Algebraic Statistics builds on the idea that statistical models can be understood via polynomials. Many statistical models are parametrized by polynomials in the model parameters; others are described implicitly by polynomial equalities and inequalities. We explore this connection for some small statistical models.

1 Introduction

From the 17th to the 22nd of April 2017, a Workshop on Algebraic Statistics took place at the Mathematisches Forschungsinstitut Oberwolfach (MFO). The weather started off cold with intermittent rain showers. The middle of the week saw snow and hail, and there were two sunny days at the end. Many of the 52 workshop participants were visiting Oberwolfach for the first time, and some of them played a game during their stay (the favorite games were Carom Billiards, Hanabi, and Resistance).

The following contingency table summarizes data we obtained from a survey of the participants. Fifty out of the 52 participants responded to the survey. The following table shows the observed counts and empirical probabilities:

<table>
<thead>
<tr>
<th>Games</th>
<th>First time at MFO</th>
<th>Been to MFO before</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liked weather</td>
<td>12 (24%)</td>
<td>5 (10%)</td>
</tr>
<tr>
<td>Disliked weather</td>
<td>5 (10%)</td>
<td>4 (8%)</td>
</tr>
<tr>
<td>No games</td>
<td>Liked weather</td>
<td>7 (14%)</td>
</tr>
<tr>
<td></td>
<td>Disliked weather</td>
<td>3 (6%)</td>
</tr>
</tbody>
</table>
For instance, a workshop participant, selected at random, was at MFO for the first time, played games and enjoyed the weather with empirical probability 0.24. The eight probabilities from the table are the joint probability distribution of three binary random variables. We represent this by the $2 \times 2 \times 2$ tensor
\[
p = \begin{bmatrix}
p_{000} & p_{010} \\
p_{100} & p_{110}
\end{bmatrix}, \quad \begin{bmatrix}
p_{001} & p_{011} \\
p_{101} & p_{111}
\end{bmatrix} = \begin{bmatrix}
0.24 & 0.1 \\
0.1 & 0.08
\end{bmatrix}, \quad \begin{bmatrix}
0.14 & 0.18 \\
0.06 & 0.1
\end{bmatrix}.
\]
The eight probabilities $p_{ijk}$ represent $P(X = i, Y = j, Z = k)$ for random variables $X$, $Y$ and $Z$. In the survey, we have three indicator random variables: ‘disliking the weather’, ‘having visited Oberwolfach before’, ‘playing no game’.

So-called marginal probabilities are obtained by ignoring (summing over) all states with some fixed indices. For example, a random participant liked the weather with probability $0.24 + 0.1 + 0.14 + 0.18 = 0.66$.

A statistical model is a collection of probability distributions that share some structure. In this article, we explore statistical models that can explain the survey results. We focus on the algebraic structure of these models.

2 Independence Models

Suppose that $X$ and $Y$ are binary variables as in our survey. Their joint probability distribution is a point in the three-dimensional simplex
\[
\Delta_3 = \{ p \in \mathbb{R}^{2 \times 2} : \sum_{ij} p_{ij} = 1, \ p_{ij} \geq 0 \},
\]
where the $p_{ij}$ represent $P(X = i, Y = j)$.

Two random variables $X$ and $Y$ are independent if the distribution of one does not change under knowledge of the other. The independence model is defined by the single quadratic equation
\[
\det \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} = p_{00}p_{11} - p_{01}p_{10} = 0.
\]  

(1)
The equation tests if the matrix of joint probabilities has rank one.

We use the independence model to study the survey data. Ignoring the games question gives the following marginal table:

<table>
<thead>
<tr>
<th></th>
<th>First time at MFO</th>
<th>Been to MFO before</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liked weather</td>
<td>19 (38%)</td>
<td>14 (28%)</td>
</tr>
<tr>
<td>Disliked weather</td>
<td>8 (16%)</td>
<td>9 (18%)</td>
</tr>
</tbody>
</table>

First time visitors to MFO were more appreciative of the weather than those who had visited the institute before. Hence the empirical distribution from the
random variables ‘liking the weather’ and ‘having visited MFO before’ are not independent. Algebraically,

$$\det \begin{bmatrix} 0.38 & 0.28 \\ 0.16 & 0.18 \end{bmatrix} = 0.0236 \neq 0.$$  

It is not surprising that the real-world data from the survey does not exactly satisfy the equation (1). The independence model occupies a volume 0% of all probability densities: it is a surface in a tetrahedron. As we explain next, 0.0236 is ‘quite close’ to 0, indicating that our table lies close to the independence model.

For most applications, the gathered data is a sample of an underlying population. We use the sample to infer properties of the population. In our survey, the population could be ‘all past, present and future MFO visitors’. Fisher’s exact test is a technique that quantifies whether the distance of the data from the model is statistically significant. Following [2, Proposition 1.1.8], we compute the probability of observing our data, under the assumption of independence and conditional on the row and column sums of the table:

$$\frac{\binom{33}{19} \binom{17}{8}}{\binom{50}{27}} = 0.1842341.$$
The probability of deviating at least as much from the statistical model as our data is called a $p$-value. It is found by adding 0.1842341 to the probability of observing all more extreme tables of hypothetical data. The $p$-value is computed to be 0.556. This exceeds the standard significance level of 5%, which means we do not reject the null hypothesis that the two variables are independent. A similar conclusion would be obtained via the asymptotics of Pearson’s $\chi^2$ test.

Now we consider probability distributions of three binary random variables $X$, $Y$ and $Z$. We shall examine various independence models on three random variables: full independence, marginal independence and conditional independence.

![Figure 2: The eight joint probabilities of three binary random variables](image)

2.1 Full Independence

The statistical model for the full independence model consists of $2 \times 2 \times 2$ tensors of rank one. This threefold is the Segre variety $\text{Seg}(\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1) \subset \mathbb{P}^7$, restricted to the 7-dimensional probability simplex. Its defining equations are

\begin{align}
 p_{000}p_{011} &= p_{010}p_{001}, \\
 p_{000}p_{101} &= p_{100}p_{001}, \\
 p_{000}p_{110} &= p_{100}p_{101}, \\
 p_{001}p_{111} &= p_{101}p_{011}, \\
 p_{010}p_{111} &= p_{110}p_{011}, \\
 p_{100}p_{111} &= p_{110}p_{101}, \\
 p_{000}p_{111} &= p_{101}p_{011}, \\
 p_{000}p_{111} &= p_{011}p_{100}.
\end{align}

The first six equations are rank one conditions on facets of the cube in Figure 2. The last three are rank one conditions on the main diagonals of the cube. We might think that the facet-independences are sufficient for full independence, but these six equations leave us with four extraneous components. These can be explored using the computer algebra software Macaulay2:

```latex
R = QQ[p_{000},p_{001},p_{010},p_{011},p_{100},p_{101},p_{110},p_{111}];
I = ideal(p_{000}*p_{011}-p_{010}*p_{001},p_{000}*p_{101}-p_{100}*p_{001},
p_{000}*p{110}-p_{100}*p_{010},
p_{001}*p_{111}-p_{101}*p_{011},p_{010}*p_{111}-p_{110}*p_{111},p_{100}*p_{111}-p_{110}*p_{101});
decompose I
```

The nine equations in (2) are not satisfied for the survey data: the three questions in the survey are not independent. As before, the model occupies
a volume 0% of distributions. We wish to assess if the data is significantly far from the independence model. We use the algebraic techniques of Markov Bases to generalize Fisher’s exact test, as explained in the Oberwolfach seminar notes [2, §1.2]. This can be done using the package \texttt{algstat} in the statistical programming language \texttt{R}. We get a $p$-value of 0.492 when measuring the significance of tables exactly, and 0.515 using asymptotics via the \chi^2 statistic.

The point on the model with the highest probability of producing the observed data is the maximum likelihood estimate (MLE). For the survey data, the MLE is

$$p = \begin{bmatrix} 0.185 & 0.158 \\ 0.095 & 0.081 \end{bmatrix}, \quad \begin{bmatrix} 0.171 & 0.146 \\ 0.088 & 0.075 \end{bmatrix} = \begin{bmatrix} 0.66 \\ 0.34 \end{bmatrix} \otimes \begin{bmatrix} 0.54 \\ 0.46 \end{bmatrix} \otimes \begin{bmatrix} 0.52 \\ 0.48 \end{bmatrix}.$$  

### 2.2 Conditional and Marginal Independence

Conditional and marginal independence are weaker forms of independence than the full independence model. Here we show the equations that describe them.

Conditional independence is the independence of two random variables, conditional on the value taken by the third. For example, conditional independence of $Y$ and $Z$ given $X$, denoted $(Y \mid| Z)\mid X$, is defined by the two equations

$$p_{000}p_{011} = p_{001}p_{010} \quad \text{and} \quad p_{100}p_{111} = p_{101}p_{110}. \quad (3)$$

Marginal independence concerns independence of two random variables after summing over the possible values taken by a third. For example, consider the marginal independence of $Y$ and $Z$ after summing over the values taken by $X$. This model is denoted $Y \mid \parallel Z$. It is defined by the polynomial equation

$$(p_{000} + p_{100})(p_{011} + p_{111}) = (p_{001} + p_{101})(p_{010} + p_{110}).$$

The interplay between conditional and marginal (in)dependence can be subtle, but algebra helps us understand what is going on. A good example is Simpson’s paradox. The apparent contradiction arises because the correlation between two variables, after \textit{marginalizing} by a third, can have a different sign than the correlation between two variables after \textit{conditioning} on a third. It highlights the importance of knowing the hidden factors influencing observations.

### 3 Latent Variable Models

We next illustrate how hidden effects may confound the relations among observed variables in the context of our survey data.

The Naïve Bayes Model consists of some observed variables, and a single hidden variable. Its joint distribution is a convex combination of independent distributions.
We seek to explain the survey data using a hidden variable. Suppose that the hidden variable is also binary. In fact, our survey had two additional questions: Do you consider yourself a "young person"? and Do you consider yourself more a "maths person" or a "stats person"?

Our $2 \times 2 \times 2$ data table kept these two binary variables hidden. The breakdown of the responses was as follows: 56% identified as ‘young people’, while 44% did not. The subject affiliations were 70% mathematics and 30% statistics. We shall not reveal the complete $2 \times 2 \times 2 \times 2 \times 2$ table of all responses.

The joint probability distributions in the Naïve Bayes Model are, up to a normalizing constant, the $2 \times 2 \times 2$ tensors of non-negative rank two:

$$ p = a_0 \otimes b_0 \otimes c_0 + a_1 \otimes b_1 \otimes c_1 \quad \text{where} \quad a_i, b_i, c_i \in \mathbb{R}^2 \text{ are non-negative.} $$

The non-negative rank is two because we marginalize out the single hidden node, which has two states. This statistical model is full-dimensional inside the probability simplex $\Delta_7$. It occupies a volume of approximately 8% of the simplex. That region is given by the polynomial inequalities found in [1]. They are the log-supermodularity conditions

$$
\begin{align*}
    p_{000}p_{011} &\geq p_{010}p_{001}, & p_{000}p_{101} &\geq p_{100}p_{001}, & p_{000}p_{110} &\geq p_{100}p_{010}, \\
    p_{001}p_{111} &\geq p_{101}p_{011}, & p_{010}p_{111} &\geq p_{110}p_{011}, & p_{100}p_{111} &\geq p_{110}p_{101}, \\
    p_{000}p_{111} &\geq p_{101}p_{010}, & p_{000}p_{111} &\geq p_{110}p_{001}, & p_{000}p_{111} &\geq p_{011}p_{100}.
\end{align*}
$$

(4)

Notice these are the same equations as in (2), but with the equalities replaced by inequalities. If all eight probabilities $p_{ijk}$ are strictly positive, it suffices to check the first six inequalities to confirm membership in the model. There are three other regions of $\Delta_7$ in the Naïve Bayes Model. Their inequalities are obtained from those above by swapping 0 and 1 in either of the three indices.

We can check to see if the above inequalities hold for survey data. It turns out that one of the six inequalities in (4) is not quite satisfied:

$$ p_{010}p_{111} - p_{110}p_{011} = -0.0044. $$
Hence the survey data cannot be perfectly explained by a single hidden binary random variable. The MLE for our data in the Naïve Bayes Model equals

\[
\hat{p} = \begin{bmatrix}
0.24 & 0.1096 \\
0.14 & 0.0704 \\
0.06 & 0.1096
\end{bmatrix}.
\]

It is found by adjusting the slice whose determinant was previously negative. The MLE \( \hat{p} \) lies in the model. It is the sum of two non-negative rank one terms:

\[
\hat{p} = \lambda \left[ \begin{array}{c}
\alpha_0 \\
1 - \alpha_0
\end{array} \right] \otimes \left[ \begin{array}{c}
\beta_0 \\
1 - \beta_0
\end{array} \right] \otimes \left[ \begin{array}{c}
\gamma_0 \\
1 - \gamma_0
\end{array} \right] + (1 - \lambda) \left[ \begin{array}{c}
\alpha_1 \\
1 - \alpha_1
\end{array} \right] \otimes \left[ \begin{array}{c}
\beta_1 \\
1 - \beta_1
\end{array} \right] \otimes \left[ \begin{array}{c}
\gamma_1 \\
1 - \gamma_1
\end{array} \right].
\]

The parameters in this maximum likelihood estimate for our data are

\[
\lambda = 0.509155 \\
(\alpha_0, \beta_0, \gamma_0) = (0.709459, 1, 0.644068) \\
(\alpha_1, \beta_1, \gamma_1) = (0.608696, 0.062840, 0.391304).
\]

The rank one terms in the MLE are similar to the data obtained after stratifying by age. This indicates the importance of the ‘young person’ variable for understanding the survey data.

## 4 Towards Deep Learning

Probability distributions on three binary random variables have only a 8% chance of factorizing according to a single hidden binary variable. Multiple hidden random variables, arranged in a layer, make a Restricted Boltzmann Machine (RBM). RBMs serve as building blocks for so-called deep belief networks. These are canonical deep learning models consisting of multiple layers of latent variables and a single layer of observed variables. Direct statistical dependence can only exist between distinct adjacent layers. Parameters of the multi-layer network can be learned greedily layer by layer: each pair of adjacent layers is an RBM.

Algebraically, RBMs correspond to Hadamard (entrywise) products of secant varieties of Segre varieties. Algebraic geometry techniques have been used to establish that the parametrization is generically identifiable. This is a recent result of Montúfar and Morton [3]. Algebraic techniques can also characterize which probability densities can be described by a particular graphical format, as we will see below.

We consider the case of two hidden binary variables. This RBM model consists of all probability distributions that can be factored as the entrywise product of two \( 2 \times 2 \times 2 \) tensors from the Naïve Bayes Model we encountered in Section 3. The graphical representation of the RBM model is as follows:
This statistical model occupies an approximate volume of 76\% of the probability simplex $\Delta_7$. It consists of six regions inside of the simplex, namely, the Hadamard products of any two of the four regions obtained from (4) by label swapping. One of the six pieces is characterized by the two quadratic inequalities

$$p_{000}p_{011} \geq p_{001}p_{010} \quad \text{and} \quad p_{100}p_{111} \geq p_{101}p_{110}. \quad (5)$$

Notice that these are the same equations as (3), with the equalities replaced by inequalities. Note also the similarity between the previous sentence and the one following (4). The five other regions in the RBM model are obtained by reversing the inequalities in (5) or by permuting indices. A cartoon of the Naïve Bayes model contained inside the RBM model is shown in Figure 4.

\textbf{Figure 3:} The four dark blobs are the pieces of the Naïve Bayes model. The striped blobs containing them are the six pieces of the Restricted Boltzmann Machine model.
The two inequalities in (5) are satisfied for our survey data: the data is independent, conditional on two hidden binary random variables. It remains to ponder what the two hidden variables are. They are the unseen factors that influence the survey data. Might young/old and maths/stats play a role?

Acknowledgements

Thanks to Guido Montúfar and Bernd Sturmfels for enjoyable conversations that helped this snapshot. Thanks to Carlos Améndola for help with algstat.

Image credits

Fig. 1. Surface 3D-printed by Thomas Kahle. Photograph by Anna Seigal.

Fig. 3. Made by Otavio Good.

Other figs. Made by Anna Seigal.

References


Anna Seigal is a graduate student in mathematics at the University of California, Berkeley.

Mathematical subjects will be filled out by the editors

Connections to other fields will be filled out by the editors

License will be filled out by the editors

DOI will be filled out by the editors

Snapshots of modern mathematics from Oberwolfach provide exciting insights into current mathematical research. They are written by participants in the scientific program of the Mathematisches Forschungsinstitut Oberwolfach (MFO). The snapshot project is designed to promote the understanding and appreciation of modern mathematics and mathematical research in the interested public worldwide. All snapshots are published in cooperation with the IMAGINARY platform and can be found on www.imaginary.org/snapshots and on www.mfo.de/snapshots.

Junior Editor
will be filled out by the editors
junior-editors@mfo.de

Senior Editor
Carla Cederbaum
senior-editor@mfo.de

Mathematisches Forschungsinstitut Oberwolfach gGmbH
Schwarzwaldstr. 9 –11
77709 Oberwolfach
Germany

Director
Gerhard Huisken

---

IMAGINARY
open mathematics