An Introduction to Algebraic Statistics

Serkan Hoşten

San Francisco State University and IMA

7 February 2007

Two messages

1. Many interesting parametric statistical models (on discrete data) are algebraic varieties
2. Geometry of these varieties has statistical relevance

Statistical Models

- State space $\mathcal{S}$ where $|\mathcal{S}| = n$

- The probability simplex
$\Delta_{\mathcal{S}} = \Delta_n = \{(p_1, \ldots, p_n) \in \mathbf{R}^n : \sum_{i=1}^n p_i = 1, \quad p_i \geq 0\}$
where $p_i = \text{Prob}(X = s_i)$

- A model on $\mathcal{S}$ is a subset $\mathcal{M} \subset \Delta_{\mathcal{S}}$

- $\mathcal{M}$ is a parametric model if there are functions
$f_1(\theta), \ldots, f_n(\theta)$ such that

$$\mathcal{M} = \text{im}(\theta \in \mathbf{R}^d \mapsto (f_1(\theta), \ldots, f_n(\theta)) \in \Delta_n)$$

Parametric models as algebraic varieties

Let $f_i(\theta)$ be polynomials in $\mathbf{R}[\theta_1, \ldots, \theta_d]$:

$$f_i = \sum_{a=(a_1,\ldots,a_d)} c_a \theta^a \quad \text{where} \quad c_a \in \mathbf{R} \quad \text{and} \quad \theta^a := \theta_1^{a_1} \cdots \theta_d^{a_d}$$

The Zariski closure of the image

$$\mathrm{im}(\theta \in \mathbf{C}^d \mapsto (f_1(\theta), \ldots, f_n(\theta)))$$

is an algebraic variety: it is defined by finitely many polynomial equations, namely, the model invariants

$$g_1(p_1, \ldots, p_n) = g_2(p_1, \ldots, p_n) = \cdots = g_k(p_1, \ldots, p_n) = 0$$

where $g_i \in \mathbf{R}[p_1, \ldots, p_n]$.

Example : Two independent random variables

Let Y and Z be two discrete random variables taking values in $[m] = \{1, \ldots, m\}$ and $[n] = \{1, \ldots, n\}$, respectively.

The joint distribution $p_{ij} = \text{Prob}(Y = i, Z = j)$ is in $\Delta_{mn} = \Delta_{\mathcal{S}}$ where $\mathcal{S} = [m] \times [n]$.

$$\begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{m1} & \cdots & p_{mn} \end{bmatrix}$$

$$\mathcal{M} \equiv Y \perp\!\!\!\perp Z$$

$$\Longleftrightarrow$$

$$\text{Prob}(Y = i, Z = j) = \text{Prob}(Y = i) \cdot \text{Prob}(Z = j)$$

$$\Longleftrightarrow$$

$$p_{ij} = \alpha_i \beta_j =: f_{ij}(\alpha, \beta)$$

$$\Longleftrightarrow$$

$$\begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{m1} & \cdots & p_{mn} \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix} \begin{bmatrix} \beta_1 & \cdots & \beta_n \end{bmatrix}$$

$$\Longleftrightarrow$$

$$p_{ij}p_{st} - p_{it}p_{sj} = 0 \quad 1 \leq i < s \leq m, \quad 1 \leq j < t \leq n$$

Hierarchical Loglinear Models

- $X_1, \ldots, X_n$ discrete random variables taking values in $[d_1], \ldots, [d_n]$, respectively. Let

$$p_{u_1 u_2 \ldots u_n} = \text{Prob}(X_j = u_j \ : \ j = 1, \ldots, n)$$

- Γ a simplicial complex on n vertices, one for each X_j.
- For each facet $F = \{X_{i_1}, \ldots, X_{i_k}\}$ of Γ introduce parameters

$$a^F_{u_{i_1}, \ldots, u_{i_k}} \text{ where } (u_{i_1}, \ldots, u_{i_k}) \in [d_{i_1}] \times \cdots \times [d_{i_k}].$$

- Define a model by the parametrization

$$p_{u_1 u_2 \ldots u_n} = \prod_F a^F_{u|_F} \quad (u_1, \ldots, u_n) \in [d_1] \times \cdots \times [d_n]$$

Examples

- $\Gamma = \{\{Y\}, \{Z\}\}$ and $d_1 = m$ and $d_2 = n$:

$$p_{ij} = a_i b_j, \;\; i = 1, \ldots, m, \;\; j = 1, \ldots, n$$

- X, Y, Z, W binary and $\Gamma = \{\{X, Y, Z\}, \{Y, Z, W\}\}$:

$$p_{ijst} = a_{ijs} b_{jst} \quad 0 \leq i, j, s, t \leq 1.$$

- X, Y, Z, W binary and
$\Gamma = \{\{X, Y\}, \{Y, Z\}, \{Z, W\}, \{X, W\}\}$:

$$p_{ijst} = a_{ij} b_{js} c_{st} d_{it} \quad 0 \leq i, j, s, t \leq 1.$$

Equations

- 2×2-minors of (p_{ij})
- 2×2-minors of (p_{ijst}) for each $0 \leq s, t \leq 1$
- The equations of the binary four-cycle are:

$$p_{1011}p_{1110} - p_{1010}p_{1111}, \quad p_{0111}p_{1101} - p_{0101}p_{1111} \quad p_{1001}p_{1100} - p_{1000}p_{1101},$$

$$p_{0110}p_{1100} - p_{0100}p_{1110}, \quad p_{0011}p_{1001} - p_{0001}p_{1011}, \quad p_{0010}p_{1000} - p_{0000}p_{1010},$$

$$p_{0011}p_{0110} - p_{0010}p_{0111}, \quad p_{0001}p_{0100} - p_{0000}p_{0101},$$

$$p_{0000}x_{0011}p_{1101}p_{1110} - p_{0001}p_{0010}p_{1100}p_{1111}, \quad p_{0000}p_{0111}p_{1001}p_{1110} - p_{0001}p_{0110}p_{1}$$

$$p_{0000}p_{0110}p_{1101}p_{1101} - p_{0010}p_{0100}p_{1001}p_{1111}, \quad p_{0001}p_{0110}p_{1010}p_{1101} - p_{0010}p_{0101}p_{1}$$

$$p_{0000}p_{0111}p_{1011}p_{1100} - p_{0011}p_{0100}p_{1000}p_{1111}, \quad p_{0010}p_{0101}p_{1011}p_{1100} - p_{0011}p_{0100}p_{1}$$

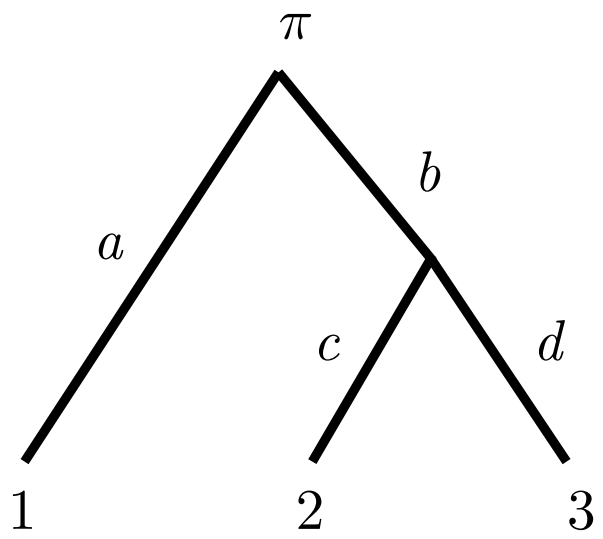
$$p_{0001}p_{0111}p_{1010}p_{1100} - p_{0011}p_{0101}p_{1000}p_{1110}, \quad p_{0100}p_{0111}p_{1001}p_{1010} - p_{0101}p_{0110}p_{1}$$

Remarks

- More on hierarchical models can be found in [Fienberg, '77].
- Popular examples include

1. No n-way interaction models (Γ is the boundary of n-simplex), [Aoki-Takemura]
2. Graphical models [Lauritzen, '96] (Γ is the clique complex of a graph), [Geiger-Meek-Sturmfels, '02]
3. Graph models [Develin-Sullivant, '03] (Γ is a graph),
4. Bayesian networks [Garcia-Stillman-Sturmfels, '05]

Markov models for phylogenetic evolution



1: `Gorilla   :  AACGTTGACCAAT ....`

2: `Human     :  ACCGTAACTAGTT ....`

3: `Chimpanzee :  ACCGCAACGAGTT ....`

The root distribution $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$
For each edge (branch) a 4×4 matrix of transition probabilities

$$M_i = \begin{bmatrix} \delta_i & \beta_i & \alpha_i & \gamma_i \\ \beta_i & \delta_i & \gamma_i & \alpha_i \\ \alpha_i & \gamma_i & \delta_i & \beta_i \\ \gamma_i & \alpha_i & \beta_i & \delta_i \end{bmatrix}$$

This is Kimura 3-parameter model. Kimura 2-parameter model is obtained when $\gamma_i = \beta_i$, and Jukes-Cantor model is obtained when $\gamma_i = \beta_i = \alpha_i$.

Jukes-Cantor Probabilities

$$
\begin{aligned}
p_{CTG} \quad = \quad & \pi_A(\alpha_a\delta_b\alpha_c\alpha_d + \alpha_a\alpha_b\alpha_c\alpha_d + \alpha_a\alpha_b\alpha_c\delta_d + \alpha_a\alpha_b\delta_c\alpha_d) + \\
& \pi_C(\delta_a\alpha_b\alpha_c\alpha_d + \delta_a\delta_b\alpha_c\alpha_d + \delta_a\alpha_b\alpha_c\delta_d + \delta_a\alpha_b\delta_c\alpha_d) + \\
& \pi_G(\alpha_a\alpha_b\alpha_c\alpha_d + \alpha_a\alpha_b\alpha_c\alpha_d + \alpha_a\delta_b\alpha_c\delta_d + \alpha_a\alpha_b\delta_c\alpha_d) + \\
& \pi_T(\alpha_a\alpha_b\alpha_c\alpha_d + \alpha_a\alpha_b\alpha_c\alpha_d + \alpha_a\alpha_b\alpha_c\delta_d + \alpha_a\delta_b\delta_c\alpha_d)
\end{aligned}
$$

Kimura 3-parameter and its specializations are group based $(\mathbf{Z}_2 \times \mathbf{Z}_2)$models, and they allow Fourier transform of the parameters and probabilities:

$$q_{ijk} = a_i c_j d_k b_{j+k} r_{i+j+k}$$

Theorem: [Sturmfels-Sullivant, '04] Let T be an arbitrary binary rooted tree. Modulo the trivial invariant $q_{00\ldots0} - 1$,

1. the ideal of the Jukes-Cantor binary model is generated by quadrics,
2. the ideal of the Jukes-Cantor DNA model is generated by linear, quadratic, and cubic polynomials,
3. the ideal of the Kimura 2-parameter model is generated by polynomials of degree 1, 2, 3 and 4,
4. the ideal of the Kimura 3-parameter model is generated by polynomials of degree 2, 3 and 4.

Each has generating sets (which are also Gröbner bases) with an explicit combinatorial description.

Remarks

• Check out Small Phylogenetic Trees website:

$$\texttt{http} : //\texttt{www.math.tamu.edu/lgp/small} - \texttt{trees}$$

• Read Algebraic Statistics for Computational Biology, edited by Pachter-Sturmfels.

• For General Markov Models, ask Allman-Rhodes.

• These models provide a general setting applicable in other sciences, such as linguistics.

• Message 1: Many interesting parametric statistical models (on discrete data) are algebraic varieties

Maximum Likelihood Estimation

A given data set is a vector $u = (u_1, \ldots, u_n)$ of non-negative integers. The problem of *maximum likelihood estimation* is to find parameters θ which best explain the data u. This leads to the following optimization problem:

$$\text{Maximize } f_1(\theta)^{u_1} f_2(\theta)^{u_2} \cdots f_n(\theta)^{u_n} \quad \text{subject to } \theta \in \mathcal{U}. \tag{1}$$

The optimal solution $\hat{\theta}$ to the problem (1) is called a maximum likelihood estimator (MLE).

Example: Two independent binary variables

$$f_1 = \theta_1\theta_2,\ f_2 = (1-\theta_1)\theta_2,\ f_3 = \theta_1(1-\theta_2),\ f_4 = (1-\theta_1)(1-\theta_2). \tag{2}$$

Write the likelihood equations :

$$\frac{u_1}{f_1}\frac{\partial f_1}{\partial \theta_1} + \frac{u_2}{f_2}\frac{\partial f_2}{\partial \theta_1} + \frac{u_3}{f_3}\frac{\partial f_3}{\partial \theta_1} + \frac{u_4}{f_4}\frac{\partial f_4}{\partial \theta_1} = \frac{u_1+u_3}{\theta_1} - \frac{u_2+u_4}{1-\theta_1} = 0$$

$$\frac{u_1}{f_1}\frac{\partial f_1}{\partial \theta_2} + \frac{u_2}{f_2}\frac{\partial f_2}{\partial \theta_2} + \frac{u_3}{f_3}\frac{\partial f_3}{\partial \theta_2} + \frac{u_4}{f_4}\frac{\partial f_4}{\partial \theta_2} = \frac{u_1+u_2}{\theta_2} - \frac{u_3+u_4}{1-\theta_2} = 0$$

... and solve:

$$\hat{\theta}_1 = \frac{u_1+u_3}{u_1+u_2+u_3+u_4} \quad \text{and} \quad \hat{\theta}_2 = \frac{u_1+u_2}{u_1+u_2+u_3+u_4}.$$

Maximum Likelihood Degree

The MLE $\hat{\theta}$ is an algebraic function of the data u. The maximum likelihood degree of the model given by $f_1, \ldots, f_n$ is the degree of that algebraic function. Equivalently, the ML degree is the number of complex solutions of the critical equations of (1), for a general vector u.

Theorem: [Catanese-H.-Khetan-Sturmfels'04]
If $f_1, \ldots, f_n$ are general polynomials of degree $b_1, \ldots, b_n$ then the maximum likelihood degree is the coefficient of z^d in the generating function

$$\frac{(1-z)^d}{(1-zb_1)(1-zb_2)\cdots(1-zb_n)}. \tag{3}$$

If the f_i are four general quadrics in two variables then the ML degree is 25 complex solutions:

$$\frac{(1-z)^2}{(1-2z)^4} \;=\; 1 + 6z + \underline{25}z^2 + 88z^3 + 280z^4 + \cdots,$$

However, for special quadrics f_i, the ML degree can be much lower than 25.

Warning: Most statistical models are not defined by generic polynomials.

Behrens-Fisher Problem

• X and Y are two p-dimensional normal random vectors with same mean $\mu \in \mathbf{R}^p$ and $p \times p$ covariance matrices Σ_X and Σ_Y.

• $X_1, \ldots, X_{N_1}$ and $Y_1, \ldots, Y_{N_2}$ are given random samples.

• Problem: Estimate the parameters μ, Σ_X, and Σ_Y by maximum likelihood method.

The MLE $(\hat{\mu}, \hat{\Sigma}_X, \hat{\Sigma}_Y)$ is a solution to the likelihood equations:

$$\hat{\Sigma}_X = \frac{1}{N_1} \sum_{i=1}^{N_1} (X_i - \hat{\mu})(X_i - \hat{\mu})^t$$

$$\hat{\Sigma}_Y = \frac{1}{N_2} \sum_{i=1}^{N_2} (Y_i - \hat{\mu})(Y_i - \hat{\mu})^t$$

$$(N_1 \hat{\Sigma}_X^{-1} + N_2 \hat{\Sigma}_Y^{-1})\hat{\mu} - (N_1 \hat{\Sigma}_X^{-1} \bar{X} + N_2 \hat{\Sigma}_Y^{-1} \bar{Y}) = 0$$

Let

$$S_X = \frac{1}{N_1} \sum_{i=1}^{N_1} (X_i - \bar{X})(X_i - \bar{X})^t \text{ and } S_Y = \frac{1}{N_2} \sum_{i=1}^{N_2} (Y_i - \bar{Y})(Y_i - \bar{Y})^t.$$

The likelihood equations are equivalent to :

$$\frac{N_1 S_X^{-1}(\bar{X}-\hat{\mu})}{1+(\bar{X}-\hat{\mu})^t S_X^{-1}(\bar{X}-\hat{\mu})}+\frac{N_2 S_Y^{-1}(\bar{Y}-\hat{\mu})}{1+(\bar{Y}-\hat{\mu})^t S_Y^{-1}(\bar{Y}-\hat{\mu})}=0$$

The above are the critical equations of $f_1(\mu)^{N_1/2} f_2(\mu)^{N_2/2}$ where

$f_1 = 1+(\bar{X}-\hat{\mu})^t S_X^{-1}(\bar{X}-\hat{\mu})$ and $f_2 = 1+(\bar{Y}-\hat{\mu})^t S_Y^{-1}(\bar{Y}-\hat{\mu})$

are generic.

Theorem: The ML degree of the Behrens-Fisher problem is the coefficient of z^p in the rational function

$$\frac{(1-z)^p}{(1-2z)^2}.$$

This coefficient is equal to

$$\sum_{i+j=p} 2^j (-1)^i \binom{p}{i} (j+1) \;=\; 2p+1$$

• How many of these $2p+1$ solutions are real? When $p=1$ all three could be real [Drton, '07]. When p =2, can all five be real?

Remarks

• When f_i are linear the ML degree is the number of bounded regions of the complement of $\bigcup_{i=1}^{n}\{f_i = 0\}$ in $\mathbf{R}^d$. [Varchenko, '97]

• A nontrivial application of the linear case is in discrete bivariate missing data problem [H.-Sullivant, '06]

• The ML degree of decomposable models is one.

• There is an algebraic algorithm to compute the ML degree. [H.-Khetan-Sturmfels, '05]

• Luis Garcia-Puente has a downloadable implementation of this algorithm, and the ML degree of small phylogenetic trees have been computed.

• The ML degrees of many hierarchical loglinear models are a mistery, i.e, we can compute them but we don't know what these numbers are.

Two messages

1. Many interesting parametric statistical models (on discrete data) are algebraic varieties
2. Geometry of these varieties has statistical relevance

THANK YOU