Tobias Boege

Reasoning in Statistics through Algebra

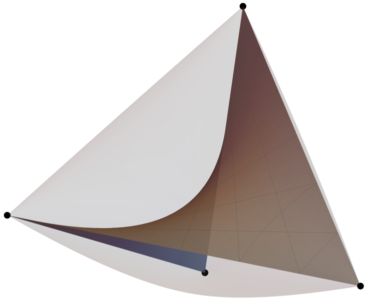
Algebraic statistics tandem, 05 October 2021, Potsdam.

Institut für Algebra und Geometrie
Otto-von-Guericke-Universität Magdeburg

DFG-Graduiertenkolleg
MATHEMATISCHE
KOMPLEXITÄTSREDUKTION

The mantra of algebraic statistics

Statistical models are semialgebraic sets*



The set of all centered, standardized Gaussian distributions parametrized by their correlation matrices

$$\Sigma = \begin{pmatrix} 1 & a & b \\ a & 1 & c \\ b & c & 1 \end{pmatrix} \in \mathrm{PD}_3$$

which satisfy the conditional independence $\xi_1 \perp\!\!\!\perp \xi_2 \mid \xi_3$, or in algebraic terms: $a = bc$.

*sometimes

Conditional independence

Conditional independence $\xi_i \perp\!\!\!\perp \xi_j \mid \xi_K$ is a notion from statistics which asserts an information-theoretical relation: if the outcome of the random variable ξ_k for all components $k \in K$ is known, then the outcome of ξ_i is independent of that of ξ_j:

$$p(\xi_i = x, \xi_j = y \mid \xi_K = z) = a(x,z) \cdot b(y,z).$$

In other words: the distribution of ξ_{ijK} factors into its marginals ξ_{iK} and ξ_{jK}.
→ complexity reduction

Conditional independence

Conditional independence $\xi_i \perp\!\!\!\perp \xi_j \mid \xi_K$ is a notion from statistics which asserts an information-theoretical relation: if the outcome of the random variable ξ_k for all components $k \in K$ is known, then the outcome of ξ_i is independent of that of ξ_j:

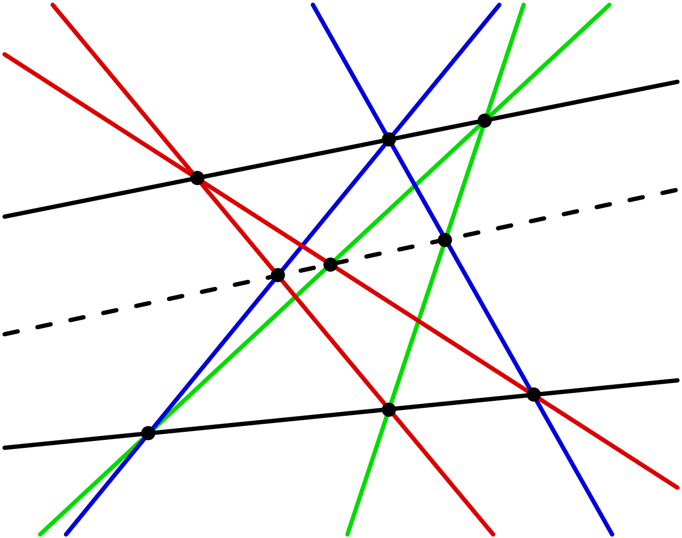
$$p(\xi_i = x, \xi_j = y \mid \xi_K = z) = a(x,z) \cdot b(y,z).$$

In other words: the distribution of ξ_{ijK} factors into its marginals ξ_{iK} and ξ_{jK}.
$\rightarrow$ complexity reduction

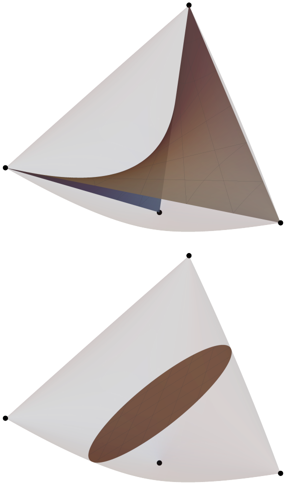
For Gaussian distributions, every conditional independence statement $\xi_i \perp\!\!\!\perp \xi_j \mid \xi_K$ corresponds to a *polynomial equation* $f_{ij|K} = 0$ on the covariance matrix, e.g.,

$$\xi_1 \perp\!\!\!\perp \xi_2 \mid \xi_3 \quad \Leftrightarrow \quad \sigma_{11} \cdot \sigma_{12} = \sigma_{13} \cdot \sigma_{23}.$$

For geometers: conditional independence ≈ collinearity

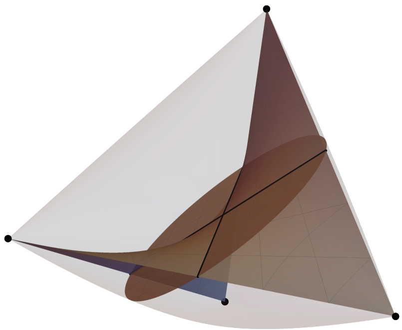


Reasoning is geometry (which is algebra)



Reasoning: if a Gaussian distribution satisfies $\xi_1 \perp\!\!\!\perp \xi_2$ and $\xi_1 \perp\!\!\!\perp \xi_2 \mid \xi_3$, then will it also satisfy $\xi_2 \perp\!\!\!\perp \xi_3$?

Reasoning is geometry (which is algebra)



Reasoning: if a Gaussian distribution satisfies $\xi_1 \perp\!\!\!\perp \xi_2$ and $\xi_1 \perp\!\!\!\perp \xi_2 \mid \xi_3$, then will it also satisfy $\xi_2 \perp\!\!\!\perp \xi_3$?

On the space of positive-definite 3×3-matrices defined by the equations

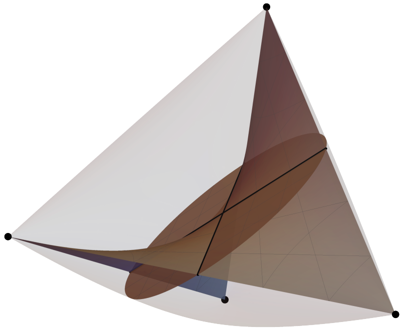
$$f_{12|\varnothing} = \sigma_{12} = 0,$$

$$f_{12|3} = \sigma_{11} \cdot \sigma_{12} - \sigma_{13} \cdot \sigma_{23} = 0$$

does the polynomial $f_{23|\varnothing} = \sigma_{23}$ vanish as well?

No (see image).

Reasoning is geometry (which is algebra)



Reasoning: if a Gaussian distribution satisfies $\xi_1 \perp\!\!\!\perp \xi_2$ and $\xi_1 \perp\!\!\!\perp \xi_2 \mid \xi_3$, then will it also satisfy $\xi_2 \perp\!\!\!\perp \xi_3$?

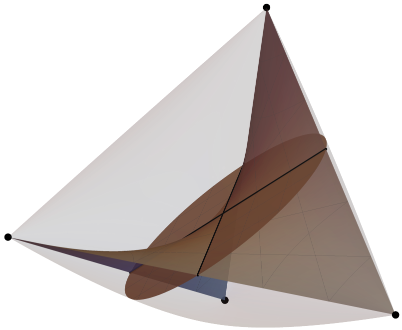
But we have

$$\sigma_{11} \cdot f_{12|\varnothing} - f_{12|3} = f_{13|\varnothing} \cdot f_{23|\varnothing}.$$

Hence algebra proves this inference rule:

$$(\xi_1 \perp\!\!\!\perp \xi_2) \wedge (\xi_1 \perp\!\!\!\perp \xi_2 \mid \xi_3) \Rightarrow (\xi_1 \perp\!\!\!\perp \xi_3) \vee (\xi_2 \perp\!\!\!\perp \xi_3).$$

Reasoning is geometry (which is algebra)



Reasoning: if a Gaussian distribution satisfies $\xi_1 \perp\!\!\!\perp \xi_2$ and $\xi_1 \perp\!\!\!\perp \xi_2 \mid \xi_3$, then will it also satisfy $\xi_2 \perp\!\!\!\perp \xi_3$?

But we have

$$\sigma_{11} \cdot f_{12|\varnothing} - f_{12|3} = f_{13|\varnothing} \cdot f_{23|\varnothing}.$$

Hence algebra proves this inference rule:

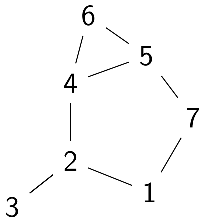
$$(\xi_1 \perp\!\!\!\perp \xi_2) \wedge (\xi_1 \perp\!\!\!\perp \xi_2 \mid \xi_3) \Rightarrow (\xi_1 \perp\!\!\!\perp \xi_3) \vee (\xi_2 \perp\!\!\!\perp \xi_3).$$

Theorem (Positivstellensatz)

Every true inference rule for Gaussians has a "proof polynomial" over $\mathbb{Z}$.

Graphical models

Graphical models are a popular tool to represent dependences among random variables. Vertices are random variables, edges and paths are dependencies (think: information is exchanged along edges).



$1 \not\perp\!\!\!\perp 2$ because there is an edge between them.

$1 \not\perp\!\!\!\perp 6$ because there is a path.

$1 \perp\!\!\!\perp 6 \mid 4, 5$ because all paths $1 \to 6$ hit 4 or 5.

$1 \perp\!\!\!\perp 6 \mid 2, 7$ for the same reason.

$1 \not\perp\!\!\!\perp 6 \mid 2, 4$ because $1 \to 7 \to 5 \to 6$ avoids 2 and 4.

The conditional independences modeled by a graph is given by all its *vertex cuts*.

Gaussian graphical models

Theorem

Let $G = (V, E)$ be an undirected graph and K a generic positive-definite adjacency matrix:

$$k_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } ij \notin E, \\ \varepsilon_{ij}, & \text{otherwise.} \end{cases}$$

Then $\Sigma = K^{-1}$ satisfies exactly the same conditional independence statements as G.

The linear concentration model specified by G consists of all matrices K above.
It is a spectrahedron. Its inverse is called the CI model $\mathcal{M}(G)$ of G.

Convexity

Theorem (Matúš 2012)

A Gaussian CI model $\mathcal{M}$ (given by any set of conditional independences $\xi_i \perp\!\!\!\perp \xi_j \mid \xi_K$) is convex if and only if $\mathcal{M} = \mathcal{M}(G)^{-1}$ for some graph G.

Thus optimizing over a linear concentration model is an instance of *semidefinite programming*:

$$\begin{aligned} \min \quad & f(\Sigma) \\ \text{s.t.} \quad & \Sigma_{ij} = 0 \text{ for } ij \notin E, \\ & \Sigma \succ 0. \end{aligned}$$

Linear concentration models are the only CI models which allow this formulation.

The following talks

Xiangying Chen: Maximum likelihood degree.

Andreas Kretschmer: Double Markovian models.

Philip Dörr: Coxeter group statistics.