Log-Linear Models, Toric Varieties, and Their Markov Bases

Seth Sullivant
Department of Mathematics and Society of Fellows, Harvard University

My goal in this chapter of lecture notes is to introduce the class of log-linear models and study them from the algebraic perspective. There are at least three different ways to introduce log-linear models: in statistics they are also called discrete exponential families and in algebraic geometry they are known as toric varieties. Our goal is to introduce these models from all three perspective and show that they are equivalent. Then we take the approach suggested in the first chapter of lecture notes and study the ideals defining the log-linear model. They are known as toric ideals. Extensive further information about toric ideals can be found in [15].

In Section 3 of this chapter, we explain how the generators of the toric ideals are useful for performing conditional inference. In particular, these generating sets can be used to perform certain random walks to generate random samples from the hypergeometric distribution, and compute p-values of various statistical tests. In this context, generating sets for toric ideals are known as Markov bases. This connection was first made in [5]. In Section 4, we discuss the particular example of hierarchical models. In Section 5 we provide a description of some of the basic results about Markov bases of the hierarchical models. In Section 6 we describe some other interesting examples of log-linear models, in particular, models for analyzing ranked data and Hardy-Weinberg equilibrium.

1 Three Definitions of a Log-Linear Model

In this section, we give three definitions of log-linear models and show that they are all equivalent to each other. The log-linear models are an important family of models, because they mimic the structure of linear models for continuous random variables in the discrete case. It turns out that they have many nice properties on the statistical side, which are reflected in the mathematics of the corresponding toric varieties.

Let $\mathcal{A} \in \mathbb{N}^{d \times m}$ be a matrix of nonnegative integers and suppose that the vector $(1, 1, \ldots, 1)$ is in the row span of $\mathcal{A}$. Let $\mathbf{h} \in \mathbb{R}^m_{>0}$ be a vector of

positive real numbers.

Definition 1.1. Let $\Theta = \mathbb{R}^d_{>0}$ and let $\phi^{\mathcal{A},\mathbf{h}}$ be the rational parametrization

$$\phi^{\mathcal{A},\mathbf{h}} : \Theta \to \mathbb{R}^m, \qquad \phi^{\mathcal{A},\mathbf{h}}_j(\theta) = h_j \cdot \prod_{i=1}^{d} \theta_i^{a_{ij}}.$$

The *toric model* is the parametric algebraic statistical model

$$\mathcal{M}_{\mathcal{A},\mathbf{h}} \quad := \quad \phi^{\mathcal{A},\mathbf{h}}(\Theta) \cap \Delta_m.$$

With our restrictions on the matrix $\mathcal{A}$, this is equivalent to the following modified definition, which contains a normalizing constant.

Definition 1.2. Let $\Theta = \mathbb{R}^d_{>0}$ and let $\hat{\phi}^{\mathcal{A},\mathbf{h}}$ be the rational parametrization

$$\hat{\phi}^{\mathcal{A},\mathbf{h}} : \Theta \to \mathbb{R}^m, \qquad \phi^{\mathcal{A},\mathbf{h}}_j(\theta) = Z(\theta)^{-1} \cdot h_j \cdot \prod_{i=1}^{d} \theta_i^{a_{ij}},$$

where

$$Z(\theta) \quad = \quad \sum_{j=1}^{m} h_j \prod_{i=1}^{d} \theta_i^{a_{ij}}$$

is the normalizing constant. The *toric model* is the parametric algebraic statistical model

$$\mathcal{M}_{\mathcal{A},\mathbf{h}} \quad = \hat{\phi}^{\mathcal{A},\mathbf{h}}(\Theta).$$

For applications in machine learning and statistical physics, the normalizing constant $Z(\theta)$ is often called the partition function. For large scale toric models, where m is very large, finding efficient means to compute the partition function or approximate it is a crucial problem. When doing our algebraic analysis of toric models we will be able to work with non-normalized probability distributions, and the partition function will not play a role.

Although these two definitions are equivalent to each other, we will often need to go back and forth between them, and, in particular, will need to use the two definitions of $\hat{\phi}^{\mathcal{A},\mathbf{h}}$ and $\phi^{\mathcal{A},\mathbf{h}}$. Our two running examples from Chapter 1 of the lecture notes, the Bernoulli random variables and independent discrete random variables, are both instances of toric models, though we will need to modify the parametrizations to realize the models as toric models.

Example 1.3 (Bernoulli Random Variable)**.** Let $\mathcal{A}$ be the $2 \times m + 1$ matrix

$$\mathcal{A} = \begin{pmatrix} 0 & 1 & 2 & \cdots & m-1 & m \\ m & m-1 & m-2 & \cdots & 1 & 0 \end{pmatrix}$$

and let $h_i = \binom{m}{i}$ for $i = 0, 1, \ldots, m$. Then the model $\mathcal{M}_{\mathcal{A},\mathbf{h}}$ is the model of a Bernoulli random variable, with m flips of a biased coin. Indeed, we have

$$\hat{\phi}_j^{\mathcal{A},\mathbf{h}} = Z(\theta)^{-1}\binom{m}{j}\theta_1^j\theta_2^{m-j}.$$

However, we know that

$$\begin{aligned} Z(\theta) &= \sum_{j=0}^{m}\binom{m}{j}\theta_1^j\theta_2^{m-j} \\ &= (\theta_1+\theta_2)^m \end{aligned}$$

Thus we can write

$$\hat{\phi}_j^{\mathcal{A},\mathbf{h}} = \binom{m}{j}\left(\frac{\theta_1}{\theta_1+\theta_2}\right)^j\left(\frac{\theta_2}{\theta_1+\theta_2}\right)^{m-j},$$

and substituting $\theta = \theta_1/(\theta_1+\theta_2)$ and $1-\theta = \theta_2/(\theta_1+\theta_2)$ yields the parametrization we saw in chapter 1. □

Example 1.4 (Independence of Two Random Variables)**.** Let $\mathcal{A}$ be the matrix $(m_1+m_2)\times m_1m_2$ matrix with columns $(e_i, e_j)^T$ where in the first entry we take e_i to be a standard unit vector in $\mathbb{R}^{m_1}$ and in the second entry we take e_j to be a standard unit vector in $\mathbb{R}^{m_2}$. For instance, if $m_1 = 2$ and $m_2 = 4$, then

$$\mathcal{A} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Set $\mathbf{h} = (1, 1, \ldots, 1)$, the all ones vector. If we label the parameters corresponding to the first m_1 rows by $\alpha_1, \ldots, \alpha_{m_1}$ and the parameters corresponding to the last m_2 rows by $\beta_1, \ldots, \beta_{m_2}$, then this yields the paramerization:

$$\hat{\phi}_{ij}^{\mathcal{A}}(\alpha,\beta) = Z(\alpha,\beta)^{-1}\alpha_i\beta_j$$

where

$$\begin{aligned} Z(\alpha,\beta) &= \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\alpha_i\beta_j \\ &= \left(\sum_{i=1}^{m_1}\alpha_i\right)\left(\sum_{j=1}^{m_2}\beta_j\right) \end{aligned}$$

from which we deduce the expression

$$\hat{\phi}_{ij}^{\mathcal{A}}(\alpha,\beta) = \frac{\alpha_i}{\sum_{i=1}^{m_1}\alpha_i}\frac{\beta_j}{\sum_{j=1}^{m_2}\beta_j}.$$

Substituting $\hat{\alpha}_i = \frac{\alpha_i}{\sum_{i=1}^{m_1}\alpha_i}$ and $\hat{\beta}_j = \frac{\beta_j}{\sum_{j=1}^{m_2}\beta_j}$, yields the parametrization from Chapter 1, with parameter set $\mathbb{R}_{>0}^{m_1+m_2}$ replaced by $\Delta_{m_1} \times \Delta_{m_2}$. □

In both of the preceding two examples we were able to replace the parameter space $\mathbb{R}_{>0}^d$ with a smaller set Θ that was a polyhedron, and so that the map $\phi : \Theta \to \Delta_m$ was a polynomial map. In particular, we could write the parametrization in such a way that no normalizing constant was needed. This behavior is the exception rather than the rule for toric models, and the normalizing constant $Z(\theta)$ is necessary to fully specify the associated distribution in the probability simplex.

Example 1.5 (No Three-Way Interaction Model)**.** Let $X = (X_1, X_2, X_3)$ be three dimensional random vector with state space $[m_1] \times [m_2] \times [m_3]$. Consider the toric parametrization $\phi^{\mathcal{A}} : \mathbb{R}^d \to \mathbb{R}^m$ with $d = m_1m_2+m_1m_3+m_2m_3$ and $m = m_1m_2m_3$, defined by

$$\phi_{ijk}^{\mathcal{A}}(\alpha,\beta,\gamma) = \alpha_{ij}\beta_{ik}\gamma_{jk}.$$

In this parametrization , α is an $m_1 \times m_2$ matrix of parameters, β is an $m_1 \times m_3$ matrix of parameters, and γ is an $m_2 \times m_3$ matrix of parameters.

This toric model is an example of a hierarchical model, to be discussed in more detail in Section 4. It is known as the no 3-way interaction model, because it contains interaction factors (potential functions) between each pair of random variables (e.g. α is the interaction factor between X_1 and X_2) but no interaction factor between all three of the random variables.

The matrix $\mathcal{A}$ is a $m_1m_2 + m_1m_3 + m_2m_3 \times m_1m_2m_3$ matrix, whose columns are the vectors (e_{ij}, e_{ik}, e_{jk}) where the first entry e_{ij} is a standard unit vector in $\mathbb{R}^{m_1\times m_2}$, the second entry e_{ik} is a standard unit vector in $\mathbb{R}^{m_1\times m_3}$, and the third entry e_{jk} is a standard unit vector in $\mathbb{R}^{m_2\times m_3}$. □

The standard description of a log-linear model involves taking the logarithm of all probabilities and paramaters involved. Then we replace the parameter space $\Theta = \mathbb{R}_{>0}^d$ with all real vectors $\mathbb{R}^d$.

Definition 1.6. Given $\mathcal{A}$ and $\mathbf{h}$, the log-linear model consists of all probability distributions satisfying

$$\log p_j = h_j + \sum_{i=1}^{d} a_{ij}\theta_i$$

with $\theta \in \mathbb{R}^d$.

The description of the log-linear model as an exponential family merely takes the exponential of the log-linear model description, but keeps the log-linear parameters. Among the many nice features of exponential families are that they apply in many different contexts, including with continuous random variables. We will set up the exponential families in this general context. See [] for more information on general

Definition 1.7. Given a function $T(x)$, a function $A(\theta)$ and $h(x)$, the exponential family consists of all joint density functions of the form:

$$f(x;\theta) = h(x)\exp(\theta^T T(x) - A(\theta))$$

where $\theta \in \Theta$, the natural parameter space of the exponential family.

When the state space for X is finite, we will see that this reduces to the description of a log-linear or toric model.

Proposition 1.8. *The toric model, log-linear model, and exponential family associated to the matrix $\mathcal{A}$ and the parameter vector $\mathbf{h}$ are all the same set of probability distributions.*

Proof. The fact that the toric model and the log-linear model are equivalent follows by taking logarithms of the toric description an replacing θ and $\mathbf{h}$ with their coordinate-wise logarithms. To see that the discrete exponential family, first we need some observations in the discrete case. Since X is a random variable with finitely many states, $f(x;\theta)$ is the joint distribution of X, so we can replace it with $p_j = f(j;\theta)$. The function $h(j)$, becomes the vector $\mathbf{h}$, and the function $T(j)$, assigns a vector of length d to each $j \in m$, so we replace $T(j)$ with the column vector $\mathbf{a}_j$, the j-th column of the matrix $\mathcal{A}$. Final $\exp(-A(\theta))$ plays the role of the normalizing constant $Z(\theta)$. Thus we have, by our identifications

$$\begin{aligned} f(x;\theta) &= h(x)\exp(\theta^T T(x) - A(\theta)) \\ p_j &= Z(\theta)h_j \exp(\theta^T \mathbf{a}_j)) \\ &= Z(\theta)h_j \prod_{i=1}^{d} (e^{\theta_i})^{a_{ij}} \end{aligned}$$

which is the parametrization of the toric model, when $\Theta = \mathbb{R}^d$. □

I will primarily use the toric model description as the standard representation of a log-linear model. However, I prefer the term "log-linear model" because it is rather more evocative of the underlying structure of the statistical model: it consists of all probability distributions whose logarithms lie in a linear space.

2 Toric Ideals

In this section we will describe properties of the vanishing ideal of a log-linear model. These ideals are called toric ideals and we illustrate some of their basic properties. Recall that $\phi^{\mathcal{A},\mathbf{h}}$ is the map defining the toric model. For the first part of this section, we will work with the case that $\mathbf{h} = (1, 1, \ldots, 1)$ is the all ones vector and drop $\mathbf{h}$ from the notation. We will explain how the general case can be derived from the special case where $\mathbf{h} = (1, 1, \ldots, 1)$ later in the section.

Denote by $\phi^{\mathcal{A}}$ the toric parametrizing map (ignoring $\mathbf{h}$) which we consider as a map from $\mathbb{K}^d \to \mathbb{K}^m$. The Zariski closure of the image $\phi^{\mathcal{A}}(\mathbb{K}^d)$ is called an *affine toric variety* and denoted $V_{\mathcal{A}}$. The vanishing ideal $I(\phi^{\mathcal{A}}(\mathbb{K}^d)) \subset \mathbb{K}[\mathbf{p}]$ is called a *toric ideal* and is denoted $I_{\mathcal{A}}$.

Proposition 2.1. *The toric ideal $I_{\mathcal{A}}$ is generated by all monomial differences $\mathbf{p}^{\mathbf{u}} - \mathbf{p}^{\mathbf{v}}$, where $\mathbf{u}, \mathbf{v} \in \mathbb{N}^m$ and $\mathcal{A}\mathbf{u} = \mathcal{A}\mathbf{v}$:*

$$I_{\mathcal{A}} = \langle \mathbf{p}^{\mathbf{u}} - \mathbf{p}^{\mathbf{v}} \mid \mathbf{u}, \mathbf{v} \in \mathbb{N}^m \text{ and } \mathcal{A}\mathbf{u} = \mathcal{A}\mathbf{v} \rangle.$$

Proposition 2.1 says that toric ideals have *binomial generating sets*. Hilbert's basis theorem guarantees that there exists a finite subset of the binomials that generates the ideal. Finding such a finite subset can be a challenging problem in general. We explore this problem for our running examples.

Example 2.2 (Independence of Two Random Variables)**.** In the case of independence of two discrete random variables, the toric ideal $I_{\mathcal{A}}$ is generated by the 2×2 minors of the point distribution matrix (p_{ij}):

$$I_{\mathcal{A}} = \langle p_{ij}p_{kl} - p_{il}p_{kj} \mid i, k \in [m_1] \text{ and } j, l \in [m_2] \rangle.$$

Example 2.3 (No Three-way Interaction Model)**.** For the model of no 3-way interaction, a minimal generating set of binomials remains unknown. From computations, we know that the generating sets of binomials grow and depend in a complicated way on the levels (m_1, m_2, m_3). This will be explained in more detail in Sections 4 and 5. In the particular case where $m_1 = m_2 = m_3 = 3$, there are 81 binomial minimal generators. Twenty-seven of these have degree four, and are equivalent to the binomial

$$p_{111}p_{122}p_{212}p_{221} - p_{112}p_{121}p_{211}p_{222}.$$

The remaining 54 binomial generators have degree six and they are equivalent to the binomial

$$p_{111}p_{122}p_{133}p_{212}p_{223}p_{231} - p_{112}p_{123}p_{131}p_{111}p_{122}p_{133}.$$

In the case where $m_1 = m_2 = m_3 = 4$, the toric ideal of the no three-way interaction model requires $148,968$ minimal generators, which fall into 15 equivalence classes modulo the symmetries of the model. This example was computed using the project-and-lift algorithm of Hemmecke and Malkin [11] using the code `4ti2` [10]. It remains a challenging problem to determine such minimal generating sets for larger sets of levels (m_1, m_2, m_3).

Denote by $I_{\mathcal{A},\mathbf{h}}$ the vanishing ideal of the general toric model $\phi^{\mathcal{A},\mathbf{h}}$. This ideal can be recovered from the ideal $I_{\mathcal{A}}$ by a simple change of coordinates, namely, consider the diagonal linear transformation $\gamma_{\mathbf{h}}$ that sends the standard unit vector $\mathbf{e}_i \mapsto h_i \mathbf{e}_i$. Then

$$I_{\mathcal{A},\mathbf{h}} = \gamma_{\mathbf{h}}^*(I_{\mathcal{A}}),$$

which means that the generators of the ideal $I_{\mathcal{A},\mathbf{h}}$ can be recovered from $I_{\mathcal{A}}$ by globally replacing the indeterminates p_i with p_i/h_i. The transformation $\gamma_{\mathbf{h}}^*$ is an example of a ring homomorphism, to be elaborated upon in Chapter 3 of the lecture notes.

Example 2.4 (Bernoulli Random Variable)**.** In the pure toric model associated to a Bernoulli random variable, the parametrization is

$$\phi_j^{\mathcal{A}}(\theta) = \theta_1^j \theta_2^{m-j}.$$

The toric ideal $I_{\mathcal{A}}$ is generated by the 2×2 minors of the $2 \times m$ Hankel matrix

$$\begin{pmatrix} p_0 & p_1 & p_2 & \cdots & p_{m-1} \\ p_1 & p_2 & p_3 & \cdots & p_m \end{pmatrix}.$$

Replacing the indeterminates p_i with $p_i/\binom{m}{j}$ yields the binomial description we saw in Chapter 1.

Since toric ideals arise frequently in computational algebra and its applications (in statistics and otherwise) there has been some considerable effort to develop fast and dedicated algorithms to compute their generating sets (and Gröbner bases, to be explained in Chapter 3 of the lecture notes). One such code, that is extremely easy to use, is called `4ti2` [10]. `4ti2` is available for download at the website `www.4ti2.de`, and is a model of a mathematical software that is simple to download and use.

Here is an example of how to perform computations with `4ti2` to compute a generating set of a toric ideal.

Example 2.5. Suppose we want to compute the generating set of the toric ideal $I_{\mathcal{A}}$ corresponding to the binomial random variable with $n = 4$ coin flips. The matrix $\mathcal{A}$ is

$$\mathcal{A} = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 & 0 \end{pmatrix}.$$

To perform this computation, we create a file called `binomial` which looks like:

```
2 5
0 1 2 3 4
4 3 2 1 0
```

The first line tells the program the `binomial` represents a 2×5 matrix, and the second and third line give the rows of that matrix. Running the command `markov binomial` produces the file `binomial.mar`, which contains the generators of the ideal $I_{\mathcal{A}}$ in vector form. This file is

```
6 5
-1 0 2 0 -1
-1 1 0 1 -1
-1 1 1 -1 0
-1 2 -1 0 0
0 -1 1 1 -1
0 0 -1 2 -1
```

The first line tells us that the generating set consists of 6 elements (modulo sign) in $\mathbb{Z}^5$. Translating into binomials, we see, for instance, that the second line corresponds to the binomial $p_2^2 - p_0 p_4$. □

The reason the command for computing minimal generating sets is `markov` will be explained in the next section.

3 Conditional Inference and Markov Bases

Suppose that we are given data, in the form of the vector of counts $\mathbf{u}$. A typical inferential question concerning the data, and a particular model $\mathcal{M} \subset \Delta_m$, is "Does this data fit my model?" More precisely, the statistical question often boils down to determining whether or not we can reject the hypothesis that the data does not fit the model.

In non-precise terms, the conditional inference tests for log-linear models amounts to asking the question "Does the data seem exceptional or generic, among all data vectors with the same *sufficient statistics* as the given data?" The goal of this section will be to make these statements precise, and relate the statistical tests to the structure of the toric ideal $I_{\mathcal{A}}$.

Inference about statistical models is often tied to the likelihood function. In the case of i.i.d. random variables, this is the function

$$\mathcal{L}(\mathbf{u}|\theta) = \binom{|\mathbf{u}|}{\mathbf{u}} \cdot \prod_{j=1}^{m} \phi_j(\theta)^{u_j}$$

which is the probability under the model $\mathcal{M}$ of observing the data $\mathbf{u}$ given the parameter θ. Here

$$\binom{|\mathbf{u}|}{\mathbf{u}} = \binom{|\mathbf{u}|}{u_1, u_2, \ldots u_m} = \frac{(u_1 + u_2 + \cdots + u_m)!}{u_1! u_2! \cdots u_m!}$$

is a multinomial coefficient. For a log-linear model, the likelihood of the data only depends on the sufficient statistics.

Definition 3.1. *Sufficient statistics* of a model are any linear function of the data $T(\mathbf{u})$ such that

$$\mathcal{L}(\mathbf{u}|\theta) = f(T(\mathbf{u}), \theta) \cdot g(\mathbf{u})$$

for all data $\mathbf{u}$. That is, the interaction in the likelihood function of θ and $\mathbf{u}$ is only through $T(\mathbf{u})$. The sufficient statistics T are *minimal sufficient statistics* if for any other sufficient statistics T', there is a linear transformation γ such that $\gamma(T'(\mathbf{u})) = T(\mathbf{u})$ for all $\mathbf{u}$.

Proposition 3.2. *The minimal sufficient statistics of a log-linear model $\mathcal{M}_{\mathcal{A},\mathbf{h}}$ are $\mathcal{A}\mathbf{u}$.*

Remark. The minimal sufficient statistics of a log-linear model $\mathcal{M}_{\mathcal{A},\mathbf{h}}$ are the image of the *moment map* of the corresponding toric variety $V_{\mathcal{A},\mathbf{h}}$. This exhibits our first close connection between the statistical properties of the log-linear model and the corresponding toric variety.

The likelihood function for a log-linear model only depends on the minimal sufficient statistics. Indeed, for such a model we have:

$$\begin{aligned} \mathcal{L}_{\mathcal{A},\mathbf{h}}(\mathbf{u}|\theta) &= \prod_{j=1}^{m} \phi_(\theta)^{u_j} \\ &= \theta^{\mathcal{A}\mathbf{u}} \cdot \binom{|\mathbf{u}|}{\mathbf{u}} \cdot \prod_{j=1}^{m} h_j^{u_j}. \end{aligned}$$

The observation underlying conditional inference is that the *conditional likelihood function* for a log-linear model $\mathcal{L}_{\mathcal{A},\mathbf{h}}(\mathbf{u}|\theta, \mathcal{A}\mathbf{u})$ does not depend on θ. Indeed, the conditional likelihood function is

$$\begin{aligned} \mathcal{L}_{\mathcal{A},\mathbf{h}}(\mathbf{u}|\theta, \mathcal{A}\mathbf{u} = \mathbf{b}) &= \frac{\theta^{\mathcal{A}\mathbf{u}} \cdot \binom{|\mathbf{u}|}{\mathbf{u}} \cdot \prod_{j=1}^{m} h_j^{u_j}}{\sum_{\mathbf{v} \in \mathcal{A}^{-1}[\mathcal{A}\mathbf{u}]} \theta^{\mathcal{A}\mathbf{u}} \cdot \binom{|\mathbf{v}|}{\mathbf{v}} \cdot \prod_{j=1}^{m} h_j^{v_j}} \\ &= \frac{\binom{|\mathbf{u}|}{\mathbf{u}} \cdot \prod_{j=1}^{m} h_j^{u_j}}{\sum_{\mathbf{v} \in \mathcal{A}^{-1}[\mathcal{A}\mathbf{u}]} \binom{|\mathbf{v}|}{\mathbf{v}} \cdot \prod_{j=1}^{m} h_j^{v_j}} \\ &\propto \binom{|\mathbf{u}|}{\mathbf{u}} \cdot \prod_{j=1}^{m} h_j^{u_j}. \end{aligned}$$

Thus, conditional likelihood inference for a log-linear model does not depend on the parameters θ. For this reason, they are sometimes called *nuisance parameters*. The resulting statistical tests depend on computing expected values of certain test functions over the fiber $\mathcal{A}^{-1}[\mathcal{A}\mathbf{u}]$ with respect to the distribution $\mathcal{L}_{\mathcal{A},\mathbf{h}}(\mathbf{u}|\mathcal{A}\mathbf{u})$. This distribution is called the *hypergeometric distribution*.

Definition 3.3. A *conditional inference p-value* is an expected value of the form

$$\mathbb{E}_{\mathbf{p}}[f(\mathbf{v})]$$

where f is a function on the fiber $\mathcal{A}^{-1}[\mathcal{A}\mathbf{u}]$, and $\mathbf{p}$ denotes the hypergeometric distribution on $\mathcal{A}^{-1}[\mathcal{A}\mathbf{u}]$.

Example 3.4 (p-value of the Exact Test)**.** Let $f = \mathbf{1}_{p(\mathbf{v})\leq p(\mathbf{u})}(\mathbf{v})$ be the indicator function of the set of $\mathbf{v} \in \mathcal{A}^{-1}[\mathcal{A}\mathbf{u}]$ which have smaller probability under the hypergeometric distribution than $\mathbf{u}$. The expect value

$$\mu = \mathbb{E}_{\mathbf{p}}[\mathbf{1}_{p(\mathbf{v})\leq p(\mathbf{u})}(\mathbf{v})]$$

is called the p-value of the exact test. It is the weight sum of all table $\mathbf{v}$ which have smaller probability of than the given table $\mathbf{u}$. □

Example 3.5 (p-value of the χ^2-test)**.** Let

$$f(\mathbf{v}) = \sum_{j=1}^{m} \frac{(\hat{u}_j - v_j)^2}{\hat{u}_j}$$

be Pearson's χ^2 statistic, where $\hat{\mathbf{u}}$ is the maximum likelihood estimate of the data under the model $\mathcal{M}_{\mathcal{A},\mathbf{h}}$. (Maximum likelihood estimates will be explained in Chapter 4). Then

$$\mu = \mathbb{E}_{\mathbf{p}}[\mathbf{1}_{f(\mathbf{v})\geq f(\mathbf{u})}(\mathbf{v})]$$

is the p-value of the χ^2 test. It measures the weighted proportion of tables $\mathbf{v}$ in the fiber $\mathcal{A}^{-1}[\mathcal{A}\mathbf{u}]$ which have greater χ^2 value, and hence, are farther from the maximum likelihood estimate. □

In general, there are many other possible p-value tests that can arise when looking at log-linear models. Instances of this appear in [1] and [4].

It is a very rare event that it is possible to compute the p-values $\mathbb{E}_p[f(\mathbf{v})]$ exactly. In general, this is only possible in situations where either m is very small (see [6]) or the fiber itself $\mathcal{A}^{-1}[\mathcal{A}\mathbf{u}]$ is small and can be explicitly list. In general, we approximate $\mathbb{E}_p[f(\mathbf{v})]$ by generating random draws $\mathbf{v}^1, \ldots, \mathbf{v}^N$ from the hypergeometric distribution and use the estimate

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{v}^i)$$

which is an unbiased estimate of $\mu = \mathbb{E}_p[f(\mathbf{v})]$.

In the remainder of this section, we describe a Markov chain Montecarlo (MCMC) algorithm for generating such random draws and show how this relates to the structure of the toric ideal $I_{\mathcal{A}}$. The goal of the algorithm is to take a random walk over the fiber $\mathcal{A}^{-1}[\mathcal{A}\mathbf{u}]$. The main question to be addressed is how to determine a set of local moves for generating such a random walk.

Algorithm 3.6. (Metropolis-Hastings)

- Input: $\mathcal{F} \subset \ker_{\mathbb{Z}}(\mathcal{A})$, and $\mathbf{u} \in \mathbb{N}^m$.
- Output: $\mathbf{v} \in \mathcal{A}^{-1}[\mathcal{A}\mathbf{u}]$
- Set: $\mathbf{v} := \mathbf{u}$
- *Until mixing has occurred* Do
  1. Select $\mathbf{f} \in \mathcal{F}$ uniformly at random
  2. If $\mathbf{f} + \mathbf{v} \in \mathbb{N}^m$, set $\mathbf{v} = \mathbf{v} + \mathbf{f}$ with probability $\min(1, \frac{p(\mathbf{v}+\mathbf{f})}{p(\mathbf{v})})$.
- Output $\mathbf{v}$.

The output $\mathbf{v}$ will be a random sample from $\mathcal{A}^{-1}[\mathcal{A}\mathbf{u}]$ provided that the Markov chain is connected, that is every vector $\mathbf{v} \in \mathcal{A}^{-1}[\mathcal{A}\mathbf{u}]$ can be reached by some sequence of moves. This leads to the following definition.

Definition 3.7. A finite set $\mathcal{F} \subset \ker_{\mathbb{Z}}(\mathcal{A})$ of moves is called a *Markov basis* for $\mathcal{A}$ if, for every pair of nonnegative integer vectors $\mathbf{u}, \mathbf{u} \in \mathbb{N}^m$ with $\mathcal{A}\mathbf{u} = \mathcal{A}\mathbf{v}$, there exist a sequence of moves $< \mathbf{f}^1, \ldots, \mathbf{f}^l > \subseteq \mathcal{F}$ such that

$$\mathbf{v} = \mathbf{u} + \sum_{i=1}^{l} \mathbf{f}^i$$

and for each $j \in [l]$,

$$\mathbf{u} + \sum_{i=1}^{j} \mathbf{f}^i \geq 0.$$

The key result about Markov bases, and probably the first theorem of algebraic statistics, is that Markov bases are generating sets of toric ideals. This observation was originally made by Diaconis and Sturmfels [5].

Theorem 3.8. *A finite set of moves $\pm\mathcal{F} \subset \ker_{\mathbb{Z}}(\mathcal{A})$ is a Markov basis for $\mathcal{A}$ if and only if the ideal $J_{\mathcal{F}}$ generated by the corresponding binomials equals the toric ideal $I_{\mathcal{A}}$; that is, if and only if*

$$\left\langle \mathbf{p}^{\mathbf{f}^+} - \mathbf{p}^{\mathbf{f}^-} \mid \mathbf{f} \in \mathcal{F} \right\rangle = I_{\mathcal{A}}.$$

In particular, Markov bases exists, since the Hilbert basis theorem implies that every ideal $I_{\mathcal{A}}$ has a finite generating set. Furthermore, this implies that computational algebra can be used to determine the Markov bases of particular log-linear models: just compute a generating set of the corresponding toric ideal can convert to vector form. This also explains the name of the command `markov` in `4ti2` for computing generating sets of toric ideals, as we saw above.

4 Hierarchical Models

One of the most studied classes of log-linear models are the hierarchical log-linear models. These models use generalizations of graphical structures to model interactions between collections of random variables. They also represent a very rich class of log-linear models, with large subclasses where it is possible to say many things about the Markov bases, though there still remain many interesting open problems about them. In this section, I will introduce these hierarchical models with some example. In Section 5 I will give an overview of what is known about their Markov bases.

To introduce the notion of a hierarchical model, we first need to explain some objects in combinatorics. An abstract set system Γ on $[n] := \{1, 2, \ldots, n\}$ is a collection of subsets of $[n]$. A set system Γ is called a *simplicial complex* if it is closed downward: that is, $S \in \Gamma$ implies that $T \in \Gamma$ for all $T \subseteq S$. The elements of Γ are called faces. The dimension of a face $T \in \Gamma$, is $\#T - 1$. The maximal faces of Γ are called facets.

Simplicial complexes have geometric realizations, and are often visualized as collections of polyhedra. This explains the terminology (face, facet, dimension) that we used in the description. Since simplicial complexes are closed downward, it suffices to describe them by only listing the facets. For instance, the simplicial complex

$$\Gamma = \{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{1,2\}, \{1,3\}, \{2,3\}, \{2,4\}, \{1,2,3\}\}$$

has three facets, namely $\{5\}, \{2,4\}$ and $\{1,2,3\}$, and can be represented as

$$\Gamma = \langle \{5\}, \{2,4\}, \{1,2,3\} \rangle . \tag{1}$$

Geometrically, we think of this simplicial complex as consisting of a single point $\{5\}$, and a segment $\{2,4\}$ attached to a triangle $\{1,2,3\}$.

An alternate notation for these simplicial complexes comes from the study of hierarchical log-linear models, where square brackets are used instead of round brackets. So in this notation, we have

$$\Gamma = [5][24][123].$$

We will tend to use these different notations interchangeably. Note in particular, however, that $S \in \Gamma$ does not necessarily mean that S is a facet of Γ. For this reason, the notation from Equation 1 can be deceptive, and we will prefer to use the square brackets notation for a simplicial complex.

Let X be a discrete random vector $X = (X_1, \ldots, X_n)$ with

$$X \in [m_1] \times [m_2] \times \cdots \times [m_n].$$

Let $\mathbf{m} = (m_1, m_2, \ldots, m_n)$ be the vector levels of the random variables. Each state of the random variable X is a vector

$$\mathbf{i} = (i_1, i_2, \ldots, i_n) \in [m_1] \times [m_2] \times \cdots \times [m_n].$$

The joint distribution of X should be thought of as n-way table or tensor $(p_{\mathbf{i}}) \in \Delta_m$ where $m = \prod_{i=1}^{n} m_i$.

For a subset $S = (s_1, s_2, \ldots, s_k) \subseteq [n]$, $\mathbf{i}_S$ denotes the subvector of $\mathbf{i}$ indexed by the elements of S:

$$\mathbf{i}_S = (i_{s_1}, i_{s_2}, \ldots i_{s_k}) \in [m_{s_1}] \times [m_{s_2}] \times \cdots \times [m_{s_k}].$$

The hierarchical log-linear model for X associated to the simplicial complex Γ is defined as follows. For each facet $S \in \Gamma$, we get a collection of parameters $\theta^S_{\mathbf{j}_S}$ where

$$\mathbf{j}_S \in [m_{s_1}] \times [m_{s_2}] \times \cdots \times [m_{s_k}].$$

For $S \subseteq [n]$ denote by m_S the product

$$m_S = \prod_{s \in S} m_s.$$

The parameter space for the model is the set

$$\Theta^{\Gamma,\mathbf{m}} = \prod_{S \in \text{facet}(F)} \mathbb{R}^{m_S}_{>0}$$

and the rational parametrization for the hierarchical log-linear model is the map

$$\phi^{\Gamma,\mathbf{m}} : \Theta^{\Gamma,\mathbf{m}} \to \Delta_m \qquad \phi^{\Gamma}_{\mathbf{i}}(\theta) = \prod_{S \in \text{facet}(F)} \theta^S_{\mathbf{i}_S}.$$

Definition 4.1. The hierarchical model determined by Γ and $\mathbf{m}$ is the log-linear model

$$\mathcal{M}_{\Gamma,\mathbf{m}} := \phi^{\Gamma,\mathbf{m}}(\Theta^{\Gamma,\mathbf{m}}) \cap \Delta_m.$$

The toric ideal of the hierarchical model is denoted $I_{\Gamma,\mathbf{m}}$.

Many of the examples we have already seen are instances of hierarchical log-linear models.

Example 4.2 (Independence of Two Random Variables)**.** Let $\Gamma = [1][2]$ be the simplicial complex consisting of two isolated points and fix $\mathbf{m} = (m_1, m_2)$. The rational parametrization $\phi^{\Gamma,\mathbf{m}}$ is defined by the rule:

$$\phi^{\Gamma,\mathbf{m}}_{\mathbf{i}}(\theta^1, \theta^2) = \theta^1_{i_1}\theta^2_{i_2}$$

which we might write in a less cumbersome way as:

$$\phi^{\Gamma,\mathbf{m}}_{ij}(\alpha, \beta) = \alpha_i\beta_j.$$

The hierarchical model $\mathcal{M}_{\Gamma,\mathbf{m}}$ consists of all probability distributions in the image of the parametrization $\phi^{\Gamma,\mathbf{m}}$. This is equal to the model of independence of two random variables $\mathcal{M}_{X_1 \perp\!\!\!\perp X_2}$. □

Example 4.3 (No Three-way Interaction Model)**.** Let $\Gamma = [12][13][23]$ be the boundary of a triangle, and fix $\mathbf{m} = (m_1, m_2, m_3)$. The rationa l parametrization $\phi^{\Gamma,\mathbf{m}}$ is defined by the rule:

$$\phi^{\Gamma,\mathbf{m}}_{\mathbf{i}}(\theta^{[12]}, \theta^{[13]}, \theta^{[23]}) = \theta^{[12]}_{i_1 i_2}\theta^{[13]}_{i_1 i_3}\theta^{[23]}_{i_2 i_3}$$

which we might write in a less cumbersome way as:

$$\phi^{\Gamma,\mathbf{m}}_{ijk}(\alpha, \beta, \gamma) = \alpha_{ij}\beta_{ik}\gamma_{jk}.$$

Thus the hierarchical model $\mathcal{M}_{\Gamma,\mathbf{m}}$ is the model of no three-way interaction of Example 1.5. □

As we have seen in the preceding examples, the precise definition of the rational parametrization, while necessary for its generality, can be unnecessarily cumbersome when working with concrete examples. Usually when presenting a specific hierarchical model we will tend to use a simplified description of its parametrization to give a presentation of the model. This is illustrated in the example of the 4-chain.

Example 4.4 (4-Chain)**.** Let $\Gamma = [12][23][34]$ be a chain of length four, and fix $\mathbf{m} = (m_1, m_2, m_3, m_4)$. The rational parametrization $\phi^{\Gamma,\mathbf{m}}$ is defined by the rule:

$$\phi^{\Gamma,\mathbf{m}}_{ijkl}(\alpha, \beta, \gamma) = \alpha_{ij}\beta_{jk}\gamma_{kl}.$$

□

As we have seen in Section 3, the role of matrix $\mathcal{A}$, that represents a log-linear model, plays an important role in inference for that model, and, in particular, defines the structure of the toric ideal $I_{\mathcal{A}}$ and the resulting

Markov bases. With this in mind, it is important to study the structure of the matrices $A_{\Gamma,\mathbf{m}}$ that represent the hierarchical log-linear models. To provide such a description, we need to introduce the notion of a marginal of a multi-way table.

Definition 4.5. Let $\mathbf{u} \in \mathbb{R}^m$ be an $m_1 \times m_2 \times \cdots \times m_n$ table, and let $S \subseteq [n]$. The S-marginal of $\mathbf{u}$, denoted $\mathbf{u}_S$ is the $|S|$-way table obtained from $\mathbf{u}$ by summing over the indices not labeled by S. If we denote by $[n] \setminus S = \{t_1, \ldots, t_k\}$ then

$$(\mathbf{u}_S)_{\mathbf{i}_S} = \sum_{i_{t_1}=1}^{m_{t_1}} \cdots \sum_{i_{t_k}=1}^{m_{t_k}} \mathbf{u}_{\mathbf{i}}.$$

Computing marginals is linear in $\mathbf{u}$. Let $\pi_{\Gamma,\mathbf{m}}$ be the linear transformation

$$\pi_{\Gamma,\mathbf{u}} : \mathbb{R}^m \to \bigotimes_{S \in \text{facet}(\Gamma)} \mathbb{R}^{m_S}, \quad \mathbf{u} \mapsto (\mathbf{u}_{S_1}, \mathbf{u}_{S_2}, \ldots, \mathbf{u}_{S_k}),$$

where $\Gamma = [S_1][S_2]\cdots[S_k]$.

Proposition 4.6. *The matrix $A_{\Gamma,\mathbf{m}}$ that encodes the hierarchical model $\mathcal{M}_{\Gamma,\mathbf{m}}$ represents the linear transformation $\pi_{\Gamma,\mathbf{m}}$. That is, the minimal sufficient statistics of the hierarchical model $\mathcal{M}_{\Gamma,\mathbf{m}}$ are the marginals of $\mathbf{u}$ encoded by the simplicial complex Γ.*

Remark. To perform conditional inference tests for hierarchical models, we must generated random elements of the fibers $A^{-1}_{\Gamma,\mathbf{m}}[A_{\Gamma,\mathbf{m}}\mathbf{u}]$. This fiber consists of all nonnegative integral tables $\mathbf{v}$ with fixed margins $A_{\Gamma,\mathbf{m}}\mathbf{u}$. The fact that these fibers have such special structure will play a role in applications to statistical disclosure limitation in Chapter 3.

Example 4.7 (Independence of Two Random Variables)**.** For the model $\mathcal{M}_{X_1 \perp\!\!\!\perp X_2}$ of the independence of two random variables, whose simplicial complex $\Gamma = [1][2]$ consists of two isolated points, the sufficient statistics are the row and columns sums of the matrix $\mathbf{u}$.

Example 4.8 (No 3-way Interaction Model)**.** For the model of no 3-way interaction, whose simplicial complex $\Gamma = [12][13][23]$ is the boundary of a a triangle, the sufficient statistics are the 2-way margins of the 3-way tensor $\mathbf{u}$.

There are many special classes of hierarchical models, which appear as special cases, that are often more familiar and widely studied. The most important such subclass is the class of undirected graphical models. These models are defined in terms of cliques of an underlying graph. If $G = (V, E)$

is a graph with vertex set V and edge set E, a clique of G is collection of vertives $S \subset V$ such that every pair $i, j \in V$ determines an edge $\{i, j\} \in E$. The set of all cliques of the graph G, forms a simplicial complex $\Gamma(G)$, which is called the clique complex of G.

Definition 4.9. Let G be an undirected graph. The undirected graphical model is the hierarchical log-linear model $\mathcal{M}_{\Gamma(G),\mathbf{m}}$.

For instance, the independence model and the 4-chain model are examples of graphical models, whereas the no 3-way interaction model is hierarchical but not graphical.

5 Markov bases of Hierarchical Models

In this section, we want to give a description of some of the theorems that are known about the sturcture of the Markov bases of hierarchical models. Some of these results are constructive, and give explicit descriptions of Markov bases (or how to determine them from other Markov bases). Other results give complexity bounds on the structures of these Markov bases and show that they have "finite" complexity (when certain parameters are allowed to vary). Finally, the last results show that it is probably impossible to give a universal simple characterization of the Markov bases of hierarchical models.

5.1 Decomposable and Reducible Models

In the literature on graphical models, especially in the machine learning community, considerable attention is paid to the class of decomposable models. These models turn out to have many nice properties such as closed form expressions for maximum likelihood estimates, and performance guarantees on message passing algorithms such as belief propagation. In this subsection, we describe the Markov bases for decomposable models, a result first obtained in [7], as well as generalizations to the broader class of reducible models, the results obtained in [8] and [12].

To introduce the notions reducible and decomposable simplicial complexes, we need some preliminary definitions. If Γ is a simplicial complex $|\Gamma|$ denotes the set of all elements of $[n]$ appearing in Γ. Hence, if $\Gamma = [12][23][467]$, then $|\Gamma| = \{1, 2, 3, 4, 6, 7\}$. A simplicial complex is called a *simplex* if it has the form $2^S = \langle S \rangle = [S]$ for some set S.

Definition 5.1. A simplicial complex Γ is called *reducible*, with decomposition (Γ_1, S, Γ_2), with $S \subset [n]$ and Γ_1, Γ_2 subcomplexes of Γ, if

1. $\Gamma_1 \cup \Gamma_2 = \Gamma$, and

2. $\Gamma_1 \cap \Gamma_2 = 2^S$ (the intersection is a simplex).

A simplicial complex Γ is called *decomposable* if it is reducible and each of Γ_1 and Γ_2 are either decomposable or a simplex.

The set S in the description of a reducible model is called a *separator* of the reducible simplicial complex.

Example 5.2. The simplicial complex [1][2] of two isolated points is decomposable with $S = \emptyset$, since both [1] and [2] are simplices. In fact, any simplicial complex with only two facets is decomposable. The boundary of a triangle [12][13][23] is not reducible. Any chain $\Gamma = [12][23][34]\cdots[n-1n]$ is decomposable, with $S = \{n-1\}$, by induction. The simplicial complex [12][13][234] is reducible, with $\Gamma_1 = [12][13][23]$, $\Gamma_2 = [234]$, and $S = \{2, 3\}$, but it is not decomposable since Γ_1 is not reducible.

To describe the Markov bases of hierarchical models, we need to introduce some extra notation, the so-called "tableau" notation. This makes it simpler to work with complicated binomials where the interesting behavior of the binomial is contained in the indices labeling the polynomial indeterminates. To each monomial

$$\mathbf{p^u} = p_{\mathbf{i}_1} p_{\mathbf{i}_2} \cdots p_{\mathbf{i}_d}$$

where each $\mathbf{i}_j \in [m_1] \times [m_2] \times \cdots \times [m_n]$, we associate the *tableau*:

$$\mathbf{p^u} = \begin{bmatrix} \mathbf{i}_1 \\ \mathbf{i}_2 \\ \vdots \\ \mathbf{i}_d \end{bmatrix}$$

which is a $d \times n$ matrix of integers. If we partition $[n]$ into blocks $B_1, \ldots B_k$, we can write these tableau as

$$\mathbf{p^u} = \begin{bmatrix} \mathbf{i}_{11} & \cdots & \mathbf{i}_{1k} \\ \mathbf{i}_{21} & \cdots & \mathbf{i}_{2k} \\ \vdots & \ddots & \vdots \\ \mathbf{i}_{d1} & \cdots & \mathbf{i}_{dk} \end{bmatrix}$$

where $\mathbf{i}_{j_1 j_2} = (\mathbf{i}_{j_1})_{B_{j_2}}$.

Let (V_1, S, V_2) be a partition of $[n]$. Consider the binomial, written in tableau notation:

$$\mathbf{f} = \begin{bmatrix} \mathbf{i}_{V_1} & \mathbf{i}_S & \mathbf{i}_{V_2} \\ \mathbf{j}_{V_1} & \mathbf{i}_S & \mathbf{j}_{V_2} \end{bmatrix} - \begin{bmatrix} \mathbf{i}_{V_1} & \mathbf{i}_S & \mathbf{j}_{V_2} \\ \mathbf{j}_{V_1} & \mathbf{i}_S & \mathbf{i}_{V_2} \end{bmatrix}$$

where $\mathbf{i}_{V_i}, \mathbf{j}_{V_i} \in \prod_{k \in V_i}[m_k]$ and $\mathbf{i}_S \in \prod_{k \in S}[m_k]$. Suppose that (V_1, S, V_2) is an *valid partition* for the simplicial complex Γ, by which we mean that when S is removed from Γ, V_1 is separated from V_2. We claim that if (V_1, S, V_2) is valid for Γ, then $\mathbf{f} \in I_{\Gamma,\mathbf{m}}$.

Definition 5.3. Let $\text{Quad}(V_1, S, V_2)$ be the set of all the quadratic binomials that arise for the partition (V_1, S, V_2).

Theorem 5.4. *[7] Let Γ be a decomposable model, and let Q consist of all quadratic binomials which arise from all valid partitions (V_1, S, V_2) for Γ; that is,*

$$Q = \cup_{(V_1,S,V_2) \in \text{valid}(\Gamma)} \text{Quad}(V_1, S, V_2).$$

Then Q generates the toric ideal of the decomposable model: $I_{\Gamma,\mathbf{m}} = \langle Q \rangle$.

These quadrics which arise for valid partitions of a decomposable model correspond to *conditional independence statements* implied by the model. We will address such conditional independence structures in later chapters. Theorem 5.4 has a converse proven by Geiger, et al. [9].

Theorem 5.5. *If $I_{\Gamma,\mathbf{m}}$ is minimally generated by quadrics then Γ is decomposable.*

In general, the elements of $\text{Quad}(V_1, S, V_2)$ are useful for building up the Markov bases of reducible models from the Markov bases of the induced submodels. We will now explain how the construction works.

Let Γ be reducible with decomposition (Γ_1, S, Γ_2), which induces a valid partition for Γ, (V_1, S, V_2). Let $\mathbf{m}^1$, and $\mathbf{m}^2$ be the induced vector of levels for Γ_1 and Γ_2, respectively. Suppose that $\mathbf{f}$ is a binomial in $I_{\Gamma_1,\mathbf{m}^1}$. We may write this in tableau notation as

$$\mathbf{f} = \begin{bmatrix} \mathbf{i}_{11} & \mathbf{i}_{12} \\ \vdots & \vdots \\ \mathbf{i}_{d1} & \mathbf{i}_{d2} \end{bmatrix} - \begin{bmatrix} \mathbf{j}_{11} & \mathbf{j}_{12} \\ \vdots & \vdots \\ \mathbf{j}_{d1} & \mathbf{j}_{d2} \end{bmatrix}$$

where $\mathbf{i}_{k1}, \mathbf{j}_{k1} \in \prod_{l \in V_1}[m_l]$ and $\mathbf{i}_{k2}, \mathbf{j}_{k2} \in \prod_{l \in S}[m_l]$. Since $\mathbf{f} \in I_{\Gamma_1,\mathbf{m}^1}$, there exists a permutation of the rows of the tableaux such that that

$$\mathbf{f} = \begin{bmatrix} \mathbf{i}_{11} & \mathbf{i}_{12} \\ \vdots & \vdots \\ \mathbf{i}_{d1} & \mathbf{i}_{d2} \end{bmatrix} - \begin{bmatrix} \mathbf{j}_{11} & \mathbf{i}_{12} \\ \vdots & \vdots \\ \mathbf{j}_{d1} & \mathbf{i}_{d2} \end{bmatrix}.$$

Now let $\mathbf{i}_{13}, \ldots \mathbf{i}_{d3} \in \prod_{l \in V_2}[m_l]$ and define the new binomial:

$$\mathbf{f}^* = \begin{bmatrix} \mathbf{i}_{11} & \mathbf{i}_{12} & \mathbf{i}_{13} \\ \vdots & \vdots & \vdots \\ \mathbf{i}_{d1} & \mathbf{i}_{d2} & \mathbf{i}_{d3} \end{bmatrix} - \begin{bmatrix} \mathbf{j}_{11} & \mathbf{i}_{12} & \mathbf{i}_{13} \\ \vdots & \vdots & \vdots \\ \mathbf{j}_{d1} & \mathbf{i}_{d2} & \mathbf{i}_{d3} \end{bmatrix} \tag{2}$$

The binomial $\mathbf{f}^*$ is easily seen to lie in $I_{\Gamma,\mathbf{m}}$. The construction of "lifted" binomials from $I_{\Gamma_1,\mathbf{m}^1}$ to $I_{\Gamma,\mathbf{m}}$ works completely analogously for binomials in $I_{\Gamma_2,\mathbf{m}^2}$

Definition 5.6. Let Γ be reducible with decomposition (Γ_1, S, Γ_2). Let $\mathcal{F}_i \subset I_{\Gamma_i,\mathbf{m}^i}$ be a collection of binomials. Define $\mathrm{Lift}(\mathcal{F}_i)$ to be the collection of all binomials arises from Equation 2.

Theorem 5.7. *Let Γ be reducible, with decomposition (Γ_1, S, Γ_2). Let $\mathcal{F}_i \subset I_{\Gamma_i,\mathbf{m}^i}$ be a binomial generating set for $I_{\Gamma_i,\mathbf{m}^i}$, $i = 1, 2$. Then*

$$\mathcal{M} = \mathrm{Lift}(\mathcal{F}_1) \cup \mathrm{Lift}(\mathcal{F}_2) \cup \mathrm{Quad}(V_1, S, V_2)$$

is a binomial generating set of $I_{\Gamma,\mathbf{m}}$.

Theorem 5.7 was originally proved in [8] and [12]. Theorem 5.7 has many corollaries and generalizations.

5.2 Markov Complexity

In general, we are often interested in finding concise descriptions of Markov bases. As we have seen in Subsection 5.1, for reducible models and especially decomposable models, there is a particularly short and concise way to describe the Markov bases. In particular, for decomposable models, the vector $\mathbf{m}$ plays essentially no role in the description of Markov bases for these models, and all the elements of the minimal Markov basis are quadratic. It is natural to ask to what extent this generalizes to more complex models.

As shown in [13], generalizing a result from [14], if we fix all entries of $\mathbf{m}$ but allow one of these entries to vary, there is still a finite description of the Markov bases of all of the resulting models.

Theorem 5.8. *Fix Γ and $m_2, m_3, \ldots, m_n$. There exists a constant $m = m(\Gamma; m_2, \ldots, m_n)$ such that any minimal Markov basis for the log-linear model $\mathcal{M}_{\Gamma,\mathbf{m}}$ consists of moves of format $m \times m_2 \times \cdots \times m_n$. In particular, if $m_2, \ldots, m_n$ and Γ are fixed, the complexity of describing the Markov basis is bounded.*

5.3 Thin 3-way Tables

In spite of the preceding results about the complexity and construction of certain Markov bases of hierarchical models, a theorem of De Loera and Onn [3] based on results in [2] implies that the Markov bases of even some simple hierarchical models can be arbitrarily complicated.

Theorem 5.9. *[3, Cor 2.1] Let $\Gamma = [12][13][23]$ be the boundary of a triangle and let $\mathbf{v} \in \mathbb{Z}^r$, be any integer vector. Then there exists $\mathbf{m} = (3, m_2, m_3)$, $\mathbf{u} \in \ker_{\mathbb{Z}}(A_{\Gamma,\mathbf{m}})$, and a coordinate projection map $\pi : \mathbb{Z}^m \to \mathbb{Z}^r$ such that:*

1. $\mathbf{u}$ *belongs to every minimal Markov basis of the log-linear model* $\mathcal{M}_{\Gamma,\mathbf{m}}$, *and*

2. $\pi(\mathbf{u}) = \mathbf{v}$.

Theorem 5.9 says that every integer vector appears as part of a Markov basis element for some $3 \times m_2 \times m_3$ table under the no-three-way interaction model. For instance, taking $r = 1$, this theorem says that for any integer value v (e.g. $v = 100$) there exists a Markov basis element for the no-three-way interaction model with a coordinate equal to v. So, in particular, Markov bases for hierarchical models can be very far from squarefree.

6 Other Log-linear Models

Aside from the hierarchical models, which we have focused on in these lectures, there are many other instances of log-linear models that appear in the literature, whose Markov bases are worthy of study. In this section, we present two such model. The first is an example of a log-linear model with coefficients, and arises in the analysis of the genomic data. The second model, which we call the Birkhoff model, arises in the statistical analysis of ranked data. In the first instance, the Markov basis is easy to describe, and in the second, it is still open to describe the conjecturally simple Markov basis.

6.1 The Birkhoff model

The *Birkhoff model* (my terminology) is a model for analyzing ranked data, and is a particular instance of a log-linear model that arises from using noncommutative Fourier analysis. The Markov bases of this model were studied in the original Markov basis paper of Diaconis and Sturmfels [5] and subsequently in [4].

The matrix $\mathcal{A}_n \in \mathbb{N}^{n^2 \times n!}$ consists of all $n \times n$ permutations matrices. The convex hull of the columns of $\mathcal{A}_n$ is called the *Birkhoff polytope* of doubly stochastic matrices. Using a Gröbner basis argument, Diaconis and Sturmfels showed that the toric ideal $I_{\mathcal{A}_n}$ is generated in degree $\leq n$. Diaconis and Eriksson [4] subsequently improved the upper bound by 1 to deduce that the ideal is generated in degree $\leq n-1$. However, they conjecture that this bound is far from best possible.

Conjecture 6.1 (Diaconis-Eriksson)**.** *The Birkhoff ideal* $I_{\mathcal{A}_n}$ *is minimally generated in degree* ≤ 3.

Diaconis and Eriksson also verified the conjecture for $n \leq 6$, using a randomized enumeration argument [4].

6.2 Hardy-Weinberg Proportions

References

[1] J. Blitzstein and P. Diaconis. A sequential importance sampling algorithm for generating random graphs with prescribed degrees, 2006. Manuscript.

[2] J. De Loera and S. Onn. All rational polytopes are transportation polytopes and all polytopal integer sets are contingency tables. Integer programming and combinatorial optimization, 338–351, Lecture Notes in Comput. Sci., 3064, Springer, Berlin, 2004.

[3] J. De Loera and S. Onn. Markov bases of three-way tables are arbitrarily complicated. *J. Symbolic Comput.* **41** (2006), no. 2, 173–181.

[4] P. Diaconis and N. Eriksson. Markov bases for noncommutative Fourier analysis of ranked data. *J. Symbolic Comput.* **41** (2006), no. 2, 182–195.

[5] P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26** (1998), no. 1, 363–397.

[6] I. H. Dinwoodie, L. Matusevich, and E. Mosteig. Transform methods for the hypergeometric distribution. *Statistics and Computing*, **14** (2004), 287–297

[7] A. Dobra. Markov bases for decomposable graphical models. *Bernoulli* **9** (2003) 1–16.

[8] A. Dobra and S. Sullivant A Divide-and-conquer algorithm for generating Markov bases of multi-way tables. *Computational Statistics* **19** (2004), 347-366

[9] D. Geiger, C. Meek, and B. Sturmfels. On the toric algebra of graphical models. Annals of Statistics **34** (2006)

[10] R. Hemmecke, R. Hemmecke, and P. Malkin, `4ti2` version 1.2: Computation of Hilbert bases, Graver basis, toric Gröbner bases, and more, 2005. Available for download at `www.4ti2.de`.

[11] R. Hemmecke and P. Malkin. Computing generating sets of lattice ideals, 2005 `math.CO/0508359`.

[12] S. Hoşten and S. Sullivant. Gröbner bases and polyhedral geometry of reducible and cyclic models. *Journal of Combinatorial Theory: Series A* **100** (2002) 277–301

[13] S. Hoşten and S. Sullivant. A finiteness theorem for Markov bases of hierarchical models. *J. Combin. Theory Ser. A*, to appear, 2006.

[14] F. Santos and B. Sturmfels. Higher Lawrence configurations. *J. Combin. Theory Ser. A* **103** (2003), no. 1, 151–164.

[15] B. Sturmfels. *Gröbner Bases and Convex Polytopes*. University Lecture Series **8**. AMS Press, Providence, RI, 1996.