Algebraic Statistics

Karl-Heinz Zimmermann

Algebraic Statistics

Hamburg University of Technology

Prof. Dr. Karl-Heinz Zimmermann
Hamburg University of Technology
21071 Hamburg
Germany

urn:nbn:de:gbv:830-88213690

For my Teachers

Thomas Beth†
Adalbert Kerber
Sun-Yuan Kung
Horst Müller

Preface

Algebraic statistics brings together ideas from algebraic geometry, commutative algebra, and combinatorics to address problems in statistics and its applications. Computer algebra provides powerful tools for the study of algorithms and software. However, these tools are rarely prepared to address statistical challenges and therefore new algebraic results need often be developed. This way of interplay between algebra and statistics fertilizes both disciplines.

Algebraic statistics is a relatively new branch of mathematics that developed and changed rapidly over the last ten years. The seminal work in this field was the paper of Diaconis and Sturmfels (1998) introducing the notion of Markov bases for toric statistical models and showing the connection to commutative algebra. Later on, the connection between algebra and statistics spread to a number of different areas including parametric inference, phylogenetic invariants, and algebraic tools for maximum likelihood estimation. These connection were highlighted in the celebrated book *Algebraic Statistics for Computational Biology* of Pachter and Sturmfels (2005) and subsequent publications.

In this report, statistical models for discrete data are viewed as solutions of systems of polynomial equations. This allows to treat statistical models for sequence alignment, hidden Markov models, and phylogenetic tree models. These models are connected in the sense that if they are interpreted in the tropical algebra, the famous dynamic programming algorithms (Needleman-Wunsch, Viterbi, and Felsenstein) occur in a natural manner. More generally, if the models are interpreted in a higher dimensional analogue of the tropical algebra, the polytope algebra, parametric versions of these dynamic programming algorithms can be established.

Markov bases allow to sample data in a given fibre using Markov chain Monte Carlo algorithms. In this way, Markov bases provide a means to increase the sample size and make statistical tests in inferential statistics more reliable. We will calculate Markov bases using Groebner bases in commutative polynomial rings.

The manuscript grew out of a lecture on algebraic statistics held for Master students of Computer Science at the Hamburg University of Technology. It appears that the first lecture held in the summer term 2008 was the first course of this kind in Germany. The current manuscript is the basis of a four-hour introductory course. The use of computer algebra systems is at the heart of the course. `Maple` is employed for symbolic computations, `Singular` for algebraic computations, and `R` for statistical computations. The monograph *Statistical Computing with R* from Maria L. Rizzo (2007) was an excellent source for implementing the `R` code in this book. The second and third editions are streamlined versions of the first one.

Hamburg, Jan. 2016 *Karl-Heinz Zimmermann*

Contents

Part II Algebraic Statistics

Part I

Algebraic and Combinatorial Methods

1

Commutative Algebra

Commutative algebra is a branch of abstract algebra that studies commutative rings and their ideals. Both algebraic geometry and algebraic number theory are built on commutative algebra. Ideals in polynomial rings are usually studied by their Groebner bases. The latter can be used to tackle important problems like testing the membership in ideals and solving polynomial equations.

1.1 Polynomial Rings

Let $\mathbb{K}$ be a field. A *monomial* in a collection of variables or unknowns $X_1, \ldots, X_n$ over $\mathbb{K}$ is a product

$$X^\alpha = X_1^{\alpha_1} \cdots X_n^{\alpha_n}, \quad \alpha_1, \ldots, \alpha_n \in \mathbb{N}_0. \tag{1.1}$$

The *total degree* of a monomial X^α is the sum of the exponents $|\alpha| = \alpha_1 + \ldots + \alpha_n$. For instance, $X_1^2 X_3^3 X_4$ is a monomial of total degree 6 in the variables X_1, X_2, X_3, X_4, since $\alpha = (2, 0, 3, 1)$ and $|\alpha| = 6$.

We can form linear combinations of monomials with coefficients in $\mathbb{K}$. The resulting objects are *polynomials* in $X_1, \ldots, X_n$ over $\mathbb{K}$. A general polynomial f in $X_1, \ldots, X_n$ with coefficients in $\mathbb{K}$ has the form

$$f = \sum_\alpha c_\alpha X^\alpha, \quad c_\alpha \in \mathbb{K}, \tag{1.2}$$

where the sum is over a finite number of elements $\alpha \in \mathbb{N}_0^n$. A nonzero product $c_\alpha X^\alpha$ involved in a polynomial is called a *term* and the scalar c_α is called the *coefficient* of the term. For instance, taking $\mathbb{K}$ to be the field $\mathbb{Q}$ of rational numbers and using the variables X, Y, Z instead of subscripts, $f = X^2 + YZ - 1$ is a polynomial containing three terms.

The set of all polynomials in $X_1, \ldots, X_n$ with coefficients in $\mathbb{K}$ is denoted by $\mathbb{K}[X_1, \ldots, X_n]$. The polynomials in $\mathbb{K}[X_1, \ldots, X_n]$ can be added and multiplied as usual,

$$\left(\sum_\alpha c_\alpha X^\alpha \right) + \left(\sum_\beta d_\beta X^\beta \right) = \sum_\alpha \left(c_\alpha + d_\alpha \right) X^\alpha, \tag{1.3}$$

$$\left(\sum_{\alpha} c_\alpha X^\alpha\right) \cdot \left(\sum_{\beta} d_\beta X^\beta\right) = \sum_{\alpha,\beta} (c_a d_b)\, X^{\alpha+\beta}. \tag{1.4}$$

Thus $\mathbb{K}[X_1, \ldots, X_n]$ forms a commutative ring with identity called *polynomial ring* in $X_1, \ldots, X_n$ over $\mathbb{K}$. Moreover, the addition of polynomials in $\mathbb{K}[X_1, \ldots, X_n]$ suggests that $\mathbb{K}[X_1, \ldots, X_n]$ forms an infinite-dimensional $\mathbb{K}$-vector space with the monomials as a $\mathbb{K}$-basis.

Each nonzero polynomial f in $\mathbb{K}[X_1, \ldots, X_n]$ has a *degree*, denoted by $\deg(f)$. This is the largest total degree of a monomial occurring in f with a nonzero coefficient. For instance, $f = 4X^3 + 3Y^5Z - Z^4$ is a polynomial of degree 6 in $\mathbb{Q}[X, Y, Z]$. The nonzero elements of $\mathbb{K}$ are the polynomials of degree 0. For any nonzero polynomials f and g in $\mathbb{K}[X_1, \ldots, X_n]$, we have $\deg(fg) = \deg(f) + \deg(g)$ by comparing monomials of largest degree. Thus the polynomial ring $\mathbb{K}[X_1, \ldots, X_n]$ is an integral domain (i.e., it has no zero divisors) and only nonzero constant polynomials have multiplicative inverses in $\mathbb{K}[X_1, \ldots, X_n]$. Hence, $\mathbb{K}[X_1, \ldots, X_n]$ is not a field.

A polynomial f in $\mathbb{K}[X_1, \ldots, X_n]$ is called *homogeneous* if all involved monomials have the same total degree. For instance, $f = 3X^4 + 5YZ^3 - X^2Z^2$ is a homogenous polynomial of total degree 4 in $\mathbb{Q}[X, Y, Z]$. It is clear that each polynomial f in $\mathbb{K}[X_1, \ldots, X_n]$ can be written as a sum of homogeneous polynomials called the *homogeneous components* of f. For instance, $f = 3X^4 + 5YZ^3 + X^2 - Y^2 - 1$ in $\mathbb{Q}[X, Y, Z]$ is a sum of homogeneous components $f^{(4)} = 3X^4 + 5YZ^3$, $f^{(2)} = X^2 - Y^2$, and $f^{(0)} = -1$,

Example 1.1 (Singular). Polynomial rings can be generated over different fields. The polynomial ring $\mathbb{Q}[X, Y, Z]$ is defined as

```
> ring r1 = 0, (x,y,z), dp;
> poly f = x2y-z2
> f*f-f;
x4y2-2x2yz2+z4-x2y+z2
```

Polynomials can be written in short (e.g., $x2y - z2$) or long (e.g., $x^2y - z^2$) notation. The definition of polynomial rings over other fields follows the same pattern such as the polynomial ring over the finite field $\mathbb{Z}_5$,

```
> ring r2 = 5, (x,y,z), dp;
```

the polynomial ring over the finite Galois field GF(8),

```
> ring r3 = (2^3,a), (x,y,z), dp;   // primitive element a
> number n = a2+1;                  // element of GF(8)
> n*n;
a5
```

and the polynomial ring over the extension field $\mathbb{Q}(a, b)$,

```
> ring r4 = (0,a,b), (x,y,z), dp;
> number n = 2a+1/b2;              // element of Q(a,b)
> n*n;
4a2b4+4ab2+1/(b4)
```

♢

1.2 Ideals

Ideals are the most prominent structures studied in polynomial rings.

A nonempty subset I of the polynomial ring $\mathbb{K}[X_1, \ldots, X_n]$ is an *ideal* if

- for each $f, g \in I$, we have $-f$ and $f + g \in I$, and
- for each $f \in I$ and $g \in \mathbb{K}[X_1, \ldots, X_n]$, we have $f \cdot g \in I$.

The first condition ensures that I is an additive subgroup of $\mathbb{K}[X_1, \ldots, X_n]$ and equals the *subgroup criterion* which says that for each $f, g \in I$, we have $f - g \in I$.

Lemma 1.2. *Let $f_1, \ldots, f_s$ be polynomials in $\mathbb{K}[X_1, \ldots, X_n]$, Then the set*

$$\langle f_1, \ldots, f_s \rangle = \left\{ \sum_{i=1}^{s} h_i f_i \mid h_1, \ldots, h_s \in \mathbb{K}[X_1, \ldots, X_n] \right\} \tag{1.5}$$

is an ideal of $\mathbb{K}[X_1, \ldots, X_n]$, the smallest ideal of $\mathbb{K}[X_1, \ldots, X_n]$ containing $f_1, \ldots, f_s$.

Proof. Let $f, g \in \langle f_1, \ldots, f_s \rangle$. Write $f = h_1 f_1 + \ldots + h_s f_s$ and $g = h'_1 f_1 + \ldots + h'_s f_s$, where $h_i, h'_i \in \mathbb{K}[X_1, \ldots, X_n]$, $1 \leq i \leq s$. Then $f - g = (h_1 - h'_1) f_1 + \ldots + (h_s - h'_s) f_s$ and thus $f - g \in \langle f_1, \ldots, f_s \rangle$. Moreover, if $h \in \mathbb{K}[X_1, \ldots, X_n]$ then $f \cdot h = (h_1 h) f_1 + \ldots + (h_s h) f_s$ and thus $f \cdot h \in \langle f_1, \ldots, f_s \rangle$. In view of the last assertion, note that each ideal of $\mathbb{K}[X_1, \ldots, X_n]$ that contains $f_1, \ldots, f_s$ must also contain $\langle f_1, \ldots, f_s \rangle$. □

The ideal $\langle f_1, \ldots, f_s \rangle$ is called the *ideal generated by* $f_1, \ldots, f_s$. The set $\{f_1, \ldots, f_s\}$ is sometimes called a *basis* of the ideal. In particular, the sets $\langle \emptyset \rangle = \{0\}$ and $\langle 1 \rangle = \mathbb{K}[X_1, \ldots, X_n]$ are the *trivial* ideals.

There are several ways to construct new ideals from given ones.

Proposition 1.3. *Let I and J be ideals of $\mathbb{K}[X_1, \ldots, X_n]$. The* sum *of I and J is the set*

$$I + J = \{f + g \mid f \in I, g \in J\}. \tag{1.6}$$

- *The sum $I + J$ is an ideal of $\mathbb{K}[X_1, \ldots, X_n]$.*
- *The sum $I + J$ is the smallest ideal containing $I \cup J$.*
- *If $I = \langle f_1, \ldots, f_r \rangle$ and $J = \langle g_1, \ldots, g_s \rangle$, then*

$$I + J = \langle f_1, \ldots, f_r, g_1, \ldots, g_s \rangle. \tag{1.7}$$

Proof. Let $f, f' \in I$ and $g, g' \in J$. Then $(f + g) - (f' + g') = (f - f') + (g - g')$ in $I + J$. Moreover, let $h \in \mathbb{K}[X_1, \ldots, X_n]$. Then $(f + g) \cdot h = (f \cdot h) + (g \cdot h) \in I + J$. Hence, $I + J$ is an ideal.

Let L be an ideal of $\mathbb{K}[X_1, \ldots, X_n]$ containing $I \cup J$. If $f \in I$ and $g \in J$ then $f + g \in L$ and thus L contains $I + J$.

Let $h \in \langle f_1, \ldots, f_r, g_1, \ldots, g_s \rangle$. Then $h = h_1 f_1 + \ldots + h_r f_r + h'_1 g_1 + \ldots + h'_s g_s$, where $h_i, h'_j \in \mathbb{K}[X_1, \ldots, X_n]$, $1 \leq i \leq r$, $1 \leq j \leq s$. Thus h is of the form $f + g$, where $f \in I$ and $g \in J$, and hence $h \in I + J$. Conversely, the ideal $\langle f_1, \ldots, f_r, g_1, \ldots, g_s \rangle$ contains $I \cup J$ and thus by the second assertion must be equal to $I + J$. □

Proposition 1.4. *Let I and J be ideals of $\mathbb{K}[X_1, \ldots, X_n]$. The* product *of I and J is the ideal*

$$I \cdot J = \langle f \cdot g \mid f \in I, g \in J \rangle. \tag{1.8}$$

- *The intersection $I \cap J$ is an ideal in $\mathbb{K}[X_1, \ldots, X_n]$.*
- *The product $I \cdot J$ is contained in the intersection $I \cap J$.*
- *If $I = \langle f_1, \ldots, f_r \rangle$ and $J = \langle g_1, \ldots, g_s \rangle$, then*

$$I \cdot J = \langle f_i \cdot g_j \mid 1 \leq i \leq r, 1 \leq j \leq s \rangle. \tag{1.9}$$

Proof. Let $f, g \in I \cap J$. Then $f - g \in I$ and $f - g \in J$ and so $f - g \in I \cap J$. Let $f \in I \cap J$ and $h \in \mathbb{K}[X_1, \ldots, X_n]$. Since I and J are ideals, $f \cdot h \in I$ and $f \cdot h \in J$. Thus $f \cdot h \in I \cap J$.

Let $f \in I$ and $g \in J$. Then $f \cdot g$ is contained in both I and J and thus belongs to $I \cap J$. That is, $I \cdot J \subseteq I \cap J$.

Since $f_i \cdot g_j$ belongs to $I \cdot J$ it follows that $I \cdot J$ contains $\langle f_i \cdot g_j \mid 1 \leq i \leq r, 1 \leq j \leq s \rangle$. Conversely, let $h \in I \cdot J$. Then h can be written in terms of generators $f \cdot g$, where $f \in I$ and $g \in I$. But the constituents of these generators f and g can be written with respect to the bases $f_1, \ldots, f_r$ and $g_1, \ldots, g_s$, respectively. Thus the polynomial h belongs to the ideal $\langle f_i \cdot g_j \mid 1 \leq i \leq r, 1 \leq j \leq s \rangle$. □

Example 1.5 (Singular). The above ideal operations in $\mathbb{Q}[X, Y, Z]$ can be defined as follows,

```
> ring r = 0, (x,y,z), dp;
> ideal i = xyz, x2-y2;
> ideal j = x2-1, y2-z2;
> i+j
_[1]=xyz
_[2]=x2-y2
_[3]=x2-1
_[4]=y2-z2
> i*j
_[1]=x3yz-xyz
_[2]=xy3z-xyz3
_[3]=x4-x2y2-x2+y2
_[4]=x2y2-y4-x2z2+y2z2
```

♢

Proposition 1.6. *Let I be an ideal of $\mathbb{K}[X_1, \ldots, X_n]$. The set*

$$\sqrt{I} = \{f \in \mathbb{K}[X_1, \ldots, X_n] \mid f^m \in I \text{ for some integer } m \geq 1\}. \tag{1.10}$$

is an ideal of $\mathbb{K}[X_1, \ldots, X_n]$ containing I called the radical *of I with $\sqrt{\sqrt{I}} = \sqrt{I}$.*

Proof. We have $I \subseteq \sqrt{I}$, since $f \in I$, i.e., $f^1 \in I$, implies $f \in \sqrt{I}$.

Claim that $\sqrt{I}$ is an ideal. Indeed, let $f, g \in \sqrt{I}$. By definition, there are positive integers k and l such that $f^k, gl \in I$. Expanding $(f + g)^{k+l-1}$ by the binomial theorem shows that each term is a multiple of some $f^m g^{m'}$ with $m + m' = k + l - 1$. Thus either $k \geq m$ or $l \geq m'$ and thus f^k or g^l is in I. Thus all terms in $(f + g)^{k+l-1}$ belong to I and hence $f + g$ lies in $\sqrt{I}$.

Let $f \in \sqrt{I}$ and $g \in \mathbb{K}[X_1, \ldots, X_n]$. By definition, $f^m \in I$ for some $m \geq 1$. Thus $(fg)^m = f^m g^m$ belongs to I and hence fg lies in $\sqrt{I}$. It follows that $\sqrt{I}$ is an ideal.

Claim that $\sqrt{\sqrt{I}} = \sqrt{I}$. Indeed, we have already shown that $\sqrt{I}$ lies in $\sqrt{\sqrt{I}}$. Conversely, let $f \in \sqrt{\sqrt{I}}$. Then $f^m \in \sqrt{I}$ for some positive integer m and thus $(f^m)^l \in I$ for some positive integer l. Thus $f \in \sqrt{I}$ and hence the claim follows. □

An ideal I is called *radical* if $\sqrt{I} = I$. For instance, the above assertion shows that $\sqrt{I}$ is radical.

Example 1.7 (Singular). The computation of the radical of an ideal requires the loading of a library.

```
> LIB "primdec.lib";              // load library for radical
> ring r = 0, (x,y,z), dp;
> ideal i = xy, x2, y3-y5;
> radical(I);
_[1]=x
_[2]=y3-y
```

♢

An ideal I of $\mathbb{K}[X_1, \ldots, X_n]$ is *prime* if $I \neq \mathbb{K}[X_1, \ldots, X_n]$ and for every pair of elements $f, g \in \mathbb{K}[X_1, \ldots, X_n]$, $fg \in I$ implies $f \in I$ or $g \in I$.

An ideal I of $\mathbb{K}[X_1, \ldots, X_n]$ is *maximal* if $I \neq \mathbb{K}[X_1, \ldots, X_n]$ and I is maximal with respect to set inclusion.

Lemma 1.8. *Each maximal ideal* $\mathfrak{m}$ *of* $\mathbb{K}[X_1, \ldots, X_n]$ *is prime.*

Proof. Let $f, g \in \mathbb{K}[X_1, \ldots, X_n]$ with $fg \in \mathfrak{m}$. Suppose $f \notin \mathfrak{m}$. Then $\mathfrak{m} \cup \langle f \rangle = \mathbb{K}[X_1, \ldots, X_n]$, since $\mathfrak{m}$ is maximal. Then $m + af = 1$ for some $m \in \mathfrak{m}$ and $a \in \mathbb{K}[X_1, \ldots, X_n]$. Thus $mg + afg = g$ and hence $g \in \mathfrak{m}$.

Example 1.9. For any field $\mathbb{K}$, every maximal ideal $\mathfrak{m}$ of $\mathbb{K}[X_1, \ldots, X_n]$ is given as follows: take a finite algebraic extension field $\mathbb{L}$ of $\mathbb{K}$ and a point $(a_1, \ldots, a_n) \in \mathbb{L}^n$, consider the ideal $\langle X_1 - a_1, \ldots, X_n - a_n \rangle$ of $\mathbb{L}[X_1, \ldots, X_n]$, and put

$$\mathfrak{m} = \langle X_1 - a_1, \ldots, X_n - a_n \rangle \cap \mathbb{K}[X_1, \ldots, X_n].$$

In particular, if $\mathbb{K}$ is algebraically closed, every maximal ideal of $\mathbb{K}[X_1, \ldots, X_n]$ has the form

$$\mathfrak{m} = \langle X_1 - a_1, \ldots, X_n - a_n \rangle$$

for some $a_1, \ldots, a_n \in \mathbb{K}$. ♢

Example 1.10. In the polynomial ring $\mathbb{K}[X_1, \ldots, X_n]$, every ideal $\langle S \rangle$ generated by a subset S of the set of variables $\{X_1, \ldots, X_n\}$ is prime; in particular, if $S = \emptyset$, then $\langle S \rangle = \{0\}$. The only maximal ideal among these prime ideals is $\langle X_1, \ldots, X_n \rangle$. ♢

1.3 Monomial Orders

We study several ways to order the terms of a polynomial. For this, we first consider orders on the set $\mathbb{N}_0^n$ of n-tuples of natural numbers. The set $\mathbb{N}_0^n$ forms a monoid with the component-wise addition

$$(\alpha_1, \ldots, \alpha_n) + (\beta_1, \ldots, \beta_n) = (\alpha_1 + \beta_1, \ldots, \alpha_n + \beta_n)$$

and the zero vector $0 = (0, \ldots, 0)$ is the identity element..

A *monomial ordering* on $\mathbb{N}_0^n$ is a total ordering $>$ on $\mathbb{N}_0^n$ satisfying the following properties:

1. If $\alpha, \beta \in \mathbb{N}_0^n$ with $\alpha > \beta$ and $\gamma \in \mathbb{N}_0^n$, then $\alpha + \gamma > \beta + \gamma$.
2. If $\alpha \in \mathbb{N}_0^n$ and $\alpha \neq 0$, then $\alpha > 0$.

The first condition shows that the ordering is compatible with the addition in $\mathbb{N}_0^n$ and the second condition means that 0 is the smallest element of the ordering. Both conditions imply that if $\alpha, \beta \in \mathbb{N}_0^n$, then $\alpha + \beta > \alpha$.

For the monoid $\mathbb{N}_0$, there is only one monomial ordering

$$0 < 1 < 2 < 3 < \dots,$$

but in monoids $\mathbb{N}_0^n$ with $n \geq 2$ there are infinitely many monomial orderings.

Example 1.11. The following orderings depend on the ordering of the variables $X_1, \dots, X_n$.

- *Lexicographical ordering (lp)*:

$$\alpha >_{\mathrm{lp}} \beta \quad :\Longleftrightarrow \quad \exists 1 \leq i \leq n : \ \alpha_1 = \beta_1, \dots, \alpha_{i-1} = \beta_{i-1}, \ \alpha_i > \beta_i.$$

- *Degree lexicographical ordering (Dp)*:

$$\alpha >_{\mathrm{Dp}} \beta \quad :\Longleftrightarrow \quad |\alpha| > |\beta| \vee (|\alpha| = |\beta| \wedge \alpha >_{\mathrm{lp}} \beta).$$

- *Degree reverse lexicographical ordering (dp)*:

$$\alpha >_{\mathrm{dp}} \beta \quad :\Longleftrightarrow \quad |\alpha| > |\beta| \vee (|\alpha| = |\beta| \wedge \exists 1 \leq i \leq n : \ \alpha_n = \beta_n, \dots, \alpha_{i+1} = \beta_{i+1}, \ \alpha_i < \beta_i).$$

In all three orderings, $(1, 0, \dots, 0), \dots, (0, \dots, 0, 1) > 0$. For instance, $(3, 0, 0) >_{\mathrm{lp}} (2, 2, 0)$ but $(2, 2, 0) >_{\mathrm{Dp}} (3, 0, 0)$ and $(2, 2, 0) >_{\mathrm{dp}} (3, 0, 0)$. Moreover, $(2, 1, 2) >_{\mathrm{Dp}} (1, 3, 1)$ but $(1, 3, 1) >_{\mathrm{dp}} (2, 1, 2)$. ♢

In the following, we require the natural component-wise ordering on $\mathbb{N}_0^n$ given by

$$(\alpha_1, \dots, \alpha_n) \leq_{\mathrm{nat}} (\beta_1, \dots, \beta_n) \quad :\Longleftrightarrow \quad \alpha_1 \leq \beta_1, \dots, \alpha_n \leq \beta_n.$$

For instance, $(1, 1, 2) \leq_{\mathrm{nat}} (2, 1, 2) \leq_{\mathrm{nat}} (2, 1, 4)$.

Theorem 1.12. (Dickson's Lemma) *Let A be a subset of $\mathbb{N}_0^n$. There is a finite subset B of A such that for each $\alpha \in A$ there is a $\beta \in B$ with $\beta \leq_{\mathrm{nat}} \alpha$.*

The set B is called a *Dickson basis* of A (Fig. 1.1).

Proof. For $n = 1$ take the smallest element of $A \subseteq \mathbb{N}_0$ as the only element of B.

For $n \geq 1$, $A \subseteq \mathbb{N}_0^{n+1}$, and $i \in \mathbb{N}_0$ define

$$A_i = \{\alpha' \in \mathbb{N}_0^n \mid (\alpha', i) \in A\} \subseteq \mathbb{N}_0^{n+1}.$$

By induction, A_i has a Dickson basis B_i. Furthermore, by induction, $\bigcup_{i \in \mathbb{N}_0} B_i$ has a Dickson basis B'. Since B' is finite, there is an index j such that $B' \subseteq B_1 \cup \dots \cup B_j$.

Claim that a Dickson basis of A is given by

$$B = \{(\beta', i) \in \mathbb{N}_0^{n+1} \mid 0 \leq i \leq j, \beta' \in B_i\}.$$

Indeed, let $(\alpha', k) \in A$. Then $\alpha' \in A_k$. Since B_k is a Dickson basis of A_k, there is an element $\beta' \in B_k$ such that $\beta' \leq_{\mathrm{nat}} \alpha'$. If $k \leq j$, then $(\beta', k) \in B$ and $(\beta', k) \leq_{\mathrm{nat}} (\alpha', k)$. Otherwise, there are $\gamma' \in B'$ and $i \leq j$ such that $\gamma' \leq_{\mathrm{nat}} \beta'$ and $(\gamma', i) \in B_i$. Then $(\gamma', i) \in B$ and $(\gamma', i) \leq_{\mathrm{nat}} (\alpha', k)$. □

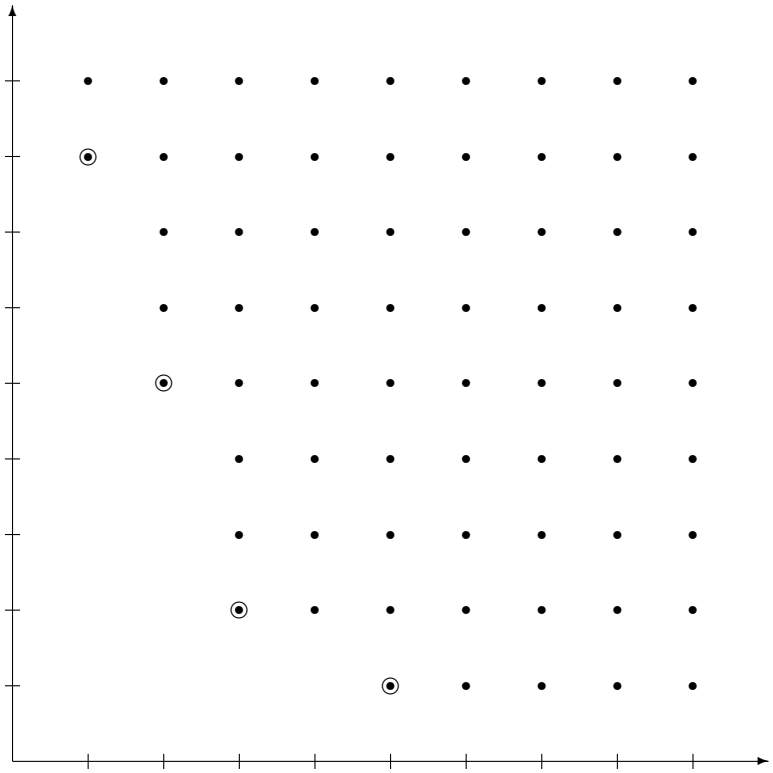


Fig. 1.1. A subset A of $\mathbb{N}_0^2$ and a Dickson set of A (encircled points).

Corollary 1.13. *Each monomial ordering on $\mathbb{N}_0^n$ is a well-ordering.*

Proof. Let $>$ be a monomial ordering on $\mathbb{N}_0^n$ and A be a nonempty subset of $\mathbb{N}_0^n$. By Dickson's lemma, the set A has a Dickson basis B. Let $\alpha \in A$. Then there is an element $\beta \in B$ with $\beta \leq_{\text{nat}} \alpha$. Thus there is an element $\gamma \in \mathbb{N}_0^n$ with $\alpha = \beta + \gamma$. Since $0 \leq \gamma$, it follows that $\beta \leq \beta + \gamma = \alpha$. Hence, the smallest element of the Dickson basis B with respect to the monomial ordering is the smallest element of A. Therefore, the monomial ordering is a well-ordering. □

Corollary 1.14. *For any monomial ordering $>$ on $\mathbb{N}_0^n$, each decreasing chain of elements of $\mathbb{N}_0^n$*

$$\alpha^{(1)} > \alpha^{(2)} > \ldots > \alpha^{(k)} > \ldots$$

becomes stationary (i.e., there is some j_0 such that $\alpha^{(j)} = \alpha^{(j_0)}$ for all $j \geq j_0$).

Proof. Put $A = \{\alpha^{(i)} \mid i \in \mathbb{N}\}$. By Corollary 1.13, A has a smallest element and hence the sequence must become stationary. □

A monomial ordering $>$ on $\mathbb{N}_0^n$ carries forward to a monomial ordering on the set of monomials of the polynomial ring $\mathbb{K}[X_1, \ldots, X_n]$. For this, define for all $\alpha, \beta \in \mathbb{N}_0$,

$$X^\alpha > X^\beta \quad :\Longleftrightarrow \quad \alpha > \beta.$$

Since any monomial ordering is total, the terms that are involved in a polynomial of $\mathbb{K}[X_1, \ldots, X_n]$ can be uniquely written in increasing or decreasing order. A polynomial f in $\mathbb{K}[X_1, \ldots, X_n]$ whose terms are written in decreasing order is in *canonical form*, i.e.,

$$f = c_0 X^{\alpha^{(0)}} + \ldots + c_m X^{\alpha^{(m)}}, \quad c_i \in \mathbb{K}^*,$$

where $\alpha^{(0)} > \ldots > \alpha^{(m)}$. Note that polynomials stored in canonical form can be efficiently tested on equality.

For polynomials in a polynomial ring $\mathbb{K}[X]$ with one unknown, there is only one monomial ordering,

$$1 < X < X^2 < X^3 < \ldots,$$

but in polynomial rings with several unknowns there are infinitely many monomial orderings.

Example 1.15 (Singular). Polynomials are stored and printed in canonical form.

```
> ring r1 = 0, (x,y,z), lp;
> poly f = x3yz+y5; f;
x3yz+y5
> ring r2 = 0, (x,y,z), Dp;
> poly f = imap(r1,f); f;
x3yz+y5
> ring r2 = 0, (x,y,z), dp;
> poly f = imap(r1,f); f;
y5+x3yz
```

♢

The leading data of a polynomial f in $\mathbb{K}[X_1, \ldots, X_n]$ are defined as follows:

- *leading term* $\mathrm{LT}_>(f) = c_0 X^{\alpha^{(0)}}$,
- *leading coefficient* $\mathrm{LC}_>(f) = c_0$, and
- *leading monomial* $\mathrm{LM}_>(f) = X^{\alpha^{(0)}}$.

A polynomial f is called *monic* if its leading coefficient is equal to 1.

Example 1.16. Consider the polynomial $f = 4XY^2Z + 4Z^2 - 5X^3 + 7X^2Z^2$ in $\mathbb{Q}[X, Y, Z]$, where X corresponds to $X^{(1,0,0)}$, Y to $X^{(0,1,0)}$, and Z to $X^{(0,0,1)}$. Thus

$$f = 4X^{(1,2,1)} + 4X^{(0,0,2)} - 5X^{(3,0,0)} + 7X^{(2,0,2)}.$$

In the `lp` ordering, $(3,0,0) \geq (2,0,2) \geq (1,2,1) \geq (0,0,2)$ and the canonical form is

$$f = -5X^3 + 7X^2Z^2 + 4XY^2Z + 4Z^2.$$

In the `Dp` ordering, $(2,0,2) \geq (1,2,1) \geq (3,0,0) \geq (0,0,2)$ and the canonical form is

$$f = 7X^2Z^2 + 4XY^2Z - 5X^3 + 4Z^2,$$

In the `dp` ordering, $(1,2,1) \geq (2,0,2) \geq (3,0,0) \geq (0,0,2)$ and the canonical form is

$$f = 4XY^2Z + 7XY^2Z - 5X^3 + 4Z^2.$$

♢

Example 1.17 (Singular). The leading data of a polynomial can be obtained as follows.

```
> ring r = 0, (x,y,z), lp;
> poly f = (xy-z)*(x2-yz);
> f;
x3y-x2z-xy2z+yz2
> leadmonom(f);
x3y
> leadexp(f);
3,1,0
> leadcoef(f);
1
> lead(f);
x3y
> f-lead(f);          // tail
-x2z-xy2z+yz2
```

♢

1.4 Division Algorithm

The ordinary division algorithm for polynomials in one variable carries forward to the multivariate case by making use of a monomial ordering.

Theorem 1.18. *Let $>$ be a monomial ordering on $\mathbb{N}_0^n$. Let f be a nonzero polynomial in $\mathbb{K}[X_1,\ldots,X_n]$ and let $F = (f_1,\ldots,f_m)$ be a sequence of nonzero polynomials in $\mathbb{K}[X_1,\ldots,X_n]$. There are polynomials $h_1,\ldots,h_m$ and r in $\mathbb{K}[X_1,\ldots,X_n]$ such that*

$$f = h_1f_1 + \cdots + h_mf_m + r \tag{1.11}$$

and either $r = 0$ or none of the terms in r is divisible by $\mathrm{LT}_>(f_1),\ldots,\mathrm{LT}_>(f_m)$. Moreover, if $h_if_i \neq 0$, then $\mathrm{LT}_>(h_if_i) \leq \mathrm{LT}_>(f)$, $1 \leq i \leq m$.

The proof is constructive and mimicks the division algorithm (Alg. 1.1).

Proof. First, put $h_1 = \ldots = h_m = 0$, $r = 0$, and $s = f$. Then we have

$$f = h_1f_1 + \cdots + h_mf_m + (r+s). \tag{1.12}$$

This equation serves as an invariant throughout the algorithm that proceeds in iterative steps. If $s = 0$, the algorithm terminates. Otherwise, there are two cases:

- *Reduction step*: If $\mathrm{LT}_>(s)$ is divisible by some $\mathrm{LT}_>(f_i)$, $1 \le i \le m$, then take the smallest index i with this property and put

$$s = s - \frac{\mathrm{LT}_>(s)}{\mathrm{LT}_> f_i} \quad \text{and} \quad h_i = h_i + \frac{\mathrm{LT}_>(s)}{\mathrm{LT}_>(f_i)}. \tag{1.13}$$

- *Shifting step*: If $\mathrm{LT}_>(s)$ is not divisible by any of the $\mathrm{LT}_>(f_i)$, $1 \le i \le m$, then put

$$r = r + \mathrm{LT}_>(s) \quad \text{and} \quad s = s - \mathrm{LT}_>(s). \tag{1.14}$$

In both cases, the equation (1.12) still holds. Moreover, if $r \neq 0$, then the assertion that no term of r is divisible by $\mathrm{LT}_>(f_i)$, $1 \le i \le m$, inductively holds. The leading term of the polynomial s is strictly decreasing with respect to the monomial ordering after each of the assignments (1.13) and (1.14). Thus the sequence formed by the leading terms of s in successive steps is strictly decreasing. By Corollary 1.13, the monomial ordering is a well-ordering and hence the sequence becomes stationary. Therefore, the division algorithm terminates with $s = 0$.

In view of the inequalities, the leading term of s decreases in each step and is either added to some product $h_i f_i$ (reduction step) or to the remainder r (shifting step). Moreover, in the reduction step, the leading term of s added to the product $h_i f_i$ is the largest term added. Since $\mathrm{LT}_>(s) = \mathrm{LT}_>(f)$ at the start of the computation, the inequalities follows. □

The remainder on the division of f by F is often denoted by $r = f^F$.

Algorithm 1.1 Division algorithm.

Require: nonzero polynomials f and $f_1, \ldots, f_m$ in $\mathbb{K}[X_1, \ldots, X_n]$
Ensure: polynomials $h_1, \ldots, h_m$ and r in $\mathbb{K}[X_1, \ldots, X_n]$ as in Thm. 1.18

```
h_1 ← 0, ..., h_m ← 0
r ← 0
s ← f
while s ≠ 0 do
  i ← 1
  division_occurred ← false
  while i ≤ m and division_occurred = false do
    if LT(f[i]) divides LT(s) then
      s ← s − LT(s)/LT(f_i) * f_i
      h_i ← h_i + LT(s)/LT(f_i)
      division_occurred ← true
    else
      i ← i + 1
    end if
  end while
  if division_occurred = false then
    r ← r + LT(s)
    s ← s − LT(s)
  end if
end while
```

Example 1.19. Consider the polynomials $f = X^2Y + XY^2 + Y^2$, $f_1 = Y^2 - 1$, and $f_2 = X - Y$ in $\mathbb{Q}[X, Y]$ using the `lp` ordering with $X > Y$. Initially, we have $h_1 = h_2 = 0$, $r = 0$, and $s = f$. First, $\text{LT}_>(s) = X^2Y$ is divisible $\text{LT}_>(f_2) = X$ and so

$$s = s - \frac{X^2Y}{X}(X - Y) = 2XY^2 + Y^2$$

and

$$h_2 = h_2 + \frac{X^2Y}{X} = XY.$$

Second, $\text{LT}_>(s) = 2XY^2$ is divisible $\text{LT}_>(f_1) = Y^2$. Thus

$$s = s - \frac{2XY^2}{Y^2}(Y^2 - 1) = Y^2 + 2X$$

and

$$h_1 = h_1 + \frac{2XY^2}{Y^2} = 2X.$$

Third, $\text{LT}_>(s) = 2X$ is divisible by $\text{LT}_>(f_2) = X$. So

$$s = s - \frac{2X}{X}(X - Y) = 2Y + Y^2$$

and

$$h_2 = h_2 + \frac{2X}{X} = XY + 2.$$

Fourth, $\text{LT}_>(s) = Y^2$ is divisible by $\text{LT}_>(f_1) = Y^2$. Thus

$$s = s - \frac{Y^2}{Y^2}(Y^2 - 1) = 2Y + 1$$

and

$$h_1 = 2X + 1.$$

Fifth, $\text{LT}_>(s) = 2Y$ is not divisible by $\text{LT}_>(f_1) = Y^2$ or $\text{LT}_>(f_2) = X$. It follows that

$$r = 2Y \quad \text{and} \quad s = 1.$$

Sixth, $\text{LT}_>(s) = 1$ is not divisible by $\text{LT}_>(f_1) = Y^2$ or $\text{LT}_>(f_2) = X$. Consequently,

$$r = 2Y + 1 \quad \text{and} \quad s = 0.$$

Therefore,

$$f = (2X + 1) \cdot (Y^2 - 1) + (XY + 2) \cdot (X - Y) + (2Y + 1) \quad \text{and} \quad f^F = 2Y + 1.$$

♢

Example 1.20 (Singular). The expression of a polynomial as a linear combination with remainder according to the division theorem is provided by the command `division`, while the command `reduce` only yields the remainder upon division.

```
> ring r = 0, (x,y), lp;
> ideal i = y2-1, x-y;
> poly f = x2y+xy2+y2;
> reduce(f,std(i));      // reduction by standard basis of i
2y+1
> division(f,i);         // division with remainder
[1]:
   _[1,1]=2x+1
   _[2,1]=xy+2
[2]:
   _[1]=2y+1
[3]:
   _[1,1]=1
```

♢

1.5 Groebner Bases

Groebner bases are specific generating sets of polynomial ideals.

Example 1.21. Take the polynomials $f = XY^2 - X$, $f_1 = Y^2 - 1$, and $f_2 = XY + 1$ in $\mathbb{Q}[X, Y]$ using the `lp` ordering with $X > Y$. First, the division of f into $F = (f_1, f_2)$ yields

$$f = X \cdot (Y^2 - 1) + 0 \cdot (XY + 1).$$

Second, the division of f into $F' = (f_2, f_1)$ gives

$$f = Y \cdot (XY + 1) + 0 \cdot (Y^2 - 1) + (-X - Y).$$

Thus the division depends on the ordering of the polynomials in the sequence F. Moreover, the first representation shows that the polynomial f lies in the ideal $I = \langle f_1, f_2 \rangle$, while this cannot be deduced from the second representation. ♢

Let $>$ be a monomial ordering on $\mathbb{K}[X_1, \ldots, X_n]$ and let I be an ideal of $\mathbb{K}[X_1, \ldots, X_n]$. A *Groebner basis* of I with respect to $>$ is a finite set of polynomials $G = \{g_1, \ldots, g_s\}$ in I such that for each nonzero polynomial $f \in I$, $\mathrm{LT}_>(f)$ is divisible by $\mathrm{LT}_>(g_i)$ for some $1 \leq i \leq s$. Groebner bases were invented by Bruno Buchberger in the 1960s and named after his advisor Walter Groebner (1899-1980).

Example 1.22 (Singular). Consider the ideal $I = \langle Y^2 - 1, XY + 1 \rangle$ in $\mathbb{Q}[X, Y]$ using the `lp` ordering with $X > Y$. A Groebner basis of the ideal I can be computed by using the command `groebner` or `std`. The latter command is more general and can be applied to calculate standard bases of polynomial ideals.

```
> ring r = 0, (x,y), lp;
> ideal i = y2-1, xy+1;
> ideal j = std(i);
> j;
j[1]=y2-1
j[2]=x+y
```

The computed Groebner basis of I is $\{Y^2+1, X+Y\}$. ♢

Theorem 1.23. *Each ideal I of $\mathbb{K}[X_1,\ldots,X_n]$ has a Groebner basis with respect to any monomial ordering.*

Proof. Let $>$ be a monomial ordering on $\mathbb{K}[X_1,\ldots,X_n]$. Consider the set

$$A = \{\alpha \in \mathbb{N}_0^n \mid X^\alpha = \text{LM}_>(f) \text{ for some } f \in I\}$$

of exponents of all leading monomials of the polynomials in the ideal I. By Dickson's lemma, the set A has a Dickson basis $B = \{\beta_1,\ldots,\beta_s\}$, where $X^{\beta_i} = \text{LM}_>(g_i)$ for some $g_i \in I$, $1 \le i \le s$. Let f be a nonzero polynomial of I with $\text{LM}_>(f) = X^\alpha$. Then $\alpha = \beta_i + \gamma$ for some $1 \le i \le s$ and $\gamma \in \mathbb{N}_0^n$. Thus $X^\alpha = X^{\beta_i}X^\gamma$ and hence the leading term of f is divisible by the leading term of g_i. It follows that $\{g_1,\ldots,g_s\}$ is a Groebner basis of I. □

Proposition 1.24 (Ideal Membership Test). *Let $>$ be a monomial ordering on $\mathbb{K}[X_1,\ldots,X_n]$ and let $I = \langle g_1,\ldots,g_s\rangle$ be an ideal of $\mathbb{K}[X_1,\ldots,X_n]$. If $G = \{g_1,\ldots,g_s\}$ is a Groebner basis of I, then for each polynomial f in $\mathbb{K}[X_1,\ldots,X_n]$, we have $f \in I$ if and only if $f^G = 0$.*

Proof. Let $f \in \mathbb{K}[X_1,\ldots,X_n]$ whose division into G yields $f = h_1g_1 + \ldots + h_sg_s + f^G$. Let $f^G = 0$. Then $f \in I$ by definition of I.

Conversely, let $f \in I$. Then $f^G = f - (h_1g_1 + \ldots + h_sg_s)$ belongs to I. Assume that $f^G \neq 0$. Then $\text{LT}_>(f^G)$ is divisible by some $\text{LT}_>(g_i)$, since G is a Groebner basis of I. But this contradicts the division algorithm, since none of the terms in the remainder f^G is divisible by any of the terms $\text{LT}_>(g_i)$. □

Corollary 1.25. *Let $>$ be a monomial ordering on $\mathbb{K}[X_1,\ldots,X_n]$ and let I be an ideal of $\mathbb{K}[X_1,\ldots,X_n]$. If $G = \{g_1,\ldots,g_s\}$ is a Groebner basis of I with respect to $>$, then $I = \langle g_1,\ldots,g_s\rangle$.*

Proof. By definition, $g_1,\ldots,g_s \in I$ and thus $\langle g_1,\ldots,g_s\rangle \subseteq I$. Conversely, let $f \in I$. The division of f into G yields $f = h_1g_1 + \ldots + h_sg_s + f^G$. Thus by Prop. 1.24, $f^G = 0$ and hence $f \in \langle g_1,\ldots,g_s\rangle$. □

Proposition 1.26. *Let $G = \{g_1,\ldots,g_s\}$ be a Groebner basis in $\mathbb{K}[X_1,\ldots,X_n]$ with respect to any monomial ordering $>$. For each polynomial f in $\mathbb{K}[X_1,\ldots,X_n]$, the remainder f^G is uniquely determined and independent of the order of the elements in G.*

Proof. Let f be a polynomial in $\mathbb{K}[X_1,\ldots,X_n]$ and let $I = \langle g_1,\ldots,g_s\rangle$. First, assume that there are two expressions $f = h_1g_1+\ldots+h_sg_s+r$ and $f = h_1'g_1+\ldots+h_s'g_s+r'$ as given by the division theorem. Then $r'-r = (h_1-h_1')g_1 + \ldots + (h_s-h_s')g_s$ lies in I. Suppose that $r'-r \neq 0$. Since G is a Groebner basis of I, the leading term of $r'-r$ is divisible by the leading term of some g_i, $1 \le i \le s$. But this contracts the fact that r and r' are remainders and so none of their terms are divisible by any of the g_i, $1 \le i \le s$.

Second, let G' be a permutation of the Groebner basis G. Then the division algorithm yields $f = h_1'g_1 + \ldots + h_s'g_s + f^{G'}$. But the remainder is uniquely determined and therefore $f^G = f^{G'}$. □

The remainder on division of a polynomial f by a Groebner basis of an ideal I is a uniquely determined *normal form* of f modulo I depending only on the monomial ordering and not how the division is performed.

Theorem 1.27. (Hilbert Basis Theorem) *Each ideal I of $\mathbb{K}[X_1,\ldots,X_n]$ is finitely generated.*

The proof follows directly from Thm. 1.23 and Cor. 1.25.

A ring is *Noetherian* if each ideal of R is finitely generated.

Theorem 1.28. *The following conditions of a ring R are equivalent:*

1. *Each ideal of R is finitely generated (that is, R is Noetherian).*
2. *Each ascending chain of ideals $I_1 \subset I_2 \subset \cdots$ in R becomes stationary (that is, there is an index j_0 such that $I_j = I_{j_0}$ for all $j \geq j_0$).*
3. *Each nonempty set of ideals in R contains a maximal element (with respect to inclusion).*

Proof. Suppose each ideal of R is finitely generated. Assume that $I_1 \subset I_2 \subset \cdots$ is an ascending chain of ideals in R. Then $I = \bigcup_j I_j$ is an ideal in R and by hypothesis has a finite generating set G. If $G \subset I_1 \cup \ldots \cup I_{j_0}$, then $I_j = I_{j_0}$ for all $j \geq j_0$.

Suppose that each ascencing chain of ideals in R become stationary. Assume that S is a nonempty set of ideals in R. If $I_1 \in S$ is not maximal in S, then there exists an ideal I_2 in S that properly contains I_1. Continuing like this gives an ascending chain of ideals in S that will become stationary. Then $I_j = I_{j_0}$ for all $j \geq j_0$ and I_{j_0} is maximal in S.

Suppose that each nonempty set of ideals in R contains a maximal element. Let I be an ideal of R, and let S be a set of ideals $J \subseteq I$ of R that are finitely generated. Then S is nonempty and by hypothesis contains a maximal element $J_0 = \langle f_1, \ldots, f_s \rangle$. Assume that $I \neq J_0$. Then there is an element $f \in I \setminus J_0$ and so $\langle f, f_1, \ldots, f_s \rangle$ will be a finitely generated ideal in I that properly contains J_0. This contradicts the maximality of J_0. Hence, I is finitely generated. □

By the Hilbert basis theorem and the above result, we obtain the following.

Corollary 1.29. *The polynomial ring $\mathbb{K}[X_1, \ldots, X_n]$ is Noetherian.*

1.6 Computation of Groebner Bases

The basic algorithm for the computation of a Groebner basis of an ideal in $\mathbb{K}[X_1, \ldots, X_n]$ is due to Buchberger.

Let $>$ be a monomial ordering on $\mathbb{K}[X_1, \ldots, X_n]$ and let $f, g \in \mathbb{K}[X_1, \ldots, X_n] \setminus \{0\}$ with $\mathrm{LM}_>(f) = X^\alpha$ and $\mathrm{LM}_>(g) = X^\beta$, respectively. The least common multiple of α and β w.r.t. the natural ordering on $\mathbb{N}_0^n$ is

$$\gamma = \mathrm{lcm}(\alpha, \beta) = (\max\{\alpha_1, \beta_1\}, \ldots, \max\{\alpha_n, \beta_n\})$$

Then the least common multiple of X^α and X^β w.r.t. the relation of divisibility is

$$X^\gamma = \mathrm{lcm}(X^\alpha, X^\beta).$$

Define the *S-polynomial* of f and g as

$$S(f, g) = \frac{X^\gamma}{\mathrm{LT}_>(f)} \cdot f - \frac{X^\gamma}{\mathrm{LT}_>(g)} \cdot g. \tag{1.15}$$

Note that $S(f, g)$ lies in the ideal generated by f and g. Moreover, in $S(f, g)$ the leading terms of f and g cancel and thus $S(f, g)$ exhibits a new leading term.

Example 1.30. Consider the polynomials $f = 2Y^2+Z^2$ and $g = 3X^2Y+YZ$ in $\mathbb{Q}[X,Y,Z]$ with respect to the `lp` ordering with $X > Y > Z$. Then $\text{LM}_>(f) = Y^2$, $\text{LM}_>(g) = X^2Y$, and $\text{lcm}(\text{LT}_>(f), \text{LT}_>(g)) = X^2Y^2$. Thus

$$S(f,g) = \frac{X^2Y^2}{2Y^2} \cdot f - \frac{X^2Y^2}{3X^2Y} \cdot g = \frac{1}{2}X^2Z^2 - \frac{1}{3}Y^2Z.$$

♢

Theorem 1.31. (Buchberger's S-Criterion) *Let $>$ be a monomial ordering on $\mathbb{K}[X_1,\ldots,X_n]$. A set $G = \{g_1,\ldots,g_s\}$ of polynomials in $\mathbb{K}[X_1,\ldots,X_n]$ is a Groebner basis of the ideal $I = \langle g_1,\ldots,g_s\rangle$ if and only if $S(g_i,g_j)^G = 0$ for all pairs $i \neq j$.*

Proof. Let G be a Groebner basis of I. Since each S-polynomial $S(g_i,g_j)$ belongs to I, it follows from Prop. 1.24 that $S(g_i,g_j)^G = 0$.

Conversely, assume that $S(g_i,g_j)^G = 0$ for all pairs $i \neq j$. Let $f \in I$. Write

$$f = h_1g_1 + \ldots + h_sg_s,$$

where $h_1,\ldots,h_s \in \mathbb{K}[X_1,\ldots,X_n]$. Let $\text{LT}_>(f) = cX^\alpha$, $\text{LT}_>(g_i) = c_iX^{\alpha_i}$, and $\text{LT}_>(h_i) = d_iX^{\beta_i}$, $1 \leq i \leq s$. Define $\delta = \max_>\{\alpha_i + \beta_i \mid 1 \leq i \leq s\}$. The above equation shows that the leading term of f is a $\mathbb{K}$-linear combination of the leading terms $\text{LT}_>(h_ig_i)$, $h_ig_i \neq 0$, and therefore $\alpha \leq \delta$.

If $\delta = \alpha$, we can assume that $\delta = \alpha_1 + \beta_1 = \ldots = \alpha_r + \beta_r$, where $r \leq s$ and $h_ig_i \neq 0$ for $1 \leq i \leq r$. Then

$$cX^\alpha = (c_1d_1 + \ldots + c_rd_r)X^\delta.$$

Thus $\text{LT}_>(g_1) = c_1X^{\alpha_1}$ divides $\text{LT}_>(f) = cX^\alpha$. If all nonzero polynomials f of I have this property, then G is a Groebner basis of I.

If $\alpha < \delta$, the maximal leading terms on the right-hand side of the representation of f must cancel. By the above notation, we obtain

$$c_1d_1 + \ldots + c_rd_r = 0. \tag{1.16}$$

Write the polynomial f in the form

$$f = C + (h_1 - \text{LT}_>(h_1))g_1 + \ldots + (h_r - \text{LT}_>(h_r))g_r + h_{r+1}g_{r+1}\ldots + h_sg_s,$$

where $C = \text{LT}_>(h_1)g_1 + \ldots + \text{LT}_>(h_r)g_r$. By putting $k_i = X^{\beta_i}g_i/c_i$, $1 \leq i \leq r$, we obtain

$$\begin{aligned} C &= c_1d_1k_1 + \ldots + c_rd_rk_r \\ &= c_1d_1(k_1 - k_2) + (c_1d_1 + c_2d_2)(k_2 - k_3) + (c_1d_1 + c_2d_2 + c_3d_3)(k_3 - k_4) + \ldots \\ &\quad \ldots + (c_1d_1 + \ldots + c_{r-1}d_{r-1})(k_{r-1} - k_r) + (c_1d_1 + \ldots + c_rd_r)k_r. \end{aligned} \tag{1.17}$$

Thus C is a linear combination of $k_i - k_j$, $1 \leq i < j \leq r$. Define $X^{\alpha_{i,j}}$ as the least common multiple of X^{α_i} and X^{α_j}. Then there exists $\xi \in \mathbb{N}_0^n$ so that $\xi + \alpha_{i,j} = \alpha_i + \beta_i = \alpha_j + \beta_j$, $1 \leq i < j \leq r$. We have

$$\begin{aligned} k_i - k_j &= \frac{X^{\beta_i}g_i}{c_i} - \frac{X^{\beta_j}g_j}{c_j} \\ &= X^\xi\left(\frac{X^{\alpha_{i,j}}g_i}{c_iX^{\alpha_i}} - \frac{X^{\alpha_{i,j}}g_j}{c_jX^{\alpha_j}}\right) \\ &= X^\xi S(g_i,g_j) \end{aligned}$$

and $\mathrm{LT}_>(k_i - k_j) < \delta$, $1 \leq i < j \leq r$. It follows from (1.16) and (1.17) that

$$C = c'_1 X^{\xi_1} S(g_1, g_2) + \ldots + c'_{r-1} X^{\xi_{r-1}} S(g_{r-1}, g_r),$$

where $c'_1, \ldots, c'_{r-1} \in \mathbb{K}$ and $\xi_1, \ldots, \xi_{r-1} \in \mathbb{N}_0^n$. By hypothesis,

$$S(g_i, g_j) = h_1^{ij} g_1 + \ldots + h_s^{ij} g_s$$

for some polynomials $h_1^{ij}, \ldots, h_s^{ij}$ with $\mathrm{LT}_>(h_l^{ij}) \leq \mathrm{LT}_>(S(g_i, g_j))$, $1 \leq i < j \leq s$ and $1 \leq l \leq s$. It follows that the polynomial C can be written as a linear combination of the polynomials $g_1, \ldots, g_s$. Thus by (1.17), the polynomial f can be expressed as a linear combination of the polynomials $g_1, \ldots, g_s$,

$$f = h'_1 g_1 + \ldots + h'_s g_s,$$

where $\max_>\{\mathrm{LT}_>(h'_i g_i) \mid h'_i g_i \neq 0,\ 1 \leq i \leq s\} < \delta$. Since each monomial ordering is a well-ordering, we obtain by continuing in this way an expression

$$f = h''_1 g_1 + \ldots + h''_s g_s,$$

where the leading monomial X^δ on the right-hand side equals $\mathrm{LT}_>(f)$. Then the case $\alpha = \delta$ will establish the result. □

Buchberger's S-criterion can be used to calculate a Groebner basis of a given ideal (Alg. 1.2).

Algorithm 1.2 Buchberger's algorithm.

Require: $I = \langle f_1, \ldots, f_m \rangle$ ideal of $\mathbb{K}[X, \ldots, X_n]$, $F = \{f_1, \ldots, f_m\}$
Ensure: Groebner basis G of I with $F \subseteq G$

```
G ← F
repeat
  G' ← G
  for each pair f ≠ g in G' do
    S ← S(f, g)^{G'}
    if S ≠ 0 then
      G ← G ∪ {S}
    end if
  end for
until G = G'
```

Theorem 1.32. *Buchberger's algorithm terminates and the output is a Groebner basis.*

Proof. First, we prove correction. Claim that at each step $G \subseteq I$. Indeed, this is true at the start of the algorithm. Suppose $G \subseteq I$ holds at the beginning of some pass and put $G = \{g_1, \ldots, g_s\}$. Then for all $f, g \in G$, $S(f, g) \in \langle f, g \rangle \subseteq I$. Moreover, the division algorithm gives $S(f, g) = h_1 g_1 + \ldots + h_s g_s + S(f, g)^G$. Thus $S(f, g)^G \in I$ and hence $G \subseteq I$ after each pass. Upon termination, the remainders of the S-polynomials divided by the current set G are 0. In this case, Buchberger's S-criterion shows that the set G is a Groebner basis.

Second, we show termination. For this, consider the ideal of leading terms of $G = \{g_1, \ldots, g_s\}$ given by

$$\langle LT(G)\rangle = \langle \mathrm{LT}(g_1), \ldots, \mathrm{LT}(g_s)\rangle.$$

In each pass, the set G is replaced by a new set G'. If $G \neq G'$, there is at least one remainder $r = S(f,g)^G$ with $f, g \in G$ which is added to G'. Since no term of r is divisible by any of the leading terms of the polynomials in G, we have

$$\langle LT(G)\rangle \subset \langle LT(G')\rangle.$$

This gives an ascending chain of ideals of $\mathbb{K}[X_1, \ldots, X_n]$. But the polynomial ring $\mathbb{K}[X_1, \ldots, X_n]$ is Noetherian and so the chain becomes stationary; that is, at some pass

$$\langle LT(G)\rangle = \langle LT(G')\rangle.$$

Thus $G = G'$ by the way G' is constructed from G and hence the algorithm stops. □

Example 1.33. Consider the ideal $I = \langle Y^2 + Z^2, X^2Y + YZ\rangle$ in $\mathbb{Q}[X, Y, Z]$ with respect to the `lp` ordering with $X > Y > Z$. The following session provides a Groebner basis of I according to Buchberger's algorithm:

```
> LIB "teachstd.lib";        // library for command spoly
> ring r = 0, (x,y,z), lp;
> ideal i = y2+z2, x2y+yz;
> reduce(spoly(y2+z2, x2y+yz), i);
x2z2+z3
> ideal j = y2+z2, x2y+yz, x2z2+z3;
> reduce(spoly(y2+z2, x2y+yz), j);
0
> reduce(spoly(y2+z2, x2z2+z3), j);
0
> reduce(spoly(x2y+yz, x2z2+z3), j);
0
```

It follows that $\{Y^2 + Z^2, X^2Y + YZ, X^2Z^2 + Z^3\}$ is a Groebner basis of I. ♢

1.7 Reduced Groebner Bases

Groebner bases are not unique since a Groeber basis remains a Groebner basis if an arbitrary polynomial is added. It will shown that reduced Groebner bases are unique.

Let $>$ be a monomial ordering on $\mathbb{K}[X_1, \ldots, X_n]$. A Groebner basis $G = \{g_1, \ldots, g_s\}$ in $\mathbb{K}[X_1, \ldots, X_n]$ is *minimal* if the polynomials $g_1, \ldots, g_s$ are monic and $\mathrm{LT}_>(g_i)$ is not divisible by $\mathrm{LT}_>(g_j)$ for any pair $i \neq j$.

Proposition 1.34. *Each nonzero ideal I of $\mathbb{K}[X_1, \ldots, X_n]$ has a minimal Groebner basis with respect to any monomial ordering.*

Proof. Let $G = \{g_1, \ldots, g_s\}$ be a Groebner basis of the ideal I with respect to the monomial ordering $>$. We may assume that each generator g_i is monic by multiplying g_i with the inverse of its leading coefficient, $1 \leq i \leq s$.

Suppose G is not minimal. We may assume that $\mathrm{LT}_>(g_1)$ is divisible by $\mathrm{LT}_>(g_i)$ for some $2 \leq i \leq s$. By reduction, the polynomial

$$h = g_1 - \frac{\mathrm{LT}_>(g_1)}{\mathrm{LT}_>(g_i)} g_i \tag{1.18}$$

belongs to I and, by Prop. 1.24, its division into G yields $h^G = 0$. But the leading term of g_1 cancels in (1.18) and is larger than the leading term of h. Thus the polynomial g_1 cannot be used during the division of h by the basis G. Hence, the polynomial h is a linear combination of $g_2, \ldots, g_s$. It follows that by (1.18), the generator g_1 is also a linear combination of $g_2, \ldots, g_s$. Therefore, $G' = \{g_2, \ldots, g_s\}$ also generates the ideal I. Moreover, G' is a Groebner basis since if the leading term of a polynomial $f \in I$ is divisible by $\mathrm{LT}_>(g_1)$, then it is also divisible by $\mathrm{LT}_>(g_i)$.

Repeating the above argument leads to a minimal Groebner basis in a finite number of steps. □

Let $>$ be a monomial ordering on $\mathbb{K}[X_1, \ldots, X_n]$. A Groebner basis $G = \{g_1, \ldots, g_s\}$ in $\mathbb{K}[X_1, \ldots, X_n]$ is *reduced* if G is a minimal Groebner basis and no term in g_i is divisible by $\mathrm{LT}_>(g_j)$ for any pair $i \neq j$.

Proposition 1.35. *Each nonzero ideal I of $\mathbb{K}[X_1, \ldots, X_n]$ has a unique reduced Groebner basis with respect to any monomial ordering.*

Proof. Let $\{f_1, \ldots, f_r\}$ and $\{g_1, \ldots, g_s\}$ be reduced Groebner bases of I with respect to the monomial ordering $>$.

Claim that $r = s$ and after reordering $\mathrm{LT}_>(f_1) = \mathrm{LT}_>(g_1), \ldots, \mathrm{LT}_>(f_s) = \mathrm{LT}_>(g_s)$. Indeed, by definition of Groebner bases, $\mathrm{LT}_>(g_1)$ is divisible by some $\mathrm{LT}_>(f_i)$, $1 \leq i \leq r$. We may assume that $i = 1$. Moreover, $\mathrm{LT}_>(f_1)$ is divisible by some $\mathrm{LT}_>(g_j)$, $1 \leq j \leq s$. Then $\mathrm{LT}_>(g_j)$ divides $\mathrm{LT}_>(g_1)$. By minimality, we have $j = 1$. Since f_1 and g_1 are monic, it follows that $\mathrm{LT}_>(f_1) = \mathrm{LT}_>(g_1)$. The same argument applies to the other generators. In this way, we obtain the desired result.

Claim that $f_1 = g_1, \ldots, f_s = g_s$. Indeed, consider the polynomial $f_1 - g_1$. The first assertion shows that the leading terms in f_1 and g_1 cancel. From this and the definition of reduced Groebner bases it follows that no term in $f_1 - g_1$ is divisible by $\mathrm{LT}_>(f_1) = \mathrm{LT}_>(g_1), \mathrm{LT}_>(f_2) = \mathrm{LT}_>(g_2), \ldots \mathrm{LT}_>(f_s) = \mathrm{LT}(g_s)$. Thus if $f_1 - g_1$ is divided into $(f_1, \ldots, f_s)$, it is already the remainder. But $f_1 - g_1 \in I$ and so it follows from Prop. 1.24 that $(f_1 - g_1)^G = 0$. Hence, $f_1 = g_1$. The same procedure applies to the other generators and the claim follows.

Finally, claim that the ideal I has a reduced Groebner basis. Indeed, the ideal I has a minimal Groebner basis $\{g_1, \ldots, g_s\}$ by Prop. 1.34. First, replace g_1 by the remainder of g_1 modulo $(g_2, \ldots, g_s)$. By the division algorithm, none of the terms of the new g_1 is divisible by $\mathrm{LT}_>(g_2), \ldots, \mathrm{LT}_>(g_s)$. Moreover, by minimality, the leading term of the original g_1 will be shifted to the new g_1. Second, substitute g_2 by the remainder of g_2 modulo $(g_1, g_3, \ldots, g_s)$. This procedure is Continued until g_s is replaced by the remainder of g_s modulo $(g_1, \ldots, g_{s-1})$. Then the leading terms of the original generators $g_1, \ldots, g_s$ will survive and thus the new generators $g_1, \ldots, g_s$ will still form a Groebner basis. Furthermore, by construction, none of the terms in g_i is divisible by $\mathrm{LT}_>(g_j)$ by any pair $i \neq j$. Hence, we end up with a reduced Groebner basis as claimed. □

Example 1.36 (Singular). The commands `groebner` and `std` compute reduced Groebner bases with respect to (global) monomial orderings.

```
> ring r = 0, (x,y,z), dp;
> ideal i = xyz, xy-yz, xz-y2;
> ideal j = std(i);
> j;
j[1]=y2-xz
j[2]=xy-yz
j[3]=yz2
j[4]=x2z-xz2
j[5]=xz3
```

♢

1.8 Toric Ideals

Toric ideals represent algebraic relations between monomials and arise naturally in algebraic statistics.

Let $\mathbb{K}$ be a field and let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be variables over $\mathbb{K}$. Let $t_1, \ldots, t_n$ be monomials in $\mathbb{K}[Y_1, \ldots, Y_m]$. Consider the $\mathbb{K}$-algebra homomorphism $\phi : \mathbb{K}[X_1, \ldots, X_n] \to \mathbb{K}[Y_1, \ldots, Y_m]$ given by

$$\phi : X_i \mapsto t_i, \quad 1 \leq i \leq n. \tag{1.19}$$

The kernel of this map, $\ker \phi = \{a \in \mathbb{K}[X_1, \ldots, X_n] \mid \phi(a) = 0\}$, is called the *toric ideal* associated with $t_1, \ldots, t_n$ and is denoted by $I(t_1, \ldots, t_n)$.

Proposition 1.37. *Let R be a ring and let $R[X_1, \ldots, X_n]$ be a polynomial ring over R. Let $t_1, \ldots, t_n \in R$ and let $\psi : R[X_1, \ldots, X_n] \to R$ be an R-homomorphism defined by $\psi(X_i) = t_i$, $1 \leq i \leq n$.*

- *For each element $f \in R[X_1, \ldots, X_n]$, there exist elements $h_1, \ldots, h_n \in R[X_1, \ldots, X_n]$ and $r \in R$ such that*
$$f = \sum_{i=1}^{n} h_i \cdot (X_i - t_i) + r.$$
- *The kernel of the R-homomorphism ψ is the ideal $\langle X_i - t_i \mid 1 \leq i \leq n \rangle$ in $R[X_1, \ldots, X_n]$.*

Proof.

- Divide the polynomial $f \in R[X_1, \ldots, X_n]$ into $X_i - t_i$, $1 \leq i \leq n$. The division yields the desired representation of f.
- We have $\psi(X_i - t_i) = \psi(X_i) - \psi(t_i) = \psi(X_i) - t_i = 0$, $1 \leq i \leq n$, and thus the ideal $\langle X_i - t_i \mid 1 \leq i \leq n \rangle$ belongs to the kernel of ψ. Conversely, assume that $f \in R[X_1, \ldots, X_n]$ lies in the kernel of ψ. By the first assertion, we have $0 = f = \sum_{i=1}^{n} h_i \cdot (X_i - t_i) + r$. Then $0 = \psi(f) = f(t_1, \ldots, t_n) = r$ and thus the polynomial f lies in the ideal $\langle X_i - t_i \mid 1 \leq i \leq n \rangle$. □

Proposition 1.38. *Let $t_1, \ldots, t_n$ be monomials in $\mathbb{K}[Y_1, \ldots, Y_m]$ and let $J = \langle X_i - t_i \mid 1 \leq i \leq n \rangle$ be an ideal of $\mathbb{K}[X_1, \ldots, X_n, Y_1, \ldots, Y_m]$.*

- *We have $I(t_1, \ldots, t_n) = J \cap \mathbb{K}[X_1, \ldots, X_n]$.*
- *If G is a Groebner basis of J with respect to an elimination ordering with $Y_1 > \ldots > Y_m > X_1 > \ldots > X_n$, then $G \cap \mathbb{K}[X_1, \ldots, X_n]$ is a Groebner basis of $I(t_1, \ldots, t_n)$.*

Proof. Take the $\mathbb{K}$-algebra homomorphism $\psi : \mathbb{K}[X_1, \ldots, X_n, Y_1, \ldots, Y_m] \to \mathbb{K}[Y_1, \ldots, Y_m]$ given by $\psi(Y_i) = Y_i$, $1 \leq i \leq m$, and $\psi(X_i) = t_i$, $1 \leq i \leq n$. Consider $\mathbb{K}[X_1, \ldots, X_n]$ as a subring of the polynomial ring $\mathbb{K}[X_1, \ldots, X_n, Y_1, \ldots, Y_m]$ and observe that the $\mathbb{K}$-algebra homomorphism ϕ given in (1.19) is the restriction of ψ onto $\mathbb{K}[X_1, \ldots, X_n]$; that is, $\phi(f) = \psi(f)$ for each $f \in \mathbb{K}[X_1, \ldots, X_n]$. Thus by Prop. 1.37, we have $\ker \psi = J$ and hence $I(t_1, \ldots, t_n) = \ker \phi = \ker \psi_{|\mathbb{K}[X_1,\ldots,X_n]} = J \cap \mathbb{K}[X_1, \ldots, X_n]$.

The second assertion follows directly from the Elimination theorem. □

A polynomial in $\mathbb{K}[X_1, \ldots, x_n]$ is a *binomial* if it is given by the difference of two monomials. Thus a binomial is of the form $X^\alpha - X^\beta$ where $\alpha, \beta \in \mathbb{N}_0^n$. A binomial $X^\alpha - X^\beta$ is called *pure* if $\gcd(X^\alpha, X^\beta) = 1$. For instance, the binomial $X_1X_2^3 - X_3X_4^2$ is pure, while $X_1X_2^3 - X_1X_3X_4^2$ is not.

Theorem 1.39. *Consider a grading on the polynomial ring $\mathbb{K}[X_1, \ldots, X_n, Y_1, \ldots, Y_m]$ where the degrees of the variables $Y_1, \ldots, Y_m$ are arbitrary and $\deg X_i = \deg t_i$, $1 \leq i \leq n$. Then the toric ideal $I = I(t_1, \ldots, t_n)$ is prime, generated by pure binomials, and homogeneous.*

Proof. Let $fg \in I$. Then $0 = \phi(fg) = \phi(f)\phi(g)$. But $\mathbb{K}[X_1, \ldots, X_n]$ is an integral domain and so $\phi(f) = 0$ or $\phi(g) = 0$. Thus $f \in I$ or $g \in I$ and hence I is a prime ideal.

The ideal $J = \langle X_i - t_i \mid 1 \leq i \leq n \rangle$ of $\mathbb{K}[X_1, \ldots, X_n, Y_1, \ldots, Y_m]$ is generated by binomials. Groebner basis theory implies that all elements in any reduced Groebner basis of J are also binomials. Thus by Prop. 1.38, the ideal I is generated by binomials. Let $X^\alpha - X^\beta$ be a binomial in I. Suppose it is not pure. Then $X^\alpha - X^\beta = X^\gamma(X^\delta - X^\epsilon)$ for some $\gamma, \delta, \epsilon \in \mathbb{N}_0^n$ and so $0 = \phi(X^\alpha - X^\beta) = t^\gamma \phi(X^\delta - X^\epsilon)$. Since $\mathbb{K}[X_1, \ldots, X_n]$ is an integral domain, $\phi(X^\delta - X^\epsilon) = 0$. Thus $X^\delta - X^\epsilon \in I$ and hence I is generated by pure binomials.

The ideal J is homogeneous, since it is generated by homogeneous polynomials. Let f be a polynomial in $J \cap \mathbb{K}[X_1, \ldots, X_n]$. Since J is homogeneous, all homogeneous components of f lie in J and therefore all homogeneous components belong to $\mathbb{K}[X_1, \ldots, X_n]$. Thus all homogeneous components are in I and hence I is homogeneous. □

Example 1.40 (Singular). The toric ideal $I = I(Y_1^3Y_2^3, Y_1^2, Y_2^2)$ is the kernel of the $\mathbb{Q}$-algebra homomorphism $\phi : \mathbb{Q}[X_1, X_2, X_3] \to \mathbb{Q}[Y_1, Y_2]$ given by

$$\phi(X_1) = Y_1^3Y_2^3, \quad \phi(X_2) = Y_1^2, \quad \text{and} \quad \phi(X_3) = Y_2^2.$$

Consider the ideal $J = \langle X_1 - Y_1^3Y_2^3, X_2 - Y_1^2, X_3 - Y_2^2 \rangle$ of $\mathbb{Q}[X_1, X_2, X_3, Y_1, Y_2]$. The following computation provides a reduced Groebner basis of J,

```
> ring r = 0, (y(1..2), x(1..3)), dp;
> ideal j = x(1)-y(1)^3*y(2)^2, x(2)-y(1)^2, x(3)-y(2)^2;
> eliminate( std(j),y(1)*y(2) );
_[1]=x(2)^3*x(3)^3-x(1)^2
```

The toric ideal $I = J \cap \mathbb{Q}[X_1, X_2, X_3]$ has the reduced Groebner basis $\{X_1^2 - X_2^3X_3^3\}$. ♢

Toric ideals often arise by using integral matrices. To see this, let $A = (a_{ij})$ be an integral $m \times n$ matrix with non-negative entries. The columns of the matrix A give rise to the monomials

$$t_i = Y_1^{a_{1i}} \cdots Y_m^{a_{mi}}, \quad 1 \leq i \leq n, \tag{1.20}$$

in the polynomial ring $\mathbb{K}[Y_1, \ldots, Y_m]$. The *toric ideal associated to A* is the toric ideal $I(t_1, \ldots, t_n)$ in $\mathbb{K}[X_1, \ldots, X_n]$, which is also denoted by $I(A)$.

Proposition 1.41. *Let $A = (a_{ij}) \in \mathbb{Z}_{\geq 0}^{m \times n}$. The toric ideal $I(A)$ equals the ideal*

$$I_A = \langle X^\alpha - X^\beta \mid A\alpha = A\beta,\ \alpha, \beta \in \mathbb{N}_0^n \rangle.$$

Proof. Let $\alpha \in \mathbb{N}_0^n$. The $\mathbb{K}$-algebra homomorphism $\phi : \mathbb{K}[X_1, \ldots, X_n] \to \mathbb{K}[Y_1, \ldots, Y_m]$ given by $\phi(X_i) = t_i$ with t_i as in (1.20), $1 \leq i \leq n$, assigns to the monomial X^α the monomial $Y^{A\alpha}$.

Let $X^\alpha - X^\beta$ lie in I_A. Then $\phi(X^\alpha - X^\beta) = \phi(X^\alpha) - \phi(X^\beta) = Y^{A\alpha} - Y^{A\beta} = 0$ and thus $X^\alpha - X^\beta$ lies in $I(A)$. Conversely, by Thm. 1.39, the toric ideal $I(A)$ is generated by binomials $X^\alpha - X^\beta$, $\alpha, \beta \in \mathbb{N}_0^n$. For each such binomial, $0 = \phi(X^\alpha - X^\beta) = \phi(X^\alpha) - \phi(X^\beta) = Y^{A\alpha} - Y^{A\beta}$. Thus $A\alpha = A\beta$ and hence $X^\alpha - X^\beta$ belongs to I_A. ♢

Example 1.42 (Singular). The integral matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 2 & 1 & 1 & 0 \end{pmatrix}$$

gives the toric ideal $I(A) = I(Y_3^2, Y_1Y_3, Y_2Y_3, Y_1Y_2)$. A reduced Groebner basis of this ideal is $\{X_1X_4 - X_2X^3\}$ as can be seen from the following calculation,

```
> ring r = 0, (x(1..4),y(1..3)), lp;
> ideal i = x(1)-y(3)^2, x(2)-y(1)*y(3), x(3)-y(2)*y(3), x(4)-y(1)*y(2);
> ideal j = std(i);
> eliminate( j, y(1)*y(2)*y(3) );
_[1]=x(1)*x(4)-x(2)*x(3)
```

♢

2

Algebraic Geometry

Algebraic geometry is the study of algebraic varieties, which are the zero sets of systems of multivariate polynomials. Algebraic varieties are geometric objects that can be described algebraically by commutative algebra. This section provides a dictionary which allows to translate geometric objects into algebraic ones.

2.1 Affine Varieties

Let $\mathbb{K}$ be a field. The set $\mathbb{K}^n = \{(a_1, \ldots, a_n) \mid a_1, \ldots, a_n \in \mathbb{K}\}$ is the *affine n-dimensional space* over $\mathbb{K}$. Each polynomial f in $\mathbb{K}[X_1, \ldots, X_n]$ defines a *polynomial function* $f : \mathbb{K}^n \to \mathbb{K}$, where the value at the point $a = (a_1, \ldots, a_n) \in \mathbb{K}^n$ is obtained by substituting $X_i = a_i$, $1 \leq i \leq n$, and evaluating the resulting expression in $\mathbb{K}$. More precisely, if $f = \sum_\alpha c_\alpha X^\alpha$, $c_\alpha \in \mathbb{K}$, then

$$f(a_1, \ldots, a_n) = \sum_\alpha c_\alpha a^\alpha, \quad a^\alpha = a_1^{\alpha_1} \cdots a_n^{\alpha_n}. \tag{2.1}$$

This amounts to a ring homomorphism which assigns to each polynomial $f \in \mathbb{K}[X_1, \ldots, X_n]$ its polynomial function $f : \mathbb{K}^n \to \mathbb{K}$.

Proposition 2.1. *Let $\mathbb{K}$ be an infinite field. A polynomial $f \in \mathbb{K}[X_1, \ldots, X_n]$ is the zero polynomial if and only if the corresponding polynomial function $f : \mathbb{K}^n \to \mathbb{K}$ is the zero function.*

Proof. The zero polynomial $f = 0$ gives rise to the zero polynomial function.

Conversely, we need to show that if f is the zero polynomial function, i.e., $f(a) = 0$ for all $a \in \mathbb{K}^n$, then f is the zero polynomial. In case of $n = 1$, the Fundamental theorem of algebra applies which says that each nonzero polynomial $f \in \mathbb{K}[X]$ of positive degree m has at most m roots in $\mathbb{K}$. Since $\mathbb{K}$ is infinite, the assumption that $f(a) = 0$ for all $a \in \mathbb{K}$ is only satisfied by the zero polynomial.

Let $n \geq 1$. Take a polynomial f in $\mathbb{K}[X_1, \ldots, X_n, X_{n+1}]$. Write f as a polynomial in X_{n+1} with coefficients in $\mathbb{K}[X_1, \ldots, X_n]$; that is,

$$f = \sum_{i=0}^{N} h_i(X_1, \ldots, X_n) X_{n+1}^i,$$

where $h_i \in \mathbb{K}[X_1, \ldots, X_n]$. Let $(a_1, \ldots, a_n) \in \mathbb{K}^n$. Then $f(a_1, \ldots, a_n, X_{n+1}) \in \mathbb{K}[X_{n+1}]$. In view of the case $n = 1$, $f(a_1, \ldots, a_n, X_{n+1})$ is the zero polynomial. Thus $h_i(a_1, \ldots, a_n) = 0$ for $0 \leq i \leq N$. Since $(a_1, \ldots, a_n)$ was chosen arbitrarily, each h_i is the zero function. By induction, each h_i is the zero polynomial. Hence, f is the zero polynomial. □

Corollary 2.2. *Let $\mathbb{K}$ be an infinite field and let $f, g \in \mathbb{K}[X_1, \ldots, X_n]$. Then $f = g$ in $\mathbb{K}[X_1, \ldots, X_n]$ if and only if f, g define the same polynomial functions.*

Proof. Suppose $f, g \in \mathbb{K}[X_1, \ldots, X_n]$ give rise to the same polynomial function. Then the polynomial $f - g$ vanishes at all points in $\mathbb{K}^n$. By Prop. 2.1, $f - g$ is the zero polynomial and hence $f = g$. The converse is clear. □

The situation is different for finite fields. For instance, all elements of the finite field $\mathbb{F}_q$ with q elements are zeros of the polynomial $X^q - X$.

The objects studied in affine algebraic geometry are the subsets of the affine space defined by one or more polynomial equations. For instance, in the Euclidean space $\mathbb{R}^3$, consider the cone given by the set of triples (x, y, z) that satisfy the equation $X^2 + Y^2 = Z$ (Fig. 2.1).

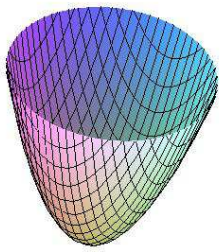


Fig. 2.1. Cone in Euclidean 3-space.

Note that any polynomial equation $f = g$ can be rewritten as $f - g = 0$. Thus it will be customary to write all equations in the form $f = 0$. More generally, the simultaneous solutions of a system of polynomial equations are considered.

Let S be a set of polynomials in $\mathbb{K}[X_1, \ldots, X_n]$. The set of all simultaneous solutions $(a_1, \ldots, a_n) \in \mathbb{K}^n$ of the system of equations

$$f(a_1, \ldots, a_n) = 0, \quad f \in S, \tag{2.2}$$

is the *affine variety* defined by S and is denoted by $\mathcal{V}(S)$. In particular, if $S = \{f_1, \ldots, f_s\}$ is a finite set, we also write $\mathcal{V}(S) = \mathcal{V}(f_1, \ldots, f_s)$. A subset V of $\mathbb{K}^n$ is an *affine variety* if $V = \mathcal{V}(S)$ for some set S of polynomials in $\mathbb{K}[X_1, \ldots, X_n]$. For instance, we have $\mathcal{V}(\{1\}) = \emptyset$ and $\mathcal{V}(\{0\}) = \mathbb{K}^n$, and thus both the empty set and the affine space $\mathbb{K}^n$ are affine varieties (Fig. 2.2).

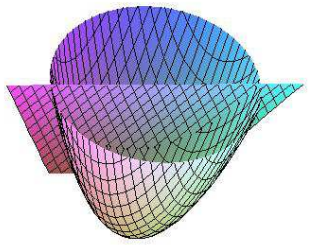


Fig. 2.2. Intersection of cone and plane in Euclidean 3-space.

Example 2.3 (Maple). The cubic plane curve $Y^2 = X^2(X+1)$ in $\mathbb{R}^2$ can be generated by using the command (Fig. 2.3)

```
> with(plots):
> plot([sqrt(x^2*(x+1),-sqrt(x^2*(x+1))], x=-1..10);
```

♢

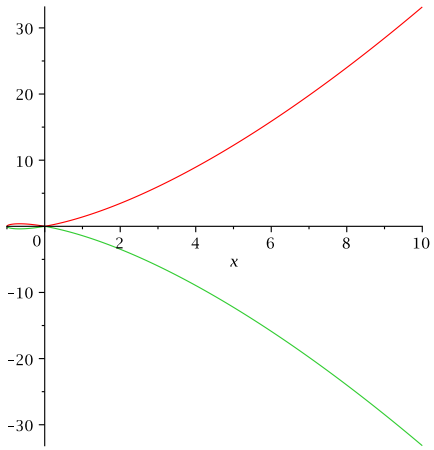


Fig. 2.3. Cubic plane curve.

Proposition 2.4. *If S and S' are subsets of $\mathbb{K}[X_1, \ldots, X_n]$ such that $S \subseteq S'$, then $\mathcal{V}(S') \subseteq \mathcal{V}(S)$.*

If there is more than one defining equation, the resulting affine variety can be considered as an intersection of other varieties.

On the other hand, the set $W = \mathbb{R}\setminus\{0,1,2,3\}$ is not an affine variety. Indeed, a polynomial $f \in \mathbb{R}[X]$ that vanishes at every point in W has infinitely many roots. Since a polynomial in $\mathbb{K}[X]$ of degree $n \geq 1$ at most n zeros in $\mathbb{K}$, the polynomial f must be the zero polynomial. Hence, the smallest affine variety in $\mathbb{R}$ that contains W is the whole real line.

The study of affine varieties depends heavily on the base field. In particular, algebraic geometry over the field of real numbers has some unpleasant surprises. For instance, we have $\mathcal{V}(X^2+1) = \emptyset$ if taken over $\mathbb{R}$. On the other hand, each polynomial in $\mathbb{C}[X]$ factors completely by the Fundamental theorem of algebra and we find that $\mathcal{V}(X^2+1) = \{\pm i\}$.

Proposition 2.5. *If V is an affine variety in $\mathbb{K}^n$, there is an ideal I of $\mathbb{K}[X_1,\ldots,X_n]$ such that $V = \mathcal{V}(I)$.*

Proof. By definition, there is a subset S of $\mathbb{K}[X_1,\ldots,X_n]$ such that $V = \mathcal{V}(S)$. Let I be the ideal of $\mathbb{K}[X_1,\ldots,X_n]$ generated by the elements of S. So each element $f \in I$ can be written as $f = h_1f_1 + \ldots + h_sf_s$, where $f_1,\ldots,f_s \in S$ and $h_1,\ldots,h_s \in \mathbb{K}[X_1,\ldots,X_n]$. Thus for each point $a \in \mathcal{V}(S)$, $f_1(a) = \ldots = f_s(a) = 0$ and thus $f(a) = 0$. Hence, $\mathcal{V}(S) \subseteq \mathcal{V}(I)$. Conversely, S is a subset of I and thus $\mathcal{V}(S) \supseteq \mathcal{V}(I)$. □

Theorem 2.6 (Weak Nullstellensatz). *Let $\mathbb{K}$ be an algebraically closed field. If I is a proper ideal of $\mathbb{K}[X_1,\ldots,X_n]$, the affine variety $\mathcal{V}(I)$ is nonempty.*

Proof. Each proper ideal I in $\mathbb{K}[X]$ is generated by a single polynomial $f \in \mathbb{K}[X]$; this can be shown by using the divison theorem. Thus we have $I = \langle f\rangle$. The Fundamental theorem of algebra says that if $\mathbb{K}$ is algebraically closed, each nonconstant polynomial f has a zero in $\mathbb{K}$. It follows that $\mathcal{V}(I) \neq \emptyset$.

Assume that the result holds for the proper ideals of $\mathbb{K}[X_2,\ldots,X_n]$. Take an ideal I of $\mathbb{K}[X_1,\ldots,X_n]$ for which $\mathcal{V}(I) = \emptyset$. By Hilbert's basis theorem, I is finitely generated and so $I = \langle f_1,\ldots,f_s\rangle$ for some $f_1,\ldots,f_s \in \mathbb{K}[X_1,\ldots,X_n]$. Suppose $f_1 \neq 0$ is not a constant polynomial; otherwise, $I = \mathbb{K}[X_1,\ldots,X_n]$. Write f_1 as a polynomial in $\mathbb{K}[X_2,\ldots,X_n][X_1]$; that is,

$$f_1(X_1,\ldots,X_n) = cX_1^N + \text{terms in which } X_1 \text{ has degree} < N,$$

where $0 \neq c \in \mathbb{K}[X_2,\ldots,X_n]$. Consider the following nonsingular linear change of coordinates,

$$\begin{aligned} X_1 &= Y_1,\\ X_2 &= Y_2 + a_2Y_1, \quad a_2 \in \mathbb{K},\\ &\vdots\\ X_n &= Y_n + a_nY_1, \quad a_n \in \mathbb{K}. \end{aligned}$$

By this setting, we obain

$$\begin{aligned} f_1(X_1,\ldots,X_n) &= f_1(Y_1, Y_2 + a_1Y_1,\ldots,Y_n + a_nY_1)\\ &= c(a_2,\ldots,a_n)Y_1^N + \text{terms in which } Y_1 \text{ has degree} < N, \end{aligned}$$

where $c(a_2,\ldots,a_n)$ is a nonzero polynomial expression in $a_2,\ldots,a_n$. Since $\mathbb{K}$ is algebraically closed, $\mathbb{K}$ is infinite. Thus by Prop. 2.1, $a_2,\ldots,a_n$ can be chosen such that $c(a_2,\ldots,a_n) \neq 0$.

Under this linear transformation, each polynomial $f \in \mathbb{K}[X_1, \ldots, X_n]$ becomes a polynomial $\hat{f} \in \mathbb{K}[Y_1, \ldots, Y_n]$. Moreover, the ideal I passes to the ideal $\hat{I} = \langle \hat{f}_1, \ldots, \hat{f}_s \rangle$ which also satisfies $\mathcal{V}(\hat{I}) = \emptyset$. Furthermore, the polynomial f_1 transforms into

$$\hat{f}_1(Y_1, \ldots, Y_n) = c(a_2, \ldots, a_n) Y_1^{\hat{N}} + \text{terms in which } Y_1 \text{ has degree} < \hat{N},$$

where $c(a_2, \ldots, c_n) \neq 0$.

Take the projection mapping $\pi_1 : \mathbb{K}^n \to \mathbb{K}^{n-1} : (a_1, a_2, \ldots, a_n) \mapsto (a_2, \ldots, a_n)$ and put $\hat{I}_1 = \hat{I} \cap \mathbb{K}[Y_2, \ldots, Y_n]$. As the leading coefficient of the polynomial $\hat{f}_1$ is a constant, the Extension theorem implies that partial solutions in $\mathbb{K}^{n-1}$ always extend; that is, $\mathcal{V}(\hat{I}_1) = \pi_1(\mathcal{V}(\hat{I}))$. It follows that $\mathcal{V}(\hat{I}_1) = \pi_1(\mathcal{V}(\hat{I})) = \pi_1(\emptyset) = \emptyset$. By induction, we have $\hat{I}_1 = \mathbb{K}[Y_2, \ldots, Y_n]$. Thus $1 \in \hat{I}_1 \subseteq \hat{I}$ and hence $I = \mathbb{K}[X_1, \ldots, X_n]$. □

Example 2.7 (Singular). The substitution defined in the proof is a ring homomorphism given as follows:

```
> ring r = 0, (x,y,z), dp;
> poly f = x2yz+xy+z2;
> ring s = 0, (u,v,w), dp;
> map F = r, u, 2u+v, 3u+w // map F from ring r to ring s
                           // x -> x, y -> 2u+v, z -> 3u+w
> poly g = F(f);           // apply F
6u^4+3u3v+2u3w+u2vw+11u2+uv+6uw+w2
```

♢

Example 2.8 (Singular). Consider the ideal $I = \langle XY - Y, Y^2 - X^2, X - Y^3 \rangle$ in $\mathbb{Q}[X, Y]$.

```
> ring r = 0, (x,y), dp;
> ideal i = xy-y, y2-x2, x-y3;
> std(i);
_[1]=x-y
_[2]=y2-y
```

The reduced Groebner basis is $G = \{X - Y, Y^2 - Y\}$ and thus the variety $\mathcal{V}(I)$ consists of three points in $\mathbb{C}^2$, namely $\mathcal{V}(I) = \mathcal{V}(G) = \{(0, 0), (1, 1), (-1, -1)\}$. ♢

2.2 Ideal-Variety Correspondence

We set up a dictionary that allows to relate geometric properties to algebraic ones.

Proposition 2.9. *If I and J are ideals of $\mathbb{K}[X_1, \ldots, X_n]$, then $\mathcal{V}(I + J) = \mathcal{V}(I) \cap \mathcal{V}(J)$.*

Proof. By Prop. 1.3, $I + J$ is an ideal. Since I and J are subsets of $I + J$, it follows from Prop. 2.4 that $\mathcal{V}(I + J) \subseteq \mathcal{V}(I)$ and $\mathcal{V}(I + J) \subseteq \mathcal{V}(J)$. Therefore, $\mathcal{V}(I + J) \subseteq \mathcal{V}(I) \cap \mathcal{V}(J)$.

Conversely, let $(a_1, \ldots, a_n) \in \mathcal{V}(I) \cap \mathcal{V}(J)$ and let $h \in I + J$. Then there are polynomials $f \in I$ and $g \in J$ such that $h = f + g$. By hypothesis, $f(a_1, \ldots, a_n) = 0$ and $g(a_1, \ldots, a_n) = 0$. Thus $h(a_1, \ldots, a_n) = 0$ and hence $(a_1, \ldots, a_n) \in \mathcal{V}(I + J)$. □

Example 2.10 (Singular). In the Euclidean space $\mathbb{R}^3$, consider the surfaces $\mathcal{V}(Y-X^2)$ and $\mathcal{V}(Z-X^3)$. Their intersection yields an interesting curve, the twisted cubic $V = \mathcal{V}(Y - X^2, Z - X^3)$ where

$$\begin{aligned} V = \mathcal{V}(Y - X^2, Z - X^3) &= \mathcal{V}(Y - X^2) \cap \mathcal{V}(Z - X^3) \\ &= \{(x, x^2, z) \mid x, z \in \mathbb{R}\} \cap \{(x, y, x^3) \mid x, y \in \mathbb{R}\} \\ &= \{(x, x^2, x^3) \mid x \in \mathbb{R}\}. \end{aligned}$$

The latter representation is a *parametrization* of V which provides a way to draw the curve (Fig. 2.4). However, not every affine variety can be parametrized in this way. ♢

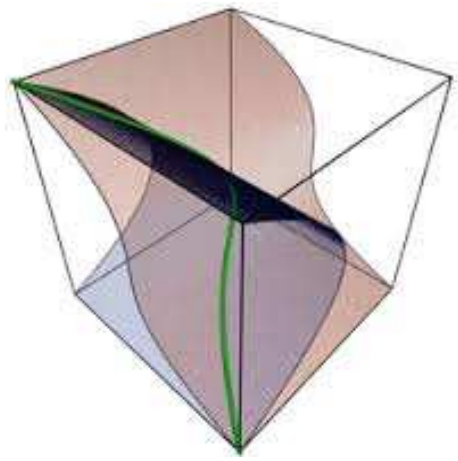


Fig. 2.4. Twisted cubic.

Proposition 2.11. *If I and J are ideals of $\mathbb{K}[X_1, \ldots, X_n]$, then $\mathcal{V}(I \cdot J) = \mathcal{V}(I) \cup \mathcal{V}(J)$.*

Proof. Let $(a_1, \ldots, a_n) \in \mathcal{V}(I \cdot J)$. By definition, the ideal $I \cdot J$ is generated by elements of the form $f \cdot g$, where $f \in I$ and $g \in J$. It follows that $(f \cdot g)(a_1, \ldots, a_n) = f(a_1, \ldots, a_n) \cdot g(a_1, \ldots, a_n) = 0$. Thus $(a_1, \ldots, a_n) \in \mathcal{V}(I)$ or $(a_1, \ldots, a_n) \in \mathcal{V}(J)$ and hence $(a_1, \ldots, a_n) \in \mathcal{V}(I) \cup \mathcal{V}(J)$.

Conversely, let $(a_1, \ldots, a_n) \in \mathcal{V}(I) \cup \mathcal{V}(J)$. Assume that $(a_1, \ldots, a_n) \in \mathcal{V}(I)$. Then $f(a_1, \ldots, a_n) = 0$ for all polynomials $f \in I$ and so $f(a_1, \ldots, a_n) \cdot g(a_1, \ldots, a_n) = 0$ for each polynomial $g \in J$. It follows that $(f \cdot g)(a_1, \ldots, a_n) = 0$. But the ideal $I \cdot J$ is generated by elements of the form $f \cdot g$, where $f \in I$ and $g \in J$. Therefore, $(a_1, \ldots, a_n) \in \mathcal{V}(I \cdot J)$. □

Example 2.12 (Singular). The ideals $I = \langle X, Y \rangle$ and $J = \langle Z \rangle$ give rise to the (x, y)-plane $\mathcal{V}(X, Y)$ and the z-axis $\mathcal{V}(Z)$, respectively. The product ideal $IJ = \langle XZ, YZ \rangle$ provides the union of the (x, y)-plane and the z-axis.

```
> ring r = 0, (x,y,z), dp;
> ideal i = x,y;
```

```
> ideal j = z;
> ideal ij = i*j;
> std(ij);
_[1]=yz
_[2]=xz
```

♢

Proposition 2.13. *If I and J are ideals of $\mathbb{K}[X_1, \ldots, X_n]$, then $\mathcal{V}(I \cap J) = \mathcal{V}(I) \cup \mathcal{V}(J)$.*

Proof. Let $(a_1, \ldots, a_n) \in \mathcal{V}(I) \cup \mathcal{V}(J)$. Assume that $(a_1, \ldots, a_n) \in \mathcal{V}(I)$. Then $f(a_1, \ldots, a_n) = 0$ for each polynomial $f \in I$. Thus $f(a_1, \ldots, a_n) = 0$ for all polynomials $f \in I \cap J$ and hence $(a_1, \ldots, a_n) \in \mathcal{V}(I \cap J)$.

Conversely, we have $I \cdot J \subseteq I \cap J$ by Prop. 1.4 and thus $\mathcal{V}(I \cap J) \subseteq \mathcal{V}(I \cdot J)$ by Prop. 2.4. But $\mathcal{V}(I \cdot J) = \mathcal{V}(I) \cup \mathcal{V}(J)$ by Prop. 2.11 and so $\mathcal{V}(I \cap J) \subseteq \mathcal{V}(I) \cup \mathcal{V}(J)$. □

Proposition 2.14. *If $\mathbb{K}$ is algebraically closed, each finite subset of the affine space $\mathbb{K}^n$ is an affine variety.*

Proof. By Ex. 1.9, each maximal ideal of $\mathbb{K}[X_1, \ldots, X_n]$ has the form $\mathfrak{m}_a = \langle X_1 - a_1, \ldots, X_n - a_n \rangle$ for some point $a = (a_1, \ldots, a_n) \in \mathbb{K}^n$. Take the ideal I given by the sum of maximal ideals $\mathfrak{m}_a$, where a runs over the elements of V. Since $\mathcal{V}(\mathfrak{m}_a) = \{a\}$ for each point $a \in \mathbb{K}^n$, we have $V = \mathcal{V}(I)$. □

We associate to each affine variety V in $\mathbb{K}^n$ the collection of polynomials $\mathcal{I}(V)$ that vanish at every point of V, i.e.,

$$\mathcal{I}(V) = \{f \in \mathbb{K}[X_1, \ldots, X_n] \mid f(a_1, \ldots, a_n) = 0 \text{ for all } (a_1, \ldots, a_n) \in V\} \tag{2.3}$$

The set $\mathcal{I}(V)$ called the *ideal of* V.

Proposition 2.15. *If V is an affine variety in $\mathbb{K}^n$, then $\mathcal{I}(V)$ is an ideal of $\mathbb{K}[X_1, \ldots, X_n]$.*

Proof. Let $f, g \in \mathcal{I}(V)$ and $h \in \mathbb{K}[X_1, \ldots, X_n]$. For each $(a_1, \ldots, a_n) \in V$, we have $(f-g)(a_1, \ldots, a_n) = f(a_1, \ldots, a_n) - g(a_1, \ldots, a_n) = 0$ and $(f \cdot h)(a_1, \ldots, a_n) = f(a_1, \ldots, a_n) \cdot h(a_1, \ldots, a_n) = 0$. Therefore, $f - g$ and $f \cdot h \in \mathcal{I}(V)$. □

For instance, we have $\mathcal{I}(\emptyset) = \mathbb{K}[X_1, \ldots, X_n]$. If $\mathbb{K}$ is infinite, then by Prop. 2.1, $\mathcal{I}(\mathbb{K}^n) = \{0\}$. If $\mathbb{K}$ is algebraically closed, for each point $(a_1, \ldots, a_n) \in \mathbb{K}^n$, $\mathcal{I}(\{(a_1, \ldots, a_n)\}) = \langle X_1 - a_1, \ldots, X_n - a_n \rangle$ by the proof of Prop. 2.14. For example, $\mathcal{I}(\{i\}) = \langle x - i \rangle$ over $\mathbb{C}$, but $\mathcal{I}(\{i\}) = \langle x^2 + 1 \rangle$ over $\mathbb{R}$.

Proposition 2.16. *If V and V' are affine varieties in $\mathbb{K}^n$ such that $V \subseteq V'$, then $\mathcal{I}(V') \subseteq \mathcal{I}(V)$.*

If $V = \mathcal{V}(I)$, is it always true that $\mathcal{I}(V) = I$? The answer is no, as the following simple example demonstrates. Consider the ideal $I = \langle X^2 \rangle$ in $\mathbb{R}[X, Y]$ that consists of all polynomials divisible by X^2. The corresponding affine variety $V = \mathcal{V}(X^2)$ is given by the x-axis. Therefore, the ideal $\mathcal{I}(V) = \langle X \rangle$ is generated by the polynomial X and hence the ideal $\mathcal{I}(\mathcal{V}(I))$ is strictly larger than I. The reason is that the ideal $\langle X^2 \rangle$ is not radical, but $\langle X \rangle$ has this property.

Proposition 2.17. *If V is an affine variety in $\mathbb{K}^n$, the ideal $\mathcal{I}(V)$ is radical.*

Proof. Write $I = \mathcal{I}(V)$. Let $f \in \mathbb{K}[X_1, \ldots, X_n]$ such that $f^m \in I$ for some integer $m \geq 1$. Then the polynomial f^m vanishes at each point of V. But $0 = f^m(a) = f(a)^m$ for each point $a \in \mathbb{K}^n$ and so f vanishes at each point of V. Thus $f \in I$ and hence the ideal I is radical. □

The properties of the field $\mathbb{K}$ also affect the relation between an ideal I in $\mathbb{K}[X_1, \ldots, X_n]$ and the corresponding ideal $\mathcal{I}(\mathcal{V}(I))$. For instance, over the field of real numbers, we have $\mathcal{V}(X^2 + 1) = \emptyset$ and thus $\mathcal{I}(\mathcal{V}(X^2+1)) = \mathbb{R}[X]$. On the other hand, if we take the field of complex numbers, each polynomial in $\mathbb{C}[X]$ factors completely by the Fundamental theorem of algebra. We find that $\mathcal{V}(X^2 + 1) = \{\pm i\}$ and thus $\mathcal{I}(\mathcal{V}(X^2 + 1)) = \langle X^2 + 1\rangle$.

Proposition 2.18. *If I is an ideal of $\mathbb{K}[X_1, \ldots, X_n]$, then $\sqrt{I} \subseteq \mathcal{I}(\mathcal{V}(I))$.*

Proof. Let $f \in \sqrt{I}$. Then there exists an integer $m \geq 1$ such that $f^m \in I$. Thus for each point $a \in \mathcal{V}(I)$, we have $0 = f^m(a) = f(a)^m$ and thus $f(a) = 0$. Hence, $f \in \mathcal{I}(\mathcal{V}(I))$. □

Theorem 2.19 (Strong Nullstellensatz). *If $\mathbb{K}$ is an algebraically closed field and I is an ideal of $\mathbb{K}[X_1, \ldots, X_n]$, then*

$$\mathcal{I}(\mathcal{V}(I)) = \sqrt{I}. \tag{2.4}$$

The proof is given by the *Rabinowitsch trick*, a short way to show the Nullstellensatz.

Proof. Let I be an ideal of $\mathbb{K}[X_1, \ldots, X_n]$. By Hilbert's basis theorem, there are polynomials $f_1, \ldots, f_s \in \mathbb{K}[X_1, \ldots, X_n]$ such that $I = \langle f_1, \ldots, f_s\rangle$. Take $f \in \mathcal{I}(\mathcal{V}(I))$. Then $f(a_1, \ldots, a_n) = 0$ for each point $(a_1, \ldots, a_n) \in \mathcal{V}(I)$. Consider the ideal $\hat{I} = \langle f_1, \ldots, f_s, 1 - Y \cdot f\rangle$ of $\mathbb{K}[X_1, \ldots, X_n, Y]$.

Claim that $\mathcal{V}(\hat{I}) = \emptyset$. Indeed, let $(a_1, \ldots, a_n, a_{n+1}) \in \mathbb{K}^{n+1}$. If $(a_1, \ldots, a_n) \in \mathcal{V}(I)$, then $f(a_1, \ldots, a_n) = 0$. It follows that $1 - Y \cdot f$, evaluated at the point $(a_1, \ldots, a_n, a_{n+1})$, has the value $1 - a_{n+1}f(a_1, \ldots, a_n) = 1$ and so $(a_1, \ldots, a_n, a_{n+1}) \notin \mathcal{V}(\hat{I})$. If $(a_1, \ldots, a_n) \notin \mathcal{V}(I)$, there is an index i, $1 \leq i \leq s$, such that $f_i(a_1, \ldots, a_n) \neq 0$. Think of f_i as a polynomial in $n + 1$ variables. Then $f_i(a_1, \ldots, a_n, a_{n+1}) \neq 0$ and hence $(a_1, \ldots, a_n, a_{n+1}) \notin \mathcal{V}(\hat{I})$. This proves the claim.

Since $\mathcal{V}(\hat{I}) = \emptyset$, the Weak Nullstellensatz implies that $1 \in \hat{I}$. Thus there are polynomials $h, h_1, \ldots, h_s$ in $\mathbb{K}[X_1, \ldots, X_n, Y]$ such that

$$1 = h_1 \cdot f_1 + \ldots + h_s \cdot f_s + h \cdot (1 - Y \cdot f).$$

Put $Y = 1/f$. Then we obtain an equation of rational functions

$$1 = \sum_i h_i(X_1, \ldots, X_n, 1/f) \cdot f_i.$$

Multiply both sides by a power f^m, where m is sufficiently large to clear the denominators. The resulting equation has the form $f^m = \sum_i \hat{h}_i f_i \in I$, where $\hat{h}_i \in \mathbb{K}[X_1, \ldots, X_n]$, $1 \leq i \leq s$. Thus f lies in the radical ideal $\sqrt{I}$ and hence we have shown that $\mathcal{I}(\mathcal{V}(I))$ is a subset of $\sqrt{I}$. By Prop. 2.18, the result follows. □

Example 2.20 (Singular). The ideal $I = \langle Y - X^2, Z - X^3\rangle$ in $\mathbb{Q}[X, Y, Z]$ defining the twisted cubic curve is radical as the following computation shows.

```
> ring r = 0, (x,y,z), dp;
> ideal i = y-x2, z-x3;
> LIB "primdec.lib";
> std(i);
_[1]=y2-xz
_[2]=xy-z
_[3]=x2-y
> ideal j = radical(std(i));
> std(j);
_[1]=y2-xz
_[2]=xy-z
_[3]=x2-y
```

♢

Theorem 2.21 (Ideal-Variety Correspondence).

- *Let $\mathbb{K}$ be an arbitrary field. The maps*

$$\textit{affine varieties in } \mathbb{K}^n \xrightarrow{\mathcal{I}} \textit{ideals in } \mathbb{K}[X_1, \ldots, X_n]$$

and

$$\textit{ideals in } \mathbb{K}[X_1, \ldots, X_n] \xrightarrow{\mathcal{V}} \textit{affine varieties in } \mathbb{K}^n$$

are inclusion-reversing, and $\mathcal{V}(\mathcal{I}(V)) = V$ for each affine variety V in $\mathbb{K}^n$.
- *Let $\mathbb{K}$ be an algebraically closed field. The maps*

$$\textit{affine varieties in } \mathbb{K}^n \xrightarrow{\mathcal{I}} \textit{radical ideals in } \mathbb{K}[X_1, \ldots, X_n]$$

and

$$\textit{radical ideals in } \mathbb{K}[X_1, \ldots, X_n] \xrightarrow{\mathcal{V}} \textit{affine varieties in } \mathbb{K}^n$$

are inclusion-reversing bijections and inverses of each other. In particular, $\mathcal{I}(\mathcal{V}(I)) = I$ for each radical ideal I of $\mathbb{K}[X_1, \ldots, X_n]$.

Proof. First, Prop. 2.4 and 2.16 show that the maps are inclusion-reversing.

Second, let V be an affine variety in $\mathbb{K}^n$. Then there is an ideal I in $\mathbb{K}[X_1, \ldots, X_n]$ such that $V = \mathcal{V}(I)$. By Hilbert's basis theorem, there are polynomials $f_1, \ldots, f_s$ in $\mathbb{K}[X_1, \ldots, X_n]$ such that $I = \langle f_1, \ldots, f_s \rangle$. Let $(a_1, \ldots, a_n) \in V$. Then $f_1(a_1, \ldots, a_n) = \cdots = f_s(a_1, \ldots, a_n) = 0$. But each element $h \in I$ is of the form $h = h_1 f_1 + \ldots + h_s f_s$, where $h_i \in \mathbb{K}[X_1, \ldots, X_n]$, $1 \leq i \leq s$. Thus $h(a_1, \ldots, a_n) = 0$ and hence I is contained in $\mathcal{I}(V)$. Then by Prop. 2.16, $\mathcal{V}(\mathcal{I}(V))$ is contained in $V = \mathcal{V}(I)$. Conversely, let $(a_1, \ldots, a_n) \in V$. Then $f(a_1, \ldots, a_n) = 0$ for each polynomial $f \in \mathcal{I}(V)$ and so $(a_1, \ldots, a_n) \in \mathcal{V}(\mathcal{I}(V))$. It follows that V is a subset of $\mathcal{V}(\mathcal{I}(V))$. Hence, $\mathcal{V}(\mathcal{I}(V)) = V$.

Third, the Strong Nullstellensatz says that $\mathcal{I}(\mathcal{V}(I)) = I$ for each radical ideal I in $\mathbb{K}[X_1, \ldots, X_n]$. Moreover, the first part exhibits that $\mathcal{V}(\mathcal{I}(V)) = V$ for each affine variety V in $\mathbb{K}^n$. This shows that the mappings are bijective and inverses of each other. □

2.3 Zariski Topology

The affine varieties of the affine space $\mathbb{K}^n$ form the closed sets of a topology on $\mathbb{K}^n$. Recall that a family of subsets Ξ of a set X is a *topology* on X if both the empty set and the whole set X are elements of Ξ, any intersection of finitely many elements of Ξ is an element of Ξ, and any union of elements of Ξ is an element of Ξ. The members of Ξ are the *open sets* in X and their complements in X are the *closed sets* in X.

Proposition 2.22. *Let I and J be ideals and let (I_j) be a family of ideals of $\mathbb{K}[X_1, \ldots, X_n]$.*

- $\mathcal{V}(0) = \mathbb{K}^n$ *and* $\mathcal{V}(\mathbb{K}[X_1, \ldots, X_n]) = \emptyset$.
- $\mathcal{V}(IJ) = \mathcal{V}(I) \cup \mathcal{V}(J)$.
- $\mathcal{V}(\sum_j I_j) = \bigcap_j \mathcal{V}(I_j)$.

Proof. The first assertion is obvious. The second assertion is Prop.2.11. Finally, the ideal $\sum_j I_j$ consists of all finite sums of the form $\sum_j f_j$, where $f_j \in I_j$. It follows that $\mathcal{V}(\sum_j I_j)$ is equal to the intersection of all affine varieties $\mathcal{V}(I_j)$. □

These assertions show that the affine varieties in $\mathbb{K}^n$ are the closed sets in $\mathbb{K}^n$: The whole space and the empty set are closed, the finite union of closed sets is closed, and the arbitrary intersection of closed sets is closed. This topology is the *Zariski topology* on $\mathbb{K}^n$. By Prop. 2.14, each finite subset of $\mathbb{K}^n$ is closed.

Proposition 2.23. *If W is a subset of $\mathbb{K}^n$, then the set $\mathcal{V}(\mathcal{I}(W))$ is the smallest affine variety that contains W.*

Proof. By Prop. 2.15, the set $\mathcal{I}(W)$ is an ideal of $\mathbb{K}[X_1, \ldots, X_n]$ and thus by definition $\mathcal{V}(\mathcal{I}(W))$ is an affine variety in $\mathbb{K}^n$.

Claim that W is a subset of $\mathcal{V}(\mathcal{I}(W))$. Indeed, if $(a_1, \ldots, a_n) \in W$, each polynomial in $\mathcal{I}(W)$ vanishes at the point $(a_1, \ldots, a_n)$ and thus the point $(a_1, \ldots, a_n)$ belongs to $\mathcal{V}(\mathcal{I}(W))$.

Claim that each affine variety V in $\mathbb{K}^n$ with $W \subseteq V$ satisfies $\mathcal{V}(\mathcal{I}(W)) \subseteq V$. Indeed, if $W \subseteq V$, then by Prop. 2.16, $\mathcal{I}(V) \subseteq \mathcal{I}(W)$ and by Prop. 2.4, $\mathcal{V}(\mathcal{I}(W)) \subseteq \mathcal{V}(\mathcal{I}(V))$. But V is an affine variety and therefore by Thm. 2.21, $\mathcal{V}(\mathcal{I}(V)) = V$. □

The *Zariski closure* of a subset W of $\mathbb{K}^n$ is the smallest affine variety containing it. By Prop. 2.23, the Zariski closure of W equals $\mathcal{V}(\mathcal{I}(W))$. For instance, we have seen that the Zariski closure of the set $W = \mathbb{R} \setminus \{0, 1, 2, 3\}$ is the real line $\mathbb{R}$. By the ideal-variety correspondence, if V is an affine variety, then $\mathcal{V}(\mathcal{I}(V)) = V$ and hence each affine variety equals its Zariski closure.

2.4 Irreducible Affine Varieties

Affine varieties can be decomposed into irreducible components which can then be studied separately. An affine variety V in the affine space $\mathbb{K}^n$ is *irreducible* if in each expression of V as a union of affine varieties $V = V_1 \cup V_2$, either $V = V_1$ or $V = V_2$.

Proposition 2.24. *An affine variety V in $\mathbb{K}^n$ is irreducible if and only if the ideal $\mathcal{I}(V)$ is prime in $\mathbb{K}[X_1, \ldots, X_n]$.*

Proof. Let V be irreducible and let $fg \in \mathcal{I}(V)$. Put $V_1 = V \cap \mathcal{V}(f)$ and $V_2 = V \cap \mathcal{V}(g)$. By Prop. 2.9, the intersection of affine varieties is an affine variety and thus V_1 and V_2 are also affine varieties. By Prop. 2.4, $\mathcal{V}(fg) \supseteq \mathcal{V}(\mathcal{I}(V))$. Moreover, since $V \subseteq \mathcal{V}(\mathcal{I}(V))$, we have $V = V \cap \mathcal{V}(fg)$. It follows by Prop. 2.11, $V = V \cap (\mathcal{V}(f) \cup \mathcal{V}(g)) = (V \cap \mathcal{V}(f)) \cup (V \cap \mathcal{V}(g)) = V_1 \cup V_2$. But V is irreducible and so $V = V_1$ or $V = V_2$. Without loss of generality, let $V = V_1$. Then the polynomial f vanishes on V and thus $f \in \mathcal{I}(V)$. Hence, the ideal $\mathcal{I}(V)$ is prime.

Conversely, suppose V is reducible. Then there are affine varieties V_1 and V_2 contained in V such that $V = V_1 \cup V_2$. By Prop. 2.16, we have $\mathcal{I}(V) \subseteq \mathcal{I}(V_1)$ and $\mathcal{I}(V) \subseteq \mathcal{I}(V_2)$. By the ideal-variety correspondence, we have $V_1 = \mathcal{V}(\mathcal{I}(V_1))$, $V_2 = \mathcal{V}(\mathcal{I}(V_2))$, and $V = \mathcal{V}(\mathcal{I}(V))$. Since $V_1 \neq V$ and $V_2 \neq V$, it follows that $\mathcal{I}(V_1) \neq \mathcal{I}(V)$ and $\mathcal{I}(V_2) \neq \mathcal{I}(V)$. Take polynomials $f \in \mathcal{I}(V_1) \backslash \mathcal{I}(V)$ and $g \in \mathcal{I}(V_2) \backslash \mathcal{I}(V)$. Then $fg \in \mathcal{I}(V_1 \cup V_2) = \mathcal{I}(V)$ and hence $\mathcal{I}(V)$ is not prime. □

Example 2.25. The ideals $I_1 = \langle X, Y \rangle$ and $I_2 = \langle Z \rangle$ are prime in $\mathbb{R}[X, Y, Z]$. Thus the corresponding affine varieties, the (x, y)-plane $\mathcal{V}(I_1) = \{(x, y, 0) \mid x, y \in \mathbb{R}\}$ and the z-axis $\mathcal{V}(I_2) = \{(0, 0, z) \mid z \in \mathbb{R}\}$, are irreducible. ♢

Proposition 2.26. *Each affine variety V in $\mathbb{K}^n$ can be uniquely written (up to permutation) in the form*

$$V = V_1 \cup \ldots \cup V_m,$$

where $V_1, \ldots, V_m$ are pairwise distinct irreducible affine varieties.

Proof. Let V be an affine variety that cannot be written as a finite union of irreducible affine varieties. Then V is reducible with $V = V_1 \cup V_1'$ such that $V_1 \neq V$ and $V_1' \neq V$. Furthermore, at least one of V_1 and V_1' cannot be described as a union of irreducible affine varieties. Assume that V_1 is not a union of irreducible affine varieties. Then again write $V_1 = V_2 \cup V_2'$, where $V_2 \neq V_1$ and $V_2' \neq V_1$. Continuing in this way, we obtain an infinite descending sequence of affine varieties $V \supset V_1 \supset V_2 \supset \ldots$. By the ideal-variety correspondence, the operator $\mathcal{I}$ is one-to-one and thus gives rise to an ascending sequence of ideals $\mathcal{I}(V) \subset \mathcal{I}(V_1) \subset \mathcal{I}(V_2) \subset \ldots$. By the ascending chain condition 1.28, this chain must become stationary contradicting the infiniteness of the sequence of affine varieties.

Assume there are two such expressions $V = U_1 \cup \ldots \cup U_r = W_1 \cup \ldots \cup W_s$. Consider $U_1 = V \cap U_1 = (W_1 \cap U_1) \cup \ldots \cup (W_s \cap U_1)$. Since U_1 is irreducible, there is an index j such that $U_1 = W_j \cap U_1$; that is, $U_1 \subseteq W_j$. Likewise, there is an index k such that $W_1 \subseteq U_k$. It follows that $U_1 \subseteq U_k$, which implies by hypothesis that $j = k$ and so $U_1 = W_j$ Continuing we see that $r = s$ and that one decomposition is only a renumbering of the other. □

Example 2.27. The affine variety $V = \{(x, y, z) \mid xz = yz = 0, x, y, z \in \mathbb{R}\}$ in $\mathbb{R}^3$ is reducible, since V decomposes into the union of the (x, y)-plane $V_1 = \{(x, y, 0) \mid x, y \in \mathbb{R}\}$ and the z-axis $V_2 = \{(0, 0, z) \mid z \in \mathbb{R}\}$. Both are irreducible affine varieties. ♢

2.5 Elimination Theory

We provide a straightforward method for solving systems of polynomial equations based on the elimination orderings.

Let I be an ideal in $\mathbb{K}[X_1, \ldots, X_n]$ and let $k \geq 0$ be an integer. The kth *elimination ideal* of I is given as

$$I_k = I \cap \mathbb{K}[X_{k+1}, \ldots, X_n]. \tag{2.5}$$

Note that the 0-th elimination ideal is $I_0 = I$. Clearly, I_k is an ideal of $\mathbb{K}[X_1, \ldots, X_n]$.

A monomial ordering $>$ on $\mathbb{K}[X_1, \ldots, X_n]$ has the *elimination property* for $X_1, \ldots, X_k$ if $f \in \mathbb{K}[X_1, \ldots, X_n]$ and $\text{LM}_>(f) \in \mathbb{K}[X_{k+1}, \ldots, X_n]$ implies $f \in \mathbb{K}[X_{k+1}, \ldots, X_n]$. That is, monomials which contain one of $X_1, \ldots, X_k$ are always larger than monomials which contain none of $X_1, \ldots, X_k$. A monomial ordering $>$ on $\mathbb{K}[X_1, \ldots, X_n]$ is an *elimination ordering* for $X_1, \ldots, X_k$ if it has the elimination property for $X_1, \ldots, X_k$.

For instance, the `lp` ordering has the elimination property for any sequence $X_1, \ldots, X_k$, $k \geq 0$. Product orderings provide a large class of elimination orderings. For this, let $>_1$ be a monomial ordering on $\mathbb{K}[X_1, \ldots, X_m]$ and $>_2$ be a monomial ordering on $\mathbb{K}[Y_1, \ldots, Y_n]$. Then the *product ordering* $>$ on $\mathbb{K}[X_1, \ldots, X_m, Y_1, \ldots, Y_n]$, denoted by $(>_1, >_2)$, is defined by

$$X^\alpha Y^\beta > X^\gamma Y^\delta \quad :\Longleftrightarrow \quad X^\alpha >_1 X^\gamma \vee (X^\alpha = X^\gamma \wedge Y^\beta >_2 Y^\delta).$$

The product ordering $>$ on $\mathbb{K}[X_1, \ldots, X_m, Y_1, \ldots, Y_n]$ is an elimination ordering for $X_1, \ldots, X_m$.

Example 2.28 (Singular). Product orderings can be specified by the ring definitions.

```
> ring r = 0, (w,x,y,z), (dp(2),Dp(2)); // mixed product ordering
> poly f = wx2z+w2x2yz+wx+yz2+y2z;
> f;
w2x2yz+wx2z+wx+y2z+yz2
```

♢

Theorem 2.29. (Elimination) *Let I be an ideal of $\mathbb{K}[X_1, \ldots, X_n]$ and let $k \geq 0$ be an integer. If G is a Groebner basis of I with respect to an elimination ordering $>$ for $X_1, \ldots, X_k$, the k-th elimination ideal I_k of I has the Groebner basis*

$$G_k = G \cap \mathbb{K}[X_{k+1}, \ldots, X_n].$$

Proof. Let $G = \{g_1, \ldots, g_s\}$ be a Groebner basis of I. Assume that the first $r \leq s$ elements of G lie in $\mathbb{K}[X_{k+1}, \ldots, X_n]$.

Claim that $G_k = \{g_1, \ldots, g_r\}$ is a generating set of I_k. Indeed, by definition, $G_k \subseteq I_k$ and so $\langle g_1, \ldots, g_r \rangle \subseteq I_k$. Conversely, let $f \in I_k$. Then divide f into $g_1, \ldots, g_s$ giving the remainder $f^G = 0$. In view of the given elimination ordering, the leading terms of the (eliminated) polynomials $g_{r+1}, \ldots, g_s$ must involve at least one of the variables $X_1, \ldots, X_k$ and these terms are greater than any term in f. It follows that the division of f into $g_1, \ldots, g_s$ does not involve $g_{r+1}, \ldots, g_s$ and therefore f is of the form

$$f = h_1 g_1 + \ldots + h_r g_r + 0 \cdot g_{r+1} + \ldots + 0 \cdot g_s + 0.$$

Hence, $f \in \langle g_1, \ldots, g_r \rangle$ as required.

Claim that $G_k = \{g_1, \ldots, g_r\}$ is a Groebner basis of I_k. Indeed, divide the S-polynomial $S(g_i, g_j) \in I_k$ into G_k for each pair $i \neq j$, $1 \leq i, j \leq r$. The previous paragraph shows that the remainder $S(g_i, g_j)^{G_k}$ is zero. Thus by the Buchberger S-criterion, G_k is a Groebner basis of I_k. □

Example 2.30 (Singular).

```
> ring r = 0, (w,x,y,z), lp;
> ideal i = w2,x4,y5,z3,wxyz;
> eliminate(i,z);
_[1]=w2
_[2]=x4
_[3]=y5
> eliminate(i,yz);
_[1]=w2
_[2]=x4
> eliminate(i,xyz);
_[1]=w2
```

♢

An ideal I of $\mathbb{K}[X_1, \ldots, X_n]$ generated by $f_1, f_2, \ldots, f_s$ provides a system of polynomial equations

$$f_1 = 0,\ f_2 = 0, \ldots, f_s = 0.$$

Any point $(a_1, \ldots, a_n) \in \mathcal{V}(I)$ is a *solution* of the system of equations, and any point $(a_{k+1}, \ldots, a_n)$ in $\mathcal{V}(I_k)$ is a *partial solution* of the system of equations. Each solution truncates to a partial solution, but not each partial solution extends to a solution. This is where the following Extension results comes into play. For this, note that each polynomial f in I_{k-1} can be written as a polynomial in X_k, whose coefficients are polynomials in $X_{k+1}, \ldots, X_n$,

$$f = cX_k^N + \text{terms in which } X_k \text{ has degree} < N, \tag{2.6}$$

where $0 \neq c \in \mathbb{K}[X_{k+1}, \ldots, X_n]$ is called the leading coefficient polynomial of f.

Theorem 2.31. (Extension) *Let $\mathbb{K}$ is an algebraically closed field. A partial solution $(a_{k+1}, \ldots, a_n)$ in $\mathcal{V}(I_k)$ extends to a partial solution $(a_k, a_{k+1}, \ldots, a_n)$ in $\mathcal{V}(I_{k-1})$ if the leading coefficient polynomials of the elements of the Groebner basis of I_{k-1} with respect to an eliminiation ordering with $X_1 > \ldots > X_n$ do not all vanish at $(a_{k+1}, \ldots, a_n)$.*

Note that the condition of the theorem is particularly fulfilled if the leading coefficient polynomials are constants in $\mathbb{K}$. The Elimination theorem shows that a Groebner basis G of the given ideal I with respect to the `lp` ordering eliminates successively more and more variables. This gives the following strategy for finding all solutions of the system of equations: Start with the polynomials in G with the fewest variables, solve them, and then extend these partial solutions to solutions of the whole system adding one variable at a time.

Example 2.32 (Singular). Consider the system of equations

$$\begin{aligned} X^2 + Y^2 + Z^2 &= 2, \\ X^2 + 2Y^2 &= 3, \\ XZ &= 1. \end{aligned}$$

Take the ideal $I = \langle X^2 + Y^2 + Z^2 - 2, X^2 + 2Y^2 - 3, XZ - 1 \rangle$ in $\mathbb{C}[X, Y, Z]$. First, compute a Groebner basis of I with respect to an elimination ordering.

```
> ring r = 0, (x,y,z), lp;   // lexicographical ordering
> ideal i = x2+y2+z2-2, x2+2y2-3, xz-1;
> ideal j = std(i); j;
j[1]=2z4-z2+1
j[2]=y2-z2-1
j[3]=x+2y2z-3z
```

The corresponding Groebner basis is $G = \{2Z^4 - Z^2 + 1, Y^2 - Z^2 - 1, X + 2Y^2Z - 3Z\}$. The second elimination ideal $I_2 = I \cap \mathbb{C}[Z]$ has the Groebner basis $G_2 = \{2Z^4 - Z^2 + 1\}$. The generating polynomial is irreducible which can be tested by `Maple`.

```
> with(PolynomialTools):
> factor( 2z^4-z^2-1+1 );
```

The zeros of this polynomial can be numerically found as follows.

```
> LIB "solve.lib";
> ring r2 = 0, (z), lp;
> ideal i2 = 2z4-z2+1;
> solve (i2,6);
[1]:
   (-0.691776+i*0.478073)
[2]:
   (0.691776-i*0.478073)
[3]:
   (0.691776-i*0.478073)
[4]:
   (-0.691776+i*0.478073)
```

However, these roots are *algebraic numbers* and can be symbolically established by `Maple`.

```
> with(PolynomialTools):
> solve( 2z^4-z^2-1+1, z );
```

This produces the four solutions

$$\pm\frac{1}{2}\sqrt{1+i\sqrt{7}},\ \frac{1}{2}\sqrt{1-i\sqrt{7}}.$$

Note that each algebraic number has a *degree* which is the degree of its minimal polynomial over $\mathbb{Q}$; for instance, the above algebraic numbers have degree 4. By elimination, the first elimination ideal $I_1 = I \cap \mathbb{C}[Y, Z]$ is generated by the polynomials $Y^2 - Z^2 - 1$ and $2Z^4 - Z^2 + 1$. The leading coefficient polynomial of $Y^2 - Z^2 - 1 \in \mathbb{C}[Z][Y]$ is the leading term of Y^2 which is a nonzero constant. Thus by extension, each partial solution in $\mathcal{V}(I_2)$ extends to a solution in $\mathcal{V}(I_1)$. There are eight such points. To find them, substitute a root of the generator $2Z^4 - Z^2 + 1$ for Z and solve the resulting equation for Y. For instance, the `Maple` command

```
> subs(Z=(1/2)*sqrt(1+I*sqrt(7)), G);
```

produces

$$\frac{3}{4} - \frac{1}{4}i\sqrt{7} + \frac{1}{8}(1+i\sqrt{7})^2, Y^2 - \frac{5}{4} - \frac{1}{4}i\sqrt{7}, X - \frac{1}{2}\sqrt{1+i\sqrt{7}} + \frac{1}{4}(1+i\sqrt{7})^{3/2}.$$

We can check that the first expression is a zero by using `Maple`

```
> evalf(%);
```

which yields as output

$$[0 + 0 \cdot i, -1.250000000 - 0.6614378278 \cdot i + Y^2, -0.9783183438 + 0.6760967252 \cdot i + X].$$

The second expression shows that

$$Y = \pm\sqrt{\frac{5}{4} + \frac{1}{4}i\sqrt{7}}.$$

Finally, the leading coefficient polynomial of $X + 2Z^3 - Z \in \mathbb{C}[Y,Z][X]$ is the coefficient of the term X which is a nonzero constant. By extension, each partial solution in $\mathcal{V}(I_1)$ can be extended to a point in $\mathcal{V}(I)$. For instance, for the above value of Z we obtain

$$X = \frac{1}{2}\sqrt{1+i\sqrt{7}} - \frac{1}{4}(1+i\sqrt{7})^{3/2}.$$

This gives rise to the following solutions of the system of equations,

$$(\frac{1}{2}\sqrt{1+i\sqrt{7}} - \frac{1}{4}(1+i\sqrt{7})^{3/2}, \pm\sqrt{\frac{5}{4} + \frac{1}{4}i\sqrt{7}}, \frac{1}{2}\sqrt{1+i\sqrt{7}}).$$

All other solutions can be derived in the same way. ♢

Example 2.33 (Singular). Consider the system of equations

$$XY = 1, \tag{2.7}$$
$$XZ = 1. \tag{2.8}$$

This gives the ideal $I = \langle XY - 1, XZ - 1\rangle$ in $\mathbb{C}[X,Y,Z]$. First, calculate a Groebner basis of I with respect to an elimination ordering.

```
> ring r  = 0, (x,y,z), lp;
> ideal i = xy-1, xz-1;
> ideal j = std(i); j;
j[1]=y-z
j[2]=xz-1
```

The associated Groebner basis is $G = \{Y - Z, XZ - 1\}$. The first elimination ideal $I_1 = I \cap \mathbb{C}[Y,Z]$ has the Groebner basis $G_1 = \{Y - Z\}$. The zeros of this generator are the pairs (a,a) with $a \in \mathbb{C}$; that is, $\mathcal{V}(I_1) = \{(a,a) \mid a \in \mathbb{C}\}$.

The leading coefficient polynomial of $XZ - 1$ equals Z. By extension, each partial solution (a,a) with $a \neq 0$ extends to a solution $(X,Y,Z) = (1/a, a, a)$ in $\mathcal{V}(I)$ and thus solves the system of equations. Note that the partial solution $(0,0)$ cannot be extended. ♢

The above examples is rather simple because the coordinates of the solutions can all be expressed in terms of roots of complex numbers. Unfortunately, general systems of polynomial equations are rarely this nice. For instance, it is known that there are no general formulae involving only the field operations in $\mathbb{K}$ and extraction of roots, forming so-called radicals, for solving single variable polynomial equations of degree 5 or higher. This is a famous result due to Evariste Galois (1811-1832). Thus if elimination leads to a one-variable equations of degree 5 or higher, we may not be able to give radical formulae for the roots.

Example 2.34 (Maple). Consider the system of equations

$$\begin{aligned} X^5 + Y^2 + Z^2 &= 2, \\ X^2 + 2Y^2 &= 3, \\ XZ &= 1. \end{aligned}$$

To solve these equations, we first compute a Groebner basis of the ideal $I = \langle X^5 + Y^2 + Z^2 - 2, X^2 + 2Y^2 - 3, XZ - 1\rangle$ with respect to the `lp` ordering.

```
> with(Groebner):
> F2 := [x^5+y^2+z^2-2, x^2+2*y^2-3, x*z-1]:
> G2 := gbasis(F2, plex(x,y,z));
```

This gives the output

$$[2Z^7 - Z^5 - Z^3 + 2, 4Y^2 - 2Z^5 + Z^3 + Z - 6, 2X + 2Z^6 - Z^4 - Z^2].$$

By elimination, the second elimination ideal $I_2 = I \cap \mathbb{C}[Z]$ is generated by the polynomial $2Z^7 - Z^5 - Z^3 + 2$. This generator is irreducible over $\mathbb{Q}$. In this situation, we need to decide what kind of answer is required.

If we want a purely algebraic description of the solutions, then `Maple` can represent solutions of systems like this by the `solve` command. Entering

```
> solve(convert(G2, set), {x,y,z});
```

gives the output

$$\begin{aligned} X = {} & \frac{1}{2} \cdot \mathrm{Root\_of}(2\_Z^7 - \_Z^5 - \_Z^3 + 2)^2 + \frac{1}{2} \cdot \mathrm{Root\_of}(2\_Z^7 - \_Z^5 - \_Z^3 + 2)^4 \\ & - \mathrm{Root\_of}(2\_Z^7 - \_Z^5 - \_Z^3 + 2)^6, \\ Y = {} & \frac{1}{2} \cdot \mathrm{Root\_of}(-6 + Y')^5, \\ Y' = {} & \mathrm{Root\_of}(2\_Z^7 - \_Z^5 - \_Z^3 + 2) + \mathrm{Root\_of}(2\_Z^7 - \_Z^5 - \_Z^3 + 2)^3 \\ & - 2 \cdot \mathrm{Root\_of}(2\_Z^7 - \_Z^5 - \_Z^3 + 2)^5 + \_Z^7, \\ Z = {} & \mathrm{Root\_of}(2\_Z^7 - \_Z^5 - \_Z^3 + 2). \end{aligned}$$

Here $\mathrm{Root\_of}(2\_Z^7 - \_Z^5 - \_Z^3 + 2)$ stands for any of the roots of the polynomial equation $2\_Z^7 - \_Z^5 - \_Z^3 + 2 = 0$ in the dummy variable $\_Z$.

On the other hand, in many practical situations where equations must be solved, knowing a numerical approximation to a real or complex solution is often more useful and perfectly acceptable provided

that the results are sufficiently accurate. The command `fsolve` finds numerical approximations to all real or complex roots of a polynomial by a combination of root location and numerical techniques. For instance,

```
> fsolve(z^7-z^5-z^3+2);
```

computes approximate values for the *real* roots of the polynomial. The output is

$$-1.160417997.$$

Using this approximate value $Z = -1.160417997$ as partial solution in $\mathcal{V}(I_2)$, we can substitute this number into the Groebner basis using

```
> L := subs (z = -1.160417997, G2);
```

and obtain

$$[-7 \cdot 10^{-9}, -4.514744785 + 4Y^2, 1.723516882 + 2X].$$

It shows that the value of the first polynomial is not exactly 0. Nevertheless, we can extend this approximate partial solution as follows

```
> y := solve(L[2]);
> x := solve(L[3]);
```

In this way, we obtain two approximate solutions of the system,

$$(x, y, z) = (-0.8617584410, \pm 1.062396440, -1.160417997).$$

Checking one of these by substituting into the Groebner basis using

```
> subs([x=-0.8617584410, y=1.062396440, z=-1.160417997], G2);
```

we find that

$$[-7 \cdot 10^{-9}, -1 \cdot 10^{-9}, 0].$$

Thus we have a reasonably good approximate solution in the sense that the values are very close to 0. The remaining solutions can be derived in the same way. ♢

2.6 Geometry of Elimination

In this section, it will be shown that elimination can be interpreted as the projection of an affine variety onto a lower-dimensional subspace. For this, take integers k, n with $0 \leq k \leq n$ and consider the *projection mapping*

$$\pi_k : \mathbb{K}^n \to \mathbb{K}^{n-k} : (a_1, \ldots, a_n) \mapsto (a_{k+1}, \ldots, a_n).$$

Lemma 2.35. *Let I be an ideal of $\mathbb{K}[X_1, \ldots, X_n]$ and let $V = \mathcal{V}(I)$ be the corresponding affine variety. For each $0 \leq k \leq n$, the k-th elimination ideal I_k of I satisfies*

$$\pi_k(V) \subseteq \mathcal{V}(I_k).$$

Proof. Let $f \in I_k$. Since $f \in I$, it follows that for any point $(a_1, \ldots, a_n) \in V$, we have $f(a_1, \ldots, a_n) = 0$. But f only involves the coordinates $X_{k+1}, \ldots, X_n$ and so $f(\pi_k(a_1, \ldots, a_n)) = f(a_{k+1}, \ldots, a_n) = 0$. Hence, f vanishes at all points of $\pi_k(V)$. □

It follows that the projection of the k-th elimination ideal I_k can be written as

$$\pi_k(V) = \{(a_{k+1}, \ldots, a_n) \in V(I_k) \mid \exists a_1, \ldots, a_k \in \mathbb{K} : (a_1, \ldots, a_n) \in V\}.$$

Thus $\pi_k(V)$ consists exactly of the partial solutions that extend to complete solutions. However, $\pi_k(V)$ is generally not an affine variety.

Example 2.36. Reconsider the system of equations $XY = 1$ and $XZ = 1$ in $\mathbb{C}[X, Y, Z]$ (Ex. 2.33). The first elimination ideal I_1 is generated by the polynomial $Y - Z$ and the associated affine variety $\mathcal{V}(I_1) = \{(a, a) \mid a \in \mathbb{C}\}$ is a line in the (y, z)-plane.

On the other hand, the projected set $\pi_1(V) = \{(a, a) \mid a \in \mathbb{C}, a \neq 0\}$ is not an affine variety. It misses the point $(0, 0)$, since there is no solution $(a, 0, 0) \in \mathcal{V}(I)$ for some $a \in \mathbb{C}$. ♢

The gap between the projected set $\pi_k(V)$ and the affine variety $\mathcal{V}(I_k)$ can be determined by using the Extension theorem.

Theorem 2.37. *Let $\mathbb{K}$ be an algebraically closed field, let I be an ideal of $\mathbb{K}[X_1, \ldots, X_n]$, and let $V = \mathcal{V}(I)$ be the corresponding affine variety. Let $G_1 = \{g_1, \ldots, g_s\}$ be a Groebner basis of the first elimination ideal I_1 of I with respect to an elimination ordering with $X_1 > \ldots > X_n$ and let h_i denote the leading coefficient polynomial of g_i, $1 \leq i \leq s$. Then we have*

$$\mathcal{V}(I_1) = \pi_1(V) \cup [\mathcal{V}(h_1, \ldots, h_s) \cap \mathcal{V}(I_1)].$$

Proof. By Lemma 2.35, the set on the right hand side lies in $\mathcal{V}(I_1)$. Conversely, let $(a_2, \ldots, a_n) \in \mathcal{V}(I_1)$. If $(a_2, \ldots, a_n) \notin \mathcal{V}(h_1, \ldots, h_s)$, then by the Extension theorem there exists $a_1 \in \mathbb{K}$ such that $(a_1, a_2, \ldots, a_n) \in \mathcal{V}(I)$ and thus $\pi_1(a_1, a_2, \ldots, a_n) = (a_2, \ldots, a_n) \in \pi_1(V)$. Otherwise, $(a_2, \ldots, a_n)$ lies in $\mathcal{V}(h_1, \ldots, h_s) \cap \mathcal{V}(I_1)$. □

The relationship between the projected set $\pi_k(V)$ and the affine variety $\mathcal{V}(I_k)$ can be explained as follows.

Theorem 2.38. (Closure) *Let $\mathbb{K}$ be an algebraically closed field, let $I = \langle f_1, \ldots, f_s \rangle$ be an ideal in $\mathbb{K}[X_1, \ldots, X_n]$, and let $V = \mathcal{V}(I)$ be the corresponding affine variety in $\mathbb{K}^n$. For each $0 \leq k \leq n$, the affine variety $\mathcal{V}(I_k)$ is the Zariski closure of $\pi_k(V)$.*

Proof. By Prop. 2.23, we have to show that $\mathcal{V}(I_k) = \mathcal{V}(\mathcal{I}(\pi_k(V)))$.

By Lemma 2.35, we have $\pi_k(V) \subseteq \mathcal{V}(I_k)$. But by Prop. 2.23, $\mathcal{V}(\mathcal{I}(\pi_k(V)))$ is the smallest affine variety containing $\pi_k(V)$ and so $\mathcal{V}(\mathcal{I}(\pi_k(V))) \subseteq \mathcal{V}(I_k)$.

Conversely, let $f \in \mathcal{I}(\pi_k(V))$; that is, $f(a_{k+1}, \ldots, a_n) = 0$ for all $(a_{k+1}, \ldots, a_n) \in \pi_k(V)$. Consider f as an element of $\mathbb{K}[X_1, \ldots, X_n]$. Then $f(a_1, \ldots, a_n) = 0$ for all $(a_1, \ldots, a_n) \in V$; that is, $f \in \mathcal{I}(\mathcal{V}(I))$. By the Strong Nullstellensatz, $f \in \sqrt{I}$. But f lies in $\mathbb{K}[X_{k+1}, \ldots, X_n]$ and so $f \in \sqrt{I}_k$. It follows that $\mathcal{I}(\pi_k(V)) \subseteq \sqrt{I_k}$. Therefore, by Prop. 2.16, $\mathcal{V}(I_k) = \mathcal{V}(\sqrt{I_k}) \subseteq \mathcal{V}(\mathcal{I}(\pi_k(V)))$. □

Example 2.39. Reconsider the system of equations $XY = 1$ and $XZ = 1$ in $\mathbb{C}[X, Y, Z]$ (Ex. 2.36). The first elimination ideal I_1 has the associated affine variety $V = \{(a, a) \mid a \in \mathbb{C}\}$ which is a line in the (y, z)-plane. On the other hand, the projected set $\pi_1(V) = \{(a, a) \mid a \in \mathbb{C}, a \neq 0\}$ is not an affine variety. By the Closure theorem, the affine variety V is the Zariski closure of $\pi_1(V)$. ♢

2.7 Implicit Representation

An affine variety is defined as the set of solutions of a system of polynomial equations. There is another way to represent an affine variety, namely, by a system of parametric equations such that its elements can be explicitly written down. This representation can be used for drawing an affine variety, but not every affine variety can be described in this way.

Let V be an affine variety in the affine space $\mathbb{K}^n$. An *implicit representation* of V describes the set V as a set of solutions a system of polynomial equations,

$$f_1 = \ldots = f_m = 0, \tag{2.9}$$

where $f_1, \ldots, f_m$ are polynomials in $\mathbb{K}[X_1, \ldots, X_n]$. On the other hand, a *parametric representation* of V describes the set V as the Zariski closure of the set

$$\{(f_1(t_1, \ldots, t_m), \ldots, f_n(t_1, \ldots, t_m)) \mid t_1, \ldots, t_m \in \mathbb{K}\}, \tag{2.10}$$

where $f_1, \ldots, f_n$ are polynomials in $\mathbb{K}[T_1, \ldots, T_m]$ or rational functions in $\mathbb{K}(T_1, \ldots, T_m)$. The implicit representation is useful to test whether or not a point lies in the variety, while the parametric representation is useful for plotting the variety.

Example 2.40. The affine variety V given by the solutions of the equation

$$X^2 - Y = 0$$

can be equivalently described by the polynomial parametrization

$$X = T, \quad Y = T^2.$$

♢

Take polynomials $f_1, \ldots, f_n \in \mathbb{K}[T_1, \ldots, T_m]$ and consider the following system of equations in $\mathbb{K}[T_1, \ldots, T_m, X_1, \ldots, X_n]$ given as

$$X_i = f_i(T_1, \ldots, T_m), \quad 1 \leq i \leq n. \tag{2.11}$$

The polynomials $f_1, \ldots, f_n$ give rise to the mapping $F : \mathbb{K}^m \to \mathbb{K}^n$ defined as

$$F : (t_1, \ldots, t_m) \mapsto (f_1(t_1, \ldots, t_m), \ldots, f_n(t_1, \ldots, t_m)). \tag{2.12}$$

The set $F(\mathbb{K}^m)$ is a subset of $\mathbb{K}^n$ that is parametrized by the equations (2.11). But $F(\mathbb{K}^m)$ may not be an affine variety and thus we search for the smallest affine variety that contains $F(\mathbb{K}^m)$; that is, the Zariski closure of $F(\mathbb{K}^m)$. For this, we relate implicitization to elimination. To this end, observe that the system of equations (2.11) defines the affine variety

$$V = \mathcal{V}(X_1 - f_1, \ldots, X_n - f_n) \subseteq \mathbb{K}^{m+n}. \tag{2.13}$$

The points of V can be written in the form

$$(t_1, \ldots, t_m, f_1(t_1, \ldots, t_m), \ldots, f_n(t_1, \ldots, t_m)), \quad t_1, \ldots, t_m \in \mathbb{K}. \tag{2.14}$$

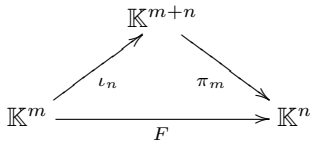
Define the embedding $\iota_n : \mathbb{K}^m \to \mathbb{K}^{m+n}$ by

$$\iota_n : (t_1, \ldots, t_m) \mapsto (t_1, \ldots, t_m, f_1(t_1, \ldots, t_m), \ldots, f_n(t_1, \ldots, t_m)), \tag{2.15}$$

and the projection $\pi_m : \mathbb{K}^{m+n} \to \mathbb{K}^n$ by

$$\pi_m : (t_1, \ldots, t_m, x_1, \ldots, x_n) \mapsto (x_1, \ldots, x_n). \tag{2.16}$$

These maps give rise to the commutative diagram



That is, the map F can be written as the composition

$$F = \pi_m \circ \iota_n. \tag{2.17}$$

By definition, we have

$$\iota_n(\mathbb{K}^m) = V. \tag{2.18}$$

Thus we obtain

$$F(\mathbb{K}^m) = \pi_m(\iota_n(\mathbb{K}^m)) = \pi_m(V). \tag{2.19}$$

Therefore, the image of the parametrization equals the projection of the affine variety. Thus the Closure theorem immediately implies the following result.

Theorem 2.41. (Polynomial Implicitization) *Let $\mathbb{K}$ be an algebraically closed field, let $F : \mathbb{K}^m \to \mathbb{K}^n$ be a map determined by the polynomial parametrization (2.11), and let $I = \langle X_1 - f_1, \ldots, X_n - f_n \rangle$ be an ideal in $\mathbb{K}[T_1, \ldots, T_m, X_1, \ldots, X_n]$. Then for the m-th elimination ideal $I_m = I \cap \mathbb{K}[X_1, \ldots, X_n]$, the affine variety $\mathcal{V}(I_m)$ is the Zariski closure of $F(\mathbb{K}^m)$.*

The following algorithm solves the *polynomial implicitication problem*: Given a system of equations

$$X_i = f_i, \quad 1 \leq i \leq n,$$

where $f_1, \ldots, f_n$ are polynomials in $\mathbb{K}[T_1, \ldots, T_m]$. Consider the ideal $I = \langle X_1 - f_1, \ldots, X_n - f_n \rangle$ in $\mathbb{K}[T_1, \ldots, T_m, X_1, \ldots, X_n]$. Compute a Groebner basis of I with respect to an elimination ordering with $T_1 > \ldots > T_m > X_1 > \ldots > X_n$. Then the elements of the Groebner basis which are not involving $T_1, \ldots, T_m$ form a Groebner basis of the m-th elimination ideal I_m. By the Implicitization theorem, this basis defines the affine variety in $\mathbb{K}^n$ containing the parametrization.

Example 2.42 (Singular). Consider the parametric surface

$$S = \{(uv, uv^2, u^2) \mid u, v \in \mathbb{C}\}$$

that is given by the system of equations

$$\begin{aligned} X &= UV, \\ Y &= UV^2, \\ Z &= U^2. \end{aligned}$$

The surface S can be illustrated by the `Maple` code (Fig. 2.5)

```
> with(plots):
> plot3d([u*v, u*v^2, u^2], u=-5..5, v=-5..5, grid=[20,20]);
```

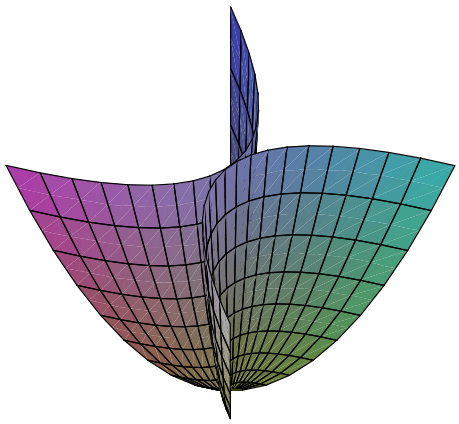


Fig. 2.5. Parametric surface.

Take the ideal $I = \langle X - UV, Y - UV^2, Z - U^2 \rangle$ in $\mathbb{C}[U, V, X, Y, Z]$ and compute a Groebner basis of I with respect to the `lp` ordering with $U > V > X > Y > Z$.

```
> ring r = 0, (u,v,x,y,z), lp:
> ideal i = x-uv, y-uv2, z-u2;
> std(i);
_[1]=x4-y2z
_[2]=vyz-x3
_[3]=vx-y
_[4]=v2z-x2
_[5]=uy-v2z
_[6]=ux-vz
_[7]=uv-x
_[8]=u2-z
```

Thus the second elimination ideal I_2 has the Groebner basis $\{X^4 - Y^2Z\}$. By the Implicitization theorem, the affine variety $V = \mathcal{V}(X^4 - Y^2Z)$ is the Zariski closure of the parametric surface S. $\diamondsuit$

Second, consider a *parametric representation* of an affine variety V in $\mathbb{K}^n$ as the Zariski closure of the set

$$\left\{ \left(\frac{f_1(t_1,\ldots,t_m)}{g_1(t_1,\ldots,t_m)}, \ldots, \frac{f_n(t_1,\ldots,t_m)}{g_1(t_1,\ldots,t_m)} \right) \mid t_1,\ldots,t_m \in \mathbb{K} \right\}, \tag{2.20}$$

where $f_1,\ldots,f_n$ and $g_1,\ldots,g_n$ are polynomials in $\mathbb{K}[T_1,\ldots,T_m]$.

Example 2.43 (Maple). Consider a curve in $\mathbb{C}^2$ parametrized by rational functions

$$X = \frac{a(T)}{c(T)}, \quad Y = \frac{b(T)}{c(T)},$$

where $a, b, c \in \mathbb{K}[T]$ are polynomials such that $c \neq 0$ and $\gcd(a,b,c) = 1$. In particular, consider the curve

$$C = \left\{ \left(\frac{2T^2+4T+5}{T^2+2T+3}, \frac{3T^2+T+4}{T^2+2T+3} \right) \mid t \in \mathbb{C} \right\}.$$

This curve can be drawn by using the `Maple` code (Fig. 2.6)

```
> with(plots):
> plot( [(2T^2+4T+5)/(T^2+2T+3), (3T^2+T+4)/(T^2+2T+3)], t=-10..10);
```

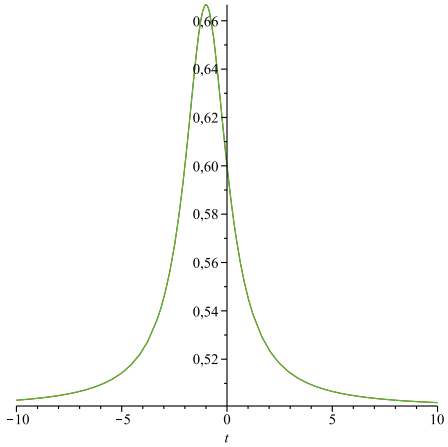


Fig. 2.6. Parametric curve.

Parametrizations of this form play an important role in computer-aided geometric design. A question of particular interest is the implicitization problem, which asks how the equation $f(X,Y) = 0$ of the underlying curve is obtained from the parametrization. ♢

Take polynomials $f_1,\ldots,f_n$ and $g_1,\ldots,g_n$ in $\mathbb{K}[T_1,\ldots,T_m]$ and consider the system of equations

$$X_i = \frac{f_i(T_1, \ldots, T_m)}{g_i(T_1, \ldots, T_m)}, \quad 1 \leq i \leq n. \tag{2.21}$$

These polynomials give rise to the mapping $F : \mathbb{K}^m \to \mathbb{K}^n$ defined as

$$F : (t_1, \ldots, t_m) \mapsto \left(\frac{f_1(t_1, \ldots, t_m)}{g_1(t_1, \ldots, t_m)}, \ldots, \frac{f_n(t_1, \ldots, t_m)}{g_1(t_1, \ldots, t_m)} \right). \tag{2.22}$$

The mapping F may not be defined at all points in $\mathbb{K}^m$ because of the denominators. Therefore, we put $W = \mathcal{V}(g_1, \ldots, g_n)$ and obtain the mapping $F : \mathbb{K}^m \setminus W \to \mathbb{K}^n$. In order to control the denominators, we put $g = g_1 \cdots g_n$, introduce an additional variable Y, and consider the ideal

$$I = \langle g_1 X_1 - f_1, \ldots, g_n X_n - f_n, Yg - 1 \rangle. \tag{2.23}$$

The equation $1 - Yg = 0$ means that the denominators $g_1, \ldots, g_n$ never vanish on $\mathcal{V}(I)$.

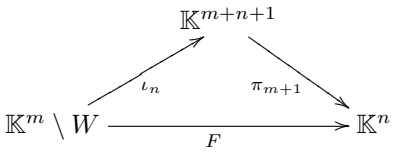
Define the embedding $\iota_n : \mathbb{K}^m \setminus W \to \mathbb{K}^{m+n+1}$ by

$$\iota_n : (t_1, \ldots, t_m) \mapsto \left(\frac{1}{g(t_1, \ldots, t_m)}, t_1, \ldots, t_m, \frac{f_1(t_1, \ldots, t_m)}{g_1(t_1, \ldots, t_m)}, \ldots, \frac{f_n(t_1, \ldots, t_m)}{g_1(t_1, \ldots, t_m)} \right), \tag{2.24}$$

and the projection $\pi_{m+1} : \mathbb{K}^{m+n+1} \to \mathbb{K}^n$ by

$$\pi_{m+1} : (y, t_1, \ldots, t_m, x_1, \ldots, x_n) \mapsto (x_1, \ldots, x_n). \tag{2.25}$$

These maps give rise to the commutative diagram



That is, the mapping F can be written as the composition $F = \pi_{m+1} \circ \iota_n$. By definition, we have $\iota_n(K^m \setminus W) = \mathcal{V}(I)$ and thus

$$F(\mathbb{K}^m \setminus W) = \pi_{m+1}(\iota_n(\mathbb{K}^m \setminus W)) = \pi_{m+1}(\mathcal{V}(I)). \tag{2.26}$$

Therefore, the image of the parametrization equals the projection of the affine variety. The Closure theorem yields the following result.

Theorem 2.44. (Rational Implicitization) *Let $\mathbb{K}$ be an algebraically closed field. Let $F : \mathbb{K}^m \setminus W \to \mathbb{K}^n$ be the mapping determined by the rational parametrization (2.21) and let $I = \langle g_1 X_1 - f_1, \ldots, g_1 X_n - f_n, 1 - Yg \rangle$ be an ideal in $\mathbb{K}[Y, T_1, \ldots, T_m, X_1, \ldots, X_n]$, where $g = g_1 \cdots g_n$. Then for the $(m+1)$-th elimination ideal $I_{m+1} = I \cap K[X_1, \ldots, X_n]$, the affine variety $V(I_{m+1})$ is the Zariski closure $F(\mathbb{K}^m \setminus W)$.*

The following algorithm solves the *rational implicitication problem*: Given a system of equations

$$X_i = \frac{f_i(T_1, \ldots, T_m)}{g_i(T_1, \ldots, T_m)}, \quad 1 \leq i \leq n,$$

where $f_1, \ldots, f_n$ and $g_1, \ldots, g_n$ are polynomials in $\mathbb{K}[T_1, \ldots, T_m]$. Take a new variable Y and consider the ideal $I = \langle g_1 X_1 - f_1, \ldots, g_n X_n - f_n, 1 - Y g \rangle$ of $\mathbb{K}[Y, T_1, \ldots, T_m, X_1, \ldots, X_n]$. Compute a Groebner basis with respect to the an elimination ordering with $Y > T_1 > \ldots > T_m > X_1 > \ldots > X_n$ By the Elimination theorem, the elements of the Groebner basis not involving $Y, T_1, \ldots, T_m$ form a Groebner basis of the $(m+1)$-th elimination ideal I_{m+1}. By the Implicitization theorem, this Groebner basis defines the affine variety in $\mathbb{K}^n$ containing the parametrization.

Example 2.45. Consider a curve in $\mathbb{K}^2$ parametrized by the rational functions

$$X = \frac{2T^2 + 4T + 5}{T^2 + 2T + 3}, \quad Y = \frac{3T^2 + T + 4}{T^2 + 2T + 3}.$$

This parametrization gives the following ideal in $\mathbb{C}[Z, T, X, Y]$

$$I = \langle ((T^2 + 2T + 3)X - (2T^2 + 4T + 5), (T^2 + 2T + 3)Y - (T^2 + 2T + 3), (T^2 + 2T + 3)^2 Z - 1 \rangle.$$

A Groebner basis of I with respect to the `lp` ordering is given by

$$XY - Y, YT^2 + 2YT + 2Y, 2XT^2 + 4XT + 5X - T^2 - 2T - 3, -4X^2 + 4X + Z - 1.$$

Thus the second elimination ideal is $I_2 = \langle XY - Y \rangle$ and underlying curve is given by $f(X, Y) = XY - Y$. $\diamondsuit$

3

Combinatorial Geometry

In this chapter we examine some interesting recently discovered connections between polynomials and the geometry of convex polytopes. This will naturally lead to the polytope algebra that can be viewed as a multi-dimensional generalization of the tropical algebra.

3.1 Tropical Algebra

A semiring is an algebraic structure similar to a ring, but without the requirement that each element must have an additive inverse. A prominent example of a semiring is the so-called tropical algebra.

A *semiring* is a non-empty set R together with two binary operations, addition $+$ and multiplication $\cdot$, such that $(R, +)$ is a commutative monoid with identity element 0, $(R, \cdot)$ is a monoid with identity element 1, multiplication distributes over addition, i.e., for all $a, b, c \in R$,

$$a \cdot (b + c) = (a \cdot b) + (a \cdot c) \quad \text{and} \quad (a + b) \cdot c = (a \cdot c) + (b \cdot c),$$

and multiplication with 0 annihilates R, i.e., for all $a \in R$, $a \cdot 0 = 0 = 0 \cdot a$. A *commutative* semiring is a semiring whose multiplication is commutative. An *idempotent* semiring is a semiring whose addition is idempotent, i.e., for all $a \in R$, $a + a = a$.

Example 3.1. Each ring is also a semiring. The set of natural numbers $\mathbb{N}_0$ forms a commutative semiring with the ordinary addition and multiplication. Likewise, the non-negative rational numbers and the non-negative real numbers form commutative semirings. ♢

Example 3.2. The set $\mathbb{R} \cup \{\infty\}$ together with the operations

$$x \oplus y = \min\{x, y\} \quad \text{and} \quad x \odot y = x + y, \quad x, y \in \mathbb{R} \cup \{\infty\}$$

forms a commutative, idempotent semiring with additive identity ∞ and multiplicative identity 0. Note that additive and multiplicative inverses may not exist. For instance, the equations $3 \oplus x = 10$ and $\infty \odot x = 1$ have no solutions $x \in \mathbb{R} \cup \{\infty\}$. ♢

This semiring is also known min-plus algebra or *tropical algebra*. The attribute "tropical" was coined by French scholars (1998) in honor of the Brazilian mathematician Imre Simon who studied the tropical semiring in the early 1960s.

Proposition 3.3. *The mapping* $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R} \cup \{\infty\} : x \mapsto -\log x$ *is an antitone, bijective mapping such that* $\phi(0) = \infty$, $\phi(1) = 0$, *and*

$$\phi(x \cdot y) = \phi(x) \odot \phi(y), \quad x, y \in \mathbb{R}_{\geq 0}. \tag{3.1}$$

Thus the mapping ϕ is a monoid isomorphism from $(\mathbb{R}_{\geq 0}, \cdot)$ onto $(\mathbb{R} \cup \{\infty\}, \odot)$. This mapping is called the *tropicalization* of the ordinary semiring $(\mathbb{R}_{\geq 0}, +, \cdot)$ (Fig. 3.1).

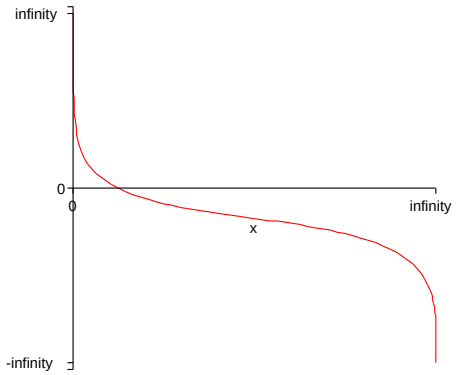


Fig. 3.1. The function $x \mapsto -\log x$.

3.2 Shortest Paths Problem

We illustrate an important problem in graph theory that makes use of the tropical algebra. For this, let $G = (V, E)$ be a digraph with vertex set $V = \{1, \ldots, n\}$. Each edge (i, j) in G has an associated length d_{ij} given by a positive real number. We put $d_{ii} = 0$, $1 \leq i \leq n$, and $d_{ij} = +\infty$ if (i, j), $i \neq j$, is not an edge in G. We represent the digraph G by the $n \times n$ adjacency matrix $D_G = (d_{ij})$. Consider the $(n-1)$th power of the matrix D_G in the tropical algebra $(\mathbb{R} \cup \{\infty\}, \oplus, \odot)$,

$$D_G^{\odot n-1} = D_G \odot D_G \odot \cdots \odot D_G \quad (n-1 \text{ times}). \tag{3.2}$$

Proposition 3.4. *Let* G *be a digraph on* n *vertices with* $n \times n$ *adjacency matrix* D_G. *The entry of the matrix* $D_G^{\odot n-1}$ *in row* i *and column* j *equals the length of the shortest path from vertex* i *to vertex* j.

Proof. Let $d_{ij}^{(r)}$ denote the minimum length of any path from vertex i to vertex j, which uses at most r edges in G. Clearly, we have $d_{ij}^{(1)} = d_{ij}$. The shortest path from vertex i to vertex j visits each vertex at most once, because the weights are assumed to be nonnegative. Thus the shortest path uses at most $n-1$ edges and hence the length of a shortest path from i to j equals $d_{ij}^{(n-1)}$.

Observe that a shortest path from vertex i to vertex j, which uses at most $r \geq 2$ edges, consists of a shortest path from vertex i to some vertex k, which uses at most $r-1$ edges, and the edge (k, j). That is, the shortest paths from vertex i to vertex j satisfy the equation

$$d_{ij}^{(r)} = \min\{d_{ik}^{(r-1)} + d_{kj} \mid 1 \leq k \leq n\}, \quad 2 \leq r \leq n-1. \tag{3.3}$$

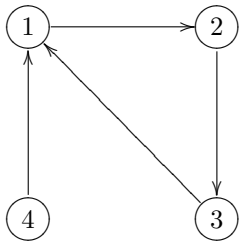
The tropicalization of this equation yields

$$d_{ij}^{(r)} = \bigoplus_{k=1}^{n} d_{ik}^{(r-1)} \odot d_{kj}, \quad 2 \leq r \leq n-1. \tag{3.4}$$

The right-hand side is the tropical product of the ith row of $D_G^{\odot r-1}$ and the jth column of D_G. Thus the left-hand side is the (i, j)th entry of the matrix $D_G^{\odot r}$. Hence, the assertion follows. □

The iterative evaluation of Eq. (3.3) is known as *Floyd-Warshall algorithm* for finding the shortest paths between each pair of vertices in a digraph.

Example 3.5. Consider the following digraph G,



The corresponding adjacency matrix is

$$D_G = \begin{pmatrix} 0 & 1 & \infty & \infty \\ \infty & 0 & 1 & \infty \\ 1 & \infty & 0 & \infty \\ 1 & \infty & \infty & 0 \end{pmatrix}$$

and the tropical matrix products are

$$D_G^{\odot 2} = \begin{pmatrix} 0 & 1 & 2 & \infty \\ 2 & 0 & 1 & \infty \\ 1 & 2 & 0 & \infty \\ 1 & 2 & \infty & 0 \end{pmatrix} \quad \text{and} \quad D_G^{\odot 3} = \begin{pmatrix} 0 & 1 & 2 & \infty \\ 2 & 0 & 1 & \infty \\ 1 & 2 & 0 & \infty \\ 1 & 2 & 3 & 0 \end{pmatrix}.$$

♢

3.3 Geometric Zoo

We consider the Euclidean n-space $\mathbb{R}^n$ equipped with the ordinary scalar product

$$\langle u, v\rangle = u_1v_1 + \cdots + u_nv_n, \quad u, v \in \mathbb{R}^n. \tag{3.5}$$

The Euclidean distance between two points u and v in $\mathbb{R}^n$ is defined as

$$\|u - v\| = \sqrt{\langle u - v, u - v\rangle}. \tag{3.6}$$

A set C in $\mathbb{R}^n$ is called *convex* if it contains the line segment connecting any two points in C. The *line segment* between two points u and v in $\mathbb{R}^n$ is given as

$$[u, v] = \{\lambda u + (1 - \lambda)v \mid 0 \leq \lambda \leq 1\}. \tag{3.7}$$

Simple examples of convex sets are the singleton sets $\{v\}$, where $v \in \mathbb{R}^n$, and the Euclidean space $\mathbb{R}^n$. There is a simple way to construct new convex sets from given ones.

Proposition 3.6. *The intersection of an arbitrary collection of convex sets is convex.*

Proof. If a line segment belongs to every set in the collection, it also belongs to the intersection. □

If a set is not itself convex, its *convex hull* is the smallest convex set containing it (Fig. 3.2). The convex hull of a set S in $\mathbb{R}^n$ is denoted by $\mathrm{conv}(S)$.

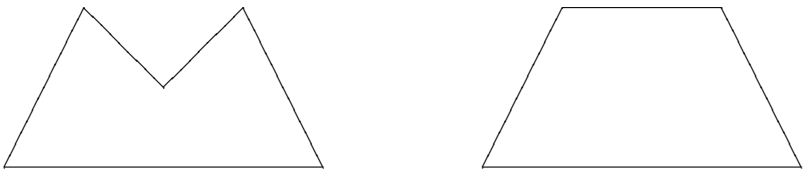


Fig. 3.2. A set and its convex hull, a convex polygon.

Proposition 3.7. *If S is a subset of $\mathbb{R}^n$, then its convex hull is*

$$\mathrm{conv}(S) = \{\lambda_1 s_1 + \ldots + \lambda_m s_m \mid m \geq 0,\ s_i \in S,\ \lambda_i \geq 0,\ \sum_i \lambda_i = 1\}. \tag{3.8}$$

Moreover,

- *for each subset of R^n,* $\mathrm{conv}(\mathrm{conv}(S)) = \mathrm{conv}(S)$.
- *If S_1 and S_2 are subsets of $\mathbb{R}^n$, then* $\mathrm{conv}(\mathrm{conv}(S_1) \cup \mathrm{conv}(S_2)) = \mathrm{conv}(S_1 \cup S_2)$, *and if $S_1 \subseteq S_2$, then* $\mathrm{conv}(S_1) \subseteq \mathrm{conv}(S_2)$.

Proof. Let $u, v \in S$. By definition, for each λ, $0 \leq \lambda \leq 1$, the point $\lambda u + (1 - \lambda)v$ belongs to $\mathrm{conv}(S)$ and so $\mathrm{conv}(S)$ is a convex set. Moreover, for each $s \in S$, $s = 1 \cdot s \in \mathrm{conv}(S)$ and thus $\mathrm{conv}(S)$ contains the set S.

Let C be a convex set in $\mathbb{R}^n$ containing S. We show that $\mathrm{conv}(S)$ lies in C by using induction on the size (i.e., number of terms) of the linear combinations. Each element of S lies in C. Let $s_1, \ldots, s_{m+1}$ be elements of S. Consider the convex combination

$$s = \lambda_1 s_1 + \ldots + \lambda_m s_m + \lambda_{m+1} s_{m+1},$$

where $\lambda_i \geq 0$, $1 \leq i \leq m+1$, and $\sum_i \lambda_i = 1$. If $\lambda_{m+1} = 1$, then $s = s_{m+1} \in C$, and if $\lambda_{m+1} = 0$, then by induction we have $s \in C$. Otherwise, we have

$$s = (1 - \lambda_{m+1}) \sum_{i=1}^{m} \frac{\lambda_i}{1 - \lambda_{m+1}} s_i + \lambda_{m+1} s_{m+1}.$$

and

$$\frac{\lambda_i}{1 - \lambda_{m+1}} \geq 0, \quad 1 \leq i \leq m, \quad \text{and} \quad \sum_{i=1}^{m} \frac{\lambda_i}{1 - \lambda_{m+1}} = 1.$$

Thus, by induction, the convex combination

$$s' = \sum_{i=1}^{m} \frac{\lambda_i}{1 - \lambda_{m+1}} s_i$$

belongs to C. Since C is convex and contains S, it follows that the element $s = (1-\lambda_{m+1})s' + \lambda_{m+1} s_{m+1}$ belongs to C, as required. The remaining assertions are left to the reader. □

A linear combinations of the form $\lambda_1 s_1 + \ldots + \lambda_m s_m$, where $s_i \in S$, $\lambda_i \geq 0$, and $\sum_i \lambda_i = 1$, is called a *convex combination*.

A *polytope* is the convex hull of a finite set in $\mathbb{R}^n$. If the set is $S = \{s_1, \ldots, s_m\}$ in $\mathbb{R}^n$, then by Prop. 3.7, the corresponding polytope can be expressed as

$$\mathrm{conv}(S) = \{\lambda_1 s_1 + \ldots + \lambda_m s_m \mid \lambda_i \geq 0, \sum_i \lambda_i = 1\}. \tag{3.9}$$

In lower dimensions, polytopes are familiar geometric figures: A polytope in $\mathbb{R}$ is a *line segment*, a polytope in $\mathbb{R}^2$ is a line segment or a *convex polygon* (Fig. 3.2), and a polytope in $\mathbb{R}^3$ is a line segment, a convex polygon lying in a plane, or a *convex polyhedron*. In particular, a *lattice polytope* is a polytope given by the convex hull of a set of integral points.

Example 3.8. The mathematical software `polymake` was designed to work with polytopes. Each polytope in `polymake` is treated as an object and is given by a file storing the data. The program `polymake` allows to construct polytopes from scratch or by applying constructions to existing polytopes.

Consider the lattice polytope given by the convex hull of the points (0,8), (0,7), (0,6), (0,5), (1,6), (1,5), (1,4), (1,3), (2,4), (2,3), and (3,2) (Fig. 5.7). In `polymake`, this polytope can be specified in a text file, say `dude`, containing the following information

```
POINTS
1 0 8
1 0 7
```

```
1 0 6
1 0 5
1 1 6
1 1 5
1 1 4
1 1 3
1 2 4
1 2 3
1 3 2
```

The points are always represented in homogeneous coordinates, where the first coordinate is used for homogenization. ♢

In particular, a *n-dimensional simplex* or *n-simplex* is the convex hull of $n+1$ points $m_1, \ldots, m_{n+1}$ in $\mathbb{R}^n$ such that the vectors $m_2 - m_1, \ldots, m_{n+1} - m_1$ form a basis of $\mathbb{R}^n$. A n-simplex can be constructed from a $(n-1)$-simplex in $\mathbb{R}^{n-1}$ by adding one point in the n-th dimension and connecting the point with all points of the $(n-1)$-simplex. In this way, one obtains inductively simplices that are singleton points, line segments, triangles, tetrahedrons (Fig. 3.3), and so on.

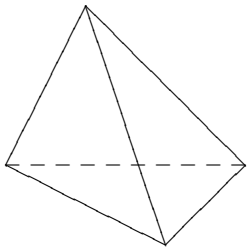


Fig. 3.3. A tetrahedron.

Each polytope has a well-defined dimension. To see this, we need to develop the theory of affine subspaces. An *affine subspace* of $\mathbb{R}^n$ is a subset A of $\mathbb{R}^n$ with the property that if $m \geq 0$ and $s_1, \ldots, s_m \in A$ then $\lambda_1 s_1 + \ldots + \lambda_m s_m \in A$ with $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$, whenever $\sum_i \lambda_i = 1$. Linear combinations of the form $\lambda_1 s_1 + \ldots + \lambda_m s_m$, where $s_i \in A$ and $\sum_i \lambda_i = 1$, are called *affine combinations*.

Given a subset S of $\mathbb{R}^n$ and a vector $v \in \mathbb{R}^n$, the *translate of S by v* is the set

$$v + S = \{v + s \mid s \in S\}. \tag{3.10}$$

Proposition 3.9. *Each affine subspace of $\mathbb{R}^n$ is a translate of a unique linear subspace of $\mathbb{R}^n$.*

Proof. Let A be an affine subspace of $\mathbb{R}^n$ and $v \in A$. Consider the translate

$$-v + A = \{\lambda_1 a_1 + \ldots + \lambda_m a_m \mid m \geq 0, a_i \in A, \sum_i \lambda_i = 0\}.$$

It is easy to check that the translate $-v + A$ is a linear subspace of $\mathbb{R}^n$. Since $A = v + (-v + A)$, it follows that A is a translate of the linear subspace $-v + A$. Moreover, if $v, w \in A$, then the above

representation of a translate shows that the linear subspaces $-v + A$ and $-w + A$ are equal. It follows that the affine subspace A is a translate of a unique linear subspace of $\mathbb{R}^n$. □

The *dimension* of an affine subspace in $\mathbb{R}^n$ is defined as the dimension of the linear subspace of $\mathbb{R}^n$ corresponding to it as in Prop. 3.9.

Proposition 3.10. *The translate of a polytope is a polytope.*

Proof. Let P be a polytope in $\mathbb{R}^n$ and let $v \in \mathbb{R}^n$. By definition, there is a subset $S = \{s_1, \ldots, s_m\}$ of $\mathbb{R}^n$ such that $P = \text{conv}(S)$. Claim that $v + \text{conv}(S) = \text{conv}(v + S)$. Indeed, let $w \in P$. Write $w = \sum_i \lambda_i s_i$, where $\lambda_i \geq 0$, $1 \leq i \leq m$, and $\sum_i \lambda_i = 1$. Then

$$v + \sum_i \lambda_i s_i = \sum_i \lambda_i \left(v + s_i\right).$$

The left-hand side is a point in $v + \text{conv}(S)$ and the right-hand side is a point in $\text{conv}(v + S)$. This proves the claim. Thus the translate $v + P$ is the convex hull of the set $v + S$ and hence $v + P$ is a polytope. □

If a set is not itself an affine subspace of $\mathbb{R}^n$, its *affine hull* is the smallest affine subspace containing it. The affine hull of a set S in $\mathbb{R}^n$ is denoted by $\text{aff}(S)$.

Proposition 3.11. *If S is a subset of $\mathbb{R}^n$, then*

$$\text{aff}(S) = \{\lambda_1 s_1 + \ldots + \lambda_m s_m \mid m \geq 0,\, s_i \in S,\, \lambda_i \in \mathbb{R},\, \sum_i \lambda_i = 1\}. \tag{3.11}$$

Moreover,

- *for each subset S of $\mathbb{R}^n$,* $\text{aff}(\text{aff}(S)) = \text{aff}(S)$.
- *If S_1 and S_2 are subsets of $\mathbb{R}^n$ such that $S_1 \subseteq S_2$, then* $\text{aff}(S_1) \subseteq \text{aff}(S_2)$.

Proof. By definition, the set $\text{aff}(S)$ is an affine subspace of $\mathbb{R}^n$. Moreover, for each $s \in S$, $s = 1 \cdot s \in \text{aff}(S)$ and thus $\text{aff}(S)$ contains the set S.

Let A be an affine subspace of $\mathbb{R}^n$ containing S. We show that $\text{aff}(S)$ lies in A by using induction on the size (i.e, number of terms) of the linear combinations. Each element of S lies in A. Let $s_1, \ldots, s_{m+1}$ be elements of S. Consider the affine combination

$$s = \lambda_1 s_1 + \ldots + \lambda_m s_m + \lambda_{m+1} s_{m+1},$$

where $\sum_i \lambda_i = 1$. If $\lambda_{m+1} = 1$, then $s = s_{m+1} \in A$, and if $\lambda_{m+1} = 0$, then by induction we have $s \in A$. Otherwise, we have

$$s = (1 - \lambda_{m+1}) \sum_{i=1}^{m} \frac{\lambda_i}{1 - \lambda_{m+1}} s_i + \lambda_{m+1} s_{m+1}.$$

and

$$\sum_{i=1}^{m} \frac{\lambda_i}{1 - \lambda_{m+1}} = 1.$$

Thus, by induction, the affine combination

$$s' = \sum_{i=1}^{m} \frac{\lambda_i}{1-\lambda_{m+1}} s_i$$

belongs to A. Since A is an affine subspace and contains S, it follows that the element $s = (1-\lambda_{m+1})s' + \lambda_{m+1}s_{m+1}$ belongs to A, as required. The remaining assertions are left to the reader. □

For instance, the affine hull of a point is a point, the affine hull of a line segment is a line, the affine hull of a convex polygon is a plane, and the affine hull of a convex polyhedron is the Euclidean 3-space.

The *dimension* of a polytope in $\mathbb{R}^n$ is defined as the dimension of its affine hull. This is the smallest affine space containing the polytope. For instance, a point has dimension 0, a line segment has dimension 1, a convex polygon has dimension 2, and a convex polyhedron has dimension 3.

Proposition 3.12. *Each n-simplex in $\mathbb{R}^n$ has dimension n.*

Proof. Let C be a n-simplex in $\mathbb{R}^n$ given by the points $m_1, \ldots, m_{n+1}$. The affine hull of C contains the points $m_2 - m_1, \ldots, m_{n+1} - m_1$. These points form a basis of $\mathbb{R}^n$ and thus the affine hull has dimension at least n. But $\mathrm{aff}(C)$ is an affine subspace of $\mathbb{R}^n$ and so has dimension at most n. □

The unbounded counterparts of polytopes are cones. A *cone* in $\mathbb{R}^n$ is a subset C of $\mathbb{R}^n$ with the property that if $m \geq 0$ and $s_1, \ldots, s_m \in C$ then $\lambda_1 s_1 + \ldots + \lambda_m s_m$, whenever $\lambda_i \geq 0$, $1 \leq i \leq m$.

If a set is not itself a cone, its *positive hull* is the smallest cone containing it. The positive hull of a set S in $\mathbb{R}^n$ is denoted by $\mathrm{pos}(S)$.

Proposition 3.13. *If S is a subset of $\mathbb{R}^n$, then the positive hull of S is*

$$\mathrm{pos}(S) = \{\lambda_1 s_1 + \ldots + \lambda_m s_m \mid m \geq 0,\ s_i \in S,\ \lambda_i \geq 0\}. \tag{3.12}$$

Proof. Let $u, v \in S$. By definition, the set $\mathrm{pos}(S)$ is a cone in $\mathbb{R}^n$. Moreover, for each $s \in S$, $s = 1 \cdot s \in \mathrm{pos}(S)$ and thus $\mathrm{pos}(S)$ contains the set S.

Finally, let C be a cone in $\mathbb{R}^n$ containing S. We show that $\mathrm{pos}(S)$ lies in C by using induction on the size (i.e., number of terms) of the linear combinations. Each element of S lies in C. Let $s_1, \ldots, s_{m+1}$ be elements of S. Consider the linear combination

$$s = \lambda_1 s_1 + \ldots + \lambda_m s_m + \lambda_{m+1} s_{m+1},$$

where $\lambda_i \geq 0$, $1 \leq i \leq m+1$. If $\lambda_{m+1} = 0$, then by induction we have $s \in C$. Otherwise, we have

$$s = \lambda_{m+1} \sum_{i=1}^{m} \frac{\lambda_i}{\lambda_{m+1}} s_i + \lambda_{m+1} s_{m+1}$$

and

$$\frac{\lambda_i}{\lambda_{m+1}} \geq 0.$$

Thus, by induction, the linear combination

$$s' = \sum_{i=1}^{m} \frac{\lambda_i}{\lambda_{m+1}} s_i$$

belongs to C. Since C is a cone and contains S, it follows that the element $s = \lambda_{m+1}s' + \lambda_{m+1}s_{m+1}$ belongs to C, as required. □

For instance, quadrants in $\mathbb{R}^2$ are cones, octants in $\mathbb{R}^3$ are cones, and half-spaces are cones.

Proposition 3.14. *The positive hull of a set is convex.*

Proof. By definition, convex combinations are nonnegative linear combinations and so the positive hull of a set is convex. □

3.4 Geometry of Polytopes

The geometric structure of polytopes will be studied in more detail.

An *affine hyperplane* is an affine subspace of codimension 1 in the Euclidean space $\mathbb{R}^n$. In Cartesian coordinates, an affine hyperplane is given by a single linear equation (with not all w_i equal to 0)

$$v_1 w_1 + \ldots + v_n w_n = \alpha.$$

More specifically, an affine hyperplane in $\mathbb{R}^n$ is defined as

$$H_{w,\alpha} = \{v \in \mathbb{R}^n \mid \langle v, w \rangle = \alpha\} \tag{3.13}$$

where $w \neq 0$ is a vector in $\mathbb{R}^n$ and α is a real number. Note that two affine hyperplanes $H_{w,\alpha}$ and $H_{w,\beta}$ for different values α and β are parallel to each other. Moreover, the sets

$$H^+_{w,\alpha} = \{v \mid \langle v, w \rangle \geq \alpha\}$$

and

$$H^-_{w,\alpha} = \{v \mid \langle v, w \rangle \leq \alpha\}$$

are called the *half-spaces* bounded by the affine hyperplane.

For instance, a point is an affine hyperplane in Euclidean 1-space, a line is an affine hyperplane in Euclidean 2-space (Fig. 3.4), and a plane is an affine hyperplane in Euclidean 3-space.

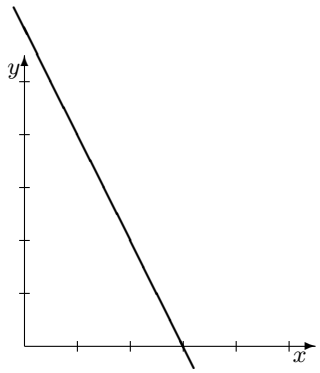


Fig. 3.4. Affine hyperplane $2x + y = 6$ in Euclidean 2-space.

Proposition 3.15. *An affine hyperplane in $\mathbb{R}^n$ is an affine subspace of $\mathbb{R}^n$ with dimension $n-1$.*

Proof. Let $H = \{v \mid \langle v, w\rangle = \alpha\}$ be an affine hyperplane in $\mathbb{R}^n$. Consider the linear mapping $\phi : \mathbb{R}^n \mapsto \mathbb{R} : v \mapsto \langle v, w\rangle$ given by the scalar product with fixed w. The kernel of this mapping is the linear subspace $U = \{u \mid \langle u, w\rangle = 0\}$ and the image is $\mathbb{R}$, since by definition $w \neq 0$. The dimension formula gives $\dim \mathbb{R}^n = \dim \ker \phi + \dim \operatorname{im} \phi = \dim U + \dim \mathbb{R}$ and so $\dim U = n-1$. But the affine hyperplane H is a translate of the linear subspace U given by $v + U$, where $v \in H$. It follows that the hyperplane H has also dimension $n-1$. □

Let P be a polytope in $\mathbb{R}^n$, let $w \neq 0$ be a vector in $\mathbb{R}^n$, and let α be a real number. Define the number

$$\rho_P(w) = \min\{\langle v, w\rangle \mid v \in P\}. \tag{3.14}$$

The number $\rho_P(w)$ always exists, since the linear map $P \to \mathbb{R} : v \mapsto \langle v, w\rangle$ given by the scalar product with fixed w is continuous and P is closed and bounded. Thus the minimum on P will be attained. The corresponding affine hyperplane

$$H_{P,w} = \{v \mid \langle v, w\rangle = \rho_P(w)\} \tag{3.15}$$

is called a *supporting hyperplane* of P, and we call w the *outward pointing normal* (Fig. 3.5).

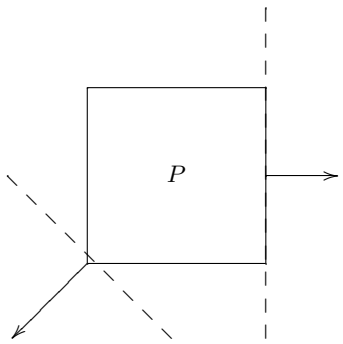


Fig. 3.5. A polytope P given as a square, two supporting hyperplanes, and associated outward pointing normals.

Proposition 3.16. *If $H_{P,w}$ is a supporting hyperplane of a polytope P in $\mathbb{R}^n$, then the intersection $P_w = P \cap H_{P,w}$ is a non-empty polytope and P is contained in the half-space $H^+_{P,w}$.*

Proof. The set P_w is non-empty, since the number $\rho_P(w)$ always exists.

Let $u, v \in P_w$. Then $u, v \in P$ and $\langle u, w\rangle = \langle v, w\rangle = \rho_P(w)$. Since P is a polytope, for each λ with $0 \leq \lambda \leq 1$, $\lambda u + (1-\lambda)v \in P$. Moreover, $\langle \lambda u + (1-\lambda)v, w\rangle = \lambda\langle u, w\rangle + (1-\lambda)\langle v, w\rangle = \rho_P(w)$ and so $\lambda u + (1-\lambda)v \in H_{P,w}$. Thus $\lambda u + (1-\lambda)v \in P_w$ and hence P_w is convex.

Assume that P is the convex hull of the points $a_1, \ldots, a_m$ in $\mathbb{R}^n$. Suppose that the points $a_1, \ldots, a_l$ are on the hyperplane $H_{P,w}$, while the points $a_{l+1}, \ldots, a_m$ are not. Equivalently, there are positive real numbers $\beta_{l+1}, \ldots, \beta_m$ such that

$$\langle a_i, w\rangle = \begin{cases} \rho_P(w), & 1 \le i \le l, \\ \rho_P(w) + \beta_i, & l+1 \le i \le m. \end{cases}$$

Let v be a point in P. By definition, the point v is a convex combination of the points $a_1, \ldots, a_m$,

$$v = \lambda_1 a_1 + \ldots + \lambda_m a_m,$$

where $\lambda_i \ge 0$, $1 \le i \le m$, and $\sum_i \lambda_i = 1$. It follows that

$$\langle v, w\rangle = \sum_{i=1}^{m} \lambda_i \langle a_i, w\rangle = \sum_{i=1}^{m} \lambda_i \rho_p(w) + \sum_{i=l+1}^{m} \lambda_i \beta_i = \rho_P(w) + \sum_{i=l+1}^{m} \lambda_i \beta_i.$$

Therefore, $v \in H_{P,w}$ if and only if $\lambda_{l+1}\beta_{l+1} + \ldots + \lambda_m \beta_m = 0$. By hypothesis, this is equivalent to $\lambda_{l+1} = \ldots = \lambda_m = 0$. Equivalently, the point v is a convex combination of the points $a_1, \ldots, a_l$. It follows that $P_w = P \cap H_{P,w} = \mathrm{conv}(\{a_1, \ldots, a_l\})$ is a polytope.

Finally, by definition, P_w belongs to the half-space $H^+_{P,w}$. □

We call the non-empty polytope P_w the *face of P determined by w*. That is, the face P_w is the set of all points in P at which the linear map $P \to \mathbb{R} : v \to \langle v, w\rangle$ given by the scalar product with fixed w attains its minimum.

Proposition 3.17. *Each polytope has only finitely many faces.*

Proof. Let P be a polytope in $\mathbb{R}^n$. By definition, there is a finite set $A = \{a_1, \ldots, a_m\}$ such that $P = \mathrm{conv}(A)$. The proof of Prop. 3.16 shows that for each supporting hyperplane H of P, the corresponding face $P \cap H$ is the convex hull of a subset of A. But there are only finitely many subsets of a finite set und so the result follows. □

Proposition 3.18. *Each face of a polytope P in $\mathbb{R}^n$ has dimension less than* $\dim P$.

Proof. We may assume that P is a polytope in $\mathbb{R}^n$ with dimension n. For each supporting hyperplane H of P, the affine space $P \cap H$ is contained in the affine space $\mathrm{aff}(H) = H$. But, by Prop. 3.15, the affine subspace H has dimension $n-1$ and so, by Prop. 3.11, the face $P \cap H$ has dimension at most $n-1$. □

Since each face is a polytope, it can be assigned a dimension. A *k-face* of a polytope P is a face of P with dimension k. A 0-face of P is called a *vertex* of P and a 1-face of P is called an *edge* of P. If P has dimension n, then a $(n-2)$-face of P is called a *ridge* of P and a $(n-1)$-face of P is called a *facet* of P. We write $f_k(P)$ for the number of k-faces of a polytope P in $\mathbb{R}^n$, $0 \le k \le n-1$. If P has dimension n, the vector $f(P) = (f_0(P), f_1(P), \ldots, f_{n-1}(P))$ is termed the *f-vector* of P. For instance, a tetrahedron (Fig. 3.3) has 4 facets, 6 edges, and 4 vertices and so its f-vector is $(4, 6, 4)$.

Lemma 3.19. *The vertices of a polytope P are precisely the points in P that cannot be written as convex combinations of other points in P.*

Proof. Let $v \in P$ be a vertex given by the supporting hyperplane $H_{P,w}$ in P. Write the point v as a convex combination $v = \sum_i \lambda_i a_i$ of points $a_i \in P$, where $\lambda_i \ge 0$ and $\sum_i \lambda_i = 1$. Then, by definition,

$$\rho_P(w) = \langle v, w\rangle = \sum_i \lambda_i \langle a_i, w\rangle \geq \sum_i \lambda_i \rho_P(w) = \rho_P(w).$$

Thus $\langle a_i, w\rangle = \rho_P(w)$ for each index i with $\lambda_i > 0$; that is, $a_i \in P \cap H_{P,w}$ for each index i with $\lambda_i > 0$. But v is a vertex of P and so $v = a_i$ for each index i with $\lambda_i > 0$.

Suppose $v \in P$ is not a vertex. Then v lies in a k-face $P_w = H \cap P$ for some $k \geq 1$ and supporting hyperplane $H = H_{P,w}$. The polytope P_w has a finite set $A = \{a_1, \ldots, a_m\}$ such that $P = \text{conv}(A)$. Write the point v as convex combination $v = \lambda_1 a_1 + \ldots + \lambda_m a_m$. If $v = a_i$ for each index i with $\lambda_i > 0$, then v will be a vertex. Thus by hypothesis, the point v can be written as nontrivial linear combination of points in P_w. □

Proposition 3.20. *Each polytope is the convex hull of its vertices.*

Proof. Let P be a polytope in $\mathbb{R}^n$. By definition, there is a finite set $A = \{a_1, \ldots, a_m\}$ such that $P = \text{conv}(A)$. The proof of Prop. 3.16 shows that each vertex of P is of the form $\{a\}$, where $a \in A$. Suppose $P_i = \{a_i\}$, $1 \leq i \leq l$, are the vertices of P. Then, by Prop. 3.7, $\text{conv}(P_1 \cup \ldots \cup P_l) = \text{conv}(\{a_1, \ldots, a_l\}) \subseteq \text{conv}(P) = P$.

Conversely, we may successively eliminate points from A that can be written as convex combinations of other points in A. For instance, assume that $a_m = \sum_{i=1}^{m-1} \lambda_i a_i$, where $\lambda_i \geq 0$, $1 \leq i \leq m-1$, and $\sum_i \lambda_i = 1$. Then for each point $v \in P$,

$$v = \sum_{i=1}^{m} \mu_i a_i = \sum_{i=1}^{m-1} (\mu_i + \lambda_i \mu_m) a_i,$$

where $\mu_i \geq 0$, $1 \leq i \leq m$, and $\sum_i \mu_i = 1$. But $\mu_i + \lambda_i \mu_m \geq 0$, $1 \leq i \leq m-1$, and $\sum_{i=1}^{m-1} \mu_i + \lambda_i \mu_m = 1$, and so the point v can be written as a convex combination of the points $a_1, \ldots, a_{m-1}$. It follows that $\text{conv}(\{a_1, \ldots, a_m\}) = \text{conv}(\{a_1, \ldots, a_{m-1}\})$.

Assume that all points are eliminated from A that can be written as convex combinations of other points in A. Claim that the remaining points in A are vertices of P. Suppose the point a_m can be written as convex combination of points $v_1, \ldots, v_k$ in P. That is, $a_m = \sum_{j=1}^{k} \mu_j v_j$, where $\mu_j \geq 0$, $1 \leq j \leq k$, and $\sum_j \mu_j = 1$. Write $v_j = \sum_{i=1}^{m} \lambda_{ji} a_i$, where $\lambda_{ji} \geq 0$, $1 \leq i \leq m$, and $\sum_i \lambda_{ji} = 1$. Then

$$a_m = \sum_{i=1}^{m} \sum_{j=1}^{k} \mu_j \lambda_{ji} a_i$$

and so

$$(1 - \sum_{j=1}^{m} \mu_j \lambda_{jm}) a_m = \sum_{i=1}^{m-1} \sum_{j=1}^{k} \mu_j \lambda_{ji} a_i.$$

But by hypothesis, the point a_m cannot be written as convex combination of the points $a_1, \ldots, a_{m-1}$ and so $\sum_{j=1}^{m} \mu_j \lambda_{jm} = 1$. Thus $\lambda_{jm} = 1$ for each index j for which $\mu_j > 0$; that is, $v_j = a_m$ for each index j for which $\mu_j > 0$. By Lemma 3.19, the points in A are the vertices of P. □

Example 3.21. Consider the set of points $A = \{(0,0), (1,1), (2,0), (0,3)\}$ in $\mathbb{R}^2$. The corresponding lattice polytope $\text{conv}(A)$ is the triangle with the vertices $(0,0)$, $(2,0)$, and $(0,3)$. The point $(1,1)$ is a convex combination of the triangle's vertices

$$(1,1) = \frac{1}{6}(0,0) + \frac{1}{2}(2,0) + \frac{1}{3}(0,3).$$

□

A *polyhedral set* in $\mathbb{R}^n$ is the intersection of a finite number of half-spaces in $\mathbb{R}^n$.

Proposition 3.22. *A bounded polyhedral set in $\mathbb{R}^n$ is a polytope in $\mathbb{R}^n$, and vice versa.*

Proof. Let P be a polyhedral set in $\mathbb{R}^n$ given by the intersection of $m \geq 1$ half-spaces. Then it is easy to check that $P = \{v \in \mathbb{R}^n \mid Av \geq b, v \geq 0\}$ for some matrix $A \in \mathbb{R}^{m \times n}$ and vector $b \in \mathbb{R}^m$. The set P is convex and since P is bounded, it is a polytope in $\mathbb{R}^n$.

Conversely, let P be an n-dimensional polytope in $\mathbb{R}^n$ with facets $F_1, \ldots, F_l$ having outward pointing normals $w_1, \ldots, w_l$, respectively. Then it is easy to check that

$$P = \{v \in \mathbb{R}^n \mid \langle v, w_j \rangle \geq \rho_P(w_j), 1 \leq j \leq l\}.$$

Hence, P is a bounded polyhedral set. □

Example 3.23. The square $P = \mathrm{conv}(\{(0,0),(0,1),(1,0),(1,1)\})$ in $\mathbb{R}^2$ has four facets that are given by the inequalities

$$\langle v, w_1 \rangle \geq 0, \quad \langle v, -w_2 \rangle \geq -1, \quad \langle v, w_3 \rangle \geq 0, \quad \langle v, -w_4 \rangle \geq -1,$$

where $e_1 = w_1 = w_2$ and $e_2 = w_3 = w_4$ are the unit vectors (Fig. 3.5). ♢

It follows that each convex polytope can be represented either as the set of convex combinations of a finite number of points (vertices), or as an intersection of a finite number of half-spaces. These representations are referred to as *V-polytopes* and *H-polytopes*, respectively. Both representations are useful in their own respect. For instance, the representation as V-polytopes is preferable if one wants to show that every projection of a polytope is a polytope. On the other hand, the representation as H-polytopes is preferable if one has to prove that every intersection of a polytope with an affine subspace is a polytope.

Example 3.24. Consider the polytope in Ex. 3.8. The vertices of this polytope are produced by the `polymake` command

```
> polymake dude VERTICES
VERTICES
1 0 8
1 0 5
1 1 3
1 3 2
```

The system also provides the vertex normals (i.e., the i-th row is the normal vector of a hyperplane separating the i-th vertex from the remaining ones),

```
VERTEX_NORMALS
0 1/3 1/3
0 -3 -1
0 1 -1
0 1 0
```

Furthermore, the updated file `dude` yields the representation of the polytope as a polyhedral set,

```
FACETS
-5 2 1
0 1 0
8 -2 -1
-7 1 2

AFFINE HULL
```

This output tells us that the polytope is defined by four linear inequalities

$$\begin{aligned} -5 + 2x_1 + x_2 &\geq 0, \\ x_1 &\geq 0, \\ 8 - 2x_1 - x_2 &\geq 0, \\ -7 + x_1 + 2x_2 &\geq 0, \end{aligned}$$

while there is no affine hull contribution which would provide additional linear equalities.

The command `DIM` confirms that the polytope is two-dimensional,

```
> polymake dude DIM
DIM
2
```

The f-vector of our polytope is

```
> polymake dude F_VECTOR
F_VECTOR
4 4
```

Inspecting the updated file `dude` again shows that each facet of the polytope is given by four lines,

```
VERTICES_IN_FACETS
{1 2}
{0 1}
{0 3}
{2 3}
```

♢

A *fan* in $\mathbb{R}^n$ is a family $\mathcal{F} = \{C_1, C_2, \ldots, C_m\}$ of nonempty cones with the following properties:

- Each non-empty face of a cone in $\mathcal{F}$ is also a cone in $\mathcal{F}$.
- The intersection of any two cones in $\mathcal{F}$ is a face of both.

A fan $\mathcal{F}$ in $\mathbb{R}^n$ is *complete* if the union $\bigcup \mathcal{F} = C_1 \cup \ldots \cup C_m$ equals $\mathbb{R}^n$. A fan $\mathcal{F}$ in $\mathbb{R}^n$ is *pointed* if $\{0\}$ is a cone in $\mathcal{F}$ and therefore a face of each cone in $\mathcal{F}$.

Example 3.25. The pointed fan in Fig. 3.6 in $\mathbb{R}^2$ has $m = 11$ cones, of which 5 are full dimensional. ♢

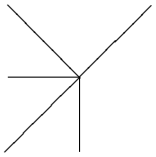


Fig. 3.6. A fan in $\mathbb{R}^2$.

Let P be a polytope in $\mathbb{R}^n$ and let F be a face of P. The *normal cone* of P at F is defined as

$$N_P(F) = \{w \in \mathbb{R}^n \mid F = P \cap H_{P,w}\}. \tag{3.16}$$

That is, $N_P(F)$ consists of all vectors $w \in \mathbb{R}^n$ with the property that F is the set of all points at which the linear map $P \to \mathbb{R} : x \mapsto \langle x, w\rangle$ given by the scalar product with fixed w attains the minimum. In particular, if $F = \{v\}$ is a vertex of P, then its normal cone $N_P(v)$ consists of all linear maps $P \to \mathbb{R} : x \mapsto \langle x, w\rangle$ that attain the minimum at the point v.

Example 3.26. Linear programming is a method to minimize or maximize a linear function over a convex set. The canonical form of a linear program in Euclidean n-space is

$$\begin{aligned} &\min\ c^T x \\ &\text{s.t.}\ \ Ax \geq b \\ &\text{and}\ \ x \geq 0 \end{aligned}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $c \in \mathbb{R}^n$ are given and x is the vector of variables. The objective function $\mathbb{R}^n \to \mathbb{R} : x \mapsto \langle c, x\rangle$ has to be minimized over the convex set $P = \{x \in \mathbb{R}^n \mid Ax \geq b, x \geq 0\}$. Suppose the minimum is attained at the face F of P. Then the normal cone $N_P(F)$ consists of all vectors c which attain the minimum at F. This amounts to an inverse problem of linear programming. ♢

Proposition 3.27. *Let P be a polytope in $\mathbb{R}^n$ and let F be a face of P. The normal cone $N_P(F)$ is a cone in $\mathbb{R}^n$ with dimension* $\dim N_P(F) = n - \dim F$.

Proof. For each $w \in \mathbb{R}^n$, let $f_w : P \to \mathbb{R} : x \mapsto \langle x, w\rangle$ be the scalar product with fixed w. Let $v, w \in \mathbb{R}^n$ such that the linear mappings f_v and f_w attain the minimum at F, and let $\lambda \geq 0$ and $\mu \geq 0$. Then the linear mapping $f_{\lambda v+\mu w}$ attains the minimum at F and so $\lambda v + \mu w$ belongs to $N_P(F)$.

Let F be a k-face. Then the face F is determined by $n - k$ linearly independent linear equations and the cone $N_P(F)$ is determined by k linearly independent linear equations. Hence, the dimension formula follows. □

Example 3.28. Consider the square P in $\mathbb{R}^2$ as given in Fig. 3.5. We may assume that it has a facet $F = \{(x, 0) \mid 0 \leq x \leq p\}$ for some real number $p > 0$. The affine subspace $U = \mathrm{aff}(F)$ is thus the real line given by the x-axis. Take the linear subspace W of $\mathbb{R}^2$ orthogonal to U; that is, $U \oplus W = \mathbb{R}^2$ as linear spaces. Then W is the real line given by the y-axis. For each $w \in W$, $\rho_P(w) = 0$ and so $N_P(F) = W$. Therefore, $\dim W = \dim \mathbb{R}^2 - \dim U = 1$, as required. ♢

The collection of all non-empty normal cones $N_P(F)$, as F runs over all faces of P, is called the *normal fan* of P and is denoted by $\mathcal{N}(P)$.

Proposition 3.29. *Let P be a polytope of $\mathbb{R}^n$. The normal fan $\mathcal{N}(P)$ is a complete fan of $\mathbb{R}^n$.*

Proof. Let $w \in \mathbb{R}^n$. If we put $F = P \cap H_{P,w}$, then $w \in N_P(F)$. It follows that the normal cones are non-empty and their union is the Euclidean n-space. □

Example 3.30. Consider the triangle P in $\mathbb{R}^2$ as given in Fig. 3.7. The normal cone of each vertex v is an cone $N_P(v)$ and the normal cone of an edge e is a half line $N_P(e)$. The normal fan consists of seven cones, of which are three full-dimensional (Fig. 3.8). ♢

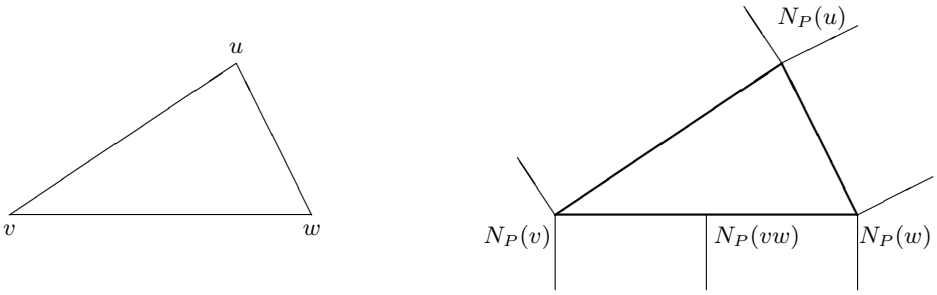


Fig. 3.7. A triangle and its normal cones.

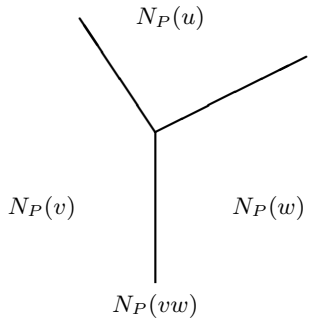


Fig. 3.8. The normal fan of the triangle in Fig. 3.7.

3.5 Polytope Algebra

We introduce important constructions of new polytopes from given ones. For this, the vector space structure of the Euclidean n-space is used.

Let P and Q be polytopes in $\mathbb{R}^n$. The *Minkowski sum* of P and Q is given as

$$P + Q = \{p + q \mid p \in P, q \in Q\}, \tag{3.17}$$

where $p + q$ denotes the addition in $\mathbb{R}^n$. Moreover, let $\lambda \geq 0$ be a real number. The polytope λP is defined by

$$\lambda P = \{\lambda p \mid p \in P\}, \tag{3.18}$$

where λp denotes the λ-multiple of $p \in P$.

Example 3.31. The Minkowski sum of the two polytopes P and Q in Fig. 3.9 can be obtained by placing a copy of P at every point of Q. This works because P contains the origin. ♢

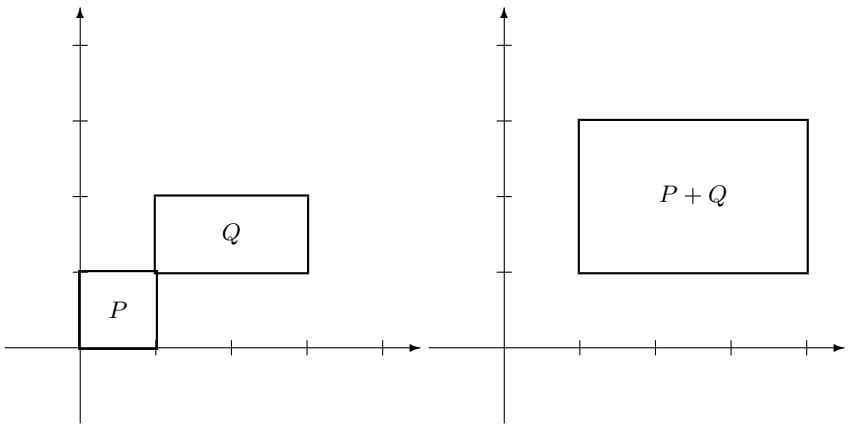


Fig. 3.9. Two polytopes and their Minkowski sum.

Proposition 3.32. *If P and Q are polytopes in $\mathbb{R}^n$, then the Minkowski sum $P + Q$ is also a polytope in $\mathbb{R}^n$.*

Proof. Let $p, p' \in P$ and $q, q' \in Q$. For each λ, $0 \leq \lambda \leq 1$, we have

$$\lambda(p + q) + (1 - \lambda)(p' + q') = [\lambda p + (1 - \lambda)p'] + [\lambda q + (1 - \lambda)q'] \in P + Q.$$

Thus the Minkowski sum $P + Q$ is convex.

Let $P = \text{conv}(A)$ for some finite subset $A = \{a_1, \ldots, a_m\}$ of $\mathbb{R}^n$, and let $p \in P$ and $q \in Q$. Write $p = \sum_i \lambda_i a_i$, where $\lambda_i \geq 0$, $1 \leq i \leq m$, and $\sum_i \lambda_i = 1$. Then

$$p + q = \left(\sum_i \lambda_i a_i\right) + q = \sum_i \lambda_i \left(a_i + q\right) \in \text{conv}\left(\bigcup_i (a_i + Q)\right).$$

Conversely, by Prop. 3.7 and the fact that $P+Q$ is convex, we obtain

$$\operatorname{conv}\left(\bigcup_i (a_i+Q)\right) \subseteq \operatorname{conv}(P+Q) = P+Q.$$

It follows that

$$P+Q = \operatorname{conv}\left(\bigcup_i (a_i+Q)\right).$$

Let $Q=\operatorname{conv}(B)$ for some finite subset B of $\mathbb{R}^n$. Then by Prop. 3.10, $a_i+Q=\operatorname{conv}(a_i+B)$, $1\le i\le m$, and by Prop. 3.7,

$$P+Q = \operatorname{conv}\left(\bigcup_i \operatorname{conv}(a_i+B)\right) = \operatorname{conv}\left(\bigcup_i (a_i+B)\right).$$

Thus $P+Q$ has a finite generating set and hence is a polytope. □

Proposition 3.33. *Let P and Q be polytopes in $\mathbb{R}^n$, and let $w\neq 0$ be a vector in $\mathbb{R}^n$. We have*

$$\rho_{P+Q}(w) = \rho_P(w)+\rho_Q(w) \quad \textit{and} \quad (P+Q)_w = P_w+Q_w.$$

Proof. First, we have

$$\begin{aligned}\rho_{P+Q}(w) &= \min\{\langle v,w\rangle \mid v\in P+Q\} = \min\{\langle p,w\rangle+\langle q,w\rangle \mid p\in P, q\in Q\}\\ &= \min\{\langle p,w\rangle \mid p\in P\}+\min\{\langle q,w\rangle \mid q\in Q\}\\ &= \rho_P(w)+\rho_Q(w).\end{aligned}$$

Second, by the first asseration,

$$\begin{aligned}(P+Q)_w &= (P+Q)\cap H_{P+Q,w} = \{p+q \mid \langle p+q,w\rangle = \rho_{P+Q}(w), p\in P, q\in Q\}\\ &= \{p+q \mid \langle p,w\rangle+\langle q,w\rangle = \rho_P(w)+\rho_Q(w), p\in P, q\in Q\}\\ &= \{p \mid \langle p,w\rangle = \rho_P(w), p\in P\}+\{q \mid \langle q,w\rangle = \rho_Q(w), q\in Q\}\\ &= (P\cap H_{P,w})+(Q\cap H_{Q,w}) = P_w+Q_w.\end{aligned}$$

□

The *polytope algebra* on $\mathbb{R}^n$ is a triple $(\mathcal{P}_n,\oplus,\odot)$ that consists of the set of all polytopes in $\mathbb{R}^n$, denoted by $\mathcal{P}_n$, and two arithmetic operations $\oplus$ and $\odot$, called addition and multiplication, defined as

$$P\oplus Q = \operatorname{conv}(P\cup Q) \quad \text{and} \quad P\odot Q = P+Q. \tag{3.19}$$

By Prop. 3.32, the multiplication is well-defined.

Proposition 3.34. *If $P=\operatorname{conv}(A)$ and $Q=\operatorname{conv}(B)$ are polytopes in $\mathbb{R}^n$, then*

$$P\oplus Q = \operatorname{conv}(A\cup B).$$

Proof. By Prop. 3.7, we have $\text{conv}(A \cup B) \subseteq \text{conv}(P \cup Q) = P \oplus Q$. Conversely, let $v \in P \oplus Q$. Write the point v as a convex combination of points in $P \cup Q$; that is, $v = \sum_i \lambda_i p_i + \sum_j \mu_j q_j$ for some points $p_i \in P$ and $q_j \in Q$. But by definition, the points p_i and q_j are convex combinations of the points in the generating sets A and B, respectively. It follows that $v \in \text{conv}(A \cup B)$. □

Example 3.35. Consider the non-collinear line segments $P = \{(x, 0) \mid 0 \leq x \leq p\}$ and $Q = \{(0, y) \mid 0 \leq y \leq q\}$ in $\mathbb{R}^2$. Their sum and product are illustrated in Fig. 3.10. ♢

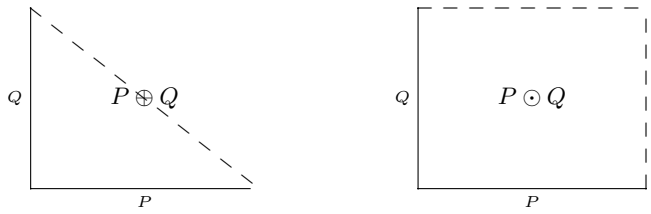


Fig. 3.10. Two line segments P and Q in $\mathbb{R}^2$ and their sum $P \oplus Q$ and product $P \odot Q$.

Proposition 3.36. *The polytope algebra $(\mathcal{P}_n, \oplus, \odot)$ on $\mathbb{R}^n$ is a commutative, idempotent semiring.*

Proof. It is easy to see that $(\mathcal{P}_n, \oplus)$ is a commutative monoid with identity element $\emptyset$ and $(\mathcal{P}_n, \odot)$ is a commutative monoid with identity element $\{0\}$. Moreover, by Prop. 3.7, the addition is idempotent and multiplication with the empty set annihilates $\mathcal{P}_n$. To see that the distributive law holds, take $p \in P$, $q \in Q$, and $r \in R$. Then for each λ with $0 \leq \lambda \leq 1$, we have

$$p + (\lambda q + (1 - \lambda)r) = \lambda(p + q) + (1 - \lambda)(p + r).$$

The left-hand side is a point in $P \odot (Q \oplus R)$ and the right-hand side is a point in $(P \odot Q) \oplus (P \odot R)$. □

Example 3.37. Consider the polytope algebra $\mathcal{P}_1$ on the Euclidean 1-space. The elements of $\mathcal{P}_1$ are exactly the line segments $[a, b] = \{\lambda a + (1 - \lambda)b \mid 0 \leq \lambda \leq 1\}$, where $a, b \in \mathbb{R}$. The sum and product of two line segments $[a, b]$ and $[c, d]$ are given by

$$[a, b] \oplus [c, d] = [\min\{a, c\}, \max\{b, d\}] \quad \text{and} \quad [a, b] \odot [c, d] = [a + c, b + d].$$

♢

Proposition 3.38. *The mapping $f : \mathcal{P}_1 \to \mathbb{R} \cup \{\infty\} : [a, b] \mapsto a$ is an epimorphism from the polytope algebra $\mathcal{P}_1$ onto the tropical algebra.*

Proof. The mapping is well-defined and we have $f(\emptyset) = \infty$ and $f(\{0\}) = f([0,0]) = 0$. For any two line segments $[a,b]$ and $[c,d]$ in $\mathcal{P}_1$,

$$f([a,b] \oplus [c,d]) = f([\min\{a,c\}, \max\{b,d\}]) = \min\{a,c\} = a \oplus c = f([a,b]) \oplus f([c,d])$$

and

$$f([a,b] \odot [c,d]) = f([a+c, b+d]) = a + c = a \odot c = f([a,b]) \odot f([c,d]).$$

□

In this way, the polytope algebra on $\mathbb{R}^n$ can be viewed as a natural higher-dimensional generalization of the tropical algebra.

3.6 Newton Polytopes

We establish an interesting connection between lattice polytopes and polynomials. For this, take a polynomial f in the polynomial ring $\mathbb{K}[X_1, \ldots, X_n]$ and write

$$f = \sum_{\alpha \in \mathbb{N}_0^n} c_\alpha X^\alpha.$$

The *Newton polytope* of f, denoted as $\mathrm{NP}(f)$, is the lattice polytope

$$\mathrm{NP}(f) = \mathrm{conv}(\{\alpha \in \mathbb{N}_0^n \mid c_\alpha \neq 0\}).$$

That is, the Newton polytope is generated by the exponents of the monomials involved in the polynomial. It is a measure of shape or sparsity of a polynomial. Note that the actual values of the coefficients do not matter in the definition of the Newton polytope.

Example 3.39. Any polynomial in $\mathbb{K}[X,Y]$ of the form

$$f = aXY + bX^2 + cY^3 + d,$$

where a, b, c, d are nonzero elements of $\mathbb{K}$, has the Newton polytope equal to the triangle

$$P = \mathrm{conv}(\{(1,1), (2,0), (0,3), (0,0)\}).$$

By Ex. 3.21, polynomials of the above form with $a = 0$ have the same Newton polytope. ♢

We can also go the other way, from exponents to polynomials. Suppose we have a finite set of exponents $A = \{\alpha_1, \ldots, \alpha_l\}$ in $\mathbb{N}_0^n$. Then let $L(A)$ be the set of all polynomials whose terms all have the exponents in A,

$$L(A) = \{c_1 X^{\alpha_1} + \cdots + c_l X^{\alpha_l} \mid c_1, \ldots, c_l \in \mathbb{K}\}.$$

Note that $L(A)$ is a vector space over $\mathbb{K}$ with basis $\{X^{\alpha_1}, \ldots, X^{\alpha_l}\}$ and dimension l. The following result is immediate from the definitions.

Proposition 3.40. *Let A be a finite subset of $\mathbb{N}_0^n$. For each polynomial f in $L(A)$, we have* $\mathrm{NP}(f) \subseteq \mathrm{conv}(A)$.

The formation of Minkowskis sums is compatible with polynomial multiplication. To see this, let $w \neq 0$ be a vector in $\mathbb{R}^n$ and $f = \sum_\alpha c_\alpha X^\alpha$ be a polynomial in $\mathbb{K}[X_1, \ldots, X_n]$. Define the number

$$\pi_f(w) = \min\{\langle \alpha, w\rangle \mid c_\alpha \neq 0\}.$$

This number exists, since each polynomial has only a finite number of terms. The *initial form* of f with respect to w is the subsum $\mathrm{in}_w(f)$ of all terms $c_\alpha X^\alpha$, $c_\alpha \neq 0$, such that $\langle \alpha, w\rangle$ is minimal. That is,

$$\mathrm{in}_w(f) = \sum\{c_\alpha X^\alpha \mid c_\alpha \neq 0, \langle \alpha, w\rangle = \pi_f(w)\}.$$

Example 3.41. Any polynomial in $\mathbb{R}[X, Y]$ of the form

$$f = aXY + bX^2 + cY^3 + d,$$

where a, b, c, d are nonzero real numbers, has with respect to $w = (-2, 2)$ the initial form

$$\mathrm{in}_w(f) = bX^2.$$

♢

Proposition 3.42. *Let f and g be polynomials in $\mathbb{K}[X_1, \ldots, X_n]$, and let $w \neq 0$ be a vector in $\mathbb{R}^n$. We have*

$$\begin{aligned} \mathrm{in}_w(f \cdot g) &= \mathrm{in}_w(f) \cdot \mathrm{in}_w(g), \\ \pi_f(w) &= \rho_{\mathrm{NP}(f)}(w), \\ \mathrm{NP}(\mathrm{in}_w(f)) &= \mathrm{NP}(f)_w. \end{aligned}$$

Proof. Let $f = \sum_\alpha c_\alpha X^\alpha$ and $g = \sum_\beta d_\beta X^\beta$ be polynomials in $\mathbb{K}[X_1, \ldots, X_n]$. First, we have

$$\begin{aligned} \pi_{f\cdot g}(w) &= \min\{\langle \alpha + \beta, w\rangle \mid c_\alpha \neq 0, d_\beta \neq 0\} \\ &= \min\{\langle \alpha, w\rangle \mid c_\alpha \neq 0\} + \min\{\langle \beta, w\rangle \mid d_\beta \neq 0\} \\ &= \pi_f(w) + \pi_g(w). \end{aligned}$$

Then we obtain

$$\begin{aligned} \mathrm{in}_w(f \cdot g) &= \sum\{c_\alpha d_\beta X^{\alpha+\beta} \mid c_\alpha \neq 0, d_\beta \neq 0, \langle \alpha + \beta, w\rangle = \pi_{f\cdot g}(w)\} \\ &= \sum\{c_\alpha X^\alpha \mid c_\alpha \neq 0, \langle \alpha, w\rangle = \pi_f(w)\} \cdot \sum\{d_\beta X^\beta \mid d_\beta \neq 0, \langle \beta, w\rangle = \pi_g(w)\} \\ &= \mathrm{in}_w(f) \cdot \mathrm{in}_w(g). \end{aligned}$$

Second, we have $\pi_f(w) \geq \rho_{\mathrm{NP}(f)}(w)$ by definition. On the other hand, the Newton polytope $\mathrm{NP}(f)$ is the convex hull of the set $A = \{\alpha \mid c_\alpha \neq 0\}$. By the proof of Prop. 3.16, the vertices of $\mathrm{NP}(f)$ belong to this set. Let $a_1, \ldots, a_m \in A$ be the vertices of $\mathrm{NP}(f)$. We may assume that $\langle a_1, w\rangle \leq \langle a_i, w\rangle$ for each i, $1 \leq i \leq m$. Let $v \in \mathrm{NP}(f)$ such that $\rho_{\mathrm{NP}(f)}(w) = \langle v, w\rangle$. By Prop. 3.20, $v = \sum_i \lambda_i a_i$, where $\lambda_i \geq 0$, $1 \leq i \leq m$, and $\sum_i \lambda_i = 1$. Then $\langle v, w\rangle = \sum_i \lambda_i \langle a_i, w\rangle \geq \sum_i \lambda_i \langle a_1, w\rangle = \langle a_1, w\rangle$. It follows that $\rho_{\mathrm{NP}(f)}(w) \geq \pi_f(w)$.

Third, by using the last assertion, we obtain

$$\begin{aligned}\mathrm{NP}(f)_w &= \mathrm{NP}(f) \cap H_{\mathrm{NP}(f),w} \\ &= \mathrm{conv}\{\alpha \mid c_\alpha \neq 0\} \cap \{v \mid v \in \mathrm{NP}(f), \langle v, w\rangle = \rho_{\mathrm{NP}(f)}(w)\} \\ &= \mathrm{conv}\{\alpha \mid c_\alpha \neq 0\} \cap \{v \mid v \in \mathrm{NP}(f), \langle v, w\rangle = \pi_f(w)\} \\ &= \mathrm{conv}\{\alpha \mid c_\alpha \neq 0, \langle \alpha, w\rangle = \pi_f(w)\} \\ &= \mathrm{NP}(\mathrm{in}_w(f)).\end{aligned}$$

□

Lemma 3.43. *If f and g are polynomials in $\mathbb{K}[X_1, \ldots, X_n]$ and $w \neq 0$ is a vector in $\mathbb{R}^n$, then*

$$\mathrm{NP}(\mathrm{in}_w(f) \cdot \mathrm{in}_w(g)) = \mathrm{NP}(f)_w + \mathrm{NP}(g)_w. \tag{3.20}$$

Proof. Let $f = \sum_\alpha c_\alpha X^\alpha$ and $g = \sum_\beta d_\beta X^\beta$ be polynomials in $\mathbb{K}[X_1, \ldots, X_n]$. We have

$$\begin{aligned}\mathrm{NP}(\mathrm{in}_w(f) \cdot \mathrm{in}_w(g)) &= \mathrm{conv}\{\alpha + \beta \mid c_\alpha \neq 0, d_\beta \neq 0, \langle \alpha, w\rangle = \pi_f(w), \langle \beta, w\rangle = \pi_g(w)\} \\ &= \mathrm{conv}\{\alpha \mid c_\alpha \neq 0, \langle \alpha, w\rangle = \pi_f(w)\} + \mathrm{conv}\{\beta \mid d_\beta \neq 0, \langle \beta, w\rangle = \pi_g(w)\} \\ &= \mathrm{NP}(f)_w + \mathrm{NP}(g)_w.\end{aligned}$$

□

Theorem 3.44. *Let f and g be polynomials in $\mathbb{K}[X_1, \ldots, X_n]$. Then*

$$\mathrm{NP}(f \cdot g) = \mathrm{NP}(f) \odot \mathrm{NP}(g).$$

Proof. Let $w \neq 0$ be a vector in $\mathbb{R}^n$. We have by Prop. 3.42, Lemma 3.43, and Prop. 3.33,

$$\begin{aligned}\mathrm{NP}(f \cdot g)_w &= \mathrm{NP}(\mathrm{in}_w(f \cdot g)) \\ &= \mathrm{NP}(\mathrm{in}_w(f) \cdot \mathrm{in}_w(g)) \\ &= \mathrm{NP}(f)_w + \mathrm{NP}(g)_w \\ &= (\mathrm{NP}(f) + \mathrm{NP}(g))_w .\end{aligned}$$

This equality shows that the polytopes $\mathrm{NP}(f \cdot g)$ and $\mathrm{NP}(f) \odot \mathrm{NP}(g)$ have the same set of vertices. But by Prop. 3.20, each polytope is the convex hull of its vertices and so the result follows. □

Theorem 3.45. *Let f and g be polynomials in $\mathbb{K}[X_1, \ldots, X_n]$. Then*

$$\mathrm{NP}(f + g) \subseteq \mathrm{NP}(f) \oplus \mathrm{NP}(g).$$

Equality holds, if all coefficients in the polynomials f and g are positive.

Proof. Let $f = \sum_\alpha c_\alpha X^\alpha$ and $g = \sum_\beta d_\beta X^\beta$ be polynomials in $\mathbb{K}[X_1, \ldots, X_n]$. By Prop. 3.34, we have

$$\begin{aligned}\mathrm{NP}(f) \oplus \mathrm{NP}(g) &= \mathrm{conv}(\mathrm{conv}(\{\alpha \mid c_\alpha \neq 0\}) \cup \mathrm{conv}(\{\beta \mid d_\beta \neq 0\})) \\ &= \mathrm{conv}(\{\alpha, \beta \mid c_\alpha \neq 0, d_\beta \neq 0\}).\end{aligned}$$

On the other hand, we have $f + g = \sum_\gamma (c_\gamma + d_\gamma) X^\gamma$ and so

$$\mathrm{NP}(f+g) = \mathrm{conv}(\{\gamma \mid c_\gamma + d_\gamma \neq 0\}).$$

Let $c_\gamma + d_\gamma \neq 0$. Then $c_\gamma \neq 0$ or $d_\gamma \neq 0$ and so $\gamma \in \{\alpha, \beta \mid c_\alpha \neq 0, d_\beta \neq 0\}$. This proves the inclusion.

Finally, suppose that all coefficients in f and g are positive. Then $c_\gamma \neq 0$ or $d_\gamma \neq 0$ implies that $c_\gamma + d_\gamma \neq 0$. Thus the other inclusion also holds and hence both polytopes are equal. □

Example 3.46. In $\mathbb{R}[X, Y]$ consider the polynomials

$$f = X^p + 1 \quad \text{and} \quad g = Y^q + 1,$$

where p and q are positive integers. The corresponding Newton polytopes are line segments in $\mathbb{R}^2$ given as

$$\mathrm{NP}(f) = \mathrm{conv}(\{(0,0),(p,0)\}) = \{(x,0) \mid 0 \leq x \leq p\}$$

and

$$\mathrm{NP}(g) = \mathrm{conv}(\{(0,0),(0,q)\}) = \{(0,y) \mid 0 \leq y \leq q\}.$$

The sum $f + g$ has the Newton polytope

$$\mathrm{NP}(f+g) = \mathrm{NP}(X^p + Y^q + 2) = \mathrm{conv}(\{(p,0),(0,q),(0,0)\}),$$

which is a triangle with vertices $(0,0)$, $(p,0)$, and $(0,q)$, and the product $f \cdot g$ has the Newton polytope

$$\mathrm{NP}(f \cdot g) = \mathrm{NP}(x^p y^q + x^p + y^q + 1) = \mathrm{conv}(\{(p,q),(p,0),(0,q),(0,0)\}),$$

which is a rectangle with vertices $(0,0)$, $(p,0)$, $(0,q)$, and (p,q) (Ex. 3.35). ♢

3.7 Parametric Shortest Path Problem

The problem of finding shortest paths in a network can be extended by making use of the polytope algebra. For this, let $G = (V, E)$ be a digraph with vertex set $V = \{1, \ldots, n\}$ and edge set E. Assume that each edge (i, j) in G has an associated polytope P_{ij} in the Euclidean d-space. We put $P_{ii} = \{0\}$, $1 \leq i \leq n$, and $P_{ij} = \emptyset$ if (i, j), $i \neq j$, is not an edge in G. We represent the digraph G by the $n \times n$ matrix $D_G = (P_{ij})$ of polytopes in the polytope algebra $\mathcal{P}_d$.

Each vector $w \in \mathbb{R}^n$ allows to assign scalar values to the edges (i, j) in G by linear programming on the polytope P_{ij}:

$$d_{ij} = d_{ij}(w) = \min\{\langle w, p\rangle \mid p \in P_{ij}\}, \quad (i,j) \in E. \tag{3.21}$$

In particular, we have $d_{ii} = 0$, $1 \leq i \leq n$, and $d_{ij} = \infty$ if (i, j), $i \neq j$, is not an edge in G. Thus each vector $w \in \mathbb{R}^n$ gives rise to an $n \times n$ adjacency matrix $D_G = D_{G,w} = (d_{ij})$ with respect to w. We show that the lengths of the shortest paths in G given by the $n \times n$ adjacency matrix D_G with respect to w can be derived by computation in the polytope algebra $\mathcal{P}_d$.

Proposition 3.47. *Let G be a digraph on n vertices with $n \times n$ adjacency matrix D_G and let $w \in \mathbb{R}^d$. The length of the shortest path from vertex i to vertex j in the digraph G is given by*

$$d_{ij}^{(n-1)} = \min\{\langle w, p\rangle \mid p \in P_{ij}^{(n-1)}\},$$

where $(P_{ij}^{(n-1)})$ is the (i,j)-th entry in the $(n-1)$th power of the matrix D_G computed in the polytope algebra $\mathcal{P}^d$.

Proof. The proof of Prop. 3.4 shows that the lengths of the shortest paths satisfy the recursion formula

$$d_{ij}^{(r)} = \min\{d_{ik}^{(r-1)} + d_{kj} \mid 1 \le k \le n\}, \quad 2 \le r \le n-1.$$

We put $P_{ij}^{(1)} = P_{ij}$, $1 \le i, j \le n$. Claim that for $1 \le r \le n-1$,

$$d_{ij}^{(r)} = \min\{\langle w, p\rangle \mid p \in P_{ij}^{(r)}\}.$$

Indeed, this assertion holds by definition for $r = 1$. For $2 \le r \le n-1$, we have

$$\begin{aligned}
d_{ij}^{(r)} &= \min\{d_{ik}^{(r-1)} + d_{kj} \mid 1 \le k \le n\} \\
&= \min\{\min\{\langle w, p\rangle \mid p \in P_{ik}^{(r-1)}\} + \min\{\langle w, p\rangle \mid p \in P_{kj}^{(1)}\} \mid 1 \le k \le n\} \\
&= \min\{\min\{\langle w, p\rangle \mid p \in P_{ik}^{(r-1)} \odot P_{kj}\} \mid 1 \le k \le n\} \\
&= \min\{\langle w, p\rangle \mid p \in \bigoplus_{k=1}^{n} P_{ik}^{(r-1)} \odot P_{kj}\} \\
&= \min\{\langle w, p\rangle \mid p \in P_{ij}^{(r)}\}.
\end{aligned}$$

The second equality follows from the induction hypothesis, the third from the definition of multiplication in the polytope algebra, and the fourth from the definition of addition in the polytope algebra and the fact that the minimum is attained at a vertex which is, by Prop. 3.34, a vertex of one of the involved polytopes. □

The Floyd-Warshall algorithm for finding shortest paths in a weighted digraph can be extended to this parametric setting. If the parameter d is kept fixed, the algorithm still runs in polynomial time.

Example 3.48. Reconsider the directed graph G in Ex. 3.5. Suppose the adjacency matrix of G is defined over the polytope algebra $\mathcal{P}_d$ as follows

$$D_G = \begin{pmatrix} \{0\} & P & \emptyset & \emptyset \\ \emptyset & \{0\} & P & \emptyset \\ P & \emptyset & \{0\} & \emptyset \\ P & \emptyset & \emptyset & \{0\} \end{pmatrix},$$

where P is a polytope in $\mathbb{R}^d$. Then we have

$$D_G^{\odot 2} = \begin{pmatrix} \{0\} & P & P^{\odot 2} & \emptyset \\ P^{\odot 2} & \{0\} & P & \emptyset \\ P & P^{\odot 2} & \{0\} & \emptyset \\ P & P^{\odot 2} & \emptyset & \{0\} \end{pmatrix} \quad \text{and} \quad D_G^{\odot 3} = \begin{pmatrix} \{0\} & P & P^{\odot 2} & \emptyset \\ P^{\odot 2} & \{0\} & P & \emptyset \\ P & P^{\odot 2} & \{0\} & \emptyset \\ P & P^{\odot 2} & P^{\odot 3} & \{0\} \end{pmatrix}.$$

♢

Part II

Algebraic Statistics

4

Basic Algebraic Statistical Models

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. Statistics builds models of the process that generated the data. In descriptive statistics, data are summarizes and measured by indexes such as mean and standard deviation, while in inferential statistics, conclusions about data are drawn subject to random variation such as confidences intervals and hypothesis testing. In this chapter, some basic algebraic statistical models are introduced which will serve as a basis for the subsequent chapters.

4.1 Introductory Example

We consider a statistical model called DiaNA that produces sequences of symbols over the DNA alphabet $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ such as

$$\mathtt{CTCACGTGATGAGAGCATTCTCAGACCGTGACGCGTGTAGCAGCGGCTC}. \tag{4.1}$$

DiaNA uses three tetrahedral dice to generate DNA sequences. The first two dice are loaded and the third die is fair (Table 4.1). DiaNA first picks one of her three dice at random, where the first die (GC-rich) is picked with probability θ_1, the second die (GC-poor) is picked with probability θ_2, and the third die is picked with probability $1 - \theta_1 - \theta_2$.

Table 4.1. The three tetrahedral dice of DiaNA.

	A	C	G	T
first die	0.15	0.33	0.36	0.16
second die	0.27	0.24	0.23	0.26
third die	0.25	0.25	0.25	0.25

DiaNA uses the following probabilities to generate the four symbols:

$$p_A = -0.10 \cdot \theta_1 + 0.02 \cdot \theta_2 + 0.25,$$

$$\begin{aligned} p_C &= 0.08 \cdot \theta_1 - 0.01 \cdot \theta_2 + 0.25, \\ p_G &= 0.11 \cdot \theta_1 - 0.02 \cdot \theta_2 + 0.25, \\ p_T &= -0.09 \cdot \theta_1 + 0.01 \cdot \theta_2 + 0.25. \end{aligned} \tag{4.2}$$

We have

$$p_A + p_C + p_G + p_T = 1,$$

and the three distributions in the rows of Table 4.1 are obtained by specializing (θ_1, θ_2) to (1,0), (0,1), and (0,0), respectively.

Consider the likelihood of observing the data (4.1). For this, note that the data contains 10 `A`'s, 14 `C`'s, 15 `G`'s, and 10 `T`'s. Assume that all symbols were independently generated. Then the likelihood of observing the data is given by

$$L = p_A^{10} \cdot p_C^{14} \cdot p_G^{15} \cdot p_T^{10}.$$

The likelihood function $L = L(\theta_1, \theta_2)$ is a real-valued function on the triangle

$$\Theta = \{(\theta_1, \theta_2) \mid \theta_1 > 0, \theta_2 > 0, \theta_1 + \theta_2 < 1\}.$$

Equivalently, the likelihood of observing the data can be described by the log-likelihood function

$$\begin{aligned} \ell(\theta_1, \theta_2) &= \log L(\theta_1, \theta_2) \\ &= 10 \cdot \log p_A(\theta_1, \theta_2) + 14 \cdot \log p_C(\theta_1, \theta_2) + 15 \cdot \log p_G(\theta_1, \theta_2) + 10 \cdot \log p_T(\theta_1, \theta_2). \end{aligned}$$

The parameters θ_1 and θ_2 can be estimated by maximizing this likelihood function. For this, we equate the two partial derivatives of the function to zero:

$$\begin{aligned} \frac{\partial \ell}{\partial \theta_1} &= \frac{10}{p_A} \cdot \frac{\partial p_A}{\partial \theta_1} + \frac{14}{p_C} \cdot \frac{\partial p_A}{\partial \theta_1} + \frac{15}{p_G} \cdot \frac{\partial p_A}{\partial \theta_1} + \frac{10}{p_T} \cdot \frac{\partial p_A}{\partial \theta_1} = 0, \\ \frac{\partial \ell}{\partial \theta_2} &= \frac{10}{p_A} \cdot \frac{\partial p_A}{\partial \theta_2} + \frac{14}{p_C} \cdot \frac{\partial p_A}{\partial \theta_2} + \frac{15}{p_G} \cdot \frac{\partial p_A}{\partial \theta_2} + \frac{10}{p_T} \cdot \frac{\partial p_A}{\partial \theta_2} = 0. \end{aligned}$$

We use `Maple` to solve these equations:

```
> pA := -0.10*x + 0.02*y + 0.25:
> pC :=  0.08*x - 0.01*y + 0.25:
> pG :=  0.11*x - 0.02*y + 0.25:
> pT := -0.09*x - 0.01*y + 0.25:

> L := pA^10 * pC^14 * pG^15 * pT^10:
> l := log( L ):

> lx := diff(l, x):
> ly := diff(l, y):

> fsolve( {lx=0, ly=0}, {x,y}, {x=0..1}, {y=0..1} );
```

The `fsolve` command provides the critical point

$$\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2) = (0.5191263945, 0.2172513326).$$

The corresponding probability distribution is

$$(\hat{p}_A, \hat{p}_C, \hat{p}_G, \hat{p}_T) = (0.202432, 0.289358, 0.302759, 0.205451).$$

This distribution lies very close to the empirical distribution

$$\frac{1}{49}(10, 14, 15, 10) = (0.204082, 0.285714, 0.306122, 0.204082).$$

To determine the nature of the critical point $\hat{\theta}$, we examine the corresponding Hessian matrix

$$H = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_2^2} \end{pmatrix}.$$

At the critical point $\theta = \hat{\theta}$, the Hessian matrix equals

$$\begin{pmatrix} -7.409465471 & 1.195056562 \\ 1.195056562 & -0.2034803046 \end{pmatrix}.$$

Since the Hessian matrix is a real-valued symmetric matrix, its eigenvalues are real-valued

$$-7.602486025, \ -0.01045975018.$$

As the eigenvalues are negative, the Hessian matrix is negative definite. Thus the critical point $\hat{\theta}$ is a local maximum of the likelihood function $\ell(\theta)$. These calculations can be carried out by `Maple` as follows:

```
> x := 0.5191263945: y := 0.2172513326:
> with( linalg ):
> H := matrix( [[ diff(diff(l,x),x), diff(diff(l,x),y) ],
                [ diff(diff(l,y),x), diff(diff(l,y),y) ]] );
> eigenvalues ( H );
```

4.2 General Algebraic Statistical Model

The above example exhibits all characteristics of an algebraic statistical model. In general, one considers a *state space* given by the first m positive integers

$$[m] := \{1, \ldots, m\}. \tag{4.3}$$

A probability distribution on the set $[m]$ is a point in the probability simplex

$$\Delta_{m-1} = \{(p_1, \ldots, p_m) \in [0,1]^m \mid \sum_i p_i = 1\}. \tag{4.4}$$

An *algebraic statistical model* is defined by a polynomial map $f : \mathbb{R}^d \to \mathbb{R}^m$ given by

$$(\theta_1, \ldots, \theta_d) \mapsto (f_1(\theta), \ldots, f_m(\theta)), \tag{4.5}$$

where $\theta_1, \ldots, \theta_d$ are the model parameters and $f_1, \ldots, f_m$ are polynomials (or rational functions) in $\mathbb{R}[X_1, \ldots, X_n]$. Note that the number of parameters d is usually much smaller than the size of the state space m.

The parameter vector $(\theta_1, \ldots, \theta_d)$ ranges over a suitable nonempty open subset Θ of $\mathbb{R}^d$, the *parameter space* of f. We assume that the parameter space Θ satisfies

$$\Theta \subseteq \{\theta \in \mathbb{R}^d \mid f_i(\theta) > 0, 1 \leq i \leq m\} \tag{4.6}$$

Thus we have

$$f(\Theta) \subseteq \Delta_{m-1} \Longleftrightarrow f_1(\theta) + \ldots + f_m(\theta_m) = 1. \tag{4.7}$$

The right-hand side is an identity of polynomial functions in which all nonconstant terms cancel and the constant terms add up to 1. If (4.7) holds, then the model is simply the set $f(\Theta)$.

However, not all algebraic statistical models satisfy (4.7). In this case, the vectors in $f(\Theta)$ can be scaled to obtain a family of probability distributions on $[m]$,

$$\frac{1}{\sum_i f_i(\theta)} \cdot (f_1(\theta), \ldots, f_m(\theta)), \quad \theta \in \Theta. \tag{4.8}$$

The denominator polynomial $\sum_i f_i(\theta)$ is known as the *partition function* of the model.

The sample data are typically given by a sequence of values from the state space,

$$i_1, i_2, i_3, \ldots, i_N. \tag{4.9}$$

The integer N is the sample size. Assume that the values are independent and identically distributed. Then the data can be summarized by the frequency vector

$$u = (u_1, u_2, \ldots, u_m), \tag{4.10}$$

where u_i is the number of occurences of $i \in [m]$ in the data, $1 \leq i \leq m$. It follows that

$$u_1 + u_2 + \ldots + u_m = N \tag{4.11}$$

and the empirical distribution corresponding to the data is given by the scaled vector

$$\frac{1}{N}(u_1, u_2, \ldots, u_m), \tag{4.12}$$

which belongs to the probability simplex Δ_{m-1}. The coordinates u_i/N are the observed relative frequencies of the outcomes.

Consider an algebraic statistical model $f : \mathbb{R}^d \to \mathbb{R}^m$ for the data. The probability of observing the data (4.9) is given by

$$L(\theta) = f_{i_1}(\theta) f_{i_2}(\theta) \cdots f_{i_N}(\theta) = f_1(\theta)^{u_1} \cdots f_{i_m}(\theta)^{u_m}. \tag{4.13}$$

If the frequency vector u is kept fixed, the likelihood function L is a function from the parameter space Θ to the positive real numbers.

By reordering the data (4.9), we obtain the same frequency vector u. Thus the probability of observing the frequency vector u is given by

$$\binom{N}{u_1, \ldots, u_m} L(\theta). \tag{4.14}$$

The frequency vector is a *sufficient statistic* for the model f since the likelihood function $L(\theta)$ depends on the data only through u (and not through the data itself).

We may run into numerical problems when multiplying many probabilities. For this, we use the log transformation and represent the likelihood function by the log-likelihood function

$$\ell(\theta) = \log L(\theta) = u_1 \cdot \log f_1(\theta) + \ldots + u_m \cdot \log f_m(\theta). \tag{4.15}$$

The log-likelihood function $\ell(\theta)$ is a function from the parameter space Θ to the negative real numbers.

The problem of *maximum likelihood estimation* is to maximize the likelihood function $L(\theta)$ or, equivalently, the scaled likelihood function (4.14), or, equivalently, the scaled log-likelihood function $\ell(\theta)$, over the parameter space:

$$\begin{aligned} &\max \ell(\theta) \\ &\text{s.t. } \theta \in \Theta \end{aligned} \tag{4.16}$$

A solution to this optimization problem is called *maximum likelihood estimate* of θ with respect to the model f and the data u. The simplest algebraic statistical models are the linear and toric models, since they easily allow to establish maximum likelihood estimates.

4.3 Linear Models

An algebraic statistical model $f : \mathbb{R}^d \to \mathbb{R}^m$ is called a *linear model* if its coordinate functions $f_i(\theta)$, $1 \leq i \leq d$, are linear functions. That is, there are real numbers $a_{i1}, \ldots, a_{id}$ and b_i, $1 \leq i \leq d$, such that

$$f_i(\theta) = \sum_{j=1}^{d} a_{ij}\theta_j + b_j, \quad 1 \leq i \leq d. \tag{4.17}$$

For instance, DiaNA is a linear model $f : \mathbb{R}^2 \to \mathbb{R}^4$ given by the coordinate functions

$$\begin{aligned} f_1(\theta) &= -0.10 \cdot \theta_1 + 0.02 \cdot \theta_2 + 0.25, \\ f_2(\theta) &= 0.08 \cdot \theta_1 - 0.01 \cdot \theta_2 + 0.25, \\ f_3(\theta) &= 0.11 \cdot \theta_1 - 0.02 \cdot \theta_2 + 0.25, \\ f_4(\theta) &= -0.09 \cdot \theta_1 + 0.01 \cdot \theta_2 + 0.25. \end{aligned}$$

Proposition 4.1. *For any linear model $f : \mathbb{R}^d \to \mathbb{R}^m$ and sufficient statistic $u \in \mathbb{N}_0^m$, the log-likelihood function*

$$\ell(\theta) = \sum_{i=1}^{m} u_i \log f_i(\theta)$$

is concave. If the linear map f is one-to-one and the data u_i, $1 \leq i \leq m$, are positive, then the log-likelihood function $\ell(\theta)$ is strictly concave.

Proof. Consider the Hessian matrix of the log-likelihood function,

$$H = \left(\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k} \right)_{j,k}.$$

The log-likelihood function $\ell(\theta)$ is concave if and only if the Hessian matrix H is negative semi-definite for each $\theta \in \Theta$. Since the Hessian matrix is real-valued and symmetric, its eigenvalues are real-valued. It follows that the Hessian matrix H is negative semi-definite if and only if the eigenvectors of H are non-positive.

Taking partial derivatives of the coordinate functions gives

$$\frac{\partial f_i}{\partial \theta_j} = a_{ij}, \quad 1 \leq i \leq m, 1 \leq j \leq d.$$

Thus the partial derivatives of the log-likelihood function are

$$\frac{\partial \ell}{\partial \theta_j} = \sum_{i=1}^{m} \frac{u_i a_{ij}}{f_i(\theta)}, \quad 1 \leq j \leq d.$$

Taking partial derivatives again yields

$$\frac{\partial \ell}{\partial \theta_j \partial \theta_k} = -\sum_{i=1}^{m} \frac{u_i a_{ij} a_{ik}}{f_i(\theta)^2}, \quad 1 \leq j, k \leq d.$$

Thus the Hessian matrix is given by the matrix product

$$H = -A^T \cdot \operatorname{diag}\left(\frac{u_1}{f_1(\theta)^2}, \ldots \frac{u_m}{f_m(\theta)^2} \right) \cdot A,$$

where A is the $m \times d$ matrix with entries a_{ij}. Hence, the eigenvalues of H are non-positive.

If the mapping f is one-to-one, the matrix A has full rank d, and if the data u_i, $1 \leq i \leq d$, are strictly positive, then by (4.18) all eigenvalues of the Hessian matrix are strictly negative. Hence, the log-likelihood function is strictly concave. □

Maximum likelihood estimates for a linear model are given by the critical points of the log-likelihood function.

Corollary 4.2. *If the linear model $f : \mathbb{R}^d \to \mathbb{R}^n$ is one-to-one and the data u_i, $1 \leq i \leq m$, are positive, then each critical point of the log-likehood function $\ell(\theta)$ is a local maximum.*

Consider the *simple linear regression model* given by n real-valued data points (x_i, y_i) for $1 \leq i \leq n$. Suppose the relation between the coordinates of these data points is described by the linear expressions

$$y_i = \theta_1 x_i + \theta_0 + \epsilon_i, \quad 1 \leq i \leq n,$$

where $\theta_0, \theta_1 \in \mathbb{R}$ and ϵ_i is an $N(0, \sigma^2)$ error. The objective is to find the equation of the straight line

$$y = \theta_1 x + \theta_0$$

which provides the best fit for the data points in the sense of least-squares minimization; i.e., $\sum_i \epsilon_i^2$ is minimal. The ordinary least-squares method gives

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2}$$

and

$$\hat{\theta}_0 = \overline{y} - \hat{\theta}_1 \overline{x}.$$

Example 4.3 (R). The computation of the statistics of a linear model in `R` can be accomplished by the function `lm` (Fig. 4.1)

```
# relation between age and specific blood value of 20 persons
> age <- c(46,20,52,30,57,25,28,36,22,43,57,33,22,63,40,48,28,49,52,58)
> bv <- c(3.5,1.9,4.0,2.6,4.5,3.0,2.9,3.8,2.1,3.8,4.1,3.0,2.5,4.6,3.2,4.2,
+ 2.3,4.0,2.3,4.0,4.3,3.9)
> srm <- lm( bv ~ age )  # linear model
> summary( srm )

Call:
lm(formula = bv ~ age)

Residuals:
     Min       1Q   Median       3Q      Max
-0.48979 -0.22844 -0.02445  0.20009  0.63844

Coefficients:
            Estimate  Std. Error  t value   P(>|t|)
(Intercept)  1.15174     0.22257    5.175    6.37e-05 ***
age          0.05583     0.00522   10.695    3.15e-09 ***
---
...
Multiple R-squared: 0.864,
...
```

The parameter estimates are $\hat{\theta}_0 = 0.56$ (intercept) and $\hat{\theta}_1 = 0.15$ (age). The standard errors of the parameter estimates are $\mathrm{se}(\hat{\beta}_0) = 0.223$ and $\mathrm{se}(\hat{\beta}_1) = 0.005$. The R^2 value of 0.86 indicates that about 86% of the variance of the blood values can be explained by the model. ♢

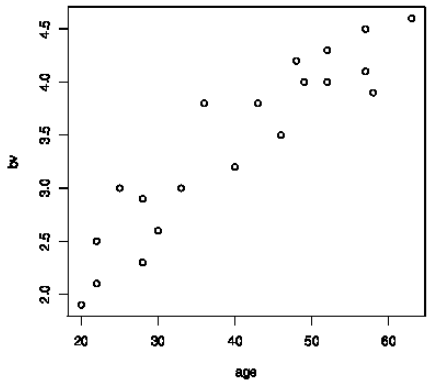


Fig. 4.1. Simple linear regression.

4.4 Toric Models

The toric models form another class of simple algebraic statistical models. To define a toric model, take a matrix $A = (\alpha_{ij}) \in \mathbb{N}_0^{d\times m}$ whose column sums are all equal:

$$\sum_{i=1}^{d} a_{i1} = \ldots = \sum_{i=1}^{d} a_{im}. \tag{4.18}$$

Let the jth column vector α_j of the matrix A represent the monomial

$$\theta^{\alpha_j} = \theta_1^{\alpha_{1j}} \cdots \theta_d^{\alpha_{dj}}, \quad 1 \le j \le m. \tag{4.19}$$

By (4.18), these monomials all have the same degree. The matrix A provides an algebraic statistical model $f : \mathbb{R}^d \to \mathbb{R}^m$ defined as

$$f : \theta \mapsto (\theta^{\alpha_1}, \ldots, \theta^{\alpha_m}), \tag{4.20}$$

which is called the *toric model* associated with A. The parameter space of this toric model is given by

$$\Theta = \{\theta \in \mathbb{R}^d \mid \theta^{\alpha_j} > 0, \sum_j \theta^{\alpha_j} = 1\}. \tag{4.21}$$

Given a frequency vector $u \in \mathbb{N}_0^m$ with sample length $N = \sum_i u_i$. The maximum likelihood function of this model has the form

$$\begin{aligned} L(\theta) &= f_1(\theta)^{u_1} \cdots f_m(\theta)^{u_m} \\ &= (\theta^{\alpha_1})^{u_1} \cdots (\theta^{\alpha_m})^{u_m} \end{aligned}$$

$$
\begin{aligned}
&= \left(\prod_{i=1}^{d} \theta_i^{\alpha_{i1}u_1}\right) \cdots \left(\prod_{i=1}^{d} \theta_i^{\alpha_{im}u_m}\right) \\
&= \prod_{i=1}^{d} \theta_i^{\alpha_{i1}u_1+\alpha_{i2}u_2+\ldots+\alpha_{im}u_m} \\
&= \theta^{Au}.
\end{aligned} \tag{4.22}
$$

The vector $b = Au$ is a sufficient statistic for the model. Maximum likelihood estimation for the toric model means solving the optimization problem

$$
\begin{aligned}
&\max \theta^b \\
&\text{s.t. } \theta \in \Theta.
\end{aligned} \tag{4.23}
$$

Proposition 4.4. *Let $f : \mathbb{R}^d \to \mathbb{R}^n$ be the toric model associated with a matrix $A \in \mathbb{N}_0^{d\times n}$ and let $u \in \mathbb{N}_0^m$ be a frequency vector. If $\hat{\theta}$ is a local maximum of the optimization problem (4.23), then*

$$
A \cdot \hat{p} = \frac{1}{N} \cdot b,
$$

where $b = Au$ is the sufficient statistic, $N = u_1 + \ldots + u_m$ is the sample size, and $\hat{p} = f(\hat{\theta})$.

Proof. We introduce a Lagrange multiplier λ. Each local optimum of (4.23) is a critical point of the following function in the variable $\theta_1, \ldots, \theta_d$ and λ,

$$
\theta^b + \lambda \cdot \left(1 - \sum_{j=1}^{m} \theta^{\alpha_j}\right).
$$

If this function is subjected to the (scaled) gradient operator

$$
\theta \cdot \nabla_\theta = \left(\theta_1 \frac{\partial}{\partial \theta_1}, \ldots, \theta_d \frac{\partial}{\partial \theta_d}\right)^T,
$$

we obtain the expression

$$
\begin{pmatrix} b_1 \cdot \theta_1^{b_1} \\ \vdots \\ b_d \cdot \theta_d^{b_d} \end{pmatrix} - \lambda \cdot \sum_{j=1}^{m} \begin{pmatrix} \alpha_{1j} \\ \vdots \\ \alpha_{dj} \end{pmatrix} \theta^{\alpha_j}.
$$

We abbreviate the left vector by the expression $\theta^b \cdot b$. If we put $p = (\theta^{\alpha_1}, \ldots, \theta^{\alpha_m})^T$, the critical equation obtained by equating (4.24) to zero becomes

$$
\theta^b \cdot b = \lambda \cdot \sum_{j=1}^{m} \theta^{\alpha_j} \cdot \alpha_j = \lambda \cdot A \cdot p.
$$

For each critical point $\hat{\theta}$ with $\hat{p} = (\hat{\theta}^{\alpha_1}, \ldots, \hat{\theta}^{\alpha_m})^T$, we obtain

$$(\hat{\theta})^b \cdot b = \lambda \cdot A \cdot \hat{p}.$$

Thus the vector $A \cdot \hat{p}$ is a scalar multiple of the vector $b = A \cdot u$. But the matrix A has the all-one vector $(1, \ldots, 1)$ in its row space and $\sum_j \hat{p}_j = 1$. Hence the scalar factor must be $1/N$. □

Example 4.5 (Maple). Take the matrix

$$A = \begin{pmatrix} 2\,1\,0 \\ 0\,1\,2 \end{pmatrix}.$$

The associated toric model is given as

$$f : \mathbb{R}^2 \to \mathbb{R}^3 : (\theta_1, \theta_2) \mapsto (\theta_1^2, \theta_1\theta_2, \theta_2^2).$$

Suppose $u = (11, 17, 23)$ is a frequency vector. The sample size is $N = 51$ and we have

$$b = Au = \begin{pmatrix} 39 \\ 63 \end{pmatrix}.$$

The problem is to maximize the likelihood function $L(\theta) = \theta_1^{39}\theta_2^{63}$ over all positive real vectors (θ_1, θ_2) that satisfy $\theta_1^2 + \theta_1\theta_2 + \theta_2^2 = 1$. For this, by Prop. 4.4, we consider the system of equations

$$\begin{pmatrix} 2\hat{\theta}_1^2 + \hat{\theta}_1\hat{\theta}_2 \\ \hat{\theta}_1\hat{\theta}_2 + 2\hat{\theta}_2^2 \end{pmatrix} = \frac{1}{51} \cdot \begin{pmatrix} 39 \\ 63 \end{pmatrix}. \tag{4.24}$$

We use `Maple` to solve these equations. To this end, the toric model is described by the matrix

```
> with(linalg}:
> d := 2: m := 3:
> A := matrix( d, m, [2,1,0,0,1,2] );
```

Take the frequency vector

```
> u := vector( [11,17,23] );
```

Using the matrix A and the vector u, we obtain

```
> N := 0: for j from 1 to m do N := N + u[j] od:
> b := scalarmul( multiply( A, u), 1/N);
> p := vector( [ x^A[1,1] * y^A[2,1],
                 x^A[1,2] * y^A[2,2],
                 x^A[1,3] * y^A[2,3] ]);
> v := multiply( A, p);
```

We solve (4.24) by the floating-point solver `fsolve`

```
> fsolve( {v[1] = b[1], v[2] = b[2]}, {x,y}, {x=0..1}, {y=0..1} );
```

and obtain the unique solution

$$\hat{\theta}_1 = 0.4718898804,\ \hat{\theta}_2 = 0.6767378939.$$

The corresponding probability distribution satisfies

$$\hat{p} = (0.2226800592, 0.3193457638, 0.4579741770),$$

which lies close to the empirical distribution

$$\frac{1}{N} \cdot u = (0.2156862745, 0.3333333333, 0.4509803922).$$

♢

A particularly simple toric model is the socalled *independence model*. To describe an independence model for two random variables, we take positive integers m_1 and m_2, and put $d = m_1 + m_2$ and $m = m_1 \cdot m_2$. Consider the $d \times m$ matrix

$$A = \begin{pmatrix} 1 \ldots 1 & & & & \\ & 1 \ldots 1 & & & \\ & & \ldots & & \\ & & & 1 \ldots 1 & \\ 1 & & 1 & & 1 \\ & \ddots & & \ddots & \ldots & \ddots \\ & & 1 & & 1 & & 1 \end{pmatrix}, \tag{4.25}$$

whose first m_1 rows successively contain the all-one vector of length m_2 and whose last m_2 rows are comprised of successive $m_2 \times m_2$ identity matrices. The column sums in this matrix are all equal to 2. Thus the matrix A defines a toric model $f : \mathbb{R}^d \to \mathbb{R}^m$ given by

$$f : (\theta_1, \ldots, \theta_d) \mapsto (\theta_i \theta_{j+m_1})_{i \in [m_1], j \in [m_2]}. \tag{4.26}$$

The probability distribution $p = (p_{ij})$ has the form

$$p_{ij} = \theta_i \theta_{j+m_1}, \quad 1 \leq i \leq m_1,\ 1 \leq j \leq m_2, \tag{4.27}$$

and satisfies

$$A \cdot p = \begin{pmatrix} \theta_1\theta_{m_1+1} + \ldots + \theta_1\theta_{m_1+m_2} \\ \vdots \\ \theta_{m_1}\theta_{m_1+1} + \ldots + \theta_{m_1}\theta_{m_1+m_2} \\ \theta_1\theta_{m_1+1} + \ldots + \theta_{m_1}\theta_{m_1+1} \\ \vdots \\ \theta_1\theta_{m_1+m_2} + \ldots + \theta_{m_1}\theta_{m_1+m_2} \end{pmatrix}. \tag{4.28}$$

This model can be considered as the independence model for two random variables. To this end, let X_1 and X_2 be random variables on the state sets $[m_1]$ and $[m_2]$, respectively. These random variables are *independent* if

$$p_{ij} = P(X_1 = i \wedge X_2 = j) = P(X_1 = i) \cdot P(X_2 = j), \quad 1 \leq i \leq m_1,\ 1 \leq j \leq m_2. \tag{4.29}$$

By putting $P(X_1 = i) = \theta_i$ and $P(X_2 = j) = \theta_{j+m_1}$, we see the analogy to the algebraic statistical model.

Given a frequency vector $u = (u_{ij}) \in \mathbb{N}_0^m$ with sample length $N = \sum_{ij} u_{ij}$. The sufficient statistic of this model is

$$b = A \cdot u = \begin{pmatrix} u_{1,1} + \ldots + u_{1,m_2} \\ \vdots \\ u_{m_1,1} + \ldots + u_{m_1,m_2} \\ u_{1,1} + \ldots + u_{m_1,1} \\ \vdots \\ u_{1,m_2} + \ldots + u_{m_1,m_2} \end{pmatrix}. \tag{4.30}$$

By Prop. 4.4, we obtain the following result.

Proposition 4.6. *Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be the independence model associated with the matrix $A \in \mathbb{N}_0^{d \times m}$ in (4.25) and let $u = (u_{ij}) \in \mathbb{N}_0^m$ be a frequency vector. A local maximum $\hat{\theta}$ for these data is given by*

$$\hat{\theta}_i = \frac{1}{N} \sum_{j=1}^{m_2} u_{ij}, \quad 1 \leq i \leq m_1,$$

and

$$\hat{\theta}_{j+m_1} = \frac{1}{N} \sum_{i=1}^{m_1} u_{ij}, \quad 1 \leq j \leq m_2.$$

Example 4.7. Consider the independence model for a binary and ternary random variable ($m_1 = 2$ and $m_2 = 3$) given by the matrix

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

The matrix A gives rise to the toric model $f : \mathbb{R}^5 \to \mathbb{R}^{2 \times 3}$ defined as

$$(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) \mapsto (\theta_1\theta_3, \theta_1\theta_4, \theta_1\theta_5, \theta_2\theta_3, \theta_2\theta_4, \theta_1\theta_5).$$

Thus we obtain

$$A \cdot p = \begin{pmatrix} \theta_1\theta_3 + \theta_1\theta_4 + \theta_1\theta_5 \\ \theta_2\theta_3 + \theta_2\theta_4 + \theta_2\theta_5 \\ \theta_1\theta_3 + \theta_2\theta_3 \\ \theta_1\theta_4 + \theta_2\theta_4 \\ \theta_1\theta_5 + \theta_2\theta_5 \end{pmatrix}.$$

Given a frequency vector $u = (u_{11}, u_{12}, u_{13}, u_{21}, u_{22}, u_{23})$, we derive the sufficient statistic

$$b = A \cdot u = \begin{pmatrix} u_{11}+u_{12}+u_{13} \\ u_{21}+u_{22}+u_{23} \\ u_{11}+u_{21} \\ u_{12}+u_{22} \\ u_{13}+u_{23} \end{pmatrix},$$

and the likelihood function

$$\begin{aligned} L(\theta) &= (\theta_1\theta_3)^{u_{11}}(\theta_1\theta_4)^{u_{12}}(\theta_1\theta_5)^{u_{13}}(\theta_2\theta_3)^{u_{21}}(\theta_2\theta_4)^{u_{24}}(\theta_2\theta_5)^{u_{23}} \\ &= \theta_1^{u_{11}+u_{12}+u_{13}}\theta_2^{u_{21}+u_{22}+u_{23}}\theta_3^{u_{11}+u_{21}}\theta_4^{u_{12}+u_{22}}\theta_5^{u_{13}+u_{23}}. \end{aligned}$$

By Prop. 4.6, the maximum likelihood estimates for the frequency vector u are

$$\begin{aligned} \hat{\theta}_1 &= \frac{1}{N}(u_{11}+u_{12}+u_{13}), \\ \hat{\theta}_2 &= \frac{1}{N}(u_{21}+u_{22}+u_{23}), \\ \hat{\theta}_3 &= \frac{1}{N}(u_{11}+u_{21}), \\ \hat{\theta}_4 &= \frac{1}{N}(u_{12}+u_{22}), \\ \hat{\theta}_5 &= \frac{1}{N}(u_{13}+u_{23}). \end{aligned}$$

♢

4.5 Markov Chain Model

A basic algebraic statistical model that is more complex than the linear and toric models is the Markov chain model.

First, we introduce the toric Markov chain model. For this, we take an alphabet Σ with l symbols and fix a positive integer n. We consider words $\tau = \tau_1 \dots \tau_n$ of length n over Σ and count the number of occurrences in τ of length-2 words $\sigma = \sigma_1\sigma_2$. The number of such occurrences is denoted by $a_{\sigma,\tau}$. For instance, we have $a_{\mathtt{CG},\mathtt{ACGACG}} = 2$ and $a_{\mathtt{CC},\mathtt{CGACG}} = 0$.

We record all possible occurences by a matrix $A_{l,n} = (a_{\sigma,\tau})$. Note that the matrix A has $d = l^2$ rows labelled by the length-2 words σ over Σ and $m = l^n$ columns labelled by the length-n words τ over Σ. The matrix $A_{l,n}$ has the property that the sum of each of its columns is $n-1$, because each word of length n consists of $n-1$ consecutive length-2 words. Thus the matrix $A_{l,n}$ defines a toric model $f = f_{l,n} : \mathbb{R}^d \to \mathbb{R}^m$ given by

$$\theta = (\theta_\sigma)_{\sigma\in\Sigma^2} \mapsto (p_\tau)_{\tau\in\Sigma^n}, \tag{4.31}$$

where

$$p_\tau = \frac{1}{l}\theta_{\tau_1\tau_2} \cdot \theta_{\tau_2\tau_3} \cdots \theta_{\tau_{n-1}\tau_n}, \quad \tau = \tau_1 \dots \tau_n \in \Sigma^n. \tag{4.32}$$

The leading coefficient indicates that we assume a uniform initial distribution on the states in the alphabet Σ as described by (7.1). The parameter space of the model is the set of positive $l \times l$ matrices $\Theta = \mathbb{R}_{>0}^{l \times l}$ and the state space is the set of all words over Σ of length n. This model is called *toric Markov chain model*.

Example 4.8. Take the binary alphabet $\Sigma = \{0, 1\}$ and $n = 4$. We have $l = 2$, $d = 2^2 = 4$, and $m = 2^4 = 16$, and the 4×16 matrix $A_{2,4}$ is defined as

$$\begin{array}{c} \\ 00 \\ 01 \\ 10 \\ 11 \end{array} \begin{array}{c} \begin{array}{cccccccccccccccc} 0000 & 0001 & 0010 & 0011 & 0100 & 0101 & 0110 & 0111 & 1000 & 1001 & 1010 & 1011 & 1100 & 1101 & 1110 & 1111 \end{array} \\ \left(\begin{array}{cccccccccccccccc} 3 & 2 & 1 & 1 & 1 & 0 & 0 & 0 & 2 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 2 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 2 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 2 & 0 & 0 & 0 & 1 & 1 & 1 & 2 & 3 \end{array} \right). \end{array}$$

The matrix $A_{2,4}$ provides the toric Markov chain model given by the mapping

$$f_{2,4} : \mathbb{R}^4 \mapsto \mathbb{R}^{16} : (\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11}) \mapsto (p_{0000}, p_{0001}, \ldots, p_{1111}),$$

where

$$p_{\tau_1\tau_2\tau_3\tau_4} = \frac{1}{2}\theta_{\tau_1\tau_2} \cdot \theta_{\tau_2\tau_3} \cdot \theta_{\tau_3\tau_4}, \quad \tau_1, \tau_2, \tau_3, \tau_4 \in \Sigma.$$

For instance, we have $p_{0000} = \frac{1}{2}\theta_{00}^3$, $p_{0001} = \frac{1}{2}\theta_{00}^2\theta_{01}$, and $p_{0110} = \frac{1}{2}\theta_{01}\theta_{11}\theta_{10}$. ♢

Second, we introduce the Markov chain model as a submodel of the toric Markov chain model. For this, the parameter space of the toric Markov chain model is restricted to the set all matrices $\theta \in \mathbb{R}_{>0}^{l \times l}$ whose rows sum up to 1. The parameter space of the Markov chain model is thus a subset Θ_1 of $\mathbb{R}_{>0}^{l \times l}$, and the number of parameters is $d = l \cdot (l - 1)$. The entries of the matrices $\theta \in \Theta_1$ can be viewed as transition probabilities. That is, $\theta_{\sigma_1\sigma_2}$ can be interpreted as the probability to transit from state σ_1 to state σ_2 in one step. The *Markov chain model* is given by the map $f_{l,n} : \mathbb{R}^d \to \mathbb{R}^{l^n}$ restricted to the parameter space Θ_1. Each point p in the image $f_{l,n}(\Theta_1)$ is called a *Markov chain*.

Example 4.9. Reconsider the toric Markov chain model in Ex. 4.8. The parameter space Θ_1 of the Markov chain model can be viewed as the set of all pairs $(\theta_0, \theta_1) \in \mathbb{R}_{>0}^2$ which give rise to the probability matrices

$$\theta = \begin{pmatrix} \theta_0 & 1 - \theta_0 \\ 1 - \theta_1 & \theta_1 \end{pmatrix}. \tag{4.33}$$

The Markov chains in $f_{2,4}(\Theta_1)$ are as follows:

$$\begin{array}{ll} p_{0000} = \frac{1}{2}\theta_0^3, & p_{0001} = \frac{1}{2}\theta_0^2(1 - \theta_0), \\ p_{0010} = \frac{1}{2}\theta_0(1 - \theta_0)(1 - \theta_1), & p_{0011} = \frac{1}{2}\theta_0(1 - \theta_0)\theta_1, \\ p_{0100} = \frac{1}{2}(1 - \theta_1)\theta_0^2, & p_{0101} = \frac{1}{2}(1 - \theta_0)^2(1 - \theta_1), \\ p_{0110} = \frac{1}{2}(1 - \theta_0)\theta_1(1 - \theta_1), & p_{0111} = \frac{1}{2}(1 - \theta_0)\theta_1^2, \\ p_{1000} = \frac{1}{2}(1 - \theta_1)\theta_0^2, & p_{1001} = \frac{1}{2}\theta_0(1 - \theta_0)(1 - \theta_1), \\ p_{1010} = \frac{1}{2}(1 - \theta_1)^2(1 - \theta_0), & p_{1011} = \frac{1}{2}(1 - \theta_0)\theta_1(1 - \theta_1), \\ p_{1100} = \frac{1}{2}\theta_1(1 - \theta_1)\theta_0, & p_{1101} = \frac{1}{2}(1 - \theta_0)\theta_1(1 - \theta_1), \\ p_{1110} = \frac{1}{2}\theta_1^2(1 - \theta_1), & p_{1111} = \frac{1}{2}\theta_1^3. \end{array}$$

♢

Let $u = (u_\tau) \in \mathbb{N}_0^m$ be a frequency vector representing N observed sequences in Σ^n. That is, $u_\tau = u_{\tau_1 \dots \tau_n}$ counts the number of times the sequence $\tau = \tau_1 \dots \tau_n$ was observed. Hence, $\sum_\tau u_\tau = N$. The sufficient statistic $v = A_{l,n} \cdot u$ can be regarded as an $l \times l$ matrix with entries v_{σ_1,σ_2}, where $\sigma_1, \sigma_2 \in \Sigma$. The entry $v_{\sigma_1\sigma_2}$ equals the number of occurrences of $\sigma_1\sigma_2 \in \Sigma^2$ as a consecutive pair in any of the N observed sequences.

Example 4.10. Reconsider the Markov chain model in Ex. 4.9. The sufficient statistic is given as

$$\begin{aligned}
v_{00} &= 3u_{0000} + 2u_{0001} + u_{0010} + u_{0011} + u_{0100} + 2u_{1000} + u_{1001} + u_{1100},\\
v_{01} &= u_{0001} + u_{0010} + u_{0011} + u_{0100} + 2u_{0101} + u_{0110} + u_{0111} + u_{1001} + u_{1010} + u_{1011} + u_{1101},\\
v_{10} &= u_{0010} + u_{0100} + u_{0101} + u_{0110} + 2u_{1000} + u_{1001} + 2u_{1010} + u_{1011} + u_{1100} + u_{1101} + u_{1110},\\
v_{11} &= u_{0011} + u_{0110} + 2u_{0111} + u_{1011} + u_{1100} + u_{1101} + 2u_{1110} + 3u_{1111}.
\end{aligned}$$

♢

Proposition 4.11. *In the Markov chain model $f_{l,n}$, the maximum likelihood estimate of the frequency data $u \in \mathbb{N}_0^{l^n}$ with sufficient statistic $v = A_{l,n} \cdot u$ is given by the $l \times l$ matrix $\hat{\theta} = (\hat{\theta}_{\sigma_1\sigma_1})$ in Θ_1 such that*

$$\hat{\theta}_{\sigma_1\sigma_2} = \frac{v_{\sigma_1\sigma_2}}{\sum_{\sigma\in\Sigma} v_{\sigma_1\sigma}}, \quad \sigma_1, \sigma_2 \in \Sigma.$$

Proof. Let $\Sigma = \{1, \dots, l\}$. The likelihood function for the toric Markov chain model is given by

$$L(\theta) = \theta^{A_{l,n}\cdot u} = \theta^v = \prod_{ij\in\Sigma^2} \theta_{ij}^{v_{ij}}$$

and so the log-likelihood function equals

$$\begin{aligned}
\ell(\theta) &= \sum_{ij\in\Sigma^2} v_{ij} \log \theta_{ij}\\
&= \sum_{i\in\Sigma} \left(v_{i1} \log \theta_{i1} + \ldots + v_{i,l-1} \log \theta_{i,l-1} + v_{il} \log \left(1 - \sum_{k=1}^{l-1} \theta_{ik} \right) \right).
\end{aligned}$$

For any length-2 word $ij \in \Sigma^2$, we obtain

$$\frac{\partial \ell}{\partial \theta_{ij}} = \frac{v_{ij}}{\theta_{ij}} - \frac{v_{il}}{1 - \sum_{k=1}^{l-1} \theta_{ik}}.$$

Equating these expressions to zero yields the unique critial point with coordinates

$$\theta_{ij} = \frac{v_{ij}}{v_{i1} + \ldots + v_{il}}, \quad ij \in \Sigma^2.$$

□

Example 4.12. Reconsider the Markov chain model in Ex. 4.9. Suppose there is a sample of length $N = 89$ given by the frequency vector

$$u = (7, 2, 8, 10, 7, 9, 7, 10, 4, 2, 5, 7, 4, 3, 2, 4)^T.$$

Then the sufficient statistic is $v = A_{2,4} \cdot u = (64, 79, 63, 67)^T$. The likelihood function is given by

$$L(\theta) = \theta_0^{64} \cdot (1-\theta_0)^{79} \cdot \theta_1^{63} \cdot (1-\theta_1)^{67}$$

and thus the log-likelihood function equals

$$\ell(\theta) = 64 \cdot \log \theta_0 + 79 \cdot \log(1-\theta_0) + 63 \cdot \log \theta_1 + 67 \cdot \log(1-\theta_1).$$

By Prop. 4.11, the maximum likelihood estimate of the data u is

$$\hat{\theta}_0 = \frac{64}{64+79} = 0.447552 \quad \text{and} \quad \hat{\theta}_1 = \frac{63}{63+67} = 0.484615.$$

♢

4.6 Maximum Likelihood Estimation

Maximum likelihood estimation is a popular statistical method used for fitting a statistical model to the data and providing estimates for the model parameters. Consider an algebraic statistical model $f : \mathbb{C}^d \to \mathbb{C}^m$ given by

$$f : \theta \mapsto (f_1(\theta), \ldots, f_m(\theta)). \tag{4.34}$$

Here the ambient spaces are taken over the complex numbers, but the coordinates $f_1, \ldots, f_m$ are polynomials in $\mathbb{Q}[\theta_1, \ldots, \theta_d]$. It is assumed that the parameter space Θ is an open subset of $\mathbb{R}^d$ and that the image $f(\Theta)$ of the parameter space is a subset of $\mathbb{R}^m_{>0}$.

Given a sample set which is summarized by the frequency vector $u = (u_1, u_2, \ldots, u_m)$ of positive integers. The probability of observing the data is given by the likelihood function

$$L_u(\theta) = f_{i_1}(\theta) f_{i_2}(\theta) \cdots f_{i_N}(\theta) = f_1(\theta)^{u_1} \cdots f_{i_m}(\theta)^{u_m}. \tag{4.35}$$

Equivalently, the likelihood function can be described by the log-likelihood function

$$\ell_u(\theta) = \log L_u(\theta) = u_1 \cdot \log f_1(\theta) + \ldots + u_m \cdot \log f_m(\theta). \tag{4.36}$$

The problem of maximum likelihood estimation is to maximize the (log) likelihood function. Each maximum of the log-likelihood function is a solution of the critical equations

$$\frac{\partial \ell_u}{\partial \theta_i} = 0, \qquad 1 \le i \le d. \tag{4.37}$$

The derivative of ℓ_u with respect to the variable θ_i is the rational function

$$\frac{\partial \ell_u}{\partial \theta_i} = \sum_{j=1}^{m} \frac{u_j}{f_j(\theta)} \frac{\partial f_j}{\partial \theta_i}, \quad 1 \le i \le d. \tag{4.38}$$

We can use Groebner bases to compute the critical points. For this, consider the polynomial ring $\mathbb{Q}[Z_1,\ldots,Z_m,\theta_1,\ldots,\theta_d]$ and take the ideal

$$J_u = \langle Z_1 f_1 - 1,\ldots,Z_m f_m - 1, \sum_{j=1}^{m} u_j Z_j \frac{\partial f_j}{\partial \theta_1},\ldots,\sum_{j=1}^{m} u_j Z_j \frac{\partial f_j}{\partial \theta_d}\rangle. \tag{4.39}$$

A point $(z,\theta) \in \mathbb{C}^{d+m}$ lies in the affine variety $\mathcal{V}(J_u)$ if and only if θ is a critical point of the log-likelihood function, where $f_j(\theta) \neq 0$ and $z_j = 1/f_j(\theta)$ for $1 \leq j \leq m$.

Consider the m-th elimination ideal of J_u with respect to an elimination ordering for $Z_1,\ldots,Z_m$; that is,

$$I_u = J_u \cap \mathbb{C}[\theta_1,\ldots,\theta_d]. \tag{4.40}$$

The ideal I_u is called the *likelihood ideal* and the variety $\mathcal{V}(I_u)$ is called the *likelihood variety* of the model f with respect to the data u.

Proposition 4.13. *A point $\theta \in \mathbb{C}^d$ with $f_j(\theta) \neq 0$ for $1 \leq j \leq m$ lies in the likelihood variety $\mathcal{V}(I_u)$ if and only if θ is a critical point of the log-likelihood function ℓ_u.*

Proof. Let θ be a critical point of ℓ_u. Put $z_j = 1/f_j(\theta)$, $1 \leq j \leq m$. Then $(z,\theta) \in \mathcal{V}(J_u)$ and so by the Closure theorem $\theta = \pi_m(z,\theta) \in \mathcal{V}(I_u)$.

Conversely, we make use the Extension theorem. First, we extend by the variable Z_1. Consider the generator $Z_1 f_1 - 1$. Since $f_1(\theta) \neq 0$ and $\theta \in \mathcal{V}(I_u)$, it follows that the solution θ can be extended to a solution (θ, z_1), $z_1 = 1/f_1(\theta)$, of the ideal $\langle Z_1 f_1 - 1\rangle + I_u$. By continuing this way, we obtain an element (θ, z) of $\mathcal{V}(J_u)$. Then θ is a critical point of ℓ_u. □

The problem of maximum likelihood estimation can be solved by computing the likelihood variety $\mathcal{V}(I_u)$ in $\mathbb{C}^d$, intersecting the variety with the preimage $f^{-1}(\Delta)$ of the probability simplex Δ_{m-1}, and identifying all local maxima among the points in $\mathcal{V}(I_u) \cap f^{-1}(\Delta)$. Equivalently, the maximum likelihood estimates can be obtained by augmenting the ideal J_u with the polynomial $f_1 + \ldots + f_m - 1$; that is,

$$J_u = \langle f_1 + f_2 + f_3 + f_4 + f_5 - 1, z_1 f_1 - 1,\ldots,z_m f_m - 1, \sum_{j=1}^{m} u_j z_j \frac{\partial f_j}{\partial \theta_1},\ldots,\sum_{j=1}^{m} u_j z_j \frac{\partial f_j}{\partial \theta_d}\rangle. \tag{4.41}$$

Then the likelihood variety $\mathcal{V}(I_u)$ is intersected with the preimage $f^{-1}(\mathbb{R}^m_{>0})$ and all local maxima among the points in $\mathcal{V}(I_u) \cap f^{-1}(\mathbb{R}^n_{>0})$ are determined.

Example 4.14 (Singular). We compute the likelihood variety of the DiaNA model. For this, take the algebraic statistical model $f : \mathbb{C}^2 \to \mathbb{C}^4$ given by

```
> ring bigring = real, (t(1..2),z(1..4), lp;
> poly f1 = -0.10*t(1) + 0.02*t(2) + 0.25;
> poly f2 =  0.08*t(1) - 0.01*t(2) + 0.25;
> poly f3 =  0.11*t(1) - 0.02*t(2) + 0.25;
> poly f4 = -0.09*t(1) + 0.01*t(2) + 0.25;
```

Suppose the frequency vector is

```
> int u1 = 10;
> int u2 = 14;
> int u3 = 15;
> int u4 = 10;
```

The ideal J_u in the big ring $\mathbb{Q}[Z_1, Z_2, Z_3, Z_4, \theta_1, \theta_2]$ is defined as

```
> ideal Ju = f1+f2+f2+f4-1,
z(1)*f1-1, z(2)*f2-1, z(3)*f3-1, z(4)*f4-1,
u1*z(1)*diff(f1,t(1)) + u2*z(2)*diff(f2,t(1))
+ u3*z(3)*diff(f3,t(1)) + u4*z(4)*diff(f4,t(1)),
u1*z(2)*diff(f1,t(2)) + u2*z(2)*diff(f2,t(2))
+ u3*z(3)*diff(f3,t(2)) + u4*z(4)*diff(f4,t(2));
```

The likelihood ideal I_u is obtained from J_u by elimination:

```
> ideal Iu = eliminate (Ju, z(1)*z(2)*z(3)*z(4));
> ring smallring = real, (t(1..2)), lp;
> ideal Iu = fetch (bigring, Iu);
> std(Iu);
Iu_[1]=t(2)^3-(8.071e+01)*t(2)^2-(3.202e+04)*t(2)+(6.959e+03)
Iu_[2]=t(1)+(2.110e-04)*t(2)^2-(1.627-01)*t(2)-(4.838-01)
```

Finally, the zeros of the reduced Groebner basis of I_u are computed as follows:

```
> LIB "solve.lib";
> solve (Iu, 10);
```

The zeros are

```
[1]:
[1]:
-27.1481605843
[2]:
-143.2004435005
[2]:
[1]:
0.5191557516
[2]:
0.2174490559
[3]:
[1]:
26.3283769148
[2]:
223.6954677454
```

The second zero is the maximum likelihood estimate (Section 4.1). ♢

4.7 Model Invariants

Each algebraic statistical model gives rise to model invariants that describe the relationships between the probabilities. Consider an algebraic statistical model $f : \mathbb{C}^d \to \mathbb{C}^m$ given by

$$f : (\theta_1, \ldots, \theta_d) \mapsto (f_1(\theta), \ldots, f_m(\theta)). \tag{4.42}$$

Here the ambient spaces are taken over the complex numbers, but the coordinates $f_1, \ldots, f_m$ are polynomials in $\mathbb{Q}[\theta_1, \ldots, \theta_d]$. We study the image $f(\mathbb{C}^d)$ by the polynomial parametrization

$$p_i = f_i(\theta_1, \ldots, \theta_d), \quad 1 \leq i \leq m. \tag{4.43}$$

The Implicitization theorem yields the following result.

Proposition 4.15. *Consider the ideal $I = \langle p_1 - f_1, \ldots, p_m - f_m \rangle$ in $\mathbb{C}[\theta_1, \ldots, \theta_d, p_1, \ldots, p_m]$. For the d-th elimination ideal $I_d = I \cap \mathbb{C}[p_1, \ldots, p_m]$, the affine variety $\mathcal{V}(I_d)$ is the Zariski closure of the image $f(\mathbb{C}^d)$.*

The polynomials in the elimination ideal I_d are called *invariants* of the model f. By the Elimination theorem, these invariants can be established by computing the reduced Groebner basis of the elimination ideal I_d with respect to an elimination ordering for $\theta_1 > \ldots > \theta_d > p_1 > \ldots > p_m$.

Example 4.16 (Singular). Consider the mapping $f : \mathbb{C}^2 \to \mathbb{C}^3 : (\theta_1, \theta_2) \mapsto (\theta_1^3, \theta_1\theta_2, \theta_1\theta_2)$. The image of f is a (dense) subset of a plane in three-space,

$$\begin{aligned} f(\mathbb{C}^2) &= \{(x, y, z) \in \mathbb{C}^3 \mid y = z \wedge (x = 0 \Rightarrow y = 0)\} \\ &= [\mathcal{V}(Y - Z) \setminus \mathcal{V}(X, Y - Z)] \cup \mathcal{V}(X, Y, Z). \end{aligned}$$

This is a Boolean combination of affine varieties, but not an affine variety. In view of the ideal $I = \langle p_1 - \theta_1^3, p_2 - \theta_1\theta_2, p_3 - \theta_1\theta_2 \rangle$ in $\mathbb{C}[\theta_1, \theta_2, p_1, p_2, p_3]$, the reduced Groebner basis with respect to the `lp` ordering with $\theta_1 > \ldots > \theta_d > p_1 > \ldots > p_m$ can be calculated as follows,

```
> ring r = 0, (t(1..2),p(1..3)), lp;
> ideal i = p(1)-t(1), p(2)-t(1)*t(2), p(3)-t(1)*t(2);
> std(i);
_[1]=p(2)-p(3)
_[2]=t(2)*p(1)-p(3)
_[3]=t(1)-p(1)
```

Thus the reduced Groebner basis of the second elimination ideal $I_2 = I \cap \mathbb{Q}[p_1, p_2, p_3]$ is $\{p_2 - p_3\}$ Hence the Zariski closure of the image $f(\mathbb{C}^d)$ is $\mathcal{V}(p_2 - p_3)$ and $p_2 - p_3$ is a model invariant for f. ♢

Example 4.17 (Singular). Reconsider the toric model $f : \mathbb{C}^2 \to \mathbb{C}^3 : (\theta_1, \theta_2) \mapsto (\theta_1^2, \theta_1\theta_2, \theta_2^2)$ studied in Ex. 4.5. Take the ideal $I = (p_1 - \theta_1^2, p_2 - \theta_1\theta_2, p_3 - \theta_2^2)$ in $\mathbb{C}[\theta_1, \theta_2, p_1, p_2, p_3]$ and calculate a Groebner basis of I with respect to the `lp` ordering with $\theta_1 > \theta_2 > p_1 > p_2 > p_3$.

```
> ring r = 0, (t(1..2),p(1..3)), lp;
> ideal i = p(1)-t(1)^2, p(2)-t(1)*t(2), p(3)-t(2)^2;
> std(i);
```

```
_[1]=p(1)*p(3)-p(2)^2
_[2]=t(2)^2-p(3)
_[3]=t(1)*p(3)-p(2)^2
_[4]=t(1)*p(2)-t(2)*p(1)
_[5]=t(1)*t(2)-p(2)
_[6]=t(1)^2-p(1)
```

The first element provides the Groebner basis of the second elimination ideal I_2 and yields the model invariant $p_1p_3 - p_2^2$. ♢

Example 4.18 (Singular). Reconsider the DiaNA model. The polynomial parametrization of the DiaNA model is given in (4.2). Take the ideal I in $\mathbb{C}[\theta_1, \theta_2, p_1, p_2, p_3, p_4]$ generated by

$$\begin{aligned} &p_1 - (-0.10 \cdot \theta_1 + 0.02 \cdot \theta_2 + 0.25), \\ &p_2 - (0.08 \cdot \theta_1 - 0.01 \cdot \theta_2 + 0.25), \\ &p_3 - (0.11 \cdot \theta_1 - 0.02 \cdot \theta_2 + 0.25), \\ &p_4 - (-0.09 \cdot \theta_1 + 0.01 \cdot \theta_2 + 0.25). \end{aligned}$$

A Groebner basis of I with respect to the `lp` ordering $\theta_1 > \theta_2 > p_1 > p_2 > p_3 > p_4$ is

$$\begin{aligned} &-0.5500000002 + 1.399999999 \cdot p_2 - 0.1999999995 \cdot p_3 + 1.0000 \cdot p_4, \\ &-0.05059523811 + 0.2380952383 \cdot p_4 + 0.9523809525 \cdot p_3 + 0.8333333334 \cdot p_1, \\ &-0.4285714291 + 0.9428571435 \cdot p_4 + 0.7714285717 \cdot p_3 + 0.6000000000e - 2 \cdot \theta_2, \\ &-1.071428573 + 2.857142859 \cdot p_4 + 1.428571429 \cdot p3 + 0.100 \cdot \theta_1. \end{aligned}$$

The first two polynomials generate the second elimination ideal I_2 and thus provide invariants of the DiaNA model. ♢

Example 4.19 (Singular). Consider the dishonest casino in Ex. 6.1. Assume that the dealer always starts with the fair coin and then switches eventually to the loaded one. If a game consists of $n = 4$ coin tosses, the probability for an outcome $\tau \in \Sigma'^4$ is

$$p_\tau = p_{FFFF,\tau} + p_{FFFL,\tau} + p_{FFLL,\tau} + p_{FLLL,\tau}.$$

The invariants for this model can be computed as follows.

```
> ring r = 0, (FF,FL,LL,Fh,Ft,Lh,Lt,p(0..15)), dp
> ideal i =
# hhhh
p(0) - Fh*FF*Fh*FF*Fh*FF*Fh - Fh*FF*Fh*FF*Fh*FL*Lh - Fh*FF*Fh*FL*Fh*LL*Lh
     - Fh*FL*Lh*LL*Lh*LL*Lh,
# hhht
p(1) - Fh*FF*Fh*FF*Fh*FF*Ft - Fh*FF*Fh*FF*Fh*FL*Lt - Fh*FF*Fh*FL*Fh*LL*Lt
     - Fh*FL*Lh*LL*Lh*LL*Lt,
# hhth
p(2) - Fh*FF*Fh*FF*Ft*FF*Fh - Fh*FF*Fh*FF*Ft*FL*Lh - Fh*FF*Fh*FL*Ft*LL*Lh
     - Fh*FL*Lh*LL*Lt*LL*Lh,
# hhtt
```

```
p(3) - Fh*FF*Fh*FF*Ft*FF*Ft - Fh*FF*Fh*FF*Ft*FL*Lt - Fh*FF*Fh*FL*Ft*LL*Lt
     - Fh*FL*Lh*LL*Lt*LL*Lt,
# hthh
p(4) - Fh*FF*Ft*FF*Fh*FF*Fh - Fh*FF*Ft*FF*Fh*FL*Lh - Fh*FF*Ft*FL*Fh*LL*Lh
     - Fh*FL*Lt*LL*Lh*LL*Lh,
# htht
p(5) - Fh*FF*Ft*FF*Fh*FF*Ft - Fh*FF*Ft*FF*Fh*FL*Lt - Fh*FF*Ft*FL*Fh*LL*Lt
     - Fh*FL*Lt*LL*Lh*LL*Lt,
# htth
p(6) - Fh*FF*Ft*FF*Ft*FF*Fh - Fh*FF*Ft*FF*Ft*FL*Lh - Fh*FF*Ft*FL*Ft*LL*Lh
     - Fh*FL*Lt*LL*Lt*LL*Lh,
# httt
p(7) - Fh*FF*Ft*FF*Ft*FF*Ft - Fh*FF*Ft*FF*Ft*FL*Lt - Fh*FF*Ft*FL*Ft*LL*Lt
     - Fh*FL*Lt*LL*Lt*LL*Lt,
# thhh
p(8) - Ft*FF*Fh*FF*Fh*FF*Fh - Ft*FF*Fh*FF*Fh*FL*Lh - Ft*FF*Fh*FL*Fh*LL*Lh
     - Ft*FL*Lh*LL*Lh*LL*Lh,
# thht
p(9) - Ft*FF*Fh*FF*Fh*FF*Ft - Ft*FF*Fh*FF*Fh*FL*Lt - Ft*FF*Fh*FL*Fh*LL*Lt
     - Ft*FL*Lh*LL*Lh*LL*Lt,
# thth
p(10)- Ft*FF*Fh*FF*Ft*FF*Fh - Ft*FF*Fh*FF*Ft*FL*Lh - Ft*FF*Fh*FL*Ft*LL*Lh
     - Ft*FL*Lh*LL*Lt*LL*Lh,
# thtt
p(11)- Ft*FF*Fh*FF*Ft*FF*Ft - Ft*FF*Fh*FF*Ft*FL*Lt - Ft*FF*Fh*FL*Ft*LL*Lt
     - Ft*FL*Lh*LL*Lt*LL*Lt,
# tthh
p(12)- Ft*FF*Ft*FF*Fh*FF*Fh - Ft*FF*Ft*FF*Fh*FL*Lh - Ft*FF*Ft*FL*Fh*LL*Lh
     - Ft*FL*Lt*LL*Lh*LL*Lh,
# ttht
p(13)- Ft*FF*Ft*FF*Fh*FF*Ft - Ft*FF*Ft*FF*Fh*FL*Lt - Ft*FF*Ft*FL*Fh*LL*Lt
     - Ft*FL*Lt*LL*Lh*LL*Lt,
# ttth
p(13)- Ft*FF*Ft*FF*Ft*FF*Fh - Ft*FF*Ft*FF*Ft*FL*Lh - Ft*FF*Ft*FL*Ft*LL*Lh
     - Ft*FL*Lt*LL*Lt*LL*Lh,
# tttt
p(15)- Ft*FF*Ft*FF*Ft*FF*Ft - Ft*FF*Ft*FF*Ft*FL*Lt - Ft*FF*Ft*FL*Ft*LL*Lt
     - Ft*FL*Lt*LL*Lt*LL*Lt;
> ideal j = std(i);
> eliminate(j, FF*FL*LL*Fh*Ft*Lh*Lt);
```

The output is a list of 53 generating invariants. ♢

4.8 Statistical Inference

We explain the concept of statistical inference for algebraic statistical models with observed and hidden random variables. In these kind of models, we know content of the observed data but nothing about the content of the hidden data. The task is then to find the most likely set of data of the hidden random variables given the set of data of the observed random variables. This problem is known as *statistical inference problem* and the most likely set of data of the hidden random variables is referred to as *explanation* of the observed data.

Consider an algebraic statistical model with hidden and observed variables such that the probability of the observed sequence τ is given by

$$p_\tau = \sum_\sigma p_{\sigma,\tau}, \tag{4.44}$$

where $p_{\sigma,\tau}$ is the probability of having the data σ at the hidden variables and the data τ at the observed variables. Thus the probability of the observed sequence τ is the marginalization over all possible values of the hidden variables. Finding an explanation of the model means identifying the set of hidden data $\bar{\sigma}$ with maximum a posteriori probability of generating the observed data τ,

$$\bar{\sigma} = \operatorname{argmax}_\sigma\{p_{\sigma,\tau}\}. \tag{4.45}$$

By putting $w_{\sigma,\tau} = -\log(p_{\sigma,\tau})$, the tropicalization of the marginal probability (4.44) yields

$$w_\tau = \bigoplus_\sigma w_{\sigma,\tau}, \tag{4.46}$$

Thus the explanation $\bar{\sigma}$ is given by evaluation in the tropical algebra,

$$\bar{\sigma} = \operatorname{argmin}_\sigma\{w_{\sigma,\tau}\}. \tag{4.47}$$

We generalize the machinery of maximum a posteriori probability estimation by allowing the observed data to include parameters. For this, consider an algebraic statistical model

$$f : \mathbb{R}^d \to \mathbb{R}^m : \theta \mapsto (f_1(\theta), \ldots, f_m(\theta)).$$

Suppose there is an associated density function

$$g(\theta) = \sum_{i=1}^{M} \theta_1^{v_{i1}} \cdots \theta_d^{v_{id}}, \tag{4.48}$$

where $v_i = (v_{i1}, \ldots, v_{id}) \in \mathbb{N}_0^d$, $1 \le i \le M$. For a fixed value of $\theta \in \mathbb{R}_{>0}^d$, the problem is to find a term $\theta_1^{v_{j1}} \cdots \theta_d^{v_{jd}}$, $1 \le j \le M$, in the expression $g(\theta)$ with maximum value

$$j = \operatorname{argmax}_i\{\theta_1^{v_{i1}} \cdots \theta_d^{v_{id}}\}. \tag{4.49}$$

Each such solution is called an *explanation* of the model. By putting $w_i = -\log\theta_i$ and $w = (w_1, \ldots, w_d)$, we obtain

$$-\log(\theta_1^{v_{i1}} \cdots \theta_d^{v_{id}}) = -[v_{i1}\log(\theta_1) + \cdots + v_{id}\log(\theta_d)] = \langle v_j, w\rangle. \tag{4.50}$$

This amounts to finding a vector v_j that minimizes the linear expression

$$\langle v_j, w\rangle = \sum_{i=1}^{d} w_i v_{ji}, \qquad 1 \leq j \leq M. \tag{4.51}$$

This minimization problem is equivalent to the linear programming problem

$$\begin{array}{c} \min\langle x, w\rangle. \\ \text{s.t. } x \in \mathrm{NP}(g) \end{array} \tag{4.52}$$

To see this, observe that the Newton polytope $\mathrm{NP}(g)$ of the polynomial g is the convex hull of the points v_i, $1 \leq i \leq M$, and the vertices of this polytope form a subset of these points. But the minimal value of a linear functional $x \mapsto \langle x, w\rangle$ over a polytope is attained a vertex of the polytope. Thus we have shown the following assertion.

Proposition 4.20. *For a fixed parameter w, the problem of solving the statistical inference problem (4.49) is equivalent to the linear programming problem of minimizing the linear functional $x \mapsto \langle x, w\rangle$ over the Newton polytope* $\mathrm{NP}(g)$.

The parametric version of this problem asks for the set of parameters w for which the vertex v_j gives the explanation. That is, we seek the set of all points w such that the linear functional $x \mapsto \langle x, w\rangle$ attains its minimum at the point v_j. By definition, this set is given by $N_{\mathrm{NP}(g)}(v_j)$, the normal cone of the polytope $\mathrm{NP}(g)$ at the vertex v_j.

Proposition 4.21. *The set of all parameters w for which the vertex v_j provides the explanation in the algebraic statistical model given by the density (4.48) is equal to the normal cone of the polytope* $\mathrm{NP}(g)$ *at the vertex v_j.*

The normal cones associated with the vertices of the Newton polytope $\mathrm{NP}(g)$ are part of the normal fan $\mathcal{N}_{\mathrm{NP}(g)}$ of the Newton polytope $\mathrm{NP}(g)$. The normal fan provides a decomposition of the parameter space into regions, but only the regions corresponding to the vertices of the polytope are relevant for statistical inference. An example of parametric statistical inference is given in the section on parametric sequence alignment.

5

Sequence Alignment

A fundamental task in computational biology is the alignment of DNA or amino acid sequences. The primary objective of biological sequence alignment is to find positions in the sequences that are homologous; that is, the symbols at those positions are derived from the same position in some anchestral sequence. It may be possible due to evolution that the two homologous positions have different states and are located at different positions. An alignment is biologically correct if it matches up all positions that are truly homologous. Unfortunately, the biological truth is in most cases unknown. Therefore, a guess is made by treating sequence alignment as an optimization problem. The corresponding objective function assigns a score to each alignment according to a scoring scheme. Then an alignment is sought that maximizes this score. But which scoring scheme should be used to predict biologically correct alignments? For this, the scoring scheme needs to be analyzed over all possible parameter values.

This chapter introduces an algebraic statistical model for pairwise sequence alignment. The interpretation of their marginal probabilities in the tropical algebra will lead to a formalization of the alignment problem as an optimization problem, and the interpretation in the polytope algebra will allow to analyze scoring schemes over all possible parameter values.

5.1 Sequence Alignment

We take a finite alphabet Σ with l letters and an additional symbol "$-$", denoted as *blank*, and call $\Sigma \cup \{-\}$ the *extended alphabet*. We consider two sequences $\sigma^1 = \sigma^1_1 \dots \sigma^1_m$ and $\sigma^2 = \sigma^2_1 \dots \sigma^2_n$ over the alphabet Σ.

An *alignment* of the sequences σ^1 and σ^2 is a pair of *aligned sequences* (μ^1, μ^2) over the extended alphabet $\Sigma \cup \{-\}$ such that both sequences μ^1 and μ^2 have the same length and are copies of σ^1 and σ^2 with inserted blanks, respectively. An alignment (μ^1, μ^2) does not allow blanks at the same position. It follows that the aligned sequences have length at most $m + n$.

Example 5.1. Consider the sequences $\sigma^1 =$ `ACGTAGC` and $\sigma^2 =$ `ACCGAGACC`. An alignment of these sequences is given by

$$\mu^1 = \texttt{A C} - \texttt{G} - \texttt{T A} - \texttt{G C}$$
$$\mu^2 = \texttt{A C C G A G A C} - \texttt{C}$$

An alignment of maximal length is

```
A C G T A G C − − − − − − − − − −
− − − − − − − A C C G A G A C C
```

♢

An alignment of a pair of sequences (σ^1, σ^2) can also be represented by a string h over the *edit alphabet* $\{H, I, D\}$. The string h is called an *edit string* and the letters of the edit alphabet stand for *homology* (H), *insertion* (I), and *deletion* (D). A letter I stands for an insertion (indel) in the first sequence σ^1, a letter D is a deletion (indel) in the first sequence σ^1, and a letter H is a character change (mutation or mismatch) including the identity change (match). We write $\#H$, $\#I$, and $\#D$ for the respective number of instances of H, I, and D in an edit string for an alignment of the pair (σ^1, σ^2). Then we have

$$\#H + \#D = m \quad \text{and} \quad \#H + \#I = n. \tag{5.1}$$

Example 5.2. Reconsider the sequences $\sigma^1 = \texttt{ACGTAGC}$ and $\sigma^2 = \texttt{ACCGAGACC}$. An alignment of these sequences including the edit string is given by

$$\begin{array}{lllllllllll} h & = H & H & I & H & I & H & H & I & D & H \\ \mu^1 & = \texttt{A} & \texttt{C} & - & \texttt{G} & - & \texttt{T} & \texttt{A} & - & \texttt{G} & \texttt{C} \\ \mu^2 & = \texttt{A} & \texttt{C} & \texttt{C} & \texttt{G} & \texttt{A} & \texttt{G} & \texttt{A} & \texttt{C} & - & \texttt{C} \end{array}$$

We have $\#H = 6$, $\#I = 3$, and $\#D = 1$. ♢

Proposition 5.3. *A string over the edit alphabet $\{H, I, D\}$ represents an alignment of an m-letter sequence σ^1 and an n-letter sequence σ^2 if and only if (5.1) holds.*

Proof. Given an alignment of the pair (σ^1, σ^2). We form an edit string h from left to right. Each symbol in σ^1 either corresponds to a symbol in σ^2, in which case we record an H in the edit string, or it gets deleted, in which case we record a D. This shows that the first equation in (5.1) holds. Each symbol in σ^2 either corresponds to a symbol in σ^1, in which case we already recorded an H in the edit string, or it gets inserted, in which case we record a I. This shows that the second equation in (5.1) holds.

Conversely, each edit string h with the property (5.1), when read from left to right, produces an alignment of the pair (σ^1, σ^2). □

We write $\mathcal{A}_{m,n}$ for the set of all strings over the edit alphabet $\{H, I, D\}$ that satisfy the equations (5.1). We call $\mathcal{A}_{m.n}$ the set of all alignments of the sequences σ^1 and σ^2 in spite of the fact that it only depends on m and n and not on the specific sequences. The cardinality of the set $\mathcal{A}_{m,n}$ is called *Delannoy number* (Fig. 5.2).

Proposition 5.4. *The cardinality of the set $\mathcal{A}_{m,n}$ can be computed as the coefficient of the monomial $x^m y^n$ in the generating function $\frac{1}{1-x-y-xy}$.*

Proof. Consider the expansion of the generating function

$$\frac{1}{1-x-y-xy} = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} a_{m,n} x^m y^n. \tag{5.2}$$

The coefficients are characterized by the linear recurrence

$$a_{m,n} = a_{m-1,n} + a_{m,n-1} + a_{m-1,n-1}, \quad m \geq 0, n \geq 0, m+n \geq 1, \tag{5.3}$$

with initial conditions $a_{0,0} = 1$, $a_{m,-1} = 0$, and $a_{-1,n} = 0$. The same recurrence holds for the cardinality of $\mathcal{A}_{m,n}$. To see this, note that for nonnegative integers m and n with $m+n \geq 1$, each string in $\mathcal{A}_{m,n}$ is either a string in $\mathcal{A}_{m-1,n-1}$ followed by an H, or a string in $\mathcal{A}_{m-1,n}$ followed by a D, or a string in $\mathcal{A}_{m,n-1}$ followed by an I (Fig. 5.1). Moreover, $\mathcal{A}_{0,0}$ has only one element, the empty string, and $\mathcal{A}_{m,n}$ is the empty set if $m < 0$ or $n < 0$. Thus the coefficient $a_{m,n}$ and the cardinality of $\mathcal{A}_{m,n}$ satisfy the same initial conditions and the same recurrence. It follows that they must be equal. □

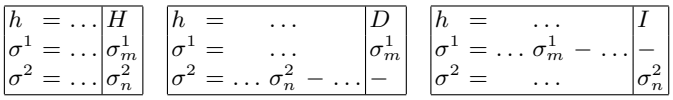


Fig. 5.1. Three possibilities for strings in $\mathcal{A}_{m,n}$.

$a_{m,n}$	0	1	2	3	4	5	6	7	8	9	10
0	1	1	1	1	1	1	1	1	1	1	1
1	1	3	5	7	9	11	13	15	17	19	21
2	1	5	13	25	41	61	85	113	145	181	221
3	1	7	25	63	129	231	377	575	833	1,159	1,561
4	1	9	41	129	321	681	1,289	2,241	3,649	5,641	8,361
5	1	11	61	231	681	1,683	3,653	7,183	13,073	22,363	36,365
6	1	13	85	377	1,289	3,653	8,989	19,825	40,081	75,517	134,245
7	1	15	113	575	2,241	7,183	19,825	48,639	108,545	224,143	433,905
8	1	17	145	833	3,649	13,073	40,081	108,545	265,729	598,417	1256,465
9	1	19	181	1,159	5,641	22,363	75,517	224,143	598,417	1,462,563	3,317,445
10	1	21	221	1,561	8,361	36,365	134,245	433,905	1,256,465	3,317,445	8,097,453

Fig. 5.2. The first hundred Delannoy numbers.

The *alignment graph* of an m-letter sequence and an n-letter sequence is a directed graph $\mathcal{G}_{m,n}$ on the set of nodes $\{0,1,\ldots,m\} \times \{0,1,\ldots,n\}$ and three classes of edges: edges $(i,j) \to (i,j+1)$ are labelled I, edges $(i,j) \to (i+1,j)$ are labelled D, and edges $(i,j) \to (i+1,j+1)$ are labelled H.

Proposition 5.5. *The set of all alignments $\mathcal{A}_{m,n}$ corresponds one-to-one with the set of all paths from the node $(0,0)$ to the node (m,n) in the alignment graph $\mathcal{G}_{m,n}$.*

Proof. Given an alignment in $\mathcal{A}_{m,n}$ by the edit string h. The string h provides a path in $\mathcal{G}_{m,n}$ starting from the node $(0,0)$. By (5.1), this path terminates at the node (m,n).

Conversely, given a path in $\mathcal{G}_{m,n}$ from $(0,0)$ to (m,n). The labelling of the path provides a string h over the edit alphabet that satisfies (5.1). By Prop. 5.3, the string h is an edit string corresponding to an alignment in $\mathcal{A}_{m,n}$. □

Example 5.6. Consider the sequences $\sigma^1 = \texttt{ACG}$ and $\sigma^2 = \texttt{ACC}$. The edit string $h = HHID$ provides the alignment

$$\begin{array}{cccc} H & H & D & I \\ \texttt{A} & \texttt{C} & \texttt{G} & - \\ \texttt{A} & \texttt{C} & - & \texttt{C} \end{array}$$

This alignment can be traced by the solid path in the alignment graph $\mathcal{G}_{3,3}$ (Fig. 5.3). ♢

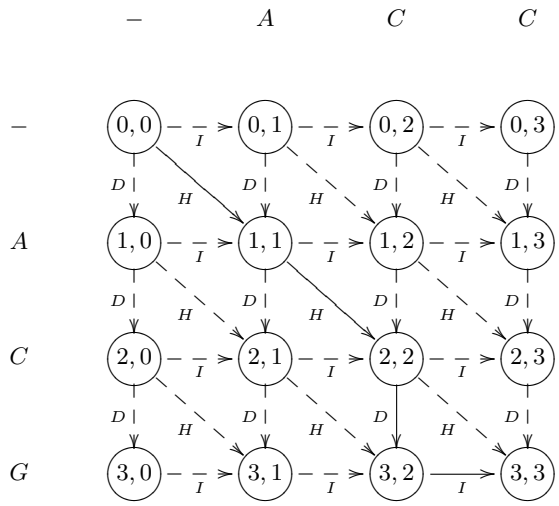


Fig. 5.3. The alignment graph $\mathcal{G}_{3,3}$ and the path corresponding to the alignment in Ex. 5.6.

5.2 Scoring Schemes

We introduce scores to alignments. For this, we need a *scoring scheme* defined by a pair of mappings

$$w : \Sigma \cup \{-\} \times \Sigma \cup \{-\} \to \mathbb{R} \quad \text{and} \quad w' : \{H, I, D\} \times \{H, I, D\} \to \mathbb{R}. \tag{5.4}$$

Take two sequences σ^1 and σ^2 over the alphabet Σ. An alignment of the pair (σ^1, σ^2) is given by a pair of sequences (μ^1, μ^2) over the extended alphabet that can be fully represented by an edit string h over the edit alphabet. The *weight* of the alignment h is defined as

$$W(h) = \sum_{i=1}^{|h|} w(\mu_i^1, \mu_i^2) + \sum_{i=2}^{|h|} w'(h_{i-1}, h_i), \tag{5.5}$$

where $|h|$ denotes the length of the string h. Thus the weight of an alignment is given by the sum of column scores of the aligned sequences and the sum of consecutive scores of the edit string.

Assume that the sequences σ^1 and σ^2 are defined over the DNA alphabet $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. We may represent the scoring scheme (w, w') by a pair of matrices,

$$w = \begin{pmatrix} w_{A,A} & w_{A,C} & w_{A,G} & w_{A,T} & w_{A,-} \\ w_{C,A} & w_{C,C} & w_{C,G} & w_{C,T} & w_{C,-} \\ w_{G,A} & w_{G,C} & w_{G,G} & w_{G,T} & w_{G,-} \\ w_{T,A} & w_{T,C} & w_{T,G} & w_{T,T} & w_{T,-} \\ w_{-,A} & w_{-,C} & w_{-,G} & w_{-,T} & \end{pmatrix} \tag{5.6}$$

and

$$w' = \begin{pmatrix} w'_{H,H} & w'_{H,I} & w'_{H,D} \\ w'_{I,H} & w'_{I,I} & w'_{I,D} \\ w'_{D,H} & w'_{D,I} & w'_{D,D} \end{pmatrix}. \tag{5.7}$$

The lower right entry $w_{-,-}$ in the matrix w is left out because it is never used by our convention. It follows that the total number of parameters in the alignment problem is $24 + 9 = 33$. We identify the parameter space with the Euclidean space $\mathbb{R}^{33}$. Thus each alignment $h \in \mathcal{A}_{m,n}$ gives rise to a mapping $W(h) : \mathbb{R}^{33} \to \mathbb{R}$.

Example 5.7. Consider the alignment of the sequences $\sigma^1 = \mathtt{ACGTAGC}$ and $\sigma^2 = \mathtt{ACCGAGACC}$ given by the edit string $h = HHIHIHHIDH$ (Ex. 5.2). The weight of this alignment is the linear expression

$$\begin{aligned} W(h) = {} & 2 \cdot w_{A,A} + 2 \cdot w_{C,C} + 1 \cdot w_{G,G} + 1 \cdot w_{T,G} + 2 \cdot w_{-,C} + 1 \cdot w_{-,A} + 1 \cdot w_{G,-} \\ & + 2 \cdot w'_{H,H} + 3 \cdot w'_{H,I} + 2 \cdot w'_{I,H} + 1 \cdot w'_{I,D} + 1 \cdot w'_{D,H}. \end{aligned}$$

♢

Example 5.8 (Maple). We compute symbolically the weight of an alignment. For simplicity, we put $w' = 0$ and consider the sequences

```
s1 := [A,C,G]: s2 := [A,C,C]:
```

The scoring scheme is given by the matrix w and can be defined as

```
w :=  array([ [ tAA, tAC, tAG, tAT, t_A ],
              [ tCA, tCC, tCG, tCT, t_C ],
              [ tGA, tGC, tGG, tGT, t_G ],
              [ tTA, tTC, tTG, tTT, t_T ],
              [ tA_, tC_, tG_, tT_,     ] ]):
```

Assume that the alignment is given by the edit string

```
h := [H,H,D,I]:
```

First, it must be checked that the string h describes an alignment. For this, we need to show that (5.1) holds

```
l := nops(h): m := nops(s1): n := nops(s2):
nH := 0: nI := 0: nD := 0:
for i from 1 to l do
  if h[i] = H then nH := nH + 1
    elif h[i] = D then nD := nD + 1
    else nI := nI + 1
  end if
end do;
if nH + nD = m and nH + nI = n then
  print("h defines alignment");
end if:
```

Second, the aligned sequences are established

```
i1 := 1: i2 := 1:
a1 := []; a2 := [];
for i from 1 to l do
  if h[i] = H then
      a1 := [op(a1), s1[i1]]; i1 := i1 + 1;
      a2 := [op(a2), s2[i2]]; i2 := i2 + 1
    elif h[i] = D then
      a1 := [op(a1), s1[i1]]; i1 := i1 + 1;
      a2 := [op(a2), _]
    else h[i] = D then
      a1 := [op(a1), _];
      a2 := [op(a2), s2[i2]]; i2 := i2 + 1
  end if
end do:
a1, a2;
```

The last command provides the aligned sequences

```
[A,C,G,_]
[A,C,_,C]
```

Third, the weight of the alignment is calculated

```
u1 := subs( { A=1, C=2, G=3, T=4 }, s1);
u2 := subs( { A=1, C=2, G=3, T=4 }, s2);
W_h := 0:
for i from 1 to l do
  W_h := W_h + w[u1[i],u2[i]]
end do:
expand (W_h);
```

The last command outputs the alignment weight

```
tAA + tCC + tG_ + t_C
```

♢

Let σ^1 and σ^2 be sequences of length m and n over the alphabet Σ, respectively. Given a scoring scheme (w, w'), the *alignment problem* is to compute alignments $h \in \mathcal{A}_{m,n}$ of the pair (σ^1, σ^2) that have minimum weight $W(h)$ among all alignment in $\mathcal{A}_{m,n}$. These alignments are called *optimal*. The problem is thus to solve the optimization problem

$$\begin{array}{l} \min W(h). \\ \text{s.t.} \;\; h \in \mathcal{A}_{m,n} \end{array} \tag{5.8}$$

Sometimes we simplify the alignment problem by assuming that $w' = 0$. Then the weight of an alignment (μ^1, μ^2) given by the edit string h is the linear functional

$$W(h) = \sum_{i=1}^{|h|} w(\mu_i^1, \mu_i^2). \tag{5.9}$$

The alignment problem can be interpreted in terms of the alignment graph. For this, the edges of the alignment graph $\mathcal{G}_{m,n}$ are weighted by scores

$$(i,j) \overset{w_{-,\sigma_{j+1}^2}}{\longrightarrow} (i, j+1), \quad (i,j) \overset{w_{\sigma_{i+1}^1,-}}{\longrightarrow} (i+1, j), \quad (i,j) \overset{w_{\sigma_{i+1}^1,\sigma_{j+1}^2}}{\longrightarrow} (i+1, j+1).$$

This decorated graph is called *weighted alignment graph* with respect to the scoring scheme $(w, w' = 0)$.

Proposition 5.9. *Let σ^1 and σ^2 be sequences over Σ of length m and n, respectively. The problem of finding the optimal alignments for the pair of sequences (σ^1, σ^2) with respect to $(w, w' = 0)$ is equivalent to finding the minimum weight paths from the node $(0,0)$ to the node (m,n) in the weighted alignment graph $\mathcal{G}_{m,n}$ with respect to $(w, w' = 0)$.*

Proof. By Prop. 5.5, the alignments of the pair of sequences (σ^1, σ^2) correspond one-to-one to the paths from $(0,0)$ to (m,n) in the graph $\mathcal{G}_{m,n}$. Moreover, by definition, the weight of an alignment given in (5.9) equals the weight of the corresponding path from $(0,0)$ to (m,n) in the weighted graph. □

Example 5.10. Take the sequences $\sigma^1 = \texttt{ACG}$ and $\sigma^2 = \texttt{ACC}$, and use the scoring scheme given by

$$w = \begin{pmatrix} 3 & -1 & -1 & -1 & -2 \\ -1 & 3 & -1 & -1 & -2 \\ -1 & -1 & 3 & -1 & -2 \\ -1 & -1 & -1 & 3 & -2 \\ -2 & -2 & -2 & -2 & \end{pmatrix} \quad \text{and} \quad w' = 0.$$

That is, matches are scored with $+3$, mismatches are scored with -1, and indels are scored with -2. The alignment $(\mu^1, \mu^2) = (\texttt{AC}-\texttt{G}, \texttt{ACC}-)$ is given by the edit string $h = HHID$ (Ex. 5.6) and has the score

$$W(h) = w(\texttt{A}, \texttt{A}) + w(\texttt{C}, \texttt{C}) + w(-, \texttt{C}) + w(\texttt{G}, -) = 3 + 3 - 2 - 2 = 2.$$

The weighted alignment graph and the path corresponding to this alignment are shown in Fig. 5.4. ♢

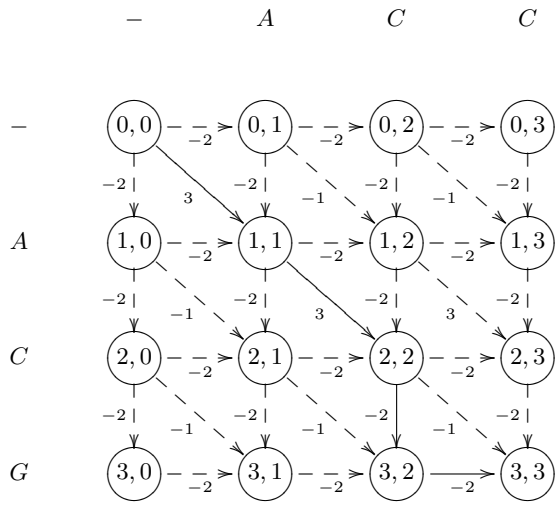


Fig. 5.4. A weighted alignment graph and the path corresponding to the alignment in Ex. 5.6.

5.3 Pair Hidden Markov Model

We show that the sequence alignment problem can be interpreted as an algebraic statistical model. For simplicity, we restrict our attention to the DNA alphabet $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. The *pair hidden Markov model* for the set of alignments $\mathcal{A}_{m,n}$ over Σ is the algebraic statistical model

$$f : \mathbb{R}^{33} \to \mathbb{R}^{4^{m+n}} : (\theta, \theta') \mapsto (f_{\sigma^1,\sigma^2}). \tag{5.10}$$

The model has 4^{m+n} states that correspond to all pairs of sequences σ^1 and σ^2 of length m and n over Σ, respectively. Moreover, the model has $24 + 9 = 33$ parameters that are written as a pair of matrices (θ, θ') as follows,

$$\theta = \begin{pmatrix} \theta_{A,A} & \theta_{A,C} & \theta_{A,G} & \theta_{A,T} & \theta_{A,-} \\ \theta_{C,A} & \theta_{C,C} & \theta_{C,G} & \theta_{C,T} & \theta_{C,-} \\ \theta_{G,A} & \theta_{G,C} & \theta_{G,G} & \theta_{G,T} & \theta_{G,-} \\ \theta_{T,A} & \theta_{T,C} & \theta_{T,G} & \theta_{T,T} & \theta_{T,-} \\ \theta_{-,A} & \theta_{-,C} & \theta_{-,G} & \theta_{-,T} & \end{pmatrix} \tag{5.11}$$

and

$$\theta' = \begin{pmatrix} \theta'_{H,H} & \theta'_{H,I} & \theta'_{H,D} \\ \theta'_{I,H} & \theta'_{I,I} & \theta'_{I,D} \\ \theta'_{D,H} & \theta'_{D,I} & \theta'_{D,D} \end{pmatrix}, \tag{5.12}$$

where the lower right entry $\theta_{-,-}$ in the matrix θ is left out as it is never used by our convention. The parameter space of the model is the product of six simplices of dimensions 15, 3, 3, 2, 2, and 2,

$$\Theta = \Delta_{15} \times \Delta_3 \times \Delta_3 \times \Delta_2 \times \Delta_2 \times \Delta_2 \subseteq \mathbb{R}^{33}. \tag{5.13}$$

The big simplex Δ_{15} consists of all non-negative 4×4 matrices $(\theta_{ij})_{i,j \in \Sigma}$ whose entries sum up to 1. The two thetrahedrons Δ_3 come from the equalities

$$\theta_{-,A} + \theta_{-,C} + \theta_{-,G} + \theta_{-,T} = \theta_{A,-} + \theta_{C,-} + \theta_{G,-} + \theta_{T,-} = 1. \tag{5.14}$$

The three triangles Δ_2 provide the equalities

$$\theta'_{H,H} + \theta'_{H,I} + \theta'_{H,D} = \theta'_{I,H} + \theta'_{I,I} + \theta'_{I,D} = \theta'_{D,H} + \theta'_{D,I} + \theta'_{D,D} = 1. \tag{5.15}$$

The coordinate function f_{σ^1,σ^2} of the pair hidden Markov model represents the marginal probability of observing the aligned pair of sequences σ^1 and σ^2 and is given by

$$f_{\sigma^1,\sigma^2} = \sum_{h \in \mathcal{A}_{m,n}} \prod_{i=1}^{|h|} \theta_{\mu_i^1,\mu_i^2} \cdot \prod_{i=2}^{|h|} \theta'_{h_{i-1},h_i}, \tag{5.16}$$

where (μ^1, μ^2) is the pair of aligned sequences over $\Sigma \cup \{-\}$ which corresponds to the edit string $h \in \mathcal{A}_{m,n}$.

Example 5.11. Consider the alignment of the sequences $\sigma^1 = \texttt{ACGTAGC}$ and $\sigma^2 = \texttt{ACCGAGACC}$ given by the edit string $h = HHIHIHHIDH$ (Ex. 5.2). The string h corresponds to the following term in the marginal probability f_{σ^1,σ^2},

$$\theta^2_{A,A} \cdot \theta^2_{C,C} \cdot \theta^2_{-,C} \cdot \theta_{G,G} \cdot \theta_{-,A} \cdot \theta_{T,G} \cdot \theta_{G,-} \cdot \theta'^2_{H,H} \cdot \theta'^3_{H,I} \cdot \theta'^2_{I,H} \cdot \theta'_{I,D} \cdot \theta'_{D,H}.$$

♢

Proposition 5.12. *The alignment problem (5.8) for the pair of sequences (σ^1, σ^2) is the tropicalization of the marginal probability f_{σ^1,σ^2} of the pair hidden Markov model.*

Proof. We apply the tropicalization map to the marginal probability f_{σ^1,σ^2}. For this, we put $w_{ij} = -\log \theta_{ij}$ and $w'_{XY} = -\log \theta'_{XY}$, where $1 \le i \le m$, $1 \le j \le n$, and $X, Y \in \{H, D, I\}$, and replace the outer sum by a tropical sum and the inner product by a tropical product. In this way, we obtain the tropical polynomial

$$\operatorname{trop}(f_{\sigma^1,\sigma^2}) = \bigoplus_{h \in \mathcal{A}_{m,n}} \bigodot_{i=1}^{|h|} w_{\mu_i^1,\mu_i^2} \odot \bigodot_{i=2}^{|h|} w'_{h_{i-1},h_i}. \tag{5.17}$$

For each alignment $h \in \mathcal{A}_{m,n}$, the corresponding tropical product in (5.17) equals the weight of the alignment h,

$$W(h) = \bigodot_{i=1}^{|h|} w_{\mu_i^1,\mu_i^2} \odot \bigodot_{i=2}^{|h|} w'_{h_{i-1},h_i}. \tag{5.18}$$

Since the tropical addition is associated with the formation of minima, the tropical polynomial $\operatorname{trop}(f_{\sigma^1,\sigma^2})$ corresponds to the alignment problem for the pair of sequences (σ^1, σ^2),

$$\operatorname{trop}(f_{\sigma^1,\sigma^2}) = \min_{h \in \mathcal{A}_{m,n}} W(h). \tag{5.19}$$

□

It follows that the evaluation of marginal probability f_{σ^1,σ^2} solves the alignment problem for the sequences σ^1 and σ^2. However, this is only practical for smaller sequences.

5.4 Sum-Product Decomposition

We show that the marginal probabilities of the pair hidden Markov model can be efficiently calculated by a sum-product decomposition. For this, let $\sigma^1 = \sigma^1_1 \dots \sigma^1_m$ and $\sigma^2 = \sigma^2_1 \dots \sigma^2_n$ be DNA sequences.

Let $\sigma^1_{\leq i}$ denote the prefix $\sigma^1_1 \dots \sigma^1_i$ of σ^1, $1 \leq i \leq m$, and let $\sigma^2_{\leq j}$ be the prefix $\sigma^2_1 \dots \sigma^2_j$ of σ^2, $1 \leq j \leq n$. Let $\mathcal{M}^X(i,j)$ be the probability of observing the aligned pair of sequences $\sigma^1_{\leq i}$ and $\sigma^2_{\leq i}$ such that X is the last symbol in the corresponding edit string. Then the marginal probability f_{σ^1,σ^2} can be decomposed as follows,

$$f_{\sigma^1,\sigma^2} = \sum_X \mathcal{M}^X(m,n), \tag{5.20}$$

where

$$\mathcal{M}^I(i,j) = \theta_{-,\sigma^2_j} \cdot \sum_X \mathcal{M}^X(i,j-1) \cdot \theta'_{X,I}, \tag{5.21}$$

$$\mathcal{M}^D(i,j) = \theta_{\sigma^1_i,-} \cdot \sum_X \mathcal{M}^X(i-1,j) \cdot \theta'_{X,D}, \tag{5.22}$$

$$\mathcal{M}^H(i,j) = \theta_{\sigma^1_i,\sigma^2_j} \cdot \sum_X \mathcal{M}^X(i-1,j-1) \cdot \theta'_{X,H}, \tag{5.23}$$

and

$$\mathcal{M}^X(0,0) = 1, \quad X \in \{H,I,D\}, \tag{5.24}$$

$$\mathcal{M}^X(0,j) = 1, \quad X \in \{H,D\}, \tag{5.25}$$

$$\mathcal{M}^X(i,0) = 1, \quad X \in \{H,I\}, \tag{5.26}$$

$$\mathcal{M}^I(0,j) = \theta_{-,\sigma^2_1} \cdot \prod_{k=2}^{j} \theta'_{I,I} \cdot \theta_{-,\sigma^2_k}, \tag{5.27}$$

$$\mathcal{M}^D(i,0) = \theta_{\sigma^1_1,-} \cdot \prod_{k=2}^{i} \theta'_{D,D} \cdot \theta_{\sigma^1_k,-}. \tag{5.28}$$

Note that three cases can occur for the alignment of the prefixes $\sigma^1_{\leq i}$ and $\sigma^2_{\leq j}$ as given by the equations (5.21)-(5.23) (see Fig. 5.5).

$h = \dots$	$h = \dots$	$h = \dots$
$\sigma^1_{\leq i} = \dots \sigma^1_i - \dots$	$\sigma^1_{\leq i} = \dots$	$\sigma^1_{\leq i} = \dots$
$\sigma^2_{\leq j} = \dots$	$\sigma^2_{\leq j} = \dots \sigma^2_j - \dots$	$\sigma^2_{\leq j} = \dots$
I	D	H
$-$	σ^1_i	σ^1_i
σ^2_j	$-$	σ^2_j

Fig. 5.5. The alignment of the prefixes $\sigma^1_{\leq i}$ and $\sigma^2_{\leq j}$.

Proposition 5.13. *The evaluation of the marginal probability f_{σ^1,σ^2} in (5.20) requires $O(mn)$ steps.*

Proof. The array $\Phi^X(i,j)$, $0 \le i \le m$, $0 \le j \le n$, $X \in \{H, D, I\}$, has $3mn$ entries and each entry is computed by a constant number of operations. □

Example 5.14 (Maple). We calculate the marginal probability f_{σ^1,σ^2} by using Maple. For simplicity, we put $w' = 0$ and consider the model $f : \mathbb{R}^{24} \to \mathbb{R}^{4^{m+n}}$. Then the matrix θ' specializes to

$$\theta' = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Take the sequences

```
s1 := [A,C,G]: s2 := [A,C,C]:
```

and provide the 24 parameters by the matrix θ defined as

```
T :=  array([ [ tAA, tAC, tAG, tAT, t_A ],
              [ tCA, tCC, tCG, tCT, t_C ],
              [ tGA, tGC, tGG, tGT, t_G ],
              [ tTA, tTC, tTG, tTT, t_T ],
              [ tA_, tC_, tG_, tT_,     ] ]):
```

We initialize

```
m := nops(s1):
n := nops(s2):
u1 := subs( { A=1, C=2, G=3, T=4 }, s1);
u2 := subs( { A=1, C=2, G=3, T=4 }, s2);
blank := 5:
```

and obtain by ordinary arithmetics on polynomials:

```
M := array( [], 0..m, 0..n):
M[0,0] := 1;
for i from 1 to m do
  M[i,0] := M[i-1,0] * T[u1[i],blank]
od:
for j from 1 to n do
  M[0,j] := M[0,j-1] * T[blank,u2[j]]
od:
for i from 1 to m do
  for j from 1 to n do
    M[i,j] := M[i-1,j] * T[u1[i],blank]
            + M[i,j-1] * T[blank,u2[j]]
            + M[i-1,j-1] * T[u1[i],u2[j]]
  od:
od:
lprint( expand(M[m,n]) );
```

This code produces the marginal probability $f_{\mathtt{ACG,ACC}}$:

```
20*tC_^2*t_A*t_C*t_G*tA_
+ 6*tC_^2*t_G*t_C*tAA
+ 3*tC_^2*t_G*t_A*tCA
+ tC_^2*t_A*t_C*tGA
+ 4*tC_*t_G*t_C*tA_*tAC
+ 7*tC_*t_G*tCC*t_A*tA_
+ 3*tC_*t_G*tCC*tAA
+ 9*tC_*tGC*t_A*t_C*tA_
+ 3*tC_*tGC*t_C*tAA
+ 2*t_C*tGC*tA_*tCA
+ t_G*tCC*tA_*tAC
+ tGC*t_C*tA_*tAC
+ 2*tGC*tCC*t_A*tA_
+ tGC*tCC*tAA.
```

This polynomial has 14 terms and each term stands for an alignment. Moreover, the sum of all coefficients equals the total number of alignments, $|\mathcal{A}_{3,3}| = 63$. For instance, the term $2 * \mathtt{\_C} * \mathtt{GC} * \mathtt{A\_} * \mathtt{CA}$ indicates the alignments

$$\begin{matrix} \mathtt{A\ C\ G} - \\ - \mathtt{\ A\ C\ C} \end{matrix} \quad \text{and} \quad \begin{matrix} \mathtt{A\ C} - \mathtt{\ G} \\ - \mathtt{\ A\ C\ C} \end{matrix}$$

♢

5.5 Optimal Alignment

The tropicalized marginal probability f_{σ^1,σ^2} of the pair hidden Markov model corresponds to the alignment problem for the pair of sequences (σ^1, σ^2). We can compute the tropicalized marginal probability f_{σ^1,σ^2} by tropicalizing its sum-product decomposition. For this, we put $\Phi^X(i,j) = -\log \mathcal{M}^X(i,j)$, $w_{ij} = -\log \theta_{ij}$, and $w'_{XY} = -\log \theta'_{XY}$, where $1 \le i \le m$, $1 \le j \le n$, and $X, Y \in \{H, D, I\}$. By replacing sums by tropical sums and products by tropical products, we obtain

$$\operatorname{trop}(f_{\sigma^1,\sigma^2}) = \bigoplus_X \Phi^X(m,n), \tag{5.29}$$

where

$$\Phi^I(i,j) = w_{-,\sigma^2_j} \odot \bigoplus_X \Phi^X(i,j-1) \odot w'_{X,I}, \tag{5.30}$$

$$\Phi^D(i,j) = w_{\sigma^1_i,-} \odot \bigoplus_X \Phi^X(i-1,j) \odot w'_{X,D}, \tag{5.31}$$

$$\Phi^H(i,j) = w_{\sigma^1_i,\sigma^2_j} \odot \bigoplus_X \Phi^X(i-1,j-1) \odot w'_{X,H}, \tag{5.32}$$

and

$$\Phi^X(0,0) = 0, \quad X \in \{H, I, D\}, \tag{5.33}$$

$$\Phi^X(0,j) = 0, \quad X \in \{H, D\}, \tag{5.34}$$

$$\Phi^X(i,0) = 0, \quad X \in \{H, I\}, \tag{5.35}$$

$$\Phi^I(0,j) = w_{-,\sigma_1^2} \odot \bigodot_{k=2}^{j} w'_{I,I} \odot w_{-,\sigma_k^2}, \tag{5.36}$$

$$\Phi^D(i,0) = w_{\sigma_1^1,-} \odot \bigodot_{k=2}^{i} w'_{D,D} \odot w_{\sigma_k^1,-}. \tag{5.37}$$

5.6 Needleman-Wunsch Algorithm

The tropicalized sum-product decomposition of the marginal probability f_{σ^1,σ^2} corresponds to the alignment problem for the pair of sequences (σ^1, σ^2). This decomposition provides directly an efficient algorithm for computing the optimal alignments of the pair (σ^1, σ^2). The Needleman-Wunsch algorithm is a special case of this algorithm by assuming that the 3×3 matrix w' is zero (Alg. 5.1).

Algorithm 5.1 Needleman-Wunsch algorithm.

Require: sequences $\sigma^1 \in \Sigma^m$, $\sigma^2 \in \Sigma^n$ and scoring scheme $w \in \mathbb{R}^{24}$
Ensure: alignment $h \in \mathcal{A}_{m,n}$ with minimal weight $W(h)$

```
M ← matrix[0..m, 0..n]
M[0,0] ← 0
for i ← 1 to m do
    M[i,0] ← M[i−1,0] + w_{σ_i^1,−}
end for
for j ← 1 to n do
    M[0,j] ← M[0,j−1] + w_{−,σ_j^2}
end for
for i ← 1 to m do
    for j ← 1 to n do
        M[i,j] ← min{M[i−1,j−1] + w_{σ_i^1,σ_j^2}, M[i−1,j] + w_{σ_i^1,−}, M[i,j−1] + w_{−,σ_j^2}}
        color the edges directed to (i,j) that attain the minimum
    end for
end for
Trace a path in the backward direction from (m,n) to (0,0) by following an arbitrary sequence of colored edges.
Output the edge labels in {H,I,D} of the given path in the forward direction.
```

Proposition 5.15. *Let σ^1 and σ^2 be sequences over Σ of length m and n, respectively. The Needleman-Wunsch algorithm computes an optimal alignment of the pair (σ^1, σ^2) with respect to the scoring scheme $(w, w' = 0)$ and has running time $O(mn)$.*

Proof. By Prop. 5.12, the alignment problem for the pair (σ^1, σ^2) equals the tropicalization of the marginal probability f_{σ^1,σ^2}. The sum-product decomposition of this tropicalized probability for the scoring scheme $(w, w' = 0)$ given by the equations (5.29) to (5.37) corresponds one-to-one with the Needleman-Wunsch algorithm.

The computation of the $(m+1) \times (n+1)$ array M requires $O(mn)$ steps. The coloring of the edges takes constant time and both the tracing of a colored path and the output of the aligned sequences take $O(m+n)$ steps. □

The Needleman-Wunsch algorithm follows the paradigm of *dynamic programming* introduced by Richard Bellman in the 1950s. This is a method for solving a complex problem by breaking it down into a collection of subproblems. It makes use of subproblem overlap and optimal substructure to solve a problem in much less time than problems that cannot take advantage of these characteristics.

Example 5.16 (Maple). Take the sequences

```
s1 := [A,C,G]: s2 := [A,C,C]:
```

and provide the 24 parameters by the matrix w defined as

```
T :=  array([ [ -3,  1,  1,  1,  2 ],
              [  1, -3,  1,  1,  2 ],
              [  1,  1, -3,  1,  2 ],
              [  1,  1,  1, -3,  2 ],
              [  2,  2,  2,  2,    ] ]):
```

We initialize

```
m := nops(s1):
n := nops(s2):
u1 := subs( { A=1, C=2, G=3, T=4 }, s1);
u2 := subs( { A=1, C=2, G=3, T=4 }, s2);
blank := 5:
```

and obtain by ordinary arithmetics

```
M := array( [], 0..m, 0..n):
M[0,0] := 0;
for i from 1 to m do
  M[i,0] := M[i-1,0] + T[u1[i],blank]
od:
for j from 1 to n do
  M[0,j] := M[0,j-1] + T[blank,u2[j]]
od:
for i from 1 to m do
  for j from 1 to n do
    M[i,j] := min( M[i-1,j] + T[u1[i],blank],
                   M[i,j-1] + T[blank,u2[j]],
                   M[i-1,j-1] + T[u1[i],u2[j]] )
  od:
od:
lprint( M[m,n] );
```

The last command provides the score -5. The entries of the table M can be read out by the loop

```
for i from 1 to m do
  for j from 1 to n do
    printf("%2d ", M[i,j]);
  od;
  printf("\n");
od;
```

The weighted alignment graph in Fig. 5.6 illustrates the table output together with the colored edges. The graph exhibits exactly one colored path from (0,0) to (3,3): $(0,0) \longrightarrow (1,1) \longrightarrow (2,2) \longrightarrow (3,3)$. This path amounts to the optimal alignment with score -5:

$$\begin{aligned} h &= H\ H\ H \\ \mu^1 &= \mathtt{A}\ \mathtt{C}\ \mathtt{G} \\ \mu^2 &= \mathtt{A}\ \mathtt{C}\ \mathtt{C} \end{aligned}$$

$\diamondsuit$

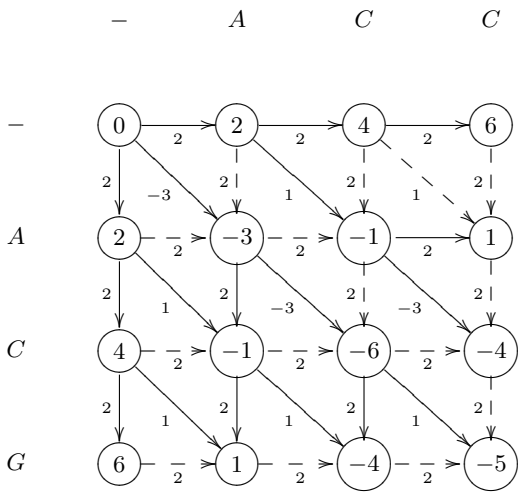


Fig. 5.6. The weighted alignment graph $\mathcal{G}_{3,3}$. The nodes provide the values of the array M and the colored edges are indicated in bold.

5.7 Parametric Sequence Alignment

We have formalized the problem of DNA sequence alignment by a pair hidden Markov model with 33 parameters. If the scoring scheme is fixed, the sum-product decomposition given by the equations (5.29) to (5.37) can be used to compute the optimal alignments for any pair of DNA sequences. Now we want

the scoring scheme to vary over all possible parameter values. We will see that the parameter space can be subdivided into regions such that parameter values in the same region give rise to the same optimal alignment. This subdivision can be attained by evaluating the coordinate functions of the pair hidden Markov model in the polytope algebra.

More concretely, we replace the parameter space $\mathbb{R}^{33}$ by the polynomial ring $\mathbb{R}[\theta, \theta']$ in the 33 variables (θ, θ') and consider each marginal probability f_{σ^1,σ^2} as a polynomial in this ring. Each such polynomial can be assigned a Newton polytope in the polytope algebra $\mathcal{P}_{33}$. Since each marginal probability f_{σ^1,σ^2} can be computed according to the sum-product decomposition given by the equations (5.20) to (5.28), the associated Newton polytope can be calculated by a corresponding sum-product decomposition in the polytope algebra $\mathcal{P}_{33}$. The evaluation of such a Newton polytope is called *polytope propagation*.

Let σ^1 and σ^2 be DNA sequences of length m and n, respectively. The polytope propagation algorithm for evaluating the marginal probability f_{σ^1,σ^2} in the polytope algebra $\mathcal{P}_{33}$ is given as follows:

$$\mathrm{NP}(f_{\sigma^1,\sigma^2}) = \bigoplus_X \mathcal{P}^X(m,n), \tag{5.38}$$

where

$$\mathcal{P}^I(i,j) = \mathrm{NP}(\theta_{-,\sigma_j^2}) \odot \bigoplus_X \mathcal{P}^X(i,j-1) \odot \mathrm{NP}(\theta'_{X,I}), \tag{5.39}$$

$$\mathcal{P}^D(i,j) = \mathrm{NP}(\theta_{\sigma_i^1,-}) \odot \bigoplus_X \mathcal{P}^X(i-1,j) \odot \mathrm{NP}(\theta'_{X,D}), \tag{5.40}$$

$$\mathcal{P}^H(i,j) = \mathrm{NP}(\theta_{\sigma_i^1,\sigma_j^2}) \odot \bigoplus_X \mathcal{P}^X(i-1,j-1) \odot \mathrm{NP}(\theta'_{X,H}), \tag{5.41}$$

and

$$\mathcal{P}^X(0,0) = \{0\}, \quad X \in \{H,I,D\}, \tag{5.42}$$

$$\mathcal{P}^X(0,j) = \{0\}, \quad X \in \{H,D\}, \tag{5.43}$$

$$\mathcal{P}^X(i,0) = \{0\}, \quad X \in \{H,I\}, \tag{5.44}$$

$$\mathcal{P}^I(0,j) = \mathrm{NP}(\theta_{-,\sigma_1^2}) \odot \bigodot_{k=2}^{j} \mathrm{NP}(\theta'_{I,I}) \odot \mathrm{NP}(\theta_{-,\sigma_k^2}), \tag{5.45}$$

$$\mathcal{P}^D(i,0) = \mathrm{NP}(\theta_{\sigma_1^1,-}) \odot \bigodot_{k=2}^{i} \mathrm{NP}(\theta'_{D,D}) \odot \mathrm{NP}(\theta_{\sigma_k^1,-}). \tag{5.46}$$

By Prop. 4.21, the normal cones of the polytope $\mathrm{NP}(f_{\sigma^1,\sigma^2})$ corresponding to the vertices provide an explanation of the marginal probability f_{σ^1,σ^2}.

Example 5.17 (Maple). We consider a simplified pair hidden Markov model for the alignment of DNA sequences with two parameters X and Y that correspond to matches, mismatches, and indels as follows:

$$\begin{aligned} \theta_{a,a} &= X, \quad a \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}, \\ \theta_{a,b} &= Y, \quad a,b \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T},-\},\ a \neq b, \\ \theta'_{X,Y} &= 1, \quad X,Y \in \{H,I,D\}. \end{aligned}$$

Let σ^1 and σ^2 be DNA sequences of length m and n, respectively. We view X and Y as variables over $\mathbb{R}$ and write P_X and P_Y for the corresponding Newton polytopes $\mathrm{NP}(X) = \{(1,0)\}$ and $\mathrm{NP}(Y) = \{(0,1)\}$, respectively. In this case, the marginal probability f_{σ^1,σ^2} can be viewed as a polynomial in the polynomial ring $\mathbb{R}[X,Y]$. The polytope propagation algorithm for evaluating the marginal probability f_{σ^1,σ^2} in the polytope algebra $\mathcal{P}_2$ has the shape

$$\mathrm{NP}(f_{\sigma^1,\sigma^2}) = \mathcal{P}(m,n), \tag{5.47}$$

where

$$\mathcal{P}(i,j) = \Big(\mathcal{P}(i-1,j-1) \odot \mathrm{NP}(\theta_{\sigma_i^1,\sigma_j^2})\Big) \oplus (\mathcal{P}(i-1,j) \odot P_X) \oplus (\mathcal{P}(i,j-1) \odot P_Y) \tag{5.48}$$

and

$$\mathcal{P}(0,0) = \{(0,0)\}, \tag{5.49}$$
$$\mathcal{P}(i,0) = \mathcal{P}(i-1,0) \odot P_X, \tag{5.50}$$
$$\mathcal{P}(0,j) = \mathcal{P}(0,j-1) \odot P_Y. \tag{5.51}$$

Note that the polytopes $\mathcal{P}(i,0)$ and $\mathcal{P}(0,j)$ are translates of polytopes by unit vectors, and the polytope $\mathcal{P}(i,j)$ is given by the convex hull of three translates of polytopes by unit vectors, $1 \le i \le m$ and $1 \le j \le n$.

We implement this polytope propagation algorithm by using `Maple`. For this, we take the sequences

```
s1 := [A,T,C,G]: s2 := [T,C,G,G]:
```

use the linear algebra package

```
with(linalg):
```

and initialize as follows:

```
m := nops( s1 ):
n := nops( s2 ):
u1 := subs( { A=1, C=2, G=3, T=4 }, s1 ):
u2 := subs( { A=1, C=2, G=3, T=4 }, s2 ):
Px := vector( [1,0] ):
Py := vector( [0,1] ):
```

Each polytope in $\mathbb{R}^2$ is given by the convex hull of a finite set of points in $\mathbb{R}^2$, and this set or any superset of it can be viewed as a *generating set* of the polytope. In the `Maple` code, we represent each polytope P by a generating set.

The operations in the polytope algebra $\mathcal{P}_2$ can be implemented by using generating sets. To see this, let P and Q be polytopes in $\mathbb{R}^2$ with generating sets A and B, respectively. The sum $P \oplus Q$ has the generating set $A \cup B$. Note that the union of two sets can be formed by the `Maple` operation `union`. The product $P \odot Q$ has the generating set $\{a+b \mid a \in A, b \in B\}$ and can be obtained as follows:

```
MinkowskiSum := proc ( P, Q )
local R:
  R := {}:
```

```
  for p in P do
    for q in Q do
      R := R union { matadd (p,q) }
    od
  od
  return R:
end proc:
```

The Newton polytope of the marginal probability $f_{\texttt{ATCG,TCGG}}$ can be established by polytope arithmetics as follows:

```
M := array( [], 0..m, 0..n ):
M[0,0] := { vector([0,0]) }:
for i from 1 to m do
  M[i,0] := M[i-1,0] union { Py }
od:
for j from 1 to n do
  M[0,j] := M[0,j-1] union { Py }
od:
for i from 1 to m do
  for j from 1 to n do
    M[i,j] := MinkowskiSum( M[i,j-1], { Py } ):
    M[i,j] := M[i,j] union MinkowskiSum( M[i-1,j], { Py } )
    if u1[i] = u2[j] then
      M[i,j] := M[i,j] union MinkowskiSum( M[i-1,j-1], { Px } )
    else
      M[i,j] := M[i,j] union MinkowskiSum( M[i-1,j-1], { Py } )
    end if
  od
od:
M[m,n];
```

The last statement prints a generating set of the polytope $P = \mathrm{NP}(f_{\texttt{ATCG,TCGG}})$:

```
[0,8],
[0,7],
[0,6], [1,6],
[0,5], [1,5],
       [1,4], [2,4],
       [1,3], [2,3],
                    [3,2].
```

The polytope P illustrated in Fig. 5.7 has the vertices $(0, 8)$, $(0, 5)$, $(1, 3)$, and $(3, 2)$. The normal cones of the vertices are exhibited in Fig. 5.8. The normal fan of the polytope decomposes the Euclidean 2-space into four cones corresponding to the vertices and four half rays associated to the edges (Fig. 5.9).

The normal cones of the vertices yield the optimal sequence alignments. For instance, the cone of the vertex $(3, 2)$ is given by the intersection of two half-spaces defined by the inequalities $-X + 2Y < 0$

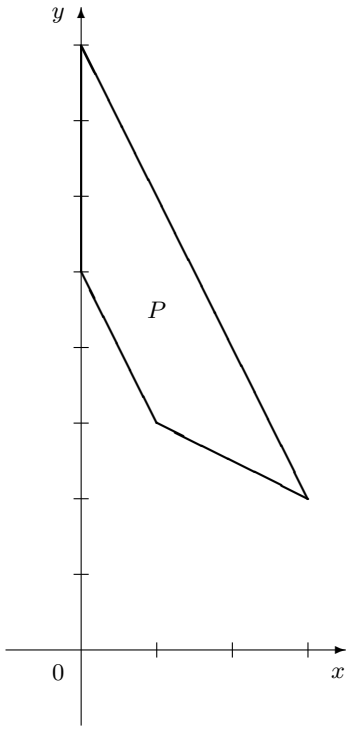


Fig. 5.7. The polytope $P = \mathrm{NP}(f_{\mathtt{ATCG},\mathtt{TCGG}})$.

and $-2X + Y < 0$. We take an arbitrary point in this cone, say $(1, 0)$, and calculate an optimal alignment for the associated scoring scheme with $X = 1$ (matches) and $Y = 0$ (mismatches and indels) by the Needleman-Wunsch algorithm. This is an optimal alignment for all scoring schemes located in this normal cone. Note that this alignment is uniquely determined up to column permutations, since alignments are represented by monomials that are elements of a commutative polynomial ring (Ex. 5.14). All four optimal alignments are illustrated in Table 5.1. ♢

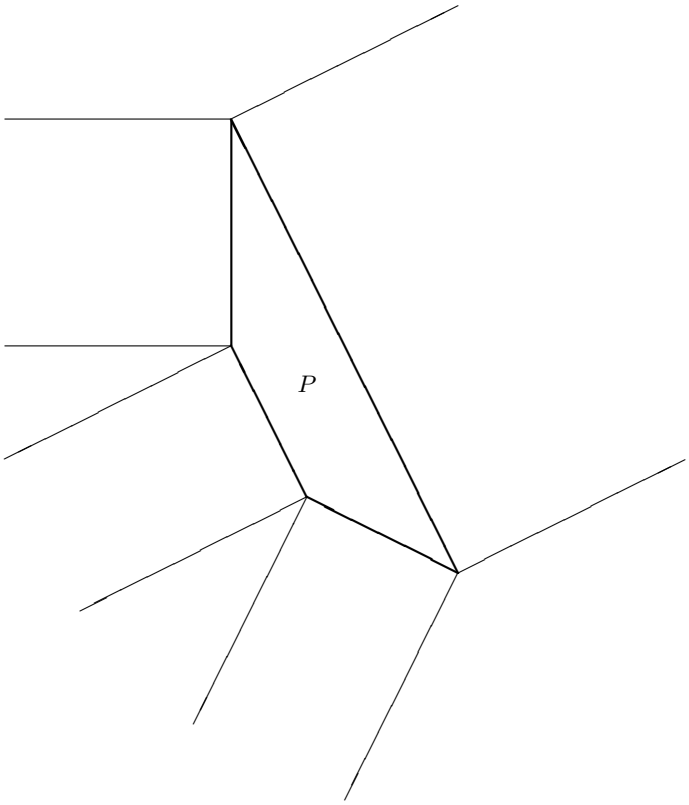


Fig. 5.8. The normal cones of the polytope $P = \mathrm{NP}(f_{\mathtt{ATCG},\mathtt{TCGG}})$ of the vertices.

Table 5.1. The optimal alignments of the sequences `ATCG` and `TCGG`.

vertex	normal cone	scoring scheme	alignment	score
$(0,8)$	$-x+2y>0,\ y>0$	(0,1)	A T C − G − T C G G	2
$(0,5)$	$-x+2y>0,\ y<0$	(-3,-1)	A T C − G − T C G G	-11
$(1,3)$	$-x+2y<0,\ -2x+y>0$	(-2,-2)	A T C G − − − − − − − − T C G G	-16
$(3,2)$	$-x+2y<0,\ -2x+y<0$	(1,0)	A T C G − − − − T C G G	0

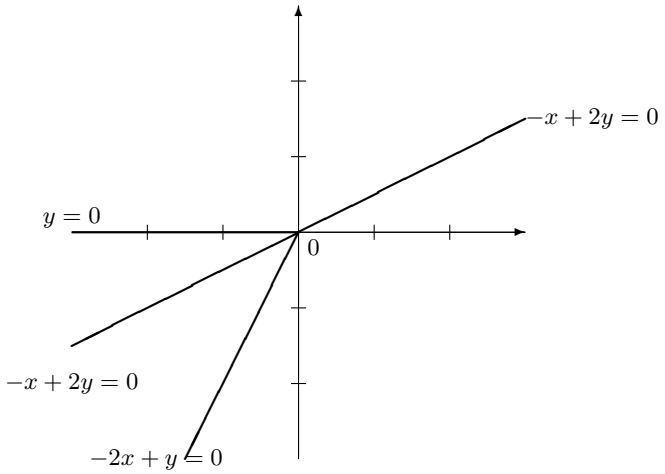


Fig. 5.9. Normal fan of the polytope $\mathrm{NP}(f_{\mathtt{ATCG},\mathtt{TCGG}})$.

6

Hidden Markov Models

The hidden Markov model is a statistical model in which the system modelled is a Markov chain with unknown parameters, and the challenge is to determine the hidden parameters from the observable data. Hidden Markov models were introduced for speech recognition in the 1960s and are now widely used in temporal pattern recognition.

We first introduce the fully observed Markov model in which the states are visible to the observer. Then we proceed to the hidden Markov model where the states are not observable, but each state has a probability distribution over the generated output data. This information can be used to determine the mostly likely state sequence that generated the output data.

6.1 Fully Observed Markov Model

We introduce a variant of the Markov chain model that will serve as a prelimiarly model for the hidden Markov model.

For this, we take an alphabet Σ with l symbols, an alphabet Σ' with l' symbols, and a positive integer n. We consider words $\sigma = \sigma_1 \dots \sigma_n$ and $\tau = \tau_1 \dots \tau_n$ over Σ and Σ' of length n, respectively. These words are used to label the entries of an integral block matrix

$$A_{(l,l'),n} = (A_{\sigma,\tau})_{\sigma \in \Sigma^n, \tau \in \Sigma'^n}, \tag{6.1}$$

whose entries $A_{\sigma,\tau}$ are pairs of matrices (w, w') such that $w = (w_{rs})$ is an $l \times l$ matrix and $w' = (w'_{rt})$ is an $l \times l'$ matrix. The entry $w_{rs} = w_{rs}(\sigma)$ counts the number of occurrences in σ of the length-2 word rs, and the entry $w'_{st} = w_{st}(\sigma, \tau)$ counts the number of indices i, $1 \le i \le n$, such that $\sigma_i = s$ and $\tau_i = t$.

We may view the matrices (w, w') as columns of the matrix $A_{(l,l'),n}$. Then the matrix $A_{(l,l'),n}$ has $d = l \cdot l + l \cdot l' = l^2 + l \cdot l'$ rows labelled by the length-2 words rs in Σ^2 and in turn by the length-2 words rt in $\Sigma \times \Sigma'$. Moreover, the matrix has $m = l^n \cdot l'^n$ columns labelled by the pairs of length-n words σ and τ over Σ and Σ', respectively. The matrix $A_{(l,l'),n}$ has the property that the sum of each of its column entries is $(n-1)+n = 2n-1$, since each word of length n has $n-1$ consecutive length-2 words and two words of length n pair in n positions. Thus the matrix $A_{(l,l'),n}$ defines a toric model $f = f_{(l,l'),n} : \mathbb{R}^d \to \mathbb{R}^m$ given as

$$(\theta, \theta') \mapsto (p_{\sigma,\tau})_{\sigma \in \Sigma^n, \tau \in \Sigma'^n}, \tag{6.2}$$

where

$$p_{\sigma,\tau} = \frac{1}{l}\theta'_{\sigma_1,\tau_1}\theta_{\sigma_1,\sigma_2}\theta'_{\sigma_2,\tau_2}\theta_{\sigma_2,\sigma_3}\cdots\theta_{\sigma_{n-1},\sigma_n}\theta'_{\sigma_n,\tau_n}. \tag{6.3}$$

Here we assume a uniform initial distribution on the states in the alphabet Σ as described by (7.1). All terms $p_{\sigma,\tau}$ have $n+(n-1) = 2n-1$ factors. The parameter space Θ of the model is the cartesian product of the set of positive $l \times l$ matrices θ and the set of positive $l \times l'$ matrices θ'; that is, $\Theta = \mathbb{R}_{>0}^{l\times l} \times \mathbb{R}_{>0}^{l\times l'}$. The matrix θ encodes a toric Markov chain, while the matrix θ' encodes the interplay between the two alphabets. The state space of the model is $\Sigma^n \times \Sigma'^n$. This model is called a *fully observed toric Markov model*.

Example 6.1. Consider a dishonest dealer in a casino tossing coins. We know that she may use a fair or loaded coin, the latter of which is supposed to have the probability of 0.75 to get heads. We also know that she does not tend to change coins, which happens with probability of 0.1 (Fig. 6.1). Given a sequence of coin tosses we wish to determine when she used the biased and fair coins.

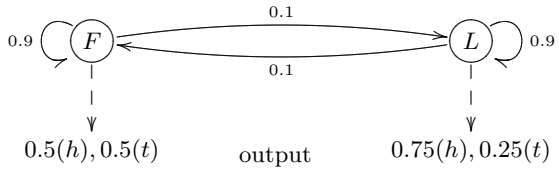


Fig. 6.1. Transition graph of casino model.

This model can be described by a toric Markov model consisting of the alphabets $\Sigma = \{F, L\}$, where F stands for fair and L stands for loaded, and $\Sigma' = \{h, t\}$, where h stands for heads and t stands for tails. The corresponding 8×256 matrix $A_{(2,2),4}$ for sequences of length $n = 4$ consists of the columns $A_{\sigma,\tau}$, where σ and τ range over all words of length 4 over Σ and Σ', respectively; e.g.,

$$A_{FFLL,htht} = \begin{matrix}\texttt{FF}\\ \texttt{FL}\\ \texttt{LF}\\ \texttt{LL}\\ \texttt{Fh}\\ \texttt{Ft}\\ \texttt{Lh}\\ \texttt{Lt}\end{matrix}\begin{pmatrix}1\\1\\0\\1\\1\\1\\1\\1\end{pmatrix} \quad \text{and} \quad A_{FFFL,hhhh} = \begin{matrix}\texttt{FF}\\ \texttt{FL}\\ \texttt{LF}\\ \texttt{LL}\\ \texttt{Fh}\\ \texttt{Ft}\\ \texttt{Lh}\\ \texttt{Lt}\end{matrix}\begin{pmatrix}2\\1\\0\\0\\3\\0\\1\\0\end{pmatrix}.$$

The model has $d = 8$ parameters given by the matrices

$$\theta = \begin{pmatrix}\theta_{FF} & \theta_{FL}\\ \theta_{LF} & \theta_{LL}\end{pmatrix} \quad \text{and} \quad \theta' = \begin{pmatrix}\theta'_{Fh} & \theta'_{Ft}\\ \theta'_{Lh} & \theta'_{Lt}\end{pmatrix}.$$

Suppose the dealer tosses four coins in a row such that we consider sequences of length $n = 4$. Then the model has $m = (2 \cdot 2)^4 = 256$ states. The fully observed toric Markov model is defined by the mapping

$$f : \mathbb{R}^8 \to \mathbb{R}^{256} : (\theta, \theta') \mapsto (p_{\sigma,\tau})_{\sigma \in \Sigma^4, \tau \in \Sigma'^4},$$

where

$$p_{\sigma_1\sigma_2\sigma_3\sigma_4,\tau_1\tau_2\tau_3\tau_4} = \frac{1}{2} \cdot \theta'_{\sigma_1,\tau_1}\theta_{\sigma_1,\sigma_2}\theta'_{\sigma_2,\tau_2}\theta_{\sigma_2,\sigma_3}\theta'_{\sigma_3,\tau_3}\theta_{\sigma_3,\sigma_4}\theta'_{\sigma_4,\tau_4}.$$

For instance, in view of the above pairs of sequences, we obtain

$$\begin{aligned} p_{FFLL,htht} &= \frac{1}{2} \cdot \theta'_{Fh}\theta_{FF}\theta'_{Ft}\theta_{FL}\theta'_{Lh}\theta_{LL}\theta'_{Lt} \\ &= \frac{1}{2} \cdot \theta_{FF}\theta_{FL}\theta_{LL}\theta'_{Fh}\theta'_{Ft}\theta'_{Lh}\theta'_{Lt} \end{aligned}$$

and

$$\begin{aligned} p_{FFFL,hhhh} &= \frac{1}{2} \cdot \theta'_{Fh}\theta_{FF}\theta'_{Fh}\theta_{FF}\theta'_{Fh}\theta_{FL}\theta'_{Lh} \\ &= \frac{1}{2} \cdot \theta^2_{FF}\theta_{FL}\theta'^3_{Fh}\theta'_{Lh}. \end{aligned}$$

♢

Second, we introduce the fully observed Markov model as a submodel of the toric Markov model. For this, the parameter space of the fully observed toric Markov model is restricted to the set of pairs of matrices (θ, θ') whose row sums are equal to 1. The parameter space of the fully observed Markov model is thus a subset Θ_1 of $\mathbb{R}_{>0}^{l\times(l-1)} \times \mathbb{R}_{>0}^{l\times(l'-1)}$, and the number of parameters is $d = (l \cdot (l-1)) \cdot (l \cdot (l'-1)) = l \cdot (l + l' - 2)$. A pair of matrices (θ, θ') in Θ_1 provides an $l \times l$ matrix θ describing *transition probabilities* and an $l \times l'$ matrix θ' providing *emission probabilities*. The value θ_{ij} represents the probability to transit from state $i \in \Sigma$ to state $j \in \Sigma$ in one step, and the value θ'_{ij} is the probability to emit the symbol $j \in \Sigma'$ in state $i \in \Sigma$. The *fully observed Markov model* is given by the mapping $f_{(l,l'),n} : \mathbb{R}^d \to \mathbb{R}^m$ restricted to the parameter space Θ_1. Each point p in the image $f_{(l,l'),n}(\Theta_1)$ is called a *marginal probability*. We assume usually that the initial distribution at the first state in Σ is uniform.

Let $u = (u_{\sigma,\tau}) \in \mathbb{N}_0^{l^n \times l'^n}$ be a frequency vector representing N observed sequence pairs in $\Sigma^n \times \Sigma'^n$. That is, $u_{\sigma,\tau}$ counts the number of times the pair (σ, τ) is observed. Thus, we have $\sum_{\sigma,\tau} u_{\sigma,\tau} = N$. The sufficient statistic $v = A_{(l,l'),n} \cdot u$ can be regarded as a pair of matrices (v, v'), where $v = (v_{rs})$ is an $l \times l$ matrix whose entries v_{rs} are the number of occurrences of $rs \in \Sigma^2$ as a consecutive pair in any of the sequences σ occurring in the observed pairs (σ, τ), and $v' = (v'_{st})$ is an $l \times l'$ matrix whose entries v'_{st} are the number of occurrences of $st \in \Sigma \times \Sigma'$ at the same position in any of the observed sequence pairs (σ, τ).

The likelihood function of the fully observed Markov model is given as

$$L(\theta, \theta') = (\theta, \theta')^{A_{(l,l'),n}\cdot u} = \theta^v \cdot (\theta')^{v'}, \quad (\theta, \theta') \in \Theta_1. \tag{6.4}$$

Proposition 6.2. *In the fully observed Markov chain model $f_{(l,l'),n}$, the maximum likelihood estimate of the frequency data $u \in \mathbb{N}_0^{l^n \times l'^n}$ with sufficient statistic $v = A_{(l,l'),n} \cdot u$ is the matrix pair $(\hat{\theta}, \hat{\theta}')$ in Θ_1 such that*

$$\hat{\theta}_{\sigma_1\sigma_2} = \frac{v_{\sigma_1\sigma_2}}{\sum_{\sigma\in\Sigma} v_{\sigma_1\sigma}} \quad \textit{and} \quad \hat{\theta}'_{\sigma_1\tau_1} = \frac{v'_{\sigma_1\tau_1}}{\sum_{\tau\in\Sigma'} v'_{\sigma_1\tau}}, \quad \sigma_1, \sigma_2 \in \Sigma,\ \tau_1 \in \Sigma'.$$

The proof is analogous to that of Prop. 4.11, since the log-likelihood function similarly decouples into independent parts.

Example 6.3 (Maple). We reconsider the dealer's example. The parameter space Θ_1 of the fully observed Markov model can be viewed as the set of all pairs of probability matrices

$$\theta = \begin{pmatrix} \theta_{FF} & 1-\theta_{FF} \\ 1-\theta_{LL} & \theta_{LL} \end{pmatrix} \quad \text{and} \quad \theta' = \begin{pmatrix} \theta'_{Fh} & 1-\theta'_{Fh} \\ 1-\theta'_{Lt} & \theta'_{Lt} \end{pmatrix},$$

where $\theta_{F,F} = \theta_{L,L} = 0.9$ is the probability to stay with a fair or loaded coin, $\theta_{F,h} = 0.5$ is the probability to observe heads for a fair coin, and $\theta_{L,h} = 0.75$ is the probability to observe heads for a loaded coin. This model has only $d = 4$ parameters.

Suppose a game involves rolling the coin $n = 4$ times. Then the model has $m = (2 \cdot 2)^4 = 256$ states. This Markov model is given by the mapping $f_{(2,2),4} : \mathbb{R}^4 \to \mathbb{R}^{256}$ with marginal probabilities

$$p_{\sigma_1\sigma_2\sigma_3\sigma_4,\tau_1\tau_2\tau_3\tau_4} = \frac{1}{2} \cdot \theta'_{\sigma_1,\tau_1}\theta_{\sigma_1,\sigma_2}\theta'_{\sigma_2,\tau_2}\theta_{\sigma_2,\sigma_3}\theta'_{\sigma_3,\tau_3}\theta_{\sigma_3,\sigma_4}\theta'_{\sigma_4,\tau_4}.$$

For instance, in a game the fair coin was used two times and then the loaded coin was taken two times, and each time heads was observed. The corresponding probability is given by

$$p_{FFLL,hhhh} = \frac{1}{2} \cdot \theta'_{Fh}\theta_{FF}\theta'_{Fh}(1-\theta_{FF})(1-\theta'_{Lt})\theta_{LL}(1-\theta'_{Lt}).$$

We compute symbolically the likelihood function. For this, we take the packages

```
with(combinat): with(linalg):
```

and initialize

```
n := 4: l := 2: l' := 2: m := (l * l')^n:
T := array([ [tFF, tFF], [tLL, tLL] ]):
E := array([ [tFh, tFh], [tLt, tLt] ]):
P := array([], 1..l^n, 1..l'^n):
```

The marginal values are computed by the following code,

```
R := powerset([1,2,3,4]):
S := powerset([1,2,3,4]):
for i from 1 to nops(R) do
  x := vector( [1,1,1,1] ):
  for u from 1 to nops(R[i]) do
    x[R[i,u]] := 2;
  od:
  for j from 1 to nops(S) do
    y := vector( [1,1,1,1] ):
    for v from 1 to nops(S[j]) do
      y[S[j,v]] := 2;
    od:
    P[i,j] := 1/2 * E[x[1],y[1]] * T[x[1],x[2]] * E[x[2],y[2]]
```

```
                    * T[x[2],x[3]] * E[x[3],y[3]] * T[x[3],x[4]]
                    * E[x[4],y[4]];
  od
od:
```

Consider a frequency vector, whose entries are randomly chosen integers between 1 and 5:

```
roll := rand (1..5):
u := randmatrix ( l^n, l'^n, entries = roll ):
```

The likelihood function can be calculated as

```
L := 1:
for i from 1 to 16 do
  for j from 1 to 16 do
    L : = L * P[i,j]^u[i,j]
  od
od:
```

The output (up to a constant) is given by the expression

```
tFh^799 tFF^577 tFt^742 tLh^806 tLF^579 tLt^753 tFL^579 tLL^590
```

Thus the maximum likelihood estimates are

$$\begin{aligned}
\hat{\theta}_{FF} &= 1 - \hat{\theta}_{FL} = \frac{577}{577+579} \\
\hat{\theta}_{LF} &= 1 - \hat{\theta}_{LL} = \frac{579}{579+590} \\
\hat{\theta}'_{Fh} &= 1 - \hat{\theta}'_{Ft} = \frac{799}{799+742} \\
\hat{\theta}'_{Lh} &= 1 - \hat{\theta}'_{Lt} = \frac{806}{806+753}.
\end{aligned}$$

♢

6.2 Hidden Markov Model

A fully observed Markov model gives rise to a hidden Markov model by only observing the emitted sequences. This can be formally described by a socalled marginalization mapping. Consider the fully observed Markov model

$$F : \mathbb{R}^{l\times(l-1)} \times \mathbb{R}^{l\times(l'-1)} \to \mathbb{R}^{l^n\times(l')^n} \tag{6.5}$$

and the *marginalization mapping*

$$\rho : \mathbb{R}^{l^n\times(l')^n} \to \mathbb{R}^{(l')^n} \tag{6.6}$$

that maps each $l^n \times (l')^n$ matrix to the vector of column sums. The algebraic statistical model given by the composition $f = \rho \circ F$ is called a *hidden Markov model*

$$f : \mathbb{R}^{l\times(l-1)} \times \mathbb{R}^{l\times(l'-1)} \to \mathbb{R}^{(l')^n}. \tag{6.7}$$

Both, the hidden Markov model and the underlying fully observed Markov model have the same parameter space $\Theta_1 \subseteq \mathbb{R}_{>0}^{l\times(l-1)} \times \mathbb{R}_{>0}^{l\times(l'-1)}$. For each pair of matrices $(\theta, \theta') \in \Theta_1$, we have

$$f(\theta, \theta') = (p_\tau)_{\tau\in\Sigma'^n}, \tag{6.8}$$

where

$$\begin{aligned} p_\tau &= \sum_{\sigma\in\Sigma^n} p_{\sigma,\tau} \\ &= \frac{1}{l} \sum_{\sigma_1\in\Sigma} \cdots \sum_{\sigma_n\in\Sigma} \theta'_{\sigma_1,\tau_1}\theta_{\sigma_1,\sigma_2}\theta'_{\sigma_2,\tau_2}\theta_{\sigma_2,\sigma_3}\cdots\theta_{\sigma_{n-1},\sigma_n}\theta'_{\sigma_n,\tau_n}. \end{aligned} \tag{6.9}$$

In the hidden Markov model, the states are not observable. What is observable is the output of the current state (Fig. 6.2).

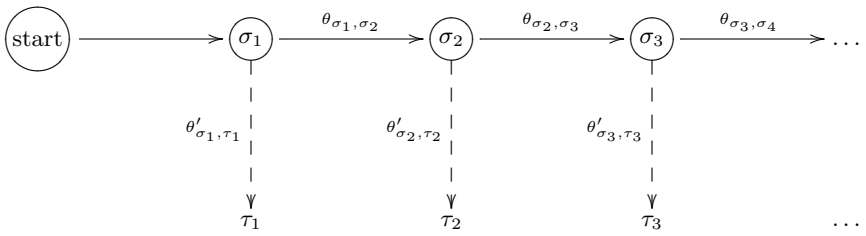


Fig. 6.2. Hidden Markov model.

Example 6.4 (Maple). We reconsider the dealer's example. In the corresponding hidden Markov model, the dealer's coin tosses are observed but not whether she chooses a fair or load coin. This model is given by the mapping ($n = 4$)

$$f : \mathbb{R}^4 \to \mathbb{R}^{16} : (\theta, \theta') \mapsto (p_\tau)_{\tau\in\Sigma'^4},$$

whose marginal probabilities are

$$p_{\tau_1\tau_2\tau_3\tau_4} = \frac{1}{2} \sum_{\sigma_1\in\Sigma} \sum_{\sigma_2\in\Sigma} \sum_{\sigma_3\in\Sigma} \sum_{\sigma_4\in\Sigma} \theta'_{\sigma_1,\tau_1}\theta_{\sigma_1,\sigma_2}\theta'_{\sigma_2,\tau_2}\theta_{\sigma_2,\sigma_3}\theta'_{\sigma_3,\tau_3}\theta_{\sigma_3,\sigma_4}\theta'_{\sigma_4,\tau_4}.$$

Suppose the game is observed N times. Let $u = (u_\tau) \in \mathbb{N}^{16}$ be the corresponding frequency vector. That is, u_τ counts the number of times the sequence $\tau \in \Sigma'^4$ is observed. Then we have $\sum_\tau u_\tau = N$. The goal is to maximize the likelihood function of the model,

$$\ell(\theta_{FF}, \theta_{LL}, \theta'_{Fh}, \theta'_{Lt}) = \prod_{\tau \in \Sigma'^4} p_\tau^{u_\tau}.$$

The `Maple` code for the computation of the likelihood function for the fully observed Markov model can be easily modified to calculate the likelihood function for the associated hidden Markov model. For this, the marginal probabilities are computed as follows,

```
M := vector (l'^n, 0):
for j from 1 to l'^n do
  M[j] := 0;
  for i from 1 to l^n do
    M[j] := M[j] + P[i,j]
  od
od:
```

By taking a randomly chosen frequency vector

```
roll := rand(1..5):
u := randvector (l'^n, entries = roll):
```

here $u = [5, 2, 5, 2, 3, 4, 4, 5, 3, 1, 5, 2, 3, 2, 2, 4]$, we obtain the likelihood function

```
L := 1:
for i from 1 to l'^n do
  L := L * M[i]^u[i]
od
simplify(L);
```

The likelihood function (up to a constant) is given as follows,

```
( rFh^4 tFF^3 + 3 tLt tLL tFh^3 tFF^2 + 3 tLt^2 tLL^2 tFh^2 tFF
+ tLt^3 tLL^3 tFh + tFh^3 tFF^3 tLt + 3 tLt^2 tLL tFh^2 tFF^2
+ 3 tLt^3 tLL^2 tFh tFF + tLt^4 tLL^3 )^52
```

♢

6.3 Sum-Product Decomposition

We show that the marginal probabilities of the hidden Markov model can be efficiently calculated by a sum-product decomposition. For this, consider a hidden Markov model of length n with state set Σ of l symbols and emission set Σ' of l' symbols. The model parameters are the transition probability matrix $\theta \in \mathbb{R}^{l \times (l-1)}$ and the emission probability matrix $\theta' \in \mathbb{R}^{l \times (l'-1)}$. If we assume a uniform initial distribution on the states, the probability of occurrence of the sequence (σ, τ), $\sigma \in \Sigma^n$ and $\tau \in \Sigma'^n$, is given as

$$p_{\sigma,\tau} = \frac{1}{l} \theta'_{\sigma_1,\tau_1} \theta_{\sigma_1,\sigma_2} \theta'_{\sigma_2,\tau_2} \theta_{\sigma_2,\sigma_3} \cdots \theta_{\sigma_{n-1},\sigma_n} \theta'_{\sigma_n,\tau_n}. \tag{6.10}$$

The marginal probability of the observed sequence τ is then given by

$$p_\tau = \sum_{\sigma \in \Sigma^n} p_{\sigma,\tau}. \tag{6.11}$$

This probability has the sum-product decomposition

$$p_\tau = \frac{1}{l} \sum_{\sigma_n \in \Sigma} \theta'_{\sigma_n,\tau_n} \left(\sum_{\sigma_{n-1} \in \Sigma} \theta_{\sigma_{n-1},\sigma_n} \theta'_{\sigma_{n-1},\tau_{n-1}} \left(\cdots \left(\sum_{\sigma_1 \in \Sigma} \theta_{\sigma_1,\sigma_2} \theta'_{\sigma_1,\tau_1} \right) \cdots \right) \right). \tag{6.12}$$

This expression can be evaluated by an $(n-1) \times l$ matrix M defined as follows,

$$\begin{aligned} M_{1,\sigma} &= \sum_{\sigma_1 \in \Sigma} \theta_{\sigma_1,\sigma} \theta'_{\sigma_1,\tau_1}, \quad \sigma \in \Sigma, \\ M_{k,\sigma} &= \sum_{\sigma_k \in \Sigma} \theta_{\sigma_k,\sigma} \theta'_{\sigma_k,\tau_k} \cdot M_{k-1,\sigma_k}, \quad 2 \le k \le n-1, \sigma \in \Sigma \\ p_\tau &= \frac{1}{l} \sum_{\sigma_n \in \Sigma} \theta'_{\sigma_n,\tau_n} \cdot M_{n-1,\sigma_n}. \end{aligned}$$

The computation of the marginal probability p_τ by using this decomposition is called the *forward algorithm* of the hidden Markov model. Its time complexity is $O(l^2 n)$, because the matrix M has $O(ln)$ entries and each entry is evaluated in $O(l)$ steps.

Example 6.5 (Maple). We reconsider the occasionally dishonest casino. We compute the marginal probability p_τ for the observed sequence $\tau = hhtt$. For this, we put

```
t := [0,0,1,1]:
```

and provide the parameters in a hidden Markov model by the matrices

```
T := array([ [tFF, tFF ], [tLL, tLL] ]:
E := array([ [tFh, tFh ], [tLt, tLt] ]:
```

We initialize

```
n := nops(t):
l := 2:
u := subs( {0=1, 1=2 }, t):
```

and obtain by symbolic computation,

```
M := array( [], 0..n, 1..l):
for i from 1 to l do
  M[0,i] := 1
od:
for k from 1 to n-1 do
  for i from 1 to l do
    for j from 1 to l do
      M[k,i] := M[k,i] + T[j,i] * E[j,u[k]] * M[k-1,j]
    od:
```

```
  od:
od:
p := 0:
for j from 1 to l do
  p := p + 1/2 * E[j,u[n]] * M[n-1,j]
od:
expand(p);
```

This code produces the marginal probability p_τ. ♢

6.4 Viterbi Algorithm

The sum-product decomposition of the marginal probabilites can be used to find an explanation for a given sequence of observed data $\tau \in \Sigma'^n$. Finding an *explanation* means identifying a hidden state sequence $\bar{\sigma}$ with maximum a posteriori probability that generated the observed data τ; that is,

$$\bar{\sigma} = \operatorname{argmax}_\sigma \{p_{\sigma,\tau}\}. \tag{6.13}$$

By putting $w_\tau = -\log p_\tau$ and $w_{\sigma,\tau} = -\log p_{\sigma,\tau}$, the tropicalization of the marginal probability (6.9) yields

$$w_\tau = \bigoplus_\sigma w_{\sigma,\tau}. \tag{6.14}$$

The explanation $\bar{\sigma}$ is given by evaluation in the tropical algebra,

$$\bar{\sigma} = \operatorname{argmin}_\sigma \{w_{\sigma,\tau}\}. \tag{6.15}$$

The value w_τ can be efficiently computed by tropicalizing the sum-product decomposition of the marginal probability p_τ. For this, we put $u_{ij} = -\log \theta_{ij}$ and $v_{ij} = -\log \theta'_{ij}$. By replacing the sums by tropical sums and the products by tropical products in the sum-product decomposition (6.12), we obtain

$$w_\tau = \bigoplus_{\sigma_n} v_{\sigma_n,\tau_n} \odot \left(\bigoplus_{\sigma_{n-1}} u_{\sigma_{n-1},\sigma_n} \odot v_{\sigma_{n-1},\tau_{n-1}} \odot \left(\cdots \left(\bigoplus_{\sigma_1} u_{\sigma_1,\sigma_2} \odot v_{\sigma_1,\tau_1} \right) \cdots \right) \right). \tag{6.16}$$

This gives us the following result.

Proposition 6.6. *Let $\tau \in \Sigma'^n$. The tropicalization w_τ of the marginal probability p_τ provides an explanation for the observed data τ.*

The tropicalized term w_τ can be computed by evaluating iteratively the parentheses in (6.16):

$$\begin{aligned} M[0,\sigma] &:= 0, \quad \sigma \in \Sigma, \\ M[k,\sigma] &:= \bigoplus_{\sigma' \in \Sigma} \left(u_{\sigma',\sigma} \odot v_{\sigma',\tau_k} \odot M[k-1,\sigma'] \right), \quad \sigma \in \Sigma,\ 1 \le k \le n-1, \\ M[n,\sigma] &:= v_{\sigma,\tau_n} \odot M[n-1,\sigma], \quad \sigma \in \Sigma, \\ w_\tau &:= \bigoplus_{\sigma \in \Sigma} M[n,\sigma]. \end{aligned} \tag{6.17}$$

This algorithm is known as *Viterbi algorithm* and the computed explanation is called *Viterbi sequence*. The Viterbi algorithm consists of a forward algorithm evaluating the data as given by Alg. 6.1 and a backward algorithm yielding an optimal state sequence that is comprised by the symbols σ which attain the minimum in each minimization step. This information can be recorded by the forward algorithm similar to the forward algorithm for sequence alignment (Alg. 5.1). The time complexity of the Viterbi algorithm is $O(l^2 n)$ as described in the previous section.

Algorithm 6.1 Viterbi forward algorithm.

Require: sequence $\tau \in \Sigma'^n$, scores (u_{ij}) and (v_{ij})
Ensure: tropicalized term w_τ
  $M \leftarrow \text{matrix}[0..n, 1..l]$
  for $\sigma \leftarrow 1$ to l **do**
    $M[0, \sigma] \leftarrow 0$
  end for
  for $k \leftarrow 1$ to $n-1$ **do**
    for $\sigma \leftarrow 1$ to l **do**
      $M[k, \sigma] \leftarrow \infty$
      for $\sigma' \leftarrow 1$ to l **do**
        $M[k, \sigma] \leftarrow \min\{M[k, \sigma], u_{\sigma', \sigma} + v_{\sigma', \tau_k} + M[k-1, \sigma']\}$
      end for
    end for
  end for
  for $\sigma \leftarrow 1$ to l **do**
    $M[n, \sigma] \leftarrow v_{\sigma, \tau_n} + M[n-1, \sigma]$
  end for
  $w_\tau \leftarrow \infty$
  for $\sigma \leftarrow 1$ to l **do**
    $w_\tau \leftarrow \min\{w_\tau, M[n, \sigma]\}$
  end for

Example 6.7. Reconsider the dealer's example. Take the weights as given in Fig. 6.3 and the output sequence $\tau = hhth$. The calculation of the Viterbi algorithm is given in Fig. 6.4. The solid lines show

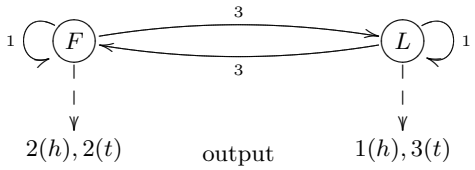


Fig. 6.3. Weighted transition graph of casino model.

where the minima are attained. Tracing back gives the explanation $\bar{\sigma} = LLLL$. ♢

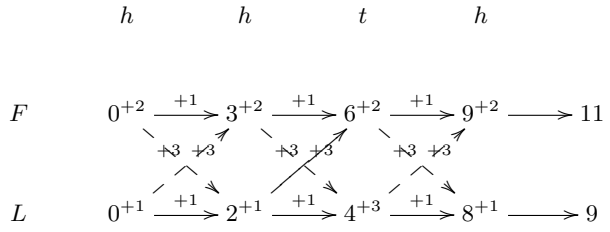


Fig. 6.4. Trellis for the output sequence $\tau = hhth$.

Example 6.8 (Maple). Reconsider the occasionally dishonest casino. We implement the Viterbi algorithm by using Maple. For this, take the observed sequence $\tau = hhtt$ encoded as

```
t := [0,0,1,1]:
```

and provide the minus-log probability matrices

```
T := array([ [-log(0.9), -log(0.1) ], [-log(0.1),  -log(0.9) ] ]:
E := array([ [-log(0.5), -log(0.5) ], [-log(0.75), -log(0.25)] ]:
```

We initialize

```
l := 2:
n := nops(t):
u := subs( {0=1, 1=2 }, t):
```

and obtain by tropical arithmetics,

```
M := array( [], 0..n, 1..l ):
for i from 1 to l do
  M[0,i] := 0
od:
for k from 1 to n-1 do
  for i from 1 to l do
      M[k,i] := min( T[1,i] + E[1,u[k]] + M[k-1,1],
                     T[2,i] + E[2,u[k]] + M[k-1,2] ):
  od:
od:
w := min( E[1,u[n]] + M[n-1,1], E[2,u[n]] + M[n-1,2] ):
print(M); print(w);
```

This code produces the table M and the minimum negative log probability $w_{\bar{\sigma},0011} = 3.088670$.

$$
\begin{aligned}
M[1,F] &= \min\{0.798508(F), 2.590267(L)\} = 0.798508\\
M[1,L] &= \min\{2.995732(F), 0.393043(L)\} = 0.393043\\
M[2,F] &= \min\{1.597015(F), 2.983310(L)\} = 1.597015\\
M[2,L] &= \min\{3.794240(F), 0.786085(L)\} = 0.786085
\end{aligned}
$$

$$\begin{aligned} M[3,F] &= \min\{2.395523(F), 4.474965(L)\} = 2.395523 \\ M[3,L] &= \min\{4.592748(F), 2.277740(L)\} = 2.277740 \\ w &= \min\{3.088670(F), 3.664034(L)\} = 3.088670. \end{aligned}$$

A corresponding Viterbi sequence $\bar{\sigma}$ can be obtained by tracing the optimal decisions made in each step the optimal decision. Here the optimal path turns out to be $M[4,F] \to M[3,F] \to M[2,F] \to M[1,F]$ giving rise to the explanation $\bar{\sigma} = FFFF$.

In view of the emission sequence $\tau = hhht$, we obtain the table

$$\begin{aligned} M[1,F] &= \min\{0.798508(F), 2.590267(L)\} = 0.798508 \\ M[1,L] &= \min\{2.995732(F), 0.393043(L)\} = 0.393043 \\ M[2,F] &= \min\{1.597015(F), 2.983310(L)\} = 1.597015 \\ M[2,L] &= \min\{3.794240(F), 0.786085(L)\} = 0.786085 \\ M[3,F] &= \min\{2.395523(F), 3.376352(L)\} = 2.395523 \\ M[3,L] &= \min\{4.592748(F), 1.179128(L)\} = 1.179128 \\ w &= \min\{3.088670(F), 2.565422(L)\} = 2.565422. \end{aligned}$$

The resulting explanation is $\bar{\sigma} = LLLL$. ♢

6.5 Expectation Maximization

The linear and toric models have the property that the likelihood function has at most one local maximum. However, this property fails for most other algebraic statistical models including those used in computational biology. In these cases, the numerical optimization technique called *expectation maximization* (EM) is widely used. It will provide under some conditions a local maximum of the likelihood function.

Consider the *hidden model* $F : \mathbb{R}^d \to \mathbb{R}^{m \times n}$ given by

$$F : (\theta_1, \ldots, \theta_d) \to (f_{ij}(\theta)) \,. \tag{6.18}$$

Assume that the sum of all the $f_{ij}(\theta)$ equals 1 and there is an open subset $\Theta \subseteq \mathbb{R}^d$ so that $f_{ij}(\theta) > 0$ for all $\theta \in \Theta$. We assume that the hidden model F has an easy and reliable algorithm for solving the maximum likelihood problem.

Consider the linear mapping that takes an $m \times n$ matrix to its vector of row sums $\rho : \mathbb{R}^{m \times n} \to \mathbb{R}^m$ given by

$$\rho : (g_{ij}) \mapsto \left(\sum_{j=1}^{n} g_{1j}, \ldots, \sum_{j=1}^{n} g_{mj} \right) . \tag{6.19}$$

The *observed model* is the composition $f = \rho \circ F : \mathbb{R}^d \to \mathbb{R}^m$ defined as

$$f : \theta \mapsto \left(\sum_{j=1}^{n} f_{1j}(\theta), \ldots, \sum_{j=1}^{n} f_{mj}(\theta) \right) . \tag{6.20}$$

We put

$$f_i(\theta) = \sum_{j=1}^{n} f_{ij}(\theta), \quad 1 \le i \le m. \tag{6.21}$$

Suppose we have a data vector $u = (u_1, \ldots, u_m) \in \mathbb{N}^m$ for the observed model. The problem is to maximize the log-likelihood function for these data with respect to the observed model,

$$\begin{aligned} &\max \ell_{\text{obs}}(\theta) = u_1 \cdot \log f_1(\theta) + \cdots + u_m \cdot \log f_m(\theta). \\ &\text{s.t. } \theta \in \Theta \end{aligned} \tag{6.22}$$

This problem is usually hard to tackle due to multiple local solutions. It would be much easier to solve the corresponding problem for the hidden model,

$$\begin{aligned} &\max \ell_{\text{hid}}(\theta) = u_{11} \cdot \log f_{11}(\theta) + \cdots + u_{mn} \cdot \log f_{mn}(\theta). \\ &\text{s.t. } \theta \in \Theta \end{aligned} \tag{6.23}$$

But here do not know the hidden data; that is, the matrix $U = (u_{ij}) \in \mathbb{N}^{m \times n}$. All we know is the marginalization data $\rho(U) = u$.

Algorithm 6.2 EM algorithm for observed model

Require: $m \times n$ matrix of polynomials $f_{ij}(\theta)$ representing the hidden model and observed data $u \in \mathbb{N}^m$
Ensure: Maximum likelihood estimate $\hat{\theta} \in \Theta$ of the log-likelihood function $\ell_{\text{obs}}(\theta)$ for observed model
[Init] Threshold $\epsilon > 0$ and parameter $\theta \in \Theta$
[E-Step] Define matrix $U = (u_{ij}) \in \mathbb{R}^{m \times n}$ with

$$u_{ij} = \frac{u_i \cdot f_{ij}(\theta)}{f_i(\theta)}$$

[M-Step] Compute solution $\theta^* \in \Theta$ of the maximation problem in the hidden model
[Comp] If $\ell_{\text{obs}}(\theta^*) - \ell_{\text{obs}}(\theta) > \epsilon$, set $\theta := \theta^*$ and resume with E-step
Output $\hat{\theta} := \theta^*$

Theorem 6.9. *During each iteration of the EM algorithm (Alg. 6.2), the value of the log-likelihood function weakly increases; that is, $\ell_{\text{obs}}(\theta^*) \ge \ell_{\text{obs}}(\theta)$. If $\ell_{\text{obs}}(\theta^*) = \ell_{\text{obs}}(\theta)$, then θ^* is a critical point of the log-likelihood function.*

Proof. We have

$$\begin{aligned} \ell_{\text{obs}}(\theta^*) - \ell_{\text{obs}}(\theta) &= \sum_{i=1}^{m} u_i \cdot [\log f_i(\theta^*) - \log f_i(\theta)] \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} u_{ij} \cdot [\log f_{ij}(\theta^*) - \log f_{ij}(\theta)] \\ &\quad + \sum_{i=1}^{m} u_i \cdot \left(\log \left(\frac{f_i(\theta^*)}{f_i(\theta)} \right) - \sum_{j=1}^{n} \frac{u_{ij}}{u_i} \cdot \log \left(\frac{f_{ij}(\theta^*)}{f_{ij}(\theta)} \right) \right). \end{aligned} \tag{6.24}$$

The first term equals $\ell_{\text{hid}}(\theta^*) - \ell_{\text{hid}}(\theta)$ and is non-negative due to the M-step. We show that the second term is also non-negative. By the E-step, the parenthesized expression equals

$$\log\left(\frac{f_i(\theta^*)}{f_i(\theta)}\right) - \sum_{j=1}^{n} \frac{u_{ij}}{u_i} \cdot \log\left(\frac{f_{ij}(\theta^*)}{f_{ij}(\theta)}\right) = \log\left(\frac{f_i(\theta^*)}{f_i(\theta)}\right) + \sum_{j=1}^{n} \frac{f_{ij}(\theta)}{f_i(\theta)} \cdot \log\left(\frac{f_{ij}(\theta)}{f_{ij}(\theta^*)}\right). \tag{6.25}$$

This expression can be rewritten as

$$\sum_{j=1}^{n} \frac{f_{ij}(\theta)}{f_i(\theta)} \cdot \log\left(\frac{f_i(\theta^*)}{f_i(\theta)}\right) + \sum_{j=1}^{n} \frac{f_{ij}(\theta)}{f_i(\theta)} \cdot \log\left(\frac{f_{ij}(\theta)}{f_{ij}(\theta^*)}\right) \tag{6.26}$$

and thus amounts to

$$\sum_{j=1}^{n} \frac{f_{ij}(\theta)}{f_i(\theta)} \cdot \log\left(\frac{f_i(\theta^*)}{f_{ij}(\theta^*)} \cdot \frac{f_{ij}(\theta)}{f_i(\theta)}\right). \tag{6.27}$$

Take the non-negative quantities

$$\pi_j = \frac{f_{ij}(\theta)}{f_i(\theta)} \quad \text{and} \quad \sigma_j = \frac{f_{ij}(\theta^*)}{f_i(\theta^*)}, \qquad 1 \le j \le n. \tag{6.28}$$

We have $\pi_1 + \ldots + \pi_n = 1 = \sigma_1 + \ldots + \sigma_n$. Thus the vectors π and σ are probability distributions on the set $[m]$. The expression (6.27) equals the *Kullback-Leibler distance* between the probability distributions π and σ,

$$\begin{aligned} H(\pi\|\sigma) &= \sum_{j=1}^{n} \pi_j \cdot \log\left(\frac{\pi_j}{\sigma_j}\right) = \sum_{j=1}^{n} (-\pi_j) \cdot \log\left(\frac{\sigma_j}{\pi_j}\right) \\ &\ge \sum_{j=1}^{n} \pi_j \cdot \left(1 - \frac{\sigma_j}{\pi_j}\right) = 0. \end{aligned} \tag{6.29}$$

where we used the inequality $\log x \le x - 1$ for all $x \in \mathbb{R}_{>0}$.

Let $\ell_{\text{obs}}(\theta^*) = \ell_{\text{obs}}(\theta)$. Then the two terms in (6.24) are both zero. Moreover, the Kullback-Leibler distance satisfies $H(\pi\|\sigma) = 0$ if and only if $\pi = \sigma$. Thus we obtain

$$\frac{f_{ij}(\theta)}{f_i(\theta)} = \frac{f_{ij}(\theta^*)}{f_i(\theta^*)}, \qquad 1 \le i \le m,\ 1 \le j \le n. \tag{6.30}$$

Therefore,

$$\begin{aligned} 0 = \frac{\partial \ell_{\text{hid}}(\theta^*)}{\partial \theta_k} &= \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{u_{ij}}{f_{ij}(\theta^*)} \cdot \frac{\partial f_{ij}(\theta^*)}{\partial \theta_k} \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{u_i}{f_i(\theta^*)} \cdot \frac{\partial f_{ij}}{\partial \theta_k}(\theta^*) = \sum_{i=1}^{m} \frac{u_i}{f_i(\theta^*)} \cdot \left(\frac{\partial}{\partial \theta_k} \sum_{j=1}^{n} f_{ij}\right)(\theta^*) \\ &= \sum_{i=1}^{m} \frac{u_i}{f_i(\theta^*)} \cdot \left(\frac{\partial}{\partial \theta_k} f_i\right)(\theta^*) = \frac{\partial \ell_{\text{obs}}(\theta^*)}{\partial \theta_k}, \quad 1 \le k \le d, \end{aligned}$$

where in the third equation we used the E-step and (6.30). □

The EM technique can be particularly used to provide maximum likelihood estimates for the hidden Markov model, because the hidden Markov model is the composition $f = \rho \cdot F$ of a fully observed toric model F and a marginalization mapping ρ. A version of the EM algorithm for the hidden Markov model is given by Alg. 6.3. In the E-step, the Viterbi algorithm can be used to compute the entities $p_{\sigma,\tau}$ and in the M-step, Prop. 6.2 is used to compute the locally maximal estimates θ^* and θ'^*.

Algorithm 6.3 EM algorithm for hidden Markov model

Require: Hidden Markov model $f : \mathbb{R}^{l\times(l-1)} \times \mathbb{R}^{l\times(l'-1)} \to \mathbb{R}^{l'^n}$ with parameter space Θ_1 and observed data $u = (u_\tau) \in \mathbb{N}^{l'^n}$

Ensure: Maximum likelihood estimate $(\hat{\theta}, \hat{\theta}') \in \Theta_1$

[Init] Threshold $\epsilon > 0$ and parameters $(\theta, \theta') \in \Theta_1$

[E-Step] Define matrix $U = (u_{\sigma,\tau}) \in \mathbb{R}^{l^n \times l'^n}$ with

$$u_{\sigma,\tau} = \frac{u_\tau \cdot p_{\sigma,\tau}(\theta, \theta')}{p_\tau(\theta, \theta')}, \quad \sigma \in \Sigma^n, \tau \in \Sigma'^n,$$

[M-Step] Compute solution $(\theta^*, \theta'^*) \in \Theta_1$ of the maximation problem in fully observed Markov model

[Comp] If $\ell(\theta^*, \theta'^*) - \ell(\theta, \theta') > \epsilon$, set $\theta := \theta^*$ and $\theta' := \theta'^*$ and resume with E-step

Output $\hat{\theta} := \theta^*$, $\hat{\theta}' := \theta^*$

6.6 Finding CpG Islands

The `CpG` sites are regions of DNA in a linear DNA strand where a cytosine is next to a guanine linked by one phosphate group. The notation "`CpG`" is used to distinguish this linear sequence from the `CG` base pairs where cytosine and guanine are on different DNA strands linked by hydrogen bonds.

Cytosines in `CpG` can be methylated to form 5-methylcytosine. In mammals, 70 % to 80 % of the `CpG` cytosines are methylated. The methylated cytosines within a gene can change the gene's expression. Gene expression is a mechanism to transcribe and translate a gene into a protein.

`CpG` islands are regions with a high frequency of `CpG` sites. An objective definition of `CpG` island is lacking. The usual formal definition is that a `CpG` island is a region with at least 200 base pairs in length, a `CG` percentage greater than 50 %, and the observed-to-expected `CpG` ratio greater than 60 %, where the observed-to-expected `CpG` ratio is given by the ratio between an observed part (i.e., number of `CpG` times length of sequence) and an expected part (i.e., number of `C` times number of `G`).

In mammals, many genes have `CpG` islands at the start of a gene (promoter regions). In mammalian genomes, `CpG` islands are usually 300 to 3,000 base pairs in length and occur in about 40 % of promoter regions. In particular, the promoter regions in human genomes have a `CpG` content of about 70 %. Over time, methylated cytosines in `CpG` sites tend to turn into thymines because of spontaneous deamination.

The methylation of `CpG` sites within promoter regions can lead to the silencing of the gene. Silencing is a phenomenon which can be found in a number of human cancers such as the silencing of tumor suppressor genes. Age has a strong impact on DNA methylation levels on tens of thousands of `CpG` sites.

In computational biology, two questions about `CpG` islands arise. First, decide whether a short stretch of a genomic linear strand lies inside of a `CpG` island. Second, find the `CgG` regions of a long stretch of a genomic linear strand.

We begin with the first question. From a set of human DNA sequences there were extracted a total of 48 putative CpG islands. From the regions labelled as CpG islands the + model was derived and from the remaining regions the − model was established. The transition probabilities for each model were calculated using the equations

$$\theta^+_{XY} = \frac{c^+_{XY}}{\sum_Z c^+_{XZ}} \quad \text{and} \quad \theta^-_{XY} = \frac{c^-_{XY}}{\sum_Z c^-_{XZ}},$$

where c^+_{XY} and c^-_{XY} are the number of times the nucleotide Y followed the nucleotide X in a CpG island and non-island, respectively. In this way, the transition probabilities of the two Markov chain models are given in Tab. 6.5.

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

−	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

Fig. 6.5. Transition probabilities for + model and − model.

Consider a DNA sequence w. We calculate in both Markov chain models the probabilities $p^+(w)$ and $p^-(w)$. For discrimination purposes, the log-odd ratio is used,

$$S(w) = \log \frac{p^+(w)}{p^-(w)} = \sum_i \log \frac{\theta^+_{w_i,w_{i+1}}}{\theta^-_{w_i,w_{i+1}}}.$$

If the value of $S(w)$ is positive, there is a high chance that the DNA sequence represents a CpG island; otherwise, it will not.

Finally, we study the second question. For this, we build a hidden Markov model for the entire DNA sequence that incorporates both of the above Markov chain models. To this end, the states are relabelled such that $\mathtt{A}_+$, $\mathtt{C}_+$, $\mathtt{G}_+$, and $\mathtt{T}_+$ determine CpG island areas, while $\mathtt{A}_-$, $\mathtt{C}_-$, $\mathtt{G}_-$, and $\mathtt{T}_-$ provide non-island areas. For simplicity, we assume that there is a uniform transition probability of switching between island and non-island. The transition probabilities are given in Tab. 6.6, where p^+ and $p^- = 1 - p^+$ are the probabilities for staying inside and outside of a CpG island.

For instance, we have $\theta_{T^+A^+} = 0.079p^+$ and $\theta_{T^+A^-} = (1 - p^+)/4$. The transition probabilities in this model can be set so that within each region they are close to the transition probabilities of the original component model, but there is also a chance of switching into the other region.

Finally, the emission probabilities are all 0 and 1. The state X^+ or X^- outputs the symbol X with certainty; that is,

$$\theta'_{X^+,Y} = \theta'_{X^-,Y} = \begin{cases} 1 \text{ if } X = Y, \\ 0 \text{ if } X \neq Y. \end{cases}$$

There are three canonical problems associated with a Hidden Markov model. First, given the parameters of the model, a DNA sequence (output), and a state sequence. Calculate the probability of

θ	$\mathtt{A}^+$	$\mathtt{C}^+$	$\mathtt{G}^+$	$\mathtt{T}^+$	$\mathtt{A}^-$	$\mathtt{C}^-$	$\mathtt{G}^-$	$\mathtt{T}^-$
$\mathtt{A}^+$	$0.180p^+$	$0.274p^+$	$0.426p^+$	$0.120p^+$	$\frac{1-p^+}{4}$	$\frac{1-p^+}{4}$	$\frac{1-p^+}{4}$	$\frac{1-p^+}{4}$
$\mathtt{C}^+$	$0.171p^+$	$0.368p^+$	$0.274p^+$	$0.188p^+$	$\frac{1-p^+}{4}$	$\frac{1-p^+}{4}$	$\frac{1-p^+}{4}$	$\frac{1-p^+}{4}$
$\mathtt{G}^+$	$0.161p^+$	$0.339p^+$	$0.375p^+$	$0.125p^+$	$\frac{1-p^+}{4}$	$\frac{1-p^+}{4}$	$\frac{1-p^+}{4}$	$\frac{1-p^+}{4}$
$\mathtt{T}^+$	$0.079p^+$	$0.355p^+$	$0.384p^+$	$0.182p^+$	$\frac{1-p^+}{4}$	$\frac{1-p^+}{4}$	$\frac{1-p^+}{4}$	$\frac{1-p^+}{4}$
$\mathtt{A}^-$	$\frac{1-p^-}{4}$	$\frac{1-p^-}{4}$	$\frac{1-p^-}{4}$	$\frac{1-p^-}{4}$	$0.300p^-$	$0.205p^-$	$0.285p^-$	$0.210p^-$
$\mathtt{C}^-$	$\frac{1-p^-}{4}$	$\frac{1-p^-}{4}$	$\frac{1-p^-}{4}$	$\frac{1-p^-}{4}$	$0.322p^-$	$0.298p^-$	$0.078p^-$	$0.302p^-$
$\mathtt{G}^-$	$\frac{1-p^-}{4}$	$\frac{1-p^-}{4}$	$\frac{1-p^-}{4}$	$\frac{1-p^-}{4}$	$0.248p^-$	$0.246p^-$	$0.298p^-$	$0.208p^-$
$\mathtt{T}^-$	$\frac{1-p^-}{4}$	$\frac{1-p^-}{4}$	$\frac{1-p^-}{4}$	$\frac{1-p^-}{4}$	$0.177p^-$	$0.239p^-$	$0.292p^-$	$0.292p^-$

Fig. 6.6. Transition probablities for hidden Markov model.

the output sequence when the model runs through the states. For instance, the DNA sequence CGCG generated by the state sequence $\mathtt{C}_+\mathtt{G}_+\mathtt{C}_+\mathtt{G}_+$ has the probability

$$p_{C_+G_+C_+G_+,CGCG} = \frac{1}{8}\theta'_{C_+,C}\theta_{C_+,G_+}\theta'_{G_+,G}\theta_{G_+,C_+}\theta'_{C_+,C}\theta_{C_+,G_+}\theta'_{G_+,G}.$$

Second, given the parameters of the model and a DNA sequence (output). Find the maximum a posteriori probability of generating the output sequence. This problem is tackled by the Viterbi algorithm that finds the most probable path. When this path goes through the + states, a CpG island is predicted. For instance, consider an output sequence and a corresponding a postoriori state sequence,

A C C C G C C G A A T A T T C G G G C C G A A T A
$\mathtt{A}^-\ \mathtt{C}^+\ \mathtt{C}^+\ \mathtt{C}^+\ \mathtt{G}^+\ \mathtt{C}^+\ \mathtt{C}^+\ \mathtt{G}^+\ \mathtt{A}^-\ \mathtt{A}^-\ \mathtt{T}^-\ \mathtt{A}^-\ \mathtt{T}^-\ \mathtt{T}^-\ \mathtt{C}^+\ \mathtt{G}^+\ \mathtt{G}^+\ \mathtt{G}^+\ \mathtt{C}^+\ \mathtt{C}^+\ \mathtt{G}^+\ \mathtt{A}^-\ \mathtt{A}^-\ \mathtt{T}^-\ \mathtt{A}^-$

The state sequence would indicate that the strand has two CpG islands.

Third, given a set of DNA sequences (output), find the most likely set of transitions probabilities. This amounts to the discovery of the parameters of the hidden Markov model given the data set. This problem can be tackled by the EM algorithm.

7

Tree Markov Models

Phylogenetics is a branch of biology that seeks to reconstruct evolutionary history. Inferring a phylogeny is an estimation procedure that is to provide the best estimate of history based on the incomplete information contained in the observed data. Ultimately, we would like to reconstruct the entire *tree of life* that describes the course of evolution leading to all present day species.

Phylogenetic reconstruction has a long history. The classical reconstruction has been based on the observation and measurement of morphological similarities between taxa with the possible adjunction of similar evidence from the fossil record. However, with the recent advances in technology for sequencing of genomic data, reconstruction based on the huge amount of available DNA sequence data and is now by far the most commonly used technique. Moreover, reconstruction from DNA sequence data can operate automatically on well-defined digital data sets that fit into the framework of classical statistics, rather than proceeding from a somewhat ill-defined mixture of qualitative and quantitative data with the need for expert oversight to adjust for difficulties such as morphological similarity.

This chapter is mainly devoted to two approaches for phylogenetic reconstruction, the maximum likelihood method that evaluates a hypothesis about evolutionary history in terms of the probability and the algebraic method of phylogenetic invariants.

7.1 Data and General Models

Phylogenetic reconstruction makes use of the structure of trees. A *tree* is a cycle-free connected graph $T = (N, E)$ with node set $N = N(T)$ and edge set $E = E(T)$.

An *unrooted tree* is considered as an undirected graph; that is, the edges are 2-subsets of the node set. Each edge $\{k, l\}$ is written as a word kl with $kl = lk$ since there is no ordering on the nodes. The edges are also called *branches*. An *unrooted binary tree* (or *trivalent tree*) contains only nodes of degree one (*terminal nodes* or *leaves*) and degree three (*internal nodes*).

A *rooted tree* is considered as a directed graph; that is, the edges are ordered pairs. Each edge (k, l) is denoted as a word kl with $kl \neq lk$ since there is an ordering on the nodes. A *rooted binary tree* contains besides nodes of degree one and three also one node of degree two, the socalled *root*. The edges are always directed away from the root.

A tree is *labelled* if its leaves are labelled. In phylogenies, trees are built from data corresponding to the leaves. These data are called *taxa*, while the data at the inner nodes are called *intermediates*. Taxa and intermediates are both called *individuals*.

In a phylogenetic tree, the arrow of time points away from the root (if any), paths down through the tree represent *lineages* (lines of descent), any point on a lineage corresponds to a point of time in the life of some ancestor of a taxon, inner nodes represent times at which lineages diverge, and the root (if any) corresponds to the most common ancestor of all the taxa.

The basic data in phylogenetics are DNA sequences corresponding one-to-one with the taxa that have been preprocessed in some suitable way. These sequences are assumed to be aligned. For simplicity, we suppose that we are dealing with segments of DNA without indels such that all taxa share the same common positions, and differences between nucleotides at these positions are due to substitutions.

Suppose we have the following four aligned DNA sequences,

taxon 1 A G A C G T T A C G T A ...
taxon 2 A G A G C A A C T T T G ...
taxon 3 A A T C G A T A C G C A ...
taxon 4 T C T A G T A A C C C C ...

A standard assumption is that the behavior at widely separated positions on the genome are statistically independent. With this assumption, the modelling problem reduces to the modelling of nucleotides observed at a given position. For this, define a *pattern* σ to be a sequence of symbols that we get when we look at a single site (column) in the aligned sequence data. For instance, the third column gives rise to the pattern AATT. A tree that might describe the course of evolution based on this pattern is shown in Fig. 7.1. Any tree with four leaves labelled by the pattern in some order is a potential candidate

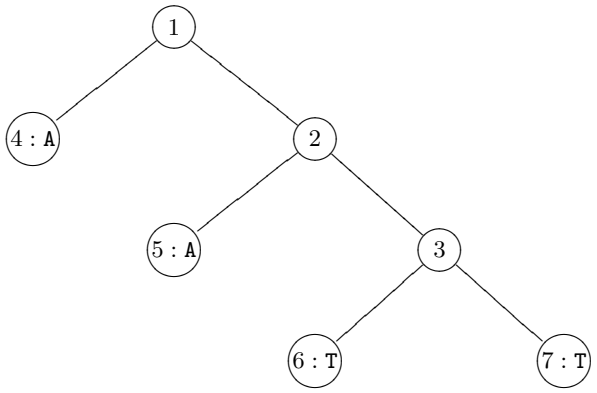


Fig. 7.1. Rooted binary tree with labelled leaves.

that describes the course of evolution. To this end, we need to capture the various tree topologies and labellings.

Two trees T and T' are *isomorphic* if there is a bijective mapping $\phi : N \to N'$ between the node sets which is compatible with the edges; that is, for each pair of nodes k, l in T, kl is an edge in T if and only if $\phi(k)\phi(l)$ is an edge in T'. The mapping ϕ is also called an *isomorphism*.

Let Σ denote the DNA alphabet (or any other alphabet used for bioinformatics data, like RNA or amino acids). Let T be a tree with set of leaves L. A *labelling* of T by DNA data is a mapping $\psi : L \to \Sigma$. A tree equipped with such a labelling is called *labelled*. Two labelled trees T and T' are *equivalent* if there is an isomorphism $\phi : N \to N'$ which is compatible with the labelling of the leaves; that is, if $\psi : L \to \Sigma$ is a labelling of T and $\psi' : L' \to \Sigma$ is a labelling of T', then $\psi(l) = \psi'(\phi(l))$ for each leaf $l \in L$.

Example 7.1. There are two non-isomorphic rooted binary trees with four leaves (Fig. 7.4). The first has twelve labelled trees (Fig. 7.2) and the second has three labelled trees (Fig. 7.3). ♢

leaves	4	5	6	7	leaves	4	5	6	7
	A	C	G	T		G	A	C	T
	A	G	C	T		G	C	A	T
	A	T	C	G		G	T	A	C
	C	A	G	T		T	A	C	G
	C	G	A	T		T	C	A	G
	C	T	A	G		T	G	A	C

Fig. 7.2. Labellings of the first rooted tree in Fig. 7.4.

leaves	4	5	6	7
	A	C	G	T
	A	G	C	T
	A	T	C	G

Fig. 7.3. Labellings of the second rooted tree in Fig. 7.4.

Proposition 7.2. *A labelled unrooted binary tree with $n \geq 3$ leaves has $n-2$ internal nodes and $2n-3$ edges. The number of inequivalent labelled unrooted binary trees with $n \geq 3$ leaves is*

$$\prod_{i=3}^{n}(2i-5).$$

Proof. Let g_n denote the number of inequivalent labelled unrooted binary trees with $n \geq 3$ leaves. There is one labelled unrooted binary tree with $n = 3$ leaves, i.e., $g_3 = 1$. This tree has $n - 2 = 1$ internal nodes and $2n - 3 = 3$ edges (Fig. 7.5).

Consider a labelled unrooted binary tree T with $n \geq 3$ leaves. Choose an edge of T, bisect it by introducing an internal node and connect this node to a new leaf (Fig. 7.6). By induction, the new tree has $n + 1$ leaves, $(n - 2) + 1 = (n + 1) - 2$ internal nodes, and $(2n - 3) + 2 = 2(n + 1) - 3$ edges. Each

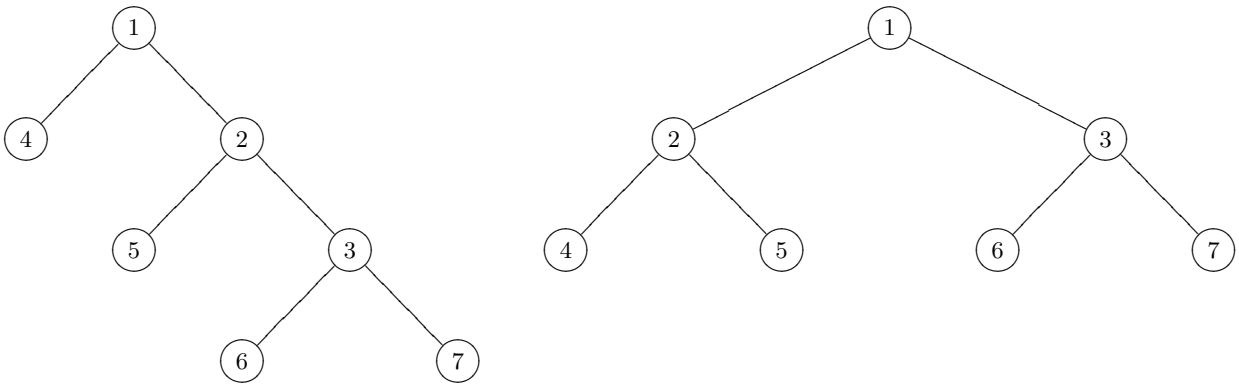


Fig. 7.4. Two rooted binary trees with four leaves.

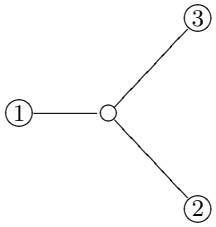


Fig. 7.5. Labelled unrooted binary tree with three leaves.

labelled unrooted binary tree with $n+1$ leaves can be constructed in this way and the constructed trees are all pairwise inequivalent. Since each labelled unrooted binary tree with n leaves has $2n-3$ nodes, we have $g_{n+1} = g_n \cdot (2n-3) = g_n \cdot (2(n+1)-5)$ as required. □

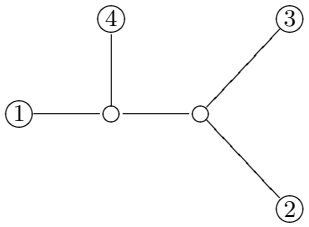


Fig. 7.6. Labelled unrooted binary tree with four leaves.

Proposition 7.3. *A labelled rooted binary tree with $n \geq 2$ leaves has $n-1$ internal nodes and $2n-2$ edges. The number of inequivalent labelled rooted binary trees with $n \geq 2$ leaves is*

$$\prod_{i=2}^{n}(2i-3).$$

Proof. Let h_n denote the number of inequivalent labelled rooted binary trees with $n \geq 2$ leaves. There is one labelled rooted binary tree with $n = 2$ leaves, i.e., $h_2 = 1$. This tree has $n - 1 = 1$ internal nodes and $2n - 2 = 2$ edges (Fig. 7.7).

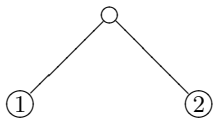


Fig. 7.7. Labelled rooted binary tree with two leaves.

Each labelled unrooted binary tree T with $n \geq 3$ leaves can be converted into a labelled rooted binary tree with n by bisecting one of its edges and taking this new node as the root (Fig. 7.8). By Prop. 7.2, each such constructed tree has $(n-2)+1 = n-1$ internal nodes and $(2n-3)+1 = 2n-2$ edges. Each labelled rooted binary tree with $n+1$ can be constructed in this way and the constructed trees are pairwise inequivalent. Since the number of labelled unrooted binary trees with n leaves is g_n and each such tree has $2n-3$ edges, the number of labelled rooted binary trees with n leaves is $h_n = g_n \cdot (2n-3) = \prod_{i=2}^{n-1}(2i-3)(2n-3) = \prod_{i=2}^{n}(2i-3)$ as required. □

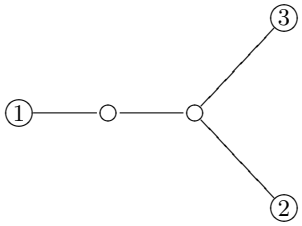


Fig. 7.8. Labelled rooted binary tree with three leaves.

The proofs are constructive and provide a method for generating both all labelled unrooted binary trees and all rooted binary trees with a small number of leaves.

7.2 Fully Observed Tree Markov Model

We assume that not only the nucleotides for the taxa can be observed but also the nucleotides for the intermediates represented by the interior nodes of a phylogenetic tree. Two individuals in a phylogenetic tree share the same lineage up to their most recent common ancestor. After the split in lineages, it is a reasonable first approximation to assume that the random mechanism by which substitutions occur

are operating independently on the genomes that are no longer shared. Equivalently, the nucleotides exhibited by the two individuals are conditionally independent of the corresponding nucleotide exhibited by their most recent common ancestor.

Example 7.4. Consider the four taxa tree in Fig. 7.1. Let X_i denote the random variable over the DNA alphabet corresponding to the individual i, $1 \leq i \leq 7$. In view of the dependence structure given by the tree, a joint probability such as

$$P(X_1 = \mathtt{A}, X_2 = \mathtt{G}, X_3 = \mathtt{T}, X_4 = \mathtt{A}, X_5 = \mathtt{A}, X_6 = \mathtt{T}, X_7 = \mathtt{T})$$

can be computed as

$$\begin{aligned} &P(X_1 = \mathtt{A}) \\ &\cdot P(X_4 = \mathtt{A} \mid X_1 = \mathtt{A}) P(X_2 = \mathtt{G} \mid X_1 = \mathtt{A}) \\ &\cdot P(X_5 = \mathtt{A} \mid X_2 = \mathtt{G}) P(X_3 = \mathtt{T} \mid X_2 = \mathtt{G}) \\ &\cdot P(X_6 = \mathtt{T} \mid X_3 = \mathtt{T}) P(X_7 = \mathtt{T} \mid X_3 = \mathtt{T}). \end{aligned} \tag{7.1}$$

Thus, for a given tree, the joint probabilities of the individuals exhibiting a particular set of nucleotides are determined by the probability distribution of the root and the transition probabilities corresponding to the edges. ♢

We describe formally the tree model. For this, let T be a rooted binary tree with root r. Each node $i \in N(T)$ corresponds to a random variable X_i with values in an alphabet Σ_i. Each edge $kl \in E(T)$ is associated to a matrix $\boldsymbol{\theta}^{kl}$ with positive entries, where the rows are indexed by Σ_k and the columns are indexed by Σ_l. The parameter space Θ of the tree model is given by the collection of the matrices $\boldsymbol{\theta}^{kl}$ with $kl \in E(T)$. Thus the dimension of the parameter space is $d = \sum_{kl \in E(T)} |\Sigma_k| \cdot |\Sigma_l|$. A state σ of the tree model is a labelling of the nodes of tree; i.e., $\sigma = (\sigma_i)_{i \in N(T)}$, where $\sigma_i \in \Sigma_i$. The state space of the model is the Cartesian product of the alphabets Σ_i with $i \in N(T)$. Thus the cardinality of the state set is $m = \prod_{i \in N(T)} |\Sigma_i|$.

The *fully observed toric tree Markov model* for the tree T is the mapping $F_T : \mathbb{R}^d \to \mathbb{R}^m$ defined as

$$F_T : \boldsymbol{\theta} = (\boldsymbol{\theta}^{kl})_{kl \in E(T)} \mapsto p = (p_\sigma), \tag{7.2}$$

where

$$p_\sigma = \frac{1}{|\Sigma_r|} \cdot \prod_{kl \in E(T)} \theta^{kl}_{\sigma_k \sigma_l} \tag{7.3}$$

for each state $\sigma = (\sigma_i)_{i \in N(T)}$. Note that the factor $1/|\Sigma_r|$ corresponds to a uniform distribution of states at the root (Ex. 7.4).

We consider a subset Θ_1 of the parameter space Θ given by all positive matrices $\boldsymbol{\theta}^{kl}$ whose row sums are 1. The matrices $\boldsymbol{\theta}^{kl}$ can then be viewed as transition probability matrices along the branches. The dimension of the parameter space Θ_1 is therefore $d = \sum_{kl \in E(T)} |\Sigma_k| \cdot (|\Sigma_l| - 1)$. The *fully observed tree Markov model* for the tree T is given by the restriction of the mapping $F_T : \mathbb{R}^d \to \mathbb{R}^m$ to the parameter space Θ_1.

Tree models in phylogenetics have usually the same alphabet Σ for the edges, but the transition probability matrices remain distinct and independent.

Example 7.5. Consider the *1,n claw tree* T in Fig. 7.9. This is a tree with no internal nodes other than the root r and n leaves; that is, $N(T) = \{r, 1, \ldots, n\}$ and $E(T) = \{r1, \ldots, rn\}$.

Assume that all nodes have the alphabet $\Sigma = \{0, 1\}$. The fully observed toric tree Markov model F_T has $d = 4 \cdot n$ parameters given by the 2×2 matrices

$$\theta^{ri} = \begin{pmatrix} \theta^{ri}_{00} & \theta^{ri}_{01} \\ \theta^{ri}_{10} & \theta^{ri}_{11} \end{pmatrix}, \qquad 1 \leq i \leq n.$$

The model has $m = 2^{n+1}$ states which are given by the binary strings $\sigma_r\sigma_1 \ldots \sigma_n \in \Sigma^{n+1}$.

The fully observed tree Markov model F_T has $d = 2 \cdot n$ parameters defined by the 2×2 matrices

$$\theta^{ri} = \begin{pmatrix} \theta^{ri}_{00} & 1 - \theta^{ri}_{00} \\ \theta^{ri}_{10} & 1 - \theta^{ri}_{10} \end{pmatrix}, \qquad 1 \leq i \leq n.$$

The coordinates of the mapping $F_T : \mathbb{R}^{2n} \to \mathbb{R}^{2^{n+1}}$ are the marginal probabilities

$$p_{\sigma_r\sigma_1 \ldots \sigma_n} = \frac{1}{2}\theta^{r1}_{\sigma_r\sigma_1} \cdots \theta^{rn}_{\sigma_r\sigma_n}.$$

♢

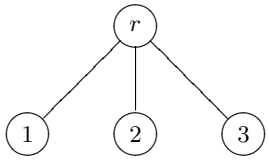


Fig. 7.9. The 1,3 claw tree.

We provide maximum likelihood estimates for the fully observed tree Markov model. For this, let T be a rooted binary tree T with n leaves. Note that the tree has $2n-1$ nodes. Assume that all individuals share a common alphabet Σ of cardinality q. Given a sequence of observations $\sigma^1, \sigma^2, \ldots, \sigma^N$ in Σ^{2n-1}. In the corresponding data vector $u = (u_\sigma)$, the entry u_σ provides the number of occurrences of the state $\sigma \in \Sigma^{2n-1}$. Thus we have $\sum_\sigma u_\sigma = N$. Let v^{kl}_{ij} denote the number of occurrences of $ij \in \Sigma^2$ as a consecutive pair for the edge kl in the tree (Fig. 7.10). The vector $v = (v_{ij})$ provides the sufficient statistic of the model (Prop. 4.11).

Proposition 7.6. *The maximum likelihood estimate of the data vector u in the fully observed tree Markov model is the parameter vector $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}^{kl}_{ij})$ in Θ_1 with coordinates*

$$\widehat{\theta}^{kl}_{ij} = \frac{v^{kl}_{ij}}{\sum_{s \in \Sigma} v^{kl}_{is}}, \quad kl \in E(T),\, ij \in \Sigma^2.$$

Proof. The log-likelihood function of the toric model can be written as follows,

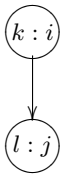


Fig. 7.10. Labelled edge $k \to l$.

$$\ell(\boldsymbol{\theta}) = \sum_{kl} \sum_{i} \left(v_{i1}^{kl} \log \theta_{i1}^{kl} + \ldots + v_{iq}^{kl} \log \theta_{iq}^{kl} \right).$$

The log-likelihood function of the fully observed tree Markov model is obtained by restriction to the set Θ_1 of positive matrices whose row sums are all equal to one. Therefore, $\ell(\boldsymbol{\theta})$ is the sum of expressions

$$v_{i1}^{kl} \log \theta_{i1}^{kl} + \ldots + v_{i,q-1}^{kl} \log \theta_{i,q-1}^{kl} + v_{iq}^{kl} \log \left(1 - \sum_{s=1}^{q-1} \theta_{is}^{kl} \right).$$

These expressions have disjoint sets of unknowns for different sets of the indices k, l, and i. To maximize $\ell(\boldsymbol{\theta})$ over Θ_1, it is sufficient to maximize the above concave functions over a $(q-1)$-dimensional simplex consisting of all non-negative vectors $(\theta_{it}^{kl})_t$ of coordinates summing to 1. By equating the partial derivatives of these expressions to zero, the unique critial point has the required coordinates. □

Example 7.7 (Maple). Consider the rooted binary tree T in Fig. 7.11. Assume that each node is associated with the binary alphabet $\Sigma = \{1, 2\}$. For this, we initialize

```
> restart: with(combinat): with(linalg):
> V12 := array(1..2,1..2);
> V13 := array(1..2,1..2);
> V24 := array(1..2,1..2);
> V25 := array(1..2,1..2);
> V36 := array(1..2,1..2);
> V37 := array(1..2,1..2);
> T12 := array([[0,0],[0,0]]);
> T13 := array([[0,0],[0,0]]);
> T24 := array([[0,0],[0,0]]);
> T25 := array([[0,0],[0,0]]);
> T36 := array([[0,0],[0,0]]);
> T37 := array([[0,0],[0,0]]);
```

generate 128 states $\sigma \in \Sigma^7$ uniformly at random,

```
digs := rand(1..2): M := randmatrix(128, 7, entries = digs):
```

and calculate the sufficient statistic

```
> for i from 1 to 128 do
>   T12[M[i,1],M[i,2]] := T12[M[i,1],M[i,2]] + 1;
```

```
>   T13[M[i,1],M[i,3]] := T13[M[i,1],M[i,3]] + 1;
>   T24[M[i,2],M[i,4]] := T24[M[i,2],M[i,4]] + 1;
>   T25[M[i,2],M[i,5]] := T25[M[i,2],M[i,5]] + 1;
>   T36[M[i,3],M[i,6]] := T36[M[i,3],M[i,6]] + 1;
>   T37[M[i,3],M[i,7]] := T37[M[i,3],M[i,7]] + 1;
> od:
> print(T12,T13,T24,T25,T36,T37);
```

$$\begin{bmatrix} 35 & 37 \\ 26 & 30 \end{bmatrix}, \begin{bmatrix} 37 & 35 \\ 28 & 28 \end{bmatrix}, \begin{bmatrix} 37 & 28 \\ 28 & 35 \end{bmatrix}, \begin{bmatrix} 23 & 42 \\ 34 & 29 \end{bmatrix}, \begin{bmatrix} 30 & 27 \\ 43 & 28 \end{bmatrix}, \begin{bmatrix} 28 & 29 \\ 34 & 37 \end{bmatrix}.$$

This allows to estimate the parameters according to Prop. 7.6,

```
V12 := array([[T12[1,1]/(T12[1,1]+T12[1,2]),
               T12[1,2]/(T12[1,1]+T12[1,2])],
              [T12[2,1]/(T12[2,1]+T12[2,2]),
               T12[2,2]/(T12[2,1]+T12[2,2])]]);
V13 := array([[T13[1,1]/(T13[1,1]+T13[1,2]),
               T13[1,2]/(T13[1,1]+T13[1,2])],
              [T13[2,1]/(T13[2,1]+T13[2,2]),
               T13[2,2]/(T13[2,1]+T13[2,2])]]);
V24 := array([[T24[1,1]/(T24[1,1]+T24[1,2]),
               T24[1,2]/(T24[1,1]+T24[1,2])],
              [T24[2,1]/(T24[2,1]+T24[2,2]),
               T24[2,2]/(T24[2,1]+T24[2,2])]]);
V25 := array([[T25[1,1]/(T25[1,1]+T25[1,2]),
               T25[1,2]/(T25[1,1]+T25[1,2])],
              [T25[2,1]/(T25[2,1]+T25[2,2]),
               T25[2,2]/(T25[2,1]+T25[2,2])]]);
V36 := array([[T36[1,1]/(T36[1,1]+T36[1,2]),
               T36[1,2]/(T36[1,1]+T36[1,2])],
              [T36[2,1]/(T36[2,1]+T36[2,2]),
               T36[2,2]/(T36[2,1]+T36[2,2])]]);
V37 := array([[T37[1,1]/(T37[1,1]+T37[1,2]),
               T37[1,2]/(T37[1,1]+T37[1,2])],
              [T37[2,1]/(T37[2,1]+T37[2,2]),
               T37[2,2]/(T37[2,1]+T37[2,2])]]);
> print(V12,V13,V24,V25,V36,V37);
```

The parameter estimates are

$$\begin{bmatrix} \frac{35}{72} & \frac{37}{72} \\ \frac{13}{28} & \frac{15}{28} \end{bmatrix}, \begin{bmatrix} \frac{37}{72} & \frac{35}{72} \\ \frac{1}{2} & \frac{14}{27} \end{bmatrix}, \begin{bmatrix} \frac{37}{65} & \frac{28}{65} \\ \frac{4}{9} & \frac{5}{9} \end{bmatrix}, \begin{bmatrix} \frac{23}{65} & \frac{42}{65} \\ \frac{34}{63} & \frac{29}{63} \end{bmatrix}, \begin{bmatrix} \frac{10}{19} & \frac{9}{19} \\ \frac{43}{71} & \frac{28}{71} \end{bmatrix}, \begin{bmatrix} \frac{28}{57} & \frac{29}{57} \\ \frac{34}{71} & \frac{37}{71} \end{bmatrix}.$$

♢

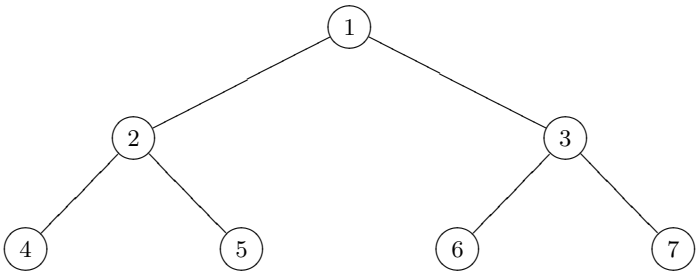


Fig. 7.11. Rooted tree with four leaves.

7.3 Hidden Tree Markov Model

A fully observed Markov tree model exhibits nucleotides of all involved individuals (taxa and intermediates). By taking the marginal probability distribution for the taxa, we obtain the corresponding hidden tree Markov model in which nucleotides are only exhibited by the taxa.

Example 7.8. Reconsider the tree with four taxa (Fig. 7.1) in Ex. 7.4. The joint probability distribution for the taxa is given by

$$P(X_4 = \mathtt{A}, X_5 = \mathtt{A}, X_6 = \mathtt{T}, X_7 = \mathtt{T}) = \\ \sum_{B_1} \sum_{B_2} \sum_{B_3} P(X_1 = B_1, X_2 = B_2, X_3 = B_3, X_4 = \mathtt{A}, X_5 = \mathtt{A}, X_6 = \mathtt{T}, X_7 = \mathtt{T}),$$

where the sums are taken over all nucleotides for the intermediates. ♢

We formally describe the hidden tree Markov model. For this, let T be a rooted binary tree with root r and leaf set $[n]$. The *hidden tree Markov model* for the tree T is obtained from the fully observed tree Markov model F_T by summing the marginal probabilities over the internal nodes of the tree. The parameter space remains the same. However, the state space is $\Sigma_1 \times \Sigma_2 \times \ldots \times \Sigma_n$, the product of the alphabets associated with the leaves of T. Thus the state space has the cardinality $m' = |\Sigma_1| \cdot |\Sigma_2| \cdots |\Sigma_n|$. The restriction of the fully observed to the hidden model is defined formally by the marginalization mapping $\rho_T : \mathbb{R}^m \to \mathbb{R}^{m'}$ which takes real-valued functions on $\prod_{i \in N(T)} \Sigma_i$ to real-valued functions on $\prod_{i=1}^n \Sigma_i$ such that the hidden tree Markov model is given by the marginalization of the fully observed tree model $f_T = \rho_T \circ F_T$.

Note that the EM technique can be used to provide maximum likelihood estimates for the hidden tree Markov model. For this, Prop. 7.6 can be employed to yield maximum likelihood estimates for the corresponding fully observed tree Markov model. Note that the hidden Markov model is a specialization of the hidden tree Markov model in which the tree is the *caterpillar tree* (Fig. 6.2).

Example 7.9 (Singular). Reconsider the $1, n$ claw tree T in Ex. 7.5. The corresponding hidden tree Markov model has $d = 2 \cdot n$ parameters and $m = 2^n$ states given by the binary strings $\sigma_1 \ldots \sigma_n \in \Sigma^n$ whose letters correspond one-to-one with the taxa. The coordinates of the mapping $f_T : \mathbb{R}^{2n} \to \mathbb{R}^{2^n}$ are the marginal probabilities

$$p_{\sigma_1 \dots \sigma_n} = \frac{1}{2}\theta^{r1}_{0\sigma_1} \cdots \theta^{rn}_{0\sigma_n} + \frac{1}{2}\theta^{r1}_{1\sigma_1} \cdots \theta^{rn}_{1\sigma_2}.$$

The corresponding EM algorithm is illustrated in (7.1). Note that the sufficient statistic $v \in \mathbb{Z}^{4n}_{\geq 0}$ can

Algorithm 7.1 EM algorithm for Markov model given by the $1, n$ claw tree.

Require: Hidden $1, n$ claw tree Markov model $f : \mathbb{R}^{2n} \to \mathbb{R}^{2^n}$ with parameter space Θ_1 and observed data $u = (u_\sigma) \in \Sigma^n$, $\Sigma = \{0, 1\}$ binary alphabet
Ensure: Maximum likelihood estimate $\hat{\theta} \in \Theta_1$
[Init] Threshold $\epsilon > 0$ and parameters $\theta \in \Theta_1$
[E-Step] Define matrix $U = (u_{\sigma_r,\sigma})_{\sigma_r \in \Sigma, \sigma \in \Sigma^n}$ with

$$u_{\sigma_r,\sigma} = \frac{u_\sigma \cdot p_{\sigma_r,\sigma}(\theta)}{p_\sigma(\theta)}, \quad \sigma_r \in \Sigma, \sigma \in \Sigma^n$$

[M-Step] Compute solution $\theta^* \in \Theta_1$ of the maximation problem in fully observed $1, n$ claw tree Markov model
[Comp] If $\ell(\theta^*) - \ell(\theta) > \epsilon$, set $\theta := \theta^*$ and resume with E-step
Output $\hat{\theta} := \theta^*$.

be established from the data $U \in \mathbb{Z}^{2^{n+1}}_{\geq 0}$ in the E-step by a linear transformation $v = AU$, where the matrix $A \in \mathbb{Z}^{4n \times 2^{n+1}}_{\geq 0}$ is given as follows ($n = 3$),

	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
θ^{r1}_{00}	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
θ^{r1}_{01}	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
θ^{r1}_{10}	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
θ^{r1}_{11}	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
θ^{r2}_{00}	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0
θ^{r2}_{01}	0	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0
θ^{r2}_{10}	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0
θ^{r2}_{11}	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1
θ^{r3}_{00}	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0
θ^{r3}_{01}	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0
θ^{r3}_{10}	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0
θ^{r3}_{11}	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1

Phylogenetic invariants of the $1, 3$ claw tree with common binary alphabet $\Sigma = \{0, 1\}$ are calculated as follows.

```
> ring r = 0, (x(1..3),y(1..3),p(1..8)), dp;
# x(i) encodes theta_00^ri, y(i) encodes theta_10^ri, i=1,2,3
# 1-x(i) encodes theta_01^ri, 1-y(i) encodes theta_11^ri, i=1,2,3
> ideal i0 = p(1) - (x(1)*x(2)*x(3)+y(1)*y(2)*y(3)); # 000
> ideal i1 = p(2) - (x(1)*x(2)*(1-x(3))+y(1)*y(2)*(1-y(3))); # 001
> ideal i2 = p(3) - (x(1)*(1-x(2))*x(3)+y(1)*(1-y(2))*y(3)); # 010
> ideal i3 = p(4) - (x(1)*(1-x(2))*(1-x(3))+y(1)*(1-y(2))*(1-y(3))); # 011
```

```
> ideal i4 = p(5) - ((1-x(1))*x(2)*x(3)+(1-y(1))*y(2)*y(3)); # 100
> ideal i5 = p(6) - ((1-x(1))*x(2)*(1-x(3))+(1-y(1))*y(2)*(1-y(3))); # 101
> ideal i6 = p(7) - ((1-x(1))*(1-x(2))*x(3)+(1-y(1))*(1-y(2))*y(3)); # 110
> ideal i7 = p(8) - ((1-x(1))*(1-x(2))*(1-x(3))+(1-y(1))*(1-y(2))*(1-y(3))); # 111
> ideal i = i1+i2+i3+i4+i5+i6+i7+i8;
> ideal j = std(i);
> ideal k = eliminate( j, x(1)*x(2)x(3)*y(1)*y(2)*y(3) );
```

The output yields two phylogenetic invariants in the ring $\mathbb{R}[p_1, \ldots, p_8]$ one of which being $p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7 + p_8 - 2$ and the other being rather large. ♢

Example 7.10 (Maple). Consider the fully observed tree Markov model for the tree T in Fig. 7.11. This model has $d = 2 \cdot 6 = 12$ parameters and $m = 2^4 = 16$ states. The coordinates of the associated mapping $f_T : \mathbb{R}^{12} \to \mathbb{R}^{16}$ are the marginal probabilities

$$p_{\sigma_4\sigma_5\sigma_6\sigma_7} = \frac{1}{2} \sum_{\sigma_1 \in \Sigma} \sum_{\sigma_2 \in \Sigma} \sum_{\sigma_3 \in \Sigma} \theta^{12}_{\sigma_1\sigma_2} \theta^{13}_{\sigma_1\sigma_3} \theta^{24}_{\sigma_2\sigma_4} \theta^{25}_{\sigma_2\sigma_5} \theta^{36}_{\sigma_3\sigma_6} \theta^{37}_{\sigma_3\sigma_7}.$$

These probabilities can be computed as follows,

```
P := vector(16):
R := powerset([1,2,3,4,5,6,7]):
for i from 1 to nops(R) do
  r := vector( [1,1,1,1,1,1,1] ):
  for j from 1 to nops(R[i]) do
    r[R[i,j]] := 2
  od
  k := 1+(r[4]-1)+(r[5]-1)*2+(r[6]-1)*2^2+(r[7]-1)*2^3:
  P[k] := 1/2*T12[r[1],r[2]]*T13[r[1],r[3]]*T24[r[2],r[4]]
          *T25[r[2],r[5]]*T36[r[3],r[6]]*T37[r[3],r[7]];
od:
```

♢

7.4 Sum-Product Decomposition

We show that the marginal probabilities of the hidden tree Markov model can be efficiently computed by a sum-product decomposition. For this, consider a rooted binary tree T with n leaves. The probability of occurrence of the pattern $\sigma \in \Sigma^n$ labelling the leaves of the tree is given by

$$p_\sigma = \sum_\tau p_{\tau,\sigma}, \tag{7.4}$$

where τ runs over all states of the internal nodes of the tree and $p_{\tau,\sigma}$ is the probability that the tree is decorated by the state (τ, σ) where τ labels the interior nodes and σ labels the leaves.

Let r be the root of the tree T and let a and b denote the decendents of the root. The marginal probability of the pattern σ then decomposes into a nested sum

$$p_\sigma = \frac{1}{|\Sigma_r|} \sum_{\tau_r} \left[\left(\sum_{\tau_a} \theta^{ra}_{\tau_r \tau_a} p^a_{\sigma_1 \dots \sigma_s} \right) \cdot \left(\sum_{\tau_b} \theta^{rb}_{\tau_r \tau_b} p^b_{\sigma_{s+1} \dots \sigma_n} \right) \right], \tag{7.5}$$

where $p^a_{\sigma_1 \dots \sigma_s}$ and $p^b_{\sigma_{s+1} \dots \sigma_n}$ denote the marginal probabilities of the subtrees of T with roots a and b and decorations of the leaves given by $\sigma, \dots, \sigma_s$ and $\sigma_{s+1}, \dots, \sigma_n$, respectively (Fig. 7.12).

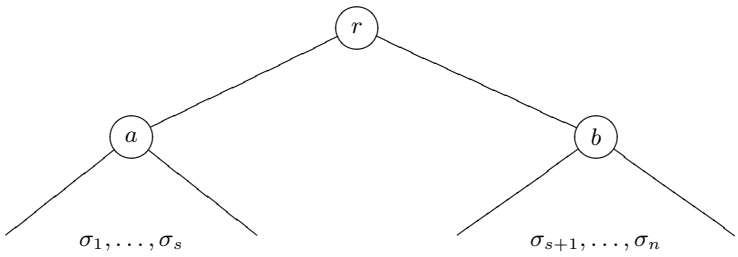


Fig. 7.12. Decomposition of a rooted binary tree into two subtrees.

Example 7.11. Reconsider the hidden tree Markov model in Ex. 7.10. The marginal probabilities are given by

$$p_{\sigma_1 \sigma_2 \sigma_3 \sigma_4} = \frac{1}{2} \cdot \sum_{\tau_1 \in \Sigma} \sum_{\tau_2 \in \Sigma} \sum_{\tau_3 \in \Sigma} \theta^{12}_{\tau_1 \tau_2} \theta^{13}_{\tau_1 \tau_3} \theta^{24}_{\tau_2 \sigma_1} \theta^{25}_{\tau_2 \sigma_2} \theta^{36}_{\tau_3 \sigma_3} \theta^{37}_{\tau_3 \sigma_4}$$

whose sum-product decomposition amounts to

$$\begin{aligned} p_{\sigma_1 \sigma_2 \sigma_3 \sigma_4} &= \sum_{\tau_1} \left[\left(\sum_{\tau_2 \in \Sigma} \theta^{12}_{\tau_1 \tau_2} p^2_{\sigma_1 \sigma_2} \right) \cdot \left(\sum_{\tau_3 \in \Sigma} \theta^{13}_{\tau_1 \tau_3} p^3_{\sigma_3 \sigma_4} \right) \right] \\ &= \frac{1}{2} \cdot \sum_{\tau_1 \in \Sigma} \left[\left(\sum_{\tau_2 \in \Sigma} \theta^{12}_{\tau_1 \tau_2} \left(\theta^{24}_{\tau_2 \sigma_1} \theta^{25}_{\tau_2 \sigma_3} \right) \right) \cdot \left(\sum_{\tau_3 \in \Sigma} \theta^{13}_{\tau_1 \tau_3} \left(\theta^{36}_{\tau_3 \sigma_3} \theta^{37}_{\tau_3 \sigma_4} \right) \right) \right]. \end{aligned} \tag{7.6}$$

♢

7.5 Felsenstein Algorithm

In the early 1980s, Felsenstein was the first who brought the maximum likelihood framework to nucleotide-based phylogenetic inference.

The Felsenstein algorithm provides an explanation for a sequence of observed data $\sigma \in \Sigma^n$ of a tree T with n leaves. Finding an explanation means identifying the hidden data $\bar{\tau}$ with maximum a posteriori probability that generated the observed data σ; that is,

$$\bar{\tau} = \operatorname{argmax}_\tau \{p_{\tau,\sigma}\}. \tag{7.7}$$

By putting $w_{\tau,\sigma} = -\log p_{\tau,\sigma}$, this equation becomes

$$\bar{\tau} = \operatorname{argmin}_{\tau}\{w_{\tau,\sigma}\}. \tag{7.8}$$

The sequence $\bar{\tau}$ provides a labelling of the internal nodes and is the solution of the marginalization (7.4) performed in the tropical algebra,

$$w_\sigma = \bigoplus_{\tau} w_{\sigma,\tau}. \tag{7.9}$$

The value w_σ can be efficiently computed by tropicalization of the sum-product decomposition of the marginal probability p_σ. For this, we put $w_{ij}^{kl} = -\log \theta_{ij}^{kl}$ for each parameter and $w_{\sigma'}^{a} = -\log p_{\sigma'}^{a}$ for each tree node a and each sequence σ' decorating the leaves of the subtree of T with root a. Then the sum-product decomposition (7.5) performed in the tropical algebra becomes

$$w_\sigma = \bigoplus_{\tau_r} \left[\left(\bigoplus_{\tau_a} w_{\tau_r\tau_a}^{ra} \odot w_{\sigma_1\ldots\sigma_s}^{a} \right) \odot \left(\bigoplus_{\tau_b} w_{\tau_r\tau_b}^{rb} \odot w_{\sigma_{s+1}\ldots\sigma_n}^{b} \right) \right]. \tag{7.10}$$

The computation of this expression is known as *Felsenstein algorithm*. It consists of a forward algorithm that evaluates the expression w_σ (Alg. 7.2) and a backward algorithm that traces back the optimal decisions made in each step; that is, the symbols for which the minimum is attained. The computed explanation $\bar{\tau}$ is called *Felsenstein sequence*.

Proposition 7.12. *The tropicalization of the marginal probability p_σ, $\sigma \in \prod_{i=1}^{n} \Sigma_i$, of the hidden tree model provides an explanation for the observed data σ in $O(l^2 n)$ steps, where l is an upper bound on the alphabet size.*

Proof. The computation can be carried out by using an $(2n-1) \times l$ table M, where

$$w_\sigma = \bigoplus_{\tau} M[r,\tau],$$

$$M[r,\tau] = \left(\bigoplus_{\tau_a} w_{\tau\tau_a}^{ra} \odot M[a,\tau_a] \right) \odot \left(\bigoplus_{\tau_b} w_{\tau\tau_b}^{rb} \odot M[b,\tau_b] \right), \quad \tau \in \Sigma,$$

and

$$M[\ell,\tau] = 0 \quad \text{for each leaf } \ell,\ \tau \in \Sigma.$$

The table M has size $O(ln)$ and each entry is evaluated in $O(l)$ time steps. □

Example 7.13. Reconsider the hidden tree model in Ex. 7.11. The tropicalization of the marginal probability (7.6) gives

$$\begin{aligned} w_{\sigma_1\sigma_2\sigma_3\sigma_4} &= \bigoplus_{\tau_1} \left[\left(\bigoplus_{\tau_2 \in \Sigma} w_{\tau_1\tau_2}^{12} \odot w_{\sigma_1\sigma_2}^{2} \right) \odot \left(\bigoplus_{\tau_3 \in \Sigma} w_{\tau_1\tau_3}^{13} \odot w_{\sigma_3\sigma_4}^{3} \right) \right] \\ &= \bigoplus_{\tau_1 \in \Sigma} \left[\left(\bigoplus_{\tau_2 \in \Sigma} w_{\tau_1\tau_2}^{12} \left(w_{\tau_2\sigma_1}^{24} \odot w_{\tau_2\sigma_3}^{25} \right) \right) \odot \left(\bigoplus_{\tau_3 \in \Sigma} w_{\tau_1\tau_3}^{13} \left(w_{\tau_3\sigma_3}^{36} \odot w_{\tau_3\sigma_4}^{37} \right) \right) \right]. \end{aligned} \tag{7.11}$$

♢

Algorithm 7.2 Felsenstein forward algorithm.

Require: Rooted binary tree T with leaf set $[n]$ and root r, observed sequence $\sigma \in \Sigma^n$, Σ common node alphabet
Ensure: Evaluation of table M
  for each leaf ℓ in T **do**
    $M[\ell, \tau] \leftarrow 0$
    mark node ℓ
  end for
  repeat
    take unmarked node v in T whose descendents a, b are already marked
    for each symbol τ in Σ **do**
      $M[v, \tau] \leftarrow \left(\bigoplus_{\tau_a} w^{va}_{\tau\tau_a} \odot M[a, \tau_a]\right) \odot \left(\bigoplus_{\tau_b} w^{vb}_{\tau\tau_b} \odot M[b, \tau_b]\right)$
    end for
    mark node v
  until root r is marked

One problem with the maximum likelihood approach is *model selection*. Suppose we have n taxa, d parameters, m states, and a vector $u \in \mathbb{N}^m$ for observed data. Then we may consider all rooted binary trees with n leaves. Each such tree leads to an algebraic statistical model $f : \mathbb{R}^d \to \mathbb{R}^m$. The task is then to select a "good" model for the data; that is, a model whose likelihood function attains the largest value. However, the number of trees grows exponentially with the number of taxa and henceforth this approach is only viable for a handful of taxa.

7.6 Evolutionary Models

The above general model for the observed nucleotides allows the substitution matrices to be arbitrary. However, there are practical reasons for constraining the form of these matrices. This leads to evolutionary models that are described by time-continuous Markov chains.

A time-continuous Markov chain is based on a matrix of rates which describes the substitution of nucleotides at an infinitesimally small time interval. A *rate matrix* is an $n \times n$ real-valued matrix $\boldsymbol{Q} = (q_{ij})$ with rows and columns indexed by a common alphabet Σ such as the DNA alphabet $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ that satisfies the following conditions,

- all off-diagonal entries are non-negative,
$$q_{ij} \geq 0, \quad i \neq j,$$
- all row sums are zero,
$$\sum_{j \in \Sigma} q_{ij} = 0, \quad i \in \Sigma,$$
- all diagonal entries are negative,
$$q_{ii} < 0, \quad i \in \Sigma.$$

A rate matrix is an *infinitesimal generator matrix* for a time-continuous Markov chain capturing the notion of instantaneous rate of mutation.

A rate matrix $\boldsymbol{Q}$ gives rise to a probability distribution matrix $\Phi(t)$, $t \geq 0$, by exponentiation. The *matrix exponential* for the matrix $\boldsymbol{Q}t$ is the $n \times n$ real-valued matrix

$$\boldsymbol{\Phi}(t) = \exp(\boldsymbol{Q}t) = \sum_{k=0}^{\infty} \frac{1}{k!} \boldsymbol{Q}^k t^k, \quad t \geq 0. \tag{7.12}$$

The entry of $\boldsymbol{\Phi}(t)$ in row i and column j equals the conditional probability that the substitution $i \to j$ occurs at a time interval of length $t \geq 0$.

Theorem 7.14. *Each rate matrix $\boldsymbol{Q}$ has the following properties:*

- *Chapman-Kolmogorov equation:*

$$\boldsymbol{\Phi}(s+t) = \boldsymbol{\Phi}(s) \cdot \boldsymbol{\Phi}(t), \quad s, t \geq 0.$$

- *The matrix $\boldsymbol{\Phi}(t)$ is the unique solution of the differential equation*

$$\boldsymbol{\Phi}'(t) = \boldsymbol{\Phi}(t) \cdot \boldsymbol{Q} = \boldsymbol{Q} \cdot \boldsymbol{\Phi}(t), \quad \boldsymbol{\Phi}(0) = \boldsymbol{I}, \quad t \geq 0.$$

- *Higher derivatives at the origin:*

$$\boldsymbol{\Phi}^{(k)}(0) = \boldsymbol{Q}^k, \quad k \geq 0.$$

- *The matrix $\boldsymbol{\Phi}(t)$ is stochastic (non-negative entries with row sums equal to one) for each $t \geq 0$.*

Proof. The matrix exponential $\exp(\boldsymbol{A})$ is well-defined because the series converges componentwise for any square matrix $\boldsymbol{A}$. If two $n \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ commute, we obtain $(AB)^k = A^k B^k$, $k \geq 0$. Then the basic exponentiation identity holds

$$\exp(\boldsymbol{A} + \boldsymbol{B}) = \exp(\boldsymbol{A}) \exp(\boldsymbol{B}). \tag{7.13}$$

But $t\boldsymbol{Q}$ and $s\boldsymbol{Q}$ commute for any values s, t and thus $\boldsymbol{\Phi}(s+t) = \boldsymbol{\Phi}(s) + \boldsymbol{\Phi}(t)$.

The matrix exponential $\exp(t\boldsymbol{Q})$ can be differentiated term-by-term, since the power series $\exp(t\boldsymbol{Q})$ has infinite radius of convergence. We have

$$\boldsymbol{\Phi}'(t) = \sum_{k=1}^{\infty} \frac{t^{k-1} \boldsymbol{Q}^k}{(k-1)!} = \boldsymbol{\Phi}(t) \cdot \boldsymbol{Q} = \boldsymbol{Q} \cdot \boldsymbol{\Phi}(t), \quad t \geq 0.$$

The uniqueness is a standard result for systems of ordinary linear differential equations. Iterated differentiation leads to the identity in item three.

Finally, when t approches 0, the Taylor series expansion gives $\boldsymbol{\Phi}(t) = \boldsymbol{I} + t\boldsymbol{Q} + O(t^2)$. Thus, when t is sufficiently small, $q_{ij}(t) \geq 0$ implies $\Phi_{ij} \geq 0$ for all $i \neq j$. But by the first part, $\boldsymbol{\Phi}(t) = \boldsymbol{\Phi}(t/m)^m$ for all m and thus $q_{ij}(t) \geq 0$ implies $\Phi_{ij} \geq 0$, $i \neq j$, for all $t \geq 0$. Moreover $\boldsymbol{Q}$ has row sums equal to zero and thus $\boldsymbol{Q}^m$ has the same property for each integer $m \geq 1$. Hence, the matrix $\boldsymbol{\Phi}(t)$ is stochastic. □

Example 7.15 (Maple). The matrix exponential for a matrix can be calculated as follows,

```
> with(LinearAlgebra):
> Q := Matrix( [[-2,1,1],[1,-2,1],[1,1,-2]] ):
> MatrixExponential(Q,t);
```

The resulting matrix is

$$\exp(Qt) = \begin{pmatrix} \frac{2}{3}e^{-3t} + \frac{1}{3} & \frac{1}{3} - \frac{2}{3}e^{-3t} & \frac{1}{3} - \frac{2}{3}e^{-3t} \\ \frac{1}{3} - \frac{2}{3}e^{-3t} & \frac{2}{3}e^{-3t} + \frac{1}{3} & \frac{1}{3} - \frac{2}{3}e^{-3t} \\ \frac{1}{3} - \frac{2}{3}e^{-3t} & \frac{1}{3} - \frac{2}{3}e^{-3t} & \frac{2}{3}e^{-3t} + \frac{1}{3} \end{pmatrix}.$$

♢

The matrix exponential $\exp(\boldsymbol{A})$ for a $n \times n$ complex-valued matrix $\boldsymbol{A}$ is invertible, since by the Chapman-Kolmogorov equation, $\exp(\boldsymbol{A}) \cdot \exp(-\boldsymbol{A}) = \exp(\boldsymbol{A} + (-\boldsymbol{A})) = \exp(\boldsymbol{0}) = \boldsymbol{I}$ and thus the inverse of $\exp(\boldsymbol{A})$ is $\exp(-\boldsymbol{A})$. The matrix exponential provides a mapping $\exp : M_n(\mathbb{C}) \to \mathrm{Gl}_n(\mathbb{C})$ from the vector space of all $n \times n$ complex matrices $(M_n(\mathbb{C}), +, \boldsymbol{0})$ to the general linear group of degree n, i.e., the group of all $n \times n$ invertible complex matrices $(\mathrm{GL}_n(\mathbb{C}), \cdot, \boldsymbol{I})$. This mapping is surjective which shows that each invertible complex matrix can be written as the exponential of some other complex matrix.

Moreover, the matrix exponential $\exp(\boldsymbol{A})$ for a $n \times n$ real-valued symmetric matrix $\boldsymbol{A}$ can be easily computed. For this, notice that any $n \times n$ real-valued symmetric matrix can be diagonalized by an orthogonal matrix; that is, there is a real-valued orthogonal matrix $\boldsymbol{U}$ such that $\boldsymbol{D} = \boldsymbol{U}^T \boldsymbol{A} \boldsymbol{U}$ is a diagonal matrix. For each diagnonal matrix $\boldsymbol{D}$ with main diagonal entries $d_1, \ldots, d_n$, the corresponding matrix exponential $\exp(\boldsymbol{D})$ is a diagonal matrix with main diagonal entries $e^{d_1}, \ldots, e^{d_n}$. Since $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$, it follows from the definitions that $\exp(\boldsymbol{A}) = \boldsymbol{U} \exp(\boldsymbol{D}) \boldsymbol{U}^T$. Since the formation of determinants commutes with matrix multiplication, we obtain

$$\det(\exp(\boldsymbol{A})) = \det(\exp(\boldsymbol{D})) = \prod_i e^{d_i} = e^{\mathrm{tr}(\boldsymbol{A})}, \tag{7.14}$$

where $\mathrm{tr}(\boldsymbol{A})$ denotes the trace of the matrix $\boldsymbol{A}$. By taking the natural logarithm, we obtain

$$\mathrm{tr}(\boldsymbol{A}) = \log(\det(\exp(\boldsymbol{A}))).$$

An *evolutionary model* for n taxa over an alphabet Σ is specified by a rooted binary tree T with n leaves together with a rate matrix $\boldsymbol{Q}$ over the alphabet Σ and an initial distribution for the root of the tree (which is often assumed to be uniform on the alphabet). The transition probability matrix for the edge kl of the tree T is given by the matrix exponential $\exp(\boldsymbol{Q} t_{kl})$, where the socalled branch lengths t_{kl} are to be determined. The corresponding algebraic statistical model $\mathbb{R}^d \to \mathbb{R}^m$ is a specialization of the hidden tree Markov model to these matrices.

The Felsenstein hierarchy is a nested family of evolutionary models. It can be seen as a cumulative result of experimentation and development of many special time-continuous Markov models with rate matrices that incorporate biologically meaningful parameters.

The simplest model is the *Jukes-Cantor model* (JC) developed in the late 1960s. This model is highly structured with equal transition probabilites among the nucleotides and with uniform root distribution. It is based on the *Jukes-Cantor rate matrix*

$$\boldsymbol{Q} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}, \tag{7.15}$$

where the parameter $\alpha > 0$ indicates that all substitutions occur at the same rate. The corresponding transition probability matrix is

$$\exp(t\boldsymbol{Q}) = \frac{1}{4}\begin{pmatrix} 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} \end{pmatrix}, \quad t \geq 0. \tag{7.16}$$

Note that all transition probabilities converge to $\frac{1}{4}$ as t approaches infinity. Thus in the equilibrium distribution all nucleotides are substituted with the same rate. Put $a = e^{-4\alpha t}$. Then the transition probability matrix has the form

$$\begin{pmatrix} 1-3a & a & a & a \\ a & 1-3\alpha & a & a \\ a & a & 1-3a & a \\ a & a & a & 1-3a \end{pmatrix}, \tag{7.17}$$

where a is the probability of a mutation from nucleotide i to another nucleotide j and $1-3a$ is the probability of staying at the nucleotide. Since the rows must sum to 1, we have $0 < a < 1/3$.

The expected number of mutations over time t is the quantity

$$3\alpha t = -\frac{1}{4} \cdot \operatorname{tr}(\boldsymbol{Q}) \cdot t = -\frac{1}{4} \cdot \log(\det(\boldsymbol{\Phi}(t))), \tag{7.18}$$

where the last equation follows from (7.14). This number is called the *branch length* of the model and is used to label the edges of a phylogenetic tree. Although the JC model does not capture the biology very well, it is easy to work with and is often used for quick calculations.

We make a change of variables in the JC model. For this, assume that the tree T has m edges and the leaves are indexed by the set $[n]$. Let $\boldsymbol{\Phi}^{(i)} = \boldsymbol{\Phi}(t_i)$ denote the transition probability matrix associated to the i-th edge, $1 \leq i \leq m$. Instead of using the parameters α_i and t_i, we introduce new parameters

$$\pi_i = \frac{1}{4}(1 - e^{-4\alpha_i t_i}) \quad \text{and} \quad \mu_i = \frac{1}{4}(1 + 3e^{-4\alpha_i t_i}). \tag{7.19}$$

The parameters are the entries of the transition probability matrix for the i-th edge,

$$\boldsymbol{\Phi}^{(i)} = \exp(t_i \boldsymbol{Q}) = \begin{pmatrix} \mu_i & \pi_i & \pi_i & \pi_i \\ \pi_i & \mu_i & \pi_i & \pi_i \\ \pi_i & \pi_i & \mu_i & \pi_i \\ \pi_i & \pi_i & \pi_i & \mu_i \end{pmatrix}. \tag{7.20}$$

The entries satisfy the linear constraint

$$\mu_i + 3\pi_i = 1. \tag{7.21}$$

The branch length of the i-th edge can be recovered from the matrix $\boldsymbol{\Phi}^{(i)}$ as follows,

$$3\alpha_i t_i = -\frac{1}{4} \cdot \log(\det(\boldsymbol{\Phi}^{(i)})). \tag{7.22}$$

The JC model has m parameters by (7.21) and 4^n states and is thus given by the mapping $f : \mathbb{R}^m \to \mathbb{R}^{4^n}$.

Example 7.16. Reconsider the 1,3 claw tree T (Fig. 7.9). We take the JC model with uniform root distribution. This model is given by the mapping $f : \mathbb{R}^3 \to \mathbb{R}^{64}$ with five distinct marginal distributions (Fig. 7.13).

The probability of observing the same symbol at all three leaves is

$$p_{123} = \frac{1}{4}(\mu_1\mu_2\mu_3 + 3\pi_1\pi_2\pi_3). \tag{7.23}$$

The probabilities of seeing the same nucleotide at two distinct leaves i, j and a different one at the third leaf are

$$p_{12} = \frac{1}{4}(\mu_1\mu_2\pi_3 + \pi_1\pi_2\mu_3 + 2\pi_1\pi_2\pi_3), \tag{7.24}$$

$$p_{13} = \frac{1}{4}(\mu_1\pi_2\mu_3 + \pi_1\mu_2\pi_3 + 2\pi_1\pi_2\pi_3), \tag{7.25}$$

$$p_{23} = \frac{1}{4}(\pi_1\mu_2\mu_3 + \mu_1\pi_2\pi_3 + 2\pi_1\pi_2\pi_3). \tag{7.26}$$

The probability of noticing three distinct letters at the leaves is

$$p_{dis} = \frac{1}{4}(\mu_1\pi_2\pi_3 + \pi_1\mu_2\pi_3 + \pi_1\pi_2\mu_3 + \pi_1\pi_2\pi_3). \tag{7.27}$$

♢

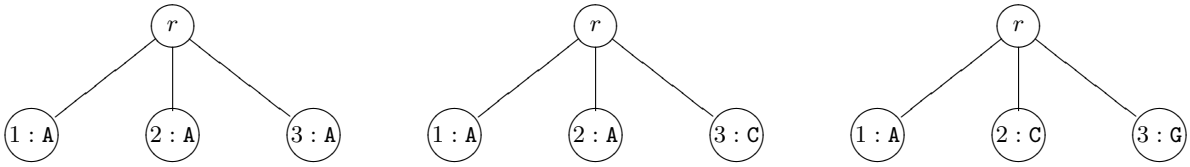


Fig. 7.13. Labellings of the 1,3 claw tree for p_{123}, p_{12}, and p_{dis}, respectively.

The nucleotides fall biochemically into two families: *purines* (adenine and guanine) and *pyrimidines* (cytosine and thymine). Substitutions within a family are called *transitions* and substitutions between families are called *transversions*.

The *Kimura two-parameter model* K2P (1980) is an extension of the Jukes-Cantor model that distinguishes between transitions and transversions by assigning a common rate to transversions and a different common rate to transitions but still assumes an equal root distribution. The model K2P is based on the rate matrix

$$\begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} . & \beta & \alpha & \beta \\ \beta & . & \beta & \alpha \\ \alpha & \beta & . & \beta \\ \beta & \alpha & \beta & . \end{pmatrix} \end{array}, \tag{7.28}$$

where $\alpha, \beta > 0$ and the diagonal entries are determined by the constraint that the row sums are equal to zero.

The *Kimura three-parameter model* K3P (1981) is a generalization of the K2P model that allows two classes of transversions to occur at different rates. Its rate matrix is of the form

$$\begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} . & \beta & \alpha & \gamma \\ \beta & . & \gamma & \alpha \\ \alpha & \gamma & . & \beta \\ \gamma & \alpha & \beta & . \end{pmatrix} \end{array}, \tag{7.29}$$

where $\alpha, \beta, \gamma > 0$. The corresponding matrix exponential is

$$\begin{pmatrix} r_t & s_t & u_t & v_t \\ s_t & r_t & v_t & u_t \\ u_t & v_t & r_t & s_t \\ v_t & u_t & s_t & r_t \end{pmatrix}, \quad t \geq 0, \tag{7.30}$$

where

$$\begin{aligned} r_t &= \frac{1}{4}\exp(-2(\alpha+\beta)t) + \frac{1}{4}\exp(-2(\alpha+\gamma)t) + \frac{1}{4}\exp(-2(\beta+\gamma)t) + \frac{1}{4} \\ s_t &= -\frac{1}{4}\exp(-2(\alpha+\beta)t) - \frac{1}{4}\exp(-2(\alpha+\gamma)t) + \frac{1}{4}\exp(-2(\beta+\gamma)t) + \frac{1}{4} \\ u_t &= -\frac{1}{4}\exp(-2(\alpha+\beta)t) + \frac{1}{4}\exp(-2(\alpha+\gamma)t) - \frac{1}{4}\exp(-2(\beta+\gamma)t) + \frac{1}{4} \\ v_t &= \frac{1}{4}\exp(-2(\alpha+\beta)t) - \frac{1}{4}\exp(-2(\alpha+\gamma)t) - \frac{1}{4}\exp(-2(\beta+\gamma)t) + \frac{1}{4}. \end{aligned}$$

The *strand symmetric model* CS05 (2004) extends the K2P model by assuming that the root distribution fulfills $\pi_{\mathtt{A}} = \pi_{\mathtt{T}}$ and $\pi_{\mathtt{C}} = \pi_{\mathtt{G}}$. The CS05 model has the rate matrix

$$\begin{pmatrix} . & \beta\pi_{\mathtt{C}} & \alpha\pi_{\mathtt{C}} & \beta\pi_{\mathtt{A}} \\ \beta\pi_{\mathtt{A}} & . & \beta\pi_{\mathtt{C}} & \alpha\pi_{\mathtt{A}} \\ \alpha\pi_{\mathtt{A}} & \beta\pi_{\mathtt{C}} & . & \beta\pi_{\mathtt{A}} \\ \beta\pi_{\mathtt{A}} & \alpha\pi_{\mathtt{C}} & \beta\pi_{\mathtt{C}} & . \end{pmatrix}, \tag{7.31}$$

where $\alpha, \beta > 0$.

The *Felsenstein model* F81 (1981) is an extension of the JC model that allows a non-uniform root distribution. The rate matrix is

$$\begin{pmatrix} . & \alpha\pi_{\mathtt{C}} & \alpha\pi_{\mathtt{G}} & \alpha\pi_{\mathtt{T}} \\ \alpha\pi_{\mathtt{A}} & . & \alpha\pi_{\mathtt{G}} & \alpha\pi_{\mathtt{T}} \\ \alpha\pi_{\mathtt{A}} & \alpha\pi_{\mathtt{C}} & . & \alpha\pi_{\mathtt{T}} \\ \alpha\pi_{\mathtt{A}} & \alpha\pi_{\mathtt{C}} & \alpha\pi_{\mathtt{G}} & . \end{pmatrix}, \tag{7.32}$$

where $\alpha > 0$.

The *Hasegawa model* HKY85 (1985) is a compromise between the models F81 and CS05. Its rate matrix is

$$\begin{pmatrix} . & \beta\pi_{\mathtt{C}} & \alpha\pi_{\mathtt{G}} & \beta\pi_{\mathtt{T}} \\ \beta\pi_{\mathtt{A}} & . & \beta\pi_{\mathtt{G}} & \alpha\pi_{\mathtt{T}} \\ \alpha\pi_{\mathtt{A}} & \beta\pi_{\mathtt{C}} & . & \beta\pi_{\mathtt{T}} \\ \beta\pi_{\mathtt{A}} & \alpha\pi_{\mathtt{C}} & \beta\pi_{\mathtt{G}} & . \end{pmatrix}, \tag{7.33}$$

where $\alpha, \beta > 0$.

The *Tamura-Nei model* TN93 (1993) is an extension of the HKY85 model including an extra parameter for the two types of transversions. It has rate matrix

$$\begin{pmatrix} . & \beta\pi_{\mathtt{C}} & \alpha\pi_{\mathtt{G}} & \beta\pi_{\mathtt{T}} \\ \beta\pi_{\mathtt{A}} & . & \beta\pi_{\mathtt{G}} & \gamma\pi_{\mathtt{T}} \\ \alpha\pi_{\mathtt{A}} & \beta\pi_{\mathtt{C}} & . & \beta\pi_{\mathtt{T}} \\ \beta\pi_{\mathtt{A}} & \gamma\pi_{\mathtt{C}} & \beta\pi_{\mathtt{G}} & . \end{pmatrix}, \tag{7.34}$$

where $\alpha, \beta, \gamma > 0$.

The *symmetric model* SYM (1994) is given by a symmetric rate matrix and assumes uniform root distribution. Its rate matrix is

$$\begin{pmatrix} . & \alpha & \beta & \gamma \\ \alpha & . & \delta & \epsilon \\ \beta & \delta & . & \phi \\ \gamma & \epsilon & \phi & . \end{pmatrix}, \tag{7.35}$$

where $\alpha, \beta, \gamma, \delta, \epsilon, \phi > 0$.

The *REV model* (1984) is the most general DNA model. Its only restriction is symmetry. The rate matrix is

$$\begin{pmatrix} . & \alpha\pi_{\mathtt{C}} & \beta\pi_{\mathtt{G}} & \gamma\pi_{\mathtt{T}} \\ \alpha\pi_{\mathtt{A}} & . & \delta\pi_{\mathtt{G}} & \epsilon\pi_{\mathtt{T}} \\ \beta\pi_{\mathtt{A}} & \delta\pi_{\mathtt{C}} & . & \phi\pi_{\mathtt{T}} \\ \gamma\pi_{\mathtt{A}} & \epsilon\pi_{\mathtt{C}} & \phi\pi_{\mathtt{G}} & . \end{pmatrix}, \tag{7.36}$$

where $\alpha, \beta, \gamma, \delta, \epsilon, \phi > 0$.

Example 7.17 (Singular). Reconsider the JC model of the 1,3 claw tree $f : \mathbb{R}^3 \to \mathbb{R}^{64}$ studied in Ex. 7.16. Since the model has only five different marginal probabilities, we may consider the JC model as the image of the simplified mapping $f' : \mathbb{R}^3 \to \mathbb{R}^5$ with marginal probabilities given by (7.23)-(7.27). The following program computes invariants of the simplified model,

```
> ring r = 0, (x(1..3),y(1..3),p(1..5)), dp;
> ideal i1 = p(1)-(x(1)*x(2)*x(3)+3*y(1)*y(2)*y(3));
> ideal i2 = p(2)-(x(1)*x(2)*y(3)+y(1)*y(2)*x(3)+2*y(1)*y(2)*y(3));
> ideal i3 = p(3)-(x(1)*y(2)*x(3)+y(1)*x(2)*y(3)+2*y(1)*y(2)*y(3));
> ideal i4 = p(4)-(y(1)*x(2)*x(3)+x(1)*y(2)*y(3)+2*y(1)*y(2)*y(3));
> ideal i5 = p(5)-(x(1)*y(2)*y(3)+y(1)*x(2)*y(3)+y(1)*y(2)*x(3)+y(1)*y(2)*y(3));
> ideal i = i1+i2+i3+i4+i5, x(1)+3*y(1)-1, x(2)+3*y(2)-1, x(3)+3*y(3)-1;
> ideal j = std(i);
> ideal k = eliminate(j,x(1)*x(2)*x(3)*y(1)*y(2)*y(3));
```

The output provides three (large) phylogenetic invariants in the ring $\mathbb{R}[p_1, \ldots, p_5]$. ♢

Given a set of observed DNA sequences. We may ask the question whether or not these sequences come from a particular evolutionary model $\mathcal{M}$ given by a tree T and a root distribution π. For this, we compute the pattern frequencies $(\hat{p}_\sigma)$ from the observed taxa. These pattern frequencies can serve as estimates for the marginal probabilities (p_σ) of the model. Then we evaluate each of the model invariants using the pattern frequencies. If the sequence data come from the model, we will expect the evaluated invariants not to differ significantly from zero. These evaluated invariants give us a score of the model. Then we may choose the evolutionary model with the minimal score among all models as the "true" evolutionary model.

7.7 Group-Based Evolutionary Models

The Jukes-Cantor model for either binary or DNA sequences and the Kimura models with two or three parameters belong to the class of group based models. These models have the property that a linear change of coordinates by using discrete Fourier transform translates the ideal of phylogenetic invariants into a toric model. More specifically, the symbols of the alphabet in a group-based model can be labelled by the elements of a finite group in such a way that the probability of translating group elements (from g to h) depends only on their difference $(g-h)$. By replacing the original coordinates $p_{i_1,\ldots,i_n}$ by Fourier coordinates $q_{i_1,\ldots,i_n}$, the ideal of phylogenetic invariantes becomes toric.

An evolutionary model on the state space $[n]$ is called *group-based* if there is an abelian group G with elements $g_1, \ldots, g_n$ and a mapping $\psi : G \to \mathbb{R}$ such that the $n \times n$ instantaneous rate matrix $\boldsymbol{Q} = (Q_{ij})$ satisfies the condition

$$Q_{ij} = \psi(g_j - g_i), \quad 1 \leq i, j \leq n. \tag{7.37}$$

Example 7.18. Consider the cyclic group $G = \mathbb{Z}_4$. The group table for the differences $g - h$ of group elements $g, h \in G$ has the form

$-$	0	1	2	3
0	0	3	2	1
1	1	0	3	2
2	2	1	0	3
3	3	2	1	0

(7.38)

and can be mapped onto the entries of the instantaneous rate matrix for the Kimura's two-parameter model K80,

$$\boldsymbol{Q} = \begin{pmatrix} . & \alpha & \beta & \alpha \\ \alpha & . & \alpha & \beta \\ \beta & \alpha & . & \alpha \\ \alpha & \beta & \alpha & . \end{pmatrix}. \tag{7.39}$$

♢

Example 7.19. Take the Klein group $G = \mathbb{Z}_2 \times \mathbb{Z}_2$. The group table for the differences $g - h$ of group elements $g, h \in G$ has the form

$$\begin{array}{c|cccc} - & (0,0) & (0,1) & (1,0) & (1,1) \\ \hline (0,0) & (0,0) & (0,1) & (1,0) & (1,1) \\ (0,1) & (0,1) & (0,0) & (1,1) & (1,0) \\ (1,0) & (1,0) & (1,1) & (0,0) & (0,1) \\ (1,1) & (1,1) & (1,0) & (0,1) & (0,0) \end{array} \tag{7.40}$$

and can be mapped onto the entries of the instantaneous rate matrix for the Kimura's three-parameter model K81,

$$\boldsymbol{Q} = \begin{pmatrix} . & \alpha & \beta & \gamma \\ \alpha & . & \gamma & \beta \\ \beta & \gamma & . & \alpha \\ \gamma & \beta & \alpha & . \end{pmatrix}. \tag{7.41}$$

♢

We show that the substitution matrices share the same dependencies. To this end, we need to summarize some basic facts about characters and discrete Fourier transforms. For this, let G be a finite abelian group. We denote the group operation by addition, the neutral element by 0, and the l-fold multiple of a group element $g \in G$ as $l \cdot g = g + \cdots + g$ (l times). Let $\mathbb{C}^*$ denote the set of non-zero complex numbers; we can regard $\mathbb{C}^*$ as an abelian group with ordinary multiplication.

A *character* of G is a group homomorphism mapping G into $\mathbb{C}^*$; that is, $\chi : G \to \mathbb{C}^*$ is a character if $\chi(g_1 + g_2) = \chi(g_1)\chi(g_2)$ for all elements $g_1, g_2 \in G$. The set of characters of G is denoted by $\hat{G}$. The set $\hat{G}$ is non-empty, since it contains the *trivial* character ϵ defined by $\epsilon(g) = 1$ for all $g \in G$.

Lemma 7.20. *The set of characters of G forms an abelian group under pointwise multiplication; that is,*

$$(\chi\chi')(g) = \chi(g)\chi'(g), \quad \chi, \chi' \in \hat{G},\ g \in G.$$

The groups G and $\hat{G}$ are isomorphic.

Proof. First, the defined operation on $\hat{G}$ is associative and commutative which follows directly from associativity and commutativity of complex multiplication. Thus $\hat{G}$ forms a commutative semigroup. The trivial character ϵ is the neutral element of $\hat{G}$ and so $\hat{G}$ forms a commutative monoid.

Suppose the group G has order n. Then for each character $\chi \in \hat{G}$, we have $\chi(g)^n = \chi(g^n) = \chi(1) = 1$. Thus the image of G under χ lies in the group of n-th roots of unity. The norm of an n-th root of unity ζ is $|\zeta| = \sqrt{\zeta\overline{\zeta}} = 1$, where $\overline{z}$ denotes the conjugate of a complex number z. Thus $\zeta\overline{\zeta} = 1$ and hence $\overline{\zeta} = \zeta^{-1}$. Therefore, the inverse of a character χ is given by $\chi^{-1}(g) = \overline{\chi(g)}$ for all $g \in G$. It follows that $\hat{G}$ forms an abelian group.

Second, the fundamental theorem of abelian groups says that each finite abelian group G is a direct sum of cyclic groups $Z_1, \ldots, Z_k$. Let z_i be a generating element of Z_i with order n_i; that is, $Z_i = \{l \cdot z_i \mid 0 \le l \le n_i - 1\}$, $1 \le i \le k$. Let ζ_i denote a primitive n_i-th root of unity; that is, $\zeta_i^{n_i} = 1$ and $\zeta_i^j \neq 1$ for $1 \le j \le n_i - 1$.

Define $\chi_i \in \hat{G}$ such that $\chi_i(l_i \cdot z_i) = \zeta_i^{l_i}$, where $0 \le l_i \le n_i - 1$, $1 \le i \le k$, and extend such that

$$\chi_i(l_1 \cdot z_1 + \cdots + l_k \cdot z_k) = \zeta_i^{l_i}, \quad 0 \le l_i \le n_i - 1,\ 1 \le i \le k.$$

Let $g \in G$. Write $g = l_1 \cdot z_1 + \cdots + l_k \cdot z_k$, where $0 \le l_j \le n_j - 1$, $1 \le j \le k$, and put

$$\phi(g) = \chi_1^{l_1} \cdots \chi_k^{l_k}. \tag{7.42}$$

For each $g, h \in G$, we have by definition $\phi(g+h) = \phi(g)\phi(h)$. Hence, ϕ is a group homomorphism.

Let $g \in G$ such that $\phi(g) = \epsilon$. Writing $g = l_1 \cdot z_1 + \cdots + l_k \cdot z_k$ gives $1 = \epsilon(z_i) = (\chi_1^{l_1} \cdots \chi_k^{l_k})(z_i) = \chi_1(z_i)^{l_1} \cdots \chi_k(z_i)^{l_k} = \chi_i(z_i)^{l_i} = \zeta_i^{l_i}$, $1 \le i \le k$. It follows that n_i is a divisor of l_i for each $1 \le i \le k$ and thus $g = 1$. Hence, the mapping is one-to-one.

Let $\chi \in \hat{G}$. Since $\chi(z_i)$ is an n_i-th root of unity, we have $\chi(z_i) = \zeta_i^{e_i} = \chi_i(z_i)^{e_i}$ for some $0 \le e_i \le n_i - 1$, $1 \le i \le k$. Thus $\chi = \chi_1^{e_1} \cdots \chi_k^{e_k}$ and hence the mapping is onto. □

The group $\hat{G}$ is called the *dual group* or *character group* of G.

Lemma 7.21. *Let G and H be finite abelian groups. The dual group of the direct product $G \times H = \{(g,h) \mid g \in G, h \in H\}$ is isomorphic to $\hat{G} \times \hat{H}$.*

Proof. Let χ be a character of $G \times H$. The restriction of χ to G is a character of G and the restriction to H is a character of H; we denote the restricted characters by χ_G and χ_H, respectively. This gives $\chi(g,h) = \chi_G(g) \cdot \chi_H(h)$ for all $(g,h) \in G \times H$. The mapping $\chi \mapsto (\chi_G, \chi_H)$ provides the required isomorphism. □

Example 7.22. The dual group of the cyclic group $G = \mathbb{Z}_n = \{0, 1, \ldots, n-1\}$ is the group $\hat{G} = \{\chi^b \mid 0 \le b \le n-1\}$ with

$$\chi^b(a) = \zeta^{ab}, \quad 0 \le a, b \le n-1, \tag{7.43}$$

where ζ is a primitive n-th root of unit. ♢

Example 7.23. The dual group of the additive group $G = \mathbb{Z}_2^k$ of order 2^k has the characters

$$(\chi_1^{a_1} \cdots \chi_k^{a_k})(b_1 \cdot z_1 + \cdots + b_k \cdot z_k) = \prod_{i=1}^{k} \chi_i^{a_i}(b_i \cdot z_i) = \prod_{i=1}^{k} (-1)^{a_i b_i} = (-1)^{\langle \boldsymbol{a}, b \rangle}. \tag{7.44}$$

where $0 \le a_i, b_i \le 1$, $1 \le i \le k$. ♢

Let G be a finite abelian group and let $L^2(G) = \{f \mid f : G \to \mathbb{C}\}$ be the set of all complex-valued functions on G. This set becomes a complex vector space by defining addition

$$(f_1 + f_2)(g) = f_1(g) + f_2(g), \quad f_1, f_2 \in L^2(G),\ g \in G, \tag{7.45}$$

and scalar multiplication

$$(af)(g) = a \cdot f(g), \quad f \in L^2(G),\ g \in G,\ a \in \mathbb{C}. \tag{7.46}$$

Define the delta functions δ_g, $g \in G$, on G by

$$\delta_g(h) = \begin{cases} 1, \text{ if } g = h, \\ 0, \text{ otherwise.} \end{cases}$$

The vector space $L^2(G)$ has the delta functions as a $\mathbb{C}$-basis, since each function $f \in L^2(G)$ has the *Fourier expansion*

$$f(g) = \sum_h f(h)\delta_h(g), \quad g \in G.$$

A multiplication on the $\mathbb{C}$-space $L^2(G)$ is given as

$$(f_1 * f_2)(g) = \sum_{h \in G} f_1(h) f_2(g-h), \quad g \in G,\ f_1, f_2 \in L^2(G).$$

This operation is associative and is called *convolution* or *Hadamard product*.

An inner product on the vector space $L^2(G)$ is defined by

$$\langle f_1, f_2 \rangle = \sum_{g \in G} f_1(g)\overline{f_2(g)}, \quad f_1, f_2 \in L^2(G). \tag{7.47}$$

The delta functions on G define an orthonormal basis of $L^2(G)$ with respect to this inner product, since we have

$$\langle \delta_g, \delta_h \rangle = \sum_l \delta_g(l)\overline{\delta_h(l)} = \begin{cases} 1, \text{ if } g = h, \\ 0, \text{ otherwise,} \end{cases} \quad g, h \in G.$$

Theorem 7.24 (Orthogonality Relations). *Let G be a finite abelian group.*

- *For all characters χ and ψ of G, we have*

$$\langle \chi, \psi \rangle = \begin{cases} |G|, \textit{ if } \chi = \psi, \\ 0, \quad \textit{otherwise.} \end{cases} \tag{7.48}$$

- *For all elements g and h in G, we have*

$$\sum_\chi \chi(g)\overline{\chi(h)} = \begin{cases} |G|, \textit{ if } g = h, \\ 0, \quad \textit{otherwise.} \end{cases} \tag{7.49}$$

Proof. First, we have

$$\langle \chi, \psi \rangle = \sum_g \chi(g)\overline{\psi(g)} = \sum_g (\chi\psi^{-1})(g)\overline{\epsilon(g)} = \langle \chi\psi^{-1}, \epsilon \rangle,$$

since $\overline{\psi} = \psi^{-1}$. Thus we can reduce to the case $\psi = \epsilon$. Put

$$S = \langle \chi, \epsilon \rangle = \sum_g \chi(g).$$

If χ is the trival character, the result will follow. Otherwise, there is a group element $h \in G$ with $\chi(h) \neq 1$. By multiplying the above equation with $\chi(h)$, we obtain

$$\chi(h)S = \chi(h)\sum_g \chi(g) = \sum_g \chi(h+g) = \sum_g \chi(g) = S.$$

Thus we have $\chi(h)S = S$ with $\chi(h) \neq 1$ and hence $S = 0$. This proves the first assertion.

Second, we have

$$\sum_\chi \chi(g)\overline{\chi(h)} = \sum_\chi \chi(g-h), \quad g, h \in G.$$

If $g = h$, the result will follow. Otherwise, there is a character ψ with $\psi(l) \neq 1$ for some $l \in G$. Define $S = \sum_\chi \chi(l)$. Thus $\psi(l)S = \sum_\chi (\psi\chi)(l) = S$ and hence $S = 0$. □

The *discrete Fourier transform* (DFT) $\mathcal{F} : L^2(G) \to L^2(\hat{G})$ assigns to each function $f \in L^2(G)$ a function $\mathcal{F}f = \hat{f}$ defined by

$$\hat{f}(\chi) = \sum_{g\in G} f(g)\overline{\chi(g)}, \quad \chi \in \hat{G}. \tag{7.50}$$

In particular, the DFT of the delta function δ_g, $g \in G$, is given by

$$\hat{\delta}_g(\chi) = \sum_h \overline{\chi(h)}\delta_g(h) = \overline{\chi(g)}, \quad \chi \in \hat{G}. \tag{7.51}$$

Theorem 7.25. *Let G be a finite abelian group.*

- *Linearity: The DFT $\mathcal{F} : L^2(G) \to L^2(\hat{G})$ is a $\mathbb{C}$-space isomorphism.*
- *Convolution: The DFT turns convolution into multiplication,*

$$\widehat{f_1 * f_2}(\chi) = \hat{f}_1(\chi) \cdot \hat{f}_2(\chi), \quad \chi \in \hat{G},\ f_1, f_2 \in L^2(G).$$

- *Inversion: For each function $f \in L^2(G)$,*

$$f(g) = \frac{1}{|G|}\sum_{\chi\in\hat{G}} \chi(g)\hat{f}(\chi), \quad g \in G.$$

- *Parseval identity: For all functions $f_1, f_2 \in L^2(G)$,*

$$\langle f_1, f_2\rangle = \frac{1}{|G|}\langle \hat{f}_1, \hat{f}_2\rangle.$$

- *Translation: For each $h \in G$ and $f \in L^2(G)$, define $f^h(g) = f(h+g)$. Then we have*

$$\widehat{f^h}(\chi) = \chi(h)\hat{f}(\chi), \quad f \in L^2(G),\ \chi \in \hat{G},\ h \in G.$$

Proof. First, the DFT provides a linear map, since

$$\begin{aligned}(\hat{f}_1 + \hat{f}_2)(\chi) = \hat{f}_1(\chi) + \hat{f}_2(\chi) &= \sum_g \overline{\chi(g)}[f_1(g) + f_2(g)] \\ &= \sum_g \overline{\chi(g)}[f_1 + f_2](g) = \widehat{f_1 + f_2}(\chi)\end{aligned}$$

and

$$\widehat{af}(\chi) = \sum_g (af)(g)\overline{\chi(g)} = \sum_g a \cdot f(g)\overline{\chi(g)} = a \cdot \hat{f}(\chi), \quad a \in \mathbb{C}.$$

The inversion formula implies that this mapping is one-to-one. Since both spaces have the same dimension, it follows by linear algebra that the mapping is also onto. Hence, the mapping is a vector space isomorphism.

Second, we have

$$\begin{aligned}
\widehat{f_1 * f_2}(\chi) &= \sum_g \overline{\chi(g)}\widehat{f_1 * f_2}(g) = \sum_g \overline{\chi(g)} \sum_h f_1(h) f_2(g-h) \\
&= \sum_h \sum_l \overline{\chi(h+l)} f_1(h) f_2(l), \quad l = g - h, \\
&= \sum_h \sum_l \overline{\chi(h)} f_1(h) \overline{\chi(l)} f_2(l) \\
&= \sum_h \overline{\chi(h)} f_1(h) \sum_l \overline{\chi(l)} f_2(l) \\
&= \hat{f}_1(\chi) \cdot \hat{f}_2(\chi).
\end{aligned}$$

Third, due to linearity, we may consider only the basis elements δ_h, $h \in G$. By the orthogonality relations (7.49) and (7.51), the right-hand side gives

$$\frac{1}{|G|} \sum_\chi \chi(g) \hat{\delta}_h(\chi) = \frac{1}{|G|} \sum_\chi \chi(g)\overline{\chi(h)} = \begin{cases} 1, \text{ if } g = h, \\ 0, \text{ otherwise,} \end{cases}$$

which is equal to $\delta_h(g)$, as required.

Fourth, the orthogonality relations (7.49) give

$$\begin{aligned}
\langle \hat{f}_1, \hat{f}_2 \rangle &= \sum_\chi \hat{f}_1(\chi)\overline{\hat{f}_2(\chi)} \\
&= \sum_\chi \left(\sum_g f_1(g)\overline{\chi(g)} \cdot \overline{\sum_h f_2(h)\overline{\chi(h)}} \right) \\
&= \sum_\chi \sum_g \sum_h f_1(g)\overline{f_2(h)\chi(g)}\chi(h) \\
&= \sum_g \sum_h f_1(g)\overline{f_2(h)} \sum_\chi \overline{\chi(g)}\chi(h) \\
&= |G| \cdot \sum_g f_1(g)\overline{f_2(g)} \\
&= |G| \cdot \langle f_1, f_2 \rangle.
\end{aligned}$$

Finally, we have

$$\begin{aligned}\widehat{f^h}(\chi) &= \sum_g f^h(g)\overline{\chi(g)} \\ &= \sum_g f(g+h)\overline{\chi(g)} \\ &= \sum_l f(l)\overline{\chi(l-h)}, \quad l = g+h, \\ &= \sum_l f(l)\overline{\chi(l)}\chi(h) \\ &= \chi(h)\hat{f}(\chi).\end{aligned}$$

□

Example 7.26. Consider the cyclic group $G = \mathbb{Z}_n$. By taking the basis of $L^2(G)$ given by the delta functions, the proof of Lemma 7.20 and (7.51) show that

$$\hat{\delta}_a(\chi^b) = \overline{\chi^b(a)} = 1/\chi^b(a) = \zeta^{-ab}, \quad 0 \le a, b \le n-1, \tag{7.52}$$

where $\zeta = \exp(2\pi i/n)$ is a primitive n-th root of unity. Thus the matrix of the DFT is given by

$$\boldsymbol{A}_n = \left(\zeta^{-(a-1)(b-1)}\right)_{a,b\in[n]}. \tag{7.53}$$

In the binary case $n = 2$, we obtain

$$\boldsymbol{A}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}. \tag{7.54}$$

The DFT of the function $f = (f_0, f_1)$ is given by

$$\hat{f} = \begin{pmatrix} f_0 + f_1 \\ f_0 - f_1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \end{pmatrix}. \tag{7.55}$$

Moreover, the inversion formula gives

$$f_0 = \frac{1}{2}(\hat{f}_0 + \hat{f}_1) \quad \text{and} \quad f_1 = \frac{1}{2}(\hat{f}_0 - \hat{f}_1). \tag{7.56}$$

In the quaternary case $n = 4$, we obtain

$$\boldsymbol{A}_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{pmatrix}. \tag{7.57}$$

♢

Example 7.27 (Maple). The constant function $f : \mathbb{Z}_n \to \mathbb{C} : a \mapsto 1$ has the DFT

$$\hat{f}(\chi^b) = \begin{cases} n \text{ if } b = 0, \\ 0 \text{ otherwise.} \end{cases}$$

To see this, we have

$$\hat{f}(\chi^b) = \sum_a f(a)\zeta^{-ab}.$$

Clearly, if $b = 0$ then $\hat{f}(1) = n$. On the other hand, if $b \neq 0$, then by the formula for geometric progression,

$$\hat{f}(\chi^b) = 1 + \zeta + \ldots + \zeta^{n-1} = \frac{1-\zeta^n}{1-\zeta} = 0.$$

Take the function $f(a) = \frac{1}{2}(\delta_1(a) + \delta_{-1}(a))$ for all $a \in \mathbb{Z}_n$; here we assume that n is odd and so we can identify $\mathbb{Z}_n$ with the set $\{-(n-1)/2, \ldots, -1, 0, 1, \ldots, (n-1)/2\}$. The DFT of the function f is

$$\hat{f}(\chi^b) = \cos\left(\frac{2\pi b}{n}\right), \quad a \in \mathbb{Z}_n.$$

Indeed, we have

$$\begin{aligned} \hat{f}(\chi^b) &= \sum_a f(a)\zeta^{-ab} = \frac{1}{2}\exp(-2\pi ib/n) + \frac{1}{2}\exp(2\pi ib/n) \\ &= \frac{1}{2}\left[\cos\left(\frac{-2\pi b}{n}\right) + i \cdot \sin\left(\frac{-2\pi b}{n}\right)\right] + \frac{1}{2}\left[\cos\left(\frac{2\pi b}{n}\right) + i \cdot \sin\left(\frac{2\pi b}{n}\right)\right] \\ &= \cos\left(\frac{2\pi b}{n}\right). \end{aligned}$$

The function $f : \mathbb{Z}_n \to \mathbb{C} : a \mapsto a^3$ has the DFT shown in Fig. 7.14. It can be computed by the following `Maple` code:

```
> with(DiscreteTransform):
> with(plots):
> Z := Vector ( 9, x -> evalf(x^3)):
> F := FourierTransform ( Z, normalization = full):
> ptlist := convert ( F, 'list'):
> complexplot ( ptlist, x = -50..225, style = point);
```

♢

Example 7.28. Consider the additive group $G = \mathbb{Z}_2^k$ of order 2^k. The DFT of a function $f \in L^2(G)$ can be specified as

$$\hat{f}(\chi_1^{a_1} \cdots \chi_k^{a_k}) = \sum_{b \in G} (-1)^{\langle \boldsymbol{a}, b \rangle} f(\boldsymbol{b}), \quad 0 \leq a_1, \ldots, a_k \leq 1. \tag{7.58}$$

This mapping is provided by the $2^k \times 2^k$ *Hadamard matrix*

$$\boldsymbol{H}_{2^k} = ((-1)^{\langle \boldsymbol{a}, b \rangle}).$$

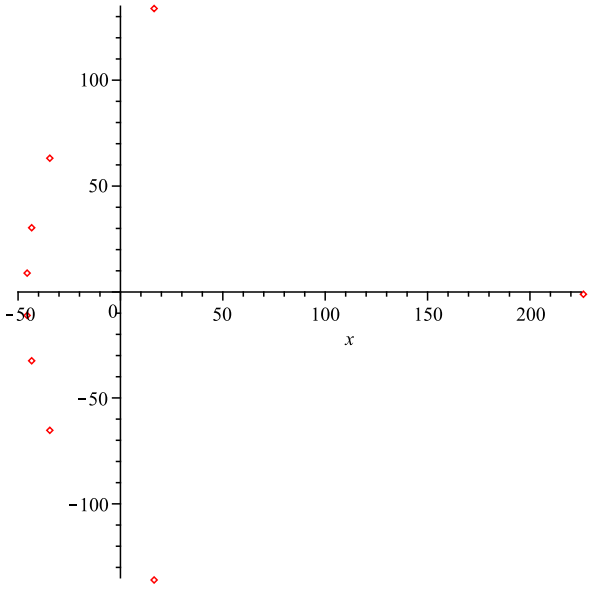


Fig. 7.14. DFT of cubic function ($n = 9$).

These matrices can be recursively defined as follows,

$$\boldsymbol{H}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad \boldsymbol{H}_{2^{k+1}} = \begin{pmatrix} \boldsymbol{H}_{2^k} & \boldsymbol{H}_{2^k} \\ \boldsymbol{H}_{2^k} & -\boldsymbol{H}_{2^k} \end{pmatrix} = \boldsymbol{H}_{2^k} \otimes \boldsymbol{H}_2, \quad k \geq 1. \tag{7.59}$$

♢

We will see that if the instantaneous rate matrix $\boldsymbol{Q}$ is group-based, the corresponding substitution matrices $\exp(\boldsymbol{Q}t)$ will also be group-based.

Lemma 7.29. *Let G be an abelian group of order n. The eigenvalues of a $n \times n$ group-based instantaneous rate matrix $\boldsymbol{Q}$ satisfying (7.37) are*

$$\lambda_\chi = \sum_{h \in G} \overline{\chi(h)} \psi(h), \quad \chi \in \hat{G}. \tag{7.60}$$

The transition probabilities of the corresponding time-continuous Markov model are

$$P_{gh}(t) = \frac{1}{|G|} \sum_{\chi \in \hat{G}} \chi(h-g) e^{\lambda_\chi t}, \quad t \geq 0. \tag{7.61}$$

Proof. First, define the $n \times n$ matrix $\boldsymbol{B} = (\chi(g))_{\chi,g}$. We have

$$\begin{aligned}(\boldsymbol{BQ})_{\chi,g} &= \sum_{h\in G}\chi(h)\psi(g-h)\\ &= \sum_{l\in G}\chi(g-l)\psi(l)\\ &= \chi(g)\sum_{l\in G}\overline{\chi(l)}\psi(l)\\ &= \chi(g)\lambda_\chi.\end{aligned}$$

Thus if we put $G=\{g_1,\ldots,g_n\}$, then

$$(\chi(g_1),\ldots,\chi(g_n))\boldsymbol{Q} = \lambda_\chi(\chi(g_1),\ldots,\chi(g_n)). \tag{7.62}$$

Hence, the rows of $\boldsymbol{B}$ are the left eigenvectors of $\boldsymbol{Q}$. This shows the first assertion.

Second, let $\boldsymbol{D}$ be the $n\times n$ diagonal matrix with diagonal entries λ_χ. By (7.62), we have $\boldsymbol{Q}=\boldsymbol{B}^{-1}\boldsymbol{D}\boldsymbol{B}$. But the orthogonality relations (7.49) imply that $\boldsymbol{B}^{-1}=\frac{1}{|G|}\boldsymbol{B}^*$, where $\boldsymbol{B}^*=(\overline{\chi(g)})_{g,\chi}$ is the Hermitian matrix of $\boldsymbol{B}$. Thus

$$\begin{aligned}(\exp(\boldsymbol{Q}t))_{g,h} &= \frac{1}{|G|}(\boldsymbol{B}^*\exp(\boldsymbol{D}t)\boldsymbol{B})_{g,h}\\ &= \frac{1}{|G|}\sum_{\chi}\overline{\chi(g)}e^{\lambda_\chi t}\chi(h)\\ &= \frac{1}{|G|}\sum_{\chi}\chi(h-g)e^{\lambda_\chi t}.\end{aligned}$$

This proves the second assertion. □

Example 7.30 (Singular). Consider the binary JC model for the 1,3 claw tree (Fig. 7.15). Let $\pi=$

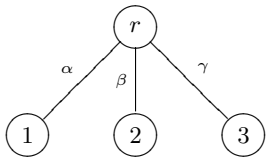


Fig. 7.15. The 1,3 claw tree.

(π_0,π_1) denote the probability distribution of the root and let the transition probability matrices along the branches be given as

$$\mathbf{P}^{(r1)}=\begin{pmatrix}\alpha_0 & \alpha_1\\ \alpha_1 & \alpha_0\end{pmatrix},\quad \mathbf{P}^{(r2)}=\begin{pmatrix}\beta_0 & \beta_1\\ \beta_1 & \beta_0\end{pmatrix},\quad \mathbf{P}^{(r3)}=\begin{pmatrix}\gamma_0 & \gamma_1\\ \gamma_1 & \gamma_0\end{pmatrix}.$$

Then the algebraic statistical model is defined by the mapping $f:\mathbb{R}^4\to\mathbb{R}^8$ with marginal probabilities

$$
\begin{aligned}
p_{000} &= \pi_0\alpha_0\beta_0\gamma_0 + \pi_1\alpha_1\beta_1\gamma_1,\\
p_{001} &= \pi_0\alpha_0\beta_0\gamma_1 + \pi_1\alpha_1\beta_1\gamma_0,\\
p_{010} &= \pi_0\alpha_0\beta_1\gamma_0 + \pi_1\alpha_1\beta_0\gamma_1,\\
p_{011} &= \pi_0\alpha_0\beta_1\gamma_1 + \pi_1\alpha_1\beta_0\gamma_0,\\
p_{100} &= \pi_0\alpha_1\beta_0\gamma_0 + \pi_1\alpha_0\beta_1\gamma_1,\\
p_{101} &= \pi_0\alpha_1\beta_0\gamma_1 + \pi_1\alpha_0\beta_1\gamma_0,\\
p_{110} &= \pi_0\alpha_1\beta_1\gamma_0 + \pi_1\alpha_0\beta_0\gamma_1,\\
p_{111} &= \pi_0\alpha_1\beta_1\gamma_1 + \pi_1\alpha_0\beta_0\gamma_0.
\end{aligned}
$$

The discrete Fourier transform gives a linear change of coordinates in the parameter space by using (7.56),

$$
\begin{aligned}
&\pi_0 = \tfrac{1}{2}(r_0 + r_1),\ \alpha_0 = \tfrac{1}{2}(a_0 + a_1),\ \beta_0 = \tfrac{1}{2}(b_0 + b_1),\ \gamma_0 = \tfrac{1}{2}(c_0 + c_1),\\
&\pi_1 = \tfrac{1}{2}(r_0 - r_1),\ \alpha_1 = \tfrac{1}{2}(a_0 - a_1),\ \beta_1 = \tfrac{1}{2}(b_0 - b_1),\ \gamma_1 = \tfrac{1}{2}(c_0 - c_1).
\end{aligned}
$$

Simultaneously, it provides a linear change of coordinates in the probability space by making use of (7.58),

$$
q_{ijk} = \sum_{r=0}^{1}\sum_{s=0}^{1}\sum_{t=0}^{1}(-1)^{ir+js+kt}p_{rst}.
$$

More specifically, we obtain

$$
\begin{aligned}
q_{000} &= p_{000} + p_{001} + p_{010} + p_{011} + p_{100} + p_{101} + p_{110} + p_{111},\\
q_{001} &= p_{000} - p_{001} + p_{010} - p_{011} + p_{100} - p_{101} + p_{110} - p_{111},\\
q_{010} &= p_{000} + p_{001} - p_{010} - p_{011} + p_{100} + p_{101} - p_{110} - p_{111},\\
q_{011} &= p_{000} - p_{001} - p_{010} + p_{011} + p_{100} - p_{101} - p_{110} + p_{111},\\
q_{100} &= p_{000} + p_{001} + p_{010} + p_{011} - p_{100} - p_{101} - p_{110} - p_{111},\\
q_{101} &= p_{000} - p_{001} + p_{010} - p_{011} - p_{100} + p_{101} - p_{110} + p_{111},\\
q_{110} &= p_{000} + p_{001} - p_{010} - p_{011} - p_{100} - p_{101} + p_{110} + p_{111},\\
q_{111} &= p_{000} - p_{001} - p_{010} + p_{011} - p_{100} + p_{101} + p_{110} - p_{111}.
\end{aligned}
$$

After these coordinate changes, the model has the monomial representation

$$
\begin{aligned}
q_{000} &= r_0a_0b_0c_0,\\
q_{001} &= r_1a_0b_0c_1,\\
q_{010} &= r_1a_0b_1c_0,\\
q_{011} &= r_0a_0b_1c_1,\\
q_{100} &= r_1a_1b_0c_0,\\
q_{101} &= r_0a_1b_0c_1,\\
q_{110} &= r_0a_1b_1c_0,\\
q_{111} &= r_1a_1b_1c_1.
\end{aligned}
$$

This model is toric and the phylogenetic invariants are given by binomials that can be established by the following program,

```
> ring r = 0, (r(0..1),a(0..1),b(0..1),c(0..1),q(0..7)), dp;
> ideal i0 = q(0)-r(0)*a(0)*b(0)*c(0);
> ideal i1 = q(1)-r(1)*a(0)*b(0)*c(1);
> ideal i2 = q(2)-r(1)*a(0)*b(1)*c(0);
> ideal i3 = q(3)-r(0)*a(0)*b(1)*c(1);
> ideal i4 = q(4)-r(1)*a(1)*b(0)*c(0);
> ideal i5 = q(5)-r(0)*a(1)*b(0)*c(1);
> ideal i6 = q(6)-r(0)*a(1)*b(1)*c(0);
> ideal i7 = q(7)-r(1)*a(1)*b(1)*c(1);
> ideal i = i0+i1+i2+i3+i4+i5+i6+i7;
> ideal j = std(i);
> eliminte (j, r(0)*r(1)*a(0)*a(1)*b(0)*b(1)*c(0)*c(1));
```

The output provides the following invariants,

```
_[1]=q(1)*q(6)-q(0)*q(7)
_[2]=q(2)*q(5)-q(0)*q(7)
_[3]=q(0)*q(4)-q(0)*q(7)
_[4]=q(2)*q(4)*q(6)-q(2)*q(6)*q(7)
_[5]=q(1)*q(4)*q(6)-q(1)*q(6)*q(7)
```

♢

Example 7.31. Consider the $1, n$ claw tree with root r and $n \geq 1$ leaves and take an abelian group G of order n. Let π denote the probability distribution of the root and let the transition probability matrices $\mathbf{P}^{(ri)}$, $1 \leq i \leq n$, along the branches be given as

$$\mathbf{P}^{(ri)}(X_i = g \mid X_r = h) = f^{(ri)}(g - h), \quad g, h \in G, \, 1 \leq i \leq n. \tag{7.63}$$

The joint probability of the group based model is then given by

$$\begin{aligned} p(g_1, \ldots, g_n) = P(X_1 = g_1, \ldots, X_n = g_n) &= \sum_{h \in G} \pi(h) \mathbf{P}^{(ri)}(X_i = g_i \mid X_r = h) \\ &= \sum_{h \in G} \pi(h) \prod_{i=1}^{n} f^{(ri)}(g_i - h). \end{aligned}$$

In order to find the discrete Fourier transform of this probability density with respect to the group G^n, the root distribution is replaced by the new function $\tilde{\pi} : G^n \to \mathbb{C}$ as follows,

$$\tilde{\pi}(h_1, \ldots, h_n) = \begin{cases} \pi(h_1), \text{ if } h_1 = \ldots = h_n, \\ 0, \qquad \text{otherwise.} \end{cases} \tag{7.64}$$

This definition gives

$$p(g_1, \ldots, g_n) = \sum_{(h_1, \ldots, h_n) \in G^n} \tilde{\pi}(h_1, \ldots, h_n) \prod_{i=1}^{n} f^{(ri)}(g_i - h_i). \tag{7.65}$$

If we define

$$f(g_1, \ldots, g_n) = \prod_{i=1}^{n} f^{(ri)}(g_i), \quad g_1, \ldots, g_n \in G, \tag{7.66}$$

the joint probability distribution p can be written as convolution of two functions on G^n,

$$p(g_1, \ldots, g_n) = (\tilde{\pi} * f)(g_1, \ldots, g_n), \quad g_1, \ldots, g_n \in G. \tag{7.67}$$

Taking the discrete Fourier transform yields

$$q(\chi_1, \ldots, \chi_n) = \hat{\tilde{\pi}}(\chi_1, \ldots, \chi_n) \cdot \hat{f}(\chi_1, \ldots, \chi_n). \tag{7.68}$$

In particular, the discrete Fourier transform of the function f has the form

$$\begin{aligned}
\hat{f}(\chi_1, \ldots, \chi_n) &= \sum_{(g_1,\ldots,g_n)\in G^n} f(g_1, \ldots, g_n)\overline{(\chi_1, \ldots, \chi_n)(g_1, \ldots, g_n)} \\
&= \sum_{(g_1,\ldots,g_n)\in G^n} \prod_{i=1}^{n} f^{(ri)}(g_i) \prod_{i=1}^{n} \overline{\chi_i(g_i)} \\
&= \prod_{i=1}^{n} \sum_{g_i \in G} f^{(ri)}(g_i)\overline{\chi_i(g_i)} \\
&= \prod_{i=1}^{n} \widehat{f^{(ri)}}(\chi_i).
\end{aligned} \tag{7.69}$$

Moreover, the discrete Fourier transform of the root distribution is

$$\begin{aligned}
\hat{\tilde{\pi}}(\chi_1, \ldots, \chi_n) &= \sum_{(g_1,\ldots,g_n)\in G^n} \tilde{\pi}(g_1, \ldots, g_n)\overline{(\chi_1, \ldots, \chi_n)(g_1, \ldots, g_n)} \\
&= \sum_{g\in G} \pi(g)\overline{(\chi_1, \ldots, \chi_n)(g, \ldots, g)} \\
&= \sum_{g\in G} \pi(g) \prod_{i=1}^{n} \overline{\chi_i(g)} \\
&= \widehat{\pi}(\chi_1 \cdots \chi_n).
\end{aligned} \tag{7.70}$$

It follows that the discrete Fourier transform of the joint probabilities has the monomial representation

$$q(\chi_1, \ldots, \chi_n) = \widehat{\pi}(\chi_1 \cdots \chi_n) \cdot \prod_{i=1}^{n} \widehat{f^{(ri)}}(\chi_i). \tag{7.71}$$

♢

This example is the base case for the induction in the general case. Given a rooted binary tree T with root r and n leaves. For each node v in T different from the root, write $a(v)$ for the unique parent of v in T. The transition from $a(v)$ to v is given by the substitution matrix $\mathbf{P}^{(v)}$. Suppose the

states of the random variables are the elements of a finite abelian group G. Then the joint probability distribution of the labelling of the leaves can be written as

$$p(g_1, \ldots, g_n) = \sum \pi(g_r) \prod_{\substack{v \in V(T) \\ v \neq r}} \mathbf{P}^{(v)}_{g_{a(v)}, g_v}, \tag{7.72}$$

where the sum extends over all states of the interior nodes of the tree T. We assume that the transition matrix entries $\mathbf{P}^{(v)}_{g_{a(v)}, g_v}$ depend only on the difference of the group elements $g_{a(v)}$ and g_v. We denote this entry by $f^{(v)}(g_{a(v)} - g_v)$. Thus the group based model has the joint probability distribution

$$p(g_1, \ldots, g_n) = \sum \pi(g_r) \prod_{\substack{v \in V(T) \\ v \neq r}} f^{(v)}(g_{a(v)} - g_v). \tag{7.73}$$

Theorem 7.32. *Given the joint probability distribution $p(g_1, \ldots, g_n)$ of a group-based model parametrized in (7.73). The corresponding discrete Fourier transform has the form*

$$q(\chi_1, \ldots, \chi_n) = \widehat{\pi}(\chi_1 \cdots \chi_n) \cdot \prod_{\substack{v \in V(T) \\ v \neq r}} \widehat{f^{(v)}}\Big(\prod_{l \in \Lambda(v)} \chi_l \Big), \tag{7.74}$$

where $\Lambda(v)$ is the set of leaves which have the node v as a common ancestor.

The formula (7.72) is a polynomial representation of the evolutionary model, while formula (7.74) provides a monomial representation of the same model. Since the groups G and $\hat{G}$ are isomorphic, the monomial representation can be rewritten as follows,

$$q_{g_1, \ldots, g_n} \mapsto \widehat{\pi}(g_1 + \ldots + g_n) \cdot \prod_{\substack{v \in V(T) \\ v \neq r}} \widehat{f^{(v)}}\Big(\sum_{l \in \Lambda(v)} g_l \Big). \tag{7.75}$$

We can regard this formula as the monomial mapping from a polynomial ring in $|G|^n$ unknowns

$$q_{g_1, \ldots, g_n} = q(g_1, \ldots, g_n) \tag{7.76}$$

to the polynomial ring in the unknowns $\widehat{\pi}(g)$ and $\widehat{f^{(v)}}(g)$, which are indexed by the nodes of T and the elements of G.

8

Computational Statistics

Computational statistics is a fast growing field in statistical research and applications. This chapter is concerned with the study of ideals defining statistical models for discrete random variables. The binomial generating sets of these ideals are known as Markov bases. Besides giving an algebraic description of the set of probability distributions coming from the model, the major use of Markov bases is as a tool in Markov chain Monte Carlo algorithms for generating random samples from a statistical model. Several important statistical models are addressed such as contingency tables, the Hardy-Weinberg law, and logistic regression.

8.1 Markov Bases

Markov bases are used in algebraic statistics to estimate the goodness of a fit of empirical data to a statistical model. In algebraic geometry Markov bases are equivalent to generating sets of toric ideals.

In order to introduce Markov bases, let $A \in \mathbb{Z}_{\geq 0}^{d \times n}$ be an integral matrix. Note that the matrix A provides a $\mathbb{Z}$-module homomorphism $\phi : \mathbb{Z}^n \to \mathbb{Z}^d : x \mapsto Ax$. The kernel of this mapping,

$$\ker \phi = \{u \in \mathbb{Z}^n \mid Au = 0\},$$

is a free $\mathbb{Z}$-submodule of $\mathbb{Z}^n$.

Let b be a vector in $\mathbb{Z}^d$. The set of all non-negative integral vectors u that have the marginal $b = Au$ is the *b-fibre* of A, denoted as $A^{-1}[b]$; i.e.,

$$A^{-1}[b] = \{u \in \mathbb{Z}_{\geq 0}^n \mid Au = b\}. \tag{8.1}$$

In statistical terms, the vector b is called a *sufficient statistic*.

A vector $m \in \mathbb{Z}^n$ is a *move* for A if $Am = 0$, i.e., if m lies in the kernel of the linear mapping induced by A.

Each integral vector $m \in \mathbb{Z}^n$ decomposes into a unique difference

$$m = m^+ - m^-,$$

where $m^+ \geq 0$ and $m^- \geq 0$ are the *positive* and *negative* parts of m, respectively. We have

$$m_i^+ = \max\{m_i, 0\} \quad \text{and} \quad m_i^- = -\min\{m_i, 0\}, \quad 1 \le i \le n. \tag{8.2}$$

For instance, $(2, -3, -1) = (2, 0, 0) - (0, 3, 1)$. If m is a move for A, then $Am = 0$ and thus the positive part m^+ and the negative part m^- have the same marginal: $Am^+ = Am^-$.

Example 8.1 (Maple). The matrix $A = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$ has the kernel $\mathbb{Z}b_1 \oplus \mathbb{Z}b_2$ with $b_1 = (-3, 0, 1)^t$ and $b_2 = (-2, 1, 0)^t$. It follows that for any marginal $b \in \mathbb{Z}$ with $Au_0 = b$ for some $u_0 \in \mathbb{Z}^3$, the b-fibre is given by all points $u = u_0 + z_1 b_1 + z_2 b_2 \in \mathbb{Z}^3_{\ge 0}$ for some $z_1, z_2 \in \mathbb{Z}$. ♢

A move m for A is called *applicable* to a vector u in the b-fibre of A if $u + m \ge 0$. Let B be a finite set of moves for A, and let u and v be elements in the b-fibre of A. We say that u is *connected to* v in the b-fibre of A if there exists a sequence of moves $m_1, \ldots, m_k$ in B such that

$$v = u + \sum_{i=1}^{k} m_i \quad \text{and} \quad u + \sum_{i=1}^{h} m_i \ge 0, \quad 1 \le h \le k. \tag{8.3}$$

That is, it is possible to pass from u to v by moves in B without causing negative entries on the way.

Obviously, the notion of connectedness is reflexive and transitive. It is symmetric if with each move m in B also $-m$ is a move in B. If connectedness by B is also symmetric, then it forms an equivalence relation on the b-fibre of A and thus the b-fibre is partitioned into disjoint equivalence classes by moves in B. These classes are called the *B-equivalence classes* of the b-fibre of A. The B-equivalence classes can be considered as a graph; this is an undirected graph $G = G_{b,B}$ with set of vertices $V = A^{-1}[b]$ and edges between nodes $u, v \in A^{-1}[b]$ if $v = u + m$ for some move $m \in B$.

Example 8.2. Take the matrix $A = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$ and the following set of moves for A,

$$B = \{\pm(0, 3, -2)^t, \pm(1, -2, 1)^t, \pm(1, 1, -1)^t, \pm(2, -1, 0)^t\}.$$

The points $u = (3, 2, 7)^t$ and $v = (6, 5, 4)^t$ are connected, since

$$v = u + (0, 3, -2)^t + (1, -2, 1)^t + 2 \cdot (1, 1, -1)^t.$$

♢

A finite set of moves B is called a *Markov basis* of A if for each marginal $b \in \mathbb{Z}^n$, the b-fibre of A itself constitutes a single B-equivalence class; that is, the graph $G_{b,B}$ is fully connected for all marginals b.

A Markov basis B is called *minimal* if no proper subset of B is a Markov basis. A minimal Markov basis always exists, because from any Markov basis we can remove redundant elements one by one, until none of the remaining elements can be removed any further.

Let B be a finite set of moves for A. Define the corresponding *ideal of moves* in the polynomial ring $\mathbb{Q}[X_1, \ldots, X_n]$ as

$$I_B = \langle X^{m^+} - X^{m^-} \mid m \in B \rangle. \tag{8.4}$$

Example 8.3. In view of the above example, X_B is generated by the binomials $X_2^3 - X_3^2$, $X_1X_3 - X_2^2$, $X_1X_2 - X_3$, and $X_1^2 - X_2$. ♢

We study the relationship between the ideal I_B and the toric ideal I_A associated with the matrix A. By Prop. 1.41, the latter ideal is defined as

$$I_A = \langle X^v - X^u \mid Av = Au,\, v, u \in \mathbb{Z}^n_{\geq 0} \rangle.$$

Proposition 8.4. *We have $I_B \subseteq I_A$.*

Proof. Let $m \in B$ be a move with $m = m^+ - m^-$. Then $0 = Am = A(m^+ - m^-) = Am^+ - Am^-$ and thus $X^{m^+} - X^{m^-}$ belongs to I_A. □

Proposition 8.5. *Let u and v be elements of $\mathbb{Z}^n_{\geq 0}$. If u and v are connected in a fibre of A, the binomial $X^u - X^v$ lies in the ideal I_B.*

Proof. Let m be a move in B such that $v = u + m$. Then $X^v - X^u = X^{u-m^-}(X^{m^+} - X^{m^-})$ lies in I_B.

Let u and v be connected as in (8.3). Then put $w = u + \sum_{i=1}^{k-1} m_i$. The points u, w and w, v are connected and thus by induction hypthesis, $X^v - X^w$ and $X^w - X^u$ lie in I_B. Hence, their sum $X^v - X^u$ belongs to I_B. □

Theorem 8.6. *If B is a Markov basis of A, then $I_B = I_A$.*

Proof. Let $u, v \in \mathbb{Z}^n_{\geq 0}$ and let $X^u - X^{\boldsymbol{v}}$ be a generating binomial of I_A. Then $Au = Av$ and thus the vectors u and v are in the same b-fibre of A. Since B is a Markov basis of A, the points u and v are connected in the fibre. Hence, by Prop. 8.5, the binomial $X^v - X^u$ belongs to I_B. The other inclusion has already been given in Prop. 8.4. □

Proposition 8.7. *Let $\{X^{m^+} - X^{m^-} \mid m \in B\}$ be a generating set of the toric ideal I_A in $\mathbb{Q}[X_1, \ldots, X_n]$. Then the set $\pm B$ is a Markov basis of A.*

Proof. Let $u, v \in \mathbb{Z}^n_{\geq 0}$ and let $X^u - X^{\boldsymbol{v}}$ be a generating binomial of I_A. By hypothesis, we can write

$$X^v - X^u = \sum_{i=1}^{k} X^{w_i}(X^{m_i^+} - X^{m_i^-}), \quad m_i \in B,\, w_i \in \mathbb{Z}^n_{\geq 0},\, 1 \leq i \leq k.$$

Claim that the vectors $m_1, \ldots, m_k$ connect u to v. Indeed, in case of $k = 1$, we have

$$X^v - X^u = X^{w_1}(X^{m_1^+} - X^{m_1^-}), \quad m_1 \in B,\, w_1 \in \mathbb{Z}^n_{\geq 0}.$$

By comparing the binomials, we have $v = w_1 + m_1^+$ and $u = w_1 + m_1^-$. Thus $w_1 = u - m_1^-$ and hence $v = u - m_1^- + m_1^+ = u + m$, as required.

Let $k > 1$. The case $k = 1$ gives $X^w - X^u = X^{w_1}(X^{m_1^+} - X^{m_1^-})$ and $w = u + m_1 \geq 0$, where $m_1 \in B$ and $w_1 \in \mathbb{Z}^n_{\geq 0}$. Thus

$$X^v - X^u = (X^w - X^u) + \sum_{i=2}^{k} X^{w_i}(X^{m_i^+} - X^{m_i^-}), \quad m_i \in B,\, w_i \in \mathbb{Z}^n_{\geq 0},\, 2 \leq i \leq k,$$

and hence

$$X^v - X^w = \sum_{i=2}^{k} X^{w_i}(X^{m_i^+} - X^{m_i^-}), \quad m_i \in B,\, w_i \in \mathbb{Z}^n_{\geq 0},\, 2 \leq i \leq k.$$

By induction hypothesis, we have $v = w + \sum_{i=2}^{k} m_i$ and $w + \sum_{i=2}^{w} m_i \geq 0$ for each $2 \leq h \leq k$. Thus m_1 connects u to w and $m_2, \ldots, m_k$ connect w to v. The claim follows. □

Finally, the problem is to compute a Markov basis of an integral matrix $A \in \mathbb{Z}_{\geq 0}^{d\times n}$. For this, consider the following ideal in $\mathbb{Q}[X_1,\ldots,X_n,Y_1,\ldots,Y_d]$,

$$J = \langle X_i - Y_1^{a_{1i}} \cdots Y_d^{a_{di}} \mid 1 \leq i \leq n\rangle. \tag{8.5}$$

In view of section 1.8, we have the following result.

Proposition 8.8. *A Markov basis of the matrix A is given by computing a Groebner basis of the ideal J with respect to an elimination ordering for $Y_1,\ldots,Y_d$ and then taking only those elements which involve only the indeterminates $X_1,\ldots,X_n$.*

Example 8.9 (Singular). Reconsider the matrix $A = \begin{pmatrix}1 & 2 & 3\end{pmatrix}$. Take the corresponding ideal $J = \langle X_1 - Y, X_2 - Y^2, X_3 - Y^3\rangle$ in $\mathbb{Q}[X_1,X_2,X_3,Y]$. A reduced Groebner basis of the elimination ideal I_A of J with respect to an elimination ordering for Y can be computed as follows,

```
> ring r = 0, (y, x(1..3)), dp;
> ideal j = x(1)-y, x(2)-y^2, x(3)-y^3;
> eliminate(std(j),y);
_[1]=x(2)^2-x(1)*x(3)
_[2]=x(1)*x(2)-x(3)
_[3]=x(1)^2-x(2)
```

The reduced Groebner basis of I_A is $G = \{X_2^2 - X_1X_3, X_1X_2 - X_3, X_1^2 - X_2\}$. Thus the associated Markov basis of A is $\{\pm(1,-2,1)^t, \pm(1,1,-1)^t, \pm(2,-1,0)^t\}$. ♢

8.2 Markov Chains

We consider discrete time, discrete state space Markov chains. A Markov chain is an infinite sequence of random variables (X_t) indexed by time $t \geq 0$. The set of all possible values of X_t is the state space. The state space is assumed to be a finite or countable set S. The sequence (X_t) is a *Markov chain* if it satisfies the *Markov property*,

$$P(X_{t+1} = j | X_0 = i_0, \ldots, X_{t-1} = i_{t-1}, X_t = i) = P(X_{t+1} = j | X_t = i), \tag{8.6}$$

for all states $i, i_0, \ldots, i_{t-1}, j \in S$ and $t \geq 0$. That is, the transition probabilities depend only on the current state, but not on the past.

A Markov chain is *homogeneous* if the probabilities of going from one state to another in one step are independent of the current step; that is,

$$P(X_{t+1} = j | X_t = i) = P(X_t = j | X_{t-1} = i), \quad i,j \in S,\ t \geq 1. \tag{8.7}$$

Let (X_t) be a homogeneous Markov chain and let the state space S be finite; we put $S = [n]$. Then the transition probabilities $P(X_{t+1}|X_t)$ can be represented by an $n \times n$ transition probability matrix $P = (p_{ij})$, where the entry p_{ij} is the probability that the chain provides a transition from state i to state j in one step. Note that the conservation of probability requires that the matrix P is row stochastic, i.e., $\sum_j p_{ij} = 1$.

Let $p_{ij}^{(k)}$ denote the probability that the chain moves from state i to state j in $k \geq 1$ steps. The k-step transition probabilities satisfy the Chapman-Kolmogorov equation

$$p_{ij}^{(k)} = \sum_{z \in S} p_{iz}^{(l)} p_{zj}^{(k-l)}, \quad i, j \in S,\ 0 < l < k. \tag{8.8}$$

It follows that the k-step transition probabilities are the entries of the power matrix P^k; i.e., $P^k = (p_{ij}^{(k)})$.

Let $Q = (Q(i))$ be a distribution on the state space. Then the distribution $Q' = (Q'(i))$ on the state space at the next time instant is given by

$$Q'(j) = \sum_{i \in S} p_{ij} Q(i), \quad j \in S, \tag{8.9}$$

or in shorthand notation,

$$Q' = PQ, \tag{8.10}$$

where P can be viewed as an operator on the space of probability distributions on the state set S.

An induction argument shows that the evolution of the Markov chain through $t \geq 1$ time steps is given by

$$P(X_t = j) = \sum_{i \in S} p_{ij}^{(t)} P(X_0 = i), \quad j \in S.$$

Starting from an initial distribution Q_0 on the state space S, the marginal distribution Q_t after t time steps is given by

$$Q_t = P^t Q_0, \quad t \geq 1. \tag{8.11}$$

Thus the operator on the space of probability distribution on the state set S is linear.

A distribution Q on the state space S is *stationary* if it satisfies the equation

$$Q = PQ. \tag{8.12}$$

Thus a stationary distribution is an eigenvector of the transition matrix with eigenvalue 1. The problem of finding the stationary distributions of a homogeneous Markov chain is a non-trivial task.

Example 8.10 (Maple). Consider the transition probability matrix for the Jukes-Cantor model

$$P = \begin{pmatrix} 1-3a & a & a & a \\ a & 1-3a & a & a \\ a & a & 1-3a & a \\ a & a & a & 1-3a \end{pmatrix}.$$

Since the matrix must be row stochastic, we have $0 < a < 1/3$. Taking $a = 0.1$, the 2-step and 16-step transition matrices are

$$P^2 = \begin{pmatrix} 0.52 & 0.16 & 0.16 & 0.16 \\ 0.16 & 0.52 & 0.16 & 0.16 \\ 0.16 & 0.16 & 0.52 & 0.16 \\ 0.16 & 0.16 & 0.2458 & 0.52 \end{pmatrix} \quad \text{and} \quad P^{16} = \begin{pmatrix} 0.2502 & 0.2499 & 0.2499 & 0.2499 \\ 0.2499 & 0.2502 & 0.2499 & 0.2499 \\ 0.2499 & 0.2499 & 0.2502 & 0.2499 \\ 0.2499 & 0.2499 & 0.2458 & 0.2502 \end{pmatrix}.$$

The transition probabilities in each row converge to the same stationary distribution π on the four states given by $\pi(i) = \frac{1}{4}$ for $1 \leq i \leq 4$ (Sect. 7.6).

In view of the initial distribution $(0.1, 0.2, 0.2, 0.5)$, we obtain after 2 steps and after 16 steps the distributions $(0.1960, 0.2320, 0.2320, 0.3400)$ and $(0.2499, 0.2499, 0.2500, 0.2500)$, respectively. The corresponding `Maple` code is

```
> with(LinearAlgebra);
> A := Matrix([1-3*a,a,a],[a,1-3*a,a],[a,a,1-3*a,a],[a,a,a,1-3*a]]);
> a := 0.1;
> A^2; A^16;
> v := Vector([0.1,0.2,0.2,0.5]);
> (A^2).v;
> (A^16).v;
```

♢

We study the convergence of Markov chains. For this, we consider a homogeneous Markov chain with finite state space S and transition probability matrix $P = (p_{ij})$. To this end, we need a metric on the space of probability distributions on the state space.

Proposition 8.11. *A metric on the space of probability distributions on the state space S is given as*

$$d(Q, Q') = \sum_{s \in S} |Q(s) - Q'(s)|. \tag{8.13}$$

Proof. Clearly, we have for all probability distributions Q, Q', and Q'' on S, $d(Q, Q) = 0$, $d(Q, Q') > 0$ if $Q \neq Q'$, and $d(Q, Q') = d(Q', Q)$. Finally, the triangle inequality $d(Q, Q') \leq d(Q, Q'') + d(Q'', Q')$ holds, since

$$\begin{aligned} d(Q, Q') &= \sum_{s \in S} |Q(s) - Q''(s) + Q''(s) - Q'(s)| \\ &\leq \sum_{s \in S} |Q(s) - Q''(s)| + \sum_{s \in S} |Q''(s) - Q'(s)| \\ &= d(Q, Q'') + d(Q'', Q'). \end{aligned}$$

♢

We assume that the transition probabilities satisfies the *strong ergodic condition*; that is, all transition probabilities p_{ij} are positive. First, we provide a fixpoint result that will guarantee the existence and uniqueness of fixed points.

Theorem 8.12. *If the transition probabilities P satisfy the strong ergodic condition, they define a contraction mapping with respect to the metric; that is, there is a non-negative real number $\alpha < 1$ such that for all probability distributions Q and Q' on the state space S,*

$$d(PQ, PQ') \leq \alpha \cdot d(Q, Q'). \tag{8.14}$$

Proof. Let Q and Q' be distinct distributions on S. Define

$$\Delta Q(s) = Q(s) - Q'(s), \quad s \in S.$$

We have

$$\begin{aligned} d(PQ, PQ') &= \sum_s |PQ(s) - PQ'(s)| \\ &= \sum_s \left| \sum_t P(s \mid t)Q(t) - P(s \mid t)Q'(t) \right| \\ &= \sum_s \left| \sum_t P(s \mid t)\Delta Q(t) \right|. \end{aligned}$$

We decompose the sum $\sum_t$ into to partial sums $\sum_{t^+} + \sum_{t^-}$ such that t^+ and t^- are the states where ΔQ is positive or negative, respectively. This gives us

$$\begin{aligned} d(PQ, PQ') &= \sum_s \left| \sum_{t^+} P(s \mid t^+)\Delta Q(t^+) + \sum_{t^-} P(s \mid t^-)\Delta Q'(t^-) \right| \\ &= \sum_s \left(\sum_{t^+} P(s \mid t^+)\Delta Q(t^+) - \sum_{t^-} P(s \mid t^-)\Delta Q'(t^-) \right) \\ &\quad -2 \sum_s \min \left\{ |\sum_{t^+} P(s \mid t^+)\Delta Q(t^+)|, |\sum_{t^-} P(s \mid t^-)\Delta Q(t^-)| \right\} \\ &= \sum_s \sum_t P(s \mid t)|\Delta Q(t)| \\ &\quad -2 \sum_s \min \left\{ |\sum_{t^+} P(s \mid t^+)\Delta Q(t^+)|, |\sum_{t^-} P(s \mid t^-)\Delta Q(t^-)| \right\} \\ &= \sum_t |\Delta Q(t)| - 2 \sum_s \min \left\{ |\sum_{t^+} P(s \mid t^+)\Delta Q(t^+)|, |\sum_{t^-} P(s \mid t^-)\Delta Q(t^-)| \right\} \\ &\leq \sum_t |\Delta Q(t)| - 2 \sum_s P^s_{\min} \min \left\{ |\sum_{t^+} \Delta Q(t^+)|, |\sum_{t^-} \Delta Q(t^-)| \right\} \end{aligned}$$

where we used $|(|x| - |y|)| = |x| + |y| - 2\min\{|x|, |y|\}$ and $P^s_{\min} = \min\{P(s \mid t) \mid t \in S\} > 0$. But we have

$$\sum_{t^+} \Delta Q(t^+) + \sum_{t^-} \Delta Q(t^-) = \sum_t Q(t) - \sum_t Q'(t) = 0$$

and thus

$$\left| \sum_{t^+} \Delta Q(t^+) \right| = \left| \sum_{t^-} \Delta Q(t^-) \right| = \frac{1}{2} \sum_t |\Delta Q(t)|.$$

It follows that

$$d(PQ, PQ') \leq \left(1 - \sum_s P^s_{\min}\right) d(Q, Q').$$

If we put $\alpha = 1 - P^s_{\min}$ for some $s \in S$, we obtain the desired result. ♢

Theorem 8.13. *Let the transition probabilities P satisfy the strong ergodic condition. For each probability distribution Q on the state space, the sequence $(Q, PQ, P^2Q, \ldots)$ has the property that for each number $\epsilon > 0$, there exists a positive integer N such that for all steps n and m larger than N,*

$$0 \leq d(P^nQ, P^mQ) < \epsilon. \tag{8.15}$$

The sequence $(Q, PQ, P^2Q, \ldots)$ converges to the probability distribution

$$Q^* = \lim_{n\to\infty} P^nQ \tag{8.16}$$

such that Q^* is the unique fixed point of P.

Proof. First, by repeated application of the triangle inequality and Thm. 8.12, we obtain

$$\begin{aligned}
d(P^nQ, P^mQ) &\leq d(P^NQ, P^nQ) + d(P^NQ, P^mQ), \quad \min\{m,\} \geq N \geq 0, \\
&\leq \sum_{k=0}^{n-N-1} d(P^{N+k}Q, P^{N+k+1}Q) + \sum_{l=0}^{m-N-1} d(P^{N+l}Q, P^{N+l+1}Q) \\
&\leq \sum_{k=0}^{n-N-1} \alpha^{N+k} d(Q, PQ) + \sum_{l=0}^{m-N-1} \alpha^{N+l} d(Q, PQ) \\
&= \alpha^N d(Q, PQ) \left(\sum_{k=0}^{n-N-1} \alpha^k + \sum_{l=0}^{m-N-1} \alpha^l \right) \\
&\leq \alpha^N d(Q, PQ) \left(\frac{2 - \alpha^{n-N} - \alpha^{m-N}}{1-\alpha} \right).
\end{aligned}$$

Thus

$$d(P^nQ, P^mQ) < \frac{2\alpha^N}{1-\alpha} d(Q, PQ).$$

Hence, if we put

$$N \geq \log_\alpha \left(\frac{\epsilon(1-\alpha)}{2d(Q, PQ)} \right),$$

we have

$$d(P^nQ, P^mQ) < \epsilon.$$

Thus the series (P^nQ) is a Cauchy sequence and hence is convergent by completeness of $\mathbb{R}^n$.

Second, let Q^* denote the limiting point. We have

$$0 \leq d(P^{n+1}Q, PQ^*) \leq \alpha \cdot d(P^nQ, Q^*), \quad n \geq 0.$$

But $\alpha \cdot d(P^n Q, Q^*)$ tends to 0 as n goes to infinity. Thus $P^n Q$ goes to PQ^* when n tends to infinity. Since limits are unique, it follows that $Q^* = PQ^*$.

Third, assume there is another distribution Q satisfying $PQ = Q$. Then $0 \leq d(Q, Q^*) = d(PQ, PQ^*) \leq \alpha \cdot d(Q, Q^*)$. Thus $0 \leq (1-\alpha) d(Q, Q^*) \leq 0$ which shows that $d(Q, Q^*) = 0$ and hence $Q = Q^*$. ♢

A probability distribution Q on the state space S is called a *fixed point* or a *stationary distribution* of the operator P provided that $PQ = Q$.

Example 8.14 (R). A one-dimensional random walk is a discrete-time Markov chain whose state space is given by the set of integers $S = \mathbb{Z}$. For some number p with $0 < p < 1$, the transition probabilities (the probability p_{ij} of moving from state i to state j) are given by

$$p_{ij} = \begin{cases} p & \text{if } j = i+1, \\ 1-p & \text{if } j = i-1, \\ 0 & \text{otherwise,} \end{cases} \quad \text{for all } i, j \in \mathbb{Z}.$$

In a random walk, at each transition a step of unit length is made at random to the right with probability p and to the left with probability $1-p$. A random walk can be interpreted as the betting of a gambler who bets 1 Euro on a sequence of p-Bernoulli trials and wins or loses 1 Euro at each transition; if $X_0 = 0$, the state of the process at time n is her gain or loss after n trials. The probability to start in state 0 and return to state 0 in $2n$ steps for the first time is

$$p_{00}^{(2n)} = \binom{2n}{n} p^n (1-p)^n.$$

It can be shown that $\sum_{n=1}^{\infty} p_{00}^{(2n)} < \infty$ if and only if $p \neq 1/2$. Thus the expected number of returns to 0 is finite if and only if $p \neq 1/2$. A random walk with Bernoulli probability $p = 0.7$ can be generated over a short time span as follows.

```
> n <- 400
# n p-Bernoulli trials
> X <- sample( c(-1,1), size=n, replace = TRUE, p=(0.3,0.7) )
# coerce to a data frame
> D <- as.integer( c(0,cumsum(X)) )
# cumulative sum with prepended 0
> plot ( D, type="l", main="", xlab = "i" )
```

A trajectory of the process starting at state 0 is given in Fig. 8.1.

Another way to define a one-dimensional random walk is to take a sequence $(X_t)_{t \geq 1}$ of independent, identically distributed random variables, where each variable has state space $\{\pm 1\}$. Put $S_0 = 0$ and consider the partial sums $S_n = \sum_{i=1}^{n} X_i$ for $n \geq 1$. The sequence $(S_t)_{t \geq 0}$ is a *simple random walk* on $\mathbb{Z}$. The series given by the sum of sequences of 1's and -1's provides the walking distance if each part of the walk is of unit length. In case of $p = 1/2$ we speak of a *symmetric random walk*. In a symmetric random walk, all states are recurrent; i.e., the chain returns to each state with probability 1. A symmetric random walk can be generated over a short time span as follows.

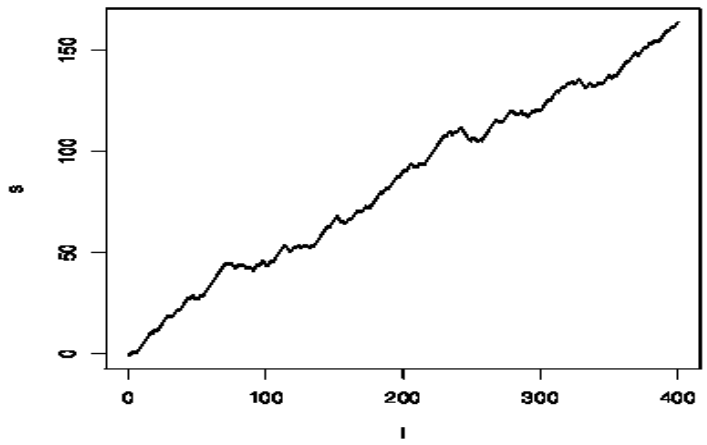


Fig. 8.1. Partial realization of a random walk with $p = 0.7$.

```
> n <- 400
# n Bernoulli trials with p=1/2
> X <- sample( c(-1,1), size=n, replace = TRUE )
# coerce to a data frame
> S <- as.integer( c(0,cumsum(X)) )
# cumulative sum with prepended 0
> plot ( S, type="l", main="", xlab = "i" )
```

A trajectory of the process starting at $S_0 = 0$ is given in Fig. 8.2. The process returns to 0 several times within the given time span. This can be seen by invoking the command `which(S==0)`. ♢

8.3 Metropolis Algorithm

The Metropolis algorithm can be used to generate random samples from a given probability distribution. The idea is to generate a Markov chain $(X_t)_{t\geq 0}$ whose stationary distribution is the target distribution. The algorithm must specify for a given state X_t how to generate the next state X_{t+1}. For this, a candidate point Y is generated from a proposed distribution $g(\cdot|X_t)$. If the candidate point is accepted, the chain moves to state Y at time $t+1$. Otherwise, the chain stays in state X_t and $X_{t+1} = X_t$.

First, we formulate the Metropolis algorithm in the context of Markov bases. For this, let $B = \{m_1, \ldots, m_l\}$ be a Markov basis for a matrix $A \in \mathbb{Z}_{\geq 0}^{d\times n}$. Given a marginal $b \in \mathbb{Z}^d$, the state set is the b-fibre of A consisting of all vectors in $\mathbb{Z}_{\geq 0}^n$ with marginal b,

$$A^{-1}[b] = \{u \in \mathbb{Z}_{\geq 0}^n \mid Au = b\}. \tag{8.17}$$

Consider the Markov chain P' given by the transition probabilities

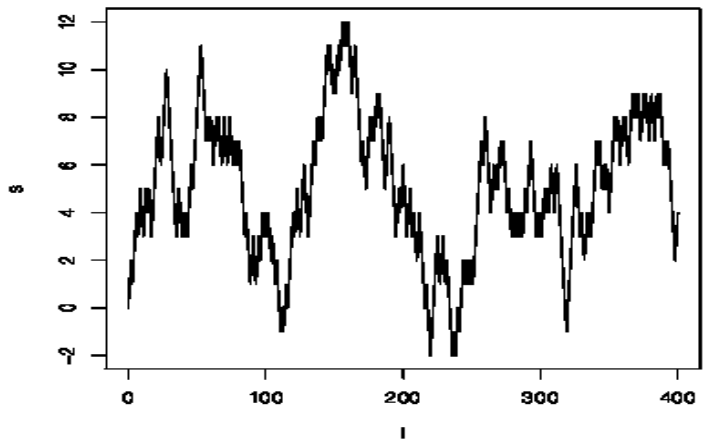


Fig. 8.2. Partial realization of a symmetric random walk.

$$p'(v|u) = \begin{cases} 1/2l \text{ if } v = u + m \geq 0 \text{ for some } m \in B, \\ 0 \quad \text{otherwise.} \end{cases} \tag{8.18}$$

Clearly, the Markov chain P' is a random walk based on the Markov basis with a uniform stationary distribution.

Fix a positive function $f : A^{-1}[b] \to \mathbb{Z}_{>0}$ on the b-fibre of A. The function f is assumed to be proportional to a given target distribution. The Metropolis construction provides a Markov chain $P = (X_t)$ on the state space $A[b^{-1}]$ with transition probabilities defined as

$$p(v|u) = \begin{cases} p'(v|u), & \text{if } v \neq u \text{ and } f(v) \geq f(u), \\ p'(v|u)\frac{f(v)}{f(u)}, & \text{if } v \neq u \text{ and } f(v) < f(u), \\ p'(u|u) + \sum_{f(v)<f(u)} p'(v|u)\left(1 - \frac{f(v)}{f(u)}\right), & \text{otherwise } (u = v). \end{cases} \tag{8.19}$$

We have $\sum_v p(v|u) = 1$ for each state u and thus P forms a Markov chain. This chain can be implemented as follows:

> Take the state u and choose a new state v from the Markov chain P'. If $f(v) \geq f(u)$, move to the state v with certainty. If $f(v) < f(u)$, perform a random experiment such that the new state v is accepted with *rejection probability* $f(v)/f(u)$. (The random experient is implemented by generating a number $r \in [0, 1]$ uniformly at random and checking if $r < f(v)/f(u)$. If so, the new state v is accepted.) Otherwise, stay at the state u.

Proposition 8.15. *The Markov chain P on the b-fibre of A defined by the transition probabilities (8.19) satisfies the* condition of detailled balance*:*

$$f(u)p(v|u) = f(v)p(u|v), \quad u, v \in A^{-1}[b]. \tag{8.20}$$

Proof. The equation is clear for $u = v$. Let $v \neq u$. If $f(v) \geq f(u)$, then $p(v|u) = p'(v|u)$ and $p(u|v) = p'(u|v)f(u)/f(v)$. Since $p'(v|u) = p'(u|v)$, the equality follows. If $f(v) < f(u)$, then $p(v|u) = p'(v|u)f(v)/f(u)$ and $p(u|v) = p'(u|v)$. Since $p'(v|u) = p'(u|v)$, the equality follows. □

In view of the algorithm, only quotients of f are considered and therefore it is sufficient to know the target distribution f up to a constant. The condition of detailled balance is sufficient such that the Markov chain becomes stationary.

Proposition 8.16. *The Markov chain P on the b-fibre of A defined by the transition probabilities (8.19) has stationary distribution proportional to f.*

Proof. Let $v \in A^{-1}[b]$. We have

$$\sum_u f(u)p(v|u) = \sum_u f(v)p(u|v) = f(v)\sum_u p(u|v) = f(v), \tag{8.21}$$

where the first equation follows from the condition of detailled balance and the last one from the conservation of probabilities. Thus the distribution proportional to $(f(u))$ is a fixed point of the transition matrix. □

Example 8.17 (R). Reconsider the matrix $A = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$. The corresponding Markov basis is $B = \{\pm(1,-2,1)^t, \pm(1,1,-1)^t, \pm(2,-1,0)^t\}$. Take the fibre $b = A(1,1,1)^t = 1+2+3 = 6$ and use as target distribution the multinomial distribution

$$f(u_1,u_2,u_3) = \frac{n!}{u_1!u_2!u_3!}p_1^{u_1}p_2^{u_2}p_3^{u_3}$$

with probabilities $p_1, p_2, p_3 > 0$ such that $p_1 + p_2 + p_3 = 1$ and $u_1, u_2, u_3 \in \mathbb{Z}_{\geq 0}$ are chosen in the fibre with $u_1 + u_2 + u_3 = 6$.

```
> B <- matrix( c(1,-2,1, 1,1,-1, 2,-1,0, -1,2,-1, -1,-1,1, -2,1,0), nrow=6, ncol=3)
> f <- function (u) {
    p <- c(0.1,0.3,0.6)
    return (p[1]^u[1]*p[2]^u[2]*p[3]^u[3])
  }
> k <- 0
> N <- 100
> z <- runif(N)   # random numbers between 0 and 1
> u <- c(1,1,1)   # starting point
> for (j in 2:N) {
# generate random integer between 1 and 6
    i <- sample(1:6, 1)
    v <- c(B[i,1]+u[1],B[i,2]+u[2],B[i,3]+u[3])
    if (f(v) > f(u))
      u <- v
    else {
      if ( z[j] - f(v)/f(u) < 0 )
         u <- v
```

```
      else
         k <- k+1   # v is rejected
    }
  }
> print(k) # number of rejections
[1] 62
```

♢

In general terms, given a target distribution f we choose a Markov chain (X_t) and a target distribution $g(\cdot|X_t)$ which fulfills some regularity conditions (irreducibility, positive recurrence, and aperiodicity). Then the Metropolis algorithm given in Fig. 8.1 will converge to the given target distribution. Note that the candidate point Y is accepted with probability

$$r(X_t, Y) = \min\{1, \frac{f(Y)g(X_t|Y)}{f(X_t)g(Y|X_t)}\}.$$

Since only quotients of the target distribution occur, it is sufficient to know the target distribution up to a constant.

Algorithm 8.1 Metropolis algorithm.

Require: S state set, f target distribution on S
Ensure: Markov chain $(X_t)_{t\geq 0}$ on S with stationary distribution f.
  Choose target distribution $g(\cdot|X_t)$ on S
  Generate X_0 from distribution g
  while (Chain not converged to stationary distribution according to some criterion) **do**
    Generate Y from $g(\cdot|X_t)$
    Generate U from uniform distribution on $[0, 1]$
    if $U \leq \frac{f(Y)g(X_t|Y)}{f(X_t)g(Y|X_t)}$ **then**
      $X_{t+1} \leftarrow Y$
    else
      $X_{t+1} \leftarrow X_t$
    end if
    $t \leftarrow t + 1$
  end while

Example 8.18 (R). A random walk can be implemented by the Metropolis algorithm. For this, a candidate point Y is generated from a symmetrical target distribution $g(Y|X_t) = g(|X_t - Y|)$ depending on the distance between the points X_t and Y. At each iteration, a random increment Z is generated from the target distribution g and the candidate point is set to $Y = X_t + Z$. The target distribution is the Student's t distribution with ν degrees of freedom and the target distribution is the normal distribution $N(X_t, \sigma^2)$. Note that the $t(\nu)$ density is proportional to

$$f(x) = (1 + x^2/\nu)^{-(\nu+1)/2}.$$

Thus by the symmetry of the target distribution, we have

$$r(x_t, y) = \frac{f(y)}{f(x_t)} = \frac{(1 + y^2/\nu)^{-(\nu+1)/2}}{(1 + x_t^2/\nu)^{-(\nu+1)/2}}.$$

The Metropolis algorithm then looks as follows.

```
> metropolis <- function(nu, sigma, x0, N) {
>   x <- numeric(N)
>   x[1] <- x0
>   z <- runif(N)
>   k <- 0
>   for (j in 2:N) {
      y <- rnorm(1, x[j-1], sigma)
      if (dt(y,nu) > dt(x[j-1],nu))
        x[j] <- y
      else {
        if ( z[j] - dt(y,nu)/dt(x[j-1],nu) < 0 )
          x[j] <- y
        else {
          x[j] <- x[j-1]
          k <- k+1
        }
      }
    }
    return(list(x=x,k=k))
  }
```

The convergence of the random walk is sensitive to the parameter choice. To this end, we have provided random walks with several choices of σ (Fig. 8.3).

```
> nu <- 4  # degrees of freedom of target Student's t distribution
> N <- 1000
> sigma <- c(0.1, 0.2, 0.5, 1.0, 1.5, 2.0, 3.0, 5.0, 10.0)
> x0 <- 20
> m1 <- metropolis( nu, sigma[1], x0, N )
> m2 <- metropolis( nu, sigma[2], x0, N )
> m3 <- metropolis( nu, sigma[3], x0, N )
> m4 <- metropolis( nu, sigma[4], x0, N )
> m5 <- metropolis( nu, sigma[5], x0, N )
> m6 <- metropolis( nu, sigma[6], x0, N )
> m7 <- metropolis( nu, sigma[7], x0, N )
> m8 <- metropolis( nu, sigma[8], x0, N )
> m9 <- metropolis( nu, sigma[9], x0, N )
# number of candidate points rejected
> print( c( m1$k, m2$k, m3$k, m4$k, m5$k, m6$k, m7$k, m8$k, m9$k ) )
[1] 1, 3, 7, 18, 28, 48, 71, 119, 169
```

♢

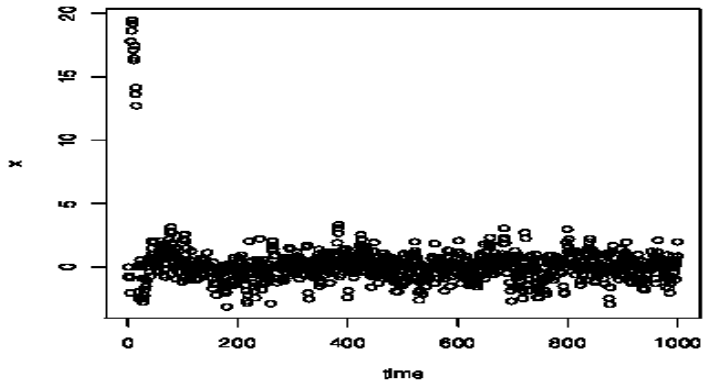


Fig. 8.3. Random walk with $\sigma = 0.5$.

8.4 Contingency Tables

In statistics, contingency tables are matrices that record the multivariate frequency distribution of two or more discrete categorial variables. They form a basic tool in business intelligence and survey research. They show a basic picture of the interrelation between two or more random variables and are used to find interactions between them.

Example 8.19. Consider the 4×4 contingency table shown in Table 8.1 that presents a classification of 592 people according to eye color and hair color. A basic question of interest for this table is whether eye color and hair color are independent features. ♢

Table 8.1. Eye color vs. hair color for 592 subjects. (The right-hand column and the bottom row contain are the marginal totals and the bottom right-hand corner cell is the grand total.)

	Hair Color				
Eye Color	Black	Brunette	Red	Blonde	Total
Brown	68	119	26	7	220
Blue	20	84	17	94	215
Hazel	15	54	14	10	93
Green	5	29	14	16	64
Total	108	286	71	127	592

Let X and Y be random variables with state sets $[r]$ and $[c]$, respectively. An $r \times c$ *contingency table* displays the frequencies of random selections from these two variables (Table 8.2). All probabilistic information about the random variables X and Y is contained in the joint probabilities

Table 8.2. Scheme of $r \times c$ contingency table for two categorial random variables.

$X \backslash Y$	$Y=1$	$\dots$	$Y=j$	$\dots$	$Y=c$	Total
$X=1$	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$X=i$	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{ic}	n_{i+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$X=r$	n_{r1}	$\dots$	n_{rj}	$\dots$	n_{rc}	n_{r+}
Total	n_{+1}	$\dots$	n_{+j}	$\dots$	n_{+c}	n_{++}

$$p_{ij} = P(X = i, Y = j), \quad 1 \le i \le r,\ 1 \le j \le c.$$

The way of which the sample data are acquired is of central importance.

- *Unrestricted sampling:* Suppose the number of observations $n = n_{++}$ is not fixed (e.g., connection between car brand and exceeding the speed limit in traffic control). Then the frequencies n_{ij} can be viewed as realizations of independent Poisson distributed random variables with mean λ_{ij},

$$P(X = i, Y = j) = e^{-\lambda_{ij}} \frac{\lambda_{ij}^{n_{ij}}}{n_{ij}!}.$$

  The maximum likelihood function is given by

$$L(\lambda, n) = \prod_{i=1}^{r} \prod_{j=1}^{c} e^{-\lambda_{ij}} \frac{\lambda_{ij}^{n_{ij}}}{n_{ij}!},$$

  and the maximum likelihood estimates of the means λ_{ij} are

$$\hat{\lambda}_{ij} = n_{ij}, \quad 1 \le i \le r,\ 1 \le j \le c.$$

- *Multinomial sampling:* Suppose the number of observations $n = n_{++}$ is fixed (e.g., connection between eye color and hair color of n persons). Then the common density is given by a multinomial distribution

$$f(n_{ij}|n) = \frac{n!}{\prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!} \prod_{i=1}^{r} \prod_{j=1}^{c} p_{ij}^{n_{ij}}.$$

  The maximum likelihood function is

$$L(p, n) = n! \prod_{i=1}^{r} \prod_{j=1}^{c} \frac{p_{ij}^{n_{ij}}}{n_{ij}!}$$

  and the maximum likelihood estimates are

$$\hat{p}_{ij} = \frac{n_{ij}}{n}, \quad 1 \le i \le r, 1 \le j \le c.$$

- *Hypergeometric sampling:* Suppose the row and column sums are fixed (e.g., classical tee test). Take the row and column sums

$$n_{i+} = \sum_{j=1}^{c} n_{ij},\ 1 \le i \le r, \quad \text{and} \quad n_{+j} = \sum_{i=1}^{r} n_{ij},\ 1 \le j \le c,$$

respectively, and the vectors of row and column sums

$$n_{.+} = (n_{1+}, \ldots, n_{r+})^t \quad \text{and} \quad n_{+.} = (n_{+1}, \ldots, n_{+c})^t,$$

respectively. Then the common density is given by the hypergeometric distribution

$$\begin{aligned} f(n_{ij}|n_{.+}, n_{+.}) &= \frac{\prod_{i=1}^{r} n_{i+}! \prod_{j=1}^{c} n_{+j}!}{n! \prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!} \\ &= \frac{\prod_{j=1}^{c} \binom{n_{+j}}{n_{1j}, \ldots, n_{rj}}}{\binom{n}{n_{1+}, \ldots, n_{r+}}}. \end{aligned} \tag{8.22}$$

In particular, in view of a 2×2 contingency table,

$$f(n_{11}|n_{.+}, n_{+.}) = \frac{\binom{n_{+1}}{n_{11}}\binom{n-n_{+1}}{n_{+1}-n_{11}}}{\binom{n}{n_{1+}}}.$$

An important problem in multivariate statistics is finding the dependence structure between categorial random variables. In inferential statistics, a *test of independence* assesses whether the observed features expressed in a contingency table are independent of each other. The null hypothesis refers to the general assertion that there is no relationship between the measured phenomena. In view of two random variables X and Y with state sets $[r]$ and $[c]$, respectively, we assume that the frequencies n_{ij} in the $r \times c$ contingency table follow a multinomial distribution. Define the *marginal probabilities*

$$p_{i+} = P(X = i), \quad 1 \le i \le r, \text{ and } p_{+j} = P(Y = j), \quad 1 \le j \le s.$$

Then the *null hypothesis* states that the random variables X and Y are independent,

$$H_0: \quad p_{ij} = P(X = i, Y = j) = p_{i+}p_{+j}, \quad 1 \le i \le r,\ 1 \le j \le c,$$

and the *alternative hypothesis* H_1 states that the opposite holds.

Under the null hypothesis, the theoretical frequencies of the outcomes are

$$\hat{p}_{i+} = \frac{n_{i+}}{n}, \quad 1 \le i \le r, \quad \text{and} \quad \hat{p}_{+j} = \frac{n_{+j}}{n}, \quad 1 \le j \le c,$$

and

$$\hat{n}_{ij} = n \cdot \hat{p}_{ij} = n\hat{p}_{i+}\hat{p}_{+j} = \frac{n_{i+}n_{+j}}{n}, \quad 1 \le i \le r,\ 1 \le j \le c.$$

Proposition 8.20. *The random variables X and Y are independent if and only if the $r \times c$ matrix $P = (p_{ij})$ has rank 1.*

Proof. Suppose X and Y are independent. Then the matrix P can be written as a product of the column vector (p_{i+}) and the row vector (p_{j-}). It follows that the matrix has rank 1.

Conversely, let P have rank 1. Then the matrix has the form $P = ab^T$, where $a \in \mathbb{R}^r$ and $b \in \mathbb{R}^c$. All entries of the matrix are non-negative and so the vectors can be chosen to have non-negative entries as well. We have $p_{ij} = a_i b_j$, $1 \le i \le r$, $1 \le j \le c$. 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 $a_+ b_+ = 1$, $1 \le i \le r$, $1 \le j \le c$. It follows that $p_{ij} = a_i b_j = a_i b_+ a_+ b_j = p_{i+} p_{+j}$, $1 \le i \le r$, $1 \le j \le c$. ♢

The test of independent can be assessed by *Pearson's chi-squared test*. For this, the statistical evaluation of the difference between the observed frequencies and the expected frequencies under the null hypothesis is based on *Pearson's cumulative test statistic*

$$\chi^2_\nu = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}. \tag{8.23}$$

This test statistic is approximately chi-squared distributed with ν degrees of freedom. The number of degrees of freedom is determined by the dependencies among the data: $\sum_{i=1}^{r} p_{i+} = 1$, $\sum_{j=1}^{c} p_{+j} = 1$, and $\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} = 1$. Thus we have

$$\nu = (r \cdot c - 1) - ((r-1) + (c-1)) = r \cdot c - r - c + 1 = (r-1)(c-1).$$

Note that for small values of n and $\hat{\chi}^2$ values < 0.1, the χ^2_ν distribution is only a coarse approximation.

The *significance level* of a statistical test is a probability threshold below which the null hypothesis will be rejected. Common values are 1% or 5%. The rejection of the null hypothesis when it is actually true is known as *type I error* or *false positive determination*. For instance, if the null hypothesis is rejected at the significance level of $\alpha = 0.05$, then on average in 5 of 100 cases a type I error will be committed.

The distribution of a test statistic $T = \chi^2$ under the null hypothesis and according the significance level α decomposes the possible values of T into two regions, one for which the null hypothesis is rejected, the so-called *critical region*, and one for which it is not. The two regions are divided according to the $(1-\alpha)$-quantile $\chi^2_{n;1-\alpha}$ of the chi-squared distribution. The probability of the critical region is α (Fig. 8.4).

Decide to reject the null hypothesis in favor of the alternative if the calculated value $\hat{\chi}^2$ of the test statistic in (8.23) is larger than the $(1-\alpha)$-quantile $\chi^2_{n;1-\alpha}$ and accept the null hypothesis otherwise (Table 8.3).

Example 8.21 (R). Reconsider the 4×4 contingency table relating eye color and hair color for $n = 592$ persons (Table 8.1). The test of independence is conducted using `R` as follows.

```
> eh <- matrix( c(68,118,26,7,20,84,17,94,15,54,14,10,5,29,14,16), ncol=4 )
> chisq.test( eh, correct=TRUE)

              Pearson's Chi-squared test

data: eh
X-squared: 138.29, df = 9, p-value < 2.2e-16
```

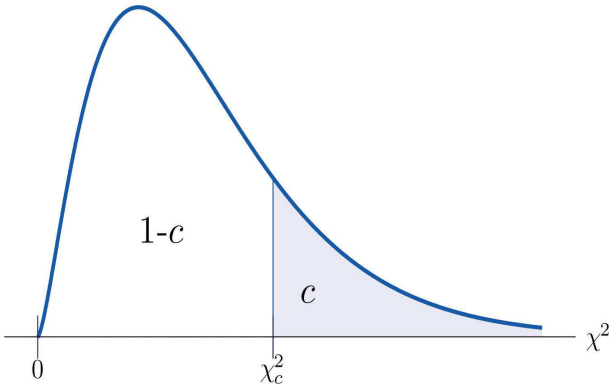


Fig. 8.4. Critical region.

Table 8.3. The 0.95-quantiles of the χ^2_ν distribution for small degrees of freedom ν with $1 \le \nu \le 9$.

ν	$\chi^2_{\nu,0.95}$
1	3.84
2	5.99
3	7.81
4	9.49
5	11.07
6	12.59
7	14.07
8	15.51
9	16.92

```
> qchisq( 0.95, 9 )
[1] 16.91898
> qchisq( 0.99, 9 )
[1] 21.66599
```

The test statistic yields $\hat{\chi}^2 = 138.29$ and the 95%-quantile of the chi-squared distribution with $\nu = 9$ degrees of freedom is $\chi^2_{9;0.95} = 16.91898$. Thus the null hypothesis must be strongly rejected at the significance level of 5%. A similar result holds for the significance level of 1%. ♢

The *p-value* is a statistic defined as the probability of obtaining a result equal to or more extreme than what was actually observed under the assumption that the null hypothesis is true. If the p-value is smaller than the significance level, the observed data are inconsistent with the null hypothesis and therefore the null hypothesis must be rejected.

Finally, we consider two-way contingency tables with fixed row and column sums. The probability of such an $r \times c$ two-way contingency table is given by (8.22). Generally, it is difficult to provide a complete list of all those tables. An alternative is sampling in the fibre of a given table by using the Metropolis algorithm and Markov bases. In order to find a Markov basis for two-way $r \times c$ contingency

tables with fixed row and column sums, consider the integral $(r+c) \times rc$ matrix

$$A_{r,c} = \begin{pmatrix} \mathbf{1}_r^T \otimes \boldsymbol{I}_c \\ \boldsymbol{I}_r \otimes \mathbf{1}_c^T \end{pmatrix} \tag{8.24}$$
$$= \begin{pmatrix} 1 \dots 1\, 0 \dots 0 \dots 0 \dots 0 \\ 0 \dots 0\, 1 \dots 1 \dots 0 \dots 0 \\ \\ 0 \dots 0\, 0 \dots 0 \dots 1 \dots 1 \\ 1 \quad 0\, 1 \quad 0 \quad 1 \quad 0 \\ \quad \ddots \quad \ddots \quad \dots \quad \ddots \\ 0 \quad 1\, 0 \quad 1 \quad 0 \quad 1 \end{pmatrix},$$

where $\mathbf{1}_r$ is the all-one vector of lenght r, $\boldsymbol{I}_r$ is the $r \times r$ identity matrix, and $\otimes$ denotes the Kronecker product.

If (n_{ij}) denotes an $r \times c$ contingency table with vectors of row and column sums $n_{\cdot +}$ and $n_{+ \cdot}$, respectively, and the table is written row-wise as a column vector, we obtain

$$A_{r,c}(n_{ij}) = \begin{pmatrix} n_{\cdot +} \\ n_{+ \cdot} \end{pmatrix}. \tag{8.25}$$

The vector on the right-hand side

$$b = \begin{pmatrix} n_{\cdot +} \\ n_{+ \cdot} \end{pmatrix}$$

forms a sufficient statistic of the model.

Proposition 8.22. *A reduced Groebner basis of the ideal $I_{A_{r,c}}$ in the polynomial ring $\mathbb{Q}[\{X_{ij}\}_{i \in [r], j \in [c]}]$ is given by*

$$G_{r,c} = \{X_{il}X_{jk} - X_{ik}X_{jl} \mid 1 \le i < j \le r,\, 1 \le k < l \le c\}.$$

Let $\boldsymbol{e}_{ij}$ denote the standard unit table, which has a 1 in the (i,j) position, and zeros elsewhere. Then by Prop. 8.7, we obtain the following result.

Proposition 8.23. *The minimal Markov basis of the matrix $A_{r,c}$ corresponding to the Groebner basis $G_{r,c}$ consists of $2 \cdot \binom{r}{2}\binom{c}{2}$ moves given by*

$$B_{r,c} = \{\pm(\boldsymbol{e}_{il} + \boldsymbol{e}_{jk} - \boldsymbol{e}_{ik} - \boldsymbol{e}_{jl}) \mid 1 \le i < j \le r,\, 1 \le k < l \le c\}.$$

It follows that the Markov basis consists of all tables which have the following 2×2 minors and zeros elsewhere,

$$\begin{matrix} +1 & -1 \\ -1 & +1 \end{matrix} \quad \text{and} \quad \begin{matrix} -1 & +1 \\ +1 & -1 \end{matrix} \tag{8.26}$$

Example 8.24 (Singular). The 3×2 contingency tables with fixed row and column sums are described by the matrix

$$A_{3,2} = \begin{pmatrix} 1\,1\,0\,0\,0\,0 \\ 0\,0\,1\,1\,0\,0 \\ 0\,0\,0\,0\,1\,1 \\ 1\,0\,1\,0\,1\,0 \\ 0\,1\,0\,1\,0\,1 \end{pmatrix}. \tag{8.27}$$

Then we have

$$A_{3,2} \begin{pmatrix} n_{11} \\ n_{12} \\ n_{21} \\ n_{22} \\ n_{31} \\ n_{32} \end{pmatrix} = \begin{pmatrix} n_{1+} \\ n_{2+} \\ n_{3+} \\ n_{+1} \\ n_{+2} \end{pmatrix}. \tag{8.28}$$

In view of the computation of a minimal Markov basis for the matrix $A_{3,2}$, consider the ideal

$$J = \langle X_{11} - Y_1Y_4, X_{12} - Y_1Y_5, X_{21} - Y_2Y_4, X_{22} - Y_2Y_5, X_{31} - Y_3Y_4, X_{32} - Y_3Y_5 \rangle.$$

A reduced Groebner basis of the ideal $I_{A_{3,2}}$ is given by

$$G_{3,2} = \{X_{12}X_{21} - X_{11}X_{22}, X_{12}X_{31} - X_{11}X_{32}, X_{22}X_{31} - X_{21}X_{32}\}$$

which can be calculated as follows,

```
> ring r = 0, (x11,x12,x21,x22,x31,x32,y(1..5)), dp;
> ideal i = x11-y(1)*y(4), x12-y(1)*y(5), x21-y(2)*y(4),
            x22-y(2)*y(6), x31-y(3)*y(4), x32-y(3)*y(6);
> ideal j = std(i);
> eliminate( j, y(1)*y(2)*y(3)*y(4)*y(5) );
_[1]=x22*x31-x21*x32
_[2]=x12*x31-x11*x32
_[3]=x12*x21-x11*x22
```

Therefore, the corresponding Markov basis is

$$B_{3,2} = \{\pm(\boldsymbol{e}_{12} + \boldsymbol{e}_{21} - (\boldsymbol{e}_{11} + \boldsymbol{e}_{22})), \pm(\boldsymbol{e}_{12} + \boldsymbol{e}_{31} - (\boldsymbol{e}_{11} + \boldsymbol{e}_{32})), \pm(\boldsymbol{e}_{22} + \boldsymbol{e}_{31} - (\boldsymbol{e}_{21} + \boldsymbol{e}_{32}))\},$$

where

$$\pm(\boldsymbol{e}_{12} + \boldsymbol{e}_{21} - \boldsymbol{e}_{11} - \boldsymbol{e}_{22}) = \pm \begin{pmatrix} -1 & +1 \\ +1 & -1 \\ 0 & 0 \end{pmatrix},$$

$$\pm(\boldsymbol{e}_{12} + \boldsymbol{e}_{31} - \boldsymbol{e}_{11} - \boldsymbol{e}_{32}) = \pm \begin{pmatrix} -1 & +1 \\ 0 & 0 \\ +1 & -1 \end{pmatrix},$$

$$\pm(\boldsymbol{e}_{22} + \boldsymbol{e}_{31} - \boldsymbol{e}_{21} - \boldsymbol{e}_{32}) = \pm \begin{pmatrix} 0 & 0 \\ -1 & +1 \\ +1 & -1 \end{pmatrix}.$$

♢

The usual approach to perform hypothesis testing for contingency tables is the asymptotic one, which involves chi-squared distributions. In many cases, especially when the table is sparse, the chi-squared approximation may not be adequate. If so, we can approximate the test statistic via the Metropolis algorithm, drawing contingency tables in the given fibre according to the hypergeometric distribution.

Given an $r \times c$ contingency table (n_{ij}) in the fibre $A^{-1}[b]$ given by the vectors $n_{.+}$ and $n_{+.}$. First, pick a Markov move $m \in B$ uniformly at random. The move gives one of two 2×2 minors

$$\begin{matrix} +1 & -1 \\ -1 & +1 \end{matrix} \quad \text{or} \quad \begin{matrix} -1 & +1 \\ +1 & -1 \end{matrix} \tag{8.29}$$

Second, calculate the new table $(n'_{ij}) = (n_{ij}) + (m_{ij}) = (n_{ij} + m_{ij})$. If the new table satisfies $(n'_{ij}) \geq 0$, move to the new table with probability $\min\{1, f(n'_{ij}|n_{.+}, n_{+.})/f(n_{ij}|n_{.+}, n_{+.})\}$; otherwise, stay at the table (n_{ij}).

The rejection probability is given as follows,

$$\begin{aligned} \frac{f(n'_{ij}|n_{.+}, n_{+.})}{f(n_{ij}|n_{.+}, n_{+.})} &= \left(\frac{\prod_{j=1}^{c} \binom{n_{+j}}{n_{1j}+m_{1j},\ldots,n_{rj}+m_{rj}}}{\binom{n}{n_{1+},\ldots,n_{r+}}} \right) \left(\frac{\prod_{j=1}^{c} \binom{n_{+j}}{n_{1j},\ldots,n_{rj}}}{\binom{n}{n_{1+},\ldots,n_{r+}}} \right)^{-1} \\ &= \prod_{j=1}^{c} \frac{\binom{n_{+j}}{n_{1j}+m_{1j},\ldots,n_{rj}+m_{rj}}}{\binom{n_{+j}}{n_{1j},\ldots,n_{rj}}} \\ &= \prod_{j=1}^{c} \frac{\prod_{i=1}^{r} n_{ij}!}{\prod_{i=1}^{r} (n_{ij} + m_{ij})!} \\ &= \prod_{\substack{i,j \\ m_{ij} \neq 0}} \frac{n_{ij}!}{(n_{ij} + m_{ij})!} \\ &= \prod_{\substack{i,j \\ m_{ij} = +1}} \frac{n_{ij}!}{(n_{ij} + 1)!} \prod_{\substack{i,j \\ m_{ij} = -1}} \frac{n_{ij}!}{(n_{ij} - 1)!} \\ &= \prod_{\substack{i,j \\ m_{ij} = +1}} (n_{ij} + 1)^{-1} \prod_{\substack{i,j \\ m_{ij} = -1}} n_{ij}, \end{aligned} \tag{8.30}$$

where the last term involves only four numbers.

Example 8.25 (Maple). Reconsider the 4×4 contingency table (n_{ij}) that provides eye color versus hair color for $N = 592$ persons (Table 8.1). As we have seen, the test statistic yields $\hat{\chi}^2 = 138.29$ and must be rejected at the significance level of 5%. Diaconis and Efron (1985) labored long and hard to determine the proportion of tables with the same row and column sums as the given table having test statistic ≤ 138.29. Their best estimate was "about 10%". The following Metropolis run illustrates that the estimate "about 20%" is more realistic. The corresponding `Maple` code is as follows (Fig. 8.5).

```
> restart:
> with(Statistics): infolevel[Statistics]:=0:
> with(RandomTools):
> X := Matrix([[68, 119, 26, 7],
```

```
               [20, 84, 17, 94],
               [15, 54, 14, 10],
               [5, 29, 14, 16]]):
> Z := Matrix([[0, 0, 0, 0],
               [0, 0, 0, 0],
               [0, 0, 0, 0],
               [0, 0, 0, 0]]):
> roll4 := rand(1..4): roll2 := rand(0..1):
> count := 0:
> L := []:
> st0 := ChiSquareIndependenceTest( X, level=0.05, output='statistic');
138.2898416
> L := [op(L), round(st0)]:
> while (nops(L)<20000) do
    i := roll4():
    j := roll4():
    if j<i then h:=j; j:=i; i:=h end if;
    u := roll4():
    v := roll4():
    if v<u then h:=v; v:=u; u:=h end if;
    r := roll2():
    if i<>j and u<>v then
      Y := Matrix([[0, 0, 0, 0],
                   [0, 0, 0, 0],
                   [0, 0, 0, 0],
                   [0, 0, 0, 0]]);
      Y[i,u] := 2*r-1;
      Y[j,v] := 2*r-1;
      Y[i,v] := -(2*r-1);
      Y[j,u] := -(2*r-1);
      Z := X + Y;
      if (Z[i,u] > 0 and Z[j,v] > 0 and Z[i,v] > 0 and Z[j,u] > 0) then
        st1 := ChiSquareIndependenceTest( Z, level=0.1, output='statistic' );
        L := [op(L),round(st1)];
        if (st1 < st0) then
          X := Z;
          count := count + 1;
        else
          rn := Generate(rational(range=0..1));
          if roll2 = 1 then mv := X[i,v]*X[j,u] / ((X[i,u]+1)*(X[j,v]+1))
            else mv := X[i,u]*X[j,v] / ((X[i,v]+1)*(X[j,u]+1))
          end if;
          if rn < mv then
            X := Z
          end if;
```

```
          end if;
        end if;
      end if;
    end:
> count;
4321
> nops(L);
20000
> evalf(count/nops(L));
0.2160500000
> Histogram(L,discrete=true);
```

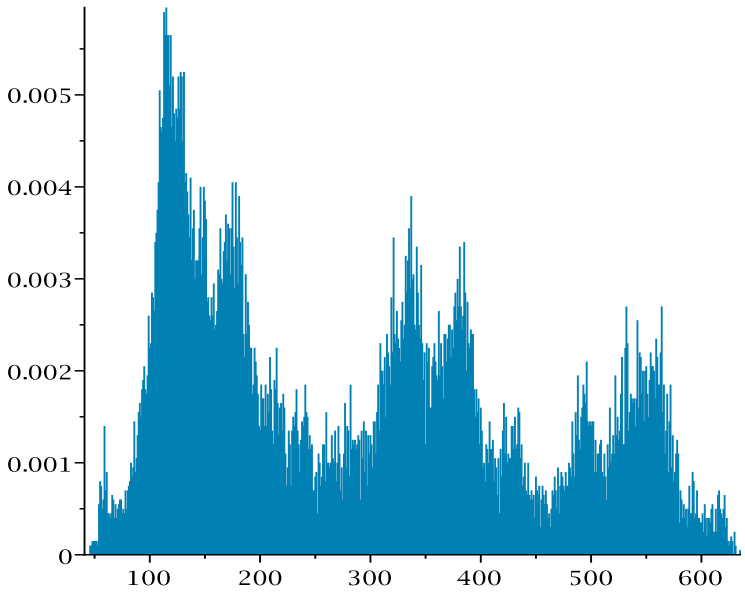


Fig. 8.5. Histogram of Metropolis run (20.000 iterations).

♢

8.5 Hardy-Weinberg Model

The Hardy-Weinberg law is a milestone of population genetics. Suppose a population of diploid organisms mates randomly and that there is no selection or mutation affecting the gene frequencies. Under these conditions, the *Hardy-Weinberg law* states that the frequency of allele combinations (genotypes) will remain constant across generations and gives a formula for these frequencies. These frequencies

apply to infinite populations, while for finite populations the question of interest is whether or not the finite population is a random subset of a population that follows the Hardy-Weinberg law. Testing whether a finite population obeys the proportions of the Hardy-Weinberg law is an important first step towards the analysis of a population.

Consider the Hardy-Weinberg model of a two-allele locus with alleles `Y` and `y`. Suppose in a population, the alleles `Y` and `y` occur with probabilities p and q, respectively. Then we have

$$p + q = 1. \tag{8.31}$$

The proportions p and q cannot be directly measured. However, we can look at the organisms in a population. For this, the Hardy-Weinberg law states that in the offspring, the genotypes `Yy` and `yY` (heterozygote) occur together with probability $2pq$, and the genotypes `YY` and `yy` (homozygote) occur with probabilities p^2 and q^2, respectively. A population with these genotype frequencies is said to be in *Hardy-Weinberg equilibrium* at the locus and the genotype frequencies are known as *Hardy-Weinberg proportions*. These genotype frequencies satisfy (Fig. 8.6).

$$1 = (p + q)^2 = p^2 + 2pq + q^2. \tag{8.32}$$

In population genetices one is particularly interested in the prevalence of a particular allele. The Hardy-Weinberg law also explains why recessive phenotype persist over time. We assume that p^2 is the probability of homozygous dominant genotype, $2pq$ is the probability of heterozygous genotype, and q^2 is the probability of homozygous recessive genotype.

	`Y`(p)	`y`(q)
`Y`(p)	`YY`(p^2)	`Yy`(pq)
`y`(q)	`yY`(qp)	`yy`(q^2)

Fig. 8.6. Hardy-Weinberg Punnett square.

Example 8.26. Consider a population of mice in Hardy-Weinberg equilibirium.

- First, suppose there are 16% of the mice in the population that are homozygous recessive. How many mice in the population are homozygous dominant? We have $q^2 = 0.16$. Then $q = 0.40$ and so $p = 1 - q = 0.60$. Thus $p^2 = 0.36$ and hence 36% of the mice are homozygous dominant.
- Second, suppose 19% of the mice in the population show the dominant phenotype. What is the dominant allele frequency? We have $p^2 + 2pq = 0.19$. Then $q^2 = 1 - (p^2 + 2pq) = 0.81$ and so $q = 0.90$. Thus $p = 1 - q = 0.10$ and hence 10% of the mice show the dominant allele frequency. ♢

Example 8.27. Suppose in the population of 300 million US Americans there are 30,000 suffering from cystic fibrosis. How many Americans are carriers? We have $q^2 = 30,000/300,000,000 = 0.0001$. Then $q = 0.01$ and so $p = 1 - q = 0.99$. Thus $2pq = 0.0198$ and hence under 2% of the US Americans are carriers; that is, $3 \cdot 10^8 \cdot 0.0198 = 5,940,000$ US Americans are carriers. ♢

More generally, let $m \geq 2$ be an integer and consider the Hardy-Weinberg model for an m-allele locus with alleles $A_1, \ldots, A_m$. Suppose the allele A_i occurs with probability p_i, $1 \leq i \leq m$, and the probabilities satisfy

$$p_1 + \ldots + p_m = 1. \tag{8.33}$$

The Hardy-Weinberg law states that in the offspring, the genotype (heterozygote) A_iA_j, $i < j$, occurs with probability $2p_ip_j$ and the genotype (homozygote) A_iA_i has probability p_i^2. It is assumed that the resulting $\binom{m+1}{2}$ genotypes are phenotypically distinguishable. A population with these genotype frequencies is said to be in Hardy-Weinberg equilibrium at the locus and the genotype frequencies are the Hardy-Weinberg proportions. These genotype frequencies fulfill

$$1 = (p_1 + \ldots + p_m)^2 = \sum_{i=1}^{m} p_i^2 + \sum_{i<j} 2p_ip_j. \tag{8.34}$$

Let $R = \{(i,j) \mid 1 \leq i \leq j \leq m\}$ and let p_{ij} be the probability that the genotype A_iA_j is observed, $1 \leq i \leq j \leq m$. Suppose that n genotypes are observed with the frequencies $u = (u_{ij})$, where u_{ij} describes the number of outcomes of the genotype A_iA_j, $(i,j) \in R$. Then we have

$$n = \sum_{(i,j)\in R} u_{ij}. \tag{8.35}$$

Since the number of observations n is fixed, the common density is given by a multinomial distribution

$$f(u|n) = \frac{n!}{\prod_{(i,j)\in R} u_{ij}!} p_{ij}^{u_{ij}}. \tag{8.36}$$

Testing the deviation from the Hardy-Weinberg proportions can be considered as a *hypothesis testing* problem. For this, the *null hypothesis* states that the population is conform with the Hardy-Weinberg proportions,

$$H_0 : \; p_{ij} = \begin{cases} p_i^2 & i = j, \\ 2p_ip_j & i \neq j. \end{cases} \tag{8.37}$$

The *alternative hypothesis* H_1 assumes that it is not. The common approach for this kind of testing is a *goodness-of-fit test*.

Example 8.28 (R). Take a random sample of 10 six-sided dice that are thrown 40 times. The outcomes are given by the vector `dice` below. We test the null hypothesis that all outcomes have equal probabilities.

```
> dice <- c( 71, 69, 74, 54, 66, 67 )
# default: uniform distribution
> sum( dice )
[1] 400
> chisq.test( dice, correct=TRUE )

        Chi-squared test for given probabilities

data: dice
X-squared: 3.53, df = 5, p-value = 0.6189

> qchisq( 0.95, 5 )
[1] 11.0705
```

The test statistic yields $\hat{\chi}^2 = 3.53$ and the 95%-quantile of the chi-squared distribution with $\nu = 5$ degrees of freedom is $\chi^2_{5;0.95} = 11.0705$. Thus the null hypothesis cannot be rejected at the significance level of 5%. ♢

Example 8.29 (R). In a course of Statistics 101, there are 26 freshmen, 33 sophomores, 20 juniors, and 22 seniors. We test the null hypothesis that freshmen, sophomores, juniors, and seniors are respresented according to the probabilities 1/3, 1/3, 1/6, and 1/6, respectively.

```
> students <- c( 26, 33, 20, 22 )
> sum(students)
[1] 101
> null.probs <- c( 1/3, 1/3, 1/6, 1/6 )
> chisq.test( students, p = null.probs, correct=TRUE )

            Chi-squared test for given probabilities

data: students
X-squared: 3.9406, df = 3, p-value = 0.268

> qchisq( 0.95, 3 )
[1] 2.365974
```

The test statistic yields $\hat{\chi}^2 = 3.9406$ and the 95%-quantile of the chi-squared distribution with $\nu = 3$ degrees of freedom is $\chi^2_{3;0.95} = 2.365974$. Thus the null hypothesis must be rejected at the significance level of 5%.

On the other hand, we test the null hypothesis that freshmen, sophomores, juniors, and seniors are respresented according to the probabilities 0.3, 0.3, 0.2, and 0.2, respectively.

```
> students <- c( 26, 33, 20, 22 )
> sum(students)
[1] 101
> null.probs <- c( 0.3, 0.3, 0.2, 0.2 )
> chisq.test( students, p = null.probs, correct=TRUE )

            Chi-squared test for given probabilities

data: students
X-squared: 1.0132, df = 3, p-value = 0.7981
> qchisq( 0.95, 3 )
[1] 2.365974
```

Here the null hypothesis cannot be rejected at the significance level of 5%. ♢

In view of the Hardy-Weinberg model, the number of alleles A_i in the sample is

$$u_{i+} = u_{1i} + \ldots + 2u_{ii} + \ldots + u_{im}, \quad 1 \le i \le m. \tag{8.38}$$

Under the null hypothesis, the theoretical frequencies of the outcomes are

$$\hat{p}_i = \frac{u_{i+}}{2n}, \quad 1 \le i \le m, \tag{8.39}$$

since each genotype is counted twice. Thus the expected number of outcomes of the genotype A_iA_j can be estimated for heterozygotes as

$$\hat{u}_{ij} = 2n\hat{p}_i\hat{p}_j, \quad 1 \le i < j \le m, \tag{8.40}$$

and for homozygotes as

$$\hat{u}_{ii} = n\hat{p}_i^2, \quad 1 \le i \le m. \tag{8.41}$$

In view of the goodness-of-fit test, the test statistic measures the statistical difference between the observed frequencies and the expected frequencies under the null hypothesis.

$$\chi^2_\nu = \sum_{(i,j)\in R} \frac{(u_{ij} - \hat{u}_{ij})^2}{\hat{u}_{ij}}. \tag{8.42}$$

This test statistic is approximately chi-squared distributed with ν degrees of freedom. The number of degrees of freedom is $\nu = \binom{m}{2}$, since the homozygotic frequences $\hat{u}_{ii}$ can be obtained from $\hat{u}_{i+}$ and the heterozygotic frequencies $\hat{u}_{ij}$.

Let $0 < \alpha < 1$. A level-α chi-squared test rejects the null hypothesis if the calculated value $\hat{\chi}^2$ of the test statistic in (8.42) satisfies $\hat{\chi}^2 > \chi^2_{\nu;1-\alpha}$, where $\chi^2_{\nu,1-\alpha}$ is the $(1-\alpha)$-quantile of the chi-squared distribution with ν degrees of freedom.

Example 8.30. Consider phenotype data from Scarlet tiger moths. After collapsing the data of $n = 1612$ moths into two groups of alleles, we observe $u_{11} = 1469$ (white-spotted), $u_{12} = 138$ (intermediate), and $u_{22} = 5$ (little spotting). The estimates of the allele probabilities are

$$\hat{p}_1 = \frac{2 \cdot 1469 + 138}{2 \cdot 1612} = 0.954 \quad \text{and} \quad \hat{p}_2 = \frac{2 \cdot 5 + 138}{2 \cdot 1612} = 0.046.$$

Then we obtain

$$\begin{aligned} \hat{u}_{11} &= n\hat{p}_1^2 = 1467.397, \\ \hat{u}_{12} &= 2n\hat{p}_1\hat{p}_2 = 141.206, \\ \hat{u}_{22} &= n\hat{p}_2^2 = 3.397. \end{aligned}$$

The test statistics gives

$$\hat{\chi}^2 = \frac{(u_{11} - \hat{u}_{11})^2}{\hat{u}_{11}} + \frac{(u_{12} - \hat{u}_{12})^2}{\hat{u}_{12}} + \frac{(u_{22} - \hat{u}_{22})^2}{\hat{u}_{22}} = 0.831.$$

At the significance level of $\alpha = 0.05$, the 0.95-quantile of the chi-squared distribution is $\chi^2_{1;0.95} = 3.84$. Thus the null hypothesis cannot be rejected at the 5% significance level for one degree of freedom. ♢

The Hardy-Weinberg model amounts to a toric statistical model that can be described by the $m \times \binom{m+1}{2}$ matrix

$$A = \begin{pmatrix} A_m & A_{m-1} & \dots & A_1 \end{pmatrix} \tag{8.43}$$

where the k-th block matrix is the $m \times k$ matrix

$$A_k = \left(\begin{array}{c|ccc} 0 & & & \\ \vdots & & O_{m-k,k-1} & \\ 0 & & & \\ \hline 2 & 1 & \dots & 1 \\ \hline 0 & & & \\ \vdots & & I_{k-1,k-1} & \\ 0 & & & \end{array}\right), \quad 1 \le k \le m. \tag{8.44}$$

In view of the observed frequencies

$$u = (u_{11}, u_{12}, \dots, u_{1m}, u_{22}, u_{23}, \dots, u_{2m}, u_{33}, \dots, u_{mm})^t \tag{8.45}$$

and the marginal frequencies

$$u_+ = (u_{1+}, \dots, u_{m+})^t, \tag{8.46}$$

we obtain

$$Au = u_+. \tag{8.47}$$

Thus the marginal frequencies u_+ form a sufficient statistic of the model.

Proposition 8.31. *A reduced Groebner basis of the ideal I_A in the polynomial ring $\mathbb{Q}[\{X_{i,j}\}_{(i,j)\in R}]$ is given as*

$$G_R = \{X_{i_1,i_2}X_{j_1,j_2} - X_{k_1,k_3}X_{k_2,k_4} \mid (k_1,k_2,k_3,k_4) = \mathrm{sort}(i_1,j_1,i_2,j_2)\},$$

where sort *denotes the sorting of quadruples over the alphabet $\{1,\dots,m\}$.*

Note that the Groebner basis of the ideal I_A consists of three types of binomials:

- $X_{i_1,i_2}X_{i_3,i_4} - X_{i_1,i_3}X_{i_2,i_4}$, where $\{i_1,\dots,i_4\}$ is a 4-element subset of [4].
- $X_{i_1,i_1}X_{i_2,i_3} - X_{i_1,i_2}X_{i_1,i_3}$, where $\{i_1,\dots,i_3\}$ is a 3-element subset of [4].
- $X_{i_1,i_1}X_{i_2,i_2} - X_{i_1,i_2}^2$, where $\{i_1,i_2\}$ is a 2-element subset of [4].

By Prop. 8.7, we obtain the following result.

Proposition 8.32. *The minimal Markov basis for the matrix A corresponding to the Groebner basis G_R is given by*

$$B_R = \{\pm(\boldsymbol{e}_{i_1,i_2} + \boldsymbol{e}_{j_1,j_2} - \boldsymbol{e}_{k_1,k_3} - \boldsymbol{e}_{k_2,k_4}) \mid (k_1,k_2,k_3,k_4) = \mathrm{sort}(i_1,j_1,i_2,j_2)\}.$$

Example 8.33 (Singular). In view of the Hardy-Weinberg model for four alleles, the toric model is given by the matrix

$$A = \left(\begin{array}{cccc|ccc|ccc|c} 2&1&1&1&0&0&0&0&0&0&0 \\ 0&1&0&0&2&1&1&0&0&0&0 \\ 0&0&1&0&0&1&0&2&1&0&0 \\ 0&0&0&1&0&0&1&0&1&2&0 \end{array}\right)$$

and we have

$$A \begin{pmatrix} u_{11} \\ u_{12} \\ u_{13} \\ u_{14} \\ u_{22} \\ u_{23} \\ u_{24} \\ u_{33} \\ u_{34} \\ u_{44} \end{pmatrix} = \begin{pmatrix} u_{1+} \\ u_{2+} \\ u_{3+} \\ u_{4+} \end{pmatrix}.$$

In view of the computation of a minimal Markov basis for the matrix A, consider the ideal

$$J = \langle X_{11} - Y_1^2, X_{12} - Y_1Y_2, X_{13} - Y_1Y_3, X_{14} - Y_1Y_4, X_{22} - Y_2^2, X_{23} - Y_2Y_3, X_{24} - Y_2Y_4, \\ X_{33} - Y_3^2, X_{34} - Y_3Y_4, X_{44} - Y_4^2 \rangle.$$

A reduced Groebner basis of the ideal I_A consists of the following elements,

$$X_{12}X_{34} - X_{13}X_{24},\ X_{12}X_{34} - X_{14}X_{23},$$

$$X_{11}X_{23} - X_{12}X_{13},\ X_{11}X_{24} - X_{12}X_{14},\ X_{11}X_{34} - X_{13}X_{14}, \\ X_{22}X_{13} - X_{12}X_{23},\ X_{22}X_{14} - X_{12}X_{24},\ X_{22}X_{34} - X_{23}X_{24}, \\ X_{33}X_{12} - X_{13}X_{23},\ X_{33}X_{14} - X_{13}X_{34},\ X_{33}X_{24} - X_{23}X_{34}, \\ X_{44}X_{12} - X_{14}X_{24},\ X_{44}X_{13} - X_{14}X_{34},\ X_{44}X_{23} - X_{24}X_{34},$$

$$X_{11}X_{22} - X_{12}^2, \quad X_{11}X_{33} - X_{13}^2, \quad X_{11}X_{44} - X_{14}^2, \\ X_{22}X_{33} - X_{23}^2, \quad X_{22}X_{44} - X_{24}^2, \quad X_{33}X_{44} - X_{34}^2.$$

This can be seen from the following computation.

```
> ring r = 0, (x11,x12,x13,x14,x22,x23,x24,x33,x34,x44,y(1..4)), dp;
> ideal i = x11-y(1)^2, x12-y(1)*y(2), x13-y(1)*y(3), x14-y(1)*y(4),
            x22-y(2)^2, x23-y(2)*y(3), x24-y(2)*y(4),
            x33-y(3)^2, x34-y(3)*y(4),
            x44-y(4)^2;
> ideal j = std(i);
> eliminate( j, y(1)*y(2)*y(3)*y(4) );
_[1]=x34^2-x33*x44
_[2]=x24*x34-x23*x44
_[3]=x14*x34-x13*x44
_[4]=x24*x33-x23*x34
_[5]=x14*x33-x13*x34
_[6]=x24^2-x22*x44
_[7]=x23*x24-x22*x34
_[8]=x14*x24-x12*x44
_[9]=x13*x24-x12*x34
```

```
_[10]=x23^2-x22*x33
_[11]=x14*x23-x12*x34
_[12]=x13*x23-x12*x33
_[13]=x14*x22-x12*x24
_[14]=x13*x22-x12*x24
_[15]=x14^2-x11*x44
_[16]=x13*x14-x11*x34
_[17]=x12*x14-x11*x24
_[18]=x13^2-x11*x33
_[19]=x12*x13-x11*x23
_[20]=x12^2-x11*x22
```

♢

For the Hardy-Weinberg model with low genotype count the asymptotic assumption of the chi-square distribution may not hold and the chi-squared goodness-of-fit tests may fail. We can obtain approximations of the test statistic by the Metropolis algorithm, drawing distributions in the fibre $A^{-1}[u_+]$ according to the hypergeometric distribution

$$P(u|u_+) = \frac{\binom{n}{\{u_{ij}\}_{(i,j)\in R}} \cdot 2^{\sum_{i<j} u_{ij}}}{\binom{2n}{u_{1+},\ldots,u_{m+}}}. \tag{8.48}$$

To see this, note that

$$P(u|n) = \binom{n}{\{u_{ij}\}_{(i,j)\in R}} \prod_{(i,j)\in R} p_{ij}^{u_{ij}}. \tag{8.49}$$

But under the null hypothesis, we have

$$\prod_{(i,j)\in R} p_{ij}^{u_{ij}} = 2^{\sum_{i<j} u_{ij}} \cdot \prod_{i=1}^{m} p_i^{u_{i+}}. \tag{8.50}$$

Moreover, we have

$$P(u_+|n) = \binom{2n}{u_{1+},\ldots,u_{m+}} \prod_{i=1}^{m} p_i^{u_{i+}} \tag{8.51}$$

Now the result follows.

8.6 Logistic Regression

In statistics, logistic regression is a regression model in which the dependent variable is categorial. We consider the case of binary dependent variables. The binary logistic model is used to estimate the probability of a binary response, like win/lose, pass/fail or alive/dead, based on one or more independent variables.

Suppose there is a binary indicator given by a random variable Y with state set $\{0,1\}$ and a set of observable covariates z that lie in a finite subset Z of $\mathbb{Z}^d$. A *logistic model* is specified by a log-linear relation of the form

$$P(Y=1|z) = \frac{e^{z\cdot\theta}}{1+e^{z\cdot\theta}} \quad \text{and} \quad P(Y=0|z) = \frac{1}{1+e^{z\cdot\theta}}, \tag{8.52}$$

where the parameter vector $\theta \in \mathbb{R}^d$ is to be estimated. For instance, if $z = (1,i)^t \in \mathbb{Z}^2$, the model becomes

$$P(Y=1|z) = \frac{e^{\theta_1+i\theta_2}}{1+e^{\theta_1+i\theta_2}} \quad \text{and} \quad P(Y=0|z) = \frac{1}{1+e^{\theta_1+i\theta_2}}.$$

This might be appropriate if the probability depends on a distance, dose or educational level occurring in equally spaced intervals.

Consider a data set of N pairs $(y_1,z_1),\ldots,(y_N,z_N)$, where $y_1,\ldots,y_N \in \{0,1\}$ and $Z = \{z_1,\ldots,z_N\} \subseteq \mathbb{Z}^d$. For each covariate $z \in Z$, let $W(z)$ denote the number of samples with $z_i = z$ and let $W_k(z)$ be the number of samples with $z_i = z$ and $y_i = k$, where $k \in \{0,1\}$. Then we have

$$W(z) = W_0(z) + W_1(z). \tag{8.53}$$

The probability of the sample set is given by the likelihood function

$$\begin{aligned} L(y_i|z_i) &= \prod_{i=1}^{N} \frac{e^{y_i(z_i\cdot\theta)}}{1+e^{z_i\cdot\theta}} \\ &= \prod_z \frac{1}{(1+e^{z\cdot\theta})^{W(z)}} \cdot e^{\theta\cdot\sum_z zW_1(z)}. \end{aligned} \tag{8.54}$$

This formula shows that a sufficient statistic is given by

$$(W(z))_z \quad \text{and} \quad \sum_z zW_1(z). \tag{8.55}$$

If $Z = \{z_1,\ldots,z_n\}$, the sample data can be summarized by a $2\times n$ matrix

$$\begin{pmatrix} W_0(z_1)\ W_0(z_2)\ \ldots\ W_0(z_n) \\ W_1(z_1)\ W_1(z_2)\ \ldots\ W_1(z_n) \end{pmatrix}. \tag{8.56}$$

Example 8.34 (R). Consider data from the US social science survey on men's response to the statement "Women should run their homes and leave men to run the country" in 1974 (Fig. 8.7). Let $Y = 1$ if the respondent "approves" and $Y = 0$ otherwise. For each respondent, the number i of years in school is reported, $0 \le i \le 12$. The proportion W_1/W seems to decrease with years of education.

It is natural to consider a logistic model of the form

$$P(Y=1 \mid i) = \frac{e^{\theta_1+i\theta_2}}{1+e^{\theta_1+i\theta_2}}, \tag{8.57}$$

where $Z = \{(1,0),(1,1),\ldots,(1,12)\}$ is a subset of $\mathbb{Z}^2$. Here the likelihood function is given as

i	0	1	2	3	4	5	6	7	8	9	10	11	12
$W_1(i)$	4	2	4	6	5	13	25	27	75	29	32	36	115
$W(i)$	6	2	4	9	10	20	34	42	124	58	77	95	360

Fig. 8.7. Men's response to "Women should run their homes and leave men to run the country" (1974).

$$\prod_{i=0}^{12} \frac{1}{(1+e^{\theta_1+i\theta_2})^{W(i)}} \cdot e^{\sum_{i=0}^{12}(\theta_1+i\theta_2)W_1(i)}$$

and thus the log-likelihood function is given by

$$\sum_{i=0}^{12}(\theta_1+i\theta_2)W_1(i) - \sum_{i=0}^{12} W(i)\log(1+e^{\theta_1+i\theta_2}).$$

The optimization of the parameters θ_1 and θ_2 can be done by maximizing the log-likelihood function. This can be achieved by using the function `optim` as shown by the following `R` code (Sect. A.9):

```
> logL <- function( theta, w, w1, i ) {
+ val <- sum( w1*(theta[1]+i*theta[2]) )
       - sum( w*log(1+exp(theta[1]+i*theta[2])) )
+ -val
}
> w1 <- c( 4,2,4,6,5,13,25,27,75,29,32,36,115 )
> w <- c( 6,2,4,9,10,20,34,42,124,58,77,95,360 )
> i <- seq( 0, 12, 1 )
> optim( c(1,1), logL, w=w, w1=w1, i=i )
$par
[1] 2.0537346 -0.2304849
$value
[1] 546.239
$counts
function gradient
      71       NA
$convergence
[1] 0
$message
NULL
```

The estimated parameters are $\hat{\theta}_1 = 2.054$ and $\hat{\theta}_2 = -0.230$.

This gives estimates $\hat{p}_i$ for the probabilities $P(Y = 1 \mid i)$, i.e.,

$$\hat{p}_i = \frac{e^{\hat{\theta}_1+i\hat{\theta}_2}}{1+e^{\hat{\theta}_1+i\hat{\theta}_2}}, \quad 0 \le i \le 12.$$

A chi-squred test statistic for goodness-of-fit compares the expected counts $W(i)\hat{p}_i$ with the observed counts $W_1(i)$ taking into account the likelihood of the joint event,

$$\chi^2_\nu = \sum_{i=0}^{12} \frac{(W(i)\hat{p}_i - W_1(i))^2}{W(i)\hat{p}_i}. \tag{8.58}$$

Here is the corresponding `R` code:

```
> p_hat <- exp(theta1+i*theta2)/ exp(1/(1+exp(theta1+i*theta2)))
> chi_hat <- sum( (w*p_hat - w1)^2 / w*p_hat )
> chi_hat
[1] 1.422446
> qchisq( 0.95, 12 )
[1] 21.02607
```

Thus the null hypothesis for the goodness-of-fit test cannot be rejected at the significance level of 5% with 12 degrees of freedom. The uneven nature of the counts with some counts small give cause of worry about the classical approximation. Therefore, approximations of the test statistic by the Metropolis algorithm lead eventually to more confident results. ♢

The logistic regression model amounts to a toric statistical model that can be described by the $(d+n) \times 2n$ matrix

$$A = \begin{pmatrix} 1 & 1 & & & & \\ & & 1 & 1 & & \\ & & & & \ddots & \\ & & & & 1 & 1 \\ 0 & z_1 & 0 & z_2 & \dots & 0 \; z_n \end{pmatrix}. \tag{8.59}$$

Let $w_{y_i}(z_i)$ denote the number of samples (y_i, z_i) in the sample set, $1 \le i \le N$. By taking the sample vector $w = (w_0(z_1), w_1(z_1), \dots, w_0(z_n), w_1(z_n))^t$, we obtain

$$Aw = \begin{pmatrix} w_0(z_1) + w_1(z_1) \\ \vdots \\ w_0(z_n) + w_1(z_n) \\ \sum_z z w_1(z) \end{pmatrix} = \begin{pmatrix} w(z_1) \\ \vdots \\ w(z_n) \\ \sum_z z w_1(z) \end{pmatrix}. \tag{8.60}$$

A simpler toric model is given by the $(d+1) \times n$ matrix

$$A = \begin{pmatrix} 1 & 1 & \dots & 1 \\ z_1 & z_2 & \dots & z_n \end{pmatrix}. \tag{8.61}$$

By taking the sample vector $w = (w_1(z_1), \dots, w_1(z_n))^t$, we obtain

$$Aw = \begin{pmatrix} \sum_z w_1(z) \\ \sum_z z w_1(z) \end{pmatrix}. \tag{8.62}$$

This toric model fixes the counts

$$\sum_z w_1(z) \quad \text{and} \quad \sum_z z w_1(z). \tag{8.63}$$

These counts also provide a sufficient statistic for the model of logistic regression when the counts $W(z)$ are fixed at the beginning, since then the original sufficient statistic (8.55) can be recovered by retrieving the counts $W_0(z)$ according to the equation $W_0(z) = W(z) - W_1(z)$

Proposition 8.35. *Given the* $3 \times (n+1)$ *matrix*

$$A = \begin{pmatrix} 1\ 1 \ldots 1 \\ 1\ 1 \ldots 1 \\ 0\ 1 \ldots n \end{pmatrix}. \tag{8.64}$$

A reduced Groebner basis G *of the ideal* I_A *in the polynomial ring* $\mathbb{Q}[X_0, X_1, \ldots, X_n]$ *consists of the binomials*

$$X_i X_l - X_j X_k,$$

where $i + l = j + k$ *and* $0 \leq i \leq j \leq k \leq l \leq n$*, and further binomials of the form* $X_i^2 - X_j X_k$*, where* $1 \leq i \leq n-1$ *and* i*,* j*,* k *are pairwise distinct.*

By Prop. 8.7, we obtain the following result.

Proposition 8.36. *A minimal Markov basis* B *for the matrix* A *in Prop. 8.35 consists of the matrices*

$$\pm(\boldsymbol{e}_i + \boldsymbol{e}_l - \boldsymbol{e}_j - \boldsymbol{e}_k),$$

where $i + l = j + k$ *and* $0 \leq i \leq j \leq k \leq l \leq n$*, and further matrices of the form* $\pm(2\boldsymbol{e}_i - \boldsymbol{e}_j \boldsymbol{e}_k)$*, where* $1 \leq i \leq n-1$ *and* i*,* j*,* k *are pairwise distinct.*

Example 8.37 (Singular). The Groebner basis for the case $n = 6$ is given by the following computation.

```
> ring r = 0, (y(1..3),x(0..6)), dp;
> ideal i = x(0)-y(1)*y(2), x(1)-y(1)*y(2)*y(3), x(2)-y(1)*y(2)*y(3)^2,
            x(3)-y(1)*y(2)*y(3)^3, x(4)-y(1)*y(2)*y(3)^4, x(5)-y(1)*y(2)*y(3)^5,
            x(6)-y(1)*y(2)*y(3)^6;
> ideal j = std(i);
> eliminate( j, y(1)*y(2)*y(3) );
_[1]=x(5)^2-x(4)*x(6)
_[2]=x(4)*x(5)-x(3)*x(6)
_[3]=x(3)*x(5)-x(2)*x(6)
_[4]=x(2)*x(5)-x(1)*x(6)
_[5]=x(1)*x(5)-x(0)*x(6)
_[6]=x(4)^2-x(2)*x(6)
_[7]=x(3)*x(4)-x(1)*x(6)
_[8]=x(2)*x(4)-x(0)*x(6)
_[9]=x(1)*x(4)-x(0)*x(5)
_[10]=x(3)^2-x(0)*x(6)
_[11]=x(2)*x(3)-x(0)*x(5)
_[12]=x(1)*x(3)-x(0)*x(4)
_[13]=x(2)^2-x(0)*x(4)
_[12]=x(1)*x(2)-x(0)*x(3)
_[15]=x(1)^2-x(0)*x(2)
```

♢

A

Computational Statistics in R

Computational statistics is a fast growing area in statistical research and applications. This supplementary chapter provides a basic introduction to computational statistics using the statistical language R. It encompasses descriptive statistics, important discrete and continuous distributions, the method of moments, and maximum likelihood estimation. Further topics of computational statistics are treated in chapter 8.

A.1 Descriptive Statistics

Descriptive statistics is a field of mathematical statistics that concentrates on the description of the main features of a collection of sample data.

Univariate data analysis focusses on the description of the distribution of a single random variable, including central tendencies like mean, median, and mode, dispersion like range and quantiles, measures of spread like, variance and standard deviation, and measures of shape like skewness and kurtosis. The characteristics of the distribution of a random variable are often described in graphical or tabular format including plots, histograms, and stem-and-leaf-displays.

The *(sample) mean* of real-valued data $x_1, \ldots, x_n$ is defined as

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The display of discrete data and the calculation of the mean shows the following R code (Fig. A.1).

```
> data <- c(45,50,55,75)
> names(data) <- c("IIW","ET","TM","CS")
> data
IIW  ET  TM  CS
 45  50  55  75
> total <- sum(data); total
[1] 225
> mean( data )
```

```
[1]  56.25
> relative <- data / total; round(relative, 2)
 IIW   ET   TM   CS
0.20 0.22 0.24 0.33
> percent <- relative * 100; round(percent, 1)
 IIW    ET    TM    CS
20.0  22.2  24.4  33.3
> pie(data)
> barplot(data)
```

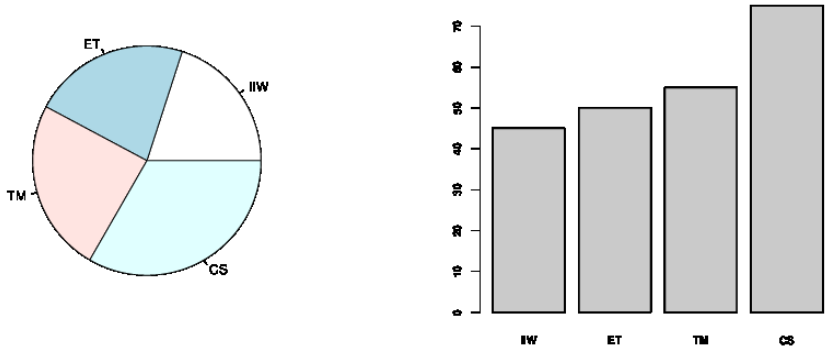


Fig. A.1. Pie plot and barplot.

The *standard deviation* of a set of real-valued data $x_1, \ldots, x_n$ is defined as

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}.$$

The square of the standard deviation s is the *variance* s^2 which can be computed as follows,

$$s^2 = \frac{1}{2n(n-1)} \sum_i \sum_j (x_i - x_j)^2.$$

The computation of mean, standard deviation, and variance of a data set shows the following R code.

```
# Body mass index (bmi) of 10 persons
> bmi <- c(25.1, 17.7, 35.5, 27.7, 28.2, 22.5, 24.3, 27.9, 21.2, 20.5)
> n <- length(bmi); n
[1] 10
> sum(bmi)
```

```
[1] 250.6
> mean(bmi)           # mean
[1] 25.06
> mean(bmi, trim=0.1) # trimmed mean by 10%
[1] 24.675
> sd(bmi)             # standard deviation
[1] 5.064956
> var(bmi)             # variance
[1] 25.65378
```

The trimmed mean is obtained from the sample mean by excluding some of the extreme values.

Quantiles are cutpoints that divide a sample set into equal sized groups. Suppose the sample data $x_1, x_2, \ldots, x_n$ is ordered such that $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. The *median* (or 2-quantile) of the data set is the middle element in the ordering if the number of data is odd. Otherwise, the median is the mean of the two middle values in the ordering. That is,

$$\tilde{x} = \begin{cases} x_{(k)} & \text{if } n \text{ is odd, } k = (n+1)/2, \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}) & \text{otherwise, } k = n/2. \end{cases}$$

More generally, for any number $0 < \alpha < 1$, the *α-quantiles* are values that partition the data set into α sublists of equal size. The 4-quantiles are the *quartiles* given by $\alpha = 1/4, 2/4, 3/4$ and the 10-quantiles are the *deciles* defined by $\alpha = k/10$ for $k = 1, \ldots, 9$. That is,

$$x_\alpha = \begin{cases} x_{(k)} & \text{if } n \cdot \alpha \text{ is not an integer, } k = \lceil n \cdot \alpha \rceil, \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}) & \text{otherwise, } k = n \cdot \alpha. \end{cases}$$

If the cumulative distribution function f of a random variable is known, the q-quantiles are given by the values of f at the images $1/q, 2/q, \ldots, (q-1)/q$. The computation of the quartiles in R illustrates the following code (Fig. A.2).

```
# Pain perception before and after therapy on a scale from 1 (low) to 10 (high)
> before <- c(3, 4, 4, 7, 9, 8, 4, 4, 7, 9)
> after <- c(2, 2, 4, 5, 6, 7, 5, 4, 2, 1)
> before; sort(before)
[1] 3 4 4 7 9 8 4 4 7 9
[1] 3 4 4 4 4 7 7 8 9 9
> after; sort(after)
[1] 2 2 4 5 6 7 5 4 2 1
[1] 1 2 2 2 4 4 5 5 6 7
> median(before); median(after)
[1] 5.5  4
> quantile(before, c(0.25,0.5,0.75))
 25%   50%   75%
4.00  5.50  7.75
> quantile(after,c(0.25,0.5,0.75))
25%   50%   75%
  2     4     5
> boxplot(before,after,names=c("before","after"))
```

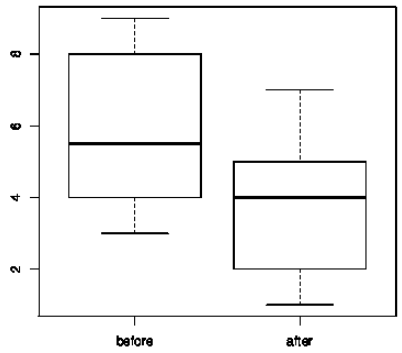


Fig. A.2. Boxplot.

A histogram is a graphical representation of the distribution of numerial data. An alternative representation is a stem-and-leaf plot. It contains two columns separated by a vertical line. The left column contains the stem and the right one the leaves. A histogram and stem-and-leaf plot are given by the following R code (Fig. A.3).

```
# Body mass index (bmi) of 10 persons
> hist(bmi, c(15,17.5,20,22.5,25,27.5,30,32.5,35,37.5,40))
> stem(bmi)

  The decimal point is 1 digit(s) to the right of the |

  1 | 8
  2 | 1134
  2 | 5888
  3 |
  3 | 6
```

Bivariate analysis involves the distribution of two random variables. The relationship between pairs of variables can be described among others by scatterplots and quantitative measures of dependence. A dot plot in R can be obtained as follows (Fig. A.4).

```
# age versus height of teenagers in African countries
> age <- c(10:19); age
[1] 10 11 12 13 14 15 16 17 18 19
> height <- c(133, 139, 147, 154, 165, 170, 175, 177, 180, 182); height
[1] 133, 139, 147, 154, 165, 170, 175, 177, 180, 182
> plot(age, height, xlim=c(10,19), ylim=c(130,190))
```

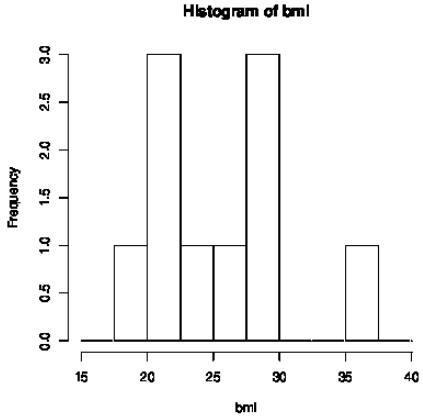


Fig. A.3. Histogram.

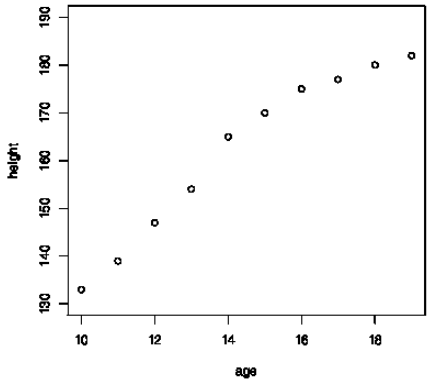


Fig. A.4. Age versus height plot.

Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be real-valued sample data from two random variables. The *covariance* is a measure of how much two random variable change together and is defined as

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}.$$

If the larger values of one variable correspond to the larger values of the other variable and the same holds for the smaller values, the covariance is positive. In the opposite case, the covariance is negative.

The *correlation coefficient* is a measure of the linear correlation between two random variables and is given by

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

The correlation coefficient gives a value between $+1$ and -1, where $+1$ is total positive correlation, 0 is no correlation, and -1 is total negative correlation.

```
# Economic growth in %
> x <- c( 2.1,2.5,4.0,3.5 )
# rate of return in %
> y <- c( 8,12,14,10 )
> cov( x, y )   # covariance
[1] 1.533333
> cor( x, y )   # correlation coefficient
[1] 0.6625739
```

Linear regression is used in statistics for modeling the relationship between an explanatory (independent) random variable and a scalar dependent variable. The following example shows a linear dependence between two data sets introduced above (Fig. A.5).

```
> lm(height ~ age)  # linear model

Coefficients:
(Intercept)        age
     79.067      5.733
```

The linear prediction function is given by $f(x) = 79.067 + 5.733 \cdot x$.

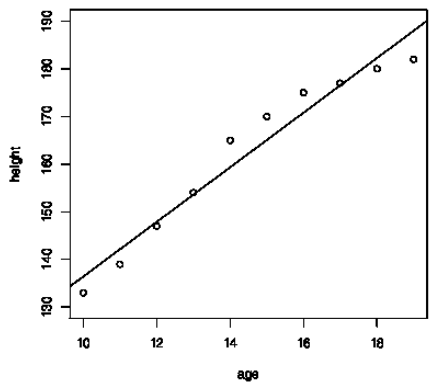


Fig. A.5. Linear regression.

A.2 Random Variables and Probability

A *random variable* is a variable whose values are subject to change due to randomness in a mathematical sense. A random variable is *discrete* if it can take on values from a finite or countable set, and a random variable is *continuous* if it can take on numerical values from an interval or a collection of intervals.

The *cumulative distribution function* (cdf) of a random variable X is F_X defined as

$$F_X(x) = P(X \leq x), \quad x \in \mathbb{R}.$$

where P denotes the probability of the argument. The subscript of F_X is omitted if it is clear from the context. The cdf of a random variable X has the following properties:

- F_X is non-decreasing.
- F_X is right-continuous; i.e.,
$$\lim_{\epsilon \to 0^+} F_X(x + \epsilon) = F_X(x), \quad x \in \mathbb{R}.$$
- F_X has the extremal values $\lim_{\epsilon \to -\infty} F_X(x) = 0$ and $\lim_{\epsilon \to \infty} F_X(x) = 1$.

A random variable X is *continuous* if the cdf F_X is a continuous function. A random variable X is *discrete* if the cdf F_X is a step function. A discrete cdf is given by a *probability mass function* (pmf) $p_X(x) = P(X = x)$. The discontinuities in the cdf are the points where the pmf is positive and $p_X(x) = F_X(x) - F_X(x^-)$.

If a random variable X is discrete, the cdf of X is given by

$$F_X(x) = P(X \leq x) = \sum_{\substack{y \leq x \\ p_X(y) > 0}} p_X(y).$$

For a continuous random variable X, the *probability density function* (pdf) of X is $f_X(x) = F'_X(x)$ for all $x \in \mathbb{R}$ if F_X is differentiable. In this case, by the fundament theorem of calculus,

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt.$$

The joint density of continuous random variables X and Y is $f_{X,Y}$ and the cdf of the pair (X, Y) is

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(s, t)dsdt.$$

The marginal probability densities of X and Y are given as

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx.$$

The formulae for the discrete random variables are analogously defined with integrals replaced by sums. In the following, f_X denotes both the pdf of X if X is continuous or the pmf of X if X is discrete.

The *mean* of a random variable X is the mathematical expectation (or expected value) of the variable and is denoted by $E[X]$. If X is continuous with pdf f_X, the mean of X is

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx.$$

If X is discrete with pmf f_X, the mean of X is

$$E[X] = \sum_{\substack{x \\ f_X(x)>0}} x f_X(x) dx.$$

We assume that $E[X]$ is finite if $E[X]$ appears in a formula. The mathematical expectation of a function $g(X)$ of a continuous random variable X with pdf f_X is

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Thus $E[X]$ is the mathematical expectation of the identity function on $\mathbb{R}$. The value $\mu_X = E[X]$ is the *first moment* of X. For any integer $r \geq 1$, the *r-th moment* of X is $E[X^r]$. Thus if X is continuous, then

$$E[X^r] = \int_{-\infty}^{\infty} x^r f_X(x) dx.$$

The *variance* of a random variable X is the second central moment given by

$$\mathrm{Var}(X) = E[(X - E[X])^2].$$

Since $E[(X - E[X])^2] = E[X^2] - (E[X])^2$, we have

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = E[X^2] - \mu_X^2.$$

The variance is also denoted by σ_X^2. The square root of the variance is the *standard deviation* and the reciprocal of the variance is the *precision*.

The mathematical expectation of the product of two continuous random variables X and Y with joint density $f_{X,Y}$ is

$$E[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x,y) dx dy.$$

The *covariance* of X and Y is

$$\mathrm{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - E[X]E[Y] = E[XY] - \mu_X \mu_Y.$$

The covariance of X and Y is also denoted by $\sigma_{X,Y}$. In particular, we have $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

The *correlation* between two continuous random variables X and Y is

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}.$$

Two random variables X and Y are *uncorrelated* if $\rho(X,Y) = 0$.

Two random variables X and Y are *independent* if

$$f_{X,Y}(x,y) = f_X(x) f_Y(y), \quad x, y \in \mathbb{R},$$

or equivalently,

$$F_{X,Y}(x,y) = F_X(x)F_Y(y), \quad x,y \in \mathbb{R}.$$

More generally, the random variables $X_1, \ldots, X_n$ are *independent* if the joint pdf f of $X_1, \ldots, X_n$ equals the product of the marginal densities; i.e.,

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i), \quad (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

If two random variables X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$ and thus $\rho(X,Y) = 0$. The converse is generally false. However, if X and Y are normally distributed random variables with $\mathrm{Cov}(X,Y) = 0$, then X and Y are independent.

The random variables $X_1, \ldots, X_n$ are a *random sample* from a distribution F_X if $X_1, \ldots, X_n$ are independent and identically distributed with distribution F_X. Thus the joint pdf of $X_1, \ldots, X_n$ is

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i), \quad (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Proposition A.1. *Let X and Y be random variables and $a, b \in \mathbb{R}$.*

1. $E[aX + bY] = aE[X] + bE[Y]$.
2. $E[aX + b] = aE[X] + b$.
3. *If X and Y are independent,* $E[XY] = E[X]E[Y]$.
4. $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.
5. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$.
6. *If X and Y are independent,* $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Corollary A.2. *If $X_1, \ldots, X_n$ are independent and identically distributed random variables, then*

$$E[X_1 + \ldots + X_n] = n\mu_X \quad \textit{and} \quad \mathrm{Var}(X_1 + \ldots + X_n) = n\sigma_X^2.$$

It follows that the *sample mean*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

has the expected value μ_X and the variance σ_X^2/n.

The conditional probability of an event A given the event B has taken place is

$$P(A \mid B) = \frac{P(AB)}{P(B)},$$

where $AB = A \cap B$ is the intersection of the events A and B. Two events A and B are *independent* if $P(AB) = P(A)P(B)$. In this case, $P(A \mid B) = P(A)$.

Let X and Y be random variables with joint density $f_{X,Y}$. Then the *conditional density* of X given $Y = y$, $y \in \mathbb{R}$, is

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \quad x \in \mathbb{R}.$$

In the same way, the conditional density of Y given $X = x$, $x \in \mathbb{R}$, is

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \quad y \in \mathbb{R}.$$

Thus the joint density $f_{X,Y}$ has the form

$$f_{X,Y}(x,y) = f_{X|Y=y}(x)f_Y(y) = f_{X|X=x}(y)f_X(x), \quad x, y \in \mathbb{R}.$$

The conditional expected value of X given $Y = y$, $y \in \mathbb{R}$, is

$$E[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y=y}(x)dx,$$

where $f_{X|Y=y}$ is assumed to be continuous.

Proposition A.3. *Let X and Y be random variables.*

1. *Conditional expectation rule:* $E[X] = E[E[X \mid Y]]$.
2. *Conditional variance formula:* $\mathrm{Var}(X) = E[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(E[X \mid Y])$.

A.3 Some Discrete Distributions

The most important discrete distributions are counting distributions which are used to model the frequencies or waiting times of events.

In the statistical language R, the probability mass functions (pmf) or densities (pdf), cumulative distribution functions (cdf), quantile functions, and random number generators of many commonly used probability distributions are made available. The first letter always denotes the function type:

d	density function
p	cumulative distribution function
q	quantile function
r	random number generator

In a *uniform distribution*, there is finite number of values that are equally likely to be observed. A uniformly distributed random variable X with state set $[m]$ has the pmf

$$P(X = x) = \frac{1}{m}, \quad x \in [m].$$

The cumulative distribution function is a step function

$$F_X(x) = \frac{x}{m}, \quad x \in [m].$$

The mean and variance of X are respectively given by

$$E[X] = \frac{m+1}{2} \quad \text{and} \quad \mathrm{Var}(X) = \frac{m^2 - 1}{12}.$$

The uniform distribution in R can be analyzed by the functions ddiscrete, pdiscrete, qdiscrete, and rdiscrete. It requires the loading of the library e1071.

Several important discrete distributions can be formulated in terms of Bernoulli trails. A *Bernoulli experiment* has two possible outcomes, success (1) or failure (0). A Bernoulli random variable X has the pmf

$$P(X = 1) = p \quad \text{and} \quad P(X = 0) = 1 - p,$$

where p is the probability of success. The mean and variance of X are respectively given by

$$E[X] = p \quad \text{and} \quad \text{Var}(X) = p(1 - p).$$

Let X be a random variable that counts the number of successes in n independent, identically distributed Bernoulli trials with success probability p. Then X has the *binomial distribution* with parameters n and p if

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad 0 \le x \le n.$$

Since the binomial variable X is the sum of n independent, identically distributed Bernoulli variables, we have

$$E[X] = np \quad \text{and} \quad \text{Var}(X) = np(1 - p).$$

The following R code provides three binomial distributions (Fig. A.6).

```
> f1 = dbinom ( 0:6, 6, 0.25 ) # pmf
> f2 = dbinom ( 0:6, 6, 0.50 )
> f3 = dbinom ( 0:6, 6, 0.75 )
> matplot( 0:6, cbind(f1,f2,f3), type="l", pch=c("red","black","green") )
> F1 = pbinom ( 0:6, 6, 0.25 ) # cdf
> F2 = pbinom ( 0:6, 6, 0.50 )
> F3 = pbinom ( 0:6, 6, 0.75 )
> matplot( 0:6, cbind(F1,F2,F3), type="l", pch=c("red","black","green") )
```

Example A.4 (R). Consider the treatment of inpatients with a specific drug. Suppose the probability of successful treatment is $p = 0.75$. Then the probability that sixteen out of twenty inpatients are successfully treated is

$$P = \binom{20}{8} 0.75^{16} 0.25^4 = 0.1896855.$$

This can be computed using R as follows,

```
> p <- 0.75
> choose(20,16) * p^16 * (1-p)^4
[1] 0.1896855
> dbinom( 16, 20, p )
[1] 0.1896855
```

The following code provides the probability of successful treatment of $n = 20$ inpatients (Fig. A.7).

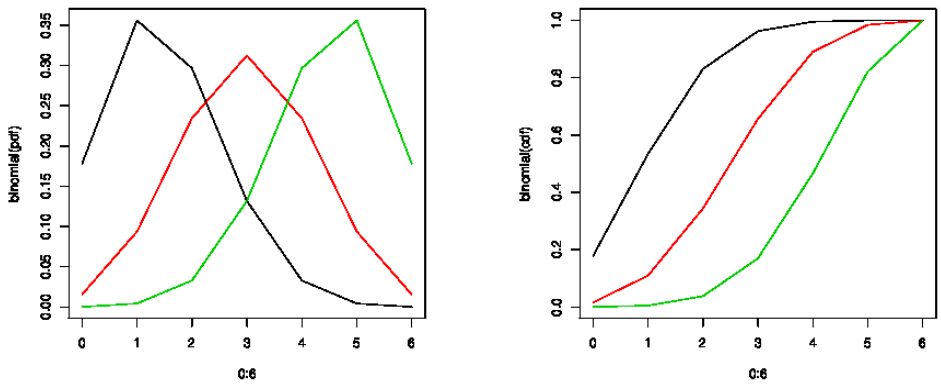


Fig. A.6. Binomial distribution.

```
> n <- 20
> p <- 0.75
> P <- rep( NA, n )
> for (i in 1:n) P[i] <- dbinom( i, n, p )
> plot( 1:n, P, ylab="binom(pmf)" )
> Q <- rep( NA, n )
> for (i in 1:n) Q[i] <- pbinom( i, n, p )
> plot( 1:n, P, type="l",  )
```

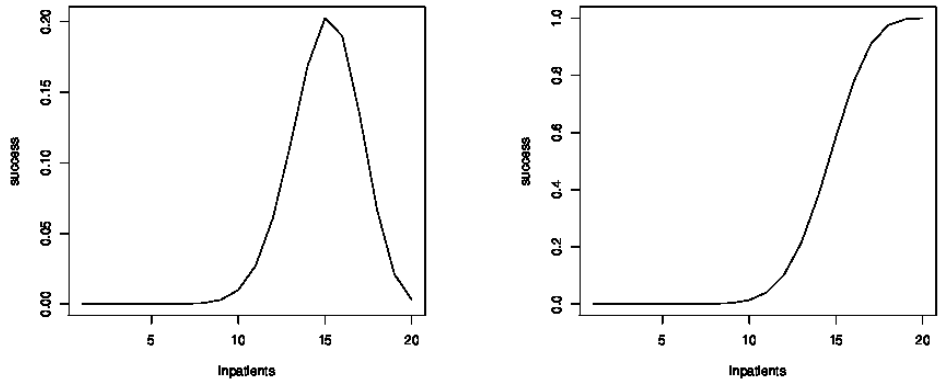


Fig. A.7. Successful treatment of twenty inpatients: pdf und cdf.

♢

The *multinomial distribution* generalizes the binomial distribution. To see this, consider k mutually exclusive and exhaustive events $A_1, \ldots, A_k$ which happen at any trial of the experiment, and each event occurs with probability $p(A_i) = p_i$, $1 \le i \le k$. Then $p_1 + \ldots + p_k = 1$. Let X_i be a random variable which counts the number of events A_i in a sequence of n independent and identical trials. Then the random variable $X = (X_1, \ldots, X_k)$ has the multinomial distribution with joint pdf

$$f(x_1, \ldots, x_k) = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \cdots p_k^{x_k}$$

where $x_1 + \ldots + x_k = n$.

For each random variable X_i, we have $E[X_i] = np_i$ and $\mathrm{Var}(X_i) = np_i(1 - p_i)$. Moreover, the covariance and the correlation of the random variables X_i and X_j for $i \neq j$ respectively are

$$\mathrm{Cov}(X_i, X_j) = -np_ip_j \quad \text{and} \quad \rho(X_i, X_j) = -\sqrt{\frac{p_ip_j}{(1-p_i)(1-p_j)}}.$$

The following example shows the use of the multinomial distribution in R.

```
> n <- 4
> p <- c(0.4,0.3,0.3)
> m <- c(0,0,4, 0,1,3, 0,2,2, 0,3,1, 0,4,0,
+        1,0,3, 1,1,2, 1,2,1, 1,3,0, 2,0,2,
+        2,1,1, 2,2,0, 3,0,1, 3,1,0, 4,0,0)
> M <- matrix( m, nrow=3 )
> ncol(M)
[1]  15
> M
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15]
[1,]    0    0    0    0    0    1    1    1    1     1     2     2     3     3     4
[2,]    0    1    2    3    4    0    1    2    3     4     1     2     0     1     0
[3,]    4    3    2    1    0    3    2    1    1     0     1     0     1     0     0
> P <- rep( NA, ncol(M) )
> for (i in 1:ncol(M)) P[i] <- dmultinom( M[,i], prob=p ) # pmf
> plot( 1:ncol(M), P, ylab="multinom(pmf)" )
```

Example A.5 (R). Consider a box with 100 colored beads of which 50 are red, 30 are green, and 20 are yellow. How large is the probability to draw 10 beads such that 4 are red, 3 are green, and 3 are yellow? The probabilities to draw one red, one green, and one yellow bead are $p_1 = 0.5$, $p_2 = 0.3$, and $p_3 = 0.2$, respectively. Thus the probability to draw 4 red beads, 3 green beads, and 3 yellow beads is

$$P = \binom{10}{4, 3, 3} 0.5^4 0.3^3 0.2^3 = 0.0567.$$

This can be computed using R as follows,

```
> dmultinom( c(4,3,3), prob=c(0.5,0.3,0.2) )
[1] 0.0567
```

♢

The *geometric distribution* is a variant of the binomial distribution. To see this, take a sequence of Bernoulli trials with success probability p. Let X be a random variable counting the number of trials until the first success happens. Then we have

$$P(X = x) = p(1-p)^{x-1}, \quad x \geq 1.$$

A random variable X with this pmf is geometrically distributed. If X has the geometric distribution with success probability p, the cdf of X is

$$F_X(x) = P(X \leq x) = 1 - (1-p)^x, \quad x \geq 1.$$

A geometrically distributed random variable X has respectively the mean and variance

$$E[X] = \frac{1-p}{p} \quad \text{and} \quad \text{Var}(X) = \frac{1-p}{p^2}.$$

Note that a geometrically distributed random variable X satisfies the relation

$$\begin{aligned} P(X = n + x | X > n) &= \frac{P(X = n + x)}{P(X > n)} \\ &= \frac{p(1-p)^{n+x-1}}{1 - (1 - p(1-p)^n)} \\ &= p(1-p)^{k-1} = P(X = x), \quad x \geq 1. \end{aligned}$$

Thus when X is interpreted as waiting time for an event to occur relative, the above property exhibits the memorylessness of the geometric distribution.

Example A.6 (R). Consider the tossing of a six-sided fair die. The probability of the first occurrence of a "six" can be described by the geometric distribution. We have $p = 1/6$ and the probability after x trails to roll a "six" is $p(1-p)^{x-1}$. The pdf and cdf can be computed by R as follows (Fig. A.8).

```
> n <- 20
> y <- dgeom( 1:n, prob=1/6 )
> plot( 1:n, y, ylab="geometric(pmf)" )
> z <- pgeom( 1:n, prob=1/6 )
> plot( 1:n, z, ylab="geometric(df)" )
```

♢

The *negative binomial distribution* is a generalization of the geometric distribution in the sense that one is interested in the number of failures until the r-th success. Let X be a random variable that counts the number of failures until the r-th success. If $X = x$ failures occur before the r-th success, the r-th success happens by the $(x + r)$-th trial and in the first $x + r - 1$ trials there are $r - 1$ successes and x failures. This takes place in $\binom{x+r-1}{r-1} = \binom{x+r-1}{x}$ different ways and each case has probability $p^r(1-p)^x$. Thus the random variable X has the pmf

$$P(X = x) = \binom{x + r - 1}{r - 1} p^r (1-p)^x, \quad x \geq 0.$$

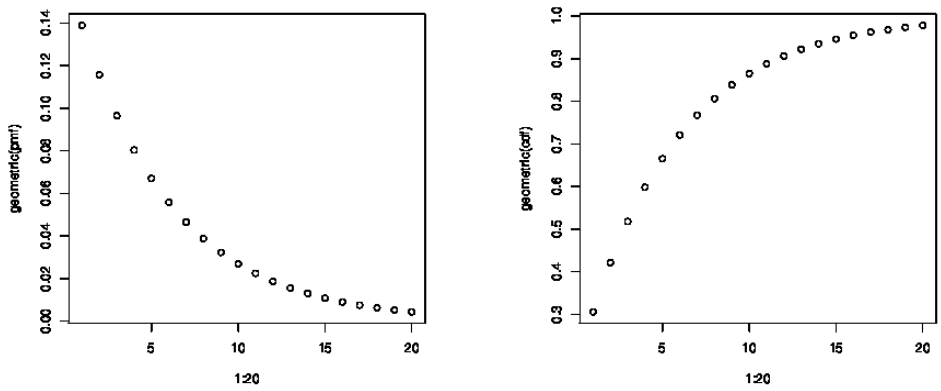


Fig. A.8. Probability of success after a number of failures: pdf and cdf.

A random variable X has the negative binomial distribution with parameters r and p if

$$P(X = x) = \frac{\Gamma(x + r)}{\Gamma(r)\Gamma(x + 1)} p^r (1 - p)^x, \quad x \geq 0,$$

where Γ denotes the complete gamma function defined as

$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t} dt, \quad r \neq 0, -1, -2, \ldots.$$

Note that $\Gamma(n) = (n - 1)!$ for each integer $n \geq 1$. It follows that both probabilities are identical when $r \geq 1$ is an integer. Let X be a random variable with a negative binomial distribution with parameters r and p. Then X is the sum of r independent, identically distributed geometric random variables with parameter p. Thus the mean and variance of X respectively are given by

$$E[X] = r\frac{1 - p}{p} \quad \text{and} \quad \text{Var}(X) = r\frac{1 - p}{p^2}.$$

Example A.7 (R). Consider the number of failures in a lottery until the third success if the probability of success is $p = 0.25$. For instance, the probability of at most n failures until the third success is given by

$$\sum_{i=0}^{n-3} \binom{i + 3 - 1}{i} 0.25^3 0.75^i.$$

The following code provides the probability distribution for at most $n = 20$ draws until the third success (Fig. A.9).

```
> n <- 20
> p <- 0.25
```

```
> P <- rep( NA, 20)
> for (i in 1:n) P[i] <- dnbinom( i, 3, p )
> plot( 1:n, P, ylab="negbinom(pmf)" )
> Q <- rep( NA, 20)
> for (i in 1:n) Q[i] <- pnbinom( i, 3, p)
> plot( 1:n, Q, ylab="negbinom(cdf)" )
```

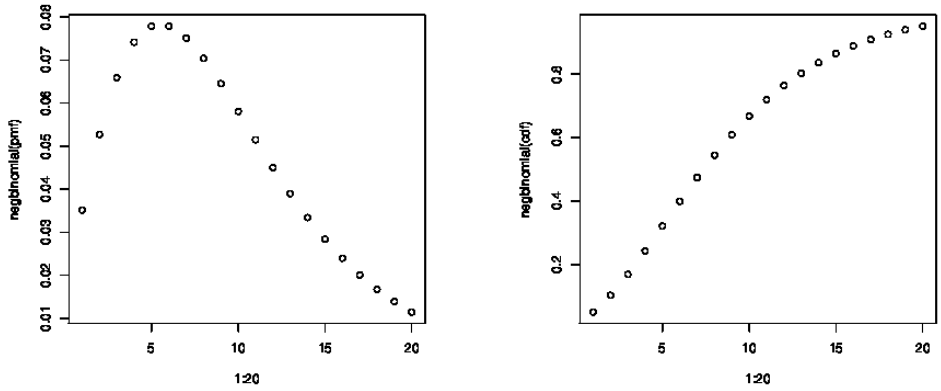


Fig. A.9. Probability of the third success after a number of failures: pdf and cdf.

♢

The *hypergeometric distribution* describes the probability of x successes in n draws without replacement from a finite population of size N which contains exactly K successes. In view of this setting, a random variable X follows the hypergeometric distribution if the pmf is given by

$$P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}.$$

The pmf is positive if $\max(0, n + K - N) \leq x \leq \min(K, n)$. The mean and variance of a hypergeometrically distributed random variable X respectively are given as

$$E[X] = n\frac{K}{N} = np \quad \text{and} \quad \text{Var}(X) = np(1-p)\frac{N-n}{N-1}.$$

Example A.8 (R). Consider a collection of 100 items which contains 5% defective items. How large is the probability of drawing 50 items of which i items are defect? The probabilities are given by a hypergeometric distribution which can be computed in R as follows (Fig. A.10).

```
# call: phyper(x,K,N-K,n)
```

```
> P  <- rep( NA, 5 )
> for (i in 1:5) P[i] <- dhyper( 50-i, 95, 5, 50 )
> plot( 1:5, P, xlab="defect items", ylab="hypergeom(pmf)" )
```

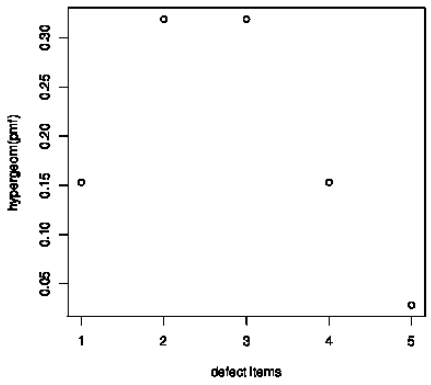


Fig. A.10. Hypergeometric distribution.

♢

The *poisson distribution* is another important discrete distribution that is applicable to systems with a large number of possible events each of which being very rare. A random variable X has the poisson distribution with parameter $\lambda > 0$ if the pmf of X is given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x \geq 0.$$

It expresses the probability of a given number of events x occurring in a fixed interval of time or space if these events happen with a known average rate and independent of time since the last event. Examples are the number of phone calls received by a call center per hour and the number of decay events per second from a radioactive source. The pmf follows the recursion $p(x+1) = p(x)\frac{\lambda}{x+1}$. If X is a random variable X with poisson distribution with parameter $\lambda > 0$, the mean and the variance of X respectively are given as

$$E[X] = \lambda \quad \text{and} \quad \text{Var}(X) = \lambda.$$

The use of the poisson distribution in R is exemplified by the following code (Fig. A.11).

```
> f1 <- dpois( 1:20, 1 ) # lambda = 1
> f2 <- dpois( 1:20, 5 )
> f3 <- dpois( 1:20, 10 )
> matplot( 1:20, cbind(f1,f2,f3), type="l", pch=c("red","green","black") )
> F1 <- ppois( 1:20, 1 ) # lambda = 1
```

```
> F2 <- ppois( 1:20, 5 )
> F3 <- ppois( 1:20, 10 )
> matplot( 1:20, cbind(F1,F2,F3), type="l", pch=c("red","green","black") )
```

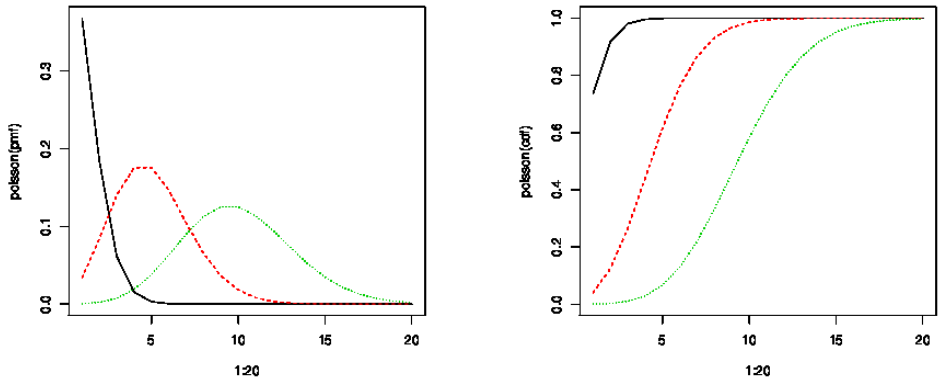


Fig. A.11. Poisson distribution.

Note that for large values of n, small values of p, and fixed value $\lambda = np$ (i.e., $n \to \infty$, $p \to 0$, and $np \to \lambda$), the binomial distribution converges to the poisson distribution with parameter λ.

Example A.9 (R). Suppose the probability that a person suffers from a drug intolerance is $p = 0.001$. How large is the probability that x persons out of $n = 2000$ persons suffer from the drug intolerance? We have $\lambda = n \cdot p = 2000 \cdot 0.001 = 2$, since $1 - p = 0.999$ is close to 1, and the probability that x persons suffer from drug intolerance is

$$P(x) = \frac{\lambda e^{-\lambda}}{x!} = \frac{2^x e^{-2}}{x!}.$$

This quantity can be computed by the following R code (Fig. A.12).

```
> l <- 2
> P <- rep( NA, 11 )
> for (x in 0:10) P[x] <- dpois( x, l )
> plot( 0:10, P, type="l", xlab="x", ylab="P(x)" )
```

♢

A.4 Some Continuous Distributions

The most basic continuous distribution is the *uniform distribution* in which all elements of an interval are equally probable. The pdf of the uniform distribution for the real-valued interval (a, b) is defined by

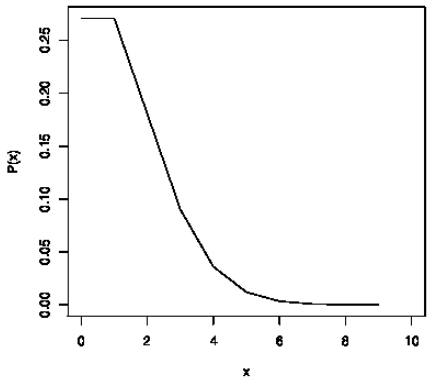


Fig. A.12. Probability of drug intolerance.

$$f(x) = \begin{cases} \frac{1}{b-a} \text{ if } a < x < b, \\ 0 \quad \text{otherwise.} \end{cases}$$

The cdf of the pdf f is given as

$$F(x) = \frac{x-a}{b-a}, \quad x \in \mathbb{R}.$$

The respective mean and variance are

$$\mu = \frac{a+b}{2} \quad \text{and} \quad \sigma^2 = \frac{(b-a)^2}{12}.$$

The *normal distribution* is one of the most important continuous distributions in statistics. They are often used to represent real-valued random variables whose distributions are unknown. The pdf of the normal distribution with mean μ and standard deviation $\sigma > 0$, denoted $N(\mu, \sigma^2)$, defined as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad x \in \mathbb{R}.$$

If $\mu = 0$ and $\sigma = 1$, the distribution is the standard normal distribution, denoted by $N(0,1)$, and is given by the standard normal pdf

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right], \quad x \in \mathbb{R}.$$

The cdf of the standard normal distribution is the integral

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt, \quad x \in \mathbb{R}.$$

Moreover, for a generic normal distribution given by pdf f with mean μ and standard deviation σ, the cdf is given by

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

The pdf of a normal distribution can be drawn as follows (Fig. A.13).

```
> mu <- 50
> sig <- 6
> low <- mu-3.5*sig; upp <- mu+3.5*sig
> x <- seq( low, upp, by=0.1 )
> f <- dnorm( x, mean=mu, sigma=sig )
> plot( x, f, type="l", xlim=c(low,upp) )
```

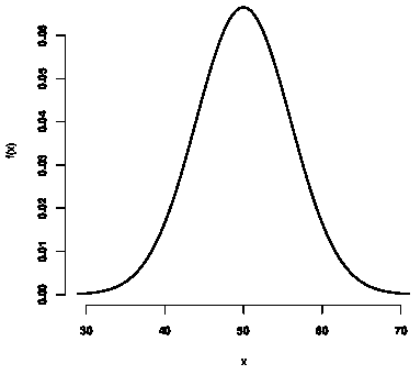


Fig. A.13. Normal distribution.

The normal distribution has several key properties. Let X be a random variable with normal distribution $N(\mu, \sigma^2)$ and let $a, b \in \mathbb{R}$. Then the linear transformation $Y = aX + b$ gives a random variable which is $N(a\mu + b, a^2\sigma^2)$. In particular, if X is $N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$ is $N(0, 1)$. Moreover, the standard normal distribution has the symmetry property

$$\Phi(-z) = P(Z \leq -z) = P(Z \geq z) = 1 - P(Z \leq z) = 1 - \Phi(z).$$

Example A.10 (R). Suppose the fasting blood sugar (mg/dl) is given by a normal random variable X with mean $\mu = 100$ and standard deviation $\sigma = 10$. How large is the probability that a randomly chosen person has fasting blood sugar (a) at most 70 mg/dl, (b) between 90 and 120 mg/dl, or (c) larger than 140 mg/dl? In view of a), we have

$$P(X \leq 70) = P(Z \leq -3) = 0.001349898,$$

In view of b), we have

$$P(X > 140) = P(Z > 4) = P(Z \leq 4) = 3.167124e - 05,$$

In view of c), we have

$$P(90 \leq X \leq 120) = P(Z \leq 2) - P(Z \leq 1) = 0.8185946.$$

The calculation in `R` is as follows.

```
> pnorm( 70, mean=100, sd=10 )
[1] 0.001349898
> pnorm( 140, mean=100, sd=10, lower.tail=FALSE )
[1] 3.167124e-05
> pnorm( 120, mean=100, sd=10 ) - pnorm( 90, mean=100, sd=10 )
[1] 0.8185946
```

♢

Let $X_1, \ldots, X_n$ be independent normal random variables, where X_i is $N(\mu_i, \sigma_i^2)$, $1 \leq i \leq n$, and let $a_1, \ldots, a_n \in \mathbb{R}$. Then the random variable given by the linear combination $Y = a_1X_1 + \ldots + a_nX_n$ has a normal distribution with mean and variance respectively given by

$$\mu = a_1\mu_1 + \ldots + a_n\mu_n \quad \text{and} \quad \sigma^2 = a_1^2\sigma_1^2 + \ldots + a_n^2\sigma_n^2.$$

In particular, if $X_1, \ldots, X_n$ are identically distributed random variables which are $N(\mu, \sigma^2)$, the random variable given by the sum $X = X_1 + \ldots + X_n$ is $N(n\mu, n\sigma^2)$. The usefulness of the normal distribution comes from the following result.

Theorem A.11 (Central Limit Theorem). *If the random variables $X_1, \ldots, X_n$ are independent and identically distributed with mean μ and variance σ^2, the limiting distribution of the random variables*

$$Z_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$$

when n becomes large is the standard normal distribution.

Continuous random variables $X_1, \ldots, X_d$ have a multivariate (or d-variate) normal distribution, abbreviated $N_d(\mu, \Sigma)$, if the joint pdf is given by

$$f(x_1, \ldots, x_d) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \left[-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right],$$

where $\Sigma = (\sigma_{ij})$ is the $d \times d$ nonsingular covariance matrix of $X_1, \ldots, X_d$, $\mu = (\mu_1, \ldots, \mu_d)^t$ is the vector of means, and $x = (x_1, \ldots, x_d)^t \in \mathbb{R}^d$. Note that the one-dimensional marginal distributions are normal with mean μ_i and variance σ_i^2, $1 \leq i \leq d$. The normal random variables $X_1, \ldots, X_d$ are independent if and only if the covariance matrix Σ is a diagonal matrix. A linear transformation of a multivariante random variable vector $X = (X_1, \ldots, X_d)$ is multivariate normal. More specifically, if A is an $l \times d$ real-valued matrix and $b = (b_1, \ldots, b_m)^t \in \mathbb{R}^l$, then $Y = AX + b$ has an l-variate normal distribution with mean vector $A\mu + b$ and covariance matrix $A\Sigma A^t$.

The *exponential distribution* is a continuous probability distribution that describes the time between events in a poisson process. This is a stochastic process in which events occur continuously and independently at a constant average rate. It is the continuous analogue of the geometric distribution and being memoryless is its key property. The pdf of a random variable X which is exponentially distributed with *rate parameter* λ is given by

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

The inverse parameter $T = 1/\lambda$ is referred to as characteristic lifetime and is called *mean time between failures*. The corresponding cdf is given as

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

A random variable X that is exponentially distributed with rate parameter λ has respectively the mean and variance

$$E[X] = \frac{1}{\lambda} \quad \text{and} \quad \text{Var}(X) = \frac{1}{\lambda^2}.$$

The use of expontially distributed random variable is shown by the following R code (Fig. A.14).

```
> x <- seq( 0, 20, by=0.1 )
> e1 <- dexp( x, rate=1 )  # lambda = 1
> e2 <- dexp( x, rate=5 )
> e3 <- dexp( x, rate=10 )
> matplot( x, cbind(f1,f2,f3), type="l", col = c("red","black","green") )
> E1 <- pexp( x, rate=1 )
> E2 <- pexp( x, rate=5 )
> E3 <- pexp( x, rate=10 )
> matplot( x, cbind(F1,F2,F3), type="l", col = c("red","black","green") )
```

Note that an exponentially distributed random variable X satisfies the relation

$$\begin{aligned} P(X > s+t | X > s) &= \frac{P(X > s+t)}{P(X > s)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\ &= e^{-\lambda t} = P(X > t), \quad s, t \geq 0. \end{aligned}$$

Thus when X is interpreted as waiting time for an event to occur relative to some initial time, the above property exhibits the memorylessness of the exponential distribution.

Example A.12. Suppose the average lifespan of a brand of light bulbs is 100 hours. How large is the probability that a randomly chosen light bulb is on light for at least 110 hours? The rate parameter is $\lambda = 0.01$ and we have

$$P(X > 110) = 1 - P(X \leq 110) = 1 - (1 - e^{-110 \cdot 0.01}) = 0.3328711.$$

The corresponding computation in R is as follows.

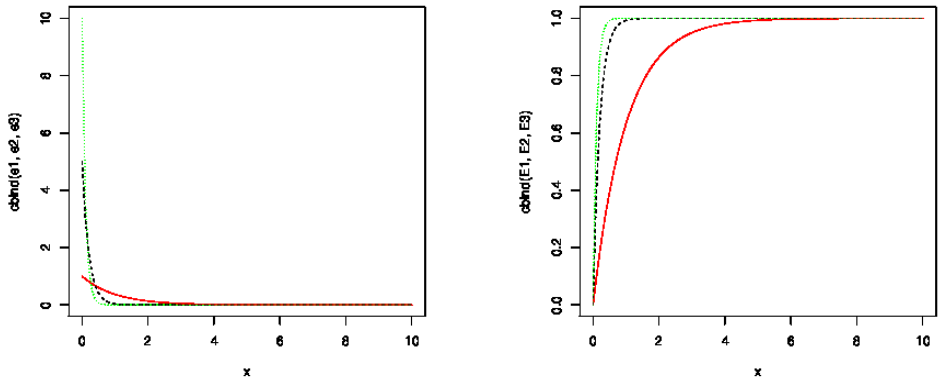


Fig. A.14. Exponential distributions.

```
> 1 - pexp( 110, rate=0.01 )
[1] 0.3328711
```

♢

The *gamma distribution* is a two-parameter family of probability distributions. It is used to model the waiting time until the occurrence of the r-th event. The density of a random variable that is gamma-distributed with shape parameter $r > 0$ and rate parameter $\lambda > 0$ is given by

$$f(x; r, \lambda) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, \quad x > 0.$$

The corresponding cdf is the regularized gamma function

$$F(x; r, \lambda) = \int_0^x f(y; r, \lambda) dy = \frac{\gamma(r, \lambda x)}{\Gamma(r)},$$

where γ is the lower incomplete gamma function. Here are some basic properties of the gamma function:

$$\begin{aligned}
\Gamma(0.5) &= \sqrt{\pi}, \\
\Gamma(1) &= \Gamma(2) = 1, \\
\Gamma(3) &= 2, \\
\Gamma(\infty) &= \infty, \\
\Gamma(x+1) &= x\Gamma(x), \quad x > 0, \\
\frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n+1)} &= \binom{m+n-1}{n}, \quad m, n \in \mathbb{N}.
\end{aligned}$$

The form of the gamma distribution depends on both, the shape and rate parameter. For $0 < r \leq 1$, the density decreased monotonously and for $r > 1$ the density is a shifted bell curve with $f(0) = 0$ and maximum value at $(r-1)/\lambda$.

If $r = 1$, the gamma density is the density of the exponential distribution. If $\lambda = 1/2$ and $r = \nu/2$ where ν is a positive integer, the gamma density is the density of the chi-squared distribution with ν degrees of freedom.

The use of gamma distribution is shown by the following R code (Fig. A.15).

```
> x <- seq( 0, 20, by=0.1 )
> g1 <- dgamma( x, shape=0.5, rate=3 )  # lambda = 3
> g2 <- dgamma( x, shape=1, rate=3 )
> g3 <- dgamma( x, shape=10, rate=3 )
> matplot( x, cbind(g1,g2,g3), type="l", col = c("red","black","green") )
> G1 <- pgamma( x, shape=0.5, rate=3 )  # lambda = 3
> G2 <- pgamma( x, shape=1, rate=3 )
> G3 <- pgamma( x, shape=10, rate=3 )
> matplot( x, cbind(F1,F2,F3), type="l", col = c("red","black","green") )
```

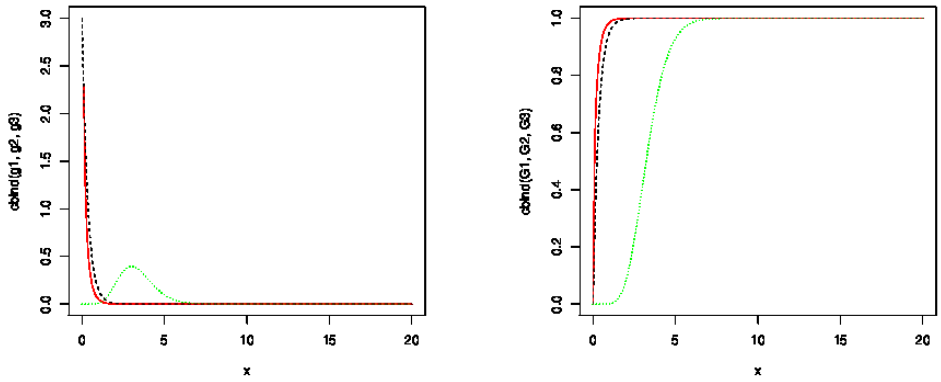


Fig. A.15. Gamma distributions with $r = 0.5$, 1, and 10, and $\lambda = 3$.

The mean and variance of a random variable X which is gamma-distributed with shape parameter $r > 0$ and rate parameter $\lambda > 0$ are respectively

$$E[X] = \frac{r}{\lambda} \quad \text{and} \quad \text{Var}(X) = \frac{r}{\lambda^2}.$$

The parameter of a gamma distribution can be estimated by the method of moments,

$$\hat{r} = \frac{n\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

and

$$\hat{\lambda} = \frac{n\overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Example A.13 (R). Consider the durability (in hours) of ten pressure vessels under working conditions. The sample data are given by the vector `time`.

```
> time <- c( 12.5, 13.7, 17.2, 10.4, 11.4, 18.1, 19.0, 20.1, 12.7, 11.9 )
> n <- length(time)
[1] 10
> m <- mean(time)
[1] 14.7
> r.hat <- (n * m^2)/(sum((time-m)^2))
[1] 19.20459
> l.hat <- (n * m)/(sum((time-m)^2))
[1] 1.306434
```

The calculation provides a gamma distribution with estimated parameters $\hat{r} = 19.20459$ and $\hat{\lambda} = 1.306434$. ♢

Finally, we consider two continuous distributions that are used for statistical testing purposes. The *chi-squared distribution* with ν degrees of freedom, denoted by $\chi^2(\nu)$, is the distribution of a sum of squares of ν independent standard normal random variables. It is one of the most widely used probability distributions in inferential statistics for hypothesis testing and in construction of confidence intervals. The pdf of a $\chi^2(\nu)$ random variable X is

$$f(x) = \frac{1}{\Gamma(\nu/2)2^{\nu/2}} x^{(\nu/2)-1} e^{-x/2}, \quad x \in \mathbb{R},\ \nu \geq 1.$$

where $\Gamma(\nu/2)$ denotes the gamma function which has closed-form values for integer arguments.

The cdf of a chi-squared pdf with ν degrees of freedom is

$$F(x, \nu) = \frac{\gamma(\frac{\nu}{2}, \frac{x}{2})}{\Gamma(\frac{\nu}{2})},$$

where γ is the lower incomplete gamma function. Tables of the chi-squared cdf are widely available in all statistical packages. For instance, the 0.95-quantiles of the chi-squared distributions with degrees for freedom ν, where $1 \leq \nu \leq 10$, can be calculated by `R` as follows,

```
> for (i in 1:10) print( qchisq( 0.95, i ))
[1] 3.841459
[1] 5.991465
[1] 7.814726
[1] 9.487729
[1] 11.0705
[1] 12.59159
[1] 14.06714
[1] 15.50731
[1] 16.91898
[1] 18.30704
```

A $\chi^2(\nu)$ random variable X has respectively the mean and variance

$$E[X] = \nu \quad \text{and} \quad \text{Var}(X) = 2\nu.$$

The use of $\chi^2(\nu)$ random variables is shown by the following R code (Fig. A.16).

```
> x <- seq( 0, 20, by=0.1 )
> f1 <- dchisq( x, 2 )  # nu = 2
> f2 <- dchisq( x, 5 )
> f3 <- dchisq( x, 10 )
> matplot( x, cbind(f1,f2,f3), type="l", col = c("red","black","green") )
> F1 <- pchisq( x, 2 )
> F2 <- pchisq( x, 5 )
> F3 <- pchisq( x, 10 )
> matplot( x, cbind(F1,F2,F3), type="l", col = c("red","black","green") )
```

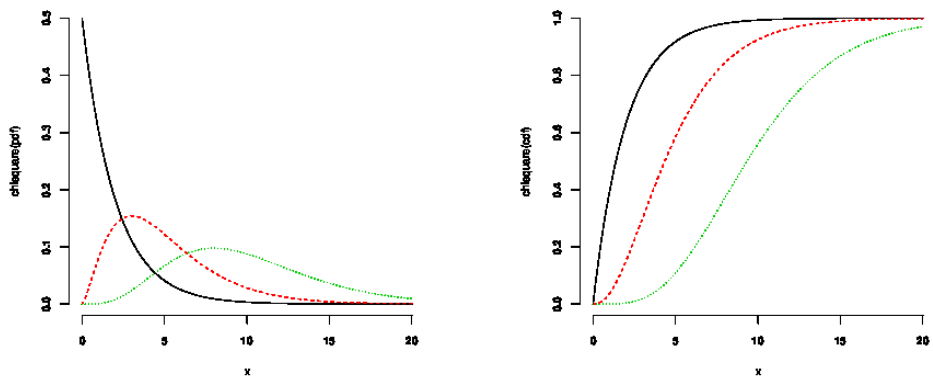


Fig. A.16. Chi-square distributions.

The sum of independent chi-squared random variables is also chi-squared distributed. More specifically, if $X_1, \ldots, X_n$ are independent chi-squared random variables with degrees of freedom $\nu_1, \ldots, \nu_n$ respectively, then $Z = X_1 + \ldots + X_n$ is a chi-squared random variable with $\nu_1 + \ldots + \nu_n$ degrees of freedom. Consider n independent and identically distributed chi-squared random variables $X_1, \ldots, X_n$ each of which with k degrees of freedom. Then the sample mean $\overline{X} = \frac{1}{n}\sum_i X_i$ is distributed according to a gamma distribution with shape $n\nu/2$ and scale $2/n$. If n goes to infinity, then by the central limit theorem, the sample mean converges towards a normal distribution with mean ν and variance 2ν.

The *Student's t distribution* emerges when the mean of a normal distributed population is to be estimated in situations where the sample size is small and the population standard deviation is unknown. Let Z be an $N(0,1)$ random variable and V be a $\chi^2(\nu)$ random variable. If Z and V are independent, the random variable

$$T = \frac{Z}{\sqrt{V/\nu}}$$

has the Student's t distribution with ν degrees of freedom, abbreviated $t(\nu)$. The density of a $t(\nu)$ random variable X is given by

$$f(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu\pi}} \frac{1}{(1+x^2/\nu)^{(\nu+1)/2}}, \quad x \in \mathbb{R},\ \nu \geq 1.$$

The mean and variance of a $t(\nu)$ random variable X are respectively

$$E[X] = 0,\ \nu > 1, \quad \text{and} \quad \mathrm{Var}(X) = \frac{\nu}{\nu-2},\ \nu > 2.$$

The use of Student's t random variables is shown by the following `R` code (Fig. A.17).

```
> x <- seq( -6, 6, by=0.1 )
> f1 <- dt( x, 3 )  # nu = 3
> f2 <- dt( x, 7 )
> f3 <- dt( x, 20 )
> matplot( x, cbind(f1,f2,f3), type="l", col = c("red","black","green") )
> F1 <- pt( x, 3 )  # nu = 3
> F2 <- pt( x, 7 )
> F3 <- pt( x, 20 )
> matplot( x, cbind(F1,F2,F3), type="l", col = c("red","black","green") )
```

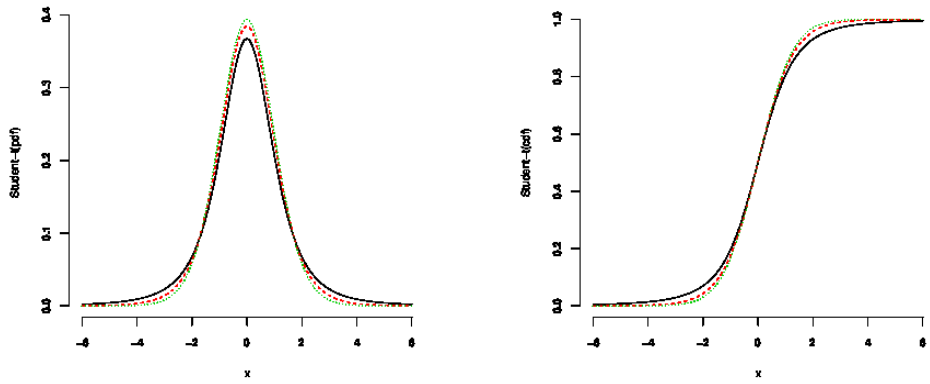


Fig. A.17. Student's t distributions.

A.5 Statistics

Let $X_1, \ldots, X_n$ be a random sample from a distribution with cdf $F_X(x) = P(X \leq x)$, pdf or pmf f_X, mean $E[X] = \mu_X$, and variance $\text{Var}(X) = \sigma_X^2$. Note that lowercase letters $x_1, \ldots, x_n$ denote an observed random sample.

A *statistic* is a function $T_n = T_n(X_1, \ldots, X_n)$ of a sample. Important statistics are the sample mean

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2,$$

and the sample standard deviation $S = \sqrt{S^2}$.

The *empirical cumulative distribution function* (ecdf) of $F_X(x) = P(X \leq x)$ is the proportion of sample points which fall into the interval $(-\infty, x]$. That is, the ecdf of an observed sample $x_1, x_2, \ldots, x_n$ with ordering $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ is given by

$$F_n(x) = \begin{cases} 0, & \text{if } x < x_{(1)}, \\ \frac{i}{n}, & \text{if } x_{(i)} \leq x < x_{(i+1)},\ 1 \leq i \leq n-1, \\ 1, & \text{if } x_{(n)} \leq x. \end{cases}$$

A statistic T_n is an *unbiased* estimator of parameter θ if $E[T_n] = \theta$. An estimator T_n is *asymptotically unbiased* for parameter θ if

$$\lim_{n \to \infty} E[T_n] = \theta.$$

The *bias* of an estimator T_n for parameter θ is given by $\text{bias}(T_n) = E[T_n] - \theta$.

Proposition A.14. *The sample mean $\overline{X}$ is an unbiased estimator of the mean $\mu = E[X]$, and the sample variance S^2 is an unbiased estimator of the variance $\sigma^2 = \text{Var}(X)$.*

Proof. Since the random variables $X_1, \ldots, X_n$ are independent and identically distributed with mean μ, in view of the sample mean we have

$$E[\overline{X}] = \frac{1}{n} \sum_{i} E[X_i] = \frac{1}{n} n\mu = \mu.$$

In view of the sample variance, we have

$$\begin{aligned} E[S^2] &= \frac{1}{n-1} E[\sum_{i=1}^{n} (X_i - \overline{X})^2] \\ &= \frac{1}{n-1} E[\sum_{i=1}^{n} (X_i - \mu)^2 - n(\overline{X} - \mu)^2] \\ &= \frac{1}{n-1} [\sum_{i=1}^{n} E[(X_i - \mu)^2] - nE[(\overline{X} - \mu)^2] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n-1}\left[n\sigma^2 - n\frac{\sigma^2}{n}\right] \\
&= \frac{1}{n-1}[(n-1)\sigma^2] \\
&= \sigma^2.
\end{aligned}
$$

□

However, the statistic S is generally a biased estimator of σ. Indeed, for any random variable X with mean μ, we have $\mathrm{Var}(X) = E[(X-\mu)^2] = E[X^2] - \mu$. Thus $\mathrm{Var}(S) = E[S^2] - E[S]^2 = \sigma^2 - E[S]^2 \geq 0$ and hence $E[S] \leq \sigma$.

The *mean-squared error* (MSE) of an estimator $T = T_n$ for parameter θ is defined as

$$\mathrm{MSE}(T) = E[(T-\theta)^2].$$

Note that for an unbiased estimator the MSE equals by definition the variance of the estimator. However, if T is a biased estimator of θ, the MSE becomes larger than the variance. To see this, we have

$$
\begin{aligned}
E[(T-\theta)^2] &= E[(T - E[T] + E[T] - \theta)^2] \\
&= E[(T-E[T])^2] + 2(E[T]-E[T])(E[T]-\theta) + (E[T]-\theta)^2 \\
&= \mathrm{Var}(T) + (E[T]-\theta)^2 \\
&= \mathrm{Var}(T) + \mathrm{bias}(T)^2,
\end{aligned}
$$

where $\mathrm{bias}(T) = (E[T]-\theta)^2$ is the bias of T and θ.

Example A.15 (R). The function `fitdistr` can be used to provide maximum likelihood fitting of the geometric distribution.

```
> library(MASS)
> set.seed(123)
# generate random numbers
> x <- rgeom( 1000, prob=0.6 )
# maximum likelihood estimation
> z <- fitdistr( x, "geometric")
    prob
 0.61050061
(0.01204868)
# bias
> (z$estimate - 0.6)^2
[1] 0.0001102628
```

♢

A.6 Method of Moments

The method of moments is a technique for the estimation of parameters. This is achieved by deriving equations that relate the estimated moments of a sample set to the parameters of interest.

Let X be a (discrete) random variable with pmf f_X and k be a positive integer. Then the *k-th moment* of X is defined as

$$\mu_k = E[X^k] = \sum_x x^k f_X(x).$$

Let $X_1, \ldots, X_n$ be a random sample of independent and identically distributed random variables with pmf f_X and realizations $x_1, \ldots, x_n$. Then the value

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

is the *k-th sample moment*, an estimate of μ_k.

For instance, the first moment $E[X]$ is the expected mean value and is estimated by

$$\hat{\mu}_1 = \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Moreover, the second moment $E[X^2]$ is estimated by

$$\hat{\mu}_2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2.$$

Thus the expected variance $\sigma^2 = \mathrm{Var}(X) = E[X^2] - E[X]^2$ is estimated by

$$\hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n} \left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right).$$

Example A.16 (R). The method of moments is used to estimate the mean value and the variance of a sample set that has a standard normal distribution.

```
> n <- 1000
> data <- rnorm( n, mean=0, sd=1 )
> m1.hat <- sum(data)/n
[1] 0.02451518
> m2.hat <- sum(data^2)/n
[2] 1.039528
> var.hat <- m2.hat - m1.hat^2
[2] 1.038927
```

The method of moments can be used as a first approximation to the solution of the likelihood equations and further improved approximations can be derived by numerical methods. On the other hand, the maximum likelihood method may lead to better results. But in some cases the maximum likelihood method may be intractable whereas estimators from the method of moments can be efficiently established.

A.7 Maximum-Likelihood Estimation

Maximum-likelihood estimation is a method of estimating the parameters of a statistical model given sample data.

Let $X_1, \ldots, X_n$ be random variables with parameter or parameter vector $\theta \in \Theta$, where Θ is the parameter space of possible parameters. The *likelihood function* $L(\theta)$ of random variables $X_1, \ldots, X_n$ with realizations $x_1, \ldots, x_n$ is defined as the joint density

$$L(\theta) = f(x_1, \ldots, x_n|\theta).$$

If $X_1, \ldots, X_n$ are a random sample of independent and identically distributed random variables with density $f(x|\theta)$, then

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

A *maximum likelihood estimate* of θ is a value or vector $\hat{\theta}$ which maximizes $L(\theta)$. That is, $\hat{\theta}$ is a solution of the maximization problem

$$L(\hat{\theta}) = f(x_1, \ldots, x_n|\hat{\theta}) = \max\{f(x_1, \ldots, x_n|\theta) \mid \theta \in \Theta\}.$$

If the estimate $\hat{\theta}$ is uniquely determined, then $\hat{\theta}$ is the *maximum likelihood estimator* (MLE) of θ.

If the parameter space Θ is an real-valued interval in $\mathbb{R}$ and the function $L(\theta)$ is differentiable and takes on a maximum on Θ, then $\hat{\theta}$ is a solution of

$$\frac{d}{d\theta} L(\theta) = 0.$$

Since the logarithm function is monotonous and differentiable, it is often easier to consider the *log-likelihood function*

$$\ell(\theta) = \log L(\theta).$$

The maximum likelihood estimates of $L(\theta)$ and $\ell(\theta)$ are the same. In particular, if $X_1, \ldots, X_n$ are a random sample of independent and identically distributed random variables with density $f(x|\theta)$, then the log-likelihood function becomes

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta).$$

The maximum likelihood estimates can often be determined analytically. If not, numerial optimization or other computational methods like heuristics can be used.

Example A.17 (R). Let $X_1, \ldots, X_n$ be a random sample with density

$$f(x|\theta) = \frac{\theta}{2} e^{-\theta x}, \quad x \in \mathbb{R},$$

and realizations $x_1, \ldots, x_n$. The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \frac{\theta}{2} e^{-\theta x_i} = \frac{\theta^n}{2^n} e^{-\theta(x_1+\ldots+x_n)}$$

and the log-likelihood function is

$$\ell(\theta) = n \log \theta - \theta(x_1 + \ldots + x_n) - \log 2^n.$$

The first derivate of $\ell(\theta)$ gives the equation

$$\frac{d}{d\theta}\ell(\theta) = \frac{n}{\theta} - (x_1 + \ldots + x_n) = 0.$$

The unique solution provides the MLE

$$\hat{\theta} = \frac{n}{x_1 + \ldots + x_n},$$

which amounts to the reciprocal sample mean. A numerical solution in R can be obtained as follows.

```
# numerical optimization
> x <- c( 0.25,0.41,0.37,0.54 )
# minus log-likelihood of density, initial value theta=1
> mlog <- function(theta=1) { return( -(length(x) * log(theta) - theta * sum(x) )) }
> library( stats4 )
> y <- mle( mlog )
> summary( y )

Call:
mle(minuslog1 = mlog)

Coefficients
   theta
2.547728
# analytic optimization (above)
> opt-theta <- length(x) / sum(x); opt-theta
[1] 2.547771
```

♢

Example A.18 (R). Let $X_1, \ldots, X_n$ be a random sample of independent and poisson identically distributed random variables with parameter $\lambda > 0$ and realizations $x_1, \ldots, x_n$. The likelihood function is

$$L(\lambda) = e^{n\lambda} \frac{\lambda^{x_1+\ldots+x_n}}{x_1! \cdots x_n!}$$

and the log-likelihood function is

$$\ell(\lambda) = \log L(\lambda) = -n\lambda + (x_1 + \ldots + x_n) \log \lambda - \log(x_1! \cdots x_n!).$$

Then

$$\frac{d}{d\lambda}\ell(\lambda) = -n + (x_1 + \ldots + x_n)\frac{1}{\lambda} = 0$$

implies

$$\hat{\lambda} = \frac{x_1 + \ldots + x_n}{n}.$$

The function `fitdistr` provides maximum likelihood fitting of the poisson distribution.

```
> library(MASS)
> set.seed(123)
# generate random numbers
> x <- rpois( 1000, lambda=5 )
# maximum likelihood estimation
> fitdistr( x, "poisson" )
# output: estimate, sd, vcov, loglik
     lambda
 5.01000000
(0.07078153)
> mean(x)
[1] 5.01
```

♢

Suppose the density is a multivariate function $f(x_1, \ldots, x_n|\theta)$, where θ is a vector in $\mathbb{R}^d$, the parameter space Θ is an open subset of $\mathbb{R}^d$, and the derivatives of the maximum likelihood function $L(\theta)$ exist in all coordinates. Then the maximum likelihood estimate $\hat{\theta}$ has to fulfill simultaneously d equations

$$\frac{\partial}{\partial\theta_i}L(\theta) = 0, \quad 1 \le i \le d.$$

This gives a system of d equations in d unknowns.

Example A.19 (R). Let $X_1, \ldots, X_n$ be a random sample of independent and identically distributed normal random variables with parameters μ and σ and let $x_1, \ldots, x_n$ be a realization. Then the likelihood function is

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

and the log-likelihood function is

$$\ell(\mu, \sigma) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Then the partial derivatives give

$$\frac{\partial}{\partial\mu}\ell = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$

and

$$\frac{\partial}{\partial\sigma}\ell = -\frac{n}{2\sigma^2} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = 0.$$

Thus the maximum likelihood estimates are

$$\hat{\mu} = \frac{x_1 + \ldots + x_n}{n}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}).$$

The function `fitdistr` provides maximum likelihood fitting of the normal distribution.

```
> library(MASS)
> set.seed(123)
# generate random numbers
> x <- rnorm( 1000, mean=80, sd=15 )
# maximum likelihood estimation
> fitdistr( x, "normal")
# output: estimate, sd, vcov, loglik
      mean           sd
80.3936556   15.1467157
(0.4789812) (0.3386909)
```

♢

A.8 Ordinary Least Squares

In statistics, ordinary reast squares (OLS) is a method for estimating the unknown parameter in a linear or nonlinear regression model.

Example A.20 (R). In the classical linear regression model, consider sample data consisting of n observations (x_i, y_i), $1 \leq i \leq n$. The response variable is a linear function of the independent variable

$$y_i = \alpha + \beta x_i, \quad 1 \leq i \leq n.$$

The ordinary least squares approximation seeks parameters α and β such that the following expression becomes minimal,

$$S(\alpha, \beta) = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2.$$

We have

$$\frac{\partial S}{\partial\alpha} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)$$

and

$$\frac{\partial S}{\partial \beta} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)x_i.$$

By setting both derivatives to zero, we obtain

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})}$$

and

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}.$$

The following data are pertubed with white noise and the parameters of the linear regression model are calculated by using a linear regression model.

```
> x <- seq( 0, 10, by=0.1 )
> n <- length( x )
> set.seed(210)
> e <- rnorm(n, mean=0, sd=1 ) # white noise
> y <- 10 + 2*x + e
> lm( y ~ x )                  # linear model

Call
lm(formula) = y ~ x)

Coefficient
(Intercept)           x
     10.160       1.971
```

Thus we obtain $\hat{\alpha} = 10.160$ and $\hat{\beta} = 1.971$ (Fig. A.18). ♢

Example A.21 (R). In the nonlinear regression model, take sample data consisting of n observations (x_i, y_i), $1 \leq i \leq n$. The response variable is an exponential function of the independent variable

$$y_i = p_1 e^{-p_2 \cdot x_i}, \quad 1 \leq i \leq n.$$

The least squares approximation seeks parameters p_1 and p_2 such that the following expression becomes minimal,

$$E(p_1, p_2) = \sum_{i=1}^{n} \left(y_i - p_1 e^{-p_2 \cdot x_i}\right)^2.$$

The following data are pertubed with white noise and the parameters of the regression model are calculated by using a nonlinear regression model.

```
> x <- seq( 0, 10, by=0.1 )
> n <- length( x )
> set.seed(210)
> e <- rnorm(n, mean=0, sd=1 )  # white noise
> y <- 10/exp(0.5*x) + e
```

```
> nls( y ~ p1/exp(p2*x) , start=list(p1=1,p2=1) ) # nonlinear model
Nonlinear regression model
   model: y ~ p1/exp(p2 * x)
    data: parent.frame()
     p1      p2
10.2135  0.1044
 residual sum-of-squares: 107.1

Number of iterations to converge: 7
Achieved convergence tolerance: 5.338e-08
```

Thus we obtain $\hat{p}_1 = 10.2135$ and $\hat{p}_2 = 0.1044$ (Fig. A.18). ♢

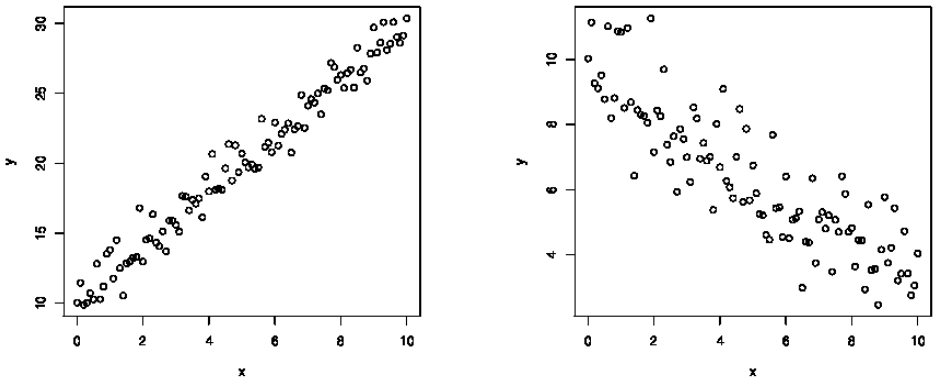


Fig. A.18. Sample data from linear and nonlinear regression model with white noise.

A.9 Parameter Optimization

The language R provides several functions for one- and two-dimensional parameter optimization. For the optimization of univariate functions, the function `optimize` can be used.

Example A.22 (R). Consider the real-valued function

$$f(x) = \frac{\log(2 + \log(x))}{\log(2 + x)}, \quad x \geq 0.$$

The graph of the function $f(x)$ is shown in Fig. A.19. The drawing of the function in the interval $[1, 10]$ can be done by the following code.

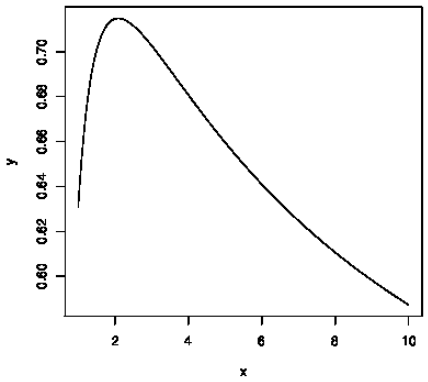


Fig. A.19. Graph of function $f(x)$.

```
> x <- seq( 1, 10, 0.01 )
> y <- log(2+log(x))/log(2+x)
> plot( x, y, type="l", xlab="x",ylab="P(x)" )
```

As can be seen, the optimum lies in the interval $[2, 4]$. Therefore, we apply the function `optimize` to this interval. The default is to minimize the function. To maximize $f(x)$, set `maximize` to `TRUE`. The following code yields the optimal parameter value $\hat{x} = 2.090302$ and the optimal function value $f(\hat{x}) = 0.714867$.

```
> f <- function( x )
+        log(2+log(x))/(log(2+x))
> optimize( f, lower=2, upper=4, maximum=TRUE )
$maximum
[1] 2.090302
$objective
[1] 0.714867
```

♢

For the optimization of bivariate functions, the function `optim` can be used.

Example A.23 (R). Let $X_1, \ldots, X_n$ be a random sample of independent and identically distributed random variables from the gamma(r, λ) distribution with shape parameter $r > 0$ and rate parameter $\lambda > 0$, and let $x_1, \ldots, x_n$ be a non-negative realization. Then the likelihood function is

$$L(r, \lambda) = \frac{\lambda^{nr}}{\Gamma(r)^n} \prod_{i=1}^{n} x_i^{r-1} \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)$$

and the log-likelihood function is

$$\ell(r,\lambda) = nr\log\lambda - n\log\Gamma(r) + (r-1)\sum_{i=1}^{n}\log x_i - \lambda\sum_{i=1}^{n} x_i.$$

The problem to maximize the log-likelihood function with respect to r and λ is a two-dimensional problem. The log-likelihood function can be implemented as follows.

```
LogL <- function( theta, sx, slogx, n ) {
+         r <- theta[1]
+         lambda <- theta[2]
+         val <- n*r*log(lambda)+(r-1)*slogx - lambda*sx - n*log(gamma(r))
+         -val
+       }
```

Note that function `optim` performs minimization by default and therefore the return value is $-\ell(r,\lambda)$. Initial values need to be chosen with care. For this problem, the method of moments can be used for the initial values of the parameters. For simplicity, the initial values are set to $r = 1$ and $\lambda = 1$. In the following, x denotes a random sample of length $n = 200$. Then the parameter estimation can be achieved as follows.

```
> n <- 200
> r <- 5; lambda <- 2
> x <- rgamma( n, shape=r, rate=lambda )
> optim( c(1,1), LogL, sx=sum(x), slogx=sum(log(x)), n=n )
$par
[1] 5.094346 2.052238
$value
[1] 289.0687
$counts
function gradient
      75       NA
$convergence
[1] 0
$message
NULL
```

The result shows that the optimization method (Nelder-Mead as default) converges successfully to the maximum likelihood estimates $\hat{r} = 5.094346$ and $\hat{\lambda} = 2.052238$. The error code `$convergence` is 0 for successful runs and otherwise indicates a problem.

Now the procedure is repeated 1,000 times to provide better parameter estimates.

```
> test <- replicate(1000, expr = {
+            x <- rgamma(200, shape=5, rate=2)
+            optim( c(1,1), LogL, sx=sum(x), slogx=sum(log(x)), n=n )$par
+          })
> colMeans(t(test))
[1] 5.042200 2.019213
```

The output exhibits the average estimated values which are slightly better than the parameter values established in one run. ♢

B

Spectral Analysis of Ranked Data

Human and particularly US Americans seem to be unable to avoid ranking things. Top Five/Ten/Twenty/Hundred lists are plentiful: Best (Worst) Dressed, Best Suburbs, Most Watched Movies, Videos, Best Football Teams, and so on. Rankings have very serious uses as well. Companies need to know what products consumers prefer, social and political leaders need to know what the society values, and elections need to be conducted. Data consisting of ranking appear in Psychology, Animal Science, Educational Testing, Sociology, Economics, and Biology. Indeed, for almost any situation where there are data it can be helpful to transform the data into ranks. This chapter is aimed at preference rankings.

B.1 Data Analysis

Data sometimes come in the form of ranks or preferences. A group of people may be asked to rank order five brands of chocolate chip cookies. Each person tastes the cookies and ranks all five. This results in a ranking $\pi(1)$, $\pi(2)$, $\pi(3)$, $\pi(4)$, and $\pi(5)$, with $\pi(i)$ the rank given the ith brand. The collection of rankings then makes up the data set. Elections are also sometimes based on rankings. For instance, the American Psychological Association asks its members to rank order five candidates for president. An analysis of the data from one such election is presented in this chapter.

Almost anyone analyzing ranked data looks at simple averages such as the proportion of times each item was ranked first or last and the average rank for each item. These are first order statistics since they are linear combinations of the number of times item i was ranked in position j. There are also natural second order statistics based on the number of times items i and i' are ranked in positions j and j'. These come in ordered and unordered modes. Similarly, there are third and higher order statistics of various types.

A basic paradigm of data analysis is to take out some found structure and to look at the rest. Thus to look at second order statistics it is natural to subtract away the observed first order structure. This leads to a natural decomposition of the original data into orthogonal parts. The decomposition is somewhat more complicated than standard analysis of variance decompositions because of the dependence inherent in the permutation structure.

Example B.1. In a study, responses of 2,262 German citizens were asked to rank order the desirability of four political goals (1989):

1. Maintain order;
2. give people more say in government;
3. fight rising prices;
4. protect freedom of speech.

The data appear as

1234	137	2134	48	3124	330	4123	21
1243	29	2143	23	3142	294	4132	30
1324	309	2314	61	3214	117	4213	29
1342	255	2341	55	3241	69	4231	52
1423	52	2413	33	3412	70	4312	35
1432	93	2431	39	3421	34	4321	27
	875		279		914		194 2262

Thus 137 people ranked (1) first, (2) second, (3) third, and (4) fourth. The marginal totals show people thought item (3) is most important (914) and ranked it first. The first order summary is the 4×4 matrix,

$$\mathbf{F} = \begin{pmatrix} 875 & 279 & 914 & 194 \\ 746 & 433 & 742 & 341 \\ 345 & 773 & 419 & 725 \\ 296 & 777 & 187 & 1002 \end{pmatrix}.$$

The first row shows the number of people ranking a given item first, while the last row exhibits the number of people ranking a given item last. The data was collected in part to study whether the population could be usefully broken into "liberals" who might favor items (2) and (4), and " conservatives" who might favor items (1) and (3). ♢

Example B.2. The American Psychological Association (APA) is a large professional organization of academicians, clinicians, and all shades in between. The APA elects a president every year by asking each member to rank order a slate of five candidates. There were about 50,000 APA members in 1980 and about 15,000 members voted. Many members cast incomplete ballots, voting for their favorite q of five candidates, $1 \leq q \leq 3$. These ballots are illustrated in Table B.1. For instance, 1022 members ranked candidate 5 first and left the other unranked, while 142 members ranked candidate 5 first and candidate 4 second; zeros and blanks indicate unranked candidates. Moreover, the 5738 complete ballots are tabulated in Table B.2. For instance, 29 members ranked candidate 5 first, candidate 4 second, candidate 3 third, candidate 2 fourth, and candidate 1 fifth. In the next section, we will mainly consider useful summaries of these data given by simple averages (first order effects).

The APA chooses a winner by the *Hare system* also known as proportional voting: If one of the five candidates is ranked first by more than half of the voters, the candidate wins. If not, the candidate with the fewest first place votes is eliminated, each of the remaining candidates is reranked in relative order and the method is inductively applied. Candidate 1 is the eventual winner here. ♢

B.2 Representation Theory for Partial Rankings

Consider a list of n items. Let λ be a partition of n. A *partial ranking of shape* λ is specified according to the following instructions: Choose your favorite λ_1 items from the list but do not bother to rank

Table B.1. APA election data; incomplete ballots; 5,141 votes ($q = 1$) and 2,462 votes ($q = 2$).

$q = 1$		$q = 2$	
ranking	votes	ranking	votes
1	1022	21	143
10	1145	12	196
100	1198	201	64
1000	881	210	48
10000	895	102	93
		120	56
		2001	70
		2010	114
		2100	89
		1002	80
		1020	87
		1200	51
		20001	117
		20010	104
		20100	547
		21000	72
		10002	72
		10020	74
		10200	302
		12000	83

within. Then choose your next λ_2 favorite items from the list but do not rank within and so on. Such partial rankings can be written as λ-tabloids such as

$$\overline{\underline{1\ 3\ 5}\ \underline{2\ 4}},$$

where the items 1, 3, 5 are ranked first and items 2, 4 are ranked second.

With this interpretation, the set of all partial rankings of shape λ forms a basis of the permutation module M^λ. Moreover, the elements of the $\mathbb{R}$-space M^λ can be considered as the set of all real-valued functions f on partial rankings given as linear combinations of partial rankings,

$$f = \sum_{\{T\}} f_{\{T\}}\{T\}, \quad f_{\{T\}} \in \mathbb{R}.$$

There is a simple method for computing the projection of a permutation module onto its isotypic subspaces. This involves the character table of the group.

Theorem B.3. *Let λ and μ be partitions of n. The orthogonal projection of an element $f \in M^\lambda$ onto the isotypic subspace V_λ^μ is the function*

$$\bar{f}_\mu(x) = \frac{\chi_\mu(id)}{n!} \sum_{\pi \in S_n} \chi_\mu(\pi) f(\pi^{-1}(x)).$$

Table B.2. APA election data; complete ballots; 5,738 votes.

ranking	votes	ranking	votes	ranking	votes	ranking	votes
54321	29	43521	91	32541	41	21543	36
54312	67	43512	84	32514	64	21534	42
54231	37	43251	30	32451	34	21453	24
54213	24	43215	35	32415	75	21435	26
54132	43	43152	38	32154	82	21354	30
54123	28	43125	35	32145	74	21345	40
53421	57	42531	58	31542	30	15432	40
53412	49	42513	66	31524	34	15423	35
53241	22	42351	24	31452	40	15342	36
53214	22	42315	51	31425	42	15324	17
53142	34	42153	52	31254	30	15243	70
53124	26	42135	40	31245	34	15234	50
52431	54	41532	50	25431	35	14532	52
52413	44	41523	45	25413	34	14523	48
52341	26	41352	31	25341	40	14352	51
52314	24	41325	23	25314	21	14325	24
52143	35	41253	22	25143	106	14253	70
52134	50	41235	16	25134	79	14235	45
51432	50	35421	71	24531	63	13542	35
51423	46	35412	61	24513	53	13524	28
51342	25	35241	41	24351	44	13452	37
51324	19	35214	27	24315	28	13425	35
51243	11	35142	45	24153	162	13254	95
51234	29	35124	36	24135	96	13245	102
45321	31	34521	107	23541	45	12543	34
45312	54	34512	133	23514	52	12534	35
45231	34	34251	62	23451	53	12453	29
45213	24	34215	28	23415	52	12435	27
45132	38	34152	87	23154	186	12354	28
45123	30	34125	35	23145	172	12345	30

B.4. **First Order Projection – Winner-takes-it-all** *The simple choice of 1 out of n leads to a partial ranking of shape* $(n-1,1)$*. The real-valued functions on the corresponding permutation module* $M^{(n-1,1)}$ *are of the form*

$$f = \sum_{i=1}^{n} f(i) \overline{\underset{\underset{i}{\frown}}{\cdot \ldots \cdot}}, \quad f(i) \in \mathbb{R}.$$

The module $M^{(n-1,1)}$ *has the splitting*

$$M^{(n-1,1)} = S^{(n)} \oplus S^{(n-1,1)}.$$

The data vector $f \in M^{(n-1,1)}$ *has the decomposition* $f = F + (f - F)$*, with* $F = (f(1) + \ldots + f(n))/n$*.*

In view of the APA election data with $q = 1$ *and* $5,141$ *items given in Table B.1, we obtain* $F = 5,141/5 = 1,028$*. The projection onto the isotypic submodule* $S^{(4,1)}$ *merely amounts to subtracting the number of rankers divided by 5 from the original data vector,*

Candidate	Projection onto $S^{(4,1)}$
1	−133
2	−147
3	170
4	117
5	−6

It follows that candidate 3 is the most popular followed by candidate 4. ♢

B.5. **Second Order Projection – Unordered Pairs** The choice of an unordered pair out of n amounts to a partial ranking of shape $(n-2,2)$. The real-valued functions on the corresponding permutation module $M^{(n-2,2)}$ are given as

$$f = \sum_{i,j} f(\{i,j\}) \overline{\underline{\cdot\ \cdot}\ \underline{\cdot\ \cdot\ \cdot}}, \quad f(\{i,j\}) \in \mathbb{R}.$$

The associated permutation module has the decomposition

$$M^{(n-2,2)} = S^{(n)} \oplus S^{(n-1,1)} \oplus S^{(n-2,2)}.$$

The projection onto the trivial submodule $S^{(n)}$ is the mean $F = \sum_{i,j} f(\{i,j\})/n(n-1)$. Moreover, the projection onto the submodule $S^{(n-1,1)}$ is given by $\bar{f}(i) = \sum_j f(\{i,j\}) - F_1$ with $F_1 = \sum_{i,j} f(\{i,j\})/n$, $1 \le i \le n$. The projection onto the submodule $S^{(n-2,2)}$ is what is left after the mean and the popularity of individual terms are taken out. The first order analysis gives

Candidate	Projection onto $S^{(n-1,1)}$
1	939
2	154
3	758
4	839
5	343

Thus there is a slight difference in the first order statistics when compared with the previous statistics. Candidate 1 is the most popular followed by candidate 4, while candidate 3 is only third ranked. ♢

B.6. **Second Order Projection – Ordered Pairs** The choice of an ordered pair out of n gives rise to a partial ranking of shape $(n-2,1^2)$. The real-valued functions on the corresponding permutation module $M^{(n-2,1^2)}$ are defined as

$$f = \sum_{i,j} f(i,j) \overline{\underline{\cdot}\ \underline{\cdot}\ \cdot}, \quad f(i,j) \in \mathbb{R}.$$

The corresponding permutation module splits as follows,

$$M^{(n-2,1^2)} = S^{(n)} \oplus 2 \cdot S^{(n-1,1)} \oplus S^{(n-2,2)} \oplus S^{(n-2,1^2)}.$$

The two copies of the space $S^{(n-1,1)}$ provide the effect of an item in first and second position. The projection onto the subspace $S^{(n-2,2)}$ describes an unordered pair effect, while the projection onto $S^{(n-2,1^2)}$ gives an ordered pair effect.

The projection onto the trivial submodule $S^{(n)}$ is the mean $F = \sum_{i,j} f(i,j)/n(n-1)$. Moreover, the projection onto the two submodules $S^{(n-1,1)}$ accounts for the position one and two. Furthermore, the projection onto the submodule $S^{(n-2,2)}$ provides an unordered pair effect, while the projection onto the submodule $S^{(n-2,1^2)}$ yields an ordered pair effect after the mean, first order and unordered pair effects were removed.

In view of the APA election data with $q = 2$ and $2,462$ items given in Table B.1, we obtain $F = 2,492/20 = 123$. The projection onto first and second submodule $S^{(4,1)}$ amounts to counting the number of votes for candidate i to be ranked first and second, respectively, and subtracting the number of rankers divided by 5,

Candidate	Projection onto 1st $S^{(n-1,1)}$	Projection onto 2nd $S^{(n-1,1)}$
1	39	348
2	−201	−135
3	291	−26
4	−29	−31
5	−97	−50

The first order statistics ranks candidate 3 first and candidate 1 second. ♢

B.7. **Complete Ranking** The partition (1^n) with all parts equal to 1 corresponds to complete rankings. The associated permutation module $M^{(1^n)}$ is isomorphic to the group algebra of the group S_n. Thus the real-valued functions on complete rankings can be written as

$$f = \sum_{\pi \in S_n} f(\pi)\pi, \quad f(\pi) \in \mathbb{R}.$$

The permutation module has the decomposition

$$M^{(1^n)} = \bigoplus_{\lambda} (\dim S^\lambda) S^\lambda.$$

For instance, the decomposition of the permutation module for the complete rankings of $n = 5$ items is given as

$$M^{(1^5)} = S^{(5)} \oplus 4 \cdot S^{(4,1)} \oplus 5 \cdot S^{(3,2)} \oplus 6 \cdot S^{(3,1^2)} \oplus 5 \cdot S^{(2^2,1)} \oplus 4 \cdot S^{(2,1^3)} \oplus S^{(1^5)}.$$

The four copies of the submodule $S^{(4,1)}$ provide the effect of an item in one of the five positions; only four of these positions are independent.

In view of the APA election data, there are 5,738 complete ballots given in Table B.2. The projection onto the trivial module $S^{(5)}$ is imply the mean, and the projection onto the four copies of the submodule $S^{(4,1)}$ providing the first order effects is given in Table B.3. This table shows the percentage of voters ranking candidate i in position j. Thus, candidate 3 is the most popular, being ranked first by 28% of the voters, but candidate 3 also had some hate vote. Candidate 1 is strongest in the second position, she

Table B.3. First order statistic: Percentage of voters ranking candidate i in position j.

	rank				
candidate	1	2	3	4	5
1	18	26	23	17	15
2	14	19	25	24	18
3	28	17	14	18	23
4	20	17	19	20	23
5	20	21	20	19	20

has no hate vote and a lower average rank than candidate 3. The voters seem indifferent on candidate 5.

♢

C

Representation Theory of the Symmetric Group

C.1 The Symmetric Group

A function from the set $[n] = \{1, \ldots, n\}$ onto itself is called a *permutation* of n, and the set of all permutations of n, together with the usual composition of functions, is the *symmetric group* of degree n, which will be denoted by S_n. The symmetric group S_n has $n!$ elements. If X is a subset of $[n]$, then S_X denotes the subgroup of S_n that fixes every number outside of X.

A permutation π of S_n is generally written in two-row format

$$\pi = \begin{pmatrix} 1 & 2 & 3 & \ldots & n \\ \pi(1) & \pi(2) & \pi(3) & \ldots & \pi(n) \end{pmatrix}.$$

By consider the orbits $\{i, \pi(i), \pi^2(i), \ldots\}$, $i \in [n]$, of the group generated by a permutation π, it follows that π can be written as a product of disjoint cycles, as in the example

$$\pi = \begin{pmatrix} 1\,2\,3\,4\,5\,6\,7 \\ 2\,1\,4\,5\,6\,3\,7 \end{pmatrix} = (12)(3456)(7).$$

We usually suppress the 1-cycles. A permutation π that interchanges different numbers a, b and leaves the other number fixed, is called a *transposition* and is written as $\pi = (ab)$. Since the cycle $(i_1 i_2 \ldots i_k)$ equals the product $(i_1 i_2)(i_1 i_3) \ldots (i_1 i_k)$, any cycle, and hence any permutation, can be written as a product of transpositions.

Moreover, if $\pi = \sigma_1 \sigma_2 \ldots \sigma_k = \tau_1 \tau_2 \ldots \tau_l$ are two ways of writing π as a product of transpositions, then it can be proved that $k - l$ is even. Hence, the *signature function* $\mathrm{sgn} : S_n \to \{\pm 1\}$ such that $\mathrm{sgn}(\pi) = (-1)^k$, when π is a product of k transpositions, is well-defined. We have $\pi(1) = 1$. For any two permutations σ and τ of n, which are products of k and l transpositions, respectively, $\mathrm{sgn}(\sigma\tau) = (-1)^{k+l} = (-1)^k \cdot (-1)^l = \mathrm{sgn}(\sigma)\mathrm{sgn}(\tau)$. Thus the sign mapping is a homomorphism. It follows that for any permutation π of n, we have $1 = \mathrm{sgn}(1) = \mathrm{sgn}(\pi^{-1}\pi) = \mathrm{sgn}(\pi^{-1})\mathrm{sgn}(\pi)$ and thus $\mathrm{sgn}(\pi)^{-1} = \mathrm{sgn}(\pi^{-1})$.

Two permutations σ and τ of n are called *conjugate* if there exists a permutation π of n such that $\tau = \pi\sigma\pi^{-1}$. Conjugation is an equivalence relation on the set of permutations of n, and the corresponding equivalence classes are called the *conjugacy classes* of n.

A sequence $\lambda = (\lambda_1, \lambda_2, \dots,)$ of non-negative integers such that $\sum_i \lambda_i = n$ is called an *improper partition* of n. An improper partition λ of n is called *proper* if the entries are monotonically decreasing; that is, $\lambda_1 \geq \lambda_2 \geq \dots$. Proper partitions are also simply called partitions. A partition is usually written as a finite sequence in which trailing zeros are deleted, such as $(4, 2, 2, 1, 0, 0, \dots) = (4, 2, 2, 1) = (4, 2^2, 1)$. If λ is a partition of n such that $\lambda_m > 0$ and $\lambda_{m+s} = 0$ for all $s \geq 1$, then λ is called a partition of n into m parts.

Let λ be a partition of n. A permutation π is said to have *cycle-type* λ if the orbits of the group generated by π have lengths $\lambda_1 \geq \lambda_2 \geq \dots$. For instance, the permutation $(12)(3456)(7)$ has the cycle-type $(4, 2, 1)$.

Let P_n denote the set of all partitions of n. Consider the map that assigns to each permutation π of n the corresponding cycle-type in P_n. This map is onto since each partition λ of n gives rise to a permutation $\pi = (1, \dots, \lambda_1)(\lambda_1 + 1, \dots, \lambda_1 + \lambda_2) \dots$ of n that has cycle-type λ. For instance, the partition $(3, 2, 1)$ gives rise to the permutation $(123)(45)(6)$. Moreover, for each cycle $(i_1, \dots, i_k)$ and each permutation π of n, we have

$$\pi(i_1, \dots, i_k)\pi^{-1} = (\pi(i_1), \dots, \pi(i_k)).$$

Thus a permutation of n and all of its conjugate permutations have the same cycle-type. Conversely, suppose σ and τ are permutations of n having the same cycle-type λ. Write $\sigma = (\dots) \dots (i_1, \dots, i_k) \dots (\dots)$ and $\tau = (\dots) \dots (j_1, \dots, j_k) \dots (\dots)$. Then the permutation π of n that assigns $i_l \mapsto j_l$ has the property $\tau = \pi\sigma\pi^{-1}$, and thus π and τ are conjugate. It follows that the map is also one-to-one. Thus we have shown the following

Proposition C.1. *The number of conjugacy classes of S_n equals the number of partitions of n.*

Let $\mathbb{K}$ be a field. The *group algebra* $\mathbb{K}S_n$ is a vector space over $\mathbb{K}$ with the elements of S_n as a basis. The elements of $\mathbb{K}S_n$ are written as linear combinations of the elements in S_n with coefficients in $\mathbb{K}$,

$$\sum_{\pi \in S_n} a_\pi \pi, \quad a_\pi \in \mathbb{K}. \tag{C.1}$$

The addition of two elements in the group algebra is given as

$$\left(\sum_{\pi \in S_n} a_\pi \pi \right) + \left(\sum_{\pi \in S_n} b_\pi \pi \right) = \sum_{\pi \in S_n} (a_\pi + b_\pi)\pi. \tag{C.2}$$

This vector space has the dimension $|S_n| = n!$. The multiplication in the group algebra is defined by linear extension of the group multiplication as

$$\left(\sum_{\pi \in S_n} a_\pi \pi \right) \left(\sum_{\sigma \in S_n} b_\sigma \sigma \right) = \sum_{\pi, \sigma \in S_n} a_\pi b_\sigma \pi\sigma = \sum_{\sigma \in S_n} \left(\sum_{\pi \in S_n} a_\pi b_{\pi^{-1}\sigma} \sigma \right). \tag{C.3}$$

This multiplication is associative and turns the group algebra $\mathbb{K}S_n$ into a unitary ring. Notice that the group algebra $\mathbb{K}G$ of any finite group G is similarly defined.

C.2 Diagrams, Tableaux, and Tabloids

Let λ be a partition of n. The *diagram* of λ is the set

$$[\lambda] = \{(i,j) \mid i \geq 1,\ 1 \leq j \leq \lambda_i\}.$$

Each element (i,j) in $[\lambda]$ is called a *node*. The kth row (column) of a diagram consists of those nodes whose first (second) coordinate is k. For instance, the diagram of the partition $\lambda = (4, 2^2, 1)$ is

$$[\lambda] = \begin{array}{cccc} \times & \times & \times & \times \\ \times & \times & & \\ \times & \times & & \\ \times & & & \end{array} .$$

Let λ and μ be partitions of n. We say that λ *dominates* μ, briefly $\lambda \trianglerighteq \mu$, provided that for all $k \geq 1$,

$$\sum_{i=1}^{k} \lambda_i \geq \sum_{i=1}^{k} \mu_i.$$

For instance, we have $(6) \trianglerighteq (5,1) \trianglerighteq (3,3) \trianglerighteq (3,2,1) \trianglerighteq (3,1^3) \trianglerighteq (2^2,1^2) \trianglerighteq (2,1^4) \trianglerighteq (1^6)$.

Moreover, we write $\lambda > \mu$ if the least number j for which $\lambda_j \neq \mu_j$ satisfies $\lambda_j > \mu_j$. This is a total order and called the *dictionary order* on partitions of n. The dictionary order contains the dominance order in the sense that $\lambda \trianglerighteq \mu$ implies $\lambda > \mu$. For instance, we have $(3,1^3) > (2^3)$ but $(3,1^3) \not\trianglerighteq (2^3)$.

Let $[\lambda]$ be a diagram. The *conjugate* diagram $[\lambda']$ is obtained by interchanging the rows and columns in the diagram $[\lambda]$. The corresponding partition λ' of n is called the partition *conjugate* to λ. For instance, if $\lambda = (4, 2, 1)$ then $\lambda' = (4, 3, 1^2)$. Notice that the part λ'_i in λ' is the number of parts λ_j in λ that are greater than or equal to i. The conjugate partition of a proper partition is also proper.

Proposition C.2. *We have $\lambda \trianglerighteq \mu$ if and only if $\mu' \trianglerighteq \lambda'$.*

Proof. Let $\lambda \trianglerighteq \mu$. There exists an index k such that $\lambda_i = \mu_i$ for $1 \leq i \leq k-1$ and $\lambda_k > \mu_k$. Then λ'_k, the number of parts λ_j with $\lambda_j \geq k$, is less than μ'_k, the number of parts μ_j with $\mu_j \geq k$. Thus $\mu' \trianglerighteq \lambda'$. The converse follows by using the identity $(\lambda')' = \lambda$. ♢

Let λ be a partition of n. A λ-*tableau* is a bijection $T : [\lambda] \to [n]$. Graphically, a λ-tableau is an array of integers obtained by replacing each node in the diagram $[\lambda]$ by an integer in $[n]$ without repeats. For instance, two $(4, 2^2, 1)$-tableaux are

$$\begin{array}{llll} 1 & 2 & 3 & 4 \\ 5 & 6 & & \\ 7 & 8 & & \\ 9 & & & \end{array} \quad \text{and} \quad \begin{array}{llll} 2 & 6 & 3 & 1 \\ 9 & 4 & & \\ 8 & 5 & & \\ 7 & & & \end{array}$$

A tableau is called *row-standard* if the entries increase along the rows. For instance, the first of the above tableaux is row-standard, the second is not.

C.3. **Basic Combinatorial Lemma** Let λ and μ be partitions of n, and suppose T is a λ-tableau and T' is a μ-tableau. If for each $i \geq 1$, the entries in the ith row of T' belong to different columns of T, then $\lambda \trianglerighteq \mu$.

Proof. Imagine we can place the μ_1 numbers from the first row of T' into the diagram $[\lambda]$ such that no two numbers are in the same column. Then $[\lambda]$ must have at least μ_1 columns; that is, $\lambda_1 \geq \mu_1$. Next we insert the μ_2 numbers from the second row of T' into different columns. To have space to do this, we require that $\lambda_1 + \lambda_2 \geq \mu_1 + \mu_2$. By continuing in this way, we obtain $\lambda \trianglerighteq \mu$. ♢

The symmetric group S_n acts on the set of λ-tableaux in a natural way. Given a λ-tableau T and a permutation π of n, the composition of the functions T and π gives the λ-tableau πT. For instance, the permutation $(1264)(3)(5978)$ sends the first of the above tableaux to the second.

Let T be a λ-tableau. The *row stabilizer* $R(T)$ of T is the subgroup of S_n that keeps the rows of T fixed setwise; that is,

$$R(T) = \{\sigma \in S_n \mid \forall i : i \text{ and } \sigma(i) \text{ belong to same row of } T\}.$$

The *column stabilizer* $C(T)$ of T is similarly defined. For instance, the tableau

$$T = \begin{array}{l} 2\,6\,3\,1 \\ 9\,4 \\ 8\,5 \\ 7 \end{array}$$

has the row stabilizer $R(T) = S_{\{1,2,3,6\}} \oplus S_{\{4,9\}} \oplus S_{\{5,8\}} \oplus S_{\{7\}}$ and the column stabilizer $C(T) = S_{\{2,7,8,9\}} \oplus S_{\{4,5,6\}} \oplus S_{\{3\}} \oplus S_{\{1\}}$.

The row and column stabilizers are Young subgroups of S_n. Generally, given a partition $X = \{X_1, \ldots, X_k\}$ of the set $[n]$ into disjoint non-empty subsets. The *Young subgroup* S_X is given by the product group

$$S_X = S_{X_1} \oplus \ldots \oplus S_{X_k},$$

where S_{X_i} is a subgroup of S_n leaving all elements outside of X_i element-wise fixed. In particular, each partition λ of n has an associated *canonical* Young subgroup

$$S_\lambda = S_{\{1,\ldots,\lambda_1\}} \oplus \ldots \oplus S_{\{\lambda_1+1,\ldots,\lambda_1+\lambda_2\}} \oplus \ldots.$$

Lemma C.4. *Let T be a λ-tableau.*

- *For each permutation π of n,*

$$C(\pi T) = \pi C(T)\pi^{-1} \quad \textit{and} \quad R(\pi T) = \pi R(T)\pi^{-1}.$$

- $C(T) \cap R(T) = 1$.
- *The row and colum stabilizer of T are subgroups of S_n that are conjugate of S_λ and $S_{\lambda'}$, respectively.*

Define an equivalence relation on the set of all λ-tableaux. Two λ-tableaux T_1 and T_2 are called equivalent, if there is some permutation $\pi \in R(T_1)$ such that $\pi T_1 = T_2$. The equivalence classes are called *tabloids*, and the equivalence class containing the tableau T is denoted as $\{T\}$. A tabloid can be considered as a tableau with totally ordered row entries. For instance, the $(3,2)$-tabloids are

$$\begin{array}{l} \underline{1\,2\,3} \\ 4\,5 \end{array}, \begin{array}{l} \underline{1\,2\,4} \\ 3\,5 \end{array}, \begin{array}{l} \underline{1\,2\,5} \\ 3\,4 \end{array}, \begin{array}{l} \underline{1\,3\,4} \\ 2\,5 \end{array}, \begin{array}{l} \underline{1\,3\,5} \\ 2\,4 \end{array},$$

$$\begin{array}{l} \underline{1\,4\,5} \\ 2\,3 \end{array}, \begin{array}{l} \underline{2\,3\,4} \\ 1\,5 \end{array}, \begin{array}{l} \underline{2\,3\,5} \\ 1\,4 \end{array}, \begin{array}{l} \underline{2\,4\,5} \\ 1\,3 \end{array}, \begin{array}{l} \underline{3\,4\,5} \\ 1\,2 \end{array}.$$

The group S_n acts on the set of λ-tabloids by

$$\pi\{T\} = \{\pi T\}, \quad \pi \in S_n,\ T\ \lambda\text{-tableau}.$$

This action is well-defined, since $\{T_1\} = \{T_2\}$ implies $T_2 = \sigma T_1$ for some permutation $\sigma \in R(T_1)$. By Lemma C.4, for each permutation $\pi \in S_n$, $\pi\sigma\pi^{-1} \in R(\pi T_1)$ and thus $\{\pi T_1\} = \{(\pi\sigma\pi^{-1})(\pi T_1)\} = \{\pi T_2\}$.

C.3 Permutation Modules

Let $\mathbb{K}$ be a field. For each partition λ of n, we consider the vector space M^λ over $\mathbb{K}$ whose basis elements are the λ-tabloids; that is,

$$M^\lambda = \bigoplus_{\{T\}} \mathbb{K}\{T\}.$$

The action of the symmetric group S_n defined on tabloids can be linearly extended on the space M^λ as follows,

$$\left(\sum_\pi k_\pi \pi\right) \cdot \left(\sum_{\{T\}} k'_{\{T\}}\{T\}\right) = \sum_\pi \sum_{\{T\}} k_\pi k'_{\{T\}} \pi\{T\}.$$

This action turns M^λ into a $\mathbb{K}S_n$-module.

Let T be a λ-tableau, whose row stabilizer is S_λ. Take a set of left coset representatives $\sigma_1, \ldots, \sigma_l$ of the subgroup S_λ in S_n. This gives the decomposition

$$S_n = \sigma_1 S_\lambda \cup \ldots \cup \sigma_l S_\lambda.$$

Form the l elements

$$\sigma_i\{T\}, \quad 1 \le i \le l.$$

Claim that these elements are linearly independent over $\mathbb{K}$. Indeed, given a linear combination $\sum_i k_i\sigma_i\{T\} = 0$, $k_i \in \mathbb{K}$, $1 \le i \le l$, each term $k_i\sigma_i\{T\}$ must be zero, since the tabloids involved are pairwisely different. Moreover, claim that these elements form a basis of M^λ. Indeed, it suffices to show that for any element $\pi \in S_n$, $\pi\sigma_i\{T\}$ is a linear combination of the elements of the set. For this, observe that $\pi\sigma_i = \sigma_j\tau$ for some $\tau \in S_\lambda$ and some unique σ_j so that $\pi\sigma_i\{T\} = \sigma_j\tau\{T\} = \sigma_j\{T\}$. This proves the claim.

Proposition C.5. *The space M^λ is a cyclic $\mathbb{K}S_n$-module generated by any one λ-tabloid, and the dimension of M^λ is*

$$\dim_{\mathbb{K}} M^\lambda = \binom{n}{\lambda_1, \lambda_2, \ldots}.$$

Proof. In view of the above basis of M^λ, we have $M^\lambda = \mathbb{K}S_n\{T\}$ for any λ-tabloid $\{T\}$; that is, M^λ is cyclic. The dimension of M^λ equals the number of coset representatives of S_λ in S_n. $\diamondsuit$

For instance, the space $M^{(3,2)}$ has dimension 10; the (3,2)-tabloids were given in the previous Section.

A bilinear form on the space M^λ can be defined by setting

$$\langle \{T_1\}, \{T_2\} \rangle = \begin{cases} 1 \ \{T_1\} = \{T_2\}, \\ 0 \text{ otherwise.} \end{cases}$$

This bilinear form is symmetric and makes the set of λ-tabloids an orthonormal basis of M^λ. This form is *non-singular* (i.e., for each non-zero element m in M there is an element m' in M with $\langle m, m' \rangle \neq 0$), and *associative* (i.e., for all tabloids $\{T_1\}, \{T_2\}$ and permutations π of n, $\langle \pi\{T_1\}, \pi\{T_2\} \rangle = \langle \{T_1\}, \{T_2\} \rangle$). It follows that the group S_n operates as a group of orthogonal transformations on the space M^λ.

C.4 Specht Modules

Let T be a λ-tableau. The *signed colum sum* of T is an element of the group algebra $\mathbb{K}S_n$ that is obtained by summing the elements in the column stabilizer of T, attaching the signature to each permutation,

$$\kappa_T = \sum_{\sigma \in C(T)} \operatorname{sgn}(\sigma)\sigma.$$

The *polytabloid* associated with the tableau T is a linear combination of λ-tabloids that is obtained by multiplying the tabloid $\{T\}$ with the signed column sum,

$$e_T = \kappa_T\{T\} = \sum_{\sigma \in C(T)} \operatorname{sgn}(\sigma)\sigma\{T\}.$$

The polytabloid e_T depends on the tableau T not just on the tabloid $\{T\}$.

Example C.6. Take the tableau $T = \begin{matrix} 1\,3\,5 \\ 2\,4 \end{matrix}$. The corresponding signed column sum is

$$\kappa_T = (1 - (12))(1 - (34)) = (1) - (12) - (34) + (12)(34)$$

and the associated polytabloid is

$$e_T = \overline{\underline{\begin{matrix} 1\,3\,5 \\ 2\,4 \end{matrix}}} - \overline{\underline{\begin{matrix} 2\,3\,5 \\ 1\,4 \end{matrix}}} - \overline{\underline{\begin{matrix} 1\,4\,5 \\ 2\,3 \end{matrix}}} + \overline{\underline{\begin{matrix} 2\,4\,5 \\ 1\,3 \end{matrix}}}.$$

♢

The practical way of writing down the polytabloid e_T is to permute the numbers in the columns of the tableau T in all possible ways, attaching the signature to each permutation, and then permuting the positions of T accordingly.

The *Specht module* for the partition λ is the submodule S^λ of M^λ spanned as a vector space by the polytabloids,

$$S^\lambda = \bigoplus_{T \ \lambda-\text{tableau}} \mathbb{K}e_T.$$

Proposition C.7. *The Specht module S^λ is a cyclic $\mathbb{K}S_n$-module generated by any one polytabloid.*

Proof. Let T be a λ-tableau. For each permutation π of n, we have

$$\begin{aligned} e_{\pi T} &= \sum_{\sigma \in C(\pi T)} \operatorname{sgn}(\sigma)\sigma\{\pi T\} \\ &= \sum_{\sigma \in C(T)} \operatorname{sgn}(\pi\sigma\pi^{-1})\pi\sigma\pi^{-1}\{\pi T\} \\ &= \pi \sum_{\sigma \in C(T)} \operatorname{sgn}(\sigma)\sigma\{T\} \\ &= \pi e_T. \end{aligned}$$

It follows that S^λ is a cyclic module. ♢

Lemma C.8. *Let λ and μ be partitions of n. If T is a λ-tableau and T' is a μ-tableau such that $\kappa_T\{T'\} = 0$ then $\lambda \trianglerighteq \mu$. In particular, if $\lambda = \mu$ then $\kappa_T\{T'\} = \pm\kappa_T\{T\} = \pm e_T$.*

Proof. Let i and j be two numbers in the same row of T'. Then we have

$$[1 - (ij)]\{T'\} = \{T'\} - (ij)\{T'\} = 0.$$

Suppose i and j are in the same column of T. Then we can find coset representatives $\sigma_1, \ldots, \sigma_k$ for the subgroup $U = \{1, (ij)\}$ of the column stabilizer of T such that

$$\kappa_T = [\operatorname{sgn}(\sigma_1)\sigma_1 + \ldots + \operatorname{sgn}(\sigma_k)\sigma_k][1 - (ij)].$$

Then it follows that $\kappa_T\{T'\} = 0$ contradicting the hypothesis. Thus we have shown that the numbers in the same row of T' belong to different columns of T. Hence, the basic combinatorial lemma gives $\lambda \trianglerighteq \mu$.

If $\lambda = \mu$, then by construction $\{T'\}$ is one of the tabloids involved in $\kappa_T\{T\}$. Thus $\{T'\} = \sigma\{T\}$ for some column permutation σ in $C(T)$. It follows that $\kappa_T\{T'\} = \kappa_T\sigma\{T\} = \pm\kappa_T\{T\}$. ♢

Corollary C.9. *If m is an element of M^λ and T is a λ-tableau, then $\kappa_T m$ is a multiple of e_T.*

Proof. By Lemma C.8, for each λ-tabloid $\{T'\}$, $\kappa_T\{T'\}$ is a multiple of e_T. But m is a linear combination of λ-tabloids and thus m is a multiple of e_T. ♢

Let $m, m' \in M^\lambda$ and T be a μ-tableau. By the properties of the bilinear form on M^λ, we have

$$\begin{aligned} \langle \kappa_T m, m' \rangle &= \sum_{\sigma \in C(T)} \langle \operatorname{sgn}(\sigma)\sigma m, m' \rangle \\ &= \sum_{\sigma \in C(T)} \langle m, \operatorname{sgn}(\sigma)\sigma^{-1} m' \rangle \\ &= \sum_{\sigma \in C(T)} \langle m, \operatorname{sgn}(\sigma)\sigma m' \rangle \\ &= \sum_{\sigma \in C(T)} \langle m, \kappa_T m' \rangle. \end{aligned} \tag{C.4}$$

C.10. **Submodule Theorem** If U is a $\mathbb{K}S_n$-submodule of M^λ, then either $U \supseteq S^\lambda$ or $U \subseteq {S^\lambda}^\perp$.

Proof. Let $u \in U$ and let T be a λ-tableau. By Corollary C.9, $\kappa_T u$ is a multiple of e_T. If $\kappa_T u \neq 0$, then e_T is an element of U. But S^λ is generated by e_T and thus $U \supseteq S^\lambda$. Otherwise, we have $\kappa_T u = 0$ for each $u \in U$ and λ-tableau T. Then we have

$$0 = \langle \kappa_T u, \{T\} \rangle = \langle u, \kappa_T \{T\} \rangle = \langle u, e_T \rangle.$$

It follows that u belongs to ${S^\lambda}^\perp$ and thus $U \subseteq {S^\lambda}^\perp$. ♢

Theorem C.11. *The Specht module S^λ over $\mathbb{Q}$ is irreducible; that is, there is no proper $\mathbb{Q}S_n$-submodule of S^λ.*

Proof. By the submodule theorem, any submodule U of S^λ is either S^λ or is contained in $S^\lambda \cap {S^\lambda}^\perp$. But $S^\lambda \cap {S^\lambda}^\perp = 0$ over $\mathbb{Q}$ and thus the result follows. ♢

Lemma C.12. *If $\phi : M^\lambda \to M^\mu$ is a $\mathbb{K}S_n$-homomorphism such that $S^\lambda \not\subset \ker \phi$, then $\lambda \trianglerighteq \mu$. In particular, if $\lambda = \mu$, then the restriction of ϕ to S^λ amounts to multiplication by a constant.*

Proof. Let T be a λ-tableau. By hypothesis, $e_T \notin \ker \phi$ and thus

$$0 \neq \phi(e_T) = \kappa_T \phi(\{T\}).$$

Thus $\phi(e_T)$ is κ_T times a linear combination of μ-tabloids. By Lemma C.8, $\lambda \trianglerighteq \mu$. In particular, if $\lambda = \mu$ then by Lemma C.8, $\phi(e_T)$ is a multiple of e_T. ♢

Corollary C.13. *If $\phi : S^\lambda \to M^\mu$ is a nonzero $\mathbb{Q}S_n$-homomorphism, then $\lambda \trianglerighteq \mu$. In particular, if $\lambda = \mu$, then ϕ amounts to multiplication by a constant.*

Proof. Over the field $\mathbb{Q}$, we have

$$S^\lambda \cap {S^\lambda}^\perp = 0 \quad \text{and} \quad M^\lambda = S^\lambda \oplus {S^\lambda}^\perp. \tag{C.5}$$

Any homomorphism $\phi : S^\lambda \to M^\mu$ can thus be extended to a homomorphism $\phi : M^\lambda \to M^\mu$ by letting it be zero on ${S^\lambda}^\perp$. Now Lemma C.12 yields the result. ♢

Theorem C.14. *The Specht modules over $\mathbb{Q}$ provide all irreducible $\mathbb{Q}S_n$-modules.*

Proof. If two Specht modules S^λ and S^μ are isomorpic, they give rise to a nonzero homomorphism $\phi : S^\lambda \to S^\mu$. Then by Corollary C.13, $\lambda \trianglerighteq \mu$. Similarly, $\mu \trianglerighteq \lambda$ and thus $\lambda = \mu$.

For any finite group G, the number of irreducible $\mathbb{Q}G$-modules equals the number of conjugacy classes of G. But, by Proposition C.1, the number of conjugacy classes of S_n equals the number of partitions of n. Thus by Theorem C.11, the Specht modules provide all irreducible $\mathbb{Q}S_n$-modules. ♢

Theorem C.15. *The permutation module M^μ is a direct sum of Specht modules S^λ where $\lambda \trianglerighteq \mu$ (possibly with repeats). In particular, the Specht module S^μ appears once in this sum.*

Proof. By Maschke's Theorem, for any finite group G, each nonzero $\mathbb{Q}G$-module M is completely reducible; that is, M can be written as a direct sum of irreducible $\mathbb{Q}G$-modules. Thus we have

$$M^\mu = \bigoplus_\lambda S^\lambda.$$

Thus for each direct summand S^λ of M^μ, there is a nonzero $\mathbb{Q}S_n$-homomorphism $\phi : S^\lambda \to M^\mu$. By Corollary C.13, it follows that $\lambda \trianglerighteq \mu$. Moreover, by Eq. (C.5), we have that S^λ appears exactly once in this decomposition. ♢

C.5 Standard Basis of Specht Modules

A λ-tableau T is called *standard* if the numbers increase along the rows and down the columns of T. A λ-tabloid $\{T\}$ is called *standard* if there is a standard tableau in its equivalence class. A λ-polytabloid e_T is called *standard* if the tableau T is standard.

Example C.16. The standard (3,2)-tableaux are

$$\begin{array}{ccccccccccc} 1\,2\,3 & & 1\,2\,4 & & 1\,2\,5 & & 1\,3\,4 & & 1\,3\,5 \\ 4\,5 & , & 3\,5 & , & 3\,4 & , & 2\,5 & , & 2\,4 \end{array}.$$

♢

Define a total ordering on the set of λ-tabloids. Let $\{T\}$ and $\{T'\}$ be λ-tabloids. Write $\{T\} < \{T'\}$ if there is a number i such that when j is a number with $j > i$, then j is in the same row of $\{T\}$ and $\{T'\}$, and the number i is in a higher row of $\{T\}$ than $\{T'\}$. For instance, we have

$$\overline{\underline{3\,4\,5}\;\underline{1\,2}} < \overline{\underline{2\,4\,5}\;\underline{1\,3}} < \overline{\underline{1\,4\,5}\;\underline{2\,3}} < \overline{\underline{2\,3\,5}\;\underline{1\,4}} < \overline{\underline{1\,3\,5}\;\underline{2\,4}}.$$

Lemma C.17. *Let $m_1, \ldots, m_k$ be elements of M^λ such that the tabloid $\{T_i\}$ is the $<$-maximal element involved in m_i, $1 \le i \le k$. If the tabloids $\{T_i\}$ are all different, then $m_1, \ldots, m_k$ are linearly independent.*

Proof. Assume that $\{T_1\} < \ldots < \{T_k\}$. If $a_1 m_1 + \ldots + a_k m_k = 0$ with $a_1, \ldots, a_k \in \mathbb{K}$ and $a_{j+1} = \ldots = a_k = 0$, $1 \le j \le m$, then $a_j = 0$, since $\{T_j\}$ is involved in m_j but not in any m_i with $i < j$. Therefore, $a_1 = \ldots = a_k = 0$. ♢

Theorem C.18. *The set of standard λ-tableaux is a basis for the Specht module S^λ.*

Proof. Let T be a standard λ-tableau. By definition of the total order on tabloids, it follows that for each tabloid $\{T'\}$ involved in e_T, we have $\{T'\} \le \{T\}$. Thus by Lemma C.17, the set of standard λ-tableaux is linearly independent.

Let T be a λ-tableau. Write $[T] = \{\sigma T \mid \sigma \in C(T)\}$ for the column equivalence class of T. The column equivalence classes are totally ordered in the same way as the row equivalence classes.

Suppose T is a non-standard λ-tableau. For each column permutation $\sigma \in C(T)$, we have $\sigma e_T = \mathrm{sgn}(\sigma) e_T$ and thus we may assume that the entries of T are increasing down the columns. Unless T is standard, there are two adjacent columns in T of the form

$$\begin{array}{cc} a_1 & b_1 \\ \vdots & \vdots \\ a_u & b_u \\ \vdots & \vdots \\ a_v & b_v \\ \vdots & \\ a_w & \end{array}$$

where $a_1 < \ldots < a_w$, $b_1 < \ldots < b_v$, and $a_u > b_u$.

Take $X = \{a_u, \ldots, a_w\}$ and $Y = \{b_1, \ldots, b_u\}$. Let $\sigma_1, \ldots, \sigma_k$ be the coset representatives for the subgroup $S_X \times S_Y$ in the group $S_{X \cup Y}$. The element $G_{X,Y} = \sum_j \operatorname{sgn}(\sigma_j)\sigma_j$ in the group algebra $\mathbb{K}S_n$ is called *Garnier element*. We have $G_{X,Y}e_T = 0$. But $G_{X,Y}e_T = \sum_j \operatorname{sgn}(\sigma_j)e_{\sigma_j T}$ and $[\sigma_j T] < [T]$ for $\sigma_j \neq 1$. Thus e_T is a linear combination of polytabloids $e_{\sigma_j T}$, $\sigma_j \neq 1$. By induction, each polytabloid $e_{\sigma_j T}$, $\sigma_j \neq 1$, can be written as a linear combination of standard polytabloids. Hence, the polytabloid e_T has the same property. ♢

Example C.19. To illustrate the proof, consider the tableau $T = \begin{array}{ll} 1 & 2 \\ 4 & 3 \\ 5 & \end{array}$. We have $X = \{4, 5\}$ and $Y = \{2, 3\}$. This gives the Garnier element

$$G_{X,Y} = 1 - (34) + (354) + (234) - (2354) + (24)(35).$$

♢

The theorem implies that the dimension of the Specht module S^λ is independent of the ground field, and equals the number of standard λ-tableaux. The proof shows that any λ-polytabloid can be written as an integral linear combination of standard λ-polytabloids.

Example C.20. By Example C.16, the Specht module $S^{(3,2)}$ has dimension 5 over any field. ♢

C.6 Young's Rule

We have seen that each permutation module M^μ is direct sum of Specht modules S^λ. By Theorem C.15, we have

$$M^\lambda = \bigoplus_{\lambda \trianglerighteq \mu} k_{\lambda,\mu} S^\mu,$$

where $k_{\lambda,\mu}$ denotes the number of irreducible submodules of M^λ that are isomorphic to S^μ. The direct sum of all submodules that are isomorphic to S^μ is called the *isotypic subspace* belonging to partition μ. This submodule is denoted by V_λ^μ. Thus, V_λ^μ is isomorphic to the $k_{\lambda,\mu}$-fold multiple of the Specht module S^μ. The multiplicities can be calculated by making use of a general result from representation theory.

Theorem C.21. *For any finite group G, the multiplicity of an irreducible $\mathbb{C}G$-module S to occur in the irreducible decomposition of a $\mathbb{C}G$-module M equals the dimension of the $\mathbb{C}$-space* $\operatorname{Hom}_{\mathbb{C}G}(S, M)$.

As $\mathbb{Q}$ is the splitting field for the group S_n, the number we seek is the dimension of the space $\operatorname{Hom}_{\mathbb{Q}S_n}(S^\lambda, M^\nu)$. A basis of the space $\operatorname{Hom}_{\mathbb{Q}S_n}(S^\lambda, M^\nu)$ can be obtained by modifying the construction of the standard basis of the Specht module.

For this, it is convenient to introduce a new copy of the permutation module. To this end, we introduce tableaux with repeated entries. For this, let λ and μ be partitions of n. A λ-tableau t has *type* μ if for each number i, the number i occurs μ_i times in t. For instance, two (4,1)-tableaux of type (3,2) are

$$\begin{array}{ll} 1 & 1 \; 2 \; 2 \\ 1 & \end{array} \quad \text{and} \quad \begin{array}{llll} 1 & 1 & 2 & 1 \\ 2 & & & \end{array}$$

The tableaux considered so far were all of type (1^n). Let $\mathcal{T}(\lambda,\mu)$ denote the set of all λ-tableaux of type μ. In the following, let T_0 be a fixed λ-tableau of type (1^n). For each tableau $t \in \mathcal{T}(\lambda,\mu)$, let $t(i)$ denote the entry in t that occurs in the same position as i in T_0. The symmetric group S_n acts on the set $\mathcal{T}(\lambda,\mu)$ by

$$(\pi t)(i) = t(\pi^{-1}(i)), \quad 1 \le i \le n,\ \pi \in S_n,\ t \in \mathcal{T}(\lambda,\mu).$$

This action is simply a place permutation. For instance, if we put

$$T_0 = \begin{array}{l} 1\,3\,4\,5 \\ 2 \end{array} \quad \text{and} \quad t = \begin{array}{l} 2\,2\,1\,1 \\ 1 \end{array}$$

then

$$(12)\,t = \begin{array}{l} 1\,2\,1\,1 \\ 2 \end{array} \quad \text{and} \quad (123)\,t = \begin{array}{l} 2\,1\,1\,1 \\ 2 \end{array}$$

The symmetric group acts transitively on the set $\mathcal{T}(\lambda,\mu)$, and there is an element whose stabilizer is the Young subgroup S_μ. Thus we may take M^μ to be the vector space spanned by the tableaux in $\mathcal{T}(\lambda,\mu)$. For instance, in view of the $(4,1)$-tableau $T_0 = \begin{array}{l} 1\,2\,3\,4 \\ 5 \end{array}$, we have the following correspondence between $(4,1)$-tabloids and $(4,1)$-tableaux of type $(3,2)$:

$$\begin{array}{l} \underline{1\,2\,3} \\ \underline{4\,5} \end{array}\ \begin{array}{l} 1\,1\,1\,2 \\ 2 \end{array}, \quad \begin{array}{l} \underline{1\,3\,4} \\ \underline{2\,5} \end{array}\ \begin{array}{l} 1\,2\,1\,1 \\ 2 \end{array}.$$

Let t_1 and t_2 be λ-tableaux. We say that t_1 and t_2 are *row equivalent* if $t_2 = \pi t_1$ for some permutation in the row stabilizer of the given λ-tableau T_0; column equivalence is similarly defined.

For each λ-tableau t of type μ, define the map φ_t by

$$\varphi_t : \kappa\{T_0\} \mapsto \kappa \sum \{t'\}, \quad \kappa \in \mathbb{K}S_n,$$

where the sum extends over all different λ-tableaux of type μ that are row equivalent to $\{t\}$. The mapping φ_t belongs to $\mathrm{Hom}_{\mathbb{K}S_n}(M^\lambda, M^\mu)$. For instance, in view of the $(4,1)$-tableau $T_0 = \begin{array}{l} 1\,3\,4\,5 \\ 2 \end{array}$ and the $(4,1)$-tableau $t = \begin{array}{l} 2\,2\,1\,1 \\ 1 \end{array}$ of type $(3,2)$,

$$\varphi_t\{T_0\} = \begin{array}{l} 2\,2\,1\,1 \\ 1 \end{array} + \begin{array}{l} 2\,1\,2\,1 \\ 1 \end{array} + \begin{array}{l} 1\,2\,2\,1 \\ 1 \end{array} + \begin{array}{l} 1\,2\,1\,2 \\ 1 \end{array} + \begin{array}{l} 1\,1\,2\,2 \\ 1 \end{array}.$$

Define $\hat{\varphi}_t$ as the restriction of φ_t to the Specht module S^λ. However, we clearly have $\kappa_{T_0} t = 0$ if and only if some column of t contains two identical numbers. Thus the homomorphism $\hat{\varphi}_t$ can sometimes be zero. To eliminate such trivial elements from the space $\mathrm{Hom}_{\mathbb{K}S_n}(S^\lambda, M^\mu)$, we consider specific tableaux.

A λ-tableau t of type μ is called *semistandard* if the numbers are nondecreasing along the rows of t and strictly increasing down the columns of t. If t is a semistandard λ-tableau of type μ, then the homomorphism $\hat{\varphi}_t$ is also called *semistandard*. For instance, there are two semistandard $(4,1)$-tableaux of type $(2^2,1)$,

$$\begin{array}{l} 1\,1\,2\,2 \\ 3 \end{array} \quad \text{and} \quad \begin{array}{l} 1\,1\,2\,3 \\ 2 \end{array}.$$

Theorem C.22. *The semistandard homomorphisms $\hat{\varphi}_t$ corresponding to the semistandard λ-tableaux t of type μ form a $\mathbb{Q}$-basis of the space* $\mathrm{Hom}_{\mathbb{Q}S_n}(S^\lambda, M^\mu)$.

Corollary C.23. *The multiplicity of the Specht module S^λ in the permutation module M^μ equals the number of semistandard λ-tableaux of type μ.*

Example C.24. The semistandard tableau of type $(3,2,2)$ are

$$\begin{array}{lll}
\begin{array}{l}1\,1\,1\,2\,2\,3\,3\end{array} & \begin{array}{l}1\,1\,1\,2\,2\,3\\3\end{array} & \begin{array}{l}1\,1\,1\,2\,3\,3\\2\end{array}\\
\begin{array}{l}1\,1\,1\,2\,2\\3\,3\end{array} & \begin{array}{l}1\,1\,1\,2\,3\\2\,3\end{array} & \begin{array}{l}1\,1\,1\,3\,3\\2\,2\end{array}\\
\begin{array}{l}1\,1\,1\,2\\2\,3\,3\end{array} & \begin{array}{l}1\,1\,1\,3\\2\,2\,3\end{array} & \begin{array}{l}1\,1\,1\,2\,3\\2\\3\end{array}\\
\begin{array}{l}1\,1\,1\,2\\2\,3\\3\end{array} & \begin{array}{l}1\,1\,1\,3\\2\,2\\3\end{array} & \begin{array}{l}1\,1\,1\\2\,2\,3\\3\end{array}\\
\begin{array}{l}1\,1\,1\\2\,2\\3\,3\end{array} & &
\end{array}$$

Thus we obtain the following decomposition of the permutation module $M^{(3,2^2)}$ into isotypic subspaces,

$$\begin{aligned} M^{(3,2^2)} &= 1\cdot S^{(7)} \oplus 2\cdot S^{(6,1)} \oplus 3\cdot S^{(5,2)} \oplus 2\cdot S^{(4,3)} \oplus 1\cdot S^{(5,1^2)} \\ &= \oplus 2\cdot S^{(4,2,1)} \oplus 1\cdot S^{(3^2,1)} \oplus 1\cdot S^{(3,2^2)}. \end{aligned}$$

♢

The semistandard λ-tableaux of type $\mu = (1^n)$ are exactly the standard λ-tableaux. But the standard λ-tableaux form a basis of the Specht module S^λ and the permutation module $M^{(1^n)}$ is isomorphic to the group algebra of the group S_n.

Corollary C.25. *The multiplicity of the Specht module S^λ in the group algebra $\mathbb{Q}S_n$ equals its dimension.*

C.7 Representations

Let G be a finite group and $\mathbb{K}$ be a field. A homomorphism of the group G into a group of $n \times n$ matrices over $\mathbb{K}$ is called a *representation of G of degree n*. This means that to each element g of G there is a matrix $\varphi(g)$ and if g and h are elements of G, then $\varphi(gh) = \varphi(g)\varphi(h)$. A representation is called *faithful* if the homomorphism is one-to-one. Frobenius posed the problem to determine all matrix representations of a finite group.

First, there is always a representation of a finite group G. For this, take $G = \{g_1, \ldots, g_n\}$ and consider the mapping φ that assigns to each group element g the linear substitution $g_i \mapsto g_i g$, $1 \leq i \leq n$. This linear substitution is represented by an $n \times n$ permutation matrix; that is, a matrix with entries 0 and 1 that has entry 1 in position (i, j) if and only if $g_i g = g_j$. Since $g_i g = g_i$ if and only if $g = 1$, the representation is faithful. The representation φ is called *regular*. At the other extreme, there is the *one-representation* ι for which $\iota(g) = 1$, for all $g \in G$.

Second, given any matrix representation of a group G we can find infinitely many. For this, let φ be a matrix representation of G of degree n and let $\boldsymbol{B}$ be an invertible $n \times n$ matrix. We can define

$$\psi(g) = \boldsymbol{B}\varphi(g)\boldsymbol{B}^{-1}, \quad g \in G.$$

Clearly ψ is a matrix representation of G, since we have

$$\begin{aligned}\psi(gh) &= \boldsymbol{B}\varphi(gh)\boldsymbol{B}^{-1} = \boldsymbol{B}\varphi(g)\varphi(h)\boldsymbol{B}^{-1} = \boldsymbol{B}\varphi(g)\boldsymbol{B}^{-1}\boldsymbol{B}\varphi(h)\boldsymbol{B}^{-1} \\ &= \psi(g)\psi(h), \quad g, h \in G.\end{aligned}$$

The new representation is obtained from the given representation by a change of basis. Representations related as φ and ψ are called *equivalent* and are regarded as essentially the same representation.

Third, let φ and ψ be matrix representations of a group G of degree m and n, respectively. Consider the matrix

$$\tau(g) = \begin{pmatrix} \varphi(g) & 0 \\ 0 & \psi(g) \end{pmatrix}$$

Clearly, τ is a matrix representation of G of degree $m + n$, since we have

$$\begin{aligned}\tau(g)\tau(h) &= \begin{pmatrix} \varphi(g) & 0 \\ 0 & \psi(g) \end{pmatrix}\begin{pmatrix} \varphi(h) & 0 \\ 0 & \psi(h) \end{pmatrix} = \begin{pmatrix} \varphi(g)\varphi(h) & 0 \\ 0 & \psi(g)\psi(h) \end{pmatrix} \\ &= \begin{pmatrix} \varphi(gh) & 0 \\ 0 & \psi(gh) \end{pmatrix} = \tau(gh), \quad g, h \in G.\end{aligned}$$

The representation τ of G is called the *direct sum* of φ and ψ, and we write $\tau = \varphi \oplus \psi$. A representation of the form $\varphi \oplus \psi$ is said to be *decomposable* with components φ and ψ. A representation that is not decomposable is called *indecomposable*.

Fourth, the property of being indecomposable depends on the field underlying the representation. For this, consider the matrix group

$$G = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix} \right\}.$$

This group is indecomposable over the rational field. To see this, observe that there is no 2×2 matrix $\boldsymbol{B}$ with rational coefficients such that

$$\boldsymbol{B}\begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix}\boldsymbol{B}^{-1} = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}, \quad a, b \in \mathbb{Q}.$$

If such a matrix would exist, then equating for the trace and determinant on both sides would give $a + b = -1$ and $ab = 1$. This would lead to the quadratic equation $a^2 + a + 1 = 0$, whose solutions are cubic roots of unity. On the other hand, over the complex field we obtain

$$\begin{pmatrix} \xi & 1 \\ \xi^2 & 1 \end{pmatrix} \begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} \xi & 1 \\ \xi^2 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} \xi & 0 \\ 0 & \xi^2 \end{pmatrix},$$

where ξ is a cubic root of unity. Thus the matrix group is decomposable over the complex field.

As a second example, consider the matrix group

$$G = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \right\},$$

whose entries lie in the field $\mathbb{Z}_2$ of characteristic 2. This matrix group G is also indecomposable; this can be proved along the same lines as in the previous example. However, unlike the previous example, the group G stays indecomposable even if the field is extended. Such a representation which remains indecomposable in any extension of the field is called *absolutely indecomposable*. Representation theory in fields of characteristic $p > 0$ is very different from the case at characteristic zero, when the order of the group is divisible by p.

Fifth, there is a weaker concept than decomposability. For this, let τ be a matrix representation of a group G such that for each group element $g \in G$,

$$\tau(g) = \begin{pmatrix} \boldsymbol{A}(g) & 0 \\ \boldsymbol{I}(g) & \boldsymbol{B}(g) \end{pmatrix},$$

where $\boldsymbol{A}(g)$ and $\boldsymbol{B}(g)$ are $s \times s$ and $t \times t$ matrices, respectively. If we define $\varphi(g) = \boldsymbol{A}(g)$ and $\psi(g) = \boldsymbol{B}(g)$ for all $g \in G$, then it is easy to see that φ and ψ are matrix representations of G of degree s and t, respectively. The representation τ and any representation equivalent to it is called *reducible*. A representation that is not reducible is called *irreducible*. The representations φ and ψ are the *constituents* of τ. If either is irreducible it is an *irreducible constituent*. A representation that is decomposible ($\boldsymbol{I}(g) = 0$) is clearly reducible. The last example above shows that a reducible representation can be indecomposable. We will see by Maschke's Theorem that this cannot happen if the characteristic of the field is zero or is not a divisor of the group order. In this case, reducible representations are decomposable.

If φ and ψ are themselves reducible, they yield constituents of smaller degree. Continuing in this way, we obtain a set of irreducible representation $\varphi_1, \ldots, \varphi_k$ as constituents of ψ. It can be shown that these representations are uniquely determined up to equivalence for each representation τ.

Sixth, representations can be derived from *representation modules*; that is, modules over group algebras. For this, let G be a finite group and $\mathbb{K}$ be a field. A *$\mathbb{K}G$-module* is a $\mathbb{K}$-vector space V on which the group operates in the same way as a scalar multiplication. Each $\mathbb{K}G$-module V gives rise to a representation of the underlying group. To this end, observe that V is a vector space of $\mathbb{K}$ and thus has a $\mathbb{K}$-basis $\{v_1, \ldots, v_k\}$. The representation corresponding to V is the mapping φ_V that assigns to each group element g the scalar matrix $\varphi_V(g) = (k_{ij}^g)$, whose entries are given by the action of g on the basis elements,

$$gv_j = \sum_{i=1}^{k} k_{ij}^g v_i, \quad 1 \le j \le k.$$

A representation module V is called *decomposable* if the corresponding representation φ_V is decomposable. Thus a decomposable representation module can be written as a direct sum of representation modules. A representation module V is called *reducible* if the associated representation φ_V is reducible;

otherwise, the module is called *irreducible*. Thus for a reducible representation module V there is a representation module that forms a proper $\mathbb{K}G$-submodule of V. Equivalently, an irreducible representation module V has only two $\mathbb{K}G$-submodules, zero module and V itself.

C.26. **Maschke's Theorem** Let G be a finite group and $\mathbb{K}$ be a field whose characteristic is zero or not a divisor of the group order. For each $\mathbb{K}G$-submodule U of a $\mathbb{K}G$-module V, there is a $\mathbb{K}G$-submodule W such that

$$V = U \oplus W.$$

Proof. Take a $\mathbb{K}$-basis $v_1, \ldots, v_r$ of V. There is a unique bilinear form ϕ on V given as

$$\phi(v_i, v_j) = \begin{cases} 1 & v_i = v_j, \\ 0 & \text{otherwise.} \end{cases}$$

A new bilinear form on V can be defined by

$$\langle u, v \rangle = \frac{1}{|G|} \sum_{g \in G} \phi(gu, gv), \quad u, v \in V.$$

This form is *G-invariant* in the sense that

$$\langle gu, gv \rangle = \langle u, v \rangle, \quad g \in G,\ u, v \in V.$$

Let U be a $\mathbb{K}G$-submodule of V. Define $U^\perp$ as the set of all elements $v \in V$ such that $\langle u, v \rangle = 0$ for all $u \in U$. Clearly, $U^\perp$ is a $\mathbb{K}$-subspace of V. For each $u \in U$, $v \in U^\perp$, and $g \in G$, we have $\langle u, gv \rangle = \langle g^{-1}u, v \rangle = 0$, since $g^{-1}u \in U$. Thus $U^\perp$ is a $\mathbb{K}G$-submodule of V. Moreover, by hypothesis on $\mathbb{K}$, for each nonzero element $u \in U$, $\langle u, u \rangle \neq 0$ $\mathbb{K}$. Hence, $U \cap U^\perp = 0$. Finally, we show that $U \oplus U^\perp = V$. ♢

Maschke's Theorem can be used to prove by induction on the dimension of $\mathbb{K}G$-modules the following

Theorem C.27. *Let G be a finite group and $\mathbb{K}$ be a field, whose characteristic is zero or not a divisor of the group order. Each $\mathbb{K}G$-module is a direct sum of irreducible $\mathbb{K}G$-submodules.*

Let H be a subgroup of a group G. Assume that $g_1, \ldots, g_k$ form a complete coset of representatives for the group H in G. Given a matrix representation φ of the group H of degree m. There is a matrix representation φ_H^G of the group G of degree $k \cdot m$ defined as

$$\varphi_H^G(g) = \begin{pmatrix} \varphi(g_1 g g_1^{-1}) & \ldots & \varphi(g_1 g g_k^{-1}) \\ \vdots & \ldots & \vdots \\ \varphi(g_k g g_1^{-1}) & \ldots & \varphi(g_k g g_k^{-1}) \end{pmatrix}, \quad g \in G,$$

where

$$\varphi(g_i g g_j^{-1}) = \begin{cases} \varphi(g_i g g_j^{-1}) & \text{if } g_i g g_j^{-1} \in H, \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

We say that the matrix representation φ_H^G of G is *induced* from the representation φ of H.

Conversely, given a matrix representation φ of G. There is a matrix representation φ_H of the subgroup H defined as

$$\varphi_H(h) = \varphi(h), \quad h \in G.$$

Finally, consider representation modules of the symmetric group. For this, let λ be a partition of n. The permutation module M^λ has the basis $\{\sigma_1\{T\}, \ldots, \sigma_l\{T\}\}$, where T is a λ-tableau with row stabilizer S_λ and the permutations $\sigma_1, \ldots, \sigma_l$ are the left coset representatives of S_λ in S_n. For each permutation $\pi \in S_n$, we have $\pi\sigma_j = \sigma_{i_j}\tau_{i_j}$ for some $\tau_{i_j} \in S_\lambda$ and a unique left coset representative σ_{i_j}, and thus $\pi\sigma_j\{T\} = \sigma_{i_j}\{T\}$. Hence, the representation of M^λ assigns to each permutation π the permutation matrix $\varphi(\pi)$ that has entries 1 in exactly the positions (i_j, j), $1 \le j \le n$. This is the reason that the representation module M^λ is referred to as a *permutation module*. For instance, the Young subgroup $S_{(2,1)} = S_{\{1,2\}} \oplus S_{\{3\}}$ has the coset representatives $\sigma_1 = (1)$, $\sigma_2 = (13)$, and $\sigma_3 = (23)$ in S_3. The permutation module M^λ can be considered as being induced from the one-dimensional $\mathbb{Q}S_\lambda$-module $\mathbb{Q}\{T\}$, where the λ-tableau T has the stabilizer S_λ.

The Specht module S^λ forms an irreducible representation module of the group S_n. A basis of this module is given by the standard λ-polytabloids $e_{T_1}, \ldots, e_{T_k}$. The representation corresponding to the Specht module S^λ is the mapping φ_λ that assigns to each group element π the matrix $\varphi_\lambda(\pi) = (k_{ij}^\pi)$, whose entries are given as

$$\pi e_{T_j} = e_{\pi T_j} = \sum_{i=1}^{k} k_{ij}^\pi e_{T_i}, \quad 1 \le j \le k.$$

As an example consider the Specht module $S^{(2,1)}$ that forms an irreducible representation module of the group S_3. The standard basis of $S^{(2,1)}$ is given by the two polytabloids

$$e_{\begin{smallmatrix}1&2\\3\end{smallmatrix}} = \frac{\overline{1\ 2}}{\underline{3}} - \frac{\overline{2\ 3}}{\underline{1}} \quad \text{and} \quad e_{\begin{smallmatrix}1&3\\2\end{smallmatrix}} = \frac{\overline{1\ 3}}{\underline{2}} - \frac{\overline{2\ 3}}{\underline{1}}.$$

For the unit element $\pi = (1)$, we have

$$(1)\, e_{\begin{smallmatrix}1&2\\3\end{smallmatrix}} = e_{\begin{smallmatrix}1&2\\3\end{smallmatrix}} \quad \text{and} \quad (1)\, e_{\begin{smallmatrix}1&3\\2\end{smallmatrix}} = e_{\begin{smallmatrix}1&3\\2\end{smallmatrix}}.$$

Thus, for the irreducible representation φ_λ,

$$\varphi_\lambda((1)) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

For the transposition $\pi = (12)$, we have

$$\begin{aligned}(12)\, e_{\begin{smallmatrix}1&2\\3\end{smallmatrix}} &= e_{\overset{(12)}{\begin{smallmatrix}1&2\\3\end{smallmatrix}}} = e_{\begin{smallmatrix}2&1\\3\end{smallmatrix}} = \frac{\overline{1\ 2}}{\underline{3}} - \frac{\overline{1\ 3}}{\underline{2}} \\ &= e_{\begin{smallmatrix}1&2\\3\end{smallmatrix}} - e_{\begin{smallmatrix}1&3\\2\end{smallmatrix}}\end{aligned}$$

and

$$(12)\,e_{\substack{1\,3\\2}} = e_{\substack{(12)1\,3\\2}} = e_{\substack{2\,3\\1}} = \overline{\underline{\underline{2\,3}}}{\underline{1}} - \overline{\underline{\underline{1\,3}}}{\underline{2}}$$
$$= -e_{\substack{1\,3\\2}}.$$

Thus, for the irreducible representation φ_λ,

$$\varphi_\lambda((12)) = \begin{pmatrix} 1 & 0 \\ -1 & -1 \end{pmatrix}.$$

Moreover, for the permutation $\pi = (123)$, we have

$$(123)\,e_{\substack{1\,2\\3}} = e_{\substack{(123)1\,2\\3}} = e_{\substack{2\,3\\1}} = \overline{\underline{\underline{2\,3}}}{\underline{1}} - \overline{\underline{\underline{1\,3}}}{\underline{2}}$$
$$= -e_{\substack{1\,3\\2}}$$

and

$$(123)\,e_{\substack{1\,3\\2}} = e_{\substack{(123)1\,3\\2}} = e_{\substack{2\,1\\3}} = \overline{\underline{\underline{2\,1}}}{\underline{3}} - \overline{\underline{\underline{3\,1}}}{\underline{2}}$$
$$= e_{\substack{1\,2\\3}} - e_{\substack{1\,3\\2}}.$$

Thus, for the irreducible representation φ_λ,

$$\varphi_\lambda((123)) = \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}.$$

C.8 Characters

The matrices representing a group allow the introduction of numerical functions on the group. These functions are called characters and play a vital role in the theory. Let $\boldsymbol{A} = (a_{ij})$ be an $n \times n$ matrix over a field $\mathbb{K}$. The *trace* of $\boldsymbol{A}$ is the sum of diagonal matrix entries,

$$\operatorname{tr} A = \sum_{i=1}^{n} a_{ii}.$$

If $\boldsymbol{B}$ is another $n \times n$ matrix over $\mathbb{K}$, then a direct calculation shows that

$$\operatorname{tr} \boldsymbol{AB} = \operatorname{tr} \boldsymbol{BA}.$$

Thus, if $\boldsymbol{B}$ is nonsingular, then

$$\operatorname{tr} \boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{B} = \operatorname{tr} \boldsymbol{A}.$$

The *character* of the representation φ of a finite group G is the function χ_φ on G to the field of representation $\mathbb{K}$ given by

$$\chi_\varphi(g) = \operatorname{tr} \varphi(g), \quad g \in G.$$

If two representations φ and ψ of a group G are equivalent, then there is a nonsingular matrix $\boldsymbol{B}$ such that $\psi(g) = \boldsymbol{B}\varphi(g)\boldsymbol{B}^{-1}$ for each element $g \in G$. But then

$$\chi_\psi(g) = \operatorname{tr} \psi(g) = \operatorname{tr} \boldsymbol{B}\varphi(g)\boldsymbol{B}^{-1} = \operatorname{tr} \varphi(g) = \chi_\varphi(g), \quad g \in G.$$

Thus, equivalent representations have the same character. If τ is a reducible representation, then for some nonsingular matrix $\boldsymbol{B}$,

$$\boldsymbol{B}\tau(g)\boldsymbol{B}^{-1} = \begin{pmatrix} \varphi(g) & 0 \\ \boldsymbol{I}(g) & \psi(g) \end{pmatrix}, \quad g \in G.$$

By taking traces on both sides, we obtain

$$\chi_\tau(g) = \chi_\varphi(g) + \chi_\psi(g), \quad g \in G.$$

Thus the character of each reducible representation is the sum of characters of its constituents. By Theorem C.27, if G is a finite group and $\mathbb{K}$ is a field, whose characteristic is zero or not a divisor of the group order, then each character of a representation of G is a sum of *irreducible characters*; that is, characters of irreducible representations.

Let φ be a representation of a finite group G, and let g and h be two conjugate elements of G; that is, $h = tgt^{-1}$ for some group element $t \in G$. We have

$$\chi_\varphi(h) = \chi_\varphi(tgt^{-1}) = \operatorname{tr} \varphi(tgt^{-1}) = \operatorname{tr} \varphi(t)\varphi(g)\varphi(t)^{-1} = \operatorname{tr} \varphi(g) = \chi_\varphi(g).$$

It follows that characters have the same value for elements of the same conjugacy class. For this reason, characters are *class functions*.

The one-representation ι of a group G equals its character which is denoted by 1_G and called the *trivial character* of G.

The *character table* of a finite group G is a matrix with columns indexed by conjugacy classes of G and rows indexed by inequivalent irreducible representations of G. The entry of the character table corresponding to irreducible representation φ and conjugacy class C is $\chi_\varphi(g)$ for some $g \in C$; that is, the value of the character of φ on the conjugacy class. One row and column of the character table can be directly filled. For this, let $C_1 = \{1\}$ be the conjugacy class of the unit element 1 of the group G. For each representation φ, the matrix $\varphi(1)$ at the unit element is the identity matrix whose size is given by the dimension of the corresponding representation module. Thus, the trace $\chi(1)$ of the identity element equals the dimension of the representation module. Moreover, let V_1 be the trivial $\mathbb{K}G$-module corresponding to the one-representation; that is, $gv = v$ for each $v \in V_1$. The character of the one-representation is the trivial character ϵ given by $\epsilon(g) = 1$ for each group element g in G. Thus the character table looks as follows,

	C_1	C_2	$\dots$	C_r
V_1	1	1		1
V_2	$\dim V_2$			
$\vdots$	$\vdots$			
V_s	$\dim V_n$			

For instance, the group S_3 has three conjugacy classes, $C_1 = \{(1)\}$, $C_2 = \{(12), (13), (23)\}$, and $C_3 = \{(123), (132)\}$. Moreover, there are three Specht modules corresponding to the partitions of 3, namely, $S^{(3)}$, $S^{(2,1)}$, and $S^{(1^3)}$. The corresponding character table is

	C_1	C_2	C_3
$S^{(3)}$	1	1	1
$S^{(2,1)}$	2	0	−1
$S^{(1^3)}$	1	−1	1

The entries in the second row were calculated in the previous section.

Let H be a subgroup of a group G. Assume that $g_1, \ldots, g_k$ form a complete coset of representatives for the group H in G. Given a matrix representation φ_H^G of G that is induced from a representation φ of H. The character of the induced representation φ_H^G can be given in terms of the character χ_H of the representation φ of H as follows,

$$\begin{aligned}\chi_H^G(g) &= \sum_{i=1}^{k} \chi'_H(g_i g g_i^{-1}), \quad g \in G,\\ &= \frac{1}{|H|} \sum_{i=1}^{k} \sum_{h \in H} \chi'_H(h g_i g g_i^{-1} h^{-1})\\ &= \frac{1}{|H|} \sum_{x \in G} \chi'_H(x g x^{-1}),\end{aligned}$$

where

$$\chi'_H(g) = \begin{cases} \chi_H(g) \text{ if } g \in H.\\ 0 \qquad \text{otherwise.}\end{cases}$$

Notice that the second equation makes use of the identity $\chi'_H(g_i g g_i^{-1}) = \chi'_H(h g_i g g_i^{-1} h^{-1})$ since $g_i g g_i^{-1} \in H$ if and only if $h g_i g g_i^{-1} h^{-1} \in H$, and the third equation uses the fact that each element in G can be uniquely written in the form $h g_i$ for some $h \in H$ and $1 \leq i \leq k$.

C.9 Characters of the Symmetric Group

We provide a procedure that allows to determine the character table of the symmetric group. For this, we define a bilinear form on the set of characters of a linear group G. To this end, we assume that the field $\mathbb{K}$ is algebraically closed with characteristic 0 or characteristic $p > 0$ such that the group order is not divisible by p. Let U and V be irreducible representation modules of G with characters χ_U and χ_V. We put

$$\langle \chi_U, \chi_V \rangle = \begin{cases} 1, U \text{ and } V \text{ are equivalent,}\\ 0, \text{ otherwise.}\end{cases}$$

Since by hypothesis, each representation module is a direct sum of irreducible representation modules, we can linearly extend the form to obtain a bilinear form on the set of characters of G.

C.28. **Frobenius Reciprocity Theorem** Let φ and ψ be absolutely irreducible representations of a group G and a subgroup H, respectively. If χ and χ' denote the respective characters corresponding to φ and ψ, then

$$\langle \chi, \chi'^G \rangle = \langle \chi_H, \chi' \rangle.$$

The theorem says that the representation ψ^G induced by ψ contains the irreducible representation φ with the same multiplicity as the representation φ_H restricted to H contains the irreducible representation ψ. The following calculations will make use of this theorem.

Let μ be a partition of n. By Theorem C.15, the permutation module M^μ decomposes into a direct sum of Specht modules

$$M^\lambda = \bigoplus_{\lambda \trianglerighteq \mu} k_{\lambda,\mu} S^\mu,$$

where $k_{\lambda,\mu} \geq 0$ denotes the multiplicity with which the module S^λ occurs in the module M^μ. In particular, we have $k_{\lambda,\lambda} = 1$. Let χ_λ denote the character of the Specht module S^λ and let $1_\lambda \uparrow S_n$ denote the character of the permutation module M^λ. It follows that

$$\langle 1_\lambda \uparrow S_n, \chi_\mu \rangle = \begin{cases} k_{\lambda,\mu}, & \lambda \trianglerighteq \mu, \\ 0, & \text{otherwise.} \end{cases}$$

Consider the matrix $\boldsymbol{K} = (k_{\lambda,\mu})$. Since the dominance order can be embedded into the dictionary order, the matrix $\boldsymbol{K}$ is a lower triangular matrix. Consequently, the matrix $\boldsymbol{B} = (b_{\lambda,\mu})$ defined by

$$b_{\lambda,\mu} = |S_\mu| \langle \chi_\lambda, 1_\mu \uparrow S_n \rangle$$

is an upper triangular matrix. In particular, we have

$$b_{\lambda,\lambda} = |S_\lambda| \langle \chi_\lambda, 1_\lambda \uparrow S_n \rangle = |S_\lambda| = \prod_i \lambda_i!.$$

Let C_μ denote the conjugacy class of S_n corresponding to the partition μ, and let $\boldsymbol{A} = (a_{\lambda,\mu})$ be the matrix given by

$$a_{\lambda,\mu} = |S_\lambda \cap C_\mu|.$$

In particular, we have

$$a_{\lambda,\lambda} = |S_\lambda \cap C_\lambda| = \prod_i (\lambda_i - 1)!,$$

since all elements of the conjugacy class C_λ can be obtained from the permutation $\pi = (1, \ldots, \lambda_1)(\lambda_1 + 1, \ldots, \lambda_1 + \lambda_2) \ldots$ Once the matrix $\boldsymbol{A}$ is known, the character table $\boldsymbol{C} = (c_{\lambda,\mu})$ of S_n can be calculated by straightforward matrix manipulation. To see this, first note that

$$\begin{aligned} \sum_\mu c_{\lambda,\mu} a_{\nu,\mu} &= \sum_\mu \chi_\lambda(C_\mu) \cdot |S_\nu \cap C_\mu| \\ &= |S_\nu| \cdot \langle \chi_\lambda \downarrow S_\nu, 1_\nu \rangle \\ &= |S_\nu| \cdot \langle \chi_\lambda, 1_\nu \uparrow S_n \rangle \\ &= b_{\lambda,\nu}. \end{aligned}$$

Therefore, $\boldsymbol{B} = \boldsymbol{C}\boldsymbol{A}^T$. Second, we have

$$\begin{aligned}\sum_{\mu} b_{\mu,\lambda} b_{\mu,\nu} &= |S_\lambda| \cdot |S_\nu| \cdot \langle 1_\lambda \uparrow S_n, 1_\nu \uparrow S_n \rangle \\ &= |S_\lambda| \cdot |S_\nu| \cdot \langle 1_\lambda \uparrow S_n \downarrow S_\nu, 1_\nu \rangle \\ &= |S_\lambda| \cdot |S_\nu| \cdot \langle 1_\lambda \uparrow S_n \downarrow S_\nu, 1_\nu \rangle \\ &= |S_\lambda| \cdot \sum_{\mu} 1_\lambda \uparrow S_n(g_\mu) \cdot |S_\nu \cap C_\mu| \\ &= \sum_{\mu} [S_n : S_\lambda] \cdot |S_\lambda \cap C_\mu| \cdot |S_\nu \cap C_\mu| \\ &= \sum_{\mu} [S_n : S_\lambda] \cdot a_{\lambda,\mu} \cdot a_{\nu,\mu}.\end{aligned}$$

Thus if the matrix $\boldsymbol{A}$ is known, the matrix $\boldsymbol{B}$ can be calculated. Moreover, the matrix $\boldsymbol{A}$ is invertible and hence we obtain the character table of S_n as

$$\boldsymbol{C} = \boldsymbol{B}\boldsymbol{A}^{T^{-1}}.$$

For instance, in case of $n = 5$, we have

$$\boldsymbol{K} = \begin{array}{c} \\ [5] \\ [4][1] \\ [3][2] \\ [3][1]^2 \\ [2]^2[1] \\ [2][1]^3 \\ [1]^5 \end{array} \begin{array}{c} \begin{array}{ccccccc} (5) & (4,1) & (3,2) & (3,1^2) & (2^2,1) & (2,1^3) & (1^5) \end{array} \\ \begin{pmatrix} 1 & & & & & & \\ 1 & 1 & & & & & \\ 1 & 1 & 1 & & & & \\ 1 & 2 & 1 & 1 & & & \\ 1 & 2 & 2 & 1 & 1 & & \\ 1 & 3 & 3 & 3 & 2 & 1 & \\ 1 & 4 & 5 & 6 & 5 & 4 & 1 \end{pmatrix} \end{array}.$$

The decompositions like $M^{(4,1)} = S^{(5)} \oplus S^{(4,1)}$ and $M^{(3,2)} = S^{(5)} \oplus S^{(4,1)} \oplus S^{(3,2)}$ can be directly obtained from Corollary C.23. But Young's ruel shows how to directly evaluate the matrix $\boldsymbol{K}$. Moreover, we have

$$\boldsymbol{A} = \begin{array}{c} \\ (5) \\ (4,1) \\ (3,2) \\ (3,1^2) \\ (2^2,1) \\ (2,1^3) \\ (1^5) \end{array} \begin{array}{c} \begin{array}{ccccccc} (5) & (4,1) & (3,2) & (3,1^2) & (2^2,1) & (2,1^3) & (1^5) \end{array} \\ \begin{pmatrix} 24 & 30 & 20 & 20 & 15 & 10 & 1 \\ & 6 & 0 & 8 & 3 & 6 & 1 \\ & & 2 & 2 & 3 & 4 & 1 \\ & & & 2 & 0 & 3 & 1 \\ & & & & 1 & 2 & 1 \\ & & & & & 1 & 1 \\ & & & & & & 1 \end{pmatrix} \end{array},$$

$$\boldsymbol{B} = \begin{array}{c} \\ (5) \\ (4)(1) \\ (3)(2) \\ (3)(1)^2 \\ (2)^2(1) \\ (2)(1)^3 \\ (1)^5 \end{array} \begin{array}{c} \begin{array}{ccccccc} (5) & (4,1) & (3,2) & (3,1^2) & (2^2,1) & (2,1^3) & (1^5) \end{array} \\ \begin{pmatrix} 120 & 24 & 12 & 6 & 4 & 2 & 1 \\ & 24 & 12 & 12 & 8 & 6 & 4 \\ & & 12 & 6 & 8 & 6 & 5 \\ & & & 6 & 4 & 6 & 6 \\ & & & & 4 & 4 & 5 \\ & & & & & 2 & 4 \\ & & & & & & 1 \end{pmatrix} \end{array},$$

and

$$\boldsymbol{C} = \begin{array}{c} \\ (5) \\ (4)(1) \\ (3)(2) \\ (3)(1)^2 \\ (2)^2(1) \\ (2)(1)^3 \\ (1)^5 \end{array} \begin{array}{c} \begin{array}{ccccccc} (5) & (4,1) & (3,2) & (3,1^2) & (2^2,1) & (2,1^3) & (1^5) \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 0 & -1 & 1 & 0 & 2 & 4 \\ 0 & -1 & 1 & -1 & 1 & 1 & 5 \\ 1 & 0 & 0 & 0 & -2 & 0 & 6 \\ 0 & 1 & -1 & -1 & 1 & -1 & 5 \\ -1 & 0 & 1 & 1 & 0 & -2 & 4 \\ 1 & -1 & -1 & 1 & 1 & -1 & 1 \end{pmatrix} \end{array},$$

where the first column lists the characters at the 5-cycles, the second at the product of 4- and 1-cycles, the third at the product of 3- and 2-cycles, the fourth at the product of 3-and two 1-cycles, the fifth at the product of two 2-and 1-cycles, the fifth at the product of 2-and three 1-cycles, and the last at identity element that has cycle type 1^5. The last column entries $\chi_\lambda(id)$ are the dimensions of the Specht modules.

C.10 Dimension of Specht Modules

Let λ be a partition of n. The permutation module M^λ decomposes into a direct sum of Specht modules

$$M^\lambda = \bigoplus_{\lambda \trianglerighteq \mu} k_{\lambda,\mu} S^\mu,$$

where $k_{\lambda,\mu} \geq 0$ denotes the multiplicity with which the module S^λ occurs in the module M^μ. In particular, we have $k_{\lambda,\lambda} = 1$. Since the matrix $\boldsymbol{K} = (k_{\lambda,\mu})$ is lower triangular with 1's down the diagonal, it can be inverted. The inverse matrix is also lower triangular with 1's down the diagonal. For this, if we write the above decomposition in the form

$$[\lambda_1][\lambda_2]\ldots = \bigoplus_{\mu} k_{\lambda,\mu}[\mu],$$

then we obtain

$$[\lambda] = \bigoplus_{\mu} (k^{-1})_{\lambda,\mu}[\mu_1][\mu_2]\ldots.$$

For instance, the inverse of the matrix $\boldsymbol{K}$ for the group S_5 is given as

$$\begin{array}{c} \\ (5) \\ (4,1) \\ (3,2) \\ (3,1^2) \\ (2^2,1) \\ (2,1^3) \\ (1^5) \end{array} \begin{array}{c} \begin{array}{ccccccc} [5] & [4,1] & [3][2] & [3][1]^2 & [2]^2[1] & [2][1]^3 & [1]^5 \end{array} \\ \begin{pmatrix} 1 & & & & & & \\ -1 & 1 & & & & & \\ 0 & -1 & 1 & & & & \\ 1 & -1 & -1 & 1 & & & \\ 0 & 1 & -1 & -1 & 1 & & \\ -1 & 1 & 2 & -1 & -2 & 1 & \\ 1 & -2 & -2 & 3 & 3 & -4 & 1 \end{pmatrix} \end{array}.$$

The inverse matrix of $\boldsymbol{K}$ can be calculated as follows.

C.29. **The Determinantal Form** If λ is a partition into n non-zero parts, then

$$[\lambda] = \det\left([\lambda_i - i + j]\right)_{i,j=1}^{n},$$

where $[m] = 0$ if $m < 0$.

The determinant for $[\lambda]$ is to put $[\lambda_1], [\lambda_2], \ldots$ in order down the diagonal, and then let the numbers increase by 1 in each row to the right of the diagonal and decrease by 1 in each row to the left of the diagonal. The element [0] serves as multiplicative identity.

For instance, we have

$$[4,1] = \det\begin{pmatrix} [4] & [5] \\ [0] & [1] \end{pmatrix} = [4][1] - [5]$$

and

$$[3,2] = \det\begin{pmatrix} [3] & [4] \\ [1] & [2] \end{pmatrix} = [3][2] - [4][1].$$

If the determinantal form holds for partitions into two non-zero parts, it can be extended to partitions into three non-zero parts, say by expanding up the last column as follows,

$$\begin{aligned} [3,1^2] &= \det\begin{pmatrix} [3] & [4] & [5] \\ [0] & [1] & [2] \\ [\text{-}1] & [0] & [1] \end{pmatrix} \\ &= [5]\det\begin{pmatrix} [0] & [1] \\ 0 & [0] \end{pmatrix} - [2]\det\begin{pmatrix} [3] & [4] \\ 0 & [0] \end{pmatrix} + [1]\det\begin{pmatrix} [3] & [4] \\ [0] & [1] \end{pmatrix} \\ &= [5] - [3][2] + [3][1]^2 - [4][1]. \end{aligned}$$

Corollary C.30. *The Specht module S^λ has the dimension*

$$\dim S^\lambda = n! \cdot \det\left(\frac{1}{\lambda_i - i + j}\right)_{i,j=1}^{n},$$

where $1/r! = 0$ if $r < 0$.

The dimension of the Specht module S^λ can be calculated by using hooks. The (i,j)-hook of the diagram $[\lambda]$ consists of the (i,j)-node along with the $\lambda_i - j$ nodes to the right of it, the hooks's *arm*, and the $\lambda'_j - i$ nodes below it, the hook's *leg*. The length of the (i,j)-hook is the number of involved

nodes, $h_{ij} = \lambda_i + \lambda'_j + 1 - i - j$. If we replace the (i,j)-node in the diagram $[\lambda]$ by the number h_{ij} for each node, we obtain the *hook graph*. For instance, the partition $\lambda = (3,2)$ corresponds to the hook graph

$$\begin{array}{ll} \mathrm{X\ X\ X} & 4\ 3\ 1 \\ \mathrm{X\ X} & 2\ 1 \end{array}$$

C.31. **The Hook Formula** The dimension of the Specht module S^λ is given by

$$\dim S^\lambda = \frac{n!}{\prod(\text{hook lengths in } [\lambda])}.$$

For instance, by the above hook graph for the partition $\lambda = (3,2)$, the Specht module $S^{(3,2)}$ has the dimension

$$\dim S^{(3,2)} = \frac{5!}{4 \cdot 3 \cdot 2} = 5.$$

Proof. We show the result for partitions into three non-zero parts. By Corollary C.30,

$$\begin{aligned}
\frac{\dim S^\lambda}{n!} &= \det \begin{pmatrix} \frac{1}{(h_{11}-2)!} & \frac{1}{(h_{11}-1)!} & \frac{1}{h_{11}!} \\ \frac{1}{(h_{21}-2)!} & \frac{1}{(h_{21}-1)!} & \frac{1}{h_{21}!} \\ \frac{1}{(h_{31}-2)!} & \frac{1}{(h_{31}-1)!} & \frac{1}{h_{31}!} \end{pmatrix} \\
&= \frac{1}{h_{11}!}\frac{1}{h_{21}!}\frac{1}{h_{31}!} \det \begin{pmatrix} h_{11}(h_{11}-1) & h_{11} & 1 \\ h_{21}(h_{21}-1) & h_{21} & 1 \\ h_{31}(h_{31}-1) & h_{31} & 1 \end{pmatrix} \\
&= \frac{(h_{11}-h_{21})(h_{11}-h_{31})(h_{21}-h_{31})}{h_{11}!h_{21}!h_{31}!} \\
&= \frac{1}{h_{11}!}\frac{1}{h_{21}!}\frac{1}{h_{31}!} \det \begin{pmatrix} (h_{11}-1)(h_{11}-2) & h_{11}-1 & 1 \\ (h_{21}-1)(h_{21}-2) & h_{21}-1 & 1 \\ (h_{31}-1)(h_{31}-2) & h_{31}-1 & 1 \end{pmatrix} \\
&= \frac{1}{h_{11}h_{21}h_{31}} \det \begin{pmatrix} \frac{1}{(h_{11}-3)!} & \frac{1}{(h_{11}-2)!} & \frac{1}{(h_{11}-1)!} \\ \frac{1}{(h_{21}-3)!} & \frac{1}{(h_{21}-2)!} & \frac{1}{(h_{21}-1)!} \\ \frac{1}{(h_{31}-3)!} & \frac{1}{(h_{31}-2)!} & \frac{1}{(h_{31}-1)!} \end{pmatrix} \\
&= \frac{1}{h_{11}h_{21}h_{31}} \cdot \frac{1}{\prod(\text{hook lengths in } [\lambda_1 - 1, \lambda_2 - 1, \ldots])} \\
&= \frac{1}{\prod(\text{hook lengths in } [\lambda])},
\end{aligned}$$

where the next to the last equation follows by making use of the induction hypothesis. ♢

Index