Algebro-geometric problems arising from statistics

Jan Draisma

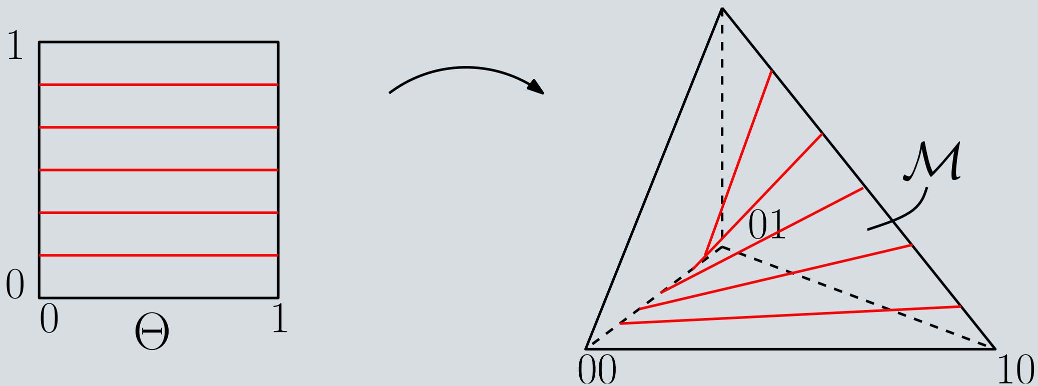
Groningen, 16 March 2010

Parametric statistical models

Θ : parameter space
Ω : sample space
$\mathcal{M} = \{P_\theta,\ \theta \in \Theta\}$ probability distributions on Ω

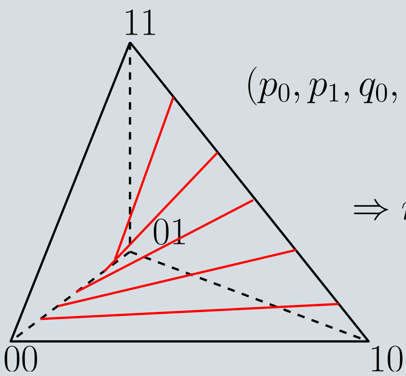
often $\Theta \subseteq \mathbb{R}^d$ semi-algebraic and $\theta \mapsto P_\theta$ "polynomial"
$\rightsquigarrow$ "algebraic statistics"

Vignette:
$(p_0, p_1, q_0, q_1) \mapsto (p_0 q_0, p_0 q_1, p_1 q_0, p_1 q_1)$



Types of algebro-geometric problems

equations for $\{P_\theta \mid \theta \in \Theta\}$?
$\rightsquigarrow$ test $P_{\text{empirical}} = P_\theta$ for some θ?
$\rightsquigarrow$ implicitisation problem



$$(p_0, p_1, q_0, q_1) \mapsto (p_0q_0, p_0q_1, p_1q_0, p_1q_1)$$
$$=: (r_{00}, r_{01}, r_{10}, r_{11})$$

$$\Rightarrow r_{00}r_{11} - r_{01}r_{10} = 0$$

identifiability?
$\rightsquigarrow$ too many parameters?

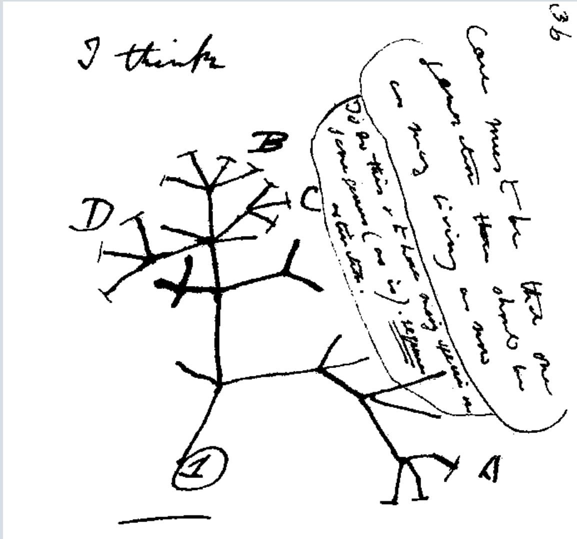
relations among models in **families**?
$\rightsquigarrow$ increasing # random variables
...

Today:
two theorems
one well-known conjecture
(perhaps one puzzle)

General Markov Model

$T = (V, E, r)$: finite (rooted) tree
$\Omega_v,\ v \in V$: finite sets
$\Theta = \{[\pi_r, (A_{u \to v})_{u \to v \in E}]\}$
$A_{u \to v}$ stochastic matrix $X_u \to X_v$
$\Omega = \prod_{v \text{ leaf of } T} \Omega_v$

$\theta = (A_{u \to v}) \mapsto P_\theta$:
π_r on Ω_r
$u \to v$ mutate according to $A_{u \to v}$



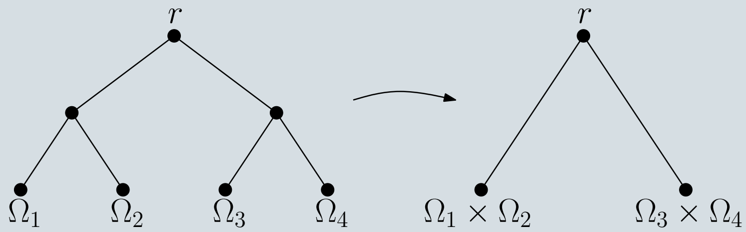
[Transmutation of species, 1837]

```
Gene: BCCIP (ENSPTRG00000003043)
Location: Chromosome 10: 126,795,545-126,826,189 forward strand.
Pan_troglodytes CGGGCCCCTGCACGCCCGCGGCCTCGG.....
Homo_sapiens    CGGGCCCCTGCACGCCCGCGGCTTCGG.....
Pongo_pygmaeus  CGGGCCCCTGCACGACCGCGGCCTCGG.....
Macaca_mulatta  TGGGCCCCTGCATGCCCGCGGCCTCGG.....
[Data from www.ensembl.org]
```

General Markov Model, continued

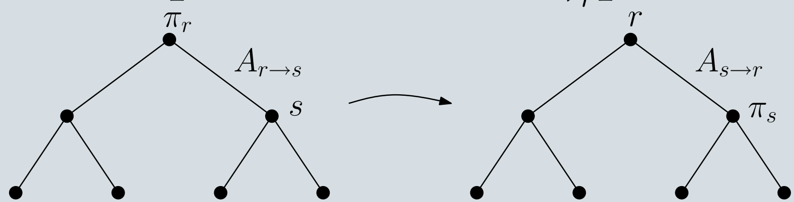
Two operations:

flattening at r
$\mathcal{M}(T) \subseteq \mathcal{M}(\flat_r T)$
$\mathrm{eqs}(\mathcal{M}(T)) \supseteq \mathrm{eqs}(\mathcal{M}(\flat_r(T)))$



T $\qquad$ $\flat_r T$

moving r:
$\mathcal{M}(T, r) = \mathcal{M}(T, s)$



T, r $\qquad$ T, s

$\rightsquigarrow \mathcal{M}(T) \subseteq \mathcal{M}(\flat_v T)$ for all $v \in V(T)$

Theorem (Allman-Rhodes 2008) (D-Kuttler 2009):
$\mathcal{M}(T) = \bigcap_{v \in V(T)} \mathcal{M}(\flat_v T)$ $\qquad$ $\mathrm{eqs}(\mathcal{M}(T)) = \sum_{v \in V(T)} \mathrm{eqs}(\mathcal{M}(\flat_v T))$

Gaussian factor analysis

$Z_1, \dots, Z_k \sim \mathcal{N}(0,1)$ independent factors
$X_1, \dots, X_n$ observed
$X_i = \sum_{j=1}^k s_{ij} Z_j + \epsilon_i$
$\epsilon_i \sim \mathcal{N}(0, v_i)$ independent noise
$\theta \mapsto P_\theta$:
$(S, v) \mapsto SS^T + \mathrm{diag}(v)$
$\mathcal{M}_{k,n} = \{\Sigma = SS^T + \mathrm{diag}(v) \mid (S,v) \in \Theta\}$

Raymond Cattell (1971): **fluid** vs **crystallised** intelligence

Proposal (Drton,Sturmfels,Sullivant 07): use polynomial relations among entries of Σ to test the model against data

Example $k = 2, n = 5$**:**
$\sum_{\pi \in \mathrm{Sym}(5)} \mathrm{sgn}\,(\pi) \sigma_{\pi(1)\pi(2)} \sigma_{\pi(2)\pi(3)} \sigma_{\pi(3)\pi(4)} \sigma_{\pi(4)\pi(5)} \sigma_{\pi(5)\pi(1)} = 0$, the **pentad**

Gaussian factor analysis, continued

Observation:
$\Sigma \in \mathcal{M}_{k,n} \Rightarrow \Sigma[I] \in \mathcal{M}_{k,\#I}$

Question:
equations for $\mathcal{M}_{k,n}$ as $n \longrightarrow \infty$?
$\exists ? n_0 \, \forall n \geq n_0$ all equations for $\mathcal{M}_{k,n}$ are generated by those for $\mathcal{M}_{k,n_0}$

Theorem:
yes for $k = 1$ ($n_0 = 4$, de Loera, Sturmfels, Thomas 1995)
yes for $k = 2$ ($n_0 = 6$, Brouwer-D, 2010)
set-theoretically yes for all k (D, 2010)

Independence and its first mixture

Independence:
$X_1, X_2, X_3, \ldots$ binary, independent
assume $\lambda := P(\forall i : X_i = 0) > 0$
$p_i := P(X_i = 1)/P(X_i = 0)$
$p_I := \lambda \prod_{i \in I} p_i,\ I \subseteq \mathbb{N}$ finite
polynomial relations:
$p_I p_J - p_K p_L = 0$ if $I \dot\cup J = K \dot\cup L$

Second copy:
$Y_1, Y_2, Y_3, \ldots$ binary, independent
$q_i := P(Y_i = 1)/P(Y_i = 0)$, q_I etc.

Mixture:
H binary, $P(H = 1) = s$

$$Z_i := \begin{cases} X_i & \text{if } H = 0 \\ Y_j & \text{if } H = 1 \end{cases}$$

Independence and its first mixture, continued

$r_I := P(Z_i = 1 \text{ if } i \in I \text{ and } 0 \text{ if } i \notin I) = (1-s)p_I + sq_I$

Polynomial relations?
Certainly $\sum_{\pi \in S_3} \operatorname{sgn}(\pi) r_{I_1 \cup J_{\pi(1)}} r_{I_2 \cup J_{\pi(2)}} r_{I_3 \cup J_{\pi(3)}}$
where $(I_1 \cup I_2 \cup I_3) \cap (J_1 \cup J_2 \cup J_3) = \emptyset$

Conjecture (Garcia, Stillman, and Sturmfels 2005):
these cubic determinants generate all equations among the r_I

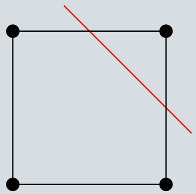
Remark:
proved set-theoretically by Landsberg and Manivel 2004

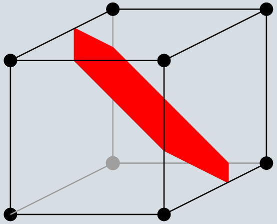
(Realistic?) Dream:
a finite computer calculation might settle GSS!

Higher mixtures of independence

Puzzle:
take $n \neq 4$
write $2^n = q(n+1) + r$ with $0 \leq r < n+1$
by repeated cutting with hyperplanes
decompose $\{0,1\}^n \subseteq \mathbb{R}^n$ into:
q affinely independent $(n+1)$-sets
1 affinely independent $2^n - k(n+1)$-set

$\rightsquigarrow$ identifiability of higher (tropical) mixtures of independence





Bibliography

[1] Elizabeth S. Allman and John A. Rhodes. Phylogenetic ideals and varieties for the general Markov model. *Adv. Appl. Math.*, 40(2):127–148, 2008.

[2] Andries E. Brouwer and Jan Draisma. Equivariant Gröbner bases and the two-factor model. *Math. Comput.*, 2010. To appear; preprint available from `http://arxiv.org/abs/0908.1530`.

[3] Jan Draisma. A tropical approach to secant dimensions. *J. Pure Appl. Algebra*, 212(2):349–363, 2008.

[4] Jan Draisma. Finiteness for the k-factor model and chirality varieties. *Adv. Math.*, 223:243–256, 2010.

[5] Jan Draisma and Jochen Kuttler. On the ideals of equivariant tree models. *Math. Ann.*, 344(3):619–644, 2009.

[6] Mathias Drton, Bernd Sturmfels, and Seth Sullivant. Algebraic factor analysis: tetrads, pentads and beyond. *Probab. Theory Relat. Fields*, 138(3–4):463–493, 2007.

[7] Mathias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on Algebraic Statistics*. Oberwolfach Seminars 39. Birkhäuser, Basel, 2009.

[8] Luis D. Garcia, Michael Stillman, and Bernd Sturmfels. Algebraic geometry of Bayesian networks. *J. Symb. Comp.*, 39(3–4):331–355, 2005.

[9] Joseph M. Landsberg and Laurent Manivel. On the ideals of secant varieties of Segre varieties. *Found. Comput. Math.*, 4(4):397–422, 2004.