Applied algebraic geometry: algebraic statistics

Jan Draisma, Universität Bern

Setting

- finite, directed acyclic graph $G = (V, E)$ (DAG)
- for every $i \in V$ a random variable X_i; for $S \subseteq V$ write X_S

We'll discuss two scenarios:

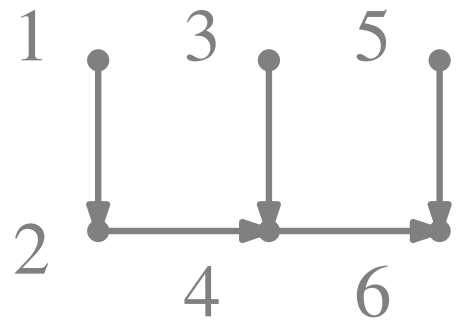
- either all X_i take finitely many values, in $[r_i]$ say
- or the X_i are jointly Gaussian with mean zero

Write $\mathrm{pa}(i)$ for the set of $j \in V$ with $j \to i$ an arrow in G—the *parents* of i, and $\mathrm{nd}(i)$ for the *non-descendents* of i.

In both scenarios we'll discuss the implications of the following hypothesis: $X_i \perp\!\!\!\perp X_{\mathrm{nd}(i)\setminus\mathrm{pa}(i)} \mid X_{\mathrm{pa}(i)}$ for all $i \in V$, and we'll give examples with latent variables.

Discrete setting

Write $Y_i = X_{\mathrm{nd}(i)\setminus\mathrm{pa}(i)}$. Then $X_i \perp\!\!\!\perp Y_i \mid X_{\mathrm{pa}(j)}$ means that $\mathrm{Prob}(X_i = x \wedge Y = y \mid X_{\mathrm{pa}(i)} = z) = \underbrace{\mathrm{Prob}(X_i = x \mid X_{\mathrm{pa}(i)} = z)}_{\theta_{i,x,z}} \cdot \mathrm{Prob}(Y_i = y \mid X_{\mathrm{pa}(i)} = z)$.



Example

$$\mathrm{Prob}(X_V = x) = \mathrm{Prob}(X_6 = x_6 \wedge X_{45} = x_{45} \wedge X_{123} = x_{123}) =$$
$$\mathrm{Prob}(X_6 = x_6 \wedge X_{123} = x_{123} \mid X_{45} = x_{45}) \cdot \mathrm{Prob}(X_{45} = x_{45}) =$$
$$\theta_{6,x_6,x_{45}} \cdot \mathrm{Prob}(X_{12345} = x_{12345}) = \theta_{6,x_6,x_{45}}\theta_{5,x_5}\theta_{4,x_4,x_{23}}\theta_{3,x_3}\theta_{2,x_2,x_1}\theta_{1,x_1}$$

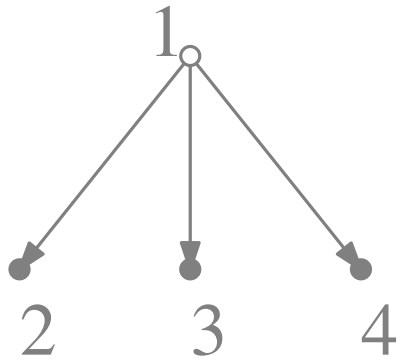
This is a proof by example of the *recursive factorisation theorem*.

Detailed study of the ideal of the image (when one imposes the condition that $\sum_x \theta_{i,x,z} = 1$) was done by Garcia-Sturmfels-Stillman.

Example of latent variables: phylogenetics

X_1, X_2, X_3, X_4 all take values in $\{A, C, G, T\}$

$\text{Prob}(X_{1234} = x_{1234}) = \theta_{2,x_2,x_1}\theta_{3,x_3,x_1}\theta_{4,x_4,x_1}\theta_{1,x_1}$



Forget about constraints on the θ. Then for fixed x_1 this is the (x_2, x_3, x_4)-entry of a general rank-1 tensor of format $4 \times 4 \times 4$.

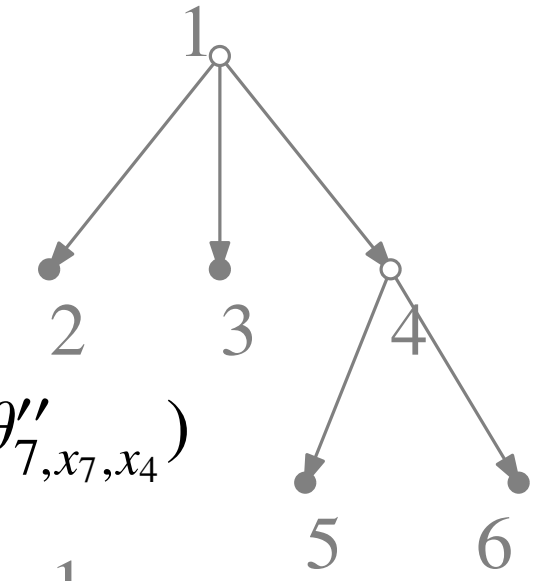
But now imagine X_4 is a random variable associated to DNA of an extinct species, and cannot be observed. Then what counts is the *marginal* distribution:

$\text{Prob}(X_{234} = x_{234}) = \sum_{x_1 \in \{A,C,G,T\}} \text{Prob}(X_{1234} = x_{1234})$, which is the x_2, x_3, x_4-entry of a general *rank-four tensor*.

Set-theoretic equations for rank-four tensors of forrmat $4 \times 4 \times 4$ were found by Friedman-Gross.

The Allman-Rhodes reduction

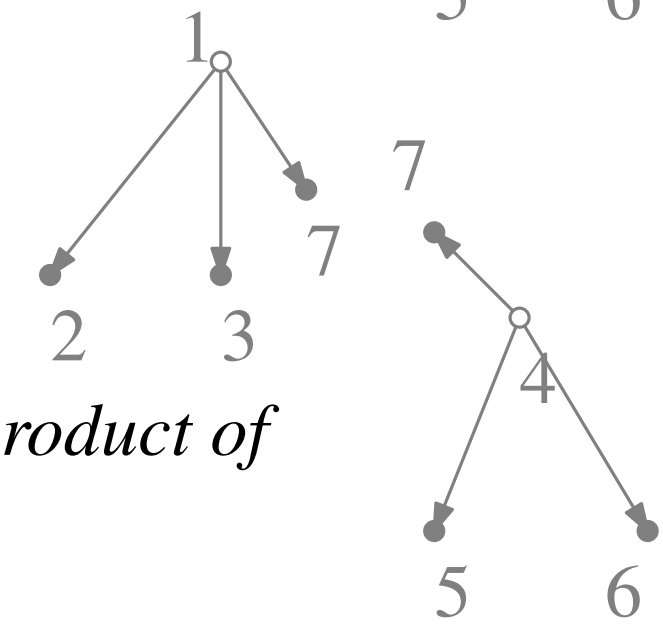
What about bigger trees?



$$\sum_{x_1,x_4} \theta_{2,x_2,x_1}\theta_{3,x_3,x_1}\theta_{4,x_4,x_1}\theta_{5,x_5,x_4}\theta_{6,x_6,x_4}$$

$$= \sum_{x_7}(\sum_{x_1} \theta_{2,x_2,x_1}\theta_{3,x_3,x_1}\theta'_{7,x_7,x_1})(\sum_{x_4} \theta_{5,x_5,x_4}\theta_{6,x_6,x_4}\theta''_{7,x_7,x_4})$$

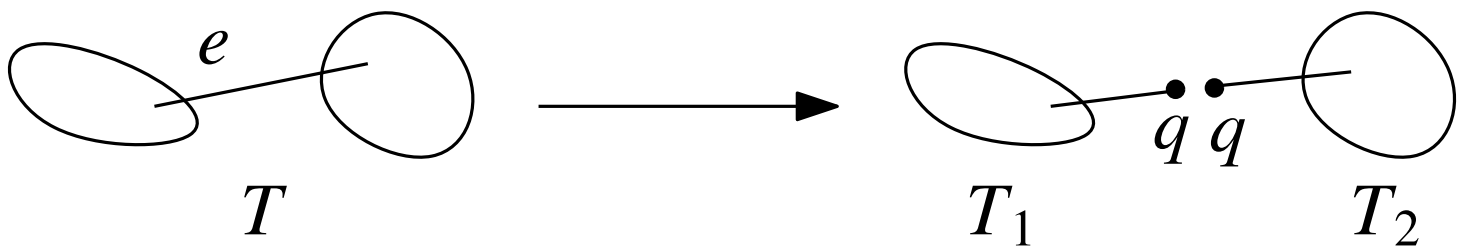
provided that $\theta_{4,x_4,x_1} = \sum_{x_7} \theta'_{7,x_7,x_1} \cdot \theta''_{7,x_7,x_4}$



The model for the bigger tree is the matrix product of the models for the smaller trees.

More precisely: for every finite tree T let X_T be the Zariski closure inside $V^{\otimes \text{leaves}(T)}$ of the set of probability tensors obtained as the parameters θ vary. Here $V = \mathbb{C}^4$ with the standard bilinear form $(.|.)$.

Pick an edge $e \in T$ and split T as follows:



Lemma

$X_T = X_{T_1} \cdot X_{T_2}$ where $\cdot$ is the bilinear map $V^{\otimes L_1} \otimes V^{\otimes \{q\}} \times V^{\otimes \{q\}} \otimes V^{\otimes L_2} \to V^{\otimes L}$ given by $(A \otimes v, w \otimes B) \mapsto (v|w)A \otimes B$

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.*