Toric Varieties in Algebraic Statistics
Math in the Mountains Tutorial

John B. Little

Department of Mathematics and Computer Science
College of the Holy Cross

July 29-31, 2013

Outline

What is algebraic statistics?

- Study of probability models and techniques for statistical inference using methods from algebra and algebraic geometry
- First occurrence of term: in the book [PRW]
- Connections especially with genomics, mathematical biology: see especially [PS]
- Now a very active field, well-represented at the SIAM conference later this week

Example 0

- Key idea: probabilities for discrete random variables often depend *polynomially* on some parameters
- So can think of parametrized families of distributions
- Example: If X is a *binomial random variable* based on n trials, with success probability θ, then X takes values in $\{0, 1, \ldots, n\}$ with probabilities given by:

$$P(X = k) = p_k(\theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

- Gives:

$$\begin{aligned} \varphi : \mathbb{R} &\to \mathbb{R}^{n+1} \\ \theta &\mapsto (p_0(\theta), p_1(\theta), \ldots, p_n(\theta)) \end{aligned}$$

Example, continued

- Since $\sum_i p_i(\theta) = 1$, the image $\varphi(\mathbb{R})$ is a *curve* in the hyperplane $\sum_i p_i = 1$
- If $\theta \in [0, 1]$, then $\varphi(\theta) \in \Delta_{n+1}$, the probability simplex defined by $\sum_i p_i = 1$, and $p_i \geq 0$ for $i = 0, \ldots, n$.
- Question: What curve is it?
- For instance, with $n = 2$, we get the curve $V(p_1^2 - 4p_0p_2, p_0 + p_1 + p_2 - 1)$, a smooth conic.
- For general n, we get a *rational normal curve* of degree n (but with not exactly the usual parametrization because of the binomial coefficient factors)

Probability models

- For the purposes of this talk, a *probability model* will be a parametrized family of probability distributions for a random variable, or joint distributions for collections
- If a (collection of) random variable(s) X with values $s \in \mathcal{S}$ has $P(X = s) = g_s(\theta_1, \ldots, \theta_n)$ for some parameters θ_j,
- then as above, we can consider the mapping

$$\varphi : \mathbb{R}^n \to \mathbb{R}^{\mathcal{S}}$$
$$\theta = (\theta_1, \ldots, \theta_n) \mapsto (g_s(\theta) : s \in \mathcal{S})$$

Probability models, cont.

- We will also assume that the g_i are *polynomial*, or at worst *rational* functions of θ.
- By standard results, this implies that $\varphi(\mathbb{R}^n)$ is a subset of some algebraic variety in $\mathbb{R}^{\mathcal{S}}$
- Given such a φ, the corresponding *model* is the set

$$\overline{\varphi(\mathbb{R}^n)} \cap \Delta$$

where Δ is the probability simplex in $\mathbb{R}^{\mathcal{S}}$.

Example 1

- Suppose X is a categorical random variable with three possible values $\{1, 2, 3\}$ and Y is a second variable of the same type.
- Assume in addition that X, Y are *independent*, which is equivalent to saying that $P(X = x \text{ and } Y = y) = P(X = x)P(Y = y)$ for all $x, y \in \{1, 2, 3\}$.
- Then writing $P(X = x) = p_x$ and $P(Y = y) = q_y$, we can arrange the 9 values needed to specify the joint probability function as a 3×3 matrix and we obtain

$$\mathcal{P} = \begin{pmatrix} p_1q_1 & p_1q_2 & p_1q_3 \\ p_2q_1 & p_2q_2 & p_2q_3 \\ p_3q_1 & p_3q_2 & p_3q_3 \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} \begin{pmatrix} q_1 & q_2 & q_3 \end{pmatrix}.$$

Example 1, cont.

- Every such matrix $\mathcal{P}$ has rank 1
- Conversely, any matrix

$$\mathcal{P} = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix}$$

  of rank 1 has this form.
- We also can see implicit equations for a variety in $M_{3\times 3}(\mathbb{R}) = \mathbb{R}^9$ containing all such matrices:

$$p_{ij}p_{kl} - p_{il}p_{kj} = 0$$

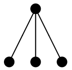
  for all pairs of rows $1 \leq i \leq k \leq 3$ and all pairs of columns $1 \leq j \leq l \leq 3$.

Example 1, concluded

- The corresponding parametrization is
$$\varphi : \mathbb{R}^4_{p_1,p_2,q_1,q_2} \to M_{3\times 3}(\mathbb{R})$$
(where $p_3 = 1 - p_1 - p_2$, and similarly $q_3 = 1 - q_1 - q_2$)
- $\overline{\varphi(\mathbb{R}^6)} \cap \Delta$ is called the 3×3 *independence model* – there are similar $k \times \ell$ independence models for all k, ℓ
- If $p_x \geq 0$ and $q_y \geq 0$ with $\sum_x p_x = 1 = \sum_y q_y$ then the sum of the entries in the 3×3 matrix is also 1.
- The variety involved in the model in the 3×3 case can be identified with *Segre embedding* of $\mathbb{P}^2 \times \mathbb{P}^2$ in $\mathbb{P}^8$, defined by the quadratic binomials given on the last slide.

Example 2 – Jukes-Cantor

- Important application of these ideas is probability models for DNA sequence evolution
- The *Jukes-Cantor* DNA model describes probabilities of changes in going from an ancestor sequence to some collection of descendant sequences.
- Model on a K_{13} "claw tree" considers 3 one-step descendant sequences: given by a mapping $\varphi : \mathbb{R}^3 \to \mathbb{R}^{64}$



- π_i, $i = 1, 2, 3$ are the probabilities of a DNA letter (A,C,G,T) in the ancestor (root) changing to a *different* letter in going from the root to descendant (leaf) i (these are *same* for all changes, but vary with i).

Example 2, cont.

- Model assumes A,C,G,T occur randomly, uniformly distributed in root sequence
- Write $\theta_i = 1 - 3\pi_i$ for the probability of not changing in descendant sequence i.
- What happens for each of the three leaves also subject to an independence assumption.
- Get probabilities for each possible collection of outcomes in the leaves. For instance,

$$\begin{aligned} P(AAA) &= P(AAA|rt = A)P(rt = A) + P(AAA|rt \neq A)P(rt \neq A) \\ &= \frac{1}{4}(\theta_1\theta_2\theta_3 + 3\pi_1\pi_2\pi_3). \end{aligned}$$

Example 2, cont.

- Among the entries of $\varphi(\pi_1, \pi_2, \pi_3)$, there are only 5 different polynomials.
- So the model has a condensed parametric form using

$$\begin{aligned}
p_{123} &= \theta_1\theta_2\theta_3 + 3\pi_1\pi_2\pi_3 \\
p_{dis} &= 6\theta_1\pi_2\pi_3 + 6\theta_2\pi_1\pi_3 + 6\theta_3\pi_1\pi_2 + 6\pi_1\pi_2\pi_3 \\
p_{12} &= 3\theta_1\theta_2\pi_3 + 3\pi_1\pi_2\theta_3 + 6\pi_1\pi_2\pi_3 \\
p_{13} &= 3\theta_1\theta_3\pi_2 + 3\pi_1\pi_3\theta_2 + 6\pi_1\pi_2\pi_3 \\
p_{23} &= 3\theta_2\theta_3\pi_1 + 3\pi_2\pi_3\theta_1 + 6\pi_1\pi_2\pi_3
\end{aligned}$$

- Here $p_{123} =$ probability of observing *the same* letter in all three descendants, $p_{dis} =$ probability of *3 distinct letters* in the descendants, and $p_{ij} =$ probability of *equal letters in descendants i, j and something different in descendant k*.

Example 2, cont.

- So we have the condensed model parametrization $\varphi : \mathbb{R}^3 \to \mathbb{R}^5$.
- Since the expressions for p_{123}, etc. are polynomials in π_i, the image is a variety of dimension 3 in a hyperplane in $\mathbb{R}^5$
- The Jukes-Cantor model is the intersection of that variety with the 4-dimensional probability simplex in that hyperplane
- From dimensional considerations, should have one equation of model in addition to $p_{123} + p_{dis} + p_{12} + p_{13} + p_{23} = 1$: it's a *complicated* polynomial of degree 3 in the variables p_{123}, p_{dis} and p_{ij}.

Did you notice something?

- In Examples 0 and 1 above, note that the implicit equations for the models we were describing were given by *binomials* of the form $x^\alpha - x^\beta$ for some multi-indices α, β
- As we know, the corresponding ideals are examples of *toric ideals* and the varieties defined by the implicit equations are *toric varieties*
- This is *not true*, though, for the Jukes-Cantor model in Example 2.
- Examples 0 and 1 are instances of *toric models*, essentially because we can give the parametric equations in *monomial form* (possibly by using “extra variables” – e.g. in Example 0, could write $p = \theta$, $q = 1 - p$, and then have nearly the standard monomial parametrization of the rational normal curve)

The general definition

- Let $\mathcal{A} = (a_{ij})$ be a $d \times m$ non-negative integer matrix, *with equal column sums* $\Leftrightarrow (1, \ldots, 1) \in \mathbb{R}^m$ is in the real rowspace of $\mathcal{A}$.
- Write A_j for the jth column of $\mathcal{A}$ and

$$\theta^{A_j} = \theta_1^{a_{1j}} \cdots \theta_d^{a_{dj}}$$

  for the corresponding monomial in parameters $\theta_1, \ldots, \theta_d$.
- Given $c_1, \ldots, c_m > 0$ consider $\varphi : \mathbb{R}^d \to \mathbb{R}^m$ defined by

$$\theta \mapsto \frac{1}{\sum_{j=1}^m c_j \theta^{A_j}} \left(c_1 \theta^{A_1}, \ldots, c_m \theta^{A_m} \right)$$

- The *toric model* associated to $\mathcal{A}$ (and c) is $\varphi(\mathbb{R}^d_{>0}) \cap \Delta$.

Comments

- Note: with this formulation, $\varphi(\mathbb{R}^d)$ is contained in the hyperplane $\sum_{j=1}^m p_j = 1$ because of the denominators in the components of φ.
- $\theta_i > 0$ implies $\varphi(\theta)$ is in the probability simplex Δ so intersecting with Δ is not actually necessary.
- The c_j are included to allow for numerical weight factors as in Example 0.
- Exercise: What is the matrix $\mathcal{A}$ for the 3×3, or more generally $k \times \ell$, independence model?

Comments, cont.

- But, the "equal column sums" assumption on $\mathcal{A}$ means that all the monomials θ^{A_j} have the same total degree and the implicit equations of the corresponding toric variety will come from homogeneous polynomials, so we can essentially ignore the denominators (or view $\varphi(\theta)$ as homogeneous coordinates of a point in $\mathbb{P}^{m-1}$).
- In a toric model, the logarithms of the probabilities (more precisely, the numerators of the components of φ) are linear functions of the $\log(\theta_i)$;
- Toric models have a long history in "mainstream" statistics; these are often called *log-linear models*.

An observation

- Even if a model is not toric, it might be possible to "make it toric" by a reparametrization. For instance, for the Jukes-Cantor model in Example 2 above – not a random fact, an application of finite Fourier transforms (!)
- Exercise: Can check that

$$\begin{aligned} q_{111} &= (\theta_1 - \pi_1)(\theta_2 - \pi_2)(\theta_3 - \pi_3) \\ q_{110} &= (\theta_1 - \pi_1)(\theta_2 - \pi_2)(\theta_3 + 3\pi_3) \\ q_{101} &= (\theta_1 - \pi_1)(\theta_2 + 3\pi_2)(\theta_3 - \pi_3) \\ q_{011} &= (\theta_1 + 3\pi_1)(\theta_2 - \pi_2)(\theta_3 - \pi_3) \\ q_{000} &= (\theta_1 + 3\pi_1)(\theta_2 + 3\pi_2)(\theta_3 + 3\pi_3) \end{aligned}$$

  are *linear combinations* of p_{123}, p_{dis}, p_{ij}, *and* monomials in linear combinations of the original parameters.
- Exercise: What is the corresponding toric variety?

The toric ideal of a model

- Let $\mathcal{A}$ be a matrix as above and consider the toric variety $X_{\mathcal{A}} = \overline{\varphi(\mathbb{R}^d_{>0})}$ with $c_i = 1$ for all i
- By general results we saw before, the vanishing ideal of $X_{\mathcal{A}}$ is the toric ideal

$$I_{\mathcal{A}} = \langle p^{e_+} - p^{e_-} : e_+, e_- \in \mathbb{N}^m, \mathcal{A}e_+ = \mathcal{A}e_- \rangle$$

  (a finite set of such binomials generates the same ideal, by the Hilbert basis theorem)
- If some $c_i \neq 1$, then the ideal of the corresponding toric variety can be found by a simple scaling (change of variables)
- Any finite set of generators for this ideal gives what is known as a *Markov basis*

“Real statistics”

- Describing a toric model (such as the binomial model from Example 0 or the 3×3 independence model from Example 1) is only the first step for statisticians
- Given data (for example some collection of sampled values of the variables involved), we could ask: Assuming the model, what parameters would best explain that data? And, perhaps: Is the corresponding model a reasonable description for the data?
- More precisely, we might want to set up a test to decide whether it is reasonable to reject the hypothesis that the model does not fit the data.

The sufficient statistics

- The *likelihood function* is the probability of observing a given collection of counts $u = (u_1, \ldots, u_m)$ under the model, as a function of the model parameters: $L(u|\theta) = \prod_{j=1}^m \varphi_j(\theta)^{u_j}$
- A collection of statistics $T(u)$ is *sufficient* for the problem of estimating θ if the interaction between θ and u in the likelihood function is entirely through $T(u)$:
- The general factorization criterion from the theory of estimators says that $T(u)$ is sufficient for θ if $L(u|\theta) = f(T(u), \theta) \cdot g(u)$
- Exercise: The factorization criterion shows that $\mathcal{A}u$ is sufficient for θ.

Maximum likelihood estimators; hypothesis testing

- A standard approach here is to look for the parameter values that *maximize the likelihood* – called the MLE parameter values, $\widehat{\theta}$.
- Given the MLE we get MLE estimates $\widehat{u}_j$ for the data values and consider the χ^2-type formula:

$$X(v) = \sum_{j=1} \frac{(\widehat{u}_j - v_j)^2}{\widehat{u}_j^2}$$

- If the proportion of the vectors v in the fiber $\mathcal{A}^{-1}(\mathcal{A}u)$ with $X(v) \geq X(u)$ is sufficiently small we would reject the hypothesis that the model does not fit the data.

How Markov bases are used

- The problem here is that for all but quite small problems, the fiber $\mathcal{A}^{-1}(\mathcal{A}u)$ is too large to enumerate explicitly.
- However given any u_0 in the fiber and a Markov basis as above $\{x^{e_+} - x^{e_-}\}$, note that $u_0 + e_+ - e_-$ is also in the fiber
- Hence, we can do a random walk through the fiber using the Markov basis in this way to estimate the necessary proportion (Metropolis-Hastings)

Toric models are “very nice” for MLE

- Main reason: Toric models give likelihood functions where all critical points $\widehat{\theta}$ have $\widehat{p} = \varphi(\widehat{\theta})$ in a very special position in the variety describing the model.
- Moreover, there can be *only one* critical point, and can show that is necessarily the MLE (a result called *Birch's theorem* in statistics)
- Moreover, and best of all, in some cases (e.g. the independence models) there are analytic expressions for the MLE's $\widehat{\theta}$ in terms of easily computable information such as marginals in the data (“contingency tables”)
- You will derive much of this in general in exercises; let's work out a simple special case to see the ideas involved.

MLE example

- Say we are working with the toric model given by the 2×4 matrix
$$\mathcal{A} = \begin{pmatrix} 3 & 2 & 1 & 0 \\ 0 & 1 & 2 & 3 \end{pmatrix}$$
- In parametric form, $\varphi(\theta_1, \theta_2) =$
$$(p_0, p_1, p_2, p_3)^t = (\theta_1^3, \theta_1^2\theta_2^2, \theta_1\theta_2^2, \theta_2^3)^t$$
- The corresponding toric variety is the cone over the standard twisted cubic:
$$V(p_0p_2 - p_1^2, p_0p_3 - p_1p_2, p_1p_3 - p_2^2)$$
- We can think of this as a model giving probability distributions for random variables with values in $\{0, 1, 2, 3\}$.

Finding the MLE

- Say we have made $N = 100$ observations and observed counts $u = (13, 35, 29, 23)^t$.
- The likelihood function here is

$$\begin{aligned} L &= (\theta_1^3)^{13}(\theta_1^2\theta_2)^{35}(\theta_1\theta_2^2)^{29}(\theta_2^3)^{23} \\ &= \theta_1^{138}\theta_2^{162}. \end{aligned}$$

- The exponents here are the entries in the vector $b = Au$ (do you see why?)
- We want to maximize this, but subject to the constraint that $\varphi(\theta_1, \theta_2)$ is a "legal" vector of probabilities:

$$q = \theta_1^3 + \theta_1^2\theta_2 + \theta_1\theta_2^2 + \theta_2^3 = 1$$

Finding the MLE, cont.

- This is a constrained optimization problem, so can use the method of Lagrange multipliers: Any critical point of L restricted to the constraint set satisfies $\frac{\partial L}{\partial \theta_i} = \lambda \frac{\partial q}{\partial \theta_i}$ for $i = 1, 2$ and some constant λ
- Because L is a monomial (hence homogeneous) and q is homogeneous, if we take the usual Lagrange equations and multiply by the first by θ_1 and the second by θ_2, we get

$$\begin{aligned} 138L &= \lambda(3\theta_1^3 + 2\theta_1^2\theta_2 + \theta_1\theta_2^2) \\ 162L &= \lambda(\theta_1^2\theta_2 + 2\theta_1\theta_2^2 + 3\theta_2^3), \end{aligned}$$

or (in vector format), writing $\widehat{p}$ for $\varphi(\widehat{\theta})$, where $\widehat{\theta}$ is the MLE for $\theta = (\theta_1, \theta_2)^t$,

$$L \cdot b = L \cdot Au = \lambda \cdot A\widehat{p} \qquad (1)$$

Finding the MLE, cont.

- Since $(1,1,1,1)$ is in the real rowspace of A, we can multiply both sides here by some vector to obtain $L \cdot 100 = \lambda$ (since $\sum u_i = N = 100$ and $(1,1,1,1)\widehat{p} = 1$).
- Substituting back into (1), we obtain

$$A\widehat{p} = \frac{1}{100} b = \left(\frac{138}{100}, \frac{162}{100}\right)$$

- Or explicitly

$$\begin{aligned} 3\theta_1^3 + 2\theta_1^2\theta_2 + \theta_1\theta_2^2 &= \frac{138}{100} \\ \theta_1^2\theta_2 + 2\theta_1\theta_2^2 + 3\theta_2^3 &= \frac{162}{100} \end{aligned}$$

Finding the MLE, cont.

- These equations can be solved numerically, yielding a unique real solution:

$$\widehat{\theta_1} \doteq .5992 \text{ and } \widehat{\theta_2} \doteq .6597$$

- The equation $\mathcal{A}p = \frac{1}{100}b$ defines a polyhedron that meets the model variety $\varphi(\mathbb{R}^2_{>0})$ in exactly the one point we found approximately above.
- The general proof for this uses the same setup and reasoning as the proof of the properties of the (“algebraic”) *moment map* for the corresponding toric variety (will discuss this in a later talk)

MLE degree of a model

How many real or complex roots can there be for the ML equations for a model? The abstract of [CHKS]: "Maximum likelihood estimation in statistics leads to the problem of maximizing a product of powers of polynomials. We study the algebraic degree of the critical equations of this optimization problem. This degree is related to the number of bounded regions in the corresponding arrangement of hypersurfaces, and to the Euler characteristic of the complexified complement. Under suitable hypotheses, the maximum likelihood degree equals the top Chern class of a sheaf of logarithmic differential forms. Exact formulae in terms of degrees and Newton polytopes are given for polynomials with generic coefficients."

References for further study

CHKS Catanese, Hosten, Khetan, and Sturmfels *The maximum likelihood degree*, Amer. J. Math. 128 (2006), 671-697

DSS Drton, Sturmfels, and Sullivant, *Lectures on Algebraic Statistics*, Springer, 2008

CLS Cox, Little, and Schenck, *Toric Varieties*, AMS, 2011

PRW Pistone, Riccomagno, and Wynn, *Algebraic Statistics: Computational Commutative Algebra in Statistics*, Chapman and Hall, 2000

PS Pachter and Sturmfels, *Algebraic Statistics for Computational Biology*, Cambridge U. Press, 2005