Chapter 1

Introduction

Algebraic statistics advocates using tools from algebraic geometry, commutative algebra, and 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 in 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 use of algebraic tools in the method of moments [**Pea94**], the constructions of different types of combinatorial designs using 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 1990s.

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.

Probability/Statistics	Algebra/Geometry
Probability distribution	Point
Statistical model	(Semi)Algebraic set
Discrete exponential family	Toric variety
Conditional inference	Lattice points in polytopes
Maximum likelihood estimation	Polynomial optimization
Model selection	Geometry of singularities
Multivariate Gaussian model	Spectrahedral geometry
Phylogenetic model	Tensor networks
MAP estimates	Tropical geometry

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 Σ, 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 $X_1 = x_1, X_2 = x_2, \ldots$, etc. 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 $x_1, \ldots, x_i \in \Sigma$

$$P(X_i = x_i | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}) = P(X_i = x_i | X_{i-1} = x_{i-1}). \tag{1.1.1}$$

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 Σ, the set of all probability distributions consistent with the condition in Definition 1.1.1. As for many statistical models, the underlying description of this model is fundamentally algebraic in character: in particular, it is described implicitly by polynomial constraints on the joint distribution. Here is 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 distribution in this context with a point in $\mathbb{R}^8$. Indeed, the joint distribution is determined by the eight 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

$$\frac{p_{ijk}}{p_{ij+}} = \frac{p_{+jk}}{p_{+j+}} \tag{1.1.2}$$

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 (1.1.2) equal to each other for the same values of j and k. Clearing denominators 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 solution sets of systems of polynomial 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

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

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 (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:

$$p_{j_1 j_2 \cdots j_m} = \pi_{j_1} \prod_{i=2}^{m} \alpha_{i, j_{i-1}, j_i}. \tag{1.1.3}$$

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

$$000, 010, 110, 000, 101, 110, 100, 010, 110, 111, 000, 000, 010$$

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 data set D given the true unknown probability distribution p is then

$$P(\mathrm{D}) = \binom{n}{u} \prod_{ijk} p_{ijk}^{u_{ijk}} =: L(p \mid u).$$

Here n denotes the sample size and $\binom{n}{u}$ is the multinomial coefficient

$$\binom{n}{u} = \frac{n!}{\prod_{ijk} u_{ijk}!}.$$

Note, since here we have 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. This probability is called the likelihood function, denoted $L(p \mid u)$. 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, whereas the probability distribution p consists of 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

$$\arg\max \; L(p \mid u) \quad \text{subject to } p \in \mathcal{M}.$$

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.,

$$\arg\max \prod_{ijk} p_{ijk}^{u_{ijk}} \text{ subject to } p_{ijk} \geq 0, \; \sum_{ijk} p_{ijk} = 1, \text{ and}$$

$$p_{000}p_{101} - p_{001}p_{100} \quad = \quad p_{010}p_{111} - p_{011}p_{110} \quad = \quad 0.$$

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

$$\arg\max \prod_{ijk} (\pi_i \alpha_{ij} \beta_{jk})^{u_{ijk}}.$$

To simplify notation we take $\alpha_{ij} = \alpha_{2,i,j}$ and $\beta_{jk} = \alpha_{3,j,k}$ as compared to the general parametrization we saw in (1.1.3). Note this optimization problem is not a completely unconstrained optimization problem, because we need π to be a probability distribution and α and β to be conditional distribution. That is, we have the constraints

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

Note that in both cases we do not need to include the multinomial coefficient $\binom{n}{u}$ because this does not affect which distribution is the optimizer.

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

$$\hat{\pi}_i = \frac{u_{i++}}{u_{+++}}, \quad \hat{\alpha}_{ij} = \frac{u_{ij+}}{u_{i++}}, \quad \hat{\beta}_{jk} = \frac{u_{+jk}}{u_{+j+}}, \tag{1.1.4}$$

where

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

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

$$\begin{aligned} \hat{p} &= \left(\frac{4 \cdot 5}{6 \cdot 13}, \frac{4 \cdot 1}{6 \cdot 13}, \frac{3 \cdot 6}{7 \cdot 13}, \frac{3 \cdot 1}{7 \cdot 13}, \frac{2 \cdot 5}{6 \cdot 13}, \frac{2 \cdot 1}{6 \cdot 13}, \frac{4 \cdot 6}{7 \cdot 13}, \frac{4 \cdot 1}{7 \cdot 13} \right) \\ &\approx (0.256, 0.051, 0.198, 0.033, 0.128, 0.026, 0.264, 0.044). \end{aligned}$$

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 p 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 p 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

$$\begin{aligned} P(u|\pi,\alpha,\beta) &= \binom{n}{u}\prod_{ijk}(\pi_i\alpha_{ij}\beta_{jk})^{u_{ijk}} \\ &= \binom{n}{u}\prod_{i,j}(\pi_i\alpha_{ij})^{u_{ij+}}\prod_{j,k}\beta_{jk}^{u_{+jk}}. \end{aligned}$$

Note that $n = u_{+++}$ is the sample size. The quantities u_{ij+}, u_{+jk} with $i,j,k \in \Sigma$ are the *sufficient statistics* of the Markov chain model of length three. 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

$$P(v|\pi,\alpha,\beta,v_{ij+}=u_{ij+},v_{+jk}=u_{+jk}) = \frac{\binom{n}{v}\prod_{i,j}(\pi_i\alpha_{ij})^{u_{ij+}}\prod_{j,k}\beta_{jk}^{u_{+jk}}}{\sum_v\binom{n}{v}\prod_{i,j}(\pi_i\alpha_{ij})^{u_{ij+}}\prod_{j,k}\beta_{jk}^{u_{+jk}}},$$

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

$$P(v|\pi,\alpha,\beta,v_{ij+}=u_{ij+},v_{+jk}=u_{+jk}) = \frac{\binom{n}{v}}{\sum_v\binom{n}{v}}$$

and, in particular, this probability does not depend on the unknown distribution p. The fact that this probability does not depend on the 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

$$\frac{\sum_{v \in \mathcal{F}(u)} \binom{n}{v} 1_{\binom{n}{v} > \binom{n}{u}}(v)}{\sum_{v \in \mathcal{F}(u)} \binom{n}{v}},$$

where $1_{\binom{n}{v} > \binom{n}{u}}(v)$ 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. $P(v) \propto \binom{n}{v}$) 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:

$$\mathcal{F}(u) = \left\{ v \in \mathbb{Z}^8 : v_{ijk} \geq 0, \quad \sum_k v_{ijk} = u_{ij+}, \quad \text{and } \sum_i v_{ijk} = u_{+jk} \right\}.$$

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 allow us 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 (1.1.4).

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 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.