Chapter 1

Markov Bases

This chapter introduces the fundamental notion of a Markov basis, which represents one of the first connections between commutative algebra and statistics. This connection was made in the paper by Diaconis and the second author [33] on contingency table analysis. Statistical hypotheses about contingency tables can be tested in an exact approach by performing random walks on a constrained set of tables with non-negative integer entries. Markov bases are of key importance to this statistical methodology because they comprise moves between tables that ensure that the random walk connects every pair of tables in the considered set.

Section 1.1 reviews the basics of contingency tables and exact tests; for more background see also the books by Agresti [1], Bishop, Holland, Fienberg [18], or Christensen [21]. Section 1.2 discusses Markov bases in the context of hierarchical log-linear models. The problem of computing Markov bases is addressed in Section 1.3, where the problem is placed into the setting of integer lattices and tied to the algebraic notion of a lattice ideal.

1.1 Hypothesis Tests for Contingency Tables

A contingency table contains counts obtained by cross-classifying observed cases according to two or more discrete criteria. Here the word 'discrete' refers to criteria with a finite number of possible levels. As an example consider the 2×2-contingency table shown in Table 1.1.1. This table, which is taken from [1, §5.2.2], presents a classification of 326 homicide indictments in Florida in the 1970s. The two binary classification criteria are the defendant's race and whether or not the defendant received the death penalty. A basic question of interest for this table is whether at the time death penalty decisions were made independently of the defendant's race. In this section we will discuss statistical tests of such independence hypotheses as well as generalizations for larger tables.

	Death Penalty		
Defendant's Race	Yes	No	Total
White	19	141	160
Black	17	149	166
Total	36	290	326

Table 1.1.1: Data on death penalty verdicts.

Classifying a randomly selected case according to two criteria with r and c levels, respectively, yields two random variables X and Y. We code their possible outcomes as $[r]$ and $[c]$, where $[r] := \{1, 2, \ldots, r\}$ and $[c] := \{1, 2, \ldots, c\}$. All probabilistic information about X and Y is contained in the *joint probabilities*

$$p_{ij} = P(X = i, Y = j), \quad i \in [r],\ j \in [c],$$

which determine in particular the *marginal probabilities*

$$p_{i+} := \sum_{j=1}^{c} p_{ij} = P(X = i), \quad i \in [r],$$

$$p_{+j} := \sum_{i=1}^{r} p_{ij} = P(Y = j), \quad j \in [c].$$

Definition 1.1.1. The two random variables X and Y are *independent* if the joint probabilities factor as $p_{ij} = p_{i+}p_{+j}$ for all $i \in [r]$ and $j \in [c]$. We use the symbol $X \perp\!\!\!\perp Y$ to denote independence of X and Y.

Proposition 1.1.2. *The two random variables X and Y are independent if and only if the $r \times c$-matrix $p = (p_{ij})$ has rank 1.*

Proof. ($\Longrightarrow$): The factorization in Definition 1.1.1 writes the matrix p as the product of the column vector filled with the marginal probabilities p_{i+} and the row vector filled with the probabilities p_{+j}. It follows that p has rank 1.

($\Longleftarrow$): Since p has rank 1, it can be written as $p = ab^t$ for $a \in \mathbb{R}^r$ and $b \in \mathbb{R}^c$. All entries in p being non-negative, a and b can be chosen to have non-negative entries as well. Let a_+ and b_+ be the sums of the entries in a and b, respectively. Then, $p_{i+} = a_i b_+$, $p_{+j} = a_+ b_j$, and $a_+ b_+ = 1$. Therefore, $p_{ij} = a_i b_j = a_i b_+ a_+ b_j = p_{i+}p_{+j}$ for all i, j. □

Suppose now that we randomly select n cases that give rise to n independent pairs of discrete random variables

$$\begin{pmatrix} X^{(1)} \\ Y^{(1)} \end{pmatrix}, \begin{pmatrix} X^{(2)} \\ Y^{(2)} \end{pmatrix}, \ldots, \begin{pmatrix} X^{(n)} \\ Y^{(n)} \end{pmatrix} \tag{1.1.1}$$

that are all drawn from the same distribution, that is,

$$P(X^{(k)} = i, Y^{(k)} = j) = p_{ij} \quad \text{for all } i \in [r],\ j \in [c],\ k \in [n].$$

The joint probability matrix $p = (p_{ij})$ for this distribution is considered to be an *unknown* element of the $rc - 1$ dimensional probability simplex

$$\Delta_{rc-1} = \left\{ q \in \mathbb{R}^{r \times c} \ : \ q_{ij} \geq 0 \text{ for all } i, j \text{ and } \sum_{i=1}^{r} \sum_{j=1}^{c} q_{ij} = 1 \right\}.$$

A *statistical model* $\mathcal{M}$ is a subset of Δ_{rc-1}. It represents the set of all candidates for the unknown distribution p.

Definition 1.1.3. The *independence model* for X and Y is the set

$$\mathcal{M}_{X \perp\!\!\!\perp Y} = \{ p \in \Delta_{rc-1} \ : \ \text{rank}(p) = 1 \}.$$

The independence model $\mathcal{M}_{X \perp\!\!\!\perp Y}$ is the intersection of the probability simplex Δ_{rc-1} and the set of all matrices $p = (p_{ij})$ such that

$$p_{ij} p_{kl} - p_{il} p_{kj} = 0 \tag{1.1.2}$$

for all $1 \leq i < k \leq r$ and $1 \leq j < l \leq c$. The solution set to this system of quadratic equations is known as the *Segre variety* in algebraic geometry. If all probabilities are positive, then the vanishing of the 2×2-minor in (1.1.2) corresponds to

$$\frac{p_{ij}/p_{il}}{p_{kj}/p_{kl}} = 1. \tag{1.1.3}$$

Ratios of probabilities being termed *odds*, the ratio in (1.1.3) is known as an *odds ratio* in the statistical literature.

The order of the observed pairs in (1.1.1) carries no information about p and we summarize the observations in a table of counts

$$U_{ij} = \sum_{k=1}^{n} 1_{\{X^{(k)}=i,\, Y^{(k)}=j\}}, \quad i \in [r],\ j \in [c]. \tag{1.1.4}$$

The table $U = (U_{ij})$ is a *two-way contingency table*. We denote the set of all contingency tables that may arise for fixed sample size n by

$$\mathcal{T}(n) := \left\{ u \in \mathbb{N}^{r \times c} \ : \ \sum_{i=1}^{r} \sum_{j=1}^{c} u_{ij} = n \right\}.$$

Proposition 1.1.4. *The random table $U = (U_{ij})$ has a* multinomial distribution, *that is, if $u \in \mathcal{T}(n)$ and n is fixed, then*

$$P(U = u) = \frac{n!}{u_{11}! u_{12}! \cdots u_{rc}!} \prod_{i=1}^{r} \prod_{j=1}^{c} p_{ij}^{u_{ij}}.$$

Proof. We observe $U = u$ if and only if the observations in (1.1.1) include each pair $(i, j) \in [r] \times [c]$ exactly u_{ij} times. The product $\prod_i \prod_j p_{ij}^{u_{ij}}$ is the probability of observing one particular sequence containing each (i, j) exactly u_{ij} times. The pre-multiplied multinomial coefficient is the number of possible sequences of samples that give rise to the counts u_{ij}. □

Consider now the *hypothesis testing* problem

$$H_0 \colon p \in \mathcal{M}_{X \perp\!\!\!\perp Y} \quad \text{versus} \quad H_1 \colon p \notin \mathcal{M}_{X \perp\!\!\!\perp Y}. \tag{1.1.5}$$

In other words, we seek to decide whether or not the contingency table U provides evidence against the *null hypothesis* H_0, which postulates that the unknown joint distribution p belongs to the independence model $\mathcal{M}_{X \perp\!\!\!\perp Y}$. This is the question of interest in the death penalty example in Table 1.1.1, and we present two common approaches to this problem.

Chi-square test of independence. If H_0 is true, then $p_{ij} = p_{i+}p_{+j}$, and the expected number of occurrences of the joint event $\{X = i, Y = j\}$ is $np_{i+}p_{+j}$. The two sets of marginal probabilities can be estimated by the corresponding empirical proportions

$$\hat{p}_{i+} = \frac{U_{i+}}{n} \quad \text{and} \quad \hat{p}_{+j} = \frac{U_{+j}}{n},$$

where the *row total*

$$U_{i+} = \sum_{j=1}^{c} U_{ij}$$

counts how often the event $\{X = i\}$ occurred in our data, and the similarly defined *column total* U_{+j} counts the occurrences of $\{Y = j\}$. We can thus estimate the expected counts $np_{i+}p_{+j}$ by $\hat{u}_{ij} = n\hat{p}_{i+}\hat{p}_{+j}$. The *chi-square statistic*

$$X^2(U) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(U_{ij} - \hat{u}_{ij})^2}{\hat{u}_{ij}} \tag{1.1.6}$$

compares the expected counts $\hat{u}_{ij}$ to the observed counts U_{ij} taking into account how likely we estimate each joint event to be. Intuitively, if the null hypothesis is true, we expect X^2 to be small since U should be close to $\hat{u}$. The *chi-square test* rejects the hypothesis H_0, if the statistic X^2 comes out to be "too large."

What is "too large"? This can be gauged using a probability calculation. Let $u \in \mathcal{T}(n)$ be a contingency table containing observed numerical values such as, for instance, Table 1.1.1. Let $X^2(u)$ be the corresponding numerical evaluation of the chi-square statistic. We would like to compute the probability that the random variable $X^2(U)$ defined in (1.1.6) takes a value greater than or equal to $X^2(u)$ provided that H_0 is true. This probability is the *p-value* of the test. If the p-value is very small, then it is unlikely to observe a table with chi-square statistic

value as large or larger than $X^2(u)$ when drawing data from a distribution in the independence model $\mathcal{M}_{X \perp\!\!\!\perp Y}$. A small p-value thus presents evidence against H_0.

Suppose the p-value for our data is indeed very small, say 0.003. Then, assuming that the model specified by the null hypothesis H_0 is true, the chance of observing data such as those we were presented with or even more extreme is only 3 in 1000. There are now two possible conclusions. Either we conclude that this rare event with probability 0.003 did indeed occur, or we conclude that the null hypothesis was wrong. Which conclusion one is willing to adopt is a subjective decision. However, it has become common practice to reject the null hypothesis if the p-value is smaller than a threshold on the order of 0.01 to 0.05. The latter choice of 0.05 has turned into a default in the scientific literature.

On the other hand, if $X^2(u)$ is deemed to be small, so that the p-value is large, the chi-square test is inconclusive. In this case, we say that the chi-square test does not provide evidence against the null hypothesis.

The above strategy cannot be implemented as such because the probability distribution of $X^2(U)$ depends on where in the model $\mathcal{M}_{X \perp\!\!\!\perp Y}$ the unknown underlying joint distribution $p = (p_{ij})$ lies. However, this problem disappears when considering limiting distributions for growing sample size n.

Definition 1.1.5. The *standard normal distribution* $\mathcal{N}(0,1)$ is the probability distribution on the real line $\mathbb{R}$ that has the density function

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

If $Z_1, \ldots, Z_m$ are independent $\mathcal{N}(0,1)$-random variables, then $Z_1^2 + \cdots + Z_m^2$ has a *chi-square distribution* with m *degrees of freedom*, which we denote by χ^2_m.

In the following proposition, we denote the chi-square statistic computed from an n-sample by $X_n^2(U)$ in order to emphasize the dependence on the sample size. A proof of this proposition can be found, for example, in [1, §12.3.3].

Proposition 1.1.6. *If the joint distribution of X and Y is determined by an $r \times c$-matrix $p = (p_{ij})$ in the independence model $\mathcal{M}_{X \perp\!\!\!\perp Y}$ and has positive entries, then*

$$\lim_{n \to \infty} P(X_n^2(U) \geq t) = P(\chi^2_{(r-1)(c-1)} \geq t) \qquad \text{for all } t > 0.$$

We denote such convergence in distribution by $X_n^2(U) \xrightarrow{D} \chi^2_{(r-1)(c-1)}$.

In this proposition, the shorthand $P(\chi^2_{(r-1)(c-1)} \geq t)$ denotes the probability $P(W \geq t)$ for a random variable W that follows a chi-square distribution with $(r-1)(c-1)$ degrees of freedom. We will continue to use this notation in subsequent statements about chi-square probabilities.

Each matrix p in the independence model $\mathcal{M}_{X \perp\!\!\!\perp Y}$ corresponds to a pair of two marginal distributions for X and Y, which are in the probability simplices Δ_{r-1} and Δ_{c-1}, respectively. Therefore, the dimension of $\mathcal{M}_{X \perp\!\!\!\perp Y}$ is $(r-1)+(c-1)$. The *codimension* of $\mathcal{M}_{X \perp\!\!\!\perp Y}$ is the difference between the dimensions of the underlying probability simplex Δ_{rc-1} and the model $\mathcal{M}_{X \perp\!\!\!\perp Y}$. We see that the degrees of freedom for the limiting chi-square distribution are given by the codimension $(rc-1)-(r-1)-(c-1)=(r-1)(c-1)$.

The convergence in distribution in Proposition 1.1.6 suggests that we gauge the size of an observed value $X^2(u)$ by computing the probability

$$P(\chi^2_{(r-1)(c-1)} \geq X^2(u)), \tag{1.1.7}$$

which is referred to as the *p-value* for the chi-square test of independence.

Example 1.1.7. For the death penalty example in Table 1.1.1, $r = c = 2$ and the degrees of freedom are $(r-1)(c-1) = 1$. The p-value in (1.1.7) can be computed using the following piece of code for the statistical software `R` [75]:

```
> u = matrix(c(19,17,141,149),2,2)
> chisq.test(u,correct=FALSE)

        Pearson's Chi-squared test

data:  u
X-squared = 0.2214, df = 1, p-value = 0.638
```

The p-value being large, there is no evidence against the independence model. □

We next present an alternative approach to the testing problem (1.1.5). This approach is *exact* in that it avoids asymptotic considerations.

Fisher's exact test. We now consider 2×2-contingency tables. In this case, the distribution of U loses its dependence on the unknown joint distribution p when we condition on the row and column totals.

Proposition 1.1.8. *Suppose $r = c = 2$. If $p = (p_{ij}) \in \mathcal{M}_{X \perp\!\!\!\perp Y}$ and $u \in \mathcal{T}(n)$, then the conditional distribution of U_{11} given $U_{1+} = u_{1+}$ and $U_{+1} = u_{+1}$ is the* hypergeometric distribution *HypGeo(n, u_{1+}, u_{+1}), that is, the probability*

$$P(U_{11} = u_{11} \,|\, U_{1+} = u_{1+}, U_{+1} = u_{+1}) \;=\; \frac{\binom{u_{1+}}{u_{11}}\binom{n-u_{1+}}{u_{+1}-u_{11}}}{\binom{n}{u_{+1}}}$$

for $u_{11} \in \{\max(0, u_{1+}+u_{+1}-n), \ldots, \min(u_{1+}, u_{+1})\}$ and zero otherwise.

Proof. Fix u_{1+} and u_{+1}. Then, as a function of u_{11}, the conditional probability in question is proportional to the joint probability

$$\begin{aligned} P(U_{11} = u_{11}, U_{1+} = u_{1+}, U_{+1} = u_{+1}) \;=\; & P(U_{11} = u_{11}, U_{12} = u_{1+} - u_{11}, \\ & U_{21} = u_{+1} - u_{11}, U_{22} = n - u_{1+} - u_{+1} + u_{11}). \end{aligned}$$

By Proposition 1.1.4 and after some simplification, this probability equals

$$\binom{n}{u_{1+}}\binom{u_{1+}}{u_{11}}\binom{n-u_{1+}}{u_{+1}-u_{11}}p_{1+}^{u_{1+}}p_{2+}^{n-u_{1+}}p_{+1}^{u_{+1}}p_{+2}^{n-u_{+1}}.$$

Removing factors that do not depend on u_{11}, we see that this is proportional to

$$\binom{u_{1+}}{u_{11}}\binom{n-u_{1+}}{u_{+1}-u_{11}}.$$

Evaluating the normalizing constant using the binomial identity

$$\sum_{u_{11}}\binom{u_{1+}}{u_{11}}\binom{n-u_{1+}}{u_{+1}-u_{11}}=\binom{n}{u_{+1}}$$

yields the claim. □

Suppose $u \in \mathcal{T}(n)$ is an observed 2×2-contingency table. Proposition 1.1.8 suggests to base the rejection of H_0 in (1.1.5) on the (conditional) p-value

$$P(X^2(U) \geq X^2(u) \,|\, U_{1+} = u_{1+}, U_{+1} = u_{+1}). \tag{1.1.8}$$

This leads to the test known as Fisher's exact test. The computation of the p-value in (1.1.8) amounts to summing the hypergeometric probabilities

$$\frac{\binom{u_{1+}}{v_{11}}\binom{n-u_{1+}}{u_{+1}-v_{11}}}{\binom{n}{u_{+1}}},$$

over all values $v_{11} \in \{\max(0, u_{1+}+u_{+1}-n), \ldots, \min(u_{1+}, u_{+1})\}$ such that the chi-square statistic for the table with entries v_{11} and $v_{12} = u_{1+} - v_{11}$, $v_{21} = u_{+1} - v_{11}$, $v_{22} = n - u_{1+} - u_{+1} + v_{11}$ is greater than or equal to $X^2(u)$, the chi-square statistic value for the observed table.

Fisher's exact test can be based on criteria other than the chi-square statistic. For instance, one could compare a random table U to the observed table u by calculating which of U_{11} and u_{11} is more likely to occur under the hypergeometric distribution from Proposition 1.1.8. The R command `fisher.test(u)` in fact computes the test in this latter form, which can be shown to have optimality properties that we will not detail here. A discussion of the differences of the two criteria for comparing the random table U with the data u can be found in [28].

As presented above, Fisher's exact test applies only to 2×2-contingency tables but the key idea formalized in Proposition 1.1.8 applies more broadly. This will be the topic of the remainder of this section.

Multi-way tables and log-linear models. Let $X_1, \ldots, X_m$ be discrete random variables with X_l taking values in $[r_l]$. Let $\mathcal{R} = \prod_{l=1}^{m}[r_l]$, and define the *joint probabilities*

$$p_i = p_{i_1 \ldots i_m} = P(X_1 = i_1, \ldots, X_m = i_m), \quad i = (i_1, \ldots, i_m) \in \mathcal{R}.$$

These form a *joint probability table* $p = (p_i \,|\, i \in \mathcal{R})$ that lies in the $\#\mathcal{R} - 1$ dimensional probability simplex $\Delta_{\mathcal{R}-1}$. (Note that, as a shorthand, we will often use $\mathcal{R}$ to represent $\#\mathcal{R}$ in superscripts and subscripts.) The interior of $\Delta_{\mathcal{R}-1}$, denoted by $\mathrm{int}(\Delta_{\mathcal{R}-1})$, consists of all strictly positive probability distributions. The following class of models provides a useful generalization of the independence model from Definition 1.1.3; this is explained in more detail in Example 1.2.1.

Definition 1.1.9. Fix a matrix $A \in \mathbb{Z}^{d\times\mathcal{R}}$ whose columns all sum to the same value. The *log-linear model* associated with A is the set of positive probability tables

$$\mathcal{M}_A = \Big\{ p = (p_i) \in \mathrm{int}(\Delta_{\mathcal{R}-1}) \ : \ \log p \in \mathrm{rowspan}(A) \Big\},$$

where $\mathrm{rowspan}(A) = \mathrm{image}(A^T)$ is the linear space spanned by the rows of A. Here $\log p$ denotes the vector whose i-th coordinate is the logarithm of the positive real number p_i. The term *toric model* was used for $\mathcal{M}_A$ in the ASCB book [73, §1.2].

Consider again a set of counts

$$U_i \ = \ \sum_{k=1}^{n} 1_{\{X_1^{(k)}=i_1,\ldots,X_m^{(k)}=i_m\}}, \quad i = (i_1,\ldots,i_m) \in \mathcal{R}, \tag{1.1.9}$$

based on a random n-sample of independent and identically distributed vectors

$$\begin{pmatrix} X_1^{(1)} \\ \vdots \\ X_m^{(1)} \end{pmatrix}, \begin{pmatrix} X_1^{(2)} \\ \vdots \\ X_m^{(2)} \end{pmatrix}, \ldots, \begin{pmatrix} X_1^{(n)} \\ \vdots \\ X_m^{(n)} \end{pmatrix}.$$

The counts U_i now form an m-way table $U = (U_i)$ in $\mathbb{N}^{\mathcal{R}}$. Let

$$\mathcal{T}(n) = \Big\{ u \in \mathbb{N}^{\mathcal{R}} \ : \ \sum_{i\in\mathcal{R}} u_i = n \Big\}.$$

Definition 1.1.10. We call the vector Au the *minimal sufficient statistics* for the model $\mathcal{M}_A$, and the set of tables

$$\mathcal{F}(u) = \big\{ v \in \mathbb{N}^{\mathcal{R}} \ : \ Av = Au \big\}$$

is called the *fiber* of a contingency table $u \in \mathcal{T}(n)$ with respect to the model $\mathcal{M}_A$.

Our definition of minimal sufficient statistics is pragmatic. In fact, sufficiency and minimal sufficiency are general statistical notions. When these are applied to the log-linear model $\mathcal{M}_A$, however, one finds that the vector Au is indeed a minimal sufficient statistic in the general sense.

Note that since the row span of A is assumed to contain the vector of 1s, the tables in the fiber $\mathcal{F}(u)$ sum to n. The next proposition highlights the special role played by the sufficient statistics and provides a generalization of Proposition 1.1.8, which drove Fisher's exact test.

Proposition 1.1.11. *If $p = e^{A^T\alpha} \in \mathcal{M}_A$ and $u \in \mathcal{T}(n)$, then*

$$P(U = u) = \frac{n!}{\prod_{i\in\mathcal{R}} u_i!}\, e^{\alpha^T(Au)},$$

and the conditional probability $P(U = u \,|\, AU = Au)$ does not depend on p.

Proof. As a generalization of Proposition 1.1.4, it holds that

$$P(U = u) = \frac{n!}{\prod_{i\in\mathcal{R}} u_i!} \prod_{i\in\mathcal{R}} p_i^{u_i} = \frac{n!}{\prod_{i\in\mathcal{R}} u_i!} \prod_{i\in\mathcal{R}} e^{(A^T\alpha)_i u_i} = \frac{n!}{\prod_{i\in\mathcal{R}} u_i!}\, e^{\alpha^T(Au)}.$$

Moreover,

$$P(U = u \,|\, AU = Au) = \frac{P(U = u)}{P(AU = Au)},$$

where

$$P(AU = Au) \;=\; \sum_{v\in\mathcal{F}(u)} \frac{n!}{\prod_{i\in\mathcal{R}} v_i!}\, e^{\alpha^T(Av)} = n!\cdot e^{\alpha^T(Au)} \sum_{v\in\mathcal{F}(u)} \left(\prod_{i\in\mathcal{R}} v_i!\right)^{-1}.$$

It follows that

$$P(U = u \,|\, AU = Au) = \frac{1/\left(\prod_{i\in\mathcal{R}} u_i!\right)}{\sum_{v\in\mathcal{F}(u)} 1/\left(\prod_{i\in\mathcal{R}} v_i!\right)}. \tag{1.1.10}$$

This expression is independent of α and hence independent of p. □

Consider the hypothesis testing problem

$$H_0 \colon p \in \mathcal{M}_A \quad \text{versus} \quad H_1 \colon p \notin \mathcal{M}_A. \tag{1.1.11}$$

Based on Proposition 1.1.11, we can generalize Fisher's exact test by computing the p-value

$$P(X^2(U) \geq X^2(u) \,|\, AU = Au). \tag{1.1.12}$$

Here

$$X^2(U) = \sum_{i\in\mathcal{R}} \frac{(U_i - \hat{u}_i)^2}{\hat{u}_i} \tag{1.1.13}$$

is the natural generalization of the *chi-square statistic* in (1.1.6). Evaluation of $X^2(U)$ requires computing the model-based expected counts $\hat{u}_i = n\hat{p}_i$, where $\hat{p}_i$ are the *maximum likelihood estimates* discussed in Section 2.1. There, it will also become clear that the estimates $\hat{p}_i$ are identical for all tables in a fiber $\mathcal{F}(u)$.

Exact computation of the p-value in (1.1.12) involves summing over all non-negative integer solutions to the system of linear equations in (1.1.10). Indeed, the p-value is equal to

$$\frac{\sum_{v\in\mathcal{F}(u)} 1_{\{X^2(v)\geq X^2(u)\}} / \left(\prod_{i\in\mathcal{R}} u_i!\right)}{\sum_{v\in\mathcal{F}(u)} 1 / \left(\prod_{i\in\mathcal{R}} v_i!\right)}.$$

In even moderately sized contingency tables, the exact evaluation of that sum can become prohibitive. However, the p-value can still be approximated using *Markov chain Monte Carlo* algorithms for sampling tables from the conditional distribution of U given $AU = Au$.

Definition 1.1.12. Let $\mathcal{M}_A$ be the log-linear model associated with a matrix A whose integer kernel we denote by $\ker_{\mathbb{Z}}(A)$. A finite subset $\mathcal{B} \subset \ker_{\mathbb{Z}}(A)$ is a *Markov basis* for $\mathcal{M}_A$ if for all $u \in \mathcal{T}(n)$ and all pairs $v, v' \in \mathcal{F}(u)$ there exists a sequence $u_1, \ldots, u_L \in \mathcal{B}$ such that

$$v' = v + \sum_{k=1}^{L} u_k \quad \text{and} \quad v + \sum_{k=1}^{l} u_k \geq 0 \quad \text{for all } l = 1, \ldots, L.$$

The elements of the Markov basis are called *moves*.

The existence and computation of Markov bases will be the subject of Sections 1.2 and 1.3. Once we have found such a Markov basis $\mathcal{B}$ for the model $\mathcal{M}_A$, we can run the following algorithm that performs a random walk on a fiber $\mathcal{F}(u)$.

Algorithm 1.1.13 (Metropolis-Hastings).
Input: *A contingency table $u \in \mathcal{T}(n)$ and a Markov basis $\mathcal{B}$ for the model $\mathcal{M}_A$.*
Output: *A sequence of chi-square statistic values $(X^2(v_t))_{t=1}^{\infty}$ for tables v_t in the fiber $\mathcal{F}(u)$.*
Step 1: *Initialize $v_1 = u$.*
Step 2: *For $t = 1, 2, \ldots$ repeat the following steps:*

(i) *Select uniformly at random a move $u_t \in \mathcal{B}$.*

(ii) *If $\min(v_t + u_t) < 0$, then set $v_{t+1} = v_t$, else set*

$$v_{t+1} = \begin{cases} v_t + u_t \\ v_t \end{cases} \qquad \text{with probability} \qquad \begin{cases} q \\ 1-q \end{cases},$$

where

$$q = \min\left\{1, \frac{P(U = v_t + u_t \,|\, AU = Au)}{P(U = v_t \,|\, AU = Au)}\right\}.$$

(iii) *Compute $X^2(v_t)$.*

An important feature of the Metropolis-Hasting algorithm is that the probability q in Step 2(ii) is defined as a ratio of two conditional probabilities. Therefore, we never need to evaluate the sum in the denominator in (1.1.10).

Theorem 1.1.14. *The output* $(X^2(v_t))_{t=1}^{\infty}$ *of Algorithm* 1.1.13 *is an aperiodic, reversible and irreducible Markov chain that has stationary distribution equal to the conditional distribution of* $X^2(U)$ *given* $AU = Au$.

A proof of this theorem can be found, for example, in [33, Lemma 2.1] or [78, Chapter 6]. It is clear that selecting the proposed moves u_t from a Markov basis ensures the irreducibility (or connectedness) of the Markov chain. The following corollary clarifies in which sense Algorithm 1.1.13 computes the p-value in (1.1.12).

Corollary 1.1.15. *With probability* 1, *the output sequence* $(X^2(v_t))_{t=1}^{\infty}$ *of Algorithm* 1.1.13 *satisfies*

$$\lim_{M\to\infty} \frac{1}{M} \sum_{t=1}^{M} 1_{\{X^2(v_t)\geq X^2(u)\}} \;=\; P(X^2(U) \geq X^2(u) \,|\, AU = Au).$$

A proof of this law of large numbers can be found in [78, Chapter 6], where heuristic guidelines for deciding how long to run Algorithm 1.1.13 are also given; compare [78, Chapter 8]. Algorithm 1.1.13 is only the most basic scheme for sampling tables from a fiber. Instead one could also apply a feasible multiple of a selected Markov basis move. As discussed in [33], this will generally lead to a better mixing behavior of the constructed Markov chain. However, few theoretical results are known about the mixing times of these algorithms in the case of hypergeometric distributions on fibers of contingency tables considered here.

1.2 Markov Bases of Hierarchical Models

Continuing our discussion in Section 1.1, with each matrix $A \in \mathbb{Z}^{d\times\mathcal{R}}$ we associate a log-linear model $\mathcal{M}_A$. This is the set of probability distributions

$$\mathcal{M}_A \;=\; \{\, p \in \operatorname{int}(\Delta_{\mathcal{R}-1}) \;:\; \log p \in \operatorname{rowspan}(A)\}.$$

We assume throughout that the sum of the entries in each column of the matrix A is a fixed value.

This section introduces the class of hierarchical log-linear models and describes known results about their Markov bases. Recall that a Markov basis is a special spanning set of the lattice $\ker_{\mathbb{Z}} A$, the integral kernel of A. The Markov basis can be used to perform irreducible random walks over the fibers $\mathcal{F}(u)$.

By a *lattice* we mean a subgroup of the additive group $\mathbb{Z}^{\mathcal{R}}$. Markov bases, and other types of bases, for general lattices will be discussed in Section 1.3. Often we will interchangeably speak of the Markov basis for $\mathcal{M}_A$, the Markov basis for the matrix A, or the Markov basis for the lattice $\ker_{\mathbb{Z}} A := \ker A \cap \mathbb{Z}^{\mathcal{R}}$. These three expressions mean the same thing, and the particular usage depends on the context. Before describing these objects for general hierarchical models, we will first focus on the motivating example from the previous section, namely, the model of independence. This is a special instance of a hierarchical model.

Example 1.2.1 (Independence). An $r \times c$ probability table $p = (p_{ij})$ is in the independence model $\mathcal{M}_{X \perp\!\!\!\perp Y}$ if and only if each p_{ij} factors into the product of the marginal probabilities p_{i+} and p_{+j}. If p has all positive entries, then

$$\log p_{ij} = \log p_{i+} + \log p_{+j}, \quad i \in [r],\ j \in [c]. \tag{1.2.1}$$

For a concrete example, suppose that $r = 2$ and $c = 3$. Then $\log p$ is a 2×3 matrix, but we write this matrix as a vector with six coordinates. Then (1.2.1) states that the vector $\log p$ lies in the row span of the matrix

$$A = \begin{array}{c} \begin{matrix} 11 & 12 & 13 & 21 & 22 & 23 \end{matrix} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

We see that the positive part of the independence model is equal to the log-linear model $\mathcal{M}_A$. For general table dimensions, A is an $(r + c) \times rc$ matrix.

Let u be an $r \times c$ table, which we again think of in "vectorized" format. The matrix A that represents the model of independence is determined by the identity

$$Au = \begin{pmatrix} u_{\cdot +} \\ u_{+ \cdot} \end{pmatrix},$$

where $u_{\cdot +}$ and $u_{+ \cdot}$ are the vectors of row and columns sums of the table u. In the particular instance of $r = 2$ and $c = 3$, the above identity reads

$$Au = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \\ u_{13} \\ u_{21} \\ u_{22} \\ u_{23} \end{pmatrix} = \begin{pmatrix} u_{1+} \\ u_{2+} \\ u_{+1} \\ u_{+2} \\ u_{+3} \end{pmatrix}.$$

The moves to perform the random walk in Fisher's exact test of independence are drawn from the lattice

$$\ker_{\mathbb{Z}} A = \left\{ v \in \mathbb{Z}^{r \times c} \ : \ \sum_{k=1}^{r} v_{kj} = 0 \text{ for all } j, \text{ and } \sum_{k=1}^{c} v_{ik} = 0 \text{ for all } i \right\},$$

which consists of all $r \times c$ integer tables whose row and column sums are zero. □

For the standard model of independence of two discrete random variables, the lattice $\ker_{\mathbb{Z}} A$ contains a collection of obvious small vectors. In the Markov basis literature, these moves are often known as *basic moves*. Let e_{ij} denote the standard unit table, which has a 1 in the (i, j) position, and zeroes elsewhere. If u is a vector or matrix, then $\|u\|_1 = \sum_{i=1}^{\mathcal{R}} |u_i|$ denotes the 1-norm of u.

Proposition 1.2.2. *The unique minimal Markov basis for the independence model $\mathcal{M}_{X \perp\!\!\!\perp Y}$ consists of the following $2 \cdot \binom{r}{2}\binom{c}{2}$ moves, each having 1-norm 4:*

$$\mathcal{B} \;=\; \big\{\pm(e_{ij} + e_{kl} - e_{il} - e_{kj}) \;:\; 1 \le i < k \le r, 1 \le j < l \le c\big\}.$$

Proof. Let $u \neq v$ be two non-negative integral tables that have the same row and column sums. It suffices to show that there is an element $b \in \mathcal{B}$, such that $u+b \geq 0$ and $\|u - v\|_1 > \|u + b - v\|_1$, because this implies that we can use elements of $\mathcal{B}$ to bring points in the same fiber closer to one another. Since u and v are not equal and $Au = Av$, there is at least one positive entry in $u - v$. Without loss of generality, we may suppose $u_{11} - v_{11} > 0$. Since $u - v \in \ker_{\mathbb{Z}} A$, there is an entry in the first row of $u - v$ that is negative, say $u_{12} - v_{12} < 0$. By a similar argument $u_{22} - v_{22} > 0$. But this implies that we can take $b = e_{12} + e_{21} - e_{11} - e_{22}$ which attains $\|u - v\|_1 > \|u + b - v\|_1$ and $u + b \geq 0$ as desired.

The Markov basis $\mathcal{B}$ is minimal because, if one of the elements of $\mathcal{B}$ is omitted, the fiber which contains its positive and negative parts will be disconnected. That this minimal Markov basis is unique is a consequence of the characterization of (non)uniqueness of Markov bases in Theorem 1.3.2 below. □

As preparation for more complex log-linear models, we mention that it is often useful to use a unary representation for the Markov basis elements. That is, we can write a Markov basis element by recording, with multiplicities, the indices of the non-zero entries that appear. This notation is called *tableau notation*.

Example 1.2.3. The tableau notation for the moves in the Markov basis of the independence model is

$$\begin{bmatrix} i & j \\ k & l \end{bmatrix} - \begin{bmatrix} i & l \\ k & j \end{bmatrix},$$

which corresponds to exchanging $e_{ij} + e_{kl}$ with $e_{il} + e_{kj}$. For the move $e_{11} + e_{12} - 2e_{13} - e_{21} - e_{22} + 2e_{23}$, which arises in Exercise 6.1, the tableau notation is

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 2 & 3 \\ 2 & 3 \end{bmatrix} - \begin{bmatrix} 1 & 3 \\ 1 & 3 \\ 2 & 1 \\ 2 & 2 \end{bmatrix}.$$

Note that the indices 13 and 23 are both repeated twice, since e_{13} and e_{23} both appear with multiplicity 2 in the move. □

Among the most important classes of log-linear models are the hierarchical log-linear models. In these models, interactions between random variables are encoded by a simplicial complex, whose vertices correspond to the random variables, and whose faces correspond to interaction factors that are also known as *potential functions*. The independence model, discussed above, is the most basic instance of a hierarchical model. We denote the power set of $[m]$ by $2^{[m]}$.

Definition 1.2.4. A *simplicial complex* is a set $\Gamma \subseteq 2^{[m]}$ such that $F \in \Gamma$ and $S \subset F$ implies that $S \in \Gamma$. The elements of Γ are called *faces* of Γ and the inclusion-maximal faces are the *facets* of Γ.

To describe a simplicial complex we need only list its facets. We will use the bracket notation from the theory of hierarchical log-linear models [21]. For instance $\Gamma = [12][13][23]$ is the bracket notation for the simplicial complex

$$\Gamma = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}\}.$$

As described above, a log-linear model is defined by a non-negative integer matrix A, and the model $\mathcal{M}_A$ consists of all probability distributions whose coordinatewise logarithm lies in the row span of A. If $\log p \in \text{rowspan}(A)$, there is an $\alpha \in \mathbb{R}^d$ such that $\log p = A^T\alpha$. Exponentiating, we have $p = \exp(A^T\alpha)$. It is natural to use this expression as a parametrization for the set of all probability distributions lying in the model, in which case we must introduce a normalizing constant $Z(\alpha)$ to guarantee that we get a probability distribution:

$$p \;=\; \frac{1}{Z(\alpha)} \exp(A^T\alpha).$$

We can make things simpler and more algebraic by avoiding the exponential notation. Instead, we will often use the equivalent *monomial notation* when writing the parametrization of a log-linear model. Indeed, setting $\theta_i = \exp(\alpha_i)$, we have

$$p_j \;=\; P(X=j) \;=\; \frac{1}{Z(\theta)} \cdot \prod_{i=1}^{d} \theta_i^{a_{ij}} \tag{1.2.2}$$

where $A = (a_{ij})$. This monomial expression can be further abbreviated as $\theta^{a_j} = \prod_{i=1}^d \theta_i^{a_{ij}}$ where a_j denotes the jth column of A.

The definition of log-linear models depends on first specifying a matrix $A = (a_{ij})$, and then describing a family of probability distributions via the parametrization (1.2.2). For many log-linear models, however, it is easiest to give the monomial parametrization first, and then recover the matrix A and the sufficient statistics. In particular, this is true for the family of hierarchical log-linear models.

We use the following convention for writing subindices. If $i = (i_1, \ldots, i_m) \in \mathcal{R}$ and $F = \{f_1, f_2, \ldots\} \subseteq [m]$ then $i_F = (i_{f_1}, i_{f_2}, \ldots)$. For each subset $F \subseteq [m]$, the random vector $X_F = (X_f)_{f\in F}$ has the state space $\mathcal{R}_F = \prod_{f\in F}[r_f]$.

Definition 1.2.5. Let $\Gamma \subseteq 2^{[m]}$ be a simplicial complex and let $r_1, \ldots, r_m \in \mathbb{N}$. For each facet $F \in \Gamma$, we introduce a set of $\#\mathcal{R}_F$ positive parameters $\theta_{i_F}^{(F)}$. The *hierarchical log-linear model* associated with Γ is the set of all probability distributions

$$\mathcal{M}_\Gamma = \left\{ p \in \Delta_{\mathcal{R}-1} \;:\; p_i = \frac{1}{Z(\theta)} \prod_{F\in \text{facet}(\Gamma)} \theta_{i_F}^{(F)} \text{ for all } i \in \mathcal{R} \right\}, \tag{1.2.3}$$

where $Z(\theta)$ is the normalizing constant (or partition function)

$$Z(\theta) = \sum_{i \in \mathcal{R}} \prod_{F \in \text{facet}(\Gamma)} \theta^{(F)}_{i_F}.$$

Example 1.2.6 (Independence). Let $\Gamma = [1][2]$. Then the hierarchical model consists of all positive probability matrices $(p_{i_1 i_2})$

$$p_{i_1 i_2} = \frac{1}{Z(\theta)} \theta^{(1)}_{i_1} \theta^{(2)}_{i_2}$$

where $\theta^{(j)} \in (0,\infty)^{r_j}$, $j = 1, 2$. That is, the model consists of all positive rank 1 matrices. It is the positive part of the model of independence $\mathcal{M}_{X \perp\!\!\!\perp Y}$, or in algebraic geometric language, the positive part of the Segre variety. □

Example 1.2.7 (No 3-way interaction). Let $\Gamma = [12][13][23]$ be the boundary of a triangle. The hierarchical model $\mathcal{M}_\Gamma$ consists of all $r_1 \times r_2 \times r_3$ tables $(p_{i_1 i_2 i_3})$ with

$$p_{i_1 i_2 i_3} = \frac{1}{Z(\theta)} \theta^{(12)}_{i_1 i_2} \theta^{(13)}_{i_1 i_3} \theta^{(23)}_{i_2 i_3}$$

for some positive real tables $\theta^{(12)} \in (0,\infty)^{r_1 \times r_2}, \theta^{(13)} \in (0,\infty)^{r_1 \times r_3}$, and $\theta^{(23)} \in (0,\infty)^{r_2 \times r_3}$. Unlike the case of the model of independence, this important statistical model does not have a correspondence with any classically studied algebraic variety. In the case of binary random variables, its implicit representation is the equation

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

That is, the log-linear model consists of all positive probability distributions that satisfy this quartic equation. Implicit representations for log-linear models will be explained in detail in Section 1.3, and a general discussion of implicit representations will appear in Section 2.2. □

Example 1.2.8 (Something more general). Let $\Gamma = [12][23][345]$. The hierarchical model $\mathcal{M}_\Gamma$ consists of all $r_1 \times r_2 \times r_3 \times r_4 \times r_5$ probability tensors $(p_{i_1 i_2 i_3 i_4 i_5})$ with

$$p_{i_1 i_2 i_3 i_4 i_5} = \frac{1}{Z(\theta)} \theta^{(12)}_{i_1 i_2} \theta^{(23)}_{i_2 i_3} \theta^{(345)}_{i_3 i_4 i_5},$$

for some positive real tables $\theta^{(12)} \in (0,\infty)^{r_1 \times r_2}, \theta^{(23)} \in (0,\infty)^{r_2 \times r_3}$, and $\theta^{(345)} \in (0,\infty)^{r_3 \times r_4 \times r_5}$. These tables of parameters represent the potential functions. □

To begin to understand the Markov bases of hierarchical models, we must come to terms with the 0/1 matrices A_Γ that realize these models in the form $\mathcal{M}_{A_\Gamma}$. In particular, we must determine what linear transformation the matrix A_Γ represents. Let $u \in \mathbb{N}^{\mathcal{R}}$ be an $r_1 \times \cdots \times r_m$ contingency table. For any subset $F = \{f_1, f_2, \ldots\} \subseteq [m]$, let $u|_F$ be the $r_{f_1} \times r_{f_2} \times \cdots$ marginal table such that $(u|_F)_{i_F} = \sum_{j \in \mathcal{R}_{[m] \setminus F}} u_{i_F, j}$. The table $u|_F$ is called the *F-marginal* of u.

Proposition 1.2.9. *Let $\Gamma = [F_1][F_2]\cdots$. The matrix A_Γ represents the linear map*

$$u \mapsto (u|_{F_1}, u|_{F_2}, \ldots),$$

and the Γ-marginals are minimal sufficient statistics of the hierarchical model $\mathcal{M}_\Gamma$.

Proof. We can read the matrix A_Γ off the parametrization. In the parametrization, the rows of A_Γ correspond to parameters, and the columns correspond to states. The rows come in blocks that correspond to the facets F of Γ. Each block has cardinality $\#\mathcal{R}_F$. Hence, the rows of A_Γ are indexed by pairs (F, i_F) where F is a facet of Γ and $i_F \in \mathcal{R}_F$. The columns of A_Γ are indexed by all elements of $\mathcal{R}$. The entry in A_Γ for row index (F, i_F) and column index $j \in \mathcal{R}$ equals 1 if $j_F = i_F$ and equals zero otherwise. This description follows by reading the parametrization from (1.2.3) down the column of A_Γ that corresponds to p_j. The description of minimal sufficient statistics as marginals comes from reading this description across the rows of A_Γ, where the block corresponding to F, yields the F-marginal $u|_F$. See Definition 1.1.10. □

Example 1.2.10. Returning to our examples above, for $\Gamma = [1][2]$ corresponding to the model of independence, the minimal sufficient statistics are the row and column sums of $u \in \mathbb{N}^{r_1 \times r_2}$. Thus we have $A_{[1][2]}u = (u|_1, u|_2)$. Above, we abbreviated these row and column sums by $u_{\cdot +}$ and $u_{+\cdot}$, respectively.

For the model of no 3-way interaction, with $\Gamma = [12][13][23]$, the minimal sufficient statistics consist of all 2-way margins of the 3-way table u. That is

$$A_{[12][13][23]}u \;=\; (u|_{12}, u|_{13}, u|_{23})$$

and $A_{[12][13][23]}$ is a matrix with $r_1r_2 + r_1r_3 + r_2r_3$ rows and $r_1r_2r_3$ columns. □

As far as explicitly writing down the matrix A_Γ, this can be accomplished in a uniform way by assuming that the rows and columns are ordered lexicographically.

Example 1.2.11. Let $\Gamma = [12][14][23]$ and $r_1 = r_2 = r_3 = r_4 = 2$. Then A_Γ equals

1111	1112	1121	1122	1211	1212	1221	1222	2111	2112	2121	2122	2211	2212	2221	2222
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0
0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0
0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1
1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0	1	1	0	0	0	0
0	0	0	0	1	1	0	0	0	0	0	0	1	1	0	0
0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	1

where the rows correspond to ordering the facets of Γ in the order listed above and using the lexicographic ordering $11 > 12 > 21 > 22$ within each facet. □

Now that we know how to produce the matrices A_Γ, we can begin to compute examples of Markov bases. The program `4ti2` [57] computes a Markov basis of a lattice $\ker_{\mathbb{Z}}(A)$ taking as input either the matrix A or a spanning set for $\ker_{\mathbb{Z}} A$. By entering a spanning set as input, `4ti2` can also be used to compute Markov bases for general lattices $\mathcal{L}$ (see Section 1.3). A repository of Markov bases for a range of widely used hierarchical models is being maintained by Thomas Kahle and Johannes Rauh at `http://mbdb.mis.mpg.de/`.

Example 1.2.12. We use `4ti2` to compute the Markov basis of the no 3-way interaction model $\Gamma = [12][13][23]$, for three binary random variables $r_1 = r_2 = r_3 = 2$. The matrix representing this model has format 12×8. First, we create a file `no3way` which is the input file consisting of the size of the matrix, and the matrix itself:

```
12 8
1 1 0 0 0 0 0 0
0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0
0 0 0 0 0 0 1 1
1 0 1 0 0 0 0 0
0 1 0 1 0 0 0 0
0 0 0 0 1 0 1 0
0 0 0 0 0 1 0 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
```

The Markov basis associated to the kernel of this matrix can be computed using the command `markov no3way`, which writes its output to the file `no3way.mar`. This file is represented in matrix format as:

```
1 8
1 -1 -1 1 -1 1 1 -1
```

The code outputs the Markov basis up to sign. In this case, the Markov basis consists of two elements, the indicated $2 \times 2 \times 2$ table, and its negative. This move would be represented in tableau notation as

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 1 \\ 2 & 2 & 2 \end{bmatrix}.$$

The move corresponds to the quartic equation at the end of Example 1.2.7. □

One of the big challenges in the study of Markov bases of hierarchical models is to find descriptions of the Markov bases as the simplicial complex Γ and the numbers of states of the random variables vary. When it is not possible to give an explicit description of the Markov basis (that is, a list of all types of moves needed in the Markov basis), we might still hope to provide structural or asymptotic information about the types of moves that could arise. In the remainder of this section, we describe some results of this type.

For a simplicial complex Γ, let $\mathcal{G}(\Gamma) = \cup_{S\in\Gamma} S$ denote the *ground set* of Γ.

Definition 1.2.13. A simplicial complex Γ is *reducible*, with reducible decomposition (Γ_1, S, Γ_2) and *separator* $S \subset \mathcal{G}(\Gamma)$, if it satisfies $\Gamma = \Gamma_1 \cup \Gamma_2$ and $\Gamma_1 \cap \Gamma_2 = 2^S$. Furthermore, we here assume that neither Γ_1 nor Γ_2 is 2^S. A simplicial complex is *decomposable* if it is reducible and Γ_1 and Γ_2 are decomposable or simplices (that is, of the form 2^R for some $R \subseteq [m]$).

Of the examples we have seen so far, the simplicial complexes [1][2] and [12][23][345] are decomposable, whereas the simplicial complex [12][13][23] is not reducible. On the other hand, the complex $\Gamma = [12][13][23][345]$ is reducible but not decomposable, with reducible decomposition $([12][13][23], \{3\}, [345])$.

If a simplicial complex has a reducible decomposition, then there is naturally a large class of moves with 1-norm equal to 4 that belong to the lattice $\ker_{\mathbb{Z}} A_\Gamma$. Usually, these moves also appear in some minimal Markov basis.

Lemma 1.2.14. *If Γ is a reducible simplicial complex with reducible decomposition (Γ_1, S, Γ_2), then the following set of moves, represented in tableau notation, belongs to the lattice $\ker_{\mathbb{Z}} A_\Gamma$:*

$$\mathcal{D}(\Gamma_1,\Gamma_2) \;=\; \left\{ \begin{bmatrix} i & j & k \\ i' & j & k' \end{bmatrix} - \begin{bmatrix} i & j & k' \\ i' & j & k \end{bmatrix} \;:\; i, i' \in \mathcal{R}_{\mathcal{G}(\Gamma_1)\setminus S}, j \in \mathcal{R}_S, \right.$$
$$\left. k, k' \in \mathcal{R}_{\mathcal{G}(\Gamma_2)\setminus S} \right\}.$$

Theorem 1.2.15 (Markov bases of decomposable models [34, 93]).
If Γ is a decomposable simplicial complex, then the set of moves

$$\mathcal{B} \;=\; \bigcup_{(\Gamma_1,S,\Gamma_2)} \mathcal{D}(\Gamma_1,\Gamma_2),$$

with the union over all reducible decompositions of Γ, is a Markov basis for A_Γ.

Example 1.2.16. Consider the 4-chain $\Gamma = [12][23][34]$. This graph has two distinct reducible decompositions with minimal separators, namely $([12], \{2\}, [23][34])$ and $([12][23], \{3\}, [34])$. Therefore, the Markov basis consists of moves of the two types $\mathcal{D}([12], [23][34])$ and $\mathcal{D}([12][23], [34])$, which in tableau notation look like

$$\begin{bmatrix} i_1 & j & i_3 & i_4 \\ i_1' & j & i_3' & i_4' \end{bmatrix} - \begin{bmatrix} i_1 & j & i_3' & i_4' \\ i_1' & j & i_3 & i_4 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} i_1 & i_2 & j & i_4 \\ i_1' & i_2' & j & i_4' \end{bmatrix} - \begin{bmatrix} i_1 & i_2 & j & i_4' \\ i_1' & i_2' & j & i_4 \end{bmatrix}.$$

Note that the decomposition ([12][23], {2, 3}, [23][34]) is also a valid reducible decomposition of Γ, but it does not produce any new Markov basis elements. □

Theorem 1.2.15 is a special case of a more general result which determines the Markov bases for reducible complexes Γ from the Markov bases of the pieces Γ_1 and Γ_2. For details see the articles [35, 59].

One of the remarkable consequences of Theorem 1.2.15 is that the structure of the Markov basis of a decomposable hierarchical log-linear model does not depend on the number of states of the underlying random variables. In particular, regardless of the sizes $r_1, r_2, \ldots, r_m$, the Markov basis for a decomposable model always consists of moves with 1-norm equal to 4, with a precise and global combinatorial description. The following theorem of De Loera and Onn [29] says that this nice behavior fails, in the worst possible way, already for the simplest non-decomposable model. We fix $\Gamma = [12][13][23]$ and consider $3 \times r_2 \times r_3$ tables, where r_2, r_3 can be arbitrary. De Loera and Onn refer to these as *slim tables*.

Theorem 1.2.17 (Slim tables). *Let $\Gamma = [12][13][23]$ be the 3-cycle and let $v \in \mathbb{Z}^k$ be any integer vector. Then there exist $r_2, r_3 \in \mathbb{N}$ and a coordinate projection $\pi : \mathbb{Z}^{3 \times r_2 \times r_3} \rightarrow \mathbb{Z}^k$ such that every minimal Markov basis for Γ on $3 \times r_2 \times r_3$ tables contains a vector u such that $\pi(u) = v$.*

In particular, Theorem 1.2.17 shows that there is no hope for a general bound on the 1-norms of Markov basis elements for non-decomposable models, even for a fixed simplicial complex Γ. On the other hand, if only one of the table dimensions is allowed to vary, then there is a bounded finite structure to the Markov bases. This theorem was first proven in [62] and generalizes a result in [81].

Theorem 1.2.18 (Long tables). *Let Γ be a simplicial complex and fix $r_2, \ldots, r_m$. There exists a number $b(\Gamma, r_2, \ldots, r_m) < \infty$ such that the 1-norms of the elements of any minimal Markov basis for Γ on $s \times r_2 \times \cdots \times r_m$ tables are less than or equal to $b(\Gamma, r_2, \ldots, r_m)$. This bound is independent of s, which can grow large.*

From Theorem 1.2.15, we saw that if Γ is decomposable and not a simplex, then $b(\Gamma, r_2, \ldots, r_m) = 4$. One of the first discovered results in the non-decomposable case was $b([12][13][23], 3, 3) = 20$, a result obtained by Aoki and Takemura [10]. In general, it seems a difficult problem to actually compute the values $b(\Gamma, r_2, \ldots, r_m)$, although some recent progress was reported by Hemmecke and Nairn [58]. The proof of Theorem 1.2.18 only gives a theoretical upper bound on this quantity, involving other numbers that are also difficult to compute.

1.3 The Many Bases of an Integer Lattice

The goal of this section is to study the notion of a Markov basis in more combinatorial and algebraic detail. In particular, we will explain the relationships between Markov bases and other classical notions of a basis of an integral lattice. In the setting of log-linear models and hierarchical models, this integral lattice would be

$\ker_{\mathbb{Z}}(A)$ as in Definition 1.1.12. One of the highlights of this section is Theorem 1.3.6 which makes a connection between Markov bases and commutative algebra.

We fix any sublattice $\mathcal{L}$ of $\mathbb{Z}^k$ with the property that the only non-negative vector in $\mathcal{L}$ is the origin. In other words, $\mathcal{L}$ is a subgroup of $(\mathbb{Z}^k, +)$ that satisfies

$$\mathcal{L} \cap \mathbb{N}^k \quad = \quad \{0\}.$$

This hypothesis holds for a lattice $\ker_{\mathbb{Z}}(A)$ given by a non-negative integer matrix A, as encountered in the previous sections, and it ensures that the fiber of any point $u \in \mathbb{N}^k$ is a finite set. Here, by the *fiber* of u we mean the set of all non-negative vectors in the same residue class modulo $\mathcal{L}$. This set is denoted by

$$\mathcal{F}(u) \quad := \quad (u + \mathcal{L}) \cap \mathbb{N}^k \quad = \quad \bigl\{ v \in \mathbb{N}^k \,:\, u - v \in \mathcal{L} \bigr\}.$$

There are four fundamental problems concerning the fibers: counting $\mathcal{F}(u)$, enumerating $\mathcal{F}(u)$, optimizing over $\mathcal{F}(u)$ and sampling from $\mathcal{F}(u)$.

The optimization problem is the *integer programming problem in lattice form*:

$$\text{minimize } w \cdot v \quad \text{subject to} \quad v \in \mathcal{F}(u). \tag{1.3.1}$$

The sampling problem asks for a random point from $\mathcal{F}(u)$, drawn according to some distribution on $\mathcal{F}(u)$. As seen in Section 1.1, the ability to sample from the hypergeometric distribution is needed for hypothesis testing, but sometimes the uniform distribution is also used [32].

These four problems can be solved if we are able to perform (random) walks that connect the fibers $\mathcal{F}(u)$ using simple steps from the lattice $\mathcal{L}$. To this end, we shall introduce a hierarchy of finite bases in $\mathcal{L}$. The hierarchy looks like this:

$$\begin{aligned} \text{lattice basis} \;\subset\; \text{Markov basis} \;\subset\; \text{Gröbner basis} \\ \subset\; \text{universal Gröbner basis} \;\subset\; \text{Graver basis}. \end{aligned}$$

The purpose of this section is to introduce these five concepts. The formal definitions will be given after the next example. Example 1.3.1 serves as a warm-up, and it shows that all four inclusions among the five different bases can be strict.

Example 1.3.1. Let $k = 4$ and consider the three-dimensional lattice

$$\mathcal{L} \quad = \quad \bigl\{ (u_1, u_2, u_3, u_4) \in \mathbb{Z}^4 \,:\, 3u_1 + 3u_2 + 4u_3 + 5u_4 \,=\, 0 \bigr\}.$$

The following three vectors form a *lattice basis* of $\mathcal{L}$:

$$(1, -1, 0, 0),\ (0, 1, -2, 1),\ (0, 3, -1, -1). \tag{1.3.2}$$

The choice of a lattice basis is not unique, but its cardinality 3 is an invariant of the lattice. Augmenting (1.3.2) by the next vector gives a *Markov basis* of $\mathcal{L}$:

$$(0, 2, 1, -2). \tag{1.3.3}$$

The Markov basis of $\mathcal{L}$ is not unique but it is "more unique" than a lattice basis. The cardinality 4 of the minimal Markov basis is an invariant of the lattice. Augmenting (1.3.2) and (1.3.3) by the following two vectors leads to a *Gröbner basis* of $\mathcal{L}$:

$$(0,1,3,-3),\ (0,0,5,-4). \tag{1.3.4}$$

This Gröbner basis is *reduced*. The reduced Gröbner basis of a lattice is not unique, but there are only finitely many distinct reduced Gröbner bases. They depend on the choice of a cost vector. Here we took $w = (100, 10, 1, 0)$. This choice ensures that the leftmost non-zero entry in each of our vectors is positive. We note that the cardinality of a reduced Gröbner basis is *not* an invariant of the lattice $\mathcal{L}$.

The *universal Gröbner basis* of a lattice is unique (if we identify each vector with its negative). The universal Gröbner basis of $\mathcal{L}$ consists of 14 vectors. In addition to the six above, it comprises the eight vectors

$$(1,0,-2,1),\ (3,0,-1,-1),\ (2,0,1,-2),\ (1,0,3,-3),$$
$$(0,4,-3,0),\ (4,0,-3,0),\ (0,5,0,-3),\ (5,0,0,-3).$$

Besides the 14 vectors in the universal Gröbner basis, the *Graver basis* of $\mathcal{L}$ contains the following additional ten vectors:

$$(1,1,1,-2)\,,\ (1,2,-1,-1)\,,\ (2,1,-1,-1)\,,\ (1,3,-3,0)\,,\ (2,2,-3,0)\,,$$
$$(3,1,-3,0)\,,\ (1,4,0,-3)\,,\ (2,3,0,-3)\,,\ (3,2,0,-3)\,,\ (4,1,0,-3).$$

The Graver basis of a lattice is unique (up to negating vectors). □

We shall now give precise definitions for the five notions in our hierarchy of bases for an integer lattice $\mathcal{L} \subset \mathbb{Z}^k$. A *lattice basis* is a subset $\mathcal{B} = \{b_1, b_2, \ldots, b_r\}$ of $\mathcal{L}$ such that every vector v in $\mathcal{L}$ has a unique representation

$$v \quad = \quad \lambda_1 b_1 + \lambda_2 b_2 + \cdots + \lambda_r b_r, \quad \text{with } \lambda_i \in \mathbb{Z}.$$

All lattice bases of $\mathcal{L}$ have the same cardinality r. Each of them specifies a particular isomorphism $\mathcal{L} \simeq \mathbb{Z}^r$. The number r is the *rank* of the lattice $\mathcal{L}$.

Consider an arbitrary finite subset $\mathcal{B}$ of $\mathcal{L}$. This subset determines an undirected graph $\mathcal{F}(u)_{\mathcal{B}}$ whose nodes are the elements in the fiber $\mathcal{F}(u)$. Two nodes v and v' are connected by an undirected edge in $\mathcal{F}(u)_{\mathcal{B}}$ if either $v - v'$ or $v' - v$ is in $\mathcal{B}$. We say that $\mathcal{B}$ is a *Markov basis* for $\mathcal{L}$ if the graphs $\mathcal{F}(u)_{\mathcal{B}}$ are connected for all $u \in \mathbb{N}^k$. (Note that this definition slightly differs from the one used in Sections 1.1 and 1.2, where it was more convenient to include both a vector and its negative in the Markov basis.) We will usually require Markov bases to be minimal with respect to inclusion. With this minimality assumption, *the Markov basis* $\mathcal{B}$ is essentially unique, in the sense made precise in Theorem 1.3.2 below.

Every vector $b \in \mathcal{L}$ can be written uniquely as the difference $b = b^+ - b^-$ of two non-negative vectors with disjoint support. The *fiber of* b is the congruence class of $\mathbb{N}^k$ modulo $\mathcal{L}$ which contains both b^+ and b^-. In symbols,

$$\text{fiber}(b) \quad := \quad \mathcal{F}(b^+) \ = \ \mathcal{F}(b^-).$$

Theorem 1.3.2. *For a minimal Markov basis $\mathcal{B}$ of a lattice $\mathcal{L}$, the multiset*

$$\{\, \text{fiber}(b) \,:\, b \in \mathcal{B} \,\} \tag{1.3.5}$$

is an invariant of the lattice $\mathcal{L} \subset \mathbb{Z}^k$ and hence so is the cardinality of $\mathcal{B}$.

Proof. We shall give an invariant characterization of the multiset (1.3.5). For any fiber $f \in \mathbb{N}^k/\mathcal{L}$ we define a graph G_f as follows. The nodes are the non-negative vectors in $\mathbb{N}^k$ which lie in the congruence class f, and two nodes u and v are connected by an edge if there exists an index i such that $u_i \neq 0$ and $v_i \neq 0$. Equivalently, $\{u, v\}$ is an edge of G_f if and only if $\text{fiber}(u - v) \neq f$.

We introduce the following multiset of fibers:

$$\{\, f \in \mathbb{N}^k/\mathcal{L} \;:\; \text{the graph } G_f \text{ is disconnected} \,\}. \tag{1.3.6}$$

The multiset structure on the underlying set is as follows. The multiplicity of f in (1.3.6) is one less than the number of connected components of the graph G_f.

We claim that the multisets (1.3.5) and (1.3.6) are equal. In proving this claim, we shall use induction on the partially ordered set (poset) $\mathbb{N}^k/\mathcal{L}$. This set inherits its poset structure from the partial order on $\mathbb{N}^k$. Namely, two fibers f and f' satisfy $f' \leq f$ if and only if there exist $u, u' \in \mathbb{N}^k$ such that

$$f = \mathcal{F}(u) \;\text{ and }\; f' = \mathcal{F}(u') \;\text{ and }\; u' \leq u \;\text{ (coordinatewise).}$$

Consider any fiber $f = \mathcal{F}(u)$ and let $C_1, \ldots, C_s$ be the connected components of G_f. Suppose that $\mathcal{B}$ is any minimal Markov basis and consider $\mathcal{B}_f = \{\, b \in \mathcal{B} : \text{fiber}(b) = f \,\}$. We will reconstruct all possible choices for $\mathcal{B}_f$. In order to prove the theorem, we must show that each of them has cardinality $s - 1$.

By induction, we may assume that $\mathcal{B}_{f'}$ has already been constructed for all fibers f' which are below f in the poset $\mathbb{N}^k/\mathcal{L}$. Let $\mathcal{B}_{<f}$ be the union of these sets $\mathcal{B}_{f'}$ where $f' < f$. The connected components of the graph $\mathcal{F}(u)_{\mathcal{B}_{<f}}$ are precisely the components $C_1, \ldots, C_s$. The reason is that any two points in the same component C_i can be connected by a sequence of moves from a smaller fiber f', but no point in C_i can be connected to a point in a different component C_j by such moves. Therefore, all the possible choices for $\mathcal{B}_f$ are obtained as follows. First we fix a spanning tree on the components $C_1, \ldots, C_s$. Second, for any edge $\{C_i, C_j\}$ in that spanning tree, we pick a pair of points $u \in C_i$ and $v \in C_j$. Finally, the desired set $\mathcal{B}_f$ consists of the resulting $s-1$ difference vectors $u-v$. This proves $\#\mathcal{B}_f = s - 1$, as desired. □

The previous proof gives a purely combinatorial algorithm which constructs the minimal Markov basis of a lattice $\mathcal{L}$. We fix a total order on the set of fibers $\mathbb{N}^k/\mathcal{L}$ which refines the natural partial order. Starting with the first fiber $f = \mathcal{F}(0) = \{0\}$ and the empty partial Markov basis $\mathcal{B}_{<0} = \emptyset$, we consider an arbitrary fiber f and the already computed partial Markov basis $\mathcal{B}_{<f}$. The steps of the algorithm are now exactly as in the proof:

1. Identify the connected components $C_1, \ldots, C_s$ of the graph G_f.
2. Pick a spanning tree on $C_1, \ldots, C_s$.
3. For any edge $\{C_i, C_j\}$ of the tree, pick points $u \in C_i$ and $v \in C_j$.
4. Define $\mathcal{B}_f$ as the set of those $s-1$ difference vectors $u - v$.
5. Move on to the next fiber (unless you are sure to be done).

Example 1.3.3. We demonstrate how this method works for the lattice in Example 1.3.1. Recall that $\mathcal{L}$ is the kernel of the linear map

$$\pi : \mathbb{Z}^4 \to \mathbb{Z}, \ (u_1, u_2, u_3, u_4) \ \mapsto \ 3u_1 + 3u_2 + 4u_3 + 5u_4.$$

The poset of fibers is a subposet of the poset of non-negative integers:

$$\mathbb{N}^4/\mathcal{L} \ = \ \pi(\mathbb{N}^4) \ = \ \{0, 3, 4, 5, 6, \ldots\} \ \subset \ \mathbb{N}.$$

The fiber 0 is trivial, so our algorithm starts with $f = 3$ and $\mathcal{B}_{<3} = \emptyset$. The graph G_3 has two connected components

$$C_1 \ = \ \big\{(1,0,0,0)\big\} \quad \text{and} \quad C_2 \ = \ \big\{(0,1,0,0)\big\},$$

so we have no choice but to take $\mathcal{B}_3 \ = \ \{\,(1,-1,0,0)\,\}$. The next steps are:

- G_4 has only one node $(0,0,1,0)$ hence $\mathcal{B}_4 = \emptyset$.
- G_5 has only one node $(0,0,0,1)$ hence $\mathcal{B}_5 = \emptyset$.
- $G_6 = \{(2,0,0,0), (1,1,0,0), (0,2,0,0)\}$ is connected hence $\mathcal{B}_6 = \emptyset$.
- $G_7 = \{(1,0,1,0), (0,1,1,0)\}$ is connected hence $\mathcal{B}_7 = \emptyset$.
- G_8 has two connected components, $C_1 = \{(1,0,0,1), (0,1,0,1)\}$ and $C_2 = \{(0,0,2,0)\}$, and we decide to take $\mathcal{B}_8 = \{(0,1,-2,1)\}$.
- G_9 has two connected components, namely $C_1 = \{(3,0,0,0), (2,1,0,0), (1,2,0,0), (0,3,0,0)\}$ and $C_2 = \{(0,0,1,1)\}$. We take $\mathcal{B}_9 = \{(0,3,-1,-1)\}$.
- G_{10} has two connected components, $C_1 = \{(2,0,1,0), (1,1,1,0), (0,2,1,0)\}$ and $C_2 = \{(0,0,0,2)\}$, and we take $\mathcal{B}_{10} = \{(0,2,1,-2)\}$.

At this stage, divine inspiration tells us that the Markov basis for $\mathcal{L}$ is already complete. So, we decide to stop and we output $\mathcal{B} = \mathcal{B}_{\leq 10}$. The multiset of Markov fibers (1.3.5) is the set $\{3, 8, 9, 10\}$, where each element has multiplicity 1. □

There are two obvious problems with this algorithm. The first is that we need a termination criterion, and the second concerns the combinatorial explosion (which becomes serious for $n - \text{rank}(\mathcal{L}) \geq 3$) of having to look at many fibers until a termination criterion kicks in. The first problem can be addressed by deriving a general bound on the sizes of the coordinates of any element in the Graver basis of

$\mathcal{L}$. Such a bound is given in [87, Theorem 4.7, p. 33]. However, a more conceptual solution for both problems can be given by recasting the Markov basis property in terms of commutative algebra [25, 87]. This will be done in Theorem 1.3.6 below.

First, however, we shall define the other three bases of $\mathcal{L}$. Fix a generic cost vector $w \in \mathbb{R}^k$. Here *generic* means that each integer program (1.3.1) has only one optimal solution. Suppose that $b \cdot w < 0$ for all $b \in \mathcal{B}$. We regard $\mathcal{F}(u)_{\mathcal{B}}$ as a directed graph by introducing a directed edge $v \rightarrow v'$ whenever $v' - v$ is in $\mathcal{B}$. In this manner, $\mathcal{F}(u)_{\mathcal{B}}$ becomes an acyclic directed graph. We say that $\mathcal{B}$ is a *Gröbner basis* of $\mathcal{L}$ if the directed graph $\mathcal{F}(u)_{\mathcal{B}}$ has a unique sink, for all $u \in \mathbb{N}^k$.

Remark 1.3.4. *If $\mathcal{B}$ is a Gröbner basis then the sink of the directed graph $\mathcal{F}(u)_{\mathcal{B}}$ is the optimal solution of the integer programming problem* (1.3.1). *For more background on the use of Gröbner bases in integer programming we refer to* [87, §5].

Among all Gröbner bases for $\mathcal{L}$ there is a distinguished *reduced Gröbner basis* which is unique when w is fixed. It consists of all vectors $b \in \mathcal{L}$ such that b^- is a sink (in its own fiber), b^+ is not a sink, but $b^+ - e_i$ is a sink for all i with $b_i > 0$.

It is known that there are only finitely many distinct reduced Gröbner bases, as w ranges over generic vectors in $\mathbb{R}^k$. The union of all reduced Gröbner bases is the *universal Gröbner basis* of $\mathcal{L}$.

All of the bases of $\mathcal{L}$ discussed so far are contained in the *Graver basis*. The Graver basis $\mathcal{G}$ of our lattice $\mathcal{L}$ is defined as follows. Fix a sign vector $\sigma \in \{-1,+1\}^k$ and consider the semigroup

$$\mathcal{L}_\sigma \quad := \quad \bigl\{ v \in \mathcal{L} \,:\, v_i \cdot \sigma_i \geq 0 \bigr\}.$$

This semigroup has a unique minimal finite generating set $\mathcal{G}_\sigma$ called the *Hilbert basis* of $\mathcal{L}_\sigma$. The *Graver basis* $\mathcal{G}$ of $\mathcal{L}$ is the union of these Hilbert bases:

$$\mathcal{G} \quad := \quad \bigcup_{\sigma \in \{-1,+1\}^k} \mathcal{G}_\sigma.$$

This set is finite because each of the Hilbert bases $\mathcal{G}_\sigma$ is finite.

Proposition 1.3.5. *The Graver basis $\mathcal{G}$ is the unique minimal subset of the lattice $\mathcal{L}$ such that every vector $v \in \mathcal{L}$ has a sign-consistent representation in terms of $\mathcal{G}$:*

$$v \;=\; \sum_{g \in \mathcal{G}} \lambda_g \cdot g \quad \text{with } \lambda_g \in \mathbb{N} \quad \text{and} \quad |v_i| \;=\; \sum_{g \in \mathcal{G}} \lambda_g \cdot |g_i| \quad \text{for all } i \in [k].$$

Markov bases, Gröbner bases, Hilbert bases, and Graver bases of integer lattices can be computed using the software `4ti2`, which was developed by Raymond Hemmecke and his collaborators [57]. Further computations with `4ti2` will be shown in the exercises in Chapter 6.

We now come to the interpretation of our bases in terms of algebraic geometry. The given lattice $\mathcal{L} \subset \mathbb{Z}^k$ is represented by the corresponding *lattice ideal*

$$I_{\mathcal{L}} \;:=\; \langle\, p^u - p^v \,:\, u, v \in \mathbb{N}^k \text{ and } u - v \in \mathcal{L} \,\rangle \quad \subset \quad \mathbb{R}[p_1, p_2, \ldots, p_k].$$

Here $p_1, \ldots, p_k$ are indeterminates, and $p^u = p_1^{u_1} p_2^{u_2} \cdots p_k^{u_k}$ denotes monomials in these indeterminates. In our applications, p_i will represent the probability of observing the ith state of a random variable with k states. Hilbert's Basis Theorem states that every ideal in the polynomial ring $\mathbb{R}[p_1, p_2, \ldots, p_k]$ is finitely generated. The finiteness of Markov bases is thus implied by the following result of [33], which was one of the starting points for the field of algebraic statistics.

Theorem 1.3.6 (Fundamental theorem of Markov bases). *A subset $\mathcal{B}$ of the lattice $\mathcal{L}$ is a Markov basis if and only if the corresponding set of binomials $\{ p^{b^+} - p^{b^-} : b \in \mathcal{B} \}$ generates the lattice ideal $I_{\mathcal{L}}$.*

The notions of Gröbner bases and Graver bases are also derived from their algebraic analogues. For a detailed account see [87]. In that book, as well as in most statistical applications, the lattice $\mathcal{L}$ arises as the kernel of an integer matrix A. The algebraic theory for arbitrary lattices is found in [70, Chapter 7]. The multiset in Theorem 1.3.2 corresponds to the multidegrees of the minimal generators of $I_{\mathcal{L}}$.

Let $A = (a_{ij}) \in \mathbb{N}^{d \times k}$ be a non-negative integer matrix. We assume that all the column sums of A are equal. The columns $a_j = (a_{1j}, a_{2j}, \ldots, a_{dj})^T$ of A represent monomials $\theta^{a_j} = \theta_1^{a_{1j}} \theta_2^{a_{2j}} \cdots \theta_d^{a_{dj}}$ in auxiliary unknowns θ_i that correspond to model parameters. The monomials θ^{a_j} all have the same degree.

The matrix A determines a monomial map

$$\phi_A \ : \ \mathbb{C}^d \ \rightarrow \ \mathbb{C}^k, \ \ \theta \ \mapsto \ (\theta^{a_1}, \theta^{a_2}, \ldots, \theta^{a_k}).$$

The closure of the image of this map is the *affine toric variety* V_A associated to the matrix A. The connection to tori arises from the fact that V_A is the closure of the image of the algebraic torus $\phi_A((\mathbb{C}^*)^d)$. If we restrict the map ϕ_A to the positive reals $\mathbb{R}^d_{>0}$, and consider the image in the probability simplex $\Delta_{k-1} = \mathbb{R}^k_{\geq 0}/\text{scaling}$, we get the log-linear model $\mathcal{M}_A$. For this reason, log-linear models are sometimes known as *toric models*. See Section 1.2 in [73] for more on toric models.

More generally, a *variety* is the solution set to a simultaneous system of polynomial equations. If I is an ideal, then $V(I)$ is the variety defined by the vanishing of all polynomials in I. Often, we might need to be more explicit about *where* the solutions to this system of equations lie, in which case we use the notation $V_*(I)$ to denote the solutions constrained by condition $*$. The different types of solution spaces will be illustrated in Example 1.3.8.

Proposition 1.3.7. *The lattice ideal $I_{\mathcal{L}}$ for $\mathcal{L} = ker_{\mathbb{Z}}(A)$ is a prime ideal. Its homogeneous elements are exactly the homogeneous polynomials in $\mathbb{R}[p_1, \ldots, p_k]$ that vanish on probability distributions in the log-linear model specified by the matrix A. In other words, the toric variety $V_A = V(I_{\mathcal{L}})$ is the* Zariski closure *of the log-linear model $\mathcal{M}_A$.*

The binomials corresponding to the Markov basis generate the ideal $I_{\mathcal{L}}$ and hence they cut out the toric variety $V_A = V(I_{\mathcal{L}})$. However, often one does not need the full Markov basis to define the toric variety set-theoretically. Finding

good choices of such partial bases is a delicate matter, as the following example demonstrates.

Example 1.3.8. Let $d = 3$, $k = 9$ and consider the matrix

$$A \;=\; \begin{array}{c} \begin{matrix} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 & p_9 \end{matrix} \\ \begin{pmatrix} 3 & 0 & 0 & 2 & 1 & 2 & 1 & 0 & 0 \\ 0 & 3 & 0 & 1 & 2 & 0 & 0 & 2 & 1 \\ 0 & 0 & 3 & 0 & 0 & 1 & 2 & 1 & 2 \end{pmatrix} \end{array} \tag{1.3.7}$$

and the associated monomial parametrization

$$\phi_A : (\theta_1, \theta_2, \theta_3) \mapsto (\theta_1^3, \theta_2^3, \theta_3^3, \theta_1^2\theta_2, \theta_1\theta_2^2, \theta_1^2\theta_3, \theta_1\theta_3^2, \theta_2^2\theta_3, \theta_2\theta_3^2). \tag{1.3.8}$$

The minimal Markov basis of the lattice $\mathcal{L} = \ker_{\mathbb{Z}}(A)$ consists of 17 vectors. These vectors correspond to the set of all 17 quadratic binomials listed in (1.3.9), (1.3.10), (1.3.11) and (1.3.12) below. We start out with the following six binomials:

$$\{\, p_1p_5 - p_4^2,\, p_2p_4 - p_5^2,\, p_1p_7 - p_6^2,\, p_3p_6 - p_7^2,\, p_2p_9 - p_8^2,\, p_3p_8 - p_9^2 \,\}. \tag{1.3.9}$$

The vectors corresponding to (1.3.9) form a basis for the kernel of A as a vector space over the rational numbers $\mathbb{Q}$ but they do not span $\mathcal{L}$ as a lattice over $\mathbb{Z}$. Nevertheless, a positive vector $p = (p_1, \ldots, p_9)$ is a common zero of these six binomials if and only if p lies in the image of a positive vector $(\theta_1, \theta_2, \theta_3)$ under the map ϕ_A. The same statement fails badly for non-negative vectors. Namely, in addition to $V_{\geq 0}(I_{\mathcal{L}})$, which is the closure of the log-linear model, the non-negative variety of (1.3.9) has seven extraneous components, which are not in the closure of the log-linear model $\mathcal{M}_A$. One such component is the three-dimensional orthant

$$\{(p_1, p_2, p_3, 0, 0, 0, 0, 0, 0) \,:\, p_1, p_2, p_3 \in \mathbb{R}_{\geq 0}\,\} \quad \subset \quad V_{\geq 0}\big((1.3.9)\big).$$

We invite the reader to find the six others. These seven extraneous components disappear again if we augment (1.3.9) by the following three binomials:

$$\{\, p_1p_2 - p_4p_5,\, p_1p_3 - p_6p_7,\, p_2p_3 - p_8p_9 \,\}. \tag{1.3.10}$$

Hence the non-negative variety defined by the nine binomials in (1.3.9) and (1.3.10) is the closure of the log-linear model. The same holds over the reals:

$$V_{\geq 0}(I_{\mathcal{L}}) \;=\; V_{\geq 0}\big((1.3.9),\, (1.3.10)\big) \quad \text{and} \quad V_{\mathbb{R}}(I_{\mathcal{L}}) \;=\; V_{\mathbb{R}}\big((1.3.9),\, (1.3.10)\big).$$

On the other hand, the varieties over the complex numbers are still different:

$$V_{\mathbb{C}}(I_{\mathcal{L}}) \;\neq\; V_{\mathbb{C}}\big((1.3.9),\, (1.3.10)\big).$$

The complex variety of the binomials in (1.3.9) and (1.3.10) breaks into three irreducible components, each of which is a multiplicative translate of the toric

variety $V_{\mathbb{C}}(I_{\mathcal{L}})$. Namely, if we start with any point p in $V_{\mathbb{C}}(I_{\mathcal{L}})$ and we replace p_4 by ηp_4 and p_5 by $\eta^2 p_5$, where $\eta = -\frac{1}{2} + \frac{\sqrt{3}}{2} i$ is a primitive cube root of unity, then the new vector is no longer in $V_{\mathbb{C}}(I_{\mathcal{L}})$ but still satisfies the nine binomials in (1.3.9) and (1.3.10). This is detected algebraically as follows. The binomial

$$p_1^3 p_8^3 - p_5^3 p_6^3 \quad = \quad (p_1 p_8 - p_5 p_6)(p_1 p_8 - \eta p_5 p_6)(p_1 p_8 - \eta^2 p_5 p_6)$$

lies in the ideal of (1.3.9) and (1.3.10) but none of its factors does. To remove the two extraneous complex components, we add six more binomials:

$$\begin{aligned} \big\{ p_1 p_8 - p_5 p_6,\, p_1 p_9 - p_4 p_7,\, p_2 p_6 - p_4 p_8,\, p_2 p_7 - p_5 p_9, \\ p_3 p_4 - p_6 p_9,\, p_3 p_5 - p_7 p_8 \big\}. \end{aligned} \tag{1.3.11}$$

Let J denote the ideal generated by the 15 binomials in (1.3.9), (1.3.10) and (1.3.11). The radical of the ideal J equals $I_{\mathcal{L}}$. This means that the complex variety of J coincides with $V_{\mathbb{C}}(I_{\mathcal{L}})$. However, the ideal J is still strictly contained in $I_{\mathcal{L}}$. To get the Markov basis, we still need to add the following two binomials:

$$\big\{ p_6 p_8 - p_4 p_9 \,,\; p_5 p_7 - p_4 p_9 \big\}. \tag{1.3.12}$$

The lattice $\mathcal{L}$ in this example has the following special property. Its Markov basis consists of quadratic binomials, but no Gröbner basis of $I_{\mathcal{L}}$ has only quadratic elements. Using the software `Gfan` [63], one can easily check that $\mathcal{L}$ has precisely $54,828$ distinct reduced Gröbner bases. Each of them contains at least one binomial of degree 3. For instance, the reduced Gröbner basis with respect to the reverse lexicographic order consists of our 17 quadrics and the two cubics $p_1 p_7 p_8 - p_4 p_6 p_9$ and $p_7^2 p_8 - p_6 p_9^2$. □

We remark that we will see quadratic binomials of the form $p_i p_j - p_k p_l$ again in Chapter 3, where they naturally correspond to conditional independence relations. The ideal of such relations will make its first appearance in Definition 3.1.5. We close the current chapter by describing a simple log-linear model in which the algebraic structure from Example 1.3.8 arises.

Example 1.3.9. Bobby and Sally play *Rock-Paper-Scissors* according to the following rules. One round consists of three games, and it is not permissible to make three different choices in one round. Should this happen then the round of three games is repeated until a valid outcome occurs. After $n = 1000$ rounds of playing, Sally decides to analyze Bobby's choices that can be summarized in the vector

$$u \quad = \quad \big(u_{rrr}, u_{ppp}, u_{sss}, u_{rrp}, u_{rpp}, u_{rrs}, u_{rss}, u_{pps}, u_{pss}\big),$$

where u_{rrr} is the number of rounds in which Bobby picks rock three times, u_{rrp} is the number of rounds in which he picks rock twice and paper once, and so on. Sally suspects that Bobby makes independent random choices picking rock

with probability θ_1, paper with probability θ_2, and scissors with probability $\theta_3 = 1 - \theta_1 - \theta_2$. Let p_{rrr}, p_{ppp}, etc. be the probabilities of Bobby's choices. Under the hypothesis of random choices, the vector of rescaled probabilities

$$(3p_{rrr}, 3p_{ppp}, 3p_{sss}, p_{rrp}, p_{rpp}, p_{rrs}, p_{rss}, p_{pps}, p_{pss})$$

is a point in the toric variety discussed in Example 1.3.8. Sally can thus use the Markov basis given there to test her hypothesis that Bobby makes random choices. All she needs to do is to run the Metropolis-Hastings Algorithm 1.1.13, and then apply the hypothesis testing framework that was outlined in Section 1.1. Note, however, that the rescaling of the probabilities leads to an adjustment of the hypergeometric distribution in (1.1.10). In this adjustment we divide the numerator of (1.1.10) by $3^{u_{rrr}+u_{ppp}+u_{sss}}$ (or multiply by $3^{u_{rrp}+u_{rpp}+u_{rrs}+u_{rss}+u_{pps}+u_{pss}}$) and apply the corresponding division (or multiplication) to each term in the sum in the denominator. □