

# Open Problems in Algebraic Statistics

BERND STURMFELS

UNIVERSITY OF CALIFORNIA, BERKELEY

*Applications of Algebraic Geometry*

Special year at IMA Minneapolis 2006-07

Workshop on **Biology, Dynamics and Statistics**

March 5, 2007

# Models with Hidden Variables

## A Concrete Problem:

Consider the variety of  $4 \times 4 \times 4$ -tables of tensor rank at most 4. Do the known polynomial invariants of degree **five** and **nine** suffice to cut out this variety?

[ASCB, Conjecture 3.24, page 102]

## The General Problem:

Study the geometry and commutative algebra of graphical models with hidden random variables.

Construct these varieties by gluing familiar secant varieties, and by applying representation theory.

# My favorite statistical model

lives in the 64-dimensional space  $\mathbb{C}^4 \otimes \mathbb{C}^4 \otimes \mathbb{C}^4$   
of  $4 \times 4 \times 4$ -tables  $(p_{ijk})$ , where  $i, j, k \in \{A, C, G, T\}$ .

It is the variety with parametric representation

$$p_{ijk} = \rho_{Ai} \cdot \sigma_{Aj} \cdot \theta_{Ak} + \rho_{Ci} \cdot \sigma_{Cj} \cdot \theta_{Ck} + \rho_{Gi} \cdot \sigma_{Gj} \cdot \theta_{Gk} + \rho_{Ti} \cdot \sigma_{Tj} \cdot \theta_{Tk}$$

Our problem is to compute its homogeneous prime ideal  $I$  in the polynomial ring with 64 unknowns,

$$\mathbb{Q}[p_{AAA}, p_{AAC}, p_{AAT}, \dots, p_{TTG}, p_{TTT}].$$

# Why (Pure) Mathematics?

“Mathematics is the Art of Giving the Same Name to Different Things” (H. Poincaré).

- the set of  $4 \times 4 \times 4$ -tables of tensor rank  $\leq 4$
- mixture of independent random variables
- the naive Bayes model
- the CI model  $[X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3 \mid Y]$ .
- third secant variety of Segre variety  $\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3$
- general Markov model for phylogenetic tree  $K_{1,3}$
- entanglement of pure states (mean fields)

# Invariants of Degree Five

Consider any  $3 \times 4 \times 4$ -subtable of  $(p_{ijk})$  and let  $A, B, C$  be the  $4 \times 4$ -slices gotten by fixing  $i$ . On our model, the following identity of  $4 \times 4$ -matrices holds:

$$A \cdot B^{-1} \cdot C = C \cdot B^{-1} \cdot A$$

After clearing the denominator  $\det(B)$ , this gives 16 quintic polynomials which lie in our prime ideal  $I$ .

**Proposition 1.** *The space of quintic polynomials in  $I$  has dimension 1728. As a  $GL(\mathbb{C}^4)^3$ -module, it equals*

$$\begin{aligned} & S_{311}(\mathbb{C}^4) \otimes S_{2111}(\mathbb{C}^4) \otimes S_{2111}(\mathbb{C}^4) \\ \oplus & S_{2111}(\mathbb{C}^4) \otimes S_{311}(\mathbb{C}^4) \otimes S_{2111}(\mathbb{C}^4) \\ \oplus & S_{2111}(\mathbb{C}^4) \otimes S_{2111}(\mathbb{C}^4) \otimes S_{311}(\mathbb{C}^4) \end{aligned}$$

# Invariants of Degree Nine

Consider any  $3 \times 3 \times 3$ -subtable  $(A, B, C)$  of  $(p_{ijk})$ .  
On our model, the following  $3 \times 3$ -matrix is singular:

$$A \cdot B^{-1} \cdot C - C \cdot B^{-1} \cdot A$$

The numerator of its determinant is a degree nine polynomial in  $I$  known as the **Strassen invariant**.

**Proposition 2.** *The  $GL(\mathbb{C}^4)^3$ -submodule of  $I_9$  generated by the Strassen invariant is not contained in  $\langle I_5 \rangle$ . This module has dimension 8000 and it equals*

$$S_{333}(\mathbb{C}^4) \otimes S_{333}(\mathbb{C}^4) \otimes S_{333}(\mathbb{C}^4).$$

References: [Garcia-Stillman-St], [Allman-Rhodes],  
[Landsberg-Manivel], [Landsberg-Weyman]

# Maximum Likelihood

## A Concrete Problem:

Characterize all projective varieties whose maximum likelihood degree is equal to one.

[HKS: *Solving the Likelihood Equations*, Problem 13]

## The General Problem:

Study the geometry of maximum likelihood estimation for algebraic statistical models.

What makes a model nice? Is the ML degree related to convergence properties of the EM algorithm?

# MLE in Projective Space

Fix projective space  $\mathbb{P}^n$  with coordinates  $(p_0 : p_1 : \cdots : p_n)$ . The  $n$ -dimensional **probability simplex** is identified with  $\mathbb{P}_{\geq 0}^n$ .

For any data vector  $(u_0, u_1, \dots, u_n) \in \mathbb{N}^{n+1}$  we consider the likelihood function

$$L = \frac{p_0^{u_0} \cdot p_1^{u_1} \cdot p_2^{u_2} \cdots p_n^{u_n}}{(p_0 + p_1 + \cdots + p_n)^{u_0 + u_1 + \cdots + u_n}}$$

This is a well-defined rational function on  $\mathbb{P}^n$ .

Q: What are the critical points of this function?

# Algebraic Statistical Models

are subvarieties of the projective space  $\mathbb{P}^n$ .

The *maximum likelihood degree* of a model  $\mathcal{M}$  is the number of critical points of the restriction to  $\mathcal{M}$  of

$$L = \frac{p_0^{u_0} \cdot p_1^{u_1} \cdot p_2^{u_2} \cdots p_n^{u_n}}{(p_0 + p_1 + \cdots + p_n)^{u_0 + u_1 + \cdots + u_n}}$$

Here we only count critical points that are not poles or zeros, and  $u_0, u_1, \dots, u_n$  are assumed to be generic.

If  $\mathcal{M}$  is smooth and the divisor of  $L$  is normal crossing then there is a formula for the ML degree.

[Catanesi-Hoşten-Khetan-St.]

Otherwise, use resolution of singularities ????

# Determinantal Varieties

are models  $\mathcal{M}$  that are specified by imposing rank conditions on a matrix of unknowns.

**Example.**  $2 \times 2$ -matrices  $\begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$  of rank  $\leq 1$ .

**Independence for two binary random variables.** The ML degree is **one**, i.e., the MLE is a rational function in the data **(Take normalized row and column sums)**

**Example.**  $3 \times 3$ -matrices  $(p_{ij})$  of rank  $\leq 2$ . **Mixture of two ternary random variables.** The ML degree is **ten**.

**Open Problem.** Determine the ML degree of the variety of  $m \times n$ -matrices of rank  $\leq r$ .

# The 100 Swiss Francs Problem

*“Our data are two aligned DNA sequences ...*

ATCACCAAACATTGGGATGCCTGTGCATTTGCAAGCGGCT

ATGAGTCTTAAACGCTGGCCATGTGCCATCTTAGACAGCG

*... test the hypothesis that these two sequences were generated by DiaNA using one biased coin and four tetrahedral dice.... ”*

[ASCB, Example 1.16]

**Model:** Positive  $4 \times 4$ -matrices  $(p_{ij})$  of rank  $\leq 2$ .

**True or False:** The matrix  $(\hat{p}_{ij}) = \frac{1}{40} \begin{pmatrix} 3 & 3 & 2 & 2 \\ 3 & 3 & 2 & 2 \\ 2 & 2 & 3 & 3 \\ 2 & 2 & 3 & 3 \end{pmatrix}$

$$\text{maximizes } L = \left( \prod_i p_{ii} \right)^4 \left( \prod_{i \neq j} p_{ij} \right)^2 \left( \sum_{i,j} p_{ij} \right)^{-40}.$$

# Gaussian CI Models

## A Concrete Problem:

Which sets of almost-principal minors can be zero for a positive definite symmetric  $5 \times 5$ -matrix?

[Matùš, Studený, Drton, Sullivant, Parrilo, ...]

## The General Problem:

Study the geometry of conditional independence models for multivariate Gaussian random variables.

Which semigraphoids can be realized by Gaussians?

# Almost-Principal Minors

A *multivariate Gaussian* with mean zero is specified by its covariance matrix  $\Sigma = (\sigma_{ij})$ . This is a **positive definite  $n \times n$ -matrix**: all principal minors are positive.

An *almost-principal minor* of  $\Sigma$  has row indices  $\{i\} \cup K$  and column indices  $\{j\} \cup K$  for some  $K \subset [n]$  and  $i, j \in [n] \setminus K$ . It is denoted  $[i \perp\!\!\!\perp j | K]$ .

**Lemma.** The determinant  $[i \perp\!\!\!\perp j | K]$  is zero if and only if the random variable  $i$  is independent of the random variable  $j$  given the joint random variable  $K$ .

# Subvarieties of the PSD cone

Let  $\text{PSD}_n$  denote the  $\binom{n+1}{2}$ -dimensional cone of positive semidefinite symmetric  $n \times n$ -matrices.

A *Gaussian conditional independence model* is a subvariety of  $\text{PSD}_n$  defined by equations

$$[i \perp\!\!\!\perp j \mid K] = 0.$$

In algebraic geometry, we would study these varieties over the complex numbers. Here we are interested in their real locus, and how it intersects the cone  $\text{PSD}_n$ .

# A Cyclic Example

Let  $n = 5$  and consider the CI model given by

$$\begin{aligned} [1 \perp\!\!\!\perp 2 \mid \{3\}] &= \sigma_{12}\sigma_{33} - \sigma_{13}\sigma_{23} \\ [2 \perp\!\!\!\perp 3 \mid \{4\}] &= \sigma_{23}\sigma_{44} - \sigma_{24}\sigma_{34} \\ [3 \perp\!\!\!\perp 4 \mid \{5\}] &= \sigma_{34}\sigma_{55} - \sigma_{35}\sigma_{45} \\ [4 \perp\!\!\!\perp 5 \mid \{1\}] &= \sigma_{45}\sigma_{11} - \sigma_{14}\sigma_{15} \\ [5 \perp\!\!\!\perp 1 \mid \{2\}] &= \sigma_{15}\sigma_{22} - \sigma_{25}\sigma_{12} \end{aligned}$$

This variety has two irreducible components

- $\langle \sigma_{12}, \sigma_{23}, \sigma_{34}, \sigma_{45}, \sigma_{15} \rangle$
- a **toric variety** of degree 31, which intersects  $\text{PSD}_5$  only in the set of **rank one matrices**.

**Corollary:** For Gaussians, the five given statements imply  $[1 \perp\!\!\!\perp 2], [2 \perp\!\!\!\perp 3], [3 \perp\!\!\!\perp 4], [4 \perp\!\!\!\perp 5], [5 \perp\!\!\!\perp 1]$ .

# The Entropy Map

The *submodular cone*  $\text{SubMod}_n$  is solution set of the following system of linear inequalities in  $\mathbb{R}^{2^n}$ :

$$H_{\{i\} \cup K} + H_{\{j\} \cup K} \leq H_{\{i,j\} \cup K} + H_K$$

Taking the logarithms of all  $2^n$  principal minors of any positive definite matrix defines the *entropy map*

$$H : \text{SDP}_n \rightarrow \text{SubMod}_n$$

**Proposition.** The Gaussian CI models are the inverse images of the *faces* of the polyhedral cone  $\text{SubMod}_n$ .

**Problem:** Characterize the image of  $H$ .

# Conclusion

This talk presented **three concrete open problems**:

1. Consider the variety of  $4 \times 4 \times 4$ -tables of tensor rank at most 4. Do the known polynomial invariants of degree five and nine suffice to cut out this variety?
2. Characterize all projective varieties whose maximum likelihood degree is equal to one.
3. Which sets of almost-principal minors can be zero for a positive definite symmetric  $5 \times 5$ -matrix?

# Bonus Problem: Rational Points

Consider  $n$  discrete random variables  $X_1, X_2, \dots, X_n$  with  $d_1, d_2, \dots, d_n$  states.

Any collection of CI statements defines a determinantal variety in the space of tables,

$$\mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2} \otimes \dots \otimes \mathbb{C}^{d_n}.$$

The corresponding *strict CI variety* is the set of tables for which the given CI statements hold but all other CI statements do not hold.

**Problem:** Does every strict CI variety have a  $\mathbb{Q}$ -rational point? What if  $d_1, d_2, \dots, d_n \gg 0$ ?

[Matúš: “Final Conclusions”, CPC 8 (1999) p. 275]