Algebraic Statistics

Meeting of the Scientific Advisory Board 2024

Marina Garrote-López & Maximilian Wiesmann
marina.garrote@mis.mpg.de maximilian.wiesmann@mis.mpg.de
Nonlinear Algebra Research Group

MAX PLANCK INSTITUTE
FOR MATHEMATICS IN THE SCIENCES

What is Algebraic Statistics?

Many statistical models can be represented as *semi-algebraic sets*. This allows studying statistical properties by means of algebraic geometry, commutative algebra and combinatorics. Algebraic Statistics has been successfully applied to problems such as **identifiability**, **model selection**, **maximum likelihood estimation** and **sampling**. Typical models studied in Algebraic Statistics include discrete and Gaussian exponential families, (hidden variable) graphical models and mixture models.

Structure	$\rightarrow$	**Model**	$\rightarrow$	**Semi-algebraic set**
e.g. causal relationships		collection of probability distr.		e.g. algebraic variety intersected with probability simplex

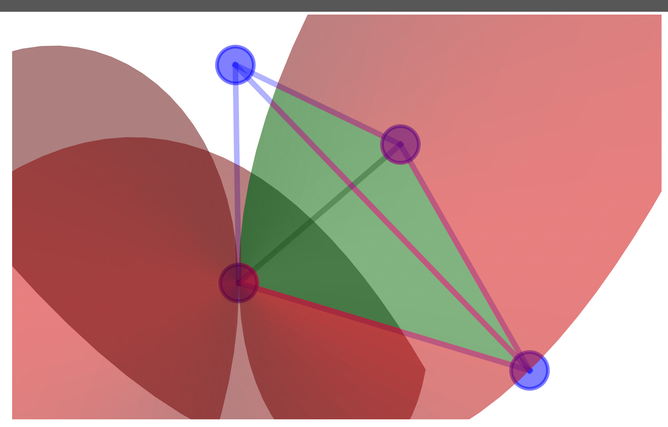


Figure 1. A model representing three (biased) coin tosses is pictured as the intersection (green) of an algebraic variety (red) with the probability simplex (blue).

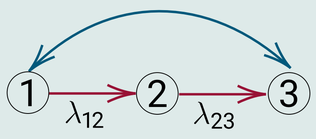
$$\rightarrow p = \Phi(x, y) = (x^3, 3x^2y, 3xy^2, y^3) \rightarrow \overline{Im(\Phi)} \cap \Delta_4$$

Graphical Models and Identifiability

Dependencies among random variables can be encoded in a graph.

Gaussian Linear Structural Equation models: statistical models that associate a family of normal distributions to the graph $G = (V, B, D)$ assuming

$$X_j = \sum_{i \to j \in D} \lambda_{ij} X_i + \epsilon_i, \text{ with } \epsilon \sim \mathcal{N}(0, \Omega).$$



$$\Lambda = \begin{pmatrix} 0 & \lambda_{12} & 0 \\ 0 & 0 & \lambda_{23} \\ 0 & 0 & 0 \end{pmatrix} \quad \Omega = \begin{pmatrix} \omega_{11} & 0 & \omega_{13} \\ 0 & \omega_{22} & 0 \\ \omega_{31} & 0 & \omega_{33} \end{pmatrix}$$

The Gaussian graphical model $\mathcal{M}_G \subseteq PD$ consists of all covariance matrices $\Sigma = (I - \Lambda)^{-T}\Omega(I - \Lambda)^{-1}$ that arise for some choice of Λ and Ω. When the structure of G can be recovered from Σ, we say that the model is **identifiable**.

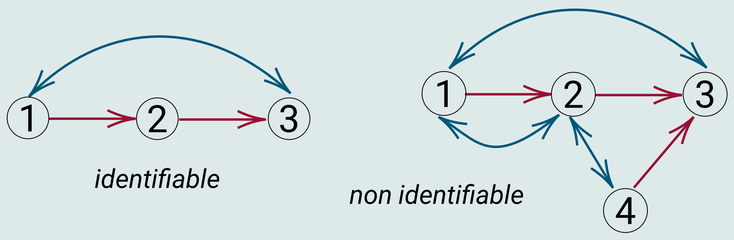


Figure 2. Algebraic Statistics provides graphical criteria for identifiability [5].

Ask us about our research

- Drton, Garrote-López, Robeva. Causal inference for linear non-Gaussian cyclic models (2024+)
- Duarte, Pavlov, Wiesmann. Algebraic Geometry of Quantum Graphical Models (2023)

Likelihood Geometry

Likelihood Geometry is the (algebro-)geometric analysis of the process of maximum likelihood estimation. This has helped understanding, among others,

- existence and uniqueness criteria for the MLE [1];
- the optimization landscape of the likelihood function, e.g. through the notion of *maximum likelihood degree* [3];
- the classification of models where MLE is computationally simple [7];
- model selection techniques for *singular* statistical models [6].

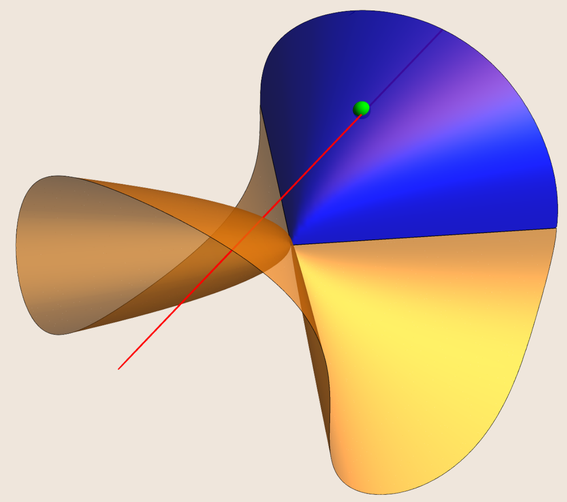


Figure 3. An illustration of *Birch's Theorem:* The MLE for a log-linear model is the unique intersection point (green) of a toric variety with a linear space (red) lying in the non-negative region (blue) of the variety.

Ask us about our research

- Garcia Puente, Garrote-López, Shehu. Computing algebraic degrees of phylogenetic varieties (2024)
- Telen, Wiesmann. Euler Stratifications of Plane Curves (2024+)

Model Selection in Phylogenetics

It is common to assume that *evolution* follows a *Markov process on a tree:*

$$M_i = \begin{pmatrix} P_{A|A} P_{C|A} P_{G|A} P_{T|A} \\ P_{A|C} P_{C|C} P_{G|C} P_{T|C} \\ P_{A|G} P_{C|G} P_{G|G} P_{T|G} \\ P_{A|T} P_{C|T} P_{G|T} P_{T|T} \end{pmatrix}$$

Given a tree T with n leaves, we get a map $\Phi^T : \Theta \rightarrow \mathbb{R}^{4^n}$ with

$$\Phi^T_{s_1,s_2,s_3} = P(X_1 = s_1, X_2 = s_2, X_3 = s_3)$$

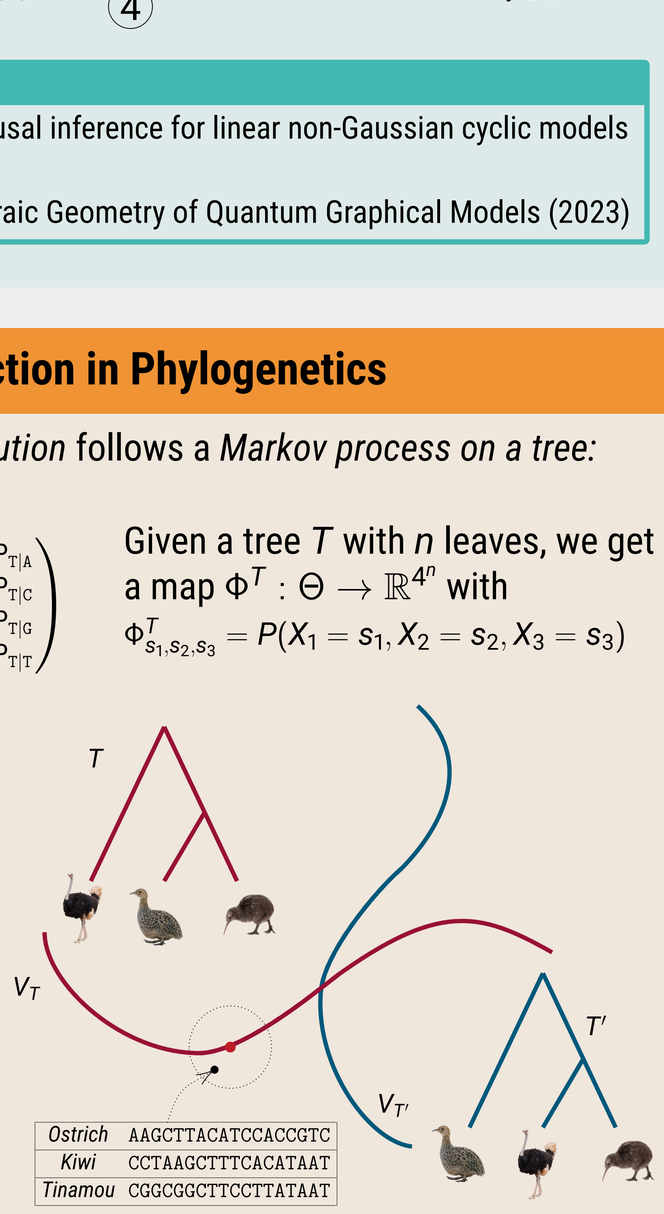


Figure 4. Different trees produce different algebraic varieties and semi-algebraic sets. This semi-algebraic description is used to select the tree that best explains the evolutionary process.

Ask us about our research

- Casanellas, Fernandez-Sanchez, Garrote-López, Sabaté-Vidales. Designing weights for quartet-based methods when data is heterogeneous across lineages (2023)

Sampling

Algebraic Statistics provides the first way to sample *contingency tables* associated to log-linear models with fixed marginals: *Markov bases* yield moves between contingency tables, e.g. to perform *Fisher's exact test* [4].

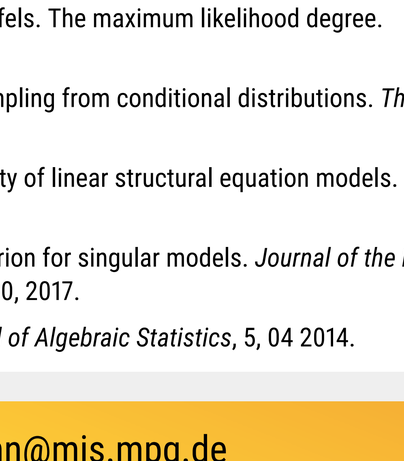


Figure 5. *Chow–Markov Chain Monte Carlo* techniques allow to sample from a variety [2].

Literature

[1] Carlos Améndola, Kathlén Kohn, Philipp Reichenbach, and Anna Seigal. Invariant theory and scaling algorithms for maximum likelihood estimation. *SIAM Journal on Applied Algebra and Geometry*, 5(2):304–337, 2021.

[2] Paul Breiding, Kathlén Kohn, and Bernd Sturmfels. Metric algebraic geometry, 2024.

[3] Fabrizio Catanese, Serkan Hoşten, Amit Khetan, and Bernd Sturmfels. The maximum likelihood degree. *American Journal of Mathematics*, 128(3):671–697, 2006.

[4] Persi Diaconis and Bernd Sturmfels. Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26(1):363–397, 1998.

[5] Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2):865–886, 2011.

[6] Mathias Drton and Martyn Plummer. A bayesian information criterion for singular models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(2):323–380, 2017.

[7] June Huh. Varieties with maximum likelihood degree one. *Journal of Algebraic Statistics*, 5, 04 2014.