' are independent and $0 \leq \lambda \leq 1$. The set of such mixtures is the *first mixture model* of the independence model.

Thus the first mixture model is the image of the parametrization

$$\begin{array}{ccc} (\Delta_{a-1} \times \Delta_{b-1} \times \Delta_{c-1})^2 \times \Delta_1 & \to & \Delta_{abc-1} \\ (\alpha', \beta', \gamma'; \alpha'', \beta'', \gamma''; \lambda) & \mapsto & \left(\lambda \alpha'_i \beta'_j \gamma'_k + (1-\lambda)\alpha''_i \beta''_j \gamma''_k\right) \end{array}$$

The first mixture model is identifiable, because the corresponding algebraic variety has the expected dimension $2a + 2b + 2c - 5$.

Secants and rank two tensors

In algebraic geometry, mixtures correspond to secant lines, and the first mixture model is known as the first secant variety of the Segre variety. Its points are the $a \times b \times c$-tables of tensor rank two.

Secants and rank two tensors

In algebraic geometry, mixtures correspond to secant lines, and the first mixture model is known as the first secant variety of the Segre variety. Its points are the $a \times b \times c$-tables of tensor rank two.

Theorem

The homogeneous prime ideal of the first mixture model is generated by cubic polynomials in the probabilities P_{ijk}.

Secants and rank two tensors

In algebraic geometry, mixtures correspond to secant lines, and the first mixture model is known as the first secant variety of the Segre variety. Its points are the $a \times b \times c$-tables of tensor rank two.

Theorem

The homogeneous prime ideal of the first mixture model is generated by cubic polynomials in the probabilities P_{ijk}.

These cubic generators are the 3×3-subdeterminants of the three matrices, of formats $(ab) \times c$, $(ac) \times b$ and $(bc) \times a$, which arise from flattening the three-dimensional table P.

This result was conjectured by [Garcia-Stillman-St 2005] and proved by [Landsberg-Manivel 2004]. A very general phylogenetic version appears in [Draisma-Kuttler 2008].

Secants and rank two tensors

In algebraic geometry, mixtures correspond to secant lines, and the first mixture model is known as the first secant variety of the Segre variety. Its points are the $a \times b \times c$-tables of tensor rank two.

Theorem

The homogeneous prime ideal of the first mixture model is generated by cubic polynomials in the probabilities P_{ijk}.

These cubic generators are the 3×3-subdeterminants of the three matrices, of formats $(ab) \times c$, $(ac) \times b$ and $(bc) \times a$, which arise from flattening the three-dimensional table P.

This result was conjectured by [Garcia-Stillman-St 2005] and proved by [Landsberg-Manivel 2004]. A very general phylogenetic version appears in [Draisma-Kuttler 2008].

Further progress on rank 4 tensors might earn you smoked salmon.

Flattening a $3 \times 2 \times 2$-table

Suppose we are given one ternary variable and two binary variables, that is, $a = 3$ and $b = c = 2$. The Landsberg-Manivel Theorem states that the first mixture model is characterized algebraically by the vanishing of the 3×3-minors of the 3×4-matrix

$$P_{\mathrm{flat}} \quad = \quad \begin{pmatrix} P_{000} & P_{001} & P_{010} & P_{011} \\ P_{100} & P_{101} & P_{110} & P_{111} \\ P_{200} & P_{201} & P_{210} & P_{211} \end{pmatrix}.$$

This matrix has rank at most two for P in the first mixture model.

Flattening a $3 \times 2 \times 2$-table

Suppose we are given one ternary variable and two binary variables, that is, $a = 3$ and $b = c = 2$. The Landsberg-Manivel Theorem states that the first mixture model is characterized algebraically by the vanishing of the 3×3-minors of the 3×4-matrix

$$P_{\text{flat}} \quad = \quad \begin{pmatrix} P_{000} & P_{001} & P_{010} & P_{011} \\ P_{100} & P_{101} & P_{110} & P_{111} \\ P_{200} & P_{201} & P_{210} & P_{211} \end{pmatrix}.$$

This matrix has rank at most two for P in the first mixture model.

Application to likelihood inference: This model has maximum likelihood degree 26. Maximizing a monomial $\prod P_{ijk}^{U_{ijk}}$ over this model reduces to solving an algebraic equation of degree 26.

Flattening a $3 \times 2 \times 2$-table

Suppose we are given one ternary variable and two binary variables, that is, $a = 3$ and $b = c = 2$. The Landsberg-Manivel Theorem states that the first mixture model is characterized algebraically by the vanishing of the 3×3-minors of the 3×4-matrix

$$P_{\text{flat}} \quad = \quad \begin{pmatrix} P_{000} & P_{001} & P_{010} & P_{011} \\ P_{100} & P_{101} & P_{110} & P_{111} \\ P_{200} & P_{201} & P_{210} & P_{211} \end{pmatrix}.$$

This matrix has rank at most two for P in the first mixture model.

Application to likelihood inference: This model has maximum likelihood degree 26. Maximizing a monomial $\prod P_{ijk}^{U_{ijk}}$ over this model reduces to solving an algebraic equation of degree 26.

The analogous computation for the variety of 4×4-matrices having rank ≤ 2 is an open problem that might earn you 100 Swiss Francs.

Bayesian inference

Given a table of data $U = (U_{ijk}) \in \mathbb{N}^{a \times b \times c}$, a central problem in Bayesian statistics is to compute the marginal likelihood integral

$$\int \prod_{i,j,k} \left(\lambda \alpha'_i \beta'_j \gamma'_k + (1-\lambda) \alpha''_i \beta''_j \gamma''_k\right)^{U_{ijk}} d\alpha \, d\beta \, d\gamma \, d\lambda$$

This integral is over the $(2a+2b+2c-5)$-dimensional polytope

$$(\Delta_{a-1} \times \Delta_{b-1} \times \Delta_{c-1})^2 \times \Delta_1$$

with respect to a probability distribution representing prior belief.

Algebraic statistics has tools for exact integration when the sample size $|U|$ is small, and for asymptotic analysis when $|U| \to \infty$.

Exact integration

Proposition (Lin-St-Xu 2008)

For uniform priors, the value of the marginal likelihood integral is a rational number. For Dirichlet priors, it is a product of special values of the Gamma function. $\longrightarrow$ *software in* `maple`

Exact integration

Proposition (Lin-St-Xu 2008)

For uniform priors, the value of the marginal likelihood integral is a rational number. For Dirichlet priors, it is a product of special values of the Gamma function. $\longrightarrow$ *software in* `maple`

Example: Consider the following $3 \times 2 \times 2$-table of data

$$U_{\text{flat}} = \begin{pmatrix} 2 & 3 & 1 & 1 \\ 2 & 1 & 3 & 1 \\ 2 & 1 & 1 & 3 \end{pmatrix}$$

Exact integration

Proposition (Lin-St-Xu 2008)

For uniform priors, the value of the marginal likelihood integral is a rational number. For Dirichlet priors, it is a product of special values of the Gamma function. $\longrightarrow$ *software in* `maple`

Example: Consider the following $3 \times 2 \times 2$-table of data

$$U_{\mathrm{flat}} = \begin{pmatrix} 2 & 3 & 1 & 1 \\ 2 & 1 & 3 & 1 \\ 2 & 1 & 1 & 3 \end{pmatrix}$$

The marginal likelihood of these data in the mixture model equals

$$\binom{|U|}{U} \cdot \int P^U dP = \frac{1000990472851655999396215 1}{958019384093441508386090262720000}$$

Here the prior on the 9-dimensional parameter polytope is uniform.

Higher mixture models

In the *r-th mixture model* we are mixing r independent distributions, so the model consists of tensors of rank r:

$$P = \mathbf{a}_1 \otimes \mathbf{b}_1 \otimes \mathbf{c}_1 + \mathbf{a}_2 \otimes \mathbf{b}_2 \otimes \mathbf{c}_2 + \cdots + \mathbf{a}_r \otimes \mathbf{b}_r \otimes \mathbf{c}_r.$$

Higher mixture models

In the *r-th mixture model* we are mixing r independent distributions, so the model consists of tensors of rank r:

$$P = \mathbf{a}_1 \otimes \mathbf{b}_1 \otimes \mathbf{c}_1 + \mathbf{a}_2 \otimes \mathbf{b}_2 \otimes \mathbf{c}_2 + \cdots + \mathbf{a}_r \otimes \mathbf{b}_r \otimes \mathbf{c}_r.$$

We consider the following class of submodels.

A *context-specific independence model* is specified by three partitions $\mathcal{A}, \mathcal{B}, \mathcal{C}$ of $\{1, \ldots, r\}$. These partitions describe how the parameters are tied together:

- $\mathbf{a}_i = \mathbf{a}_j$ if i and j are in the same block in $\mathcal{A}$,
- $\mathbf{b}_i = \mathbf{b}_j$ if i and j are in the same block in $\mathcal{B}$,
- $\mathbf{c}_i = \mathbf{c}_j$ if i and j are in the same block in $\mathcal{C}$.

[B. Georgi and A. Schliep: *Context-specific independence mixture modeling for positional weight matrices*, Bioinformatics, 2006]

Context-specific independence

Let $r = 3$ and fix the three partitions
$\mathcal{A} = \{\{1,2\},\{3\}\}$, $\mathcal{B} = \{\{1,3\},\{2\}\}$, and $\mathcal{C} = \{\{2,3\},\{1\}\}$.

This CSI model has the parametric representation

$$P_{ijk} \;=\; \lambda \cdot \alpha_i \beta_j \phi_k \;+\; \mu \cdot \alpha_i \epsilon_j \gamma_k \;+\; (1-\lambda-\mu) \cdot \delta_i \beta_j \gamma_k$$

Equivalently, in tensor notation:

$$P \;=\; \mathbf{a} \otimes \mathbf{b} \otimes \mathbf{f} + \mathbf{a} \otimes \mathbf{e} \otimes \mathbf{c} + \mathbf{d} \otimes \mathbf{b} \otimes \mathbf{c}$$

Context-specific independence

Let $r = 3$ and fix the three partitions
$\mathcal{A} = \{\{1,2\},\{3\}\}$, $\mathcal{B} = \{\{1,3\},\{2\}\}$, and $\mathcal{C} = \{\{2,3\},\{1\}\}$.

This CSI model has the parametric representation

$$P_{ijk} \;=\; \lambda \cdot \alpha_i \beta_j \phi_k \,+\, \mu \cdot \alpha_i \epsilon_j \gamma_k \,+\, (1-\lambda-\mu) \cdot \delta_i \beta_j \gamma_k$$

Equivalently, in tensor notation:

$$P \;=\; \mathbf{a} \otimes \mathbf{b} \otimes \mathbf{f} + \mathbf{a} \otimes \mathbf{e} \otimes \mathbf{c} + \mathbf{d} \otimes \mathbf{b} \otimes \mathbf{c}$$

Theorem

The Zariski closure of this CSI model is the tangential variety of the Segre variety. Its homogeneous prime ideal is generated by all $2 \times 2 \times 2$-hyperdeterminants in the $a \times b \times c$-table P together with all 3×3-determinants obtained by flattening P.

A small example

Let $a = 3, b = 2, c = 2$ and fix the CSI model specified by $\mathcal{A} = \{\{1,2\},\{3\}\}$, $\mathcal{B} = \{\{1,3\},\{2\}\}$ and $\mathcal{C} = \{\{2,3\},\{1\}\}$.

This model lies in Δ_{11}. It has dimension 8 and degree 16.
It is not identifiable because there are 10 natural parameters.

A small example

Let $a = 3, b = 2, c = 2$ and fix the CSI model specified by $\mathcal{A} = \{\{1,2\},\{3\}\}$, $\mathcal{B} = \{\{1,3\},\{2\}\}$ and $\mathcal{C} = \{\{2,3\},\{1\}\}$.

This model lies in Δ_{11}. It has dimension 8 and degree 16.
It is not identifiable because there are 10 natural parameters.

Its ideal is generated by the four 3×3-subdeterminants of

$$P_{\text{flat}} \quad = \quad \begin{pmatrix} P_{000} & P_{001} & P_{010} & P_{011} \\ P_{100} & P_{101} & P_{110} & P_{111} \\ P_{200} & P_{201} & P_{210} & P_{211} \end{pmatrix}.$$

and six $2 \times 2 \times 2$-hyperdeterminants, such as

$$\begin{gathered} p_{000}^2 p_{111}^2 + p_{010}^2 p_{101}^2 + p_{011}^2 p_{100}^2 + p_{001}^2 p_{110}^2 \\ -2p_{010}p_{011}p_{100}p_{101} - 2p_{001}p_{011}p_{100}p_{110} - 2p_{001}p_{010}p_{101}p_{110} \\ -2p_{000}p_{011}p_{100}p_{111} - 2p_{000}p_{010}p_{101}p_{111} - 2p_{000}p_{001}p_{110}p_{111} \\ +4p_{000}p_{011}p_{101}p_{110} + 4p_{001}p_{010}p_{100}p_{111}. \end{gathered}$$

Conclusion

Algebraic Statistics is both cool and useful.

Conclusion

Algebraic Statistics is both cool and useful.

For further reading see, e.g.,

[M. Drton, B. Sturmfels and S. Sullivant:
Lectures on Algebraic Statistics, Oberwolfach Seminars Series,
Vol. 40, Approx. 175 p., Softcover, Birkhäuser, Basel, 2009]