Part I

Introduction to the four themes

Part I of this book is devoted to outlining the basic principles of algebraic statistics and their relationship to computational biology. Although some of the ideas are complex, and their relationships intricate, the underlying philosophy of our approach to biological sequence analysis is summarized in the cartoon on the cover of the book. The fictional character is DiaNA, who appears throughout the book, and is the statistical surrogate for our biological intuition. In the cartoon, DiaNA is walking randomly on a graph and is tossing tetrahedral dice that can land on one of the letters `A,C,G` or `T`. A key feature of the tosses is that the outcome depends on the direction of her route. We, the observers, record the letters that appear on the successive throws, but are unable to see the path that DiaNA takes on her graph. Our goal is to guess DiaNA's path from the die roll outcomes. That is, we wish to make an inference about missing data from certain observations.

In this book, the observed data are DNA sequences. A standard problem of computational biology is to infer an optimal alignment for two given DNA sequences. We shall see that this problem is precisely our example of guessing DiaNA's path. In Chapter 4 we give an introduction to the relevant biological concepts, and we argue that our example is not just a toy problem but is fundamental for designing efficient algorithms for analyzing real biological data.

The tetrahedral shape of DiaNA's dice hint at convex polytopes. We shall see in Chapter 2 that polytopes are geometric objects which play a key role in statistical inference. Underlying the whole story is computational algebra, featured in Chapter 3. Algebra is a universal language with which to describe the process at the heart of DiaNA's randomness.

Chapter 1 offers a fairly self-contained introduction to algebraic statistics. Many concepts of statistics have a natural analog in algebraic geometry, and there is an emerging dictionary which bridges the gap between these disciplines:

Statistics		**Algebraic Geometry**
independence	=	Segre variety
log-linear model	=	toric variety
curved exponential family	=	manifold
mixture model	=	join of varieties
MAP estimation	=	tropicalization
... ...	=	... ...

Table 0.1. *A glimpse of the statistics – algebraic geometry dictionary.*

This dictionary is far from being complete, but it already suggests that al-

gorithmic tools from algebraic geometry, most notably Gröbner bases, can be used for computations in statistics that are of interest for computational biology applications. While we are well aware of the limitations of algebraic algorithms, we nevertheless believe that computational biologists might benefit from adding the techniques described in Chapter 3 to their tool box. In addition, we have found the algebraic point of view to be useful in unifying and developing many computational biology algorithms. For example, the results on parametric sequence alignment in Chapter 7 do not require the language of algebra to be understood or utilized, but were motivated by concepts such as the Newton polytope of a polynomial. Chapter 2 discusses discrete algorithms which provide efficient solutions to various problems of statistical inference. Chapter 4 is an introduction to the biology, where we return to many of the examples in Chapter 1, illustrating how the statistical models we have discussed play a prominent role in computational biology.

We emphasize that Part I serves mainly as an introduction and reference for the chapters in Part II. We have therefore omitted many topics which are rightfully considered to be an integral part of computational biology. In particular, we have restricted ourselves to the topic of biological sequence analysis, and within that domain have focused on eukaryotic genome analysis. Readers may be interested in referring to [Durbin *et al.*, 1998] or [Ewens and Grant, 2005], our favorite introductions to the area of biological sequence analysis. Also useful may be a text on molecular biology with an emphasis on genomics, such as [Brown, 2002]. Our treatment of computational algebraic geometry in Chapter 3 is only a sliver taken from a mature and developed subject. The excellent book by [Cox *et al.*, 1997] fills in many of the details missing in our discussions.

Because Part I covers a wide range of topics, a comprehensive list of prerequisites would include a background in computer science, familiarity with molecular biology, and introductory courses in statistics and abstract algebra. Direct experience in computational biology would also be desirable. Of course, we recognize that this is asking too much. Real-life readers may be experts in one of these subjects but completely unfamiliar with others, and we have taken this into account when writing the book.

Various chapters provide natural points of entry for readers with different backgrounds. Those wishing to learn more about genomes can start with Chapter 4, biologists interested in software tools can start with Section 2.5, and statisticians who wish to brush up their algebra can start with Chapter 3.

In summary, the book is not meant to serve as the definitive text for algebraic statistics or computational biology, but rather as a first invitation to biology for mathematicians, and conversely as a mathematical primer for biologists. In other words, it is written in the spirit of interdisciplinary collaboration that is highlighted in the article *Mathematics is Biology's Next Microscope, Only Better; Biology is Mathematics' Next Physics, Only Better* [Cohen, 2004].

1

Statistics

Lior Pachter

Bernd Sturmfels

Statistics is the science of data analysis. The data to be encountered in this book are derived from *genomes*. Genomes consist of long chains of DNA which are represented by sequences in the letters `A`, `C`, `G` or `T`. These abbreviate the four nucleic acids `A`denine, `C`ytosine, `G`uanine and `T`hymine, which serve as fundamental building blocks in molecular biology.

What do statisticians do with their data? They build models of the process that generated the data and, in what is known as *statistical inference*, draw conclusions about this process. Genome sequences are particularly interesting data to draw conclusions from: they are the blueprint for life, and yet their function, structure, and evolution are poorly understood. Statistical models are fundamental for genomics, a point of view that was emphasized in [Durbin *et al.*, 1998].

The inference tools we present in this chapter look different from those found in [Durbin *et al.*, 1998], or most other texts on computational biology or mathematical statistics: ours are written in the language of abstract algebra. The algebraic language for statistics clarifies many of the ideas central to the analysis of discrete data, and, within the context of biological sequence analysis, unifies the main ingredients of many widely used algorithms.

Algebraic Statistics is a new field, less than a decade old, whose precise scope is still emerging. The term itself was coined by Giovanni Pistone, Eva Riccomagno and Henry Wynn, with the title of their book [Pistone *et al.*, 2000]. That book explains how polynomial algebra arises in problems from experimental design and discrete probability, and it demonstrates how computational algebra techniques can be applied to statistics.

This chapter takes some additional steps along the algebraic statistics path. It offers a self-contained introduction to algebraic statistical models, with the aim of developing inference tools relevant for studying genomes. Special emphasis will be placed on (hidden) Markov models and graphical models.

1.1 Statistical models for discrete data

Imagine a fictional character named DiaNA who produces sequences of letters over the four-letter alphabet $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. An example of such a sequence is

$$\mathtt{CTCACGTGATGAGAGCATTCTCAGACCGTGACGCGTGTAGCAGCGGCTC}. \tag{1.1}$$

The sequences produced by DiaNA are called *DNA sequences*. DiaNA generates her sequences by some random process. When modeling this random process we make assumptions about part of its structure. The resulting *statistical model* is a family of probability distributions, one of which governs the process by which DiaNA generates her sequences. In this book we consider parametric statistical models, which are families of probability distributions that can be parameterized by finitely many parameters. One important task is to estimate DiaNA's parameters from the sequences she generates. Estimation is also called *learning* in the computer science literature.

DiaNA uses *tetrahedral dice* to generate DNA sequences. Each die has the shape of a tetrahedron, and its four faces are labeled with the letters `A`, `C`, `G` and `T`. If DiaNA rolls a fair die then each of the four letters will appear with the same probability 1/4. If she uses a loaded tetrahedral die then the four probabilities can be any four non-negative numbers that sum to one.

Example 1.1 Suppose that DiaNA uses three tetrahedral dice. Two of her dice are loaded and one die is fair. The probabilities of rolling the four letters are known to us. They are the numbers in the rows of the following table:

	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

(1.2)

DiaNA generates each letter in her DNA sequence independently using the following process. She first picks one of her three dice at random, where her first die is picked with probability θ_1, her second die is picked with probability θ_2, and her third die is picked with probability $1 - \theta_1 - \theta_2$. The probabilities θ_1 and θ_2 are unknown to us, but we do know that DiaNA makes one roll with the selected die, and then she records the resulting letter, `A`, `C`, `G` or `T`.

In the setting of biology, the first die corresponds to DNA that is `G` + `C` rich, the second die corresponds to DNA that is `G` + `C` poor, and the third is a fair die. We got the specific numbers in the first two rows of (1.2) by averaging the rows of the two tables in [Durbin *et al.*, 1998, page 50] (for more on this example and its connection to `CpG` island identification see Chapter 4).

Suppose we are given the DNA sequence of length $N = 49$ shown in (1.1). One question that may be asked is whether the sequence was generated by DiaNA using this process, and, if so, which parameters θ_1 and θ_2 did she use?

Let $p_{\mathtt{A}}$, $p_{\mathtt{C}}$, $p_{\mathtt{G}}$ and $p_{\mathtt{T}}$ denote the probabilities that DiaNA will generate any of her four letters. The statistical model we have discussed is written in

algebraic notation as

$$\begin{aligned} p_{\mathtt{A}} &= -0.10 \cdot \theta_1 + 0.02 \cdot \theta_2 + 0.25, \\ p_{\mathtt{C}} &= 0.08 \cdot \theta_1 - 0.01 \cdot \theta_2 + 0.25, \\ p_{\mathtt{G}} &= 0.11 \cdot \theta_1 - 0.02 \cdot \theta_2 + 0.25, \\ p_{\mathtt{T}} &= -0.09 \cdot \theta_1 + 0.01 \cdot \theta_2 + 0.25. \end{aligned}$$

Note that $p_{\mathtt{A}} + p_{\mathtt{C}} + p_{\mathtt{G}} + p_{\mathtt{T}} = 1$, and we get the three distributions in the rows of (1.2) by specializing (θ_1, θ_2) to $(1, 0)$, $(0, 1)$ and $(0, 0)$ respectively.

To answer our questions, we consider the *likelihood* of observing the particular data (1.1). Since each of the 49 characters was generated independently, that likelihood is the product of the probabilities of the individual letters:

$$L \quad = \quad p_{\mathtt{C}} p_{\mathtt{T}} p_{\mathtt{C}} p_{\mathtt{A}} p_{\mathtt{C}} p_{\mathtt{G}} \cdots p_{\mathtt{A}} \quad = \quad p_{\mathtt{A}}^{10} \cdot p_{\mathtt{C}}^{14} \cdot p_{\mathtt{G}}^{15} \cdot p_{\mathtt{T}}^{10}.$$

This expression is the *likelihood function* of DiaNA's model for the data (1.1). To stress the fact that the parameters θ_1 and θ_2 are unknowns we write

$$L(\theta_1, \theta_2) \quad = \quad p_{\mathtt{A}}(\theta_1, \theta_2)^{10} \cdot p_{\mathtt{C}}(\theta_1, \theta_2)^{14} \cdot p_{\mathtt{G}}(\theta_1, \theta_2)^{15} \cdot p_{\mathtt{T}}(\theta_1, \theta_2)^{10}.$$

This likelihood function is a real-valued function on the triangle

$$\Theta \quad = \quad \bigl\{ (\theta_1, \theta_2) \in \mathbb{R}^2 \,:\, \theta_1 > 0 \text{ and } \theta_2 > 0 \text{ and } \theta_1 + \theta_2 < 1 \bigr\}.$$

In the *maximum likelihood* framework we estimate the parameter values that DiaNA used by those values which make the likelihood of observing the data as large as possible. Thus our task is to maximize $L(\theta_1, \theta_2)$ over the triangle Θ. It is equivalent but more convenient to maximize the *log-likelihood function*

$$\begin{aligned} \ell(\theta_1, \theta_2) &= \log\bigl(L(\theta_1, \theta_2)\bigr) \\ &= 10 \cdot \log(p_{\mathtt{A}}(\theta_1, \theta_2)) + 14 \cdot \log(p_{\mathtt{C}}(\theta_1, \theta_2)) \\ &\quad + 15 \cdot \log(p_{\mathtt{G}}(\theta_1, \theta_2)) + 10 \cdot \log(p_{\mathtt{T}}(\theta_1, \theta_2)). \end{aligned}$$

The solution to this optimization problem can be computed in closed form, by equating the two partial derivatives of the log-likelihood function to zero:

$$\begin{aligned} \frac{\partial \ell}{\partial \theta_1} &= \frac{10}{p_{\mathtt{A}}} \cdot \frac{\partial p_{\mathtt{A}}}{\partial \theta_1} + \frac{14}{p_{\mathtt{C}}} \cdot \frac{\partial p_{\mathtt{C}}}{\partial \theta_1} + \frac{15}{p_{\mathtt{G}}} \cdot \frac{\partial p_{\mathtt{G}}}{\partial \theta_1} + \frac{10}{p_{\mathtt{T}}} \cdot \frac{\partial p_{\mathtt{T}}}{\partial \theta_1} = 0, \\ \frac{\partial \ell}{\partial \theta_2} &= \frac{10}{p_{\mathtt{A}}} \cdot \frac{\partial p_{\mathtt{A}}}{\partial \theta_2} + \frac{14}{p_{\mathtt{C}}} \cdot \frac{\partial p_{\mathtt{C}}}{\partial \theta_2} + \frac{15}{p_{\mathtt{G}}} \cdot \frac{\partial p_{\mathtt{G}}}{\partial \theta_2} + \frac{10}{p_{\mathtt{T}}} \cdot \frac{\partial p_{\mathtt{T}}}{\partial \theta_2} = 0. \end{aligned}$$

Each of the two expressions is a rational function in (θ_1, θ_2), i.e. it can be written as a quotient of two polynomials. By clearing denominators and by applying the algebraic technique of *Gröbner bases* (Section 3.1), we can transform the two equations above into the equivalent equations

$$\begin{aligned} 13003050 \cdot \theta_1 + 2744 \cdot \theta_2^2 - 2116125 \cdot \theta_2 - 6290625 &= 0, \\ 134456 \cdot \theta_2^3 - 10852275 \cdot \theta_2^2 - 4304728125 \cdot \theta_2 + 935718750 &= 0. \end{aligned} \tag{1.3}$$

The second equation has a unique solution $\hat{\theta}_2$ between 0 and 1. The corresponding value of $\hat{\theta}_1$ is obtained by solving the first equation. We find

$$(\hat{\theta}_1, \hat{\theta}_2) \quad = \quad \bigl(0.5191263945,\, 0.2172513326 \bigr).$$

The log-likelihood function attains its maximum value at this point:

$$\ell(\hat{\theta}_1, \hat{\theta}_2) \quad = \quad -67.08253037.$$

The corresponding probability distribution

$$(\hat{p}_{\mathtt{A}}, \hat{p}_{\mathtt{C}}, \hat{p}_{\mathtt{G}}, \hat{p}_{\mathtt{T}}) \quad = \quad \big(0.202432,\, 0.289358,\, 0.302759,\, 0.205451\big) \tag{1.4}$$

is very close to the empirical distribution

$$\frac{1}{49}(10, 14, 15, 10) \quad = \quad \big(0.204082,\, 0.285714,\, 0.306122,\, 0.204082\big). \tag{1.5}$$

We conclude that the proposed model is a good fit for the data (1.1). To make this conclusion precise we would need to employ a technique like the χ^2 test [Bickel and Doksum, 2000], but we keep our little example informal and simply assert that our calculation suggests that DiaNA used the probabilities $\hat{\theta}_1$ and $\hat{\theta}_2$ for choosing among her dice. □

We now turn to our general discussion of statistical models for discrete data. A *statistical model* is a family of probability distributions on some state space. In this book we assume that the state space is finite, but possibly quite large. We often identify the state space with the set of the first m positive integers,

$$[m] \quad := \quad \{1, 2, \ldots, m\}. \tag{1.6}$$

A probability distribution on the set $[m]$ is a point in the *probability simplex*

$$\Delta_{m-1} \quad := \quad \Big\{ (p_1, \ldots, p_m) \in \mathbb{R}^m \,:\, \sum_{i=1}^{m} p_i = 1 \text{ and } \; p_j \geq 0 \;\text{ for all } j \Big\}. \tag{1.7}$$

The index $m-1$ indicates the dimension of the simplex Δ_{m-1}. We write Δ for the simplex Δ_{m-1} when the underlying state space $[m]$ is understood.

Example 1.2 The state space for DiaNA's dice is the set $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ which we identify with the set $[4] = \{1, 2, 3, 4\}$. The simplex Δ is a tetrahedron. The probability distribution associated with a fair die is the point $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, which is the centroid of the tetrahedron Δ. Equivalently, we may think about our model via the concept of a *random variable*: that is, a function X taking values in the state space $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Then the point corresponding to a fair die gives the probability distribution of X as $\text{Prob}(X = \mathtt{A}) = \frac{1}{4}$, $\text{Prob}(X = \mathtt{C}) = \frac{1}{4}$, $\text{Prob}(X = \mathtt{G}) = \frac{1}{4}$, $\text{Prob}(X = \mathtt{T}) = \frac{1}{4}$. All other points in the tetrahedron Δ correspond to loaded dice. □

A statistical model for discrete data is a family of probability distributions on $[m]$. Equivalently, a statistical model is simply a subset of the simplex Δ. The ith coordinate p_i represents the probability of observing the state i, and in that capacity p_i must be a non-negative real number. However, when discussing algebraic computations (as in Chapter 3), we sometimes relax this requirement and allow p_i to be negative or even a complex number.

An *algebraic statistical model* arises as the image of a polynomial map

$$\mathbf{f} \,:\, \mathbb{R}^d \,\rightarrow\, \mathbb{R}^m \,, \quad \theta = (\theta_1, \ldots, \theta_d) \,\mapsto\, \bigl(f_1(\theta), f_2(\theta), \ldots, f_m(\theta)\bigr). \tag{1.8}$$

The unknowns $\theta_1, \ldots, \theta_d$ represent the model parameters. In most cases of interest, d is much smaller than m. Each coordinate function f_i is a polynomial in the d unknowns, which means it has the form

$$f_i(\theta) \quad = \quad \sum_{a \in \mathbb{N}^d} c_a \cdot \theta_1^{a_1} \theta_2^{a_2} \cdots \theta_d^{a_d}, \tag{1.9}$$

where all but finitely many of the coefficients $c_a \in \mathbb{R}$ are zero. We use $\mathbb{N}$ to denote the non-negative integers: that is, $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$.

The parameter vector $(\theta_1, \ldots, \theta_d)$ ranges over a suitable non-empty open subset Θ of $\mathbb{R}^d$ which is called the *parameter space* of the model $\mathbf{f}$. We assume that the parameter space Θ satisfies the condition

$$f_i(\theta) \,>\, 0 \qquad \text{for all } \; i \in [m] \;\text{ and }\; \theta \in \Theta. \tag{1.10}$$

Under these hypotheses, the following two conditions are equivalent:

$$\mathbf{f}(\Theta) \,\subseteq\, \Delta \qquad \Longleftrightarrow \qquad f_1(\theta) + f_2(\theta) + \cdots + f_m(\theta) \,=\, 1. \tag{1.11}$$

This is an identity of polynomial functions, which means that all non-constant terms of the polynomials f_i cancel, and the constant terms add up to 1. If (1.11) holds, then our model is simply the set $\mathbf{f}(\Theta)$.

Example 1.3 DiaNA's model in Example 1.1 is a *mixture model* which mixes three distributions on $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Geometrically, the image of DiaNA's map

$$\mathbf{f} : \mathbb{R}^2 \rightarrow \mathbb{R}^4 \,, \quad (\theta_1, \theta_2) \mapsto (p_{\mathtt{A}}, p_{\mathtt{C}}, p_{\mathtt{G}}, p_{\mathtt{T}})$$

is the plane in $\mathbb{R}^4$ which is cut out by the two linear equations

$$p_{\mathtt{A}} \,+\, p_{\mathtt{C}} \,+\, p_{\mathtt{G}} \,+\, p_{\mathtt{T}} \,=\, 1 \qquad \text{and} \qquad 11\, p_{\mathtt{A}} \,+\, 15\, p_{\mathtt{G}} \,=\, 17\, p_{\mathtt{C}} \,+\, 9\, p_{\mathtt{T}}. \tag{1.12}$$

These two linear equations are *algebraic invariants* of the model. The plane they define intersects the tetrahedron Δ in the quadrangle whose vertices are

$$\left(0, 0, \frac{3}{8}, \frac{5}{8}\right) \,, \quad \left(0, \frac{15}{32}, \frac{17}{32}, 0\right) \,, \quad \left(\frac{9}{20}, 0, 0, \frac{11}{20}\right) \text{ and } \quad \left(\frac{17}{28}, \frac{11}{28}, 0, 0\right). \tag{1.13}$$

Inside this quadrangle is the triangle $\mathbf{f}(\Theta)$ whose vertices are the three rows of the table in (1.2). The point (1.4) lies in that triangle and is near (1.5). □

Some statistical models are given by a polynomial map $\mathbf{f}$ for which (1.11) does not hold. If this is the case then we scale each vector in $\mathbf{f}(\Theta)$ by the positive quantity $\sum_{i=1}^m f_i(\theta)$. Regardless of whether (1.11) holds or not, our model is the family of all probability distributions on $[m]$ of the form

$$\frac{1}{\sum_{i=1}^m f_i(\theta)} \cdot \bigl(f_1(\theta), f_2(\theta), \ldots, f_m(\theta)\bigr) \qquad \text{where} \quad \theta \in \Theta. \tag{1.14}$$

We generally try to keep things simple and assume that (1.11) holds. However,

there are some cases, such as the general toric model in the next section, when the formulation in (1.14) is more natural. It poses no great difficulty to extend our theorems and algorithms from polynomials to rational functions.

Our *data* are typically given in the form of a sequence of observations

$$i_1, i_2, i_3, i_4, \ldots, i_N. \tag{1.15}$$

Each data point i_j is an element from our state space $[m]$. The integer N, which is the length of the sequence, is called the *sample size*. Assuming that the observations (1.15) are independent and identically distributed samples, we can summarize the data (1.15) in the *data vector* $u = (u_1, u_2, \ldots, u_m)$ where u_k is the number of indices $j \in [N]$ such that $i_j = k$. Hence u is a vector in $\mathbb{N}^m$ with $u_1 + u_2 + \cdots + u_m = N$. The *empirical distribution* corresponding to the data (1.15) is the scaled vector $\frac{1}{N}u$ which is a point in the probability simplex Δ. The coordinates u_i/N of this vector are the *observed relative frequencies* of the various outcomes.

We consider the model $\mathbf{f}$ to be a "good fit" for the data u if there exists a parameter vector $\theta \in \Theta$ such that the probability distribution $\mathbf{f}(\theta)$ is very close, in a statistically meaningful sense [Bickel and Doksum, 2000], to the empirical distribution $\frac{1}{N}u$. Suppose we draw N times at random (independently and with replacement) from the set $[m]$ with respect to the probability distribution $\mathbf{f}(\theta)$. Then the probability of observing the sequence (1.15) equals

$$L(\theta) \quad = \quad f_{i_1}(\theta) f_{i_2}(\theta) \cdots f_{i_N}(\theta) \quad = \quad f_1(\theta)^{u_1} \cdots f_m(\theta)^{u_m}. \tag{1.16}$$

This expression depends on the parameter vector θ as well as the data vector u. However, we think of u as being fixed and then L is a function from Θ to the positive real numbers. It is called the *likelihood function* to emphasize that it is a function that depends on θ. Note that any reordering of the sequence (1.15) leads to the same data vector u. Hence the probability of observing the data vector u is equal to

$$\frac{(u_1 + u_2 + \cdots + u_m)!}{u_1! u_2! \cdots u_m!} \cdot L(\theta). \tag{1.17}$$

The vector u plays the role of a *sufficient statistic* for the model $\mathbf{f}$. This means that the likelihood function $L(\theta)$ depends on the data (1.15) only through u. In practice one often replaces the likelihood function by its logarithm

$$\ell(\theta) \; = \; \log L(\theta) \; = \; u_1 \cdot \log(f_1(\theta)) + u_2 \cdot \log(f_2(\theta)) + \cdots + u_m \cdot \log(f_m(\theta)). \tag{1.18}$$

This is the *log-likelihood function*. Note that $\ell(\theta)$ is a function from the parameter space $\Theta \subset \mathbb{R}^d$ to the negative real numbers $\mathbb{R}_{<0}$.

The problem of *maximum likelihood estimation* is to maximize the likelihood function $L(\theta)$ in (1.16), or, equivalently, the scaled likelihood function (1.17), or, equivalently, the log-likelihood function $\ell(\theta)$ in (1.18). Here θ ranges over the parameter space $\Theta \subset \mathbb{R}^d$. Formally, we consider the optimization problem:

$$\text{Maximize } \ell(\theta) \quad \text{subject to } \theta \in \Theta. \tag{1.19}$$

A solution to this optimization problem is denoted $\hat{\theta}$ and is called a *maximum likelihood estimate* of θ with respect to the model $\mathbf{f}$ and the data u.

Sometimes, if the model satisfies certain properties, it may be that there always exists a unique maximum likelihood estimate $\hat{\theta}$. This happens for *linear models* and *toric models*, due to the concavity of their log-likelihood function, as we shall see in Section 1.2. For most statistical models, however, the situation is not as simple. First, a maximum likelihood estimate need not exist (since we assume Θ to be open). Second, even if $\hat{\theta}$ exists, there can be more than one global maximum, in fact, there can be infinitely many of them. And, third, it may be very difficult to find any one of these global maxima. In that case, one may content oneself with a local maximum of the likelihood function. In Section 1.3 we shall discuss the *EM algorithm* which is a numerical method for finding solutions to the maximum likelihood estimation problem (1.19).

1.2 Linear models and toric models

In this section we introduce two classes of models which, under weak conditions on the data, have the property that the likelihood function has exactly one local maximum $\hat{\theta} \in \Theta$. Since the parameter spaces of the models are convex, the *maximum likelihood estimate* $\hat{\theta}$ can be computed using any of the hill-climbing methods of convex optimization, such as the gradient ascent algorithm.

1.2.1 Linear models

An algebraic statistical model $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ is called a *linear model* if each of its coordinate polynomials $f_i(\theta)$ is a linear function. Being a linear function means that there exist real numbers $a_{i1}, \ldots, a_{1d}$ and b_i such that

$$f_i(\theta) \quad = \quad \sum_{j=1}^{d} a_{ij}\theta_j + b_i. \tag{1.20}$$

The m linear functions $f_1(\theta), \ldots, f_m(\theta)$ have the property that their sum is the constant function 1. DiaNA's model studied in Example 1.1 is a linear model. For the data discussed in that example, the log-likelihood function $\ell(\theta)$ had a unique local maximum on the parameter triangle Θ. The following proposition states that this desirable property holds for every linear model.

Proposition 1.4 *For any linear model $\mathbf{f}$ and data $u \in \mathbb{N}^m$, the log-likelihood function $\ell(\theta) = \sum_{i=1}^{m} u_i \log(f_i(\theta))$ is concave. If the linear map $\mathbf{f}$ is one-to-one and all u_i are positive then the log-likelihood function is strictly concave.*

Proof Our assertion that the log-likelihood function $\ell(\theta)$ is concave states that the *Hessian matrix* $\left(\frac{\partial^2 \ell}{\partial\theta_j\,\partial\theta_k}\right)$ is negative semi-definite for every $\theta \in \Theta$. In other words, we need to show that every eigenvalue of this symmetric matrix is non-positive. The partial derivative of the linear function $f_i(\theta)$ in (1.20) with

respect to the unknown θ_j is the constant a_{ij}. Hence the partial derivative of the log-likelihood function $\ell(\theta)$ equals

$$\frac{\partial \ell}{\partial \theta_j} \quad = \quad \sum_{i=1}^{m} \frac{u_i a_{ij}}{f_i(\theta)}. \tag{1.21}$$

Taking the derivative again, we get the following formula for the Hessian matrix

$$\left(\frac{\partial^2 \ell}{\partial \theta_j \, \partial \theta_k} \right) \quad = \quad -A^{\mathrm{T}} \cdot \mathrm{diag}\left(\frac{u_1}{f_1(\theta)^2}, \frac{u_2}{f_2(\theta)^2}, \ldots, \frac{u_m}{f_m(\theta)^2} \right) \cdot A. \tag{1.22}$$

Here A is the $m \times d$ matrix whose entry in row i and column j equals a_{ij}. This shows that the Hessian (1.22) is a symmetric $d \times d$ matrix each of whose eigenvalues is non-positive.

The argument above shows that $\ell(\theta)$ is a concave function. Moreover, if the linear map $\mathbf{f}$ is one-to-one then the matrix A has rank d. In that case, provided all u_i are strictly positive, all eigenvalues of the Hessian are strictly negative, and we conclude that $\ell(\theta)$ is strictly concave for all $\theta \in \Theta$. □

The critical points of the likelihood function $\ell(\theta)$ of the linear model $\mathbf{f}$ are the solutions to the system of d equations in d unknowns which are obtained by equating (1.21) to zero. What we get are the *likelihood equations*

$$\sum_{i=1}^{m} \frac{u_i a_{i1}}{f_i(\theta)} \;=\; \sum_{i=1}^{m} \frac{u_i a_{i2}}{f_i(\theta)} \;=\; \cdots \;=\; \sum_{i=1}^{m} \frac{u_i a_{id}}{f_i(\theta)} \;=\; 0. \tag{1.23}$$

The study of these equations involves the combinatorial theory of hyperplane arrangements. Indeed, consider the m hyperplanes in d-space $\mathbb{R}^d$ which are defined by the equations $f_i(\theta) = 0$ for $i = 1, 2, \ldots, m$. The complement of this arrangement of hyperplanes in $\mathbb{R}^d$ is the set of parameter values

$$\mathcal{C} \quad = \quad \bigl\{ \theta \in \mathbb{R}^d \,:\, f_1(\theta) f_2(\theta) f_3(\theta) \cdots f_m(\theta) \;\neq\; 0 \bigr\}.$$

This set is the disjoint union of finitely many open convex polyhedra defined by inequalities $f_i(\theta) > 0$ or $f_i(\theta) < 0$. These polyhedra are called the *regions* of the arrangement. Some of these regions are bounded, and others are unbounded. The natural parameter space of the linear model coincides with exactly one bounded region. The other bounded regions would give rise to negative probabilities. However, they are relevant for the algebraic complexity of our problem. Let μ denote the number of bounded regions of the arrangement.

Theorem 1.5 (Varchenko's Formula) *If the u_i are positive, then the likelihood equations* (1.23) *of the linear model $\mathbf{f}$ have precisely μ distinct real solutions, one in each bounded region of the hyperplane arrangement $\{f_i = 0\}_{i \in [m]}$. All solutions have multiplicity one and there are no other complex solutions.*

This result first appeared in [Varchenko, 1995]. The connection to maximum likelihood estimation was explored in [Catanese *et al.*, 2005].

We already saw one instance of Varchenko's Formula in Example 1.1. The four lines defined by the vanishing of DiaNA's probabilities $p_{\mathtt{A}}$, $p_{\mathtt{C}}$, $p_{\mathtt{G}}$ or $p_{\mathtt{T}}$