Snapshots of modern mathematics
from Oberwolfach

№ 1/2018

The Algebraic Statistics of an Oberwolfach Workshop

Anna Seigal

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

1 Introduction

From 17th to 22nd 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.

We conducted a survey of the participants and 50 out of the 52 in attendance responded. The following table summarizes the data we obtained. It counts the observed answers to three questions, along with the empirical probabilities that are calculated from the counts. Such a table is called a *contingency table*.

		First time at MFO	Been to MFO before
Games	Liked weather	12 (24%)	5 (10%)
	Disliked weather	5 (10%)	4 (8%)
No games	Liked weather	7 (14%)	9 (18%)
	Disliked weather	3 (6%)	5 (10%)

For instance, a workshop participant selected at random enjoyed the weather, was at MFO for the first time and played games during the week with empirical probability 0.24 (or 24%) because, in our sample of 50 participants, 12 of them satisfy all three criteria. The eight probabilities given as percentages in the table form the joint probability distribution of three binary random variables, 'disliking the weather', 'having visited Oberwolfach before', 'playing no game'. These variables are *binary*: they take two values ("yes" or "no"). We can write the probabilities in an array of numbers, called a *tensor*, of size 2 × 2 × 2:

$$p = \left[\begin{bmatrix} p_{000} & p_{010} \\ p_{100} & p_{110} \end{bmatrix}, \quad \begin{bmatrix} p_{001} & p_{011} \\ p_{101} & p_{111} \end{bmatrix} \right] = \left[\begin{bmatrix} 0.24 & 0.1 \\ 0.1 & 0.08 \end{bmatrix}, \quad \begin{bmatrix} 0.14 & 0.18 \\ 0.06 & 0.1 \end{bmatrix} \right].$$

The eight numbers p_{ijk} represent the probabilities $P(X = i, Y = j, Z = k)$ for random variables X, Y and Z. In the survey, the three indicator random variables X, Y and Z are 'disliking the weather', 'having visited Oberwolfach before', and 'playing no game'. Above we noticed that, for example, $p_{000} = 0.24$. So-called *marginal probabilities* can be obtained by choosing one variable and ignoring the others. This corresponds to summing over all states for all but the chosen variable. For example, to find out the probability that a random participant liked the weather, we fix the first index and calculate the probability in the following way:

$$p_{000} + p_{010} + p_{001} + p_{011} = 0.24 + 0.1 + 0.14 + 0.18 = 0.66.$$

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. Maybe people who had not been in Oberwolfach before were more likely to play games, or perhaps a more subtle dependency can be found in the survey results. We focus on the *algebraic* structure of the statistical 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 3-dimensional simplex

$$\Delta_3 := \left\{ p \in \mathbb{R}^{2\times 2} : \sum_{i,j\in\{0,1\}} p_{ij} = 1, \quad p_{ij} \geq 0 \right\},$$

where the p_{ij} represent $P(X = i, Y = j)$. The simplex Δ_3 is a subset of the four-dimensional space $\mathbb{R}^{2\times 2}$ but, since all points also lie in the hyperplane of points satisfying $\sum_{i,j\in\{0,1\}} p_{ij} = 1$, it is a 3-dimensional object, which turns out to be a tetrahedron.

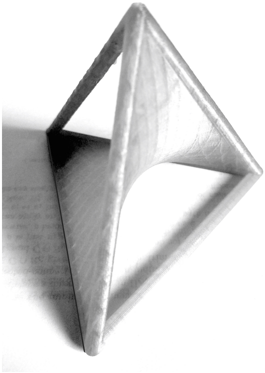

Figure 1: The 3D simplex Δ_3 is a tetrahedron. The independence model is the set of all points inside the tetrahedron for which (1) holds. This set is a surface inside the tetrahedron, parameterized by rank one matrices with non-negative entries. A photograph of a 3D printed version of this surface is shown above. For algebraists, it is the restriction of the Segre variety $\mathrm{Seg}(\mathbb{P}^1 \times \mathbb{P}^1)$ from $\mathbb{P}^3$ to the tetrahedron.

Two random variables X and Y are called *independent* if the distribution of one does not change under knowledge of the other. They lie in the *independence model* which consists of distributions which satisfy the single quadratic equation

$$\det \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} = p_{00}p_{11} - p_{01}p_{10} = 0. \tag{1}$$

The equation tests if the matrix of joint probabilities has rank one. This is the case if and only if one row in the matrix can be written as a multiple of the other row.

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

	First time at MFO	Been to MFO before
Liked weather	19 (38%)	14 (28%)
Disliked weather	8 (16%)	9 (18%)

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 we can express this by

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

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 of 0% of the space of all possible probability densities: it is a surface in a tetrahedron, see Figure 1. As we explain next, 0.0236 is “quite close” to 0, indicating that our table lies close to the independence model. Therefore, this model is a good explanation of the data observed.

For most applications, the gathered data is a sample of a larger population. We use the sample to infer properties of the whole population. In our survey, the population could be ‘all past, present and future MFO visitors’. To use such a model it is important to know whether the distance of the data from the model is statistically significant, that is, whether the data is further from the model than would be expected by chance. To quantify this, we can use the technique of Fisher’s exact test. Following [2, Proposition 1.1.8], we compute the probability of observing our data under the null hypothesis, which assumes independence of the two variables, and fixes 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, that is tables with larger determinant. The p-value exceeds the standard significance level of 5%, which means we do not reject the null hypothesis that the two variables are independent. There are other statistical tests that we could use to reach a similar conclusion, for example the asymptotics of Pearson’s χ^2 test.

Now we consider probability distributions of three binary random variables X, Y and Z instead of two variables. We shall examine the following three independence models on three random variables:

- full independence,
- marginal independence, and
- conditional independence.

2.1 Full Independence

In the situation for two random variables above, we considered a 2 × 2 matrix, and the statistical model consisted of all distributions whose matrices had rank one. We want to generalize this to three random variables. The statistical model for the full independence model for three random variables consists of $2 \times 2 \times 2$ tensors of rank one. That is, tensors which can be written $p = a \otimes b \otimes c$ for some vectors a, b, c. The tensor product “$\otimes$” of the vectors works as follows:

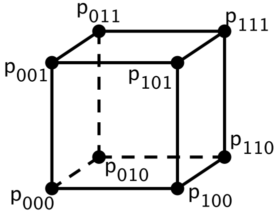

Figure 2: The eight joint probabilities of three binary random variables

to get the p_{ijk} entry of the tensor p, we multiply together the ith entry of the vector a, the jth entry of the vector b, and the kth entry of the vector c.

We consider all points in the 7-dimensional probability simplex

$$\Delta_7 := \left\{ p \in \mathbb{R}^{2\times 2\times 2} : \sum_{i,j,k\in\{0,1\}} p_{ijk} = 1, \quad p_{ijk} \geq 0 \right\},$$

which satisfy the system of equations that define the statistical model,

$$\begin{array}{lll} p_{000}p_{011} = p_{010}p_{001}, & p_{000}p_{101} = p_{100}p_{001}, & p_{000}p_{110} = p_{100}p_{010}, \\ p_{001}p_{111} = p_{101}p_{011}, & p_{010}p_{111} = p_{110}p_{011}, & p_{100}p_{111} = p_{110}p_{101}, \\ p_{000}p_{111} = p_{101}p_{010}, & p_{000}p_{111} = p_{110}p_{001}, & p_{000}p_{111} = p_{011}p_{100}. \end{array} \tag{2}$$

The first six equations are rank one conditions on facets of the cube in Figure 2. The last three are rank one conditions that involve the main diagonal from p_{000} to p_{111} and one of the three other main diagonals. We might think that the facet-independences are sufficient for full independence, since they describe independence between any pair of variables after fixing the value of the third variable. But just the six equations are not sufficient to fully describe the statistical model. This can be explored using the computer algebra software `Macaulay2`, using the code:

```
R = QQ[p000,p001,p010,p011,p100,p101,p110,p111];
I = ideal(p000*p011-p010*p001,p000*p101-p100*p001,p000*p110-p100*p010,
          p001*p111-p101*p011,p010*p111-p110*p011,p100*p111-p110*p101);
decompose I
```

Since 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 of 0% of the distributions: it is something 3-dimensional in a 7-dimensional space. 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 `algstat` in the statistical programming language `R`. We get a p-value of 0.492 when measuring the significance of tables exactly, and 0.515 using asymptotics via the χ^2 statistic. This means we do not reject the hypothesis that the three random variables are independent.

Each point on the model is a tuple of probabilities with particular structure. For each point in the model, we can compute the *likelihood* of observing our data under those probabilities. The point on the model with the highest likelihood is the maximum likelihood estimate (MLE). For the survey data, the MLE is

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

It can be computed by hand using the row and column sums of the data.

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 some random variables, after fixing the value of other random variables. The statistical model in which Y and Z are conditionally independent given X (denoted $(Y \perp\!\!\!\perp Z)|X$) consists of all distributions which satisfy the two equations

$$p_{000}p_{011} = p_{001}p_{010} \qquad \text{and} \qquad p_{100}p_{111} = p_{101}p_{110}. \tag{3}$$

For the survey data, this is the statistical model in which ‘having visited Oberwolfach before’ and ‘playing no game’ are independent, after we account for whether a participant liked the weather.

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 $Y \perp\!\!\!\perp Z$. It is defined by the polynomial equation

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

The two-dimensional example from the beginning of this section is an example of a marginal independence model, since we ignored the games variable.

The interplay between conditional and marginal (in)dependence can be subtle, as we can see in the famous example of Simpson’s paradox. Here, we have a certain correlation between two variables but when we fix the value of a third variable, the sign of the correlation changes. For a detailed explanation of the phenomenon see [4, Chapter 6]. For example, it could happen that those who liked the weather were more likely to play games, but that among the first time visitors to MFO there was a negative correlation between liking the weather and playing games, and likewise among those who had visited MFO before. The sign of the correlation after *marginalizing* by a third variable, can have a different sign than the correlation after *conditioning* on a third variable. This appears to be a contradiction, although there is no mathematical inconsistency. It highlights the importance of knowing the hidden variables, which influence the observations but have not been directly measured.

3 Latent Variable Models

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

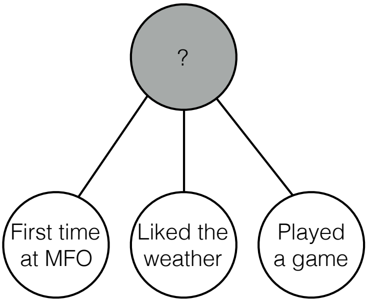

Figure 3: Three known and one hidden variable

The Naïve Bayes Model consists of some observed variables, and a single hidden variable. The statistical model is given by distributions on the observed variables, after marginalizing out the hidden variable. Its joint distribution is a convex combination of independent distributions.

We explore how to model the survey data using a hidden variable. In fact, our survey had two additional questions:

- *Do you consider yourself a "young person" or and "old 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. The rank of a tensor p is the minimal number of rank one tensors that must be summed to make p. The non-negative rank imposes the condition that the rank one tensors we sum up have non-negative entries. The tensors in our statistical model are those that can be written in the form

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

The non-negative rank is two because we summed over the values taken by the single hidden variable, which has two states. This statistical model is full-dimensional inside the probability simplex Δ_7. It occupies a volume of approximately 8% of the simplex. That region is given by the polynomial inequalities found in [1]. This region has 4 parts; one of which is given by what are called the *log-supermodularity conditions*

$$\begin{array}{lll} p_{000}p_{011} \geq p_{010}p_{001}, & p_{000}p_{101} \geq p_{100}p_{001}, & p_{000}p_{110} \geq p_{100}p_{010}, \\ p_{001}p_{111} \geq p_{101}p_{011}, & p_{010}p_{111} \geq p_{110}p_{011}, & p_{100}p_{111} \geq p_{110}p_{101}, \\ p_{000}p_{111} \geq p_{101}p_{010}, & p_{000}p_{111} \geq p_{110}p_{001}, & p_{000}p_{111} \geq p_{011}p_{100}. \end{array} \tag{4}$$

Notice these are the same equations as in (2), but with the equalities replaced by inequalities. There are three other regions of Δ_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 inequalities in (4) is not quite satisfied:

$$p_{010}p_{111} - p_{110}p_{011} = -0.0044. \tag{5}$$

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

$$\hat{p} = \left[\begin{bmatrix} 0.24 & 0.1096 \\ 0.1 & 0.1704 \end{bmatrix}, \begin{bmatrix} 0.14 & 0.0704 \\ 0.06 & 0.1096 \end{bmatrix} \right].$$

The MLE $\hat{p}$ is a point in the model, written as the sum of two non-negative rank one terms

$$\hat{p} = \lambda \begin{bmatrix} \alpha_0 \\ 1-\alpha_0 \end{bmatrix} \otimes \begin{bmatrix} \beta_0 \\ 1-\beta_0 \end{bmatrix} \otimes \begin{bmatrix} \gamma_0 \\ 1-\gamma_0 \end{bmatrix} + (1-\lambda) \begin{bmatrix} \alpha_1 \\ 1-\alpha_1 \end{bmatrix} \otimes \begin{bmatrix} \beta_1 \\ 1-\beta_1 \end{bmatrix} \otimes \begin{bmatrix} \gamma_1 \\ 1-\gamma_1 \end{bmatrix}.$$

The parameters in the maximum likelihood estimate for our data are

$$\begin{aligned} \lambda &= 0.509155 \\ (\alpha_0, \beta_0, \gamma_0) &= (0.709459, 1, 0.644068) \\ (\alpha_1, \beta_1, \gamma_1) &= (0.608696, 0.062840, 0.391304). \end{aligned}$$

The MLE in this particular case is found by considering the probabilities that occur in (5) as a distribution on two random variables X and Z (the value for Y is fixed in the expression), and finding the MLE for this conditional distribution using the row and column sums of its 2×2 matrix of probabilities.

The rank one terms in the MLE are two independent distributions that are found after fixing the value of the hidden variable. In our case, they are similar to the data we obtained after fixing the response to the age question. This indicates the importance of the age variable for understanding the survey data. It seems very plausible that being an 'old person' influences whether someone has been at MFO before.

4 Towards Deep Learning

Probability distributions on three binary random variables have only a 8% chance of being completely explained by a single hidden binary variable. Multiple hidden random variables arranged in a layer, such that all are connected to the observed variables but there are no direct edges between them, make a Restricted Boltzmann Machine (RBM). RBMs are building blocks for so-called deep belief networks. These are canonical deep learning models consisting of multiple layers of hidden 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 one layer at a time, and each pair of adjacent layers is an RBM. This is an example of a greedy algorithm, one in which a locally optimal solution is taken at several successive steps to reach the global solution.

Distributions in an RBM model can be written as the (entrywise) product of one Naïve Bayes model for each hidden variable. For algebraists, they are given by Hadamard products of secant varieties of Segre varieties. Algebraic geometry techniques have been used to establish that the parametrization is generically identifiable, meaning that there are only finitely many possible parameter values for each distribution in the model. This is a recent result of Montúfar and Morton [3]. Algebraic techniques can also tell us which probability distributions can be described by a particular arrangement of hidden variables. We see an example of this, below.

We consider the case of two hidden binary variables. This RBM model consists of all probability distributions that can be written as the entrywise product of two $2 \times 2 \times 2$ tensors from the Naïve Bayes Model that we encountered in section 3. The graphical representation of the RBM model is as follows:

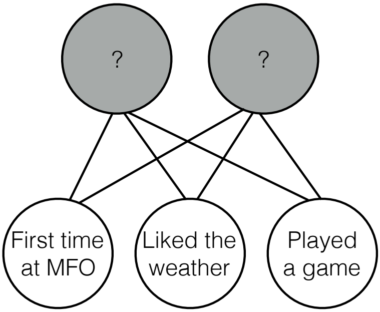

Figure 4: Three random variables and a layer of two hidden variables

This statistical model occupies an approximate volume of 76% of the probability simplex Δ_7. It consists of six regions inside of the simplex, namely, the Hadamard (entrywise) 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

$$p_{000}p_{011} \geq p_{001}p_{010} \qquad \text{and} \qquad p_{100}p_{111} \geq p_{101}p_{110}. \tag{6}$$

Notice that these are the same equations as (3), with the equalities replaced by inequalities. And recall that the inequalities in (4) were the same as the equations in (2), but with their equalities replaced by inequalities. The five other regions in the RBM model are obtained by reversing the inequalities in (6) or by permuting indices. A sketch of the Naive Bayes model contained inside the RBM model is shown in Figure 3.

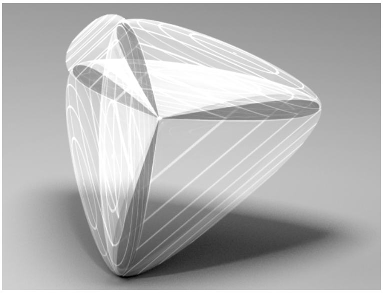

Figure 5: The four dark blobs are the pieces of the Naïve Bayes model. The striped blobs containing them are the six pieces of the Restricted Boltzmann Machine model.

The two inequalities in (6) 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

Figure 1 Surface 3D-printed by Thomas Kahle. Photograph by the author.

Figure 3 Made by Otavio Good.

All other figures made by the author.

References

[1] E. Allman, J. Rhodes, B. Sturmfels and P. Zwiernik: *Tensors of nonnegative rank two*, Linear Algebra and its Applications **473** (2015) 37–53.

[2] M. Drton, B. Sturmfels and S. Sullivant: *Lectures on Algebraic Statistics.* Oberwolfach Seminars, Vol 40, Birkhäuser, Basel, 2009.

[3] G. Montúfar and J. Morton: *Dimension of Marginals of Kronecker Product Models*, SIAM Journal on Applied Algebra and Geometry. **1** (2017), no. 1, 126–151.

[4] L. Schneps and C. Colmez, *Math on trial: how numbers get used and abused in the courtroom*, Wiley Online Library, 2013.

[5] S. Sullivant: *Algebraic Statistics*, draft copy, http://www4.ncsu.edu/~smsulli2/Pubs/asbook.pdf.

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

Mathematical subjects
Algebra and Number Theory, Probability Theory and Statistics

DOI
10.14760/SNAP-2018-001-EN

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

Director
Gerhard Huisken

Mathematisches
Forschungsinstitut
Oberwolfach

Member of
Leibniz
Leibniz
Association

IMAGINARY
open mathematics