Some Experiments with Algebraic Statistics

Jerome Le Ny

May 23, 2006

Abstract

In this report, I give a brief introduction to the ideas of algebraic statistics and look at some computational examples involving Gröbner basis techniques. The main motivation was to have a better idea of the scalability of these methods, but as we will see, the results are not very encouraging in the general case. I do not cover more advanced work touching more closely to algebraic geometry and polyhedral combinatorics.

Contents

1 Introduction

The term algebraic statistics was coined by Pistone et al. and used as the title of their book [PRW00]: the idea is to use methods from computational commutative algebra and algebraic geometry to answer questions arising in parametric statistics. One can view a graphical model as the image of a polynomial map from the space of parameters to the space of joint probability distributions on the observed random variables. The topic is now a popular subject of research. Moreover, the questions to be answered already raised new questions in algebraic geometry, in particular in tropical algebraic geometry, and in real algebraic geometry.

This report is just a very basic introduction to the field of algebraic statistics. The recent book [PS05] in particular has served as a reference. After introducing the necessary definitions, I cover some elements of (i) the search for model invariants to test the validity of a given model and (ii) maximum-likelihood estimation of the model parameters. Most of the current applications seem to be currently in computational biology, but my interest lies more in engineering applications. I show some computational examples on hidden Markov models (HMMs), which are extremely important in practice; a classical example is their use for speech recognition [Rab89]. I have tried to focus on the issue of the complexity and scalability of the methods proposed. I will come back to this point in the conclusion of this report.

2 Algebraic Statistical Models for Discrete Data

A *statistical model* is a family of probability distributions on some state space. Here we assume the state space to be finite, and denote it by $[m] = \{1, \ldots, m\}$. Then a probability distribution on $[m]$ is a point in the $(m-1)$-dimensional probability simplex

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

and *a statistical model is a subset of the simplex.* We will write Δ instead of Δ_{m-1} when the underlying state space is understood. Note that a requirement is that the coordinates p_i must be nonnegative real numbers. When we use algebraic computations however, we will allow p_i to be a complex number and deal with the real positivity independently. We denote the polynomial ring $\mathbb{C}[p] = \mathbb{C}[p_1, \ldots, p_n]$, where the variables are the state probabilities.

We restrict the family of statistical models considered to *algebraic statistical models.* They arise as the image of a polynomial map

$$\mathbf{f} \ : \mathbb{R}^d \to \mathbb{R}^m, \ \theta = (\theta_1, \ldots, \theta_d) \mapsto (f_1(\theta), \ldots, f_m(\theta)).$$

The unknowns $\theta_1, \ldots, \theta_d$ represent the model parameters, and in most cases of interest, d is much smaller than m.

Let $\mathbb{N} = \{0, 1, 2, \ldots\}$ denote the non-negative integers. Each f_i is a polynomial in the d unknowns, which means it has the the form

$$f_i(\theta) = \sum_{\alpha \in \mathbb{N}^d} c_\alpha \, \theta_1^{\alpha_1} \cdots \theta_d^{\alpha_d},$$

where only finitely many of the coefficients c_α are non-zero. We assume that the parameter vector θ ranges over an open subset $\Theta \subset \mathbb{R}^d$, which is called the *parameter space* of the model $\mathbf{f}$. We also assume that the parameter space satisfy the condition

$$f_i(\theta) > 0 \quad \text{for all } i \in [m] \text{ and } \theta \in \Theta.$$

Under these hypotheses, the following two conditions are equivalent:

$$\mathbf{f}(\Theta) \subset \Delta \Leftrightarrow f_1(\theta) + \ldots + f_m(\theta) = 1, \tag{1}$$

which means that all non-constant terms of the polynomials f_i cancel and the constant terms add up to 1. If (1) holds, then the model is simply the set $\mathbf{f}(\Theta)$. More generally, the model is the family of all probability distributions on $[m]$ of the form

$$\frac{1}{\sum_{i=1}^m f_i(\theta)} \left(f_1(\theta), \ldots, f_m(\theta)\right), \quad \text{where } \theta \in \Theta,$$

i.e., it is the image of a set of rational functions.

We now introduce two classes of algebraic models, which turn out to be important in applications.

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 polynomial $f_i(\theta)$ is a linear function, i.e., there exist vectors $a_i \in \mathbb{R}^d$ and real numbers b_i such that

$$f_i(\theta) = a_i^T \theta + b_i = \sum_{j=1}^d a_{ij}\theta_j + b_i, \quad \forall i \in \{1, \ldots, d\}.$$

For this model, it is convenient to take the m linear functions $f_1(\theta), \ldots, f_m(\theta)$ such that their sum is the constant function 1.

2.2 Log-Linear or Toric Models

The second class of models that we introduce are the toric models, or in more standard statistical terms log-linear models. They are important for studying of Markov chains.

Let $A = (a_{ij})$ be a non-negative integer $d \times m$ matrix with the property that all column sums are equal:

$$\sum_{i=1}^d a_{i1} = \ldots = \sum_{i=1}^d a_{im}. \tag{2}$$

The jth column vector a_j of the matrix A represents the monomial

$$\theta^{a_j} = \prod_{i=1}^d \theta_i^{a_{ij}}, \quad j = 1, \ldots, m.$$

Note that with assumption (2) all these monomials have the same degree. The toric model of A is the image of the positive orthant $\Theta = \mathbb{R}^d_{>0}$ under the map

$$\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m, \quad \theta \mapsto \frac{1}{\sum_{j=1}^m c_j \theta^{a_j}} \, (c_1\theta^{a_1}, \ldots, c_m\theta^{a_m}),$$

with $c_1, \ldots, c_m$ positive constants. We will only consider the case where $c_i = 1$ for all i. Note that we can scale the parameter vector without changing the image, i.e., $\mathbf{f}(\theta) = \mathbf{f}(\lambda\theta)$, and so the dimension of the toric model $\mathbf{f}(\mathbb{R}^d_{>0})$ is at most $d-1$.

The name log-linear model comes from the fact that logarithms of the probabilities are linear functions in the logarithms of the parameters θ_i. The adjective "toric" is used because the image of the map $\mathbf{f}$ is a toric variety.

Example 2.1 (Independence Model)**.** Let X_1 be a random variable on $[m_1]$, and X_2 a random variable on $[m_2]$. Then the state space for the random vector $X = (X_1, X_2)$ is of size $m = m_1 m_2$. Assume the two random variables are independent and denote

$$\begin{aligned} p_{ij} &= P(X_1 = i, X_2 = j) = P(X_1 = i)P(X_2 = j) \\ &= \underbrace{\left(\sum_{k=1}^{m_2} p_{ik}\right)}_{\theta_i} \underbrace{\left(\sum_{l=1}^{m_1} p_{lj}\right)}_{\theta_{j+m_1}}, \quad \text{for all } i \in [m_1], j \in [m_2]. \end{aligned}$$

The independence model is a toric model with $m = m_1 m_2$ and $d = m_1 + m_2$. The polynomial map is

$$\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m, \quad \theta \mapsto \frac{1}{\sum_{i,j} \theta_i \theta_{j+m_1}} \, (\theta_i \theta_{j+m_1})_{i \in [m_1], j \in [m_2]}.$$

A point $p \in \Delta_{m-1}$ lies in the image of $\mathbf{f}$ if and only if X_1 and X_2 are independent, i.e., if and only if the $m_1 \times m_2$ matrix (p_{ij}) has rank one. The matrix A has entries in $\{0, 1\}$ and exactly two 1's per column.

For instance, with $m_1 = 2, m_2 = 3$, the matrix $A =$

$$\begin{array}{c|cccccc} & p_{11} & p_{12} & p_{13} & p_{21} & p_{22} & p_{23} \\ \hline \theta_1 & 1 & 1 & 1 & 0 & 0 & 0 \\ \theta_2 & 0 & 0 & 0 & 1 & 1 & 1 \\ \theta_3 & 1 & 0 & 0 & 1 & 0 & 0 \\ \theta_4 & 0 & 1 & 0 & 0 & 1 & 0 \\ \theta_5 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}$$

encodes the rational map $\mathbf{f} : \mathbb{R}^5 \to \mathbb{R}^{2\times 3}$ given by

$$(\theta_1, \theta_2; \theta_3, \theta_4, \theta_5) \mapsto \frac{1}{(\theta_1 + \theta_2)(\theta_3 + \theta_4 + \theta_5)} \begin{pmatrix} \theta_1\theta_3 & \theta_1\theta_4 & \theta_1\theta_5 \\ \theta_2\theta_3 & \theta_2\theta_4 & \theta_2\theta_5 \end{pmatrix}.$$

$\mathbf{f}(\mathbb{R}^5_{>0})$ consists of all positive 2×3 matrices of rank 1 whose entries sum to 1.

2.3 Markov Models

2.3.1 Toric Markov Chains

We fix an alphabet Σ with l letters, and we fix a positive integer n. We shall define a toric model whose state space is the set Σ^n of all words of length n. The model is parametrized by the set Θ of positive $l \times l$

matrices. Thus $d = l^2$ and $m = l^n$. The $d \times m$ matrix A with integer entries will be denoted by $A_{l,n}$. Its rows are indexed by Σ^2, its columns by Σ^n. The entry of $A_{l,n}$ in the row indexed by the pair $\sigma_1\sigma_2 \in \Sigma^2$ and the column indexed by the word $\pi_1\pi_2 \ldots \pi_n \in \Sigma^n$ is the number of occurrences of the pair inside the word, i.e., the number of indices $i \in \{1, \ldots, n-1\}$ such that $\sigma_1\sigma_2 = \pi_i\pi_{i+1}$. The *toric Markov chain model* is the toric model specified by the matrix $A_{l,n}$.

For example, let us consider words of length $n = 4$ over the binary alphabet $\Sigma = \{0, 1\}$ so that $l = 2, d = 4, m = 16$. The matrix $A_{2,4}$ is the following 4×16 matrix:

$$\begin{array}{c|cccccccccccccccc|}
 & 0000 & 0001 & 0010 & 0011 & 0100 & 0101 & 0110 & 0111 & 1000 & 1001 & 1010 & 1011 & 1100 & 1101 & 1110 & 1111 \\
00 & 3 & 2 & 1 & 1 & 1 & 0 & 0 & 0 & 2 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
01 & 0 & 1 & 1 & 1 & 1 & 2 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\
10 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 2 & 1 & 1 & 1 & 1 & 0 \\
11 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 2 & 0 & 0 & 0 & 1 & 1 & 1 & 2 & 3
\end{array}$$

The parameters can be organized in a 2×2 matrix

$$\theta = \begin{pmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{pmatrix}$$

and the parameter space $\Theta = \mathbb{R}_{>0}^{2\times 2}$. The toric Markov chain model of length $n = 4$ for the binary alphabet is the image of Θ under the monomial map

$$\mathbf{f}_{2,4} : \mathbb{R}^{2\times 2} \to \mathbb{R}^{16}, \quad \theta \mapsto \frac{1}{\sum_{ijkl} p_{ijkl}} (p_{0000}, p_{0001}, \ldots, p_{1111}),$$
$$\text{where} \quad p_{i_1i_2i_3i_4} = \theta_{i_1i_2}\theta_{i_2i_3}\theta_{i_3i_4} \quad \text{for all } i_1i_2i_3i_4 \in \{0,1\}^4.$$

The map $\mathbf{f}_{l,n} : \mathbb{R}^d \to \mathbb{R}^m$ is defined analogously for larger alphabets and longer sequences.

2.3.2 Markov Chains

The Markov chain model is a submodel of the toric Markov chain model. Let Θ_1 denote the subset of all matrices $\theta \in \mathbb{R}_{>0}^{l\times l}$ whose rows sum to one. The Markov chain model is the image of Θ_1 under the map $\mathbf{f}_{l,n}$. A Markov chain is any point p in the model $\mathbf{f}_{l,n}(\Theta_1)$. An entry θ_{ij} of the matrix θ represents the probability of transitioning from state $i \in \Sigma$ to $j \in \Sigma$. This definition agrees with the familiar decription of a Markov chain in terms of its transition probabilities, except that here we require the initial distribution at the first state to be uniform.

To define a *fully observed Markov model*, we fix the sequence length n and we consider a first alphabet Σ with l letters and a second alphabet Σ' with l' letters. The observable states in this model are pairs $(\sigma, \tau) \in \Sigma^n \times (\Sigma')^n$ of words of length n. A sequence of N observations in this model is summarized in a matrix $u \in \mathbb{N}^{l^n \times (l')^n}$ where $u_{(\sigma,\tau)}$ is the number of times the pair (σ, τ) was observed. Hence in this model $m = (l\,l')^n$. The fully observed Markov model is parametrized by a pair of matrices (θ, θ') where θ is an $l \times l$ matrix and θ' is an $l \times l'$ matrix; as before, these matrices have rows which sum to one. The matrix

θ encodes a Markov chain as before. The entry θ'_{ij} represents the probability of outputting symbol $j \in \Sigma'$ when the Markov chain is in state $i \in \Sigma$. With the constraints on the sum of the rows of (θ, θ'), we have $d = l(l + l' - 2)$.

Let Θ_1 denote the set of matrices $(\theta, \theta') \in \mathbb{R}^{l\times l}_{>0} \times \mathbb{R}^{l\times l'}_{>0}$ whose rows sum to one. The fully observed Markov model is the restriction to Θ_1 of the toric model

$$\begin{aligned}&F : \mathbb{R}^d \to \mathbb{R}^m, \quad (\theta, \theta') \mapsto p = (p_{\sigma,\tau})\\ &\text{where} \quad p_{\sigma,\tau} = \frac{1}{l}\theta'_{\sigma_1\tau_1}\theta_{\sigma_1\sigma_2}\theta'_{\sigma_2\tau_2}\theta_{\sigma_2\sigma_3}\dots\theta_{\sigma_{n-1}\sigma_n}\theta'_{\sigma_n\tau_n}.\end{aligned}$$

2.3.3 Hidden Markov Models

Finally, the *hidden Markov model* (HMM) $\mathbf{f}$ is derived from the fully observed Markov model F by summing out the first indices $\sigma \in \Sigma^n$. More precisely, consider the map

$$\rho : \mathbb{R}^{l^n \times (l')^n} \to \mathbb{R}^{(l')^n}$$

obtained by taking the column sums of the matrix with l^n rows and $(l')^n$ columns. The hidden Markov model is the algebraic statistical model defined by composing the fully observed Markov model F with the marginalization map ρ:

$$\mathbf{f} = \rho \circ F : \Theta_1 \subset \mathbb{R}^{l(l-1)} \times \mathbb{R}^{l(l'-1)} \to \mathbb{R}^{(l')^n}.$$

The degree of $\mathbf{f}$ in the entries of θ is $n - 1$ and in the entries of θ' is n.

A more standard way of describing an HMM is to say that this model has n observed variables $Y_1, \dots Y_n$ taking on l' possible values and n hidden variables $X_1, \dots, X_n$ taking on l possible values. The HMM is then characterized by the following conditional independence statements for $i = 1, \dots, n$:

$$\begin{aligned}P(X_i|X_1, X_2, \dots, X_{i-1}) &= P(X_i|X_{i-1}),\\ P(Y_i|X_1, \dots, X_i, Y_1, \dots, Y_{i-1}) &= P(Y_i|X_i).\end{aligned}$$

These transition probabilities are given by the matrices θ and θ'.

Example 2.2. Consider the classical example of the "occasionally dishonest casino". In that casino, they use a fair die most of the time, but occasionally they switch to a loaded die. Our two alphabets are $\Sigma = \{\text{fair,loaded}\}$ and $\Sigma' = \{1, 2, 3, 4, 5, 6\}$ for the six possible outcomes of rolling a die. Suppose a particular game involves rolling the die $n = 4$ times. This hidden Markov model has $d = 12$ parameters:

$$\theta = \begin{array}{c|cc} & \text{fair} & \text{loaded} \\ \text{fair} & x & 1-x \\ \text{loaded} & 1-y & y \end{array} \quad \text{and} \quad \theta' = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \text{fair} & f_1 & f_2 & f_3 & f_4 & f_5 & 1-\sum_{i=1}^5 f_i \\ \text{loaded} & l_1 & l_2 & l_3 & l_4 & l_5 & 1-\sum_{i=1}^5 l_i \end{array}.$$

This hidden Markov model has $m = 6^4 = 1296$ possible outcomes, namely, all the words $\tau = \tau_1\tau_2\tau_3\tau_4 \in (\Sigma')^4$. The coordinates of the map $\mathbf{f} : \mathbb{R}^{12} \to \mathbb{R}^{1296}$ are polynomials of degree $7 = 3+4$ (degree 3 in the two unknowns x and y, degree 4 in the ten unknowns $f_1, \dots, l_5$):

$$p_{\tau_1\tau_2\tau_3\tau_4} = \frac{1}{2}\sum_{\sigma_1,\sigma_2,\sigma_3,\sigma_4\in\Sigma} \theta'_{\sigma_1\tau_1}\theta_{\sigma_1\sigma_2}\theta'_{\sigma_2\tau_2}\theta_{\sigma_2\sigma_3}\theta'_{\sigma_3\tau_3}\theta_{\sigma_3\sigma_4}\theta'_{\sigma_4\tau_4}.$$

We finish this section by a remark on an apparent obvious limitation of the algebraic point of view: even if the number of parameters of the Markov model is limited, the number of variables in the polynomial rings involved grows exponentially with the length n of the chain considered. Considering for example a simple isolated word recognition system as described in [Rab89], we can have $l' = 5, n = 40$, and hence a polynomial ring with 5^{40} variables p_i's, which is not representable on any computer explicitely. As we will see in the next section, Gröbner basis computations are used relatively intensively in algebraic statistics, which means that we would have to restrict computations to models far smaller than those of interest in engineering applications. At this point, it is not clear to me if such problems do not arise in the current applications of algebraic statistics in computational biology, or if the problem is recognized and methods are developped with this in mind. This would seem natural, since most of the existing algorithms for graphical models are designed with a perspective on the possible computational explosion if the structure is not properly exploited, a classical example being the Viterbi algorithm.

3 Computing Polynomial Invariants

The polynomial functions that vanish on the image of an algebraic statistical model are called *invariants of the model*. There is an issue here because in general, the image of a polynomial map $\mathbf{f} : \mathbb{C}^d \to \mathbb{C}^m$ is not an algebraic variety. Therefore, the convention is to use the expression "image of the polynomial" map to mean the the Zariski closure $\overline{\mathbf{f}(\mathbb{C}^d)}$ of the image in $\mathbb{C}^m$, which is the smallest variety containing $\mathbf{f}(\mathbb{C}^d)$ ([CLO97], chapter 3). The potential points $p \in \overline{\mathbf{f}(\mathbb{C}^d)} \setminus \mathbf{f}(\mathbb{C}^d)$ are then disregarded. In the following, we will take the polynomials with rational coefficients. Let $I_{\mathbf{f}} \subset \mathbb{Q}[p_1, \ldots, p_m]$ be the ideal representing the variety $\overline{\mathbf{f}(\mathbb{C}^d)}$. A polynomial $h \in \mathbb{Q}[p_1, \ldots, p_m]$ lies in the ideal $I_{\mathbf{f}}$ is an only if

$$h(f_1(t), \ldots, f_m(t)) = 0 \quad \text{for all } t = (t_1, \ldots, t_d) \in \mathbb{R}^d,$$

where we can replace $\mathbb{R}^d$ by any open subset $\Theta \subset \mathbb{R}^d$ (our parameter space) and obtain an equivalent condition. The ideal $I_{\mathbf{f}}$ is prime, and the polynomials in this ideal are called the model invariants.

By plugging the empirical frequencies of the observed data into these invariants (or some of them if we don't know all of them), and observing if the result is close to zero, it can be checked whether the model is appropriate. To compute the invariants, i.e., generators of $I_{\mathbf{f}}$, we need to eliminate the parameters of the equations defining the polynomial map, which is the *implicitization problem*. One way to do this is by using Gröbner basis, but as we will see in the following example, this is possible only for small models (another popular method is to use multipolynomial resultants). The method using Gröbner basis is based on the following theorem ([CLO97], p. 126):

Theorem 3.1. *(Polynomial Implicitization) Let k be an infinite field, and let $F : k^n \to k^m$ be a polynomial map $F(t_1, \ldots, t_m) = (f_1, (t_1, \ldots, t_m), \ldots, f_n(t_1, \ldots, t_m))$. Let I be the ideal $I = \langle x_1 - f_1, \ldots, x_n - f_n \rangle \subset k[t_1, \ldots, t_m, x_1, \ldots, x_n]$ and let $I_m = I \cap k[x_1, \ldots, x_n]$ be the mth elimination ideal. Then $V(I_m)$ is the smallest variety in k^n containing $F(k^m)$.*

This gives an algorithm for the implicitization problem: we start by finding a Gröbner basis with respect to an elimination ordering where every t_i is greater than every x_i. By the elimination theorem ([CLO97], p.113), the elements of the Gröbner basis not involving $t_1, \dots, t_m$ form a basis of I_m, and by the theorem above, they define the smallest variety in k^n containing the parametrization.

Example 3.1 ([PS04]). Consider an HMM with $n = 3$ and binary random variables ($l = l' = 2$). It is defined by the map given by the equations (we scaled the 8 joint probabilities by a factor of 2)

$$\begin{aligned} p_{\tau_1\tau_2\tau_3} = \theta'_{0\tau_1}\theta_{00}\theta'_{0\tau_2}\theta_{00}\theta'_{0\tau_3} + \theta'_{0\tau_1}\theta_{00}\theta'_{0\tau_2}\theta_{01}\theta'_{1\tau_3} + \theta'_{0\tau_1}\theta_{01}\theta'_{1\tau_2}\theta_{10}\theta'_{0\tau_3} + \theta'_{0\tau_1}\theta_{01}\theta'_{1\tau_2}\theta_{11}\theta'_{1\tau_3} \\ +\theta'_{1\tau_1}\theta_{10}\theta'_{0\tau_2}\theta_{00}\theta'_{0\tau_3} + \theta'_{1\tau_1}\theta_{10}\theta'_{0\tau_2}\theta_{01}\theta'_{1\tau_3} + \theta'_{1\tau_1}\theta_{11}\theta'_{1\tau_2}\theta_{10}\theta'_{0\tau_3} + \theta'_{1\tau_1}\theta_{11}\theta'_{1\tau_2}\theta_{11}\theta'_{1\tau_3}, \quad \tau_1\tau_2\tau_3 \in \{0,1\}^3, \end{aligned}$$

We can try to compute the invariants of this HMM. The equations above (the $x_i's$ in the theorem are now the $p_{\tau_1\tau_2\tau_3}$) in the 16 variables (the $p_{\tau_1\tau_2\tau_3}$, $\theta_{ij}, \theta'_{ij}$, without taking into account the constraints that the sum of the rows of the transition matrix sums to one) define the ideal I. We can try to eliminate the variables θ_{ij} and θ'_{ij} as described in the algorithm above. For this purpose, we can use Singular [GP02] to compute the reduced Gröbner basis of I with the command `std(I)`, once we have fixed an elimination ordering. It is known that the lexicographic ordering is usually expensive. Example 1.8.4 in [GP02] uses the product ordering of two degrevlex orderings `(dp(s),dp(n-s))` to eliminate the first s variables. Singular also provides a special command `eliminate`. In any case, on a relatively recent desktop, with 1 GB of RAM, the computation runs out of memory after running more than 20 hours... In their paper [PS04], Pachter and Sturmfels simply state that the image of the map was found by Gröbner basis computation to be the zero set of the single quartic polynomial

$$\begin{aligned} & p_{011}^2 p_{100}^2 - p_{001}^2 p_{110}^2 + p_{000}p_{011}p_{101}^2 - p_{000}p_{101}^2 p_{110} + p_{000}p_{011}p_{110}^2 - p_{001}p_{010}^2 p_{111} + p_{001}^2 p_{100}p_{111} \\ & + p_{010}^2 p_{100}p_{111} - p_{000}p_{100}^2 p_{111} - p_{000}p_{011}^2 p_{110} - p_{001}p_{011}p_{100}p_{101} - p_{010}p_{011}p_{100}p_{101} \\ & + p_{001}p_{010}p_{011}p_{110} - p_{010}p_{011}p_{100}p_{110} + p_{001}p_{010}p_{101}p_{110} + p_{001}p_{100}p_{101}p_{110} + p_{000}p_{010}p_{011}p_{111} \\ & - p_{000}p_{011}p_{100}p_{111} - p_{000}p_{001}p_{101}p_{111} + p_{000}p_{100}p_{101}p_{111} + p_{000}p_{001}p_{110}p_{111} - p_{000}p_{010}p_{110}p_{111}. \end{aligned} \tag{3}$$

What is not mentioned in that paper is that the computation apparently took about seven hours on a dual 2.8 GHz, 4 GB RAM machine running Singular, and was actually the largest model which could be handled by a direct application of the Gröbner basis method ([PS05], chapter 11). The Gröbner basis computation is very sensitive to the number of variables. If now I add explicitely the constraints that $\theta_{i1} = 1 - \theta_{i0}$ and $\theta'_{i1} = 1 - \theta'_{i0}$, I reduce the number of variables to 12, 4 of which have to eliminated, and now the computation is almost instantaneous: the reduced Gröbner basis of the image variety for degrevlex has now 14 polynomials (the result is too long to be reproduced here). It seems that the reason this was not done in [PS04] was to obtain an ideal of invariants that can be described more simply in terms of just the polynomial (3).

It appears that the extension to longer chains ($n \geq 4$) of the computation of invariants using direct Gröbner basis computation becomes quickly prohibitive. There are conjectures about the structure of the ideal of invariants of Markov models (see for example conjecture 13 in [PS04]), but in general the problem is hard and the ideal may remain unknown.

3.1 The Implicitization Problem Using Linear Algebra

In [PS05] (chapter 11), it is conjectured that the ideal of a binary HMM of any length is generated by polynomials of low degree. Then, by limiting the search to those generators of the ideal which have degree less than some bound, we may consider the implicitization problem as a linear algebra problem.

We discuss this point first for a general algebraic statistical model. We consider a polynomial map $\mathbf{f} : \mathbb{C}[\theta_1, \dots, \theta_d] \to \mathbb{C}[p_1, \dots, p_m]$. We let $I_{\mathbf{f}}$ be the ideal of polynomial invariants associated to $\mathbf{f}$, and $I_{\mathbf{f},\delta}$ this ideal restricted to polynomials of degree at most δ. Consider first polynomial relations between the p_i's of degree at most 1. We can expand $1, p_0, p_1, \dots, p_m$, in terms of the parameters $\theta_j, j = 1, \dots, d$; denote $f_0 = 1, f_i, i = 1, \dots, m$, these expansions. Let $\mathcal{M} = \{m_i\}_{i=1}^k$ be the vector of all monomials in the θ_j occuring in the f_i's, we have $f_i = \sum_j \beta_{ij} m_j$ for each i. Then an invariant of degree at most 1 is a linear relation

$$\sum_{i=1}^m \alpha_i \left(\sum_{j=1}^k \beta_{ij} m_j \right) = \sum_{j=1}^k \left(\sum_{i=1}^m \beta_{ij} \alpha_i \right) m_i = 0.$$

This polynomial equals the zero polynomial if and only if the coefficients of each monomial is zero. Hence generators for $I_{\mathbf{f},1}$ can be found by computing a linear basis for the kernel of the matrix $B = (\beta_{ij})$.

This method generalizes to $I_{\mathbf{f},\delta}$ for higher δ in a straightfoward manner. We consider $\mathcal{P}$, the vector of all monomials in the unknowns $p_1, \dots, p_m$ of degree at most δ, and let $\mathcal{M}$ be the vector consisting of all monomials in θ_j appearing in the expansions of the monomials in $\mathcal{P}$. Then as in the case $\delta = 1$ above, computing the relations of degree at most δ becomes computing the kernel of a large matrix whose (i, j)th entry is the coefficient of $\mathcal{M}_j$ in the expansion of $\mathcal{P}_i$.

The software written by Nicolas Bray and Jason Morton [BM05] is build on refinements of this principle. The first refinement is that when computing the generators for increasing δ, we may eliminate from $\mathcal{P}$ at a given step the monomials p^α which lie in the initial ideal generated by the generators already computed at a previous step. Additionally, there is an interesting usage of the trivial invariant $1 - \sum_{i=1}^m p_i$:

Proposition 3.2. *Suppose* $\mathbf{f}$ *is an algebraic statistical model, and* $I_{\mathbf{f}}$ *is its ideal of invariants. Then there exists a set* $\mathcal{L}$ *of homogeneous polynomials in the* p_i *such that* $\{1 - \sum_{i=1}^m p_i\} \cup \mathcal{L}$ *is a basis for* $I_{\mathbf{f}}$*.*

Proof. Let $\mathcal{B}$ be a finite basis for $I_{\mathbf{f}}$ (which exist by Hilbert's basis theorem). Take $g \in \mathcal{B}$ a non-homogeneous polynomial (if no such g exist we are done), and let δ be the smallest degree of a monomial in g. Let g_δ be the degree δ part of g. Since $1 - \sum_i p_i \in I_{\mathbf{f}}$, so is $(1 - \sum_i p_i) g_\delta$. Then we can replace $g \in \mathcal{B}$ with $g - (1 - \sum_i p_i) g_\delta$ to obtain $\mathcal{B}'$ such that $\mathcal{B}' \cup \{1 - \sum_{i=1}^m p_i\}$ still generates $I_{\mathbf{f}}$. Now the minimum degree of a monomial occuring in g has increased by at least one. Repeating this finitely many times, we have the required $\mathcal{L}$. □

Using proposition 3.2, we may restrict the search for invariants to homogeneous polynomials. The refinements above are valid for any algebraic model. Bray and Morton develop more involved refinements specific to HMMs in chapter 11 of [PS05]. Using their software [BM05], Bray and Morton report that they could compute the Gröbner basis (3) in less than one minute instead of the seven hours reported in example 3.1. However, I was not able to compile their software and test more complex models. Without a result bounding the degree of these generators, we cannot be sure that we obtain all the invariants describing the ideal $I_{\mathbf{f}}$; however, even if we obtain only a subset of all the invariants, the method is of interest if it is scalable. Unfortunately, this seems to be still far from any real-world engineering application, since the authors report that the program took 159 minutes to for an HMM with binary variables of length 5, providing 249 invariants for the model. The result for length 6 was better (14 minutes), returning 692 invariants. For length 7, only 40 invariants of degree 1 were found, and no other invariant of degree less than or equal to 4 (which seems to be the maximum degree tested): in that case, the computation took only $17s$.

Clearly, in applications we are probably not interested in obtaining hundreds of polynomials to test our model. It would be interesting to understand what it means for the data to satisfy a certain number of invariants (and what is the degree of these invariants), in terms of reliability of the model. Then if we can restrict the computation to finding a small number of interesting polynomials, the method has chances to be scalable.

Moreover, if we are interested only in obtaining some elements in the ideal obtained after elimination of the variables θ_i, it is possible that methods based on resultants might be more interesting than methods based on Gröbner basis, which describe the ideal entirely. For example, if $f, g \in \mathbb{C}[x, y]$, then we know that $\mathrm{Res}(f, g, x)$ belongs to the first elimination ideal $< f, g > \cap \mathbb{C}[y]$ ([CLO97], section 3.5 proposition 9). Now in the elimination procedure, we have more than two polynomials and one variable to eliminate so we would have to introduce multipolynomial resultants. One way of doing this is explained in [CLO97], section 3.6, in terms of the generalized resultants of a set of polynomials, but there is much more theory related to multipolynomial resultants, for example in [GKZ94], which might be useful in this context.

4 Maximum Likelihood Estimation

Another important applications of algebraic statistics is in solving the likelihood equations (also called "maximum" likelihood equations in the litterature, but here I follow Dudley [Dud]): for algebraic statistical models, this boils down to solving polynomial equations. First we recall some basic definitions.

Consider a family of laws $\{P_\theta, \theta \in \Theta\}$ on a measurable space $(X, \mathcal{B})$, all absolutely continuous with respect to a σ-finite measure μ (which will be the counting measure for us since we consider only discrete models). Then there exist Radon-Nikodym derivatives $dP/d\mu$, which give us what we call *likelihood functions* $f(\theta, x) := (dP_\theta/d\mu)(x), \theta \in \Theta, x \in X$. For each $x \in X$, a *maximum likelihood estimate* (MLE) of θ is any $\hat{\theta} = \hat{\theta}(x)$ such that $f(\hat{\theta}, x) = \sup\{f(\phi, x) : \phi \in \Theta\} > 0$.

Specialized to the algebraic models above, x was called i and $X = \{1, \dots, m\} = [m]$, the set of possible outcomes, so that the likelihood functions are precisely the $f_i(\theta)$. Now we can consider these functions as corresponding to a single experiment, and repeating the experiment N times independently to obtain a sequence of i.i.d. observations $x = (O_1, \dots, O_N)$. The corresponding likelihood functions (for the vector of observations), which we denote $L(\theta, x)$ to avoid confusion, becomes:

$$L(\theta, x) = f_{O_1}(\theta) \dots f_{O_N}(\theta) = f_1(\theta)^{u_1} \dots f_m(\theta)^{u_m},$$

where u_k is the number of indices $j \in [N]$ such that $O_j = k$ (note here that from the factorization theorem, it is clear that $(u_1, \dots, u_m)$ is a sufficient statistic for the model $\mathbf{f}$). In practice it is convenient to consider the *log-likelihood function*

$$l(\theta, x) = \log L(\theta) = u_1 \log(f_1(\theta)) + \dots + u_m \log(f_m(\theta)),$$

which is a function of the parameter space $\Theta \subset \mathbb{R}^d$ to the negative real numbers $\mathbb{R}_{<0}$. In the following, since we think of x as being fixed, we will drop it and write $L(\theta)$, $l(\theta)$.

Since we consider polynomial likelihood functions, a *necessary* condition for θ to be an MLE is that it satisfies the *likelihood equations*

$$\frac{\partial L(\theta)}{\partial \theta_j} = 0, \quad \text{for } j = 1, \ldots, d,$$

or equivalently

$$\frac{\partial l(\theta)}{\partial \theta_j} = 0, \quad \text{for } j = 1, \ldots, d.$$

There are lots of potential difficulties with maximum likelihood estimates, such as non-existence (even in an exponential family), likelihood equations corresponding to a minimum, a local maximum or a saddle point of the likelihood function, but we will not consider these here (see chapter 3 of [Dud] for some examples). Indeed, solving the likelihood equations can be in itself a challenging problem, and we focus on some associated computational issues in the following, in the case of algebraic models.

Example 4.1 (linear models)**.** Recall that for a linear model $\mathbf{f}$, we have $f_i(\theta) = \sum_{j=1}^d a_{ij}\theta_j + b_i$. The (log-)likelihood equations are

$$\sum_{i=1}^m \frac{u_i a_{i1}}{f_i(\theta)} = \ldots = \sum_{i=1}^m \frac{u_i a_{id}}{f_i(\theta)} = 0.$$

Studying these equations involves the combinatorial theory of hyperplane arrangements. Each equation $f_i(\theta) = 0$ defines a hyperplane, and the set

$$\mathcal{C} = \{\theta \in \mathbb{R}^d : f_1(\theta) \cdots f_m(\theta) \neq 0\}$$

is a disjoint union of finitely many open convex polyhedra defined by inequalities $f_i(\theta) > 0$ or $f_i(\theta) < 0$. It turns out that the natural parameter space of the linear model (where the probabilities are nonnegative) coincides with exactly one bounded region. Results in this theory, such as Varchenko's formula, have consequences on the characterization of the solutions of the likelihood equations [CHS06].

4.1 Likelihood Equations and Rational Implicitization

In the case of an algebraic model $\mathbf{f} = (f_1, \ldots, f_m)$, the log-likelihood equations can be rewritten, for a vector of observations $u = (u_1, \ldots, u_m)$:

$$\frac{\partial l}{\partial \theta_i} = \frac{u_1}{f_1(\theta)} \frac{\partial f_1(\theta)}{\partial \theta_i} + \ldots + \frac{u_m}{f_m(\theta)} \frac{\partial f_m(\theta)}{\partial \theta_i}, \quad i = 1, \ldots, d. \tag{4}$$

We would like to compute all solutions $\theta \in \mathbb{C}^d$ of these equations. Note that each function in (4) is a rational function, and so the set of critical points is an algebraic variety outside the locus where the denominators of these rational functions are zero. The Zariski closure of the set of critical points is an algebraic variety in $\mathbb{C}^d$, called the *likelihood variety* of the model $\mathbf{f}$ with respect to the data u. Then the problem becomes to compute this likelihood variety, and this can be done via elimination theory once again.

We can follow the rational implicitization procedure similar to the one described in [CLO97], section 3.3, which requires a slight modification to the algorithm we considered previously for polynomial implicitization.

We introduce m new unknowns $z_1, \ldots, z_m$ where z_i represents the inverse of $f_i(\theta)$, and an ideal generated by $m + d$ polynomials in the ring $\mathbb{Q}[\theta, z]$:

$$J = \langle z_1 f_1(\theta) - 1, \ldots, z_m f_m(\theta) - 1, \sum_{j=1}^{m} u_j z_j \frac{\partial f_j}{\partial \theta_1}, \ldots, \sum_{j=1}^{m} u_j z_j \frac{\partial f_j}{\partial \theta_d} \rangle.$$

A point $(\theta, z) \in \mathbb{C}^{d+m}$ lies in the variety $V(J)$ of this ideal if and only if θ is a solution of the likelihood equations with $f_j(\theta) \neq 0$ and $z_j = 1/f_j(\theta)$ for all j. To solve our problem, we need to compute the elimination ideal

$$I = J \cap \mathbb{Q}[\theta_1, \ldots, \theta_d].$$

This can be done as before by using an eliminationg ordering where $z_i > \theta_j$, computing a Gröbner basis with respect to this ordering for J, and keeping only the elements of the basis not involving the variables z_i. Finally we keep only the values of the parameters which have their image by $\mathbf{f}$ in the probability simplex. Also, to verify that a given parameter corresponds indeed to a maximum, we would have to compute the Hessian matrix $\partial^2 l/\partial\theta^2$.

Example 4.2 (Jukes-Cantor model)**.** The Jukes-Cantor model is a class of multilinear models which apparently arises in computational biology. One instance given in example 1.7 in [PS05] is the following:

$$\begin{aligned}
f_1(\theta) &= -24\theta_1\theta_2\theta_3 + 9\theta_1\theta_2 + 9\theta_1\theta_3 + 9\theta_2\theta_3 - 3\theta_1 - 3\theta_2 - 3\theta_3 + 1,\\
f_2(\theta) &= -48\theta_1\theta_2\theta_3 + 6\theta_1\theta_2 + 6\theta_1\theta_3 + 6\theta_2\theta_3,\\
f_3(\theta) &= 24\theta_1\theta_2\theta_3 + 3\theta_1\theta_2 - 9\theta_1\theta_3 - 9\theta_2\theta_3 + 3\theta_3,\\
f_4(\theta) &= 24\theta_1\theta_2\theta_3 - 9\theta_1\theta_2 + 3\theta_1\theta_3 - 9\theta_2\theta_3 + 3\theta_2,\\
f_5(\theta) &= 24\theta_1\theta_2\theta_3 - 9\theta_1\theta_2 - 9\theta_1\theta_3 + 3\theta_2\theta_3 + 3\theta_1.
\end{aligned}$$

Suppose we fix $\theta_3 = 1/10$ and we want to solve the likelihood equations as explained above to obtain all the possible solutions θ_1 and θ_2. We carry out the process above with Singular, using for example the command `eliminate` (the description of the commands to follow is given in example 3.26 of [PS05]), it turns out that after eliminating the five variables $z_1, \ldots, z_5$, we obtain the reduced Gröbner basis of the likelihood varieties very easily (it has six complicated polynomials). We can verify that this variety is of dimension 0 because two polynomials of the basis have pure powers of t_1 and t_2 as their initial terms, and in fact, it has exactly 16 points in $\mathbb{C}^2$, only one of which actually maps to the probability simplex. It turns out that this solution is verified to be a local maximum.

Note that this example is smaller than example 3.1 (the HMM of length 3), not so much in terms of the number of polynomials or variables to eliminate, but in terms of the number of total variables involved when creating the ring in Singular (7 instead of 16). It seems that this number of variables is critical in the scalability of the computation with Gröbner basis.

5 Conclusion

It is difficult to give a conclusion about a field that is evolving quite fast and after being able to only scratch the surface and understanding the easiest techniques. However, at this point I would say that the methods of algebraic statistics are not yet ready to be applied in the engineering sciences; or at least, let me say that the interesting methods are not easy to recognize without more background in the subject than I had. First of all, a large number of papers in the field present examples of computations using Gröbner basis techniques: it is then not purely for teaching purposes that the examples proposed are very small. As we have seen for example in the case of hidden Markov models, the number of variables in the polynomial ring involved grows exponentially with the length of the model, and extremely simple models are already beyond what can be done using these techniques.

In general, an advantage of using algebraic methods is to solve parametric problems ([PS04]), and obtain globally optimum solutions of polynomial equations. For example, we have seen that in principle, we could obtain all the solutions of the likelihood equations by characterizing the likelihood variety. Note that for most models, statisticians have to rely on the expectation-maximization algorithm because the problem of finding only one local maximum likelihood is already a difficult nonlinear programming problem. It is not clear to me at this point that the methods of algebraic statistics are anywhere close to improving on this algorithm for realistic models, when computational considerations are taken into account. However, algebra and combinatorics are useful for more theoretical questions, for example bounding the number of solutions of these equations.

The question of the size of the models is important. It may be that in computational biology, some small graphical models with just a few variables are of interest, but graphical models in general provide a useful framework mostly for very large problems with say millions of variables and a local description of the joint probability function. Then it is interesting to bring in algorithms from graph theory.

Last, the geometry of maximum-likelihood estimation has been adressed earlier using tools from differential geometry [Kas89]. In the algebraic geometric framework, this is where tropical geometry comes in, as well as techniques from polyhedral combinatorics [PS04]. Again it is not easy there to isolate the scalable computational techniques from the theoretical developments.

References

[BM05] N. Bray and J. Morton. Implicitizer. Available at http://bio.math.berkeley.edu/ascb/chapter11/, 2005.

[CHS06] F. Catanese, S. Hosten, and B. Sturmfels. The maximum likelihood degree. *Americal Journal of Mathematics*, 2006. To appear.

[CLO97] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra.* Springer, 2nd edition, 1997.

[Dud] R.M. Dudley. *Mathematical Statistics.* In Preparation. Available on OpenCourseWare.

[GKZ94] I. Gelfand, M. Kapranov, and A. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkha"auser, Basel-Boston-Berlin, 1994.

[GP02] G.-M. Greuel and G. Pfister. *A Singular Introduction to Commutative Algebra*. Springer, 2002.

[Kas89] R.E. Kass. The geometry of asymptotic inference. *Statistical Science*, 4:188–234, 1989.

[PRW00] G. Pistone, E. Riccomagno, and H.P. Wynn. *Algebraic Statistics*. Chapman & Hall, 2000.

[PS04] L. Pachter and B. Sturmfels. Tropical geometry of statistical models. *Proceedings of the National Academy of Sciences*, 101(46), November 2004.

[PS05] L. Pachter and B. Sturmfels, editors. *Algebraic Statistics for Computational Biology*. Cambridge University Press, 2005.

[Rab89] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 1989.