Algebraic Statistics

Luis David Garcia–Puente

lgp@math.tamu.edu

Texas A&M University

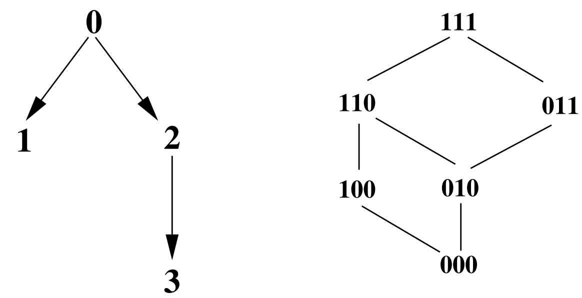
Algebraic Statistics

“The emerging field of **algebraic statistics** advocates the use of polynomial algebra as a tool for statistical inference. The core principle in algebraic statistics is that most statistical models for discrete random variables are algebraic varieties, and that understanding the structure of these varieties can be useful in statistics...”

- Consider n binary random variables $X_1, \ldots, X_n$ each indicating the occurrence of an event.
- A *mutagenetic tree* T on n events is a connected branching on the set of nodes $\{0\} \cup [n]$, rooted at node 0.
- T defines a statistical model as follows. With each edge $(\mathrm{pa}(v), v), v \in [n]$, we associate a parameter $t^v \in [0, 1]$ and the transition matrix $\theta^v = \begin{pmatrix} 1 & 0 \\ 1 - t^v & t^v \end{pmatrix}$.
- The (a, b)-entry of this matrix represents the conditional probability $Pr(X_v = b \mid X_{\mathrm{pa}(v)} = a)$.
- The n-dimensional mutagenetic tree model $\mathcal{T}$ is the image of the polynomial map given by

$$f : [0, 1]^n \longrightarrow \Delta_{2^n - 1}, \quad \theta \longmapsto (f_i(\theta))_{i \in 2^{[n]}}$$

$$f_i(\theta) = \prod_{v=1}^{n} \theta^v_{i_{\mathrm{pa}(v)}, i_v}$$



$$f_{000}(\theta) = (1-t^1)(1-t^2),\ f_{001}(\theta) = 0, \qquad f_{010}(\theta) = (1-t^1)t^2(1-t^3),$$
$$f_{011}(\theta) = (1-t^1)t^2t^3, \qquad f_{100}(\theta) = t^1(1-t^2),\ f_{101}(\theta) = 0,$$
$$f_{110}(\theta) = t^1t^2(1-t^3), \qquad f_{111}(\theta) = t^1t^2t^3.$$

Theorem[Beerenwinkel-Drton 2005, Hibi 1987] The ideal of polynomial invariants $I_{\mathcal{T}}$ of the mutagenetic tree model $\mathcal{T}$ is generated by

$$\{p_ip_j - p_{i\vee j}p_{i\wedge j} \mid i,j \in C(\mathcal{T}),\ i\wedge j < i < j < i \vee j\}\cup$$
$$\{p_i \mid i \notin C(\mathcal{T})\} \cup \{\sum_{i\in 2^{[n]}} p_i - 1\}$$

- The K-mutagenetic trees mixture model $(\mathcal{T}_1, \dots, \mathcal{T}_K)$ is the image of the map $f^{(\mathcal{T}_1, \dots, \mathcal{T}_K)} : \Delta_{K-1} \times \theta^K \longrightarrow \Delta_{2^n-1}$ given by

$$(\lambda, \theta^{(1)}, \dots, \theta^{(K)}) \longmapsto \sum_{i=1}^{K} \lambda_i f^{(\mathcal{T}_i)}(\theta^{(i)}).$$

- In general, mixture models correspond to joins and secant varieties.
- The join of two varieties $X * Y$ is the Zariski closure of the union of all lines spanned by a point in X and a point in Y. The join variety of X with itself is the secant variety of X.
- The K-mutagenetic trees mixture model $(\mathcal{T}_1, \dots, \mathcal{T}_K)$ is the intersection of the join of K toric varieties with the probability simplex Δ_{2^n-1}.

Model Selection: A Bayesian approach

- Choose the appropriate model M that best fits a given set of observations D.
- Choose M that maximizes the marginal likelihood:

$$p(D|M) = \mathbb{I}[N, Y_D, M] = \int_{\Omega} e^{N\mathcal{L}(Y_D|\omega)}\mu(\omega)d\omega.$$

- Ω denotes the domain of the model parameters ω.
- $\mu(\omega)$ is the prior parameter density, $N = |D|$.
- Y_D is the averaged sufficient statistics.
- $\mathcal{L}$ is the log–likelihood function of M.

BIC Score

- The quantity $\ln \mathbb{I}[N, Y_D, M]$ is called the Bayesian Information Criterion (BIC) for choosing a model M.
- In many cases, the BIC score gives an asymptotic approximation to this quantity [Schwarz 1978, Haughton 1988]

$$BIC = N \cdot \ln P(Y_D \mid w_{ML}) - \frac{d}{2} \ln N + O(1).$$

- $\ln P(Y_D \mid w_{ML})$ is the log-likelihood of Y_D given the ML parameters of the model.
- d is the number of independent parameters.

Asymptotic Approximation for the Marginal Likelihood

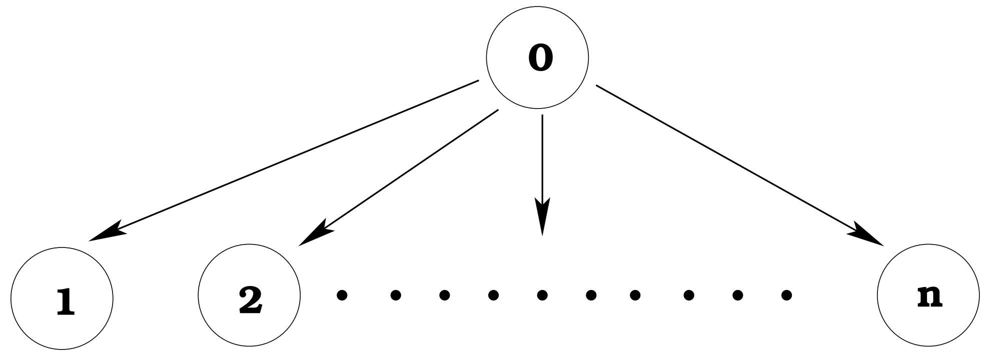
Theorem [Watanabe 2001]

Let $I(N) = \int_{W_\epsilon} e^{-Nf(w)}\mu(w)dw$ where W_ϵ is some closed ϵ-box around w_0, which is a minimum point of f in W_ϵ, and $f(w_0) = 0$. Assume that f and μ are analytic functions, $\mu(w_0) \neq 0$. Then,

$$\ln I(N) = \lambda_1 \ln N + (m_1 - 1) \ln \ln N + O(1)$$

where the rational number $\lambda_1 < 0$ and m_1 are the largest pole and its multiplicity of the analytic continuation of

$$J(\lambda) = \int_{f(w)<\epsilon} f(w)^\lambda \mu(w) dw \qquad Re(\lambda) > 0$$



- The star tree model is the **Segre** variety

$$S_{1,1,\ldots,1} \quad := \quad \mathbb{P}^1 \times \mathbb{P}^1 \times \cdots \times \mathbb{P}^1 \quad \subset \quad \mathbb{P}^{2^n-1}.$$

- The mixture of two copies of the star tree model is the secant variety of the Segre product of n projective spaces $\mathbb{P}^1$, denoted $S^2_{1,1,\ldots,1}$.

Theorem [Geiger and Rusakov 2002]

Let M be the mixture of two copies of the star tree model and Y be the sufficient statistics. Then for $n \geq 3$ as $N \longrightarrow \infty$:

- If Y is a smooth point of $S_{1,1,\ldots,1}$

$$\ln I[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{2n+1}{2} \ln N + O(1),$$

- If $Y \in S^2_{1,1} \times S_{\underbrace{1,\ldots,1}_{n-2}}$ (singularity)

$$\ln I[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{2n-1}{2} \ln N + O(1),$$

- If $Y \in S_{1,\ldots,1}$ (deepest singularity)

$$\ln I[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{n+1}{2} \ln N + O(1),$$

Road Map

- Asymptotic model selection for naive Bayesian networks, Rusakov-Geiger, UAI 2002.
- Automated analytic asymptotic evaluation of the marginal likelihood for latent models, Rusakov-Geiger, UAI 2003.
  - Algebraic analysis for nonidentifiable learning machines, Watanabe, Neural Computation 2001.
  - Automated resolution of singularities for hypersurfaces, Bodnar-Schicho, JSC 2000.
- Algebraic statistics in model selection, Garcia, UAI 2004.
- Mutagenetic tree models, Beerenwinkel-Drton, ASCB 2005.
- Secant varieties of toric varieties, Cox-Sidman, 2005.
- Combinatorial secant varieties, Sullivant-Sturmfels, 2005.
- Join varieties of toric varieties, in progress
- Algebraic Statistics for Computational Biology, Pachter-Sturmfels, eds. Cambridge 2005.
- Catalog of small trees, Casanellas-Garcia-Sullivant, ASCB 2005.