Rend. Istit. Mat. Univ. Trieste
Vol. XXXVII, 39–70 (2005)

Introductory Notes to Algebraic Statistics

Serkan Hoşten and Suela Ruffa [(*)]

Contribution to "School (and Workshop) on Computational Algebra for Algebraic Geometry and Statistics", Torino, September 2004.

Summary. - *These are the notes of a short course on algebraic statistics, a new discipline across the fields of statistical modeling and computational commutativa algebra. The basics of the theory are provided together with brief reference to applications to design of experiments, to exponential and graphical models, and to computational biology.*

1. Introduction

These notes are based on five lectures that the first author has given in September 2004 at Politecnico di Torino during the school on Computational Algebra for Algebraic Geometry and Statistics. Hence the first author is grateful to the organizers for having been given this opportunity. The second author organized a draft of this lecture notes as part of her Laurea Magistralis project.

The notes are basic and informal. They are intended to give the reader a first glimpse of algebraic statistics, and some of the exciting developments. The interested reader has now many different sources to extend his/her knowledge in this area. The (by now classic) book [13] is a good starting source. The recent book [12] has

[(*)] Authors' addresses: Serkan Hoşten, San Francisco State University, USA, e-mail: serkan@math.sfsu.edu
Suela Ruffa, Politecnico di Torino, corso Duca degli Abruzzi, 24, 10129 Torino, Italia, e-mail: suela.ruffa@polito.it

been a tremendous addition. In particular, the first four chapters of this book give an excellent introduction to algebraic statistics with a view towards computational biology. The lectures were also partially based on the lectures notes by Bernd Sturmfels of his John von Neumann Lecture in 2003 (see `http://www-m10.ma.tum.de/neumann/`). For more advanced topics, Journal of Symbolic Computation is about to publish a special issue (Volume 41, Issue 2) on *Computational Algebraic Statistics* that contains a dozen articles.

We hope algebraic statistics will become an influential subject.

2. Algebraic Varieties and Statistical Models

2.1. Ideals, varieties, and Gröbner bases

By a polynomial f in n indeterminates $p_1, \ldots, p_n$ we mean a **finite** linear combination of *monomials* $p^\alpha = p_1^{\alpha_1} \cdot \cdots \cdot p_n^{\alpha_n}$

$$f(p) = \sum_{\alpha \in A} c_\alpha p^\alpha .$$

In most applications the coefficients c_α one uses are rational numbers. For the development of the theory $\mathbb{Q}$ could be replaced by any *field* k, such as the real numbers $\mathbb{R}$, complex numbers $\mathbb{C}$, or any finite field. Algebraic geometry has a long tradition of using $\mathbb{C}$ and other algebraically closed fields. The set of all polynomials $k[p_1, \ldots, p_n]$ with the indeterminates $p_1, \ldots, p_n$ and coefficients in the field k is a *ring* under the usual addition and multiplication of polynomials. Given a set of polynomials $f_1, \ldots, f_m$ in $k[p_1, \ldots, p_n]$, the set of simultaneous solutions

$$V(f_1, \ldots, f_m) = \{a \in k^n \mid f_1(a) = \cdots = f_m(a) = 0\,\}$$

is called the *algebraic variety* defined by $f_1, \ldots, f_m$. Many applications of algebraic geometry are concerned with computing, describing, or understanding such a variety.

Definition 2.1. *A nonempty set $I \subseteq k[p_1, \ldots, p_n]$ is called an* ideal*, if*

(i) $f + g \in I$ for each $f, g \in I$, and

(ii) $h \cdot f \in I$ *for each* $f \in I$ *and* $h \in k[p_1, \ldots, p_n]$.

One can easily show that for the ideal $I = \langle f_1, \ldots, f_m \rangle$ the set of solutions

$$V(\langle f_1, \ldots, f_m \rangle) = \{a \in k^n \mid f(a) = 0 \ \forall \ f \in I\}$$

equals the variety $V(f_1, \ldots, f_m)$. This means that if a second set of polynomials $g_1, \ldots, g_t$ generates the same ideal $I = \langle f_1, \ldots, f_m \rangle = \langle g_1, \ldots, g_t \rangle$ then the varieties $V(f_1, \ldots, f_m)$ and $V(g_1, \ldots, g_n)$ are equal. All what matters is the ideal and not the individual polynomials in the polynomial system. Consequently, we introduce the notion of the variety of an ideal.

Definition 2.2. *Given an ideal* $I \subseteq k[p_1, \ldots, p_n]$, *the set of its solutions*

$$V(I) = \{a \in \mathbb{C}^n \mid f(a) = 0 \forall \ f \in I\}$$

is called the affine variety defined by I.

The natural question arises whether there are ideals which are not related to polynomial system solving because we cannot find finitely many polynomials $f_1, \ldots, f_m$ so that $I = \langle f_1, \ldots, f_m \rangle$. Luckily, this is not the case. Here we will prove one of the pillars of algebraic geometry and commutative algebra, namely *Hilbert's basis theorem.*

Theorem 2.3. *Every ideal* I *in* $k[p_1, \ldots, p_n]$ *is finitely generated.*

Proof. The proof is done by induction on n, the number of indeterminates. We note that $k[p_1, \ldots, p_n] = k[p_1, \ldots, p_{n-1}][p_n]$. In other words, the polynomials in $k[p_1, \ldots, p_n]$ can be viewed as univariate polynomials in the variable p_n whose coefficients are polynomials themselves in the remaining $n-1$ indeterminates.

The case for $n = 1$ is easy, since in this case $k[p_1] = k[p]$ is a *principal ideal domain*, i.e., every ideal is generated by a single polynomial: If I is the zero ideal, it is generated by the zero polynomial, and we are done. Otherwise, let $f \in I$ be a nonzero polynomial with least degree. We claim that $I = \langle f \rangle$. Suppose $g \in I$. Then there are polynomials q and r in $k[p]$ such that $g = fq + r$ with the property that the degree of r is strictly smaller than the degree of f. If $r \neq 0$, then we conclude that r is in I, and that would contradict the minimality of $\deg(f)$. So r has to be the zero polynomial, and this finishes the case for $n = 1$.

Now suppose that $n > 1$. We construct a sequence of polynomials from I as follows. Let $f_1 \in I$ be a polynomial with minimal degree in the variable p_n. Then for $j \geq 1$, if $\langle f_1, \ldots, f_j \rangle$ is strictly smaller than I we let $f_{j+1} \in I \setminus \langle f_1, \ldots, f_j \rangle$ be a polynomial of minimal degree in p_n. We also let $u_j \in k[p_1, \ldots, p_{n-1}]$ be the leading coefficient of f_j. The ideal in $\mathbb{Q}[p_1, \ldots, p_{n-1}]$ generated by these leading coefficients is finitely generated by induction. We can assume that $u_1, \ldots, u_m$ generate this ideal. Now we claim that $I = \langle f_1, \ldots, f_m \rangle$. If this is not the case, we must have picked a polynomial $f_{m+1} \in I \setminus \langle f_1, \ldots, f_m \rangle$. Its leading coefficient u_{m+1} is in $\langle u_1, \ldots, u_m \rangle$, and therefore $u_{m+1} = \sum_{j=1}^{m} r_j u_j$ for $r_j \in k[p_1, \ldots, p_{n-1}]$. By our construction $d := \deg_{p_n}(f_{m+1})$ is greater than or equal to the degrees of $f_1, \ldots, f_m$ in p_n, and we could define the following polynomial

$$g := f_{m+1} - \sum_{j=1}^{m} r_j f_j p_n^{d-d_j}$$

where $d_j := \deg_{p_n}(f_j)$. The degree of g is strictly smaller than the degree of f_{m+1}. But note that g is also in $I \setminus \langle f_1, \ldots, f_m \rangle$. This contradicts the minimality of $\deg(f_{m+1})$, and we are done. □

Monomial orders and Gröbner bases

Definition 2.4. *A* monomial order $\leq$ *is a total order of all monomials* $p^u \in k[p_1, \ldots, p_n]$ *such that*

(i) $1 = p^0 \leq p^u$ *for all monomials* p^u*, and*

(ii) $p^u \leq p^v$ *implies* $p^u p^w \leq p^v p^w$ *for all* p^w*.*

Typical examples of monomial orders are the *lexicographic* and the *graded reverse lexicographic orders*.

Definition 2.5. *Fix an order on the variables* $p_1 > p_2 > \cdots p_n$*. Then* $p^u <_{lex} p^v$ *in the lexicographic order, if there exists* $j \in \{1, \ldots, n\}$ *such that* $u_i = v_i$ *for* $i = 1, \ldots, j-1$ *and* $u_j < v_j$*.*

Definition 2.6. *Fix an order on the variables* $p_1 > p_2 > \cdots > p_n$. *Then* $p^u <_{grevlex} p^v$ *with respect to the graded reverse lexicographic order, if* $\deg(p^u) < \deg(p^v)$ *or* $\deg(p^u) = \deg(p^v)$ *and there exists* $j \in \{1, \ldots, n\}$ *such that* $u_i = v_i$ *for* $i = j+1, \ldots, n$ *and* $u_j > v_j$.

In the second definition $\deg(p^u)$ is the degree of the monomial p^u and is equal to $u_1 + \cdots + u_n$. Monomial orders allow us to compare monomials. In particular, we can compare the monomials in a given polynomial and determine the largest one. This leads us to the definition of the *initial monomial* and *initial term.*

Definition 2.7. *Let* $\leq$ *be a monomial order on* $k[p_1, \ldots, p_n]$ *and let*

$$f \quad = \quad \sum_{\alpha \in A} c_\alpha p^\alpha$$

be a nonzero polynomial. Then $in(f)$, *the* initial term *of* f, *is the term* $c_\beta p^\beta$ *such that* $p^\beta = \max_{\alpha \in A} p^\alpha$. *The monomial* p^β *itself is called the* initial monomial *of* f.

Example 2.8. *Let* $f = 3p^3qr^2 + 6p^2q^2r^3 - 5pq^3r^2 \in \mathbb{Q}[p,q,r]$. *We define the lex and grevlex orders with respect to* $x > y > z$. *Then* $in_{lex}(f) = 3p^3qr^2$ *and* $in_{grevlex}(f) = 6p^2q^2r^3$.

Definition 2.9. *Let* I *be a nonzero ideal in* $k[p_1, \ldots, p_n]$ *and let* $\leq$ *be a monomial order. Then the* initial ideal *of* I *with respect to* $\leq$ *is the ideal*

$$in_{\leq}(I) \quad := \quad \langle in_{\leq}(f) \, : \, f \in I \rangle$$

generated by the initial terms of the polynomials in I.

Recall that by the Hilbert's basis theorem $\mathrm{in}(I)$ is finitely generated. Moreover, the proof of Theorem 2.3 shows that the generators of $\mathrm{in}(I)$ can be chosen to be monomials. Hence computing the initial ideal of I returns a *monomial ideal*, an ideal generated by monomials.

Definition 2.10. *Given a monomial order* $\leq$ *on* $k[p_1, \ldots, p_n]$, *a set of polynomials* $g_1, \ldots, g_m$ *generating the ideal* $I = \langle g_1, \ldots, g_m \rangle$ *is called a* Gröbner basis *with respect to* $\leq$ *if*

$$in_{\leq}(I) = \langle in_{\leq}(g_1), \ldots, in_{\leq}(g_m) \rangle.$$

Furthermore, if

(i) the coefficient of each initial term $in(g_i)$ is equal to one, and

(ii) no term of g_i is divisible by $\{in(g_1), \ldots, in(g_m)\} \setminus \{in(g_i)\}$,

then such a Gröbner basis is called a reduced Gröbner basis *of I.*

It is not hard to show that the reduced Gröbner basis of I with respect to a term order is unique; see [2, Chapter 2, §7, Proposition 6].

Example 2.11. *Let's take the ideal $I = \langle pq - rs, pr - s^2 \rangle \subset \mathbb{Q}[p, q, r, s]$ and the grevlex order $p > q > r > s$. The reduced Gröbner basis is*

$$\{\underline{pr} - s^2,\ \underline{pq} - rs,\ \underline{r^2 s} - qs^2\},$$

where the underlined terms are the initial terms. If we change the monomial order to the lex $s > r > q > p$ then the reduced Gröbner basis changes to

$$\{\underline{rs} - pq,\ \underline{s^2} - pr,\ \underline{pqs} - pr^2,\ \underline{pr^3} - p^2q^2\}.$$

We will not delve into the details of how one can compute Gröbner bases as in the above example. The standard algorithm is known as the Buchberger's algorithm (see [2, Chapter 2]), and it is implemented in all computational algebra systems such Maple, Mathematica, CoCoA, Singular, Macaulay 2 etc.

2.2. Parametric versus implicit descriptions

Let $g_1, \ldots, g_n \in k[\theta_1, \ldots, \theta_d]$ be n polynomials in d indeterminates. We can define the map

$$\begin{aligned} \varphi : k^d &\to k^n \\ (\theta_1, \ldots, \theta_d) &\mapsto (g_1(\boldsymbol{\theta}), \ldots, g_n(\boldsymbol{\theta})) \end{aligned}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$. We refer to the space k^d as the *parameter space* and the points $\boldsymbol{\theta}$ as the parameters. The image $\mathrm{im}(\varphi)$ of this polynomial map is said to be given parametrically. In general $\mathrm{im}(\varphi)$ is not an affine variety but we can look at the smallest affine variety containing $\mathrm{im}(\varphi)$. This is called the Zariski closure of $\mathrm{im}(\varphi)$, denoted by $\overline{\mathrm{im}(\varphi)}$. To compute $\overline{\mathrm{im}(\varphi)}$ we need the elimination theorem (see [2, Chapter 2, §1, Theorem 2, §3, Theorem 1]).

Theorem 2.12. *Let $J = \langle p_1 - g_1(\boldsymbol{\theta}), \ldots, p_n - g_n(\boldsymbol{\theta})\rangle$ be an ideal in the polynomial ring $k[\theta_1, \ldots, \theta_d, p_1, \ldots, p_n]$. Then*

$$\overline{\operatorname{im}(\varphi)} = V(J \cap k[p_1, \ldots, p_n]).$$

The equations defining the $\overline{\operatorname{im}(\varphi)}$ are called the *implicit equations* of this variety, and the process of finding these equations from a parametric representation is called *implicitization*. What concerns us here is the algorithmic method of computing the intersection $J \cap k[p_1, \ldots, p_n]$. This is accomplished via Gröbner bases using a special type of monomial order called an *elimination order*: a monomial order of $k[\theta_1, \ldots, \theta_d, p_1, \ldots, p_n]$ where $\{\theta_1, \ldots, \theta_d\} > \{p_1, \ldots, p_n\}$.

Theorem 2.13. *Let I be an ideal of $k[\theta_1, \ldots, \theta_d, p_1, \ldots, p_n]$ and let $\leq$ be an elimination term order where $\{\theta_1, \ldots, \theta_d\} > \{p_1, \ldots, p_n\}$. If $\mathcal{G}$ is the reduced Gröbner basis of I then $\mathcal{G} \cap k[p_1, \ldots, p_n]$ is the reduced Gröbner basis of $I \cap k[p_1, \ldots, p_n]$.*

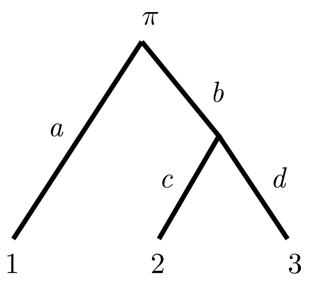
Proof. We need to show that the initial ideal of $I \cap k[p_1, \ldots, p_n]$ is generated by the initial terms of $\mathcal{G} \cap k[p_1, \ldots, p_n]$. If $f \in I \cap k[p_1, \ldots, p_n]$, then its initial term has to be divisible by the initial term of some $g \in \mathcal{G}$. Hence this initial term is a monomial of $k[p_1, \ldots, p_n]$. But then the elimination term order implies that $g \in \mathcal{G} \cap k[p_1, \ldots, p_n]$ and we are done. □

Phylogenetic invariants

We present an application of polynomial implicitization in the rapidly growing field of algebraic phylogenetics. The specific example below is taken from [18].

Phylogenetics is the study of the historical evolution of a set of species from a common ancestor using certain characteristics one can observe today. Typically, this evolution is represented by a *phylogenetic tree*. In the contemporary phylogenetics the characteristics one uses to reconstruct a phylogenetic tree is pieces of DNA and RNA sequences (sometimes the whole genome) of the organisms in question. To give a concrete example let's look at the following tree on $n = 3$ leaves pictured below.

Each node of the tree T (the leaves and non-leaves altogether) is a random variable. We will treat one of the simplest (but commonly



used) cases where all of these five random variables are binary, that is, they take values 0 or 1. The random variables at the leaves are *observed*, and they correspond to a binary characteristic we can observe of three species. The random variables at the interior nodes are *hidden*. These in turn correspond to the same binary characteristic of two ancestral species which we cannot observe. The probability distribution at the root is an unknown vector (π_0, π_1), where π_0 represents the probability of observing 0 at the root. For each of the four edges of the tree, we have the same 2×2-transition matrix:

$$M_a = M_b = M_c = M_d = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix}.$$

The four entries in this matrix are also unknown, and each entry a_{ij} represents the probability that an $i \in \{0,1\}$ changes to a $j \in \{0,1\}$ as the evolutionary clock is ticking. Together with the two root distribution parameters we get six *model parameters* describing our *model of evolution*. Assuming that all transitions on the edges are independent events, the monomial $\pi_u a_{ui} a_{uv} a_{vj} a_{vk}$ represents the probability of observing u at the root, v at the interior node, i at the leaf 1, j at the leaf 2, and k at the leaf 3, where $u, v, i, j, k \in \{0,1\}$. Since the probabilities at the root and the interior node are hidden, while the probabilities at the three leaves are observed we can write a polynomial that represents the probability of observing i, j, and k at the leaves 1, 2, and 3 respectively. In the language we developed in this section we have a polynomial map $\phi : \mathbb{R}^6 \longrightarrow \mathbb{R}^8$ defined by

$$\phi(\pi_0, \phi_1, a_{00}, a_{01}, a_{10}, a_{11}) = \\ = (p_{000}, p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, p_{111})$$

where

$$\begin{aligned}
p_{000} &= \pi_0 a_{00}^4 + \pi_0 a_{00} a_{01} a_{10}^2 + \pi_1 a_{10}^2 a_{00}^2 + \pi_1 a_{10}^3 a_{11}\\
p_{001} &= \pi_0 a_{00}^3 a_{01} + \pi_0 a_{00} a_{01} a_{10} a_{11} + \pi_1 a_{10}^2 a_{00} a_{01} + \pi_1 a_{10}^2 a_{11}^2\\
p_{010} &= \pi_0 a_{00}^3 a_{01} + \pi_0 a_{00} a_{01} a_{10} a_{11} + \pi_1 a_{10}^2 a_{00} a_{01} + \pi_1 a_{10}^2 a_{11}^2\\
p_{011} &= \pi_0 a_{00}^2 a_{01}^2 + \pi_0 a_{00} a_{01} a_{11}^2 + \pi_1 a_{10}^2 a_{01}^2 + \pi_1 a_{10} a_{11}^3\\
p_{100} &= \pi_0 a_{00}^3 a_{01} + \pi_0 a_{01}^2 a_{10}^2 + \pi_1 a_{11} a_{10} a_{00}^2 + \pi_1 a_{10}^2 a_{11}^2\\
p_{101} &= \pi_0 a_{00}^2 a_{01}^2 + \pi_0 a_{01}^2 a_{10} a_{11} + \pi_1 a_{11} a_{10} a_{00} a_{01} + \pi_1 a_{10} a_{11}^3\\
p_{110} &= \pi_0 a_{00}^2 a_{01}^2 + \pi_0 a_{01}^2 a_{10} a_{11} + \pi_1 a_{11} a_{10} a_{00} a_{01} + \pi_1 a_{10} a_{11}^3\\
p_{111} &= \pi_0 a_{01}^3 a_{00} + \pi_0 a_{01}^2 a_{11}^2 + \pi_1 a_{11} a_{10} a_{01}^2 + \pi_1 a_{11}^4.
\end{aligned}$$

Using Theorem 2.13 we can compute the implicit equations of $\overline{\mathrm{im}(\varphi)}$. This variety is defined by $p_{101} - p_{110}$ and $p_{001} - p_{010}$ together with a single polynomial of degree seven

$$\begin{aligned}
&p_{000}^2 p_{011}^4 p_{110} + 2p_{000}^2 p_{011}^2 p_{110}^3 + p_{000}^2 p_{110}^5 - 2p_{000} p_{010}^2 p_{011}^3 p_{110}\\
&- 2p_{000} p_{010}^2 p_{011} p_{110}^3 + 2p_{000} p_{010} p_{011}^3 p_{100} p_{110} - 2_{p000} p_{010} p_{011}^2 p_{100} p_{110}^2\\
&+ 2p_{000} p_{010} p_{011} p_{100} p_{110}^3 - 2p_{000} p_{010} p_{100} p_{110}^4 - p_{000} p_{011}^4 p_{100}^2\\
&- p_{000} p_{011}^3 p100^2 p_{110} - p_{000} p_{011}^2 p_{100}^2 p_{110}^2 + p_{010}^4 p_{011}^2 p_{110}\\
&- 2p_{010}^3 p_{011}^2 p_{100} p_{110} + 2p_{010}^3 p_{011} p_{100} p_{110}^2 + p_{010}^2 p_{011}^3 p_{100}^2\\
&+ p_{010}^2 p_{011}^2 p_{100}^2 p_{110} - 2p_{010}^2 p_{011} p_{100}^2 p_{110}^2\\
&+ p_{010}^2 p_{100}^2 p_{110}^3 + p_{010} p_{011}^2 p_{100}^3 p_{110}.
\end{aligned}$$

2.3. Statistical models as algebraic varieties

Let X be a discrete random variable taking values in $\{1, \dots, n\}$. In many examples in statistics the probabilities $\mathrm{P}(X = i)$ are given parametrically by polynomial $g_i(\theta_1, \dots, \theta_d)$ where $\theta_1, \dots, \theta_d$ are the parameters. Then the map

$$\begin{aligned}
\varphi : \mathbb{R}^d &\to \mathbb{R}^n\\
(\theta_1, \dots, \theta_d) &\mapsto (g_1(\boldsymbol{\theta}), \dots, g_n(\boldsymbol{\theta})).
\end{aligned}$$

describes the probability distributions on the state space $\{1, \dots, n\}$ prescribed by the polynomials $g_1, \dots, g_n$ as $\boldsymbol{\theta}$ varies in the parameter space. Usually the parametrization map φ is restricted to an open subset $U \subset \mathbb{R}^d$, and hence what we are really interested in is $\varphi(U) \cap \Delta_n$, where $\Delta_n = \{(p_1, \dots, p_n) : p_i \geq 0,\ p_1 + \cdots + p_n = 1\}$ is the $(n-1)$-dimensional probability simplex.

Definition 2.14. *A statistical model is* $\overline{\mathrm{im}(\varphi)} \cap \Delta_n$.

Example 2.15. *We consider a gene with two alleles A and a. X is the random variable taking values in $\{AA,\ Aa,\ aa\}$. Let θ be the probability of observing the allele A and $1-\theta$ the probability of observing the allele a. We mean*

$$\begin{aligned} p_1 &= P(X = AA) = \theta^2 \\ p_2 &= P(X = Aa) = 2\theta(1-\theta) \\ p_3 &= P(X = aa) = (1-\theta)^2. \end{aligned}$$

We want to describe $\overline{\mathrm{im}(\varphi)}$*, where*

$$\begin{aligned} \varphi : (0,1) &\to \mathbb{R}^3 \\ \theta &\mapsto (g_1(\theta), g_2(\theta), g_3(\theta)) \end{aligned}$$

with $g_1(\theta) = \theta^2$, $g_2(\theta) = \theta(1-\theta)$, $g_3(\theta) = (1-\theta)^2$.
Using Theorem 2.13, we find $\overline{\mathrm{im}(\varphi)} = V(\langle p_2^4 - 4p_1p_2,\ p_1+p_2+p_3-1\rangle)$

Example 2.16. *Suppose X and Y are two* independent *random variables described as follows. We let X be "the time that a man watches soccer on TV in Italy" and Y be "the amount of hair that a man has on his head". These variables take three possible values (time $<$ 1 hour, 1 hour $<$ time $<$ 3 hours, time $>$ 3 hours), with probability p_1, p_2, p_3 respectively, and (bold, receding, full), with probability q_1, q_2, q_3 respectively. Because these random variables are assumed to be independent the joint probability $P(X = i,\ Y = j) = p_{ij}$ is given by*

$$p_{ij} = p_i q_j$$

and the corresponding map is

$$\begin{aligned} \varphi : \mathbb{R}^6 &\to \mathbb{R}^{3\times 3} \\ (p_1, p_2, p_3, q_1, q_2, q_3) &\mapsto (p_i q_j \ \ i,j = 1,2,3) \end{aligned}$$

We can look at the image as the 3×3 *matrix obtained by the product:*

$$\begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} = \begin{pmatrix} p_1 & p_2 & p_3 \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix}.$$

This matrix has at most rank one; that is all 2×2 *minors are 0 and*

$$\overline{\operatorname{im}(\varphi)} = V(I_{2\times 2})$$

where

$$I_{2\times 2} = \langle p_{ij}p_{kl} - p_{il}p_{jk} \ 1 \leq i \leq k \leq 3, \ 1 \leq j \leq l \leq 3 \rangle.$$

In our example the statistical model is $V(I_{2\times 2(p_{ij})_{ij}}) \cap \Delta_9$.

3. Linear and Exponential Models

3.1. Linear models

A linear model is a statistical model where the polynomials $g_1, \ldots, g_n$ are *linear* polynomials in the parameters $\theta_1, \ldots, \theta_d$:

$$g_i(\theta_1, \ldots, \theta_d) = c_{i0} + c_{i1}\theta_1 + c_{i2}\theta_2 + \cdots + c_{id}\theta_d.$$

Hence the algebraic variety $\overline{\operatorname{im}(\varphi)}$ is an affine subspace of $\mathbb{R}^n$. The description of a linear model as it stands looks quite simple. However, there are many interesting questions one can ask in an applied context. For this we will use a widely studied example, namely, the design of experiments. In fact, this topic is one of the first statistics topics that pioneered the use of computational algebra. For a background and more details we refer the reader to the book by [13] and the papers [14] and [15].

Design of experiments

We start with an example inspired by an example in [14] to clarify the mathematical framework.

Example 3.1. *Suppose that a company is getting ready to launch a new product, say, a shaving gel for men. There are certain features of the product that this company sees as the most important attributes which determine the success among the consumers. For instance, these attributes could be the color of the gel, the perfume used in it, how soft it feels when shaving, the size of the packaging etc. The important thing is that there is a finite list* $A_1, A_2, \ldots, A_k$ *of* k *attributes. Furthermore, each attribute* A_i *can take a value from a finite list of values* $D_i = \{d_{i1}, d_{i2}, \ldots, d_{ij}\}$*. In our example, the color of the gel might be of three types: white, blue, and green. One way of doing a market research on this product is to present this shaving gel with various values of its attributes to consumers and ask them to rate the product from* 1 *(very poor) to* 5 *(excellent). Although this is a simple idea, there are many difficulties conducting this experiment. In our example, if the four attributes each have three values, then there are* 81 *different shaving gels a consumer has to try to evaluate all possible products, a costly and arduous task. The* design *of the experiment is then used to make a smart choice of a subset of possible products to infer consumer preferences that one wants to know.*

Now we will place the problems in the above example into a mathematical framework. In the situation as above the values d_{ij} can be taken to be numerical values. For instance, the three colors white, blue, and green would be coded by the integers $0, 1,$ and 2. If we assume all such values are numerical values than a *design point*, i.e., any one of the possible shaving gels in Example 3.1 is a point in $\mathbb{R}^k$. Let $D_1, D_2, \ldots, D_k$ be finite subsets of $\mathbb{R}$. These sets are associated to each of the k attributes A_i which take values in the set D_i. If we consider every attribute we obtain a full factorial design $D = D_1 \times \cdots \times D_k$, whereas if we consider a subset F of D we obtain a fraction. The points in F are called design points and indicated with $u_i \in \mathbb{R}^k$ $i = 1, \ldots, n$, where $n = |F|$ is the cardinality of F.

The totality of the design points is then an affine variety $F = \{u_1, \ldots, u_n\} \subset \mathbb{R}^k \subset \mathbb{C}^k$ consisting of finitely many points. The experiment itself is a function

$$f \,:\, F \longrightarrow \mathbb{R}.$$

In Example 3.1 the function f maps design points (different shaving gels) to some value given by the consumers, a real number between 1 and 5.

We now take the fraction $F \subset D$ and we define the model

$$p(\theta_1, \ldots, \theta_d, x_1, \ldots, x_k) = \sum_{\alpha \in S} \theta_\alpha x^\alpha$$

where S is the support set and d is the cardinality of S. Note that this is a linear model defined by

$$g_i(\theta_1, \ldots, \theta_d) := p(\theta_1, \ldots, \theta_d, u_i) = \sum_{\alpha \in S} \theta_\alpha x^\alpha(u_i)$$

where $x^\alpha(u_i) = u_{i1}^{\alpha_1} \cdots u_{ik}^{\alpha_k}$. With this we can formulate two problems:

1. **Direct problem**: Given a fraction $F = \{u_1, \ldots, u_n\}$ what are the linear models which can be identified by F? In other words, given the support S and the responses $f(u_i) = y_i$ is there a unique $(\theta_1, \ldots, \theta_d)$ so that $y_i = p(\theta_1, \ldots, \theta_d, u_i)$ for $i = 1, \ldots, n$?

2. **Inverse problem**: Given a model $p(\theta_1, \ldots, \theta_d, x_1, \ldots, x_k) = \sum_{\alpha \in S} \theta_\alpha x^\alpha$ what are the minimal fraction $F \subset D$ that identify the model?

Ideals of points

Before we continue, let's make an observation: the function $f : F \longrightarrow \mathbb{R}$ is a polynomial function, i.e. $f(x_1, \ldots, x_k)$ is a polynomial. Using Lagrange's interpolation we can find such a polynomial. However, this polynomial is not unique. In fact, two polynomials f_1 and f_2 give the same function f if and only if $f_1 - f_2$ vanishes on all of the points in F. More generally we make the following definition. Here and elsewhere, we set $k[x] = k[x_1, \ldots, x_n]$.

Definition 3.2. *Let $V \subset k^n$ be an affine variety. The* ideal *of V is*

$$I(V) := \{f \in k[x] \,:\, f(a) = 0, \text{ for all } a \in V\}.$$

It is not hard to see that $I(V)$ is indeed an ideal. Now two polynomials f and g in $k[x]$ will be the same functions when restricted to V if and only if their difference vanishes on V, in other words, $f - g$ is in $I(V)$. Therefore, the polynomial functions on V can be identified with the elements of the quotient ring $k[x]/I(V)$.

Proposition 3.3. *Let $V \subset k^n$ be an affine variety. Then the ring of polynomial functions on V is the quotient ring $k[x]/I(V)$.*

What we will do next depends on our ability to do computations in quotient rings such as $k[x]/I(V)$. We review this topic briefly.

Definition 3.4. *Let $M \subset k[x]$ be a monomial ideal. Then monomials not in M*

$$\{x^\alpha \, : \, x^\alpha \notin M\}$$

are called the standard monomials of M.

Lemma 3.5. *Let $G = \{g_1, \ldots, g_s\}$ be the reduced Gröbner basis of the ideal I with respect to a monomial order $\leq$. Then $f = g$ in $k[x]/I$ where g is the unique remainder of f obtained by reduction on G. This remainder is a k-linear combination of the standard monomials of $in_{\leq}(I)$.*

Proof. The first statement follows from the fact that when we do long division with respect to Gröbner basis we get unique remainders. The division algorithm guarantees that no term of g is divisible by the initial term of any g_i. Since $\text{in}(I) = \langle \text{in}(g_1), \ldots, \text{in}(g_s) \rangle$ the monomials appearing in g are standard monomials of $\text{in}(I)$. □

Theorem 3.6. *Let I be an ideal in $k[x]$ and $\leq$ a monomial order. Then the standard monomials of $in_{\leq}(I)$ form a k-basis for $k[x]/I$. Moreover, $k[x]/I$ and $k[x]/in(I)$ are isomorphic as vector spaces.*

Proof. Lemma 3.5 shows that the standard monomials of $\text{in}_{\leq}(I)$ span $k[x]/I$. To show linear independence, suppose $f = c_1 x^{\alpha_1} + \cdots c_t x^{\alpha_t}$ where $x^{\alpha_1}, \ldots, x^{\alpha_t}$ is a subset of the standard monomials is zero in $k[x]/I$. This means $f \in I$, and $\text{in}_{\leq}(f) \in \text{in}(I)$. Since the monomials in f are all standard monomials we conclude that $f = 0$, and hence $c_1 = \cdots = c_t = 0$. □

When we are working over an algebraically closed field such as $\mathbb{C}$ we can use *Hilbert's Nullstellensatz* [2, Chapter 4, §1, Theorem 2] to prove the following.

PROPOSITION 3.7. *Let $V = \{P_1, \dots, P_n\} \subset \mathbb{C}^k$ where $P_i = (p_{i1}, \dots, p_{ik})$ for $i = 1, \dots, n$. Then*

$$I(V) \quad = \quad \bigcap_{i=1}^{n} \langle x_1 - p_{i1}, \dots, x_k - p_{ik} \rangle .$$

Using the proposition we can compute $I(V)$, and then compute a Gröbner basis to get the standard monomials. In the case of a full factorial design D where $D_i = \{0, 1, \dots, d_i\}$ both $I(D)$ and *any* Gröbner basis of $I(D)$ is easy to compute.

PROPOSITION 3.8. *Let $D \subset \mathbb{R}^k$ be a full factorial design. Then $I(D)$ is generated by k polynomials g_i where*

$$g_i \;=\; x_i(x_i - 1)(x_i - 2) \cdots (x_i - d_i).$$

Furthermore, the set $\{g_1, \dots, g_k\}$ forms a reduced Gröbner basis for any monomal order.

Proof. The first statement follows from Proposition 3.7. For the second statement observe that for any monomial order $\text{in}(g_i) = x_i^{d_i+1}$. The cardinality of the standard monomials of $\langle \text{in}(g_i) : i = 1, \dots, k \rangle$ is equal to that of the standard monomials of $\text{in}(I(D))$. Hence the two ideals have to be the same, and this proves the second statement. □

There is one result remaining that allows us to answer Question 1.

THEOREM 3.9. *Let $V = \{P_1, \dots, P_n\} \subset \mathbb{C}^k$ be an affine variety consisting of finitely many points, and let $\leq$ be a monomial order. Then the dimension of $\mathbb{C}[x]/I(V)$ as a $\mathbb{C}$-vector space is equal to n. This number is in turn the number of standard monomials of $in_{\leq}(I(V))$.*

Proof. We have already established that $\dfrac{\mathbb{C}[x]}{I(V)}$ and $\dfrac{\mathbb{C}[x]}{\text{in}_{\leq}(I(V))}$ are isomorphic as vector spaces. The standard monomials of $\text{in}_{\leq}(I(V))$

is a basis for both vector spaces, and in the case of a finite affine variety V the number of the standard monomials, and hence the dimension of these two vector spaces, is finite. We let m denote this dimension. We leave it as an execise to show that there exists polynomials $f_1, \ldots, f_n \in \mathbb{C}[x]$ such that $f_i(p_i) = 1$ and $f_i(p_j) = 0$ for $i \neq j$. We define a linear transformation $\phi : \mathbb{C}^m = \mathbb{C}[x]/I(V) \longrightarrow \mathbb{C}^n$ by $\phi(f) = (f(p_1), \ldots, f(p_n))$. The existence of $f_1, \ldots, f_n$ above implies that ϕ is surjective, and hence also injective. This shows that $m = n$. □

Now given a fraction $F = \{u_1, \ldots, u_d\}$, we consider $\mathbb{R}[F] := \mathbb{R}[x_1, \ldots x_k]/I(F)$, all polynomial functions on F. The above results show that we can compute a reduced Groebner basis of $I(F)$ with respect to some term ordering $\leq$. Let this basis be $\{g_1, \ldots, g_s\}$. The ideal of the initial terms $\text{in}_{\leq}(I(F))$ is generated by

$$\{\text{in}_{\leq}(g_1), \ldots, \text{in}_{\leq}(g_s)\}.$$

We take the support set S to be the standard monomials of the ideal $\text{in}_{\leq}(I(F))$. The estimates of parameters $\theta_1, \ldots, \theta_d$ are the solution of

$$\begin{pmatrix} x^{\alpha_1}(u_1) & x^{\alpha_2}(u_1) & \cdots & x^{\alpha_d}(u_1) \\ x^{\alpha_1}(u_2) & x^{\alpha_2}(u_2) & \cdots & x^{\alpha_d}(u_2) \\ \vdots & \vdots & \ddots & \vdots \\ x^{\alpha_1}(u_d) & x^{\alpha_2}(u_d) & \cdots & x^{\alpha_d}(u_d) \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} = \begin{pmatrix} p_{u_1} \\ p_{u_2} \\ \vdots \\ p_{u_d} \end{pmatrix}. \quad (1)$$

The rank of the matrix in (1) is d. To show this suppose we reason by contradiction. Suppose that the columns are linearly dependent, that is there exist $\theta_1, \ldots, \theta_d \in \mathbb{R}$ not all 0 such that

$$p(\theta_1, \ldots, \theta_d, x_1, \ldots, x_k) = \sum_{x^\alpha \in S} \theta_\alpha x^\alpha = 0.$$

This means that $p(\cdot, \cdot)$ vanishes on F. Thus $p(\theta_1, \ldots, \theta_d, x_1, \ldots, x_k)$ is in $I(F)$. But this is not possible because all terms of $p(\cdot, \cdot)$ are standard monomials and none of them is divisible by $(\text{in}_{\leq}(g_i))$.

Example 3.10. *$D_1 = D_2 = \{0,1\}$ and $I(D) = \langle x^2 - x, y^2 - y\rangle$. In this case the standard monomials are $\{1, x, y, xy\}$. Let us take the model*

$$p(a, b, c, d, x, y) = a + bx + cy + dxy.$$

We have the system

$$\begin{cases} p_{00} = a \\ p_{10} = a + b \\ p_{01} = a + c \\ p_{11} = a + b + c + d \end{cases}$$

that has a unique solution.
A modified example would be to assume that there is no "interaction" between the x and y variables. Thus the model is

$$p(a, b, c, d, x, y) = a + bx + cy.$$

If we denote with:

$$A^T = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix} \qquad p = \begin{pmatrix} p_{00} \\ p_{10} \\ p_{01} \\ p_{11} \end{pmatrix} \qquad \theta = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

the system is

$$A^T\theta = p.$$

This system has solution if $p_{00} - p_{10} - p_{01} + p_{11} = 0$ that is if $p \in \mathrm{im}(A^T)$ is orthogonal to $\ker(A)$ (observe that a basis for $\ker(A)$ is $\langle(1, -1, -1, 1)\rangle$).

3.2. Exponential models

The easiest introduction to exponential models is to use the last example above. Instead of the linear model we have used we will use an exponential model for a fraction $F \subset \mathbb{N}^k$. It is defined as

$$p(\theta_1, \ldots, \theta_d, x_1, \ldots, x_k) = \exp\Big(\sum_{x^\alpha \in S} \theta_\alpha x^\alpha\Big)$$

and if we apply the change of coordinates

$$z_i = \exp(\theta_i)$$

we obtain

$$p(\theta_1, \dots, \theta_d, x_1, \dots, x_k) = z_1^{x^{\alpha_1(u)}} \cdots z_d^{x^{\alpha_d(u)}}$$

which is a product of monomials.

EXAMPLE 3.11. *The exponential model for the first part of the Example 3.10 is*

$$p = \exp(a + bx + cy + dxy).$$

In order to find a, b, c, d we have to solve the system

$$\begin{cases} p_{00} = \exp a = t \\ p_{10} = \exp(a+b) = ts \\ p_{01} = \exp(a+c) = tu \\ p_{11} = \exp(a+b+c+d) = tsuv \end{cases}$$

where $e^a = t,\ e^b = s,\ e^c = u,\ e^d = v$.
In the second situation of Example 3.10 we have:

$$p = \exp(a + bx + cy)$$

and the system

$$\begin{cases} p_{00} = \exp a = t \\ p_{10} = \exp(a+b) = ts \\ p_{01} = \exp(a+c) = tu \\ p_{11} = \exp(a+b+c) = tsu \end{cases} .$$

In this case we conclude that $p_{00}p_{11} - p_{01}p_{10} = 0$.

3.2.1. Exponential models and toric varieties

Let A be a $d \times n$ integer matrix with columns $a_1, \dots, a_n$ with $a_i \in \mathbb{Z}^d$. The class of models to be considered consists of discrete probability distributions defined via the matrix A. Let $\psi : \mathbb{C}^d \mapsto \mathbb{C}^n$ be the monomial map given by

$$(z_1, \dots, z_d) \to \Big(\prod_{i=1}^{d} z_i^{a_{1i}}, \dots, \prod_{i=1}^{d} z_i^{a_{ni}} \Big).$$

The *toric variety* defined by A is the Zariski closure of the image of this monomial map, i.e. $X_A := \overline{\text{im}(\psi)}$.

Proposition 3.12. *The toric variety X_A is defined by the ideal*

$$I_A = \langle p^u - p^v : u, v \in \mathbb{N}^n \text{ and } u - v \in \ker(A) \rangle.$$

Proof. It is easy to check the binomials in I_A vanish on X_A. Now let $f(p) = \sum_\alpha c_\alpha p^\alpha$ such that $\psi(f(p)) = 0$. If $f(p)$ is not the zero polynomial it has to have two terms $c_u p^u$ and $c_v p^v$, such that $\psi(p^u) = \psi(p^v)$. This means that $Au = Av$, and the polynomial $f(p) - c_u(p^u - p^v)$ vanishes on X_A and has one fewer term. Induction on the number of terms of $f(p)$ gives the result. □

In the monomial map ψ which defined the toric variety X_A we have used complex spaces as the domain and the range of this map. In order to work with probability distributions we will restrict ψ to the real map φ

$$\varphi : \mathbb{R}^d_{\geq 0} \to \mathbb{R}^n_{\geq 0}.$$

It is an interesting question to figure out whether a given probability distribution $p \in \Delta^n$ is in the image of φ. Such a probability distribution is said to *factor according to* A. The following theorem is Theorem 3 of [7] which characterizes when p factors according to A. We will not give the proof here. The statement of the theorem contains a combinatorial condition on the matrix A: A subset $F \subset \{1, \dots, n\}$ is said to be *nice* if for every $j \notin F$ the support of the j-th column a_j of the matrix A is not contained in $\bigcup_{k \in F} \text{supp}(a_k)$.

Theorem 3.13. *The probability distribution p factors according to A if and only if p is in $X_A^{\geq 0} := X_A \cap \mathbb{R}^n_{\geq 0}$ and the support of p is nice.*

Example 3.14. *From [7] we consider the following example. Consider the matrix*

$$A = \begin{bmatrix} 3 & 0 & 0 & 2 & 1 & 2 & 1 & 0 & 0 \\ 0 & 3 & 0 & 1 & 2 & 0 & 0 & 2 & 1 \\ 0 & 0 & 3 & 0 & 0 & 1 & 2 & 1 & 2 \end{bmatrix},$$

that defines the map

$$\varphi : (z_1, z_2, z_3) \mapsto (z_1^3, z_2^3, z_3^3, z_1^2 z_2, z_1 z_2^2, z_1^2 z_3, z_1 z_3^2, z_2^2 z_3, z_2 z_3^2).$$

We eliminate the variables z_1, z_2, z_3 from the ideal

$$\begin{aligned} I =&(p_1 - z_1^3, p_2 - z_2^3, p_3 - z_3^3, p_4 - z_1^2 z_2, p_5 - z_1 z_2^2, \\ &p_6 - z_1^2 z_3, p_7 - z_1 z_3^2, p_8 - z_2^2 z_3, p_9 - z_2 z_3^2) \end{aligned}$$

to obtain the sets of polynomials:

$$p_1p_5 - p_4^2, p_2p_4 - p_5^2, p_1p_7 - p_6^2, p_3p_6 - p_7^2, p_2p_9 - p_8^2, p_3p_8 - p_9^2 \quad (2)$$

$$p_1p_2 - p_4p_5, p_1p_3 - p_6p_7, p_3p_2 - p_8p_9 \quad (3)$$

$$p_1p_8 - p_5p_6, p_1p_9 - p_4p_7, p_2p_6 - p_4p_8, p_2p_7 - p_5p_9, p_3p_4 - p_6p_9, p_3p_5 - p_7p_8 \quad (4)$$

$$p_6p_8 - p_4p_9, p_5p_7 - p_4p_9 \quad (5)$$

$$p_1p_7p_8 - p_4p_6p_9, p_7^2p_8 - p_6p_9^2. \quad (6)$$

The polynomials in (2) define $X_A^{>0}$, and the ones in (2) and (3) define distributions that factor according to A. The polynomials in the first three sets give the complex toric variety $X_A^{\mathbb{C}}$. The toric ideal I_A is the union of the polynomials in (2)–(5). Finally all the polynomials form a reduced Gröbner basis for I_A.

4. Graphical models

4.1. Construction of a graphical model

Let X_i be a discrete random variable taking values in $D_i = \{1, \ldots, d_i\}$ for $i = 1, \ldots, n$. Let G be a graph with n vertices that are associated to each random variable X_i. If $C_1, \ldots, C_k$ are the maximal cliques of the graph G, then for each clique $C = \{X_{i_1}, \ldots, X_{i_s}\}$ we consider a set of variables $\psi_C(u_{i_1}, \dots, u_{i_s})$, where $u_i \in D_i$. The joint probability $p(X_1 = u_1, \dots, X_n = u_n) = \mathrm{P}_{u_1 u_2 \dots u_n}$ is

$$\mathrm{P}_{u_1 u_2 \dots u_n} = \psi_{C_1}(\boldsymbol{u}) \cdots \psi_{C_k}(\boldsymbol{u}). \qquad (7)$$

The definition of the joint probabilities is given by a monomial in the variables $\psi_{C_i}(\cdots)$ and hence is an exponential model. Such an exponential model is called a *graphical model* (see [11]).

EXAMPLE 4.1. *In this example we illustrate the matrix $A(G)$ which defines a graphical model. We consider four binary random variables $X_1, \dots, X_4$ and the graph in Figure 1.*

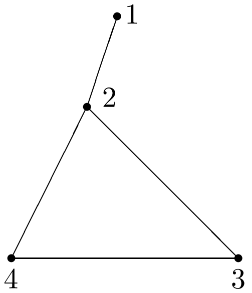


Figure 1: Graph for Example 4.1.

There are two maximal cliques, namely, $C_1 = \{1,2\}$ and $C_2 = \{2,3,4\}$. The variables associated to the first one are

$$\psi_{\{1,2\}}(0,0),\ \psi_{\{1,2\}}(0,1),\ \psi_{\{1,2\}}(1,0),\ \psi_{\{1,2\}}(1,1),$$

and those corresponding to the second one are

$$\psi_{\{2,3,4\}}(0,0,0),\ \psi_{\{2,3,4\}}(1,0,0), \dots, \psi_{\{2,3,4\}}(1,1,1).$$

From (7) we know that $\mathrm{P}_{ijkl} = \psi_{C_1}(ij)\psi_{C_2}(jkl)$ for all $i,j,k,l = 0,1$. For instance, $\mathrm{P}_{0101} = \psi_{C_1}(01)\psi_{C_2}(101)$.

We can represent joint probabilities in the matrix $A(G)$ (Table 4.1), whose columns are indexed by the joint probabilities and rows are indexed by the set of possible values obtained by the variables on each clique. The entry in the column corresponding to p_{ijkl} and in the row corresponding to $\psi_{C_1}(ij)$ and $\psi_{C_2}(jkl)$ are 1, and the other entries are 0.

	p_{0000}	p_{0001}	$\cdots$	$\cdots$	$\cdots$	p_{1111}
$\psi_{1,2}(00)$	1	1	$\cdots$	$\cdots$	$\cdots$	0
$\psi_{1,2}(01)$	0	0	$\cdots$	$\cdots$	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$				$\vdots$
$\psi_{1,2}(11)$	0	0				1
$\psi_{2,3,4}(000)$	1	0				0
$\psi_{2,3,4}(001)$	0	1				0
$\vdots$	$\vdots$					$\vdots$
$\psi_{2,3,4}(111)$	0	0	$\cdots$	$\cdots$	$\cdots$	1

Table 1: Matrix corresponding to the graph in Figure 1.

4.2. Independence models

Now let $X_1, \ldots, X_n$ be discrete random variables. Given three subsets X, Y, Z of $\{X_1, \ldots, X_n\}$, the symbols $X \perp\!\!\!\perp Y \,|\, Z$ denotes the conditional independence statement that the variables in X are independent of the variables in Y given those in Z. These variables satisfy

$$\begin{aligned}&\mathrm{P}(X=\boldsymbol{a}, Y=\boldsymbol{b}, Z=\boldsymbol{c})\mathrm{P}(X=\boldsymbol{a'}, Y=\boldsymbol{b'}, Z=\boldsymbol{c}) = \\ &\mathrm{P}(X=\boldsymbol{a}, Y=\boldsymbol{b'}, Z=\boldsymbol{c})\mathrm{P}(X=\boldsymbol{a'}, Y=\boldsymbol{b}, Z=\boldsymbol{c})\end{aligned}$$

for all vectors $\boldsymbol{c}$ and for all distinct vectors $\boldsymbol{a}, \boldsymbol{a'}, \boldsymbol{b}, \boldsymbol{b'}$. An independence statement as $X \perp\!\!\!\perp Y \,|\, Z$ gives rise to polynomial equations and these polynomial define the *independence ideal* $I_{X \perp\!\!\!\perp Y|Z}$.

Example 4.2. *Let X, Y, Z be three binary random variables, and suppose we have the independence statement $X \perp\!\!\!\perp Y|Z$. We get*

$$\begin{aligned}&\mathrm{P}(X=a, Y=b, Z=c)\mathrm{P}(X=a', Y=b', Z=c) = \\ &\mathrm{P}(X=a, Y=b', Z=c)\mathrm{P}(X=a', Y=b, Z=c)\end{aligned} \tag{8}$$

for distinct a, a', b, b' and for all c.
If we consider $a = b = 0,\ a' = b' = 1,\ c = 0, 1$ the ideal representing the independence property is generated by two quadratic binomials

$$I_{X \perp\!\!\!\perp Y|Z} = \langle p_{000}p_{110} - p_{010}p_{100},\ p_{001}p_{111} - p_{011}p_{101} \rangle$$

and the independence variety is $V_{X \perp\!\!\!\perp Y|Z} = V(I_{X \perp\!\!\!\perp Y|Z})$.

If we have more than one set of independence statements on $\{X_1, \ldots, X_n\}$, say,

$$X^{(1)} \perp\!\!\!\perp Y^{(1)}|Z^{(1)}, \ldots, X^{(k)} \perp\!\!\!\perp Y^{(k)}|Z^{(k)},$$

then the *independence ideal* is defined as

$$I = I_{X^{(1)} \perp\!\!\!\perp Y^{(1)}|Z^{(1)}} + \ldots + I_{X^{(k)} \perp\!\!\!\perp Y^{(k)}|Z^{(k)}}.$$

The variety defined by this ideal is

$$V = V_{X^{(1)} \perp\!\!\!\perp Y^{(1)}|Z^{(1)}} \bigcap \cdots \bigcap V_{X^{(k)} \perp\!\!\!\perp Y^{(k)}|Z^{(k)}}$$

and it is called an *independence model.* If $X \cup Y \cup Z = \{X_1, \ldots, X_n\}$ for each independence statement we have a *saturated independence model.*

4.3. Hammersley-Clifford Theorem

The defining ideal of a graphical model may not be easily computed. However, in these exponential models we can introduce two independence models which in certain cases describe the whole graphical models. These are the *pairwise independence* model and the *global independence* model.

Definition 4.3. *1. The pairwise independence model is an independence model with the further conditions*

$$X_i \perp\!\!\!\perp Y_j|\{X_1, \ldots, X_n\} \backslash \{X_i, X_j\}$$

where (i, j) *is not an edge in* G.

2. *The global independence model is an independence model with the further conditions*

$$Y \perp\!\!\!\perp Z | W$$

where $Y, Z, W \subseteq \{X_1, \dots, X_n\}$ and Y, Z will be disconnected when W is removed.

Following Definition 4.3 we introduce two ideals I_{pairwise} and I_{global} and the corresponding varieties $V(I_{\text{pairwise}})$ and $V(I_{\text{pairwise}})$. Clearly $I_{\text{pairwise}} \subset I_{\text{global}}$ and

$$X_{\text{pairwise}} = V(I_{\text{pairwise}}) \supset V(I_{\text{global}}) = X_{\text{global}}.$$

If we consider the map $\varphi_{A(G)} : \mathbb{R}^d_{>0} \to \mathbb{R}^n_{>0}$ then

$$\operatorname{im}(\varphi_{A(G)}) \subseteq X_{A(G)} \subseteq X_{\text{global}} \subseteq X_{\text{pairwise}}.$$

These bring us to one of the highlights of probability theory connected to graphical models, namely Hammersley-Clifford theorem. We state this theorem in the language we have developed so far.

Theorem 4.4 (Hammersley-Clifford).

$$\operatorname{im}(\varphi_{A(G)}) = X^{>0}_{A(G)} = X^{>0}_{\text{pairwise}}. \tag{9}$$

4.4. Decomposable models

Among all classes of graphical models there is a particular class which has been used extensively. For this we need a definition.

Definition 4.5. *G is called* chordal *if every cycle of length ≥ 4 has a chord.*

The smallest nontrivial example of a chordal graph G is in Figure 2.

Chordal graphs are "treelike" and have nice decomposition properties, as you can see in Figure 3.

Graphical models where G is chordal graph are called *decomposable models*. These models are "easy" from many perspectives. In the algebraic language we have seen we summarize this idea in the following theorem.

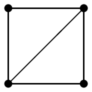


Figure 2: A chordal graph.

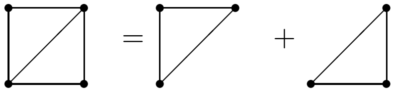


Figure 3: Decomposition properties of a chordal graph.

Theorem 4.6. *The following statements are equivalent*

(i) *G is chordal*

(ii) *I_G is quadratically generated*

(iii) *I_G has a quadratic Gröbner basis.*

This theorem and its proof appears in [7]. The papers [8] and [5] also contain one direction of the same theorem, namely that if G is chordal then I_G has a quadratic Gröbner basis. A thorough study of binary *graph* models has been done in [3]. Also there is a directed version of the graphical models known as *Bayesian networks*, see [6].

5. Maximum likelihood estimation

5.1. Definitions

Suppose that we have the statistical model:

$$\varphi : \mathbb{R}^d_{>0} \to \mathbb{R}^n_{>0}$$
$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \to (g_1(\boldsymbol{\theta}), \ldots, g_n(\boldsymbol{\theta}))$$

which define the family of probability densities $p_i = g_i(\boldsymbol{\theta})$.
If the nonnegative integer vector $u = (u_1, \ldots, u_n)$ is the data set

(where u_1 counts how many times you observe the variable assume its first value, u_2 counts how many times you observe the variable assume its second value, etc.) then we want to find what the best parameters $\boldsymbol{\theta}$ are that will explain the data. One way of doing this is to solve the following optimization problem

$$\max_{\boldsymbol{\theta}} \left(g_1^{u_1}(\boldsymbol{\theta}) \cdots g_n^{u_n}(\boldsymbol{\theta}) \right).$$

A solution $\hat{\boldsymbol{\theta}}$ of this estimation problem is called a *maximum likelihood estimate.* An equivalent problem is:

$$\max_{p \in V} \left(p_1^{u_1} \cdots p_n^{u_n} \right) \quad \text{with the condition} \quad \sum_{i=1}^{n} p_i = 1$$

Example 5.1. *Consider two independent binary variables with joint probabilities*

$$\begin{array}{ll} \mathrm{p}_{00} = \theta_1\theta_2 & \mathrm{p}_{01} = \theta_1(1-\theta_2) \\ \mathrm{p}_{10} = (1-\theta_1)\theta_2 & \mathrm{p}_{11} = (1-\theta_1)(1-\theta_2). \end{array}$$

To find the maximum likelihood estimate we have to solve

$$\max_{\theta_1,\theta_2}((\theta_1\theta_2)^{u_{00}}(\theta_1(1-\theta_2))^{u_{01}}((1-\theta_1)\theta_2)^{u_{10}}((1-\theta_1)(1-\theta_2))^{u_{11}})$$

where $u_{00}, \dots, u_{11}$ are the data.
Equivalently taking logarithms, we maximize the function

$$\log[(\theta_1\theta_2)^{u_{00}}(\theta_1(1-\theta_2))^{u_{01}}((1-\theta_1)\theta_2)^{u_{10}}((1-\theta_1)(1-\theta_2))^{u_{11}}].$$

By taking the derivatives we obtain the system of critical equations

$$\begin{cases} \dfrac{u_{00}+u_{01}}{\theta_1} - \dfrac{u_{10}+u_{11}}{1-\theta_1} = 0 \\ \dfrac{u_{00}+u_{10}}{\theta_2} - \dfrac{u_{01}+u_{11}}{1-\theta_2} = 0. \end{cases}$$

The solutions of this system of linear equations are $\hat{\theta}_1, \hat{\theta}_2$, the maximum likelihood estimates of θ_1, θ_2 where

$$\hat{\theta}_1 = \frac{u_{00}+u_{01}}{u_{00}+u_{01}+u_{10}+u_{11}} \qquad \hat{\theta}_2 = \frac{u_{00}+u_{01}}{u_{00}+u_{01}+u_{10}+u_{11}}.$$

5.2. MLEs for linear and toric models

We treat two general models for which the MLE can be found relatively easily. The first case is the linear models, i.e. the case when $g_i = a_{i0} + a_{i1}\theta_1 + \ldots + a_{id}\theta_d$ for $i = 1, \ldots, n$. This linear model can be described as

$$V_{\text{lin}} = \{p \in \mathbb{R}^n_{\geq 0} \, : \, Ap = b, \sum_{i=1}^{n} p_i = 1\}$$

for a matrix A of rank d. The maximum likelihood problem is then

$$\text{maximize} \sum_{i=1}^{n} u_i \log p_i \quad \text{such that} \quad p \in V_{\text{lin}}.$$

Using Lagrange multipliers the critical equations are

$$u_i/p_i \quad = \quad \sum_{j=1}^{d} \lambda_j a_{ij} + \mu, \quad i = 1, \ldots, n,$$

where λ_j and μ are Lagrange multipliers. The Hessian of the objective function is a diagonal matrix with the diagonal entries $-u_i/p_i^2$, and hence it is negative definite. This implies that linear models have a unique MLE.

Theorem 5.2. *The maximum likelihood estimate for a linear model is unique.*

Now we look at toric models. Given a matrix $A \in \mathbb{Z}^{d \times m}$, we assume that the vector $(1, \ldots, 1)$ is in the space generated by the rows of A (that is there exist $(s_1, \ldots, s_d)$ such that $(s_1, \ldots, s_d)A = (1, \ldots, 1)$).
In terms of ideals this means:

$$\begin{aligned} p^u - p^v \in I_A \quad &\Leftrightarrow Au = Av \\ p^u - p^v \in I_A \quad &\Leftrightarrow (s_1, \ldots, s_d)Au = (s_1, \ldots, s_d)Av \end{aligned}$$

and this implies $\sum_{i=1}^{n} u_i = \sum_{i=1}^{n} v_i$. Hence all polynomials in I_A are homogenous and thus X_A is a projective toric variety. Now let

$(u_1, \ldots, u_n)$ be the observed data set and $N = \sum_{i=1}^{n} u_i$ the sample size, then the ML problem becomes

$$\max \left(z_1^{a_1}\right)^{u_1} \cdots \left(z_n^{a_n}\right)^{u_n} \quad \text{subject to } z_1^{a_1} + \cdots + z_n^{a_n} = 1 \tag{10}$$

where $(a_1, \ldots, a_n)$ are the columns of A.

PROPOSITION 5.3. *Let $(\hat{z_1}, \ldots, \hat{z_d})$ be a solution of (10) where $\hat{p_i} = \hat{z}^{a_i} \in X_A$, then*

$$A\hat{p} = \frac{1}{N} b \tag{11}$$

where $b = Au$.

Proof. We introduce a Lagrange multiplier λ. The solutions of (10) must satisfy the d conditions

$$\frac{\partial}{\partial z_i} z^b = \lambda\left(\frac{\partial}{\partial z_i}\left(\sum_{j=1}^{n} z^{a_j} - 1\right)\right) \text{ for } i = 1, \ldots, n.$$

Then if we multiply every condition by z_i we obtain

$$b_i z^b = \lambda\left(\sum_{j=1}^{n} a_{ji} z^{a_j} - 1\right) \text{ for } i = 1, \ldots, n.$$

In vector form for the optimal solution we have

$$\hat{z}^b = \lambda A\hat{p}$$
$$Au = b = \overline{\lambda} A\hat{p} \quad \text{and}$$
$$(s_1, \ldots, s_d)Au = \overline{\lambda}(s_1, \ldots, s_d)A\hat{p}$$
$$(1, \ldots, 1)u = \overline{\lambda}(1, \ldots, 1)\hat{p}$$

thus

$$\sum_{i=1}^{n} u_i = N = \overline{\lambda} \sum_{i=1}^{n} \hat{p}_i$$
$$\overline{\lambda} = N.$$

□

The above development implies the following theorem where we let $P_A(b) = \{x \in \mathbb{R}^n \mid Ax = b, \;\; x \geq 0\}$.

THEOREM 5.4. *For a toric model defined by the matrix A given data $u_1, \ldots, u_n$, there is a unique point in $X_A \cap P_A\left(\frac{b}{N}\right)$ and this is the maximum likelihood estimate.*

5.3. Maximum likelihood degree

Recall that the MLE problem is

$$\max_{\boldsymbol{\theta}} \left(g_1^{u_1}(\boldsymbol{\theta}) \cdots g_n^{u_n}(\boldsymbol{\theta}) \right)$$

and using the logarithms

$$\max_{\boldsymbol{\theta}} \left(u_1 g_1(\boldsymbol{\theta}) + \cdots + u_n g_n(\boldsymbol{\theta}) \right).$$

Let $g(\boldsymbol{\theta}) = g_1^{u_1}(\boldsymbol{\theta}) \ldots g_n^{u_n}(\boldsymbol{\theta})$, then the critical equations for this problem are

$$\frac{\partial}{\partial \theta_i} g(\boldsymbol{\theta}) = 0 \; \forall \; i = 1, \ldots, d \tag{12}$$

We know that an MLE is a solution to these critical equations. However, in general the MLE is not unique, and we can even look at the complex solutions for equations (12).

Definition 5.5. *The number of complex solution to the MLE problem for a general data vector u is the* maximum likelihood degree *of the model.*

The following theorem is the main result in [9].

Theorem 5.6. *Let $g_1, \ldots, g_n$ be polynomials of degrees $b_1, \ldots, b_n$ in d unknowns. If the ML degree is finite then it is less than or equal to the coefficient of t^d in the function*

$$\frac{(1-t)^d}{(1-tb_1)\cdots(1-tb_n)}.$$

If the g_i's are sufficiently generic this coefficient is equal to the ML degree.

Example 5.7. *We consider a model with $d = 2$, $n = 4$, where $g_1(\theta_1, \theta_2)$, $\ldots, g_4(\theta_1, \theta_2)$ are quadratic polynomials. Given data u_1, $\ldots, u_4$, the critical equations are*

$$\begin{cases} \dfrac{u_1}{g_1(\boldsymbol{\theta})} \dfrac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_1} + \cdots + \dfrac{u_4}{g_4(\boldsymbol{\theta})} \dfrac{\partial g_4(\boldsymbol{\theta})}{\partial \theta_1} = 0 \\ \dfrac{u_1}{g_1(\boldsymbol{\theta})} \dfrac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_2} + \cdots + \dfrac{u_4}{g_4(\boldsymbol{\theta})} \dfrac{\partial g_4(\boldsymbol{\theta})}{\partial \theta_2} = 0. \end{cases}$$

We want to find ML degree by using Theorem 5.6. In this case $b_1 = b_2 = b_3 = b_4 = 2$ *and*

$$\frac{(1-t)^d}{(1-2t)\cdots(1-2t)} = 1 + 6t + 25t^2 + 88t^3 + \cdots$$

and hence the ML degree is 25. However, we note that for non-generic polynomials g_i *the ML degree can be much lower than 25.*

5.4. Solving likelihood equations

Let the model variety be a projective variety $V \subseteq \mathbb{P}^n$ with coordinates $(p_0 : p_1 : \ldots : p_n)$ and let the probability simplex $\Delta_n = \{(p_0, p_1, \ldots, p_n) \in \mathbb{R}^{n+1} \mid p_i > 0 \ \sum p_i = 1\}$. We use $V_{>0}$ to denote the intersection between V and Δ_n ($V_{>0} = V \cap \Delta_n$). The maximum likelihood problem is to find a point p=$(p_0 : p_1 : \ldots : p_n)$ which best explains data vectors $u_o, \ldots, u_n$. Then we have to solve the following optimization problem

$$\max_p L = \frac{p_0^{u_0} \cdots p_n^{u_n}}{(p_0 + \cdots + p_n)^{u_0 + \cdots + u_n}} \qquad \text{subject to } p_i \in V_{>0}.$$

Let V_{sing} denote the singular locus of the variety V and set $V_{\text{reg}} := V \backslash V_{\text{sing}}$. We will compute critical points in the complex projective variety, in fact on

$$U = V_{\text{reg}} \backslash V(p_0 \cdots p_n \cdot (\sum p_i))$$

Now we assume that V is generated by r homogeneous polynomials $f_1, \ldots, f_r$. Let be

$$J = \begin{pmatrix} p_0 & p_1 & \cdots & p_n \\ p_0 \frac{\partial f_1}{\partial p_0} & p_1 \frac{\partial f_1}{\partial p_1} & \cdots & p_n \frac{\partial f_1}{\partial p_n} \\ \vdots & \vdots & \ddots & \vdots \\ p_n \frac{\partial f_r}{\partial p_n} & p_1 \frac{\partial f_r}{\partial p_1} & \cdots & p_n \frac{\partial f_r}{\partial p_n} \end{pmatrix}.$$

Now we assume that we are given a prime ideal $P \subseteq \mathbb{R}[p_0, \ldots, p_n]$ defining V and a data vector $(u_0, \ldots, u_n)$. Now we write the steps of the algorithm to find the solutions to the ML problem (see [10].

1. Compute $c = \operatorname{codim}(V)$. Let Q be the ideal generated by $c \times c$ minors of the jacobian of P (Q is the ideal of the singular locus of V).

2. Compute the syzygy module M of J over $\mathbb{R}[V] = \mathbb{R}[p_0, \ldots, p_n]/P$.

3. Let I'_u be the ideal in $\mathbb{R}[V]$ generated by polynomials $\sum_{i=0}^{n} u_i \phi_i$ where ϕ_i runs over the minimal generators of M.

4. Compute the saturation ideal[1]
$$I_u = I'_u : ((p_0 \cdots p_n) \cdot (\sum p_i)Q)^\infty.$$

5. Find the solutions to I_u.

6. Check which of those solutions are local maxima.

Observation

- Note that the steps 3,4,5,6 depend on the data while step 1 and step 2 does not.

- The first four steps have as output the likelihood equations. The last two steps have as output the local maxima of the likelihood functions.

- This algorithm has been implemented by Luis Garcia as a SINGULAR code. See `http://www.math.tamu.edu/~lgp`.

References

[1] S. Aoki and A. Takemura, *Minimal basis for connected markov chain over $3 \times 3 \times k$ contingency table with fixed two dimensional marginals*, Tech. report, University of Tokyo, February 2002.

[2] D. Cox, J. Little and D. O'Shea, *Using algebraic geometry*, Springer Verlag, 1996.

[1] Let $I \in k[x_1, \ldots, x_n]$ be an ideal and fix $f \in k[x_1, \ldots, x_n]$. Then the saturation ideal of I with respect to f is $I : f^\infty = \{g \in k[x_1, \ldots, x_n] : \; f^m g \in I$ for some $m > 0\}$.

[3] M. Develin and S. Sullivant, *Markov bases of binary graph models*, Ann. Comb. **7** (2003), 441–466.

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

[5] A. Dobra and S. Sullivant, *A divide-and-conquer algorithm for generating markov bases of multi-way tables*, Comp. Stat. **19** (2004), 347–366.

[6] L.D. Garcia, M. Stillman and B. Sturmfels, *Algebraic geometry of bayesian networks*, math.AG/0301255.

[7] D. Geiger, C. Meek and B. Sturmfels, *On the toric algebra of graphical models*, Ann. Stat., in press.

[8] S. Hoşten and S. Sullivant, *Gröbner bases and polyhedral geometry of reducible and cyclic models*, J. Comb. Theory, Ser. A **100** (2002), 277–301.

[9] A. Khetan, F. Catanese, S. Hoşten and B. Sturmfels, *The maximum likelihood degree*, Amer. J. Math., in press.

[10] A. Khetan, S. Hoşten and B. Sturmfels, *Solving the likelihood equation*, Foundations Comp. Math., in press.

[11] S.L. Lauritzen, *Graphical models*, Claredon Press, Oxford, 1996.

[12] L. Pachter and B. Sturmfels, *Algebraic statistics for computational biology*, Cambridge, 2005.

[13] E. Riccomagno, G. Pistone and H.P. Wynn, *Algebraic statistics*, Chapmann & Hall/CRC, Boca Raton, 2001.

[14] L. Robbiano, *Gröbner bases and statistics*, Gröbner bases and applications (Linz, 1998), London Math. Soc. Lecture Note Ser., vol. 251, Cambridge Univ. Press, Cambridge, 1998, pp. 179–204.

[15] L. Robbiano and M.P. Rogantin, *Full factorial designs and distracted fractions*, Gröbner bases and applications (Linz, 1998), London Math. Soc. Lecture Note Ser., vol. 251, Cambridge Univ. Press, Cambridge, 1998, pp. 473–482.

[16] F. Santos and B. Sturmfels, *Higher lawrence confgurations*, J. Comb. Theory, Ser. A **103** (2003), 151–164.

[17] B. Sturmfels, *John von neumann lectures 2003 at the technical university munchen given by bernd sturmfels, uc berkeley*, Tech. report, 2003.

[18] B. Sturmfels, N. Eriksson, K. Ranestad and S. Sullivant, *Phylogenetic algebraic geometry*, Proceedings of “Projective Varieties with Unexpected Properties”, Siena, Italy, to appear.

Received December 6, 2005.