Set-theoretic finiteness for the k-factor model

Jan Draisma

Technische Universiteit
Eindhoven
University of Technology

Bremen, 20 January 2009

Statistics $\rightsquigarrow$ algebra

Phylogenetics

genetic data $\rightsquigarrow$ evolutionary tree

Factor analysis

many test results $\rightsquigarrow$ few types of intelligence

Message

Statistics $\rightsquigarrow$ beautiful algebra challenges ($\rightsquigarrow$ statistics)

Phylogenetics

Statistical problem

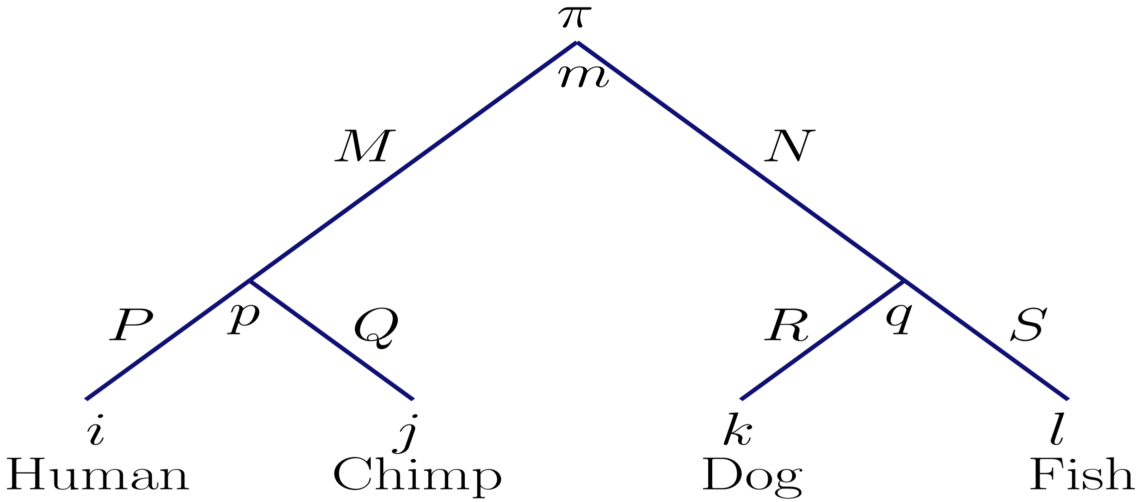
Test whether n aligned strings of DNA *match* a hypothetical tree T.

Approach

Data $\rightsquigarrow$ *empirical distribution* P on $\{A, C, G, T\}^n$
Tree $\rightsquigarrow$ *parameterised family* F_T of distributions
To test: $P \in F_T$?

Phylogenetics, continued

Parameterisation



$$\Pr(i,j,k,l) = \Sigma_{m,p,q} \pi_m M_{pm} P_{ip} Q_{jp} N_{qm} R_{kq} S_{lq}$$

Classical approach

Maximum-likelihood estimates for parameters

Phylogenetics, continued

Observation

F is parameterised by polynomials, hence an *algebraic variety*.

"Ideal approach to biology"

[Barry Cipra, *SIAM News*, Summer 2007]

Find *polynomial equations* for F_T (and test on P).

Phylogenetics, selected results

[Allman-Rhodes 2004, Sturmfels-Sullivant 2005, Casanellas-Sullivant 2005, Draisma-Kuttler 2008]

Theorem

1. Equations for *stars* $\rightsquigarrow$ equations for arbitrary trees
2. Equations for stars in certain models with symmetry

Allman's fish problem

Find all equations for $\mathrm{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$ and win a Smoked Copper River Salmon.

The k-factor model

Statistical problem

Test whether n (large) observed, jointly Gaussian variables, are pairwise independent given k (small) hidden Gaussian variables.

Approach

Data $\rightsquigarrow$ $n \times n$ empirical covariance matrix Σ
k-factor model $\rightsquigarrow$ family $F_{k,n}$ of $n \times n$-matrices
To test: $\Sigma \in F_{k,n}$?

TU/e Technische Universiteit Eindhoven University of Technology

The k-factor model, continued

Parameterisation

$$F_{k,n} = \{Y = D + SS^T \mid D > 0 \text{ diagonal and } S \in \mathbb{R}^{k \times n}\}$$

Classical approach

Maximum-likelihood estimates for D, S

Ideal approach

Find *polynomial equations* of $F_{k,n}$ (and test on Σ).

The k-factor model, some results

Theorem

[De Loera-Sturmfels-Thomas 1995]
For all n, ideal of $F_{1,n}$ is generated by off-diagonal 2×2-minors $y_{ij}y_{kl} - y_{il}y_{kj}$ (*tetrads*).

Theorem

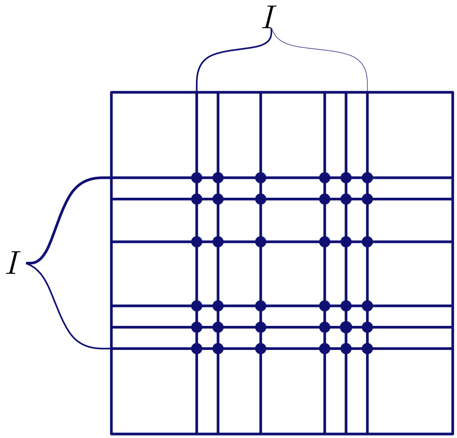
[Drton-Sturmfels-Sullivant 2007]
For $5 \leq n \leq 9$ ideal of $F_{2,n}$ is generated by off-diagonal 3×3-minors and *pentads*:

$$\sum_{\pi \in \mathrm{Sym}(5)} \mathrm{sgn}\,(\pi) y_{\pi(1)\pi(2)} y_{\pi(2)\pi(3)} \cdots y_{\pi(5)\pi(1)}.$$

The 2-factor model, finiteness

Observation

$I \subseteq \{1,\ldots,n\}$ and $Y \in F_{k,n} \rightsquigarrow Y[I] \in F_{k,|I|}$



$D[I] + S[I]S[I]^T$

Theorem

[Drton-Xiao 2008]
For $n \geq 6$:
$Y \in F_{2,n} \Leftrightarrow \forall_{I,|I|=6} Y[I] \in F_{2,n}$.

The k-factor model, fin(iten)esses?

Questions

1. Does there exist $n_0 = n_0(k)$ such that $\forall n \geq n_0$: $Y \in F_{k,n} \Leftrightarrow \forall_{I,|I|=n_0} Y[I] \in F_{k,n_0}$?
2. Same question for the *Zariski closure* of $F_{k,n}$
3. Same question for the *scheme* $F_{k,n}$

Aside

From *Cambridge Advanced Learner's Dictionary*:
finesse verb [T]: to deal with a situation or a person in a skilful and often slightly dishonest way.

TU/e Technische Universiteit Eindhoven University of Technology

The k-factor model, finiteness

Theorem

[Draisma 2008]
For all k there exists an $n_0(k)$ such that $\forall n \geq n_0$:
$Y \in \overline{F_{k,n}} \Leftrightarrow \forall_{I,|I|=n_0} Y[I] \in \overline{F_{k,n_0}}$.

Disclaimer

1. No obvious bound for n_0
2. No obvious implication for $F_{k,n}$
3. Not scheme-theoretically

G-Noetherianity

Definition

G group acting on ring R
R is G-*Noetherian* if every chain

$$I_1 \subseteq I_2 \subseteq \ldots$$

of G-stable ideals stabilises.

Theorem

[Aschenbrenner-Hillar 2007, Hillar-Sullivant 2008]
$\mathbb{R}[u_1, u_2, \ldots, v_1, v_2, \ldots, \ldots, z_1, z_2, \ldots]$ is $\mathrm{Sym}(\mathbb{N})$-Noetherian.

Technische Universiteit Eindhoven University of Technology

G-Noetherianity, continued

Idea

Prove that $\overline{F_{k,\infty}}$ is a $\mathrm{Sym}(\mathbb{N})$-stable subvariety of some $\mathrm{Sym}(\mathbb{N})$-Noetherian variety, which itself is finitely characterised.

Wishful thinking

1. Is $\mathbb{R}[y_{ij} \mid i, j = 1, 2, \ldots]$ $\mathrm{Sym}(\mathbb{N})$-Noetherian? NO!
2. Is $\mathbb{R}[y_{ij}]/\langle$off-diagonal $(k+1) \times (k+1)$-minors$\rangle$ $\mathrm{Sym}(\mathbb{N})$-Noetherian?
   Perhaps, but not even clear for $k = 1$.

The k-factor model, proof idea

Theorem

The $\mathbb{N} \times \mathbb{N}$-matrices over $\mathbb{R}$ defined by the off-diagonal $(k+1) \times (k+1)$-minors form a $\mathrm{Sym}(\mathbb{N})$-Noetherian topological space.

Proof uses Aschenbrenner-Hillar-Sullivant, induction on k, and the following lemma.

Lemma

If X an H-Noetherian H-space and $G \supseteq H$, then $G \times_H X$ is G-Noetherian.

Conclusions

Algebra challenges from statistics

1. Implicitisation: parameterisation $\rightsquigarrow$ equations
2. Finiteness: small $\rightsquigarrow$ large models
3. Maximum-likelihood equations

Relevant tools

1. Invariant theory
2. G-Noetherianity
3. Computational algebra
4. Toric and tropical geometry
5. Secant varieties