Applied algebraic geometry: algebraic statistics

Jan Draisma, Universität Bern

Theorem (Allman-Rhodes, Draisma-Kuttler)
Let $X \subseteq \mathbb{C}^{m\times k}$ and $Y \subseteq \mathbb{C}^{k\times n}$ be closed subvarieties, both GL_k-stable (via right- and left- multiplication, respectively). Then $\overline{X \cdot Y} = \{z \in \mathbb{C}^{m\times n} \mid z \cdot \mathbb{C}^{n\times k} \subseteq X \text{ and } \mathbb{C}^{k\times m} \cdot z \subseteq Y \text{ and } \mathrm{rk} z \leq k\}$, and the corresponding statement holds at the level of ideals, as well.

This reduces the study of equations for phylogenetic models on general trees to that of claw trees. For trivalent trees with all nodes taking 4 values, the Friedman-Gross result gives equations for all tree models.

Proof
Suppose f vanishes on $X \cdot Y$. Write $M = \mathbb{C}^{m\times k}$ and $N = \mathbb{C}^{k\times n}$ and $P = \mathbb{C}^{m\times n}$, and let $\mu : M \times N \to P$ denote multiplication. Set $\mu^* f =: h \in \mathbb{C}[M \times N]$. Then $h = h_1 + h_2$ with $h_1 \in I(X \times N)$ and $h_2 \in I(M \times Y)$. Note that h is GL_k-invariant.

Now apply the Reynolds operator $\rho : \mathbb{C}[M \times N] \to \mathbb{C}[M \times N]^{\mathrm{GL}_k}$

This yields $h = h'_1 + h'_2$ where $h'_1 \in I(X \times N)^{\mathrm{GL}_k}$ and $h'_2 \in I(M \times Y)^{\mathrm{GL}_k}$ (here we use that X, Y are GL_k-stable, hence so are the ideals $I(X \times N)$ and $I(M \times Y)$).

Now $\mu^* : \mathbb{C}[P] \to \mathbb{C}[M \times N]^{\mathrm{GL}_k}$ is surjective (First Fundamental Theorem in Invariant Theory), so $h'_i = \mu^*(f_i)$, i.e., $f_i(ab) = h'_i(a, b)$ for all matrices a, b.

Now $f_1(X \cdot N) = \{0\} = f_2(M \cdot y)$, and on the other hand $\mu^*(f - f_1 - f_2)$ vanishes identically, so $f - f_1 - f_2$ lies in the ideal generated by $(k+1) \times (k+1)$-determinants. (SFT) □

Jointly Gaussian random variables

Start with n independent, standard-normal distributed random variables $Y_1, \ldots, Y_n$. Their joint distribution is given by the density function $f_Y(y) := \prod_{i=1}^{k} g(y_i)$ where $g(y) := \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$ is the density of a standard normal scalar random variable.

Rewrite this as $f_Y(y) = \frac{1}{-(2\pi)^{n/2}} e^{y^T y/2}$.

Now let $A \in \mathrm{GL}_n(\mathbb{R})$. What is the density of the random variable $Z := AY$? It is the unique continuous function f_Z such that for all measurable $U \subseteq \mathbb{R}^n$ we have $\int_{z \in U} f_Z(z) dz = \int_{y \in A^{-1}U} f_Y(y) dy$ (lhs is $\mathrm{Prob}(z \in U)$ and rhs is $\mathrm{Prob}(Y \in A^{-1}U)$.

Change of variables formula yields: $f_Z(z) = \frac{1}{|\det(A)|} f_Y(A^{-1}z) = \frac{1}{|\det(A)|(2\pi)^{n/2}} e^{-z^T A^{-T} A^{-1} z/2} = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-z^T \Sigma^{-1} z/2}$ where $\Sigma = AA^T$ is the positive definite *covariance matrix* of Z.

Algebraic Gaussian models arise by putting algebraic conditions on Σ. For a subset $S \subseteq [n]$ write Z_S. This vector is multivariate Gaussian with covariant matrix Σ_S, the $S \times S$-submatrix of Σ.

Take a DAG $G = ([n], E)$ as before. Now $Z_i \perp\!\!\!\perp Z_{\mathrm{nd}(i)\setminus \mathrm{pa}(i)} \mid \mathrm{pa}(i)$ means the following (write $N = \mathrm{nd}(i) \setminus \mathrm{pa}(i)$ and $P = \mathrm{pa}(i)$):

$$f_{Z_{\{i\}\cup N}|Z_P=z_P}(z_{\{i\}\cup N}) = f_{Z_i|Z_P=z_P}(z_i) f_{Z_N|Z_P=z_P}(z_N)$$

One can check that this holds if and only if the X admit the following parameterisation: write $X_j = \sum_{i\in \mathrm{pa}(j)} \lambda_{ij} X_i + \epsilon_j$ where the ϵ_j are independent, Gaussian distributed, with variance ω_j.

Then $X = \Lambda^T X + \epsilon$ where $\lambda_{ij} = 0$ if $i \notin \mathrm{pa}(j)$, and hence $X = (I - \Lambda)^{-T}\epsilon$, with covariance matrix $\Sigma = (I - \Lambda)^{-T}\Omega(I - \Lambda)^{-1}$ where $\Omega = \mathrm{diag}(\omega_1, \ldots, \omega_n)$

A *trek* in G is a sequence $i_0 \leftarrow \cdots \leftarrow i_m \rightarrow i_{m+1} \rightarrow \cdots i_{m+l}$. Both m and l may be zero.

The corresponding *trek monomial* is
$\lambda_{i_1,i_0} \cdots \lambda_{i_m,i_{m-1}} \omega_{i_m} \lambda_{i_m,i_{m+1}} \cdots \lambda_{i_{m+l-1},i_{m+l}}$

Prop
σ_{ij} is the sum of all track monomials corresponding to treks from i to j, each multiplied with $(-1)^{m+l}$ (where $m + l$ is the length of the trek).

Theorem (Sullivant, Talaska)
Let $A, B \subseteq [n]$ of equal cardinality k. Then $\det(\Sigma_{A,B})$ is identically zero on the model if and only if there are no k *tracks* starting within A and ending within B and without *sided intersection*: any two of the treks are vertex disjoint in their up-parts as well as in their down-parts.