Mathematisches Forschungsinstitut Oberwolfach

Report No. 20/2017

DOI: 10.4171/OWR/2017/20

Algebraic Statistics

Organised by
Mathias Drton, Seattle
Thomas Kahle, Magdeburg
Bernd Sturmfels, Berkeley
Caroline Uhler, Cambridge MA

16 April – 22 April 2017

Abstract. Algebraic Statistics is concerned with the interplay of techniques from commutative algebra, combinatorics, (real) algebraic geometry, and related fields with problems arising in statistics and data science. This workshop was the first at Oberwolfach dedicated to this emerging subject area. The participants highlighted recent achievements in this field, explored exciting new applications, and mapped out future directions for research.

Mathematics Subject Classification (2010): Primary: 13P25, 62E99; Secondary: 15A72, 52A40, 92B05.

Introduction by the Organisers

The Oberwolfach workshop *Algebraic statistics* was organized by Mathias Drton, Thomas Kahle, Bernd Sturmfels, and Caroline Uhler and ran April 17-21, 2017. Algebraic statistics is a rather new field, about two decades old. The field emerged from two lines of work: Diaconis and Sturmfels introduced algebraic tools to categorical data analysis and suggested the construction of Markov bases to perform exact goodness-of-fit tests for such data. This got algebraists, combinatorialists, and algebraic geometers interested in problems in statistics. Significant contributions from Japanese statisticians resulted in a book on Markov bases in algebraic statistics. Through recent work Markov bases have also found applications to disclosure limitation and genetics. The second source, which coined the term 'algebraic statistics', is a book explaining how Gröbner basis methods can be used in experimental design. A recent direction that emerged from this is the use of

commutative algebra for experimental design in system reliability. An Oberwolfach seminar, taught in 2008 by Seth Sullivant and two of the organizers, led to a lecture notes volume that is widely used and helped shape the field.

Since its beginnings in the late 1990s, the field of algebraic statistics has grown rapidly. The development of new theory and algorithms for data analysis inspired by algebra, combinatorics and algebraic geometry has brought together previously disconnected communities of algebraists and statisticians. By now, algebraic methods have touched on virtually all major themes in statistics, including parameter identifiability, parameter estimation, hypothesis testing, model selection, and Bayesian inference. Conversely, problems and models from statistics have inspired significant new developments in algebraic combinatorics, high-dimensional commutative algebra, convex geometry, and computational algebraic geometry.

The workshop brought together established and young researchers interested in solving problems from statistics using algebraic approaches. We put a particular emphasis on involving statisticians in the development of the field to effectively communicate the developments so far, have an impact in the statistical community, and ensure that the algebraic statistics community works on important questions in both mathematics and statistics.

Anna Seigal prepared a delightful article for the Snapshots series that describes some basic aspects of algebraic statistics for a general audience. For the development of her article, she took a short survey of the 52 participants of this Oberwolfach workshop. In that survey, 30 per cent of the participants identified themselves as statisticians, the rest as mathematicians. This shows that a healthy balance between the two groups was achieved.

The program consisted of 27 formal talks, splitting into 13 longer and 14 shorter talks. There were two special evening sessions. On Monday, five graduate students introduced themselves and presented some of their work in short talks of twenty minutes. This gave them an opportunity to interact right away with senior participants. The second evening session took place on Wednesday, after an excellent hike through typical April weather. That event, titled *Statistics on interesting spaces*, was hosted by Ruriko Yoshida and Stephan Huckemann. It introduced the participants to potential new application areas for algebraic statistics, for example, models for observations that take their values in non-Euclidean spaces.

The workshop featured many spontaneous interactions and self organized research activities surrounding new connections between participants. For instance, David Gross' talk highlighted the importance of a detailed understanding of the geometry of latent variable models for falsification of theories in physics. The models he discussed were closely related to those treated in Robin Evans' talk. Throughout, different groups of participants discussed these issues in the context of specific surprisingly subtle problems involving just a few binary variables.

The week ended with a wrap-up session on Friday afternoon where groups of participants that had worked on different problems during the week updated everyone on their progress. In addition, the workshop participants proposed and discussed further open problems in the area of algebraic statistics.

Acknowledgement: The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-1049268, "US Junior Oberwolfach Fellows". Moreover, the MFO and the workshop organizers would like to thank the Simons Foundation for supporting Seth Sullivant in the "Simons Visiting Professors" program at the MFO.

Workshop: Algebraic Statistics

Table of Contents

Abstracts

Bounds on the Expected Size of the Maximum Agreement Subtree

Seth Sullivant
(joint work with Daniel I. Bernstein, Lam Si Tung Ho, Colby Long, Mike Steel, Katherine St. John)

A rooted binary tree T with leaf label set X and other vertices unlabeled is called a binary X-tree. Given a subset $S \subseteq X$, the restriction tree $T|_S$ denotes the binary tree obtained by restricting to the leaf set S and contracting all nonroot vertices of degree two. Given two X-trees T_1 and T_2, an *agreement subset* is any set $S \subseteq X$ such that $T_1|_S = T_2|_S$. The resulting tree $T_1|_S = T_2|_S$ is called an *agreement subtree*. A *maximum agreement subtree* is any agreement subtree with the largest number of leaves. Let $\mathrm{MAST}(T_1, T_2)$ be the size of a maximum agreement subtree of T_1 and T_2.

The size of a maximum agreement subtree is a tool that is used to measure dissimilarity between trees, especially when testing for coevolution. In this context the tree T_1 might be a tree of host species, the tree T_2 might be a tree of parasite species. These trees have the same leaf label set because each parasite is paired with its host species. Roughly speaking, these two phylogenetic histories are said of have underdone coevolution if they evolved together in some way, that is the evolutionary history of the hosts affected the evolutionary history of the parasites, or vice versa. Large $\mathrm{MAST}(T_1, T_2)$ indicates that the trees might have undergone coevolution whereas as small $\mathrm{MAST}(T_1, T_2)$ seems to indicate independence of these two processes. To make "large" and "small" in the preceding sentence precise involves understanding the distribution of $\mathrm{MAST}(T_1, T_2)$ for random trees under a suitable distribution. This leads to the following problem:

Problem 1. *What is the distribution of* $\mathrm{MAST}(T_1, T_2)$ *asymptotically as* $|X| \to \infty$ *for commonly used probability distributions for random binary trees such as the uniform distribution or the Yule-Harding distribution?*

Bryant, McKenzie, and Steel [2] performed simulations that suggest the following conjecture:

Conjecture 2. *For either the uniform distribution or the Yule-Harding distribution on binary trees* $\mathbb{E}[\mathrm{MAST}(T_1, T_2)] = \Theta(\sqrt{n})$ *where* $n = |X|$.

Those authors also proved an upper bound that $\mathbb{E}[\mathrm{MAST}(T_1, T_2)] = O(\sqrt{n})$ in the case of the uniform distribution. The main results of our paper [1] are to produce new bounds on the expected value of the maximum agreement subtree. Generalizing the proof of Bryant, McKenzie, and Steel [2], we showed the following general upper bound:

Theorem 3. *For any exchangeable sampling consistent distribution on random trees* $\mathbb{E}[\mathrm{MAST}(T_1, T_2)] = O(\sqrt{n})$.

Note that this Theorem includes the case of the Yule-Harding distribution and the uniform distribution. The difficult part of analyzing the expected size of the maximum agreement subtree seems to be proving lower bounds. We proved the following results, which gave the first nontrivial lower bounds on the expected size.

Theorem 4. *Under the uniform distribution on rooted binary trees,*

$$\mathbb{E}[\mathrm{MAST}(T_1, T_2)] = \Omega(n^{1/8}).$$

Theorem 5. *Let α be the unique positive root of the equation $2^{2-\alpha} = (\alpha + 1)(\alpha + 2)$ ($\alpha \approx .34184$). Under the Yule-Harding distribution on rooted binary trees $\mathbb{E}[\mathrm{MAST}(T_1, T_2)] = \Omega(n^{\alpha})$.*

References

[1] Bernstein, Daniel Irving; Ho, Lam Si Tung; Long, Colby; Steel, Mike; St. John, Katherine; Sullivant, Seth. Bounds on the expected size of the maximum agreement subtree. *SIAM J. Discrete Math.* **29** (2015), no. 4, 2065–2074.

[2] Bryant, David; McKenzie, Andy; Steel, Mike. The size of a maximum agreement subtree for random binary trees. *Bioconsensus (Piscataway, NJ, 2000/2001)*, 55–65, *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*,**61**, *Amer. Math. Soc., Providence, RI,* 2003.

Graphical Models, Model Selection and Tangent Spaces

Robin J. Evans

Model selection is a task of fundamental importance in statistics, and advances in high-dimensional model selection have been one of the major areas of progress over the past 20 years. Examples include covariate selection in linear regression, and models based on patterns of zeros in the inverse covariance matrix. Much of this progress has been due to penalized methods such as the lasso, and efficient methods for solving the relevant convex optimization problems.

However in other classes, such as directed graphical models, correct model selection is provably computationally hard [1]. In this talk we give a geometric explanation for why standard convex penalized methods cannot be adapted to directed graphs, based on the local geometry of the different models at points of intersection. These results also show that it is 'statistically' as well as computationally hard to learn these models, and that much larger samples will typically be needed for moderate effect sizes.

This has implications for other types of graphical model selection, including ancestral graph models [2] and nested Markov models [3], as well as time series models. We provide some relevant heuristics that give insights into the feasibility of model selection in various cases.

References

[1] M. Chickering, Learning Bayesian networks is NP-complete, in *Learning from Data*, Springer New York, 1996.

[2] T.S. Richardson and P. Spirtes, *Ancestral graph Markov models*, Annals of Statistics **30**(4) 962–1030, 2002.

[3] T.S. Richardson, R.J. Evans, J.M. Robins and I. Shpitser, *Nested Markov Properties for Acyclic Directed Mixed Graphs*, arXiv:1701.06686, 2017.

Geometry of Log-Concave Density Estimation

Elina Robeva

(joint work with Bernd Sturmfels, Caroline Uhler)

Shape-constrained density estimation is an important topic in mathematical statistics. Let $X = (x_1, \dots, x_n)$ be a configuration of n distinct labeled points in $\mathbb{R}^d$, and let $w = (w_1, \dots, w_n)$ be a vector of positive weights that satisfy $w_1 + \cdots + w_n = 1$. The pair (X, w) is our dataset. Our aim is to estimate the density function $p : \mathbb{R}^d \to R$ from which these points were sampled. If the weights $w_1, \dots, w_n$ are all equal, then one can think of observing each of the points in X exactly once. If the weights $w_1, \dots, w_n$ vary, then there are several ways of interpreting their statistical meaning. One can think of observing several points around each of x_i depending on the size of w_i, or one can think of having a prior on the data, i.e. knowing that x_i is "good" for estimating the unknown density with probability w_i.

In order the estimate p, we maximize the *log-likelihood of the data*:

$$\sum_{i=1}^{n} w_i p(x_i)$$

under the condition that p is a density. If we do not impose any additional constraints on the function p, then the likelihood function is unbounded, and in particular it is infinite at $p = \sum_{i=1}^{n} w_i \delta_{x_i}$. Therefore, we need to impose additional constraints on the function p.

In this work we focus on densities on $p : \mathbb{R}^d \to \mathbb{R}$ that are *log-concave*, ilel their logarithm is concave. Such densities include Gaussians, uniform distributions, beta distributions, gamma distributions, and others. The optimization problem that we are interested in then becomes

$$\begin{aligned} \text{maximize}_p \quad & \sum_{i=1}^{n} w_i \log(p(x_i)) \\ \text{s.t.} \quad & p \text{ is a density} \\ \text{and} \quad & p \text{ is log-concave.} \end{aligned} \tag{1}$$

Shape-constrained density estimation has been studied in the past. This line of research started with Grenander [11], who analyzed the case when the density is monotonically decreasing. Another popular shape constraint is convexity of the density [12]. Log-concave density estimation has been studied, among other, by Richard Samworth and his collaborators. A solution to the optimization problem (1) was given by Cule, Samworth, and Strewart in [6]. An efficient algorithm

for solving it is implemented in the `R` package `LogConcDEAD` due to Cule, Gramacy and Samworth [5].

It turns out that (1) is equivalent to a finite-dimensional convex optimization problem, which is what allows for a nice algorithm for solving it. More precisely, the optimal density p^* is a *tent function* supported on the data X. Given a vector of (tent pole) heights $y = (y_1, \ldots, y_n) \in \mathbb{R}$, we define the tent function $h_{X,y} : \mathbb{R}^d \to \mathbb{R}$ to be the smallest concave function $h : \mathbb{R}^d \to \mathbb{R}$ such that $h(x_i) \geq y_i$ for all $i = 1, \ldots, n$. Note that h is equal to $-\infty$ at the points outside the convex hull of X.

Now, the optimal solution p^* equals to $\log(h_{X,y^*})$ for some optimal y^*, and instead of solving the infinite-dimensional problem (1) over p, we can solve the *finite-dimensional convex* problem

$$\text{maximize}_{y \in \mathbb{R}^n} \quad \sum_{i=1}^{n} w_i y_i \tag{2}$$

$$\text{s.t.} \quad \int \exp(h_{X,y}(t))dt = 1.$$

Every tent function $h_{X,y}$ induces a *regular subdivision* Δ of the configuration X. A regular (polyhedral) subdivision $\Delta = \{I_1, \ldots, I_k\} \subseteq 2^{[n]}$ is the collection of cells $\text{conv}\{x_i : i \in I\}$ on which the tent function $h_{X,y}$ is affine linear. These cells are always polytopes whose vertices are a subset of $\{x_1, \ldots, x_n\}$, so they can be encoded by subsets $[n]$. A *regular triangulation* is a regular subdivision Δ all of whose cells are simplices.

In this work we study the subdivisions Δ that can be induced by the optimal heights y^* for a starting configuration X, as we vary the weights w. Our optimization problem defines a map from the space of weights to the set of heights where each set of weights w is mapped to the optimal heights that solve the problem (2). We prove that this map is surjective. In fact, every regular subdivision arises in the MLE for some set of weights with positive probability, but coarser subdivisions appear to be more likely to arise than finer ones.

Theorem 1. *For a fixed configuration X and for any vector $y \in \mathbb{R}^n$ such that $\int \exp(h_{X,y}(t))dt = 1$, there exist weights $w \in \mathbb{R}^n$ such that y maximizes (2).*

Theorem 2. *Let Δ be any regular subdivision of the configuration X. There exists a non-empty open subset $\mathcal{U}_\Delta$ in $\mathbb{R}^n$ such that, for every $w \in \mathcal{U}_\Delta$, the optimal solution $\hat{f}$ to (1) is a piecewise log-linear function whose regions of linearity are the cells of Δ.*

To quantify these results, we introduce a continuous version of the secondary polytope, whose dual we name the *Samworth body*.

$$\mathcal{S}(X) \quad = \quad \{\, y \in \mathbb{R}^n \,:\, \int_P \exp(h_{X,y}(t))dt \leq 1 \,\}.$$

The Samworth body $\mathcal{S}(X)$ is a full-dimensional closed convex set in $\mathbb{R}^n$. Note that the boundary $\partial\mathcal{S}(X)$ consists of those vectors y for which the integral equals

exactly 1, i.e. those for which $\exp(h_{X,y})$ is a density. The boundary $\partial\mathcal{S}(X)$ is smooth at those points $y \in \partial\mathcal{S}(X)$ for which $y \in \partial\mathcal{S}(X)$ which induce a regular triangulation. For such points $y \in \partial\mathcal{S}(X)$ there is a unique set of weights $w \in \partial\mathcal{S}(X)^*$ which induce y after running our optimization problem (2). At points $y \in \partial\mathcal{S}(X)$ which induce a regular subdivision which is not a triangulation, the Samworth body is not smooth. For such points there is a whole face F of the boundary of the dual body $\partial\mathcal{S}(X)^*$ such that all weights $w \in F$ induce y after running (2). This gives a strong analogy between the *secondary polytope* $\Sigma(X)$ and the dual of the Samworth body $\mathcal{S}(X)^*$, and between the secondary fan of X and the Samworth body $\mathcal{S}(X)$.

For the case of unit weights $w = \frac{1}{n}(1, \ldots, 1)$ we show that the minimal number of points in X that one needs in order to ensure that the optimal density is not log-linear is $d + 3$.

Theorem 3. *Let X be a configuration of $n = d + 2$ points that affinely span $\mathbb{R}^d$. For $w = \frac{1}{n}(1, \ldots, 1)$, the optimal density $\hat{f}$ is log-linear, so the optimal subdivision of X is trivial.*

Theorem 4. *For any integer $d \geq 2$, there exists a configuration of $n = d+3$ points in $\mathbb{R}^d$ for which the optimal subdivision with respect to unit weights is non-trivial.*

Our work establishes a new link between geometric combinatorics and nonparametric statistics, and it also suggests a number of open problems at the interplay between these fields.

- Design a test-statistic for log-concavity based on the optimal subdivision Δ.
- What is the smallest size $n(c, d)$ such that there exists a configuration X in $\mathbb{R}^d$ the optimal subdivision with unit weights has at least c cells? (Theorems 3 and 4 show that $n(2, d) = d + 3$ for $d \geq 2$.)
- Classify subdivisions that can be realized by points in $\mathbb{R}^d$ with unit weights.
- For a fixed w and a fixed combinatorial type of subdivision Δ, study the semianalytic set of all configurations X such that Δ is the optimal subdivision for the data (X, w).

References

[1] K. Adiprasito, E. Nevo and J. Samper: *A geometric lower bound theorem*, Geom. Funct. Anal. **26** (2016) 359–378.

[2] M.Y. An: *Log-concave probability distributions: theory and statistical testing*, Duke University, Department of Economics Working Paper No. 95-03.

[3] M.Y. An: *Log-concavity versus log-convexity: a complete characterization*, Journal of Economic Theory **80** (1998) 350–369.

[4] A. Barvinok: *Computing the volume, counting integral points, and exponential sums*, Discrete Comput. Geom. **10** (1993) 123–141.

[5] M. Cule, R.B. Gramacy and R. Samworth: LogConcDEAD: an R package for maximum likelihood estimation of a multivariate log-concave density. J. Statist. Software **29** (2009) Issue 2.

[6] M. Cule, R. Samworth and M. Stewart: *Maximum likelihood estimation of a multi-dimensional log-concave density*, J. R. Stat. Soc. Ser. B Stat. Methodol. **72** (2010) 545–607.
[7] J. De Loera, S. Hoşten, F. Santos and B. Sturmfels: *The polytope of all triangulations of a point configuration*, Documenta Mathematica **1** (1996) 103–119.
[8] J. De Loera, J. Rambau and F. Santos: *Triangulations. Structures for Algorithms and Applications*, Algorithms and Computation in Mathematics **25**, Springer-Verlag, Berlin, 2010.
[9] L. Dümbgen and K. Rufibach: *Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency*, Bernoulli **15** (2009) 40–68.
[10] I.M. Gel'fand, M.M. Kapranov and A.V. Zelevinsky: *Discriminants, Resultants and Multi-dimensional Determinants*, Birkhäuser, Boston, 1994.
[11] U. Grenander: *On the theory of mortality measurement II*, Skandinavisk Aktuarietidskrift **39** (1956) 125–153.
[12] P. Groeneboom, G. Jongbloed and J. A. Wellner: *Estimation of a convex function: Characterizations and asymptotic theory*, Annals of Statistics **29** (2001) 1653–1698.
[13] K. Gross and D. Richards: *Total positivity, spherical series, and hypergeometric functions of matrix argument*, Journal of Approximation Theory **59** (1989) 224–246
[14] E. Miller and B. Sturmfels: *Combinatorial Commutative Algebra*, Graduate Texts in Mathematics, Vol. 227, Springer Verlag, New York, 2004.
[15] R. Thomas: *Lectures in Geometric Combinatorics*, Student Mathematical Library **33**, IAS/Park City Mathematical Subseries, American Mathematical Society, Providence, RI, 2006.
[16] G. Walther: *Inference and modeling with log-concave distributions*, Statistical Science **24** (2009) 319–327.

Multivariate Total Positivity and Conditional Independence

Steffen Lauritzen

(joint work with Shaun Fallat, Kayvan Sadeghi, Caroline Uhler, Nanny Wermuth, and Piotr Zwiernik)

The lecture summarizes results in [1] and [2]. We analyze distributions that are multivariate totally positive of order 2 (MTP_2) and discuss various properties of such distributions. MTP_2 distributions appear in the context of positive dependence, ferromagnetism in the Ising model and various latent models.

A multivariate real-valued distribution with density f w.r.t. a product measure μ is *multivariate totally positive of order 2* (MTP_2) if the density satisfies

$$f(x)f(y) \leq f(x \wedge y)f(x \vee y).$$

Regular multivariate Gaussian distributions are MTP_2 if and only if their inverse covariance matrix (concentration matrix) K is an M-matrix, i.e. iff all off-diagonal elements are non-positive.

It is shown that MTP_2 distributions have specific Markov properties. More precisely, *any MTP_2 distribution with strictly positive density is faithful to its pairwise independence graph.* Thus if we define the graph $G(P)$ by

$$u \sim v \iff X_u \perp\!\!\!\perp X_v \,|\, X_{V\setminus\{u,v\}}$$

it holds for arbitrary disjoint subsets $A, B, C \subseteq V$ of the (finite) set of variables V that

$$X_A \perp\!\!\!\perp X_B \,|\, X_C \iff A \perp_{G(P)} B \,|\, C$$

where $A \perp_{G(P)} B \,|\, C$ means that A is separated from B in $G(P)$.

Further, we show that the maximum likelihood problem in the case of a multivariate Gaussian distribution is a convex optimization problem having a unique solution whenever the number of observations is at least two: the solution $\hat{K} = \hat{\Sigma}^{-1}$ is then determined by the equation system

$$\hat{k}_{uv} \leq 0 \text{ for all } u \neq v, \tag{1}$$

$$\hat{\sigma}_{vv} - s_{vv} = 0 \text{ for all } v \in V, \tag{2}$$

$$(\hat{\sigma}_{uv} - s_{uv}) \geq 0 \text{ for all } u \neq v, \tag{3}$$

$$(\hat{\sigma}_{uv} - s_{uv})\hat{k}_{uv} = 0 \text{ for all } u \neq v, , \tag{4}$$

where $S = \{s_{uv}\}$ is the sample covariance matrix.

The condition (4) ensures that the MLE $\hat{K}$ is automatically sparse. If we let $\hat{G}$ denote the graph induced by non-zero entries of $\hat{K}$, we also show that the *maximum weight spanning forest MWSF(R) of the correlation matrix is a subgraph of* $G(\hat{K})$.

References

[1] Fallat, S., Lauritzen, S., Sadeghi, K., Uhler, C., Wermuth, N., and Zwiernik, P. (2016). Total positivity in Markov structures. arXiv:1702.04031. To appear in the*Annals of Statistics*.

[2] Lauritzen, S., Uhler, C., and Zwiernik, P. (2017). Maximum likelihood estimation in Gaussian models under total positivity. arXiv:1702.04031.

Positive Polynomials and Matrices

Daniel Plaumann

This was a survey talk on some recent results in real algebraic geometry deemed interesting in the context of algebraic statistics, with an emphasis on rank constraints for positive semidefinite matrices in linear spaces.

A *spectrahedron* is the intersection $S = L \cap \mathrm{Sym}_n^+(\mathbb{R})$ of the cone $\mathrm{Sym}_n^+(\mathbb{R})$ of positive semidefinite matrices with an affine linear subspace L in the real vector space $\mathrm{Sym}_n(\mathbb{R})$ of real symmetric matrices. Spectrahedra arise in the study of positive polynomials as the parameter spaces of sum-of-squares representations of real polynomials in several variables, the so-called *Gram-spectrahedra* [4]. In algebraic statistics, they arise naturally in the context of Gaussian graphical models [10]. While the facial structure of the positive semidefinite cone is simple, with its extreme rays being precisely the matrices of rank 1, it is often much harder to understand the structure of the boundary of a spectrahedron. We assume in the following that S contains a positive definite matrix.

(1) If L has dimension k and R is an extreme point of S of rank r, then $\binom{r+1}{2} + k \leq \binom{n+1}{2}$ (an upper bound on r).

(2) If L is *generic*, then $k \geq \binom{n-r+1}{2}$ (a lower bound on r). This follows from (1) through an application of convex duality (see [6]).
(3) If S is the *Gram spectrahedron* of a real polynomial, the rank of a matrix in S is the length of the corresponding sum-of-squares representation. This is an interesting class of non-generic spectrahedra for which the possible ranks on the boundary have been studied extensively, with considerable progress in recent years (see [1], [4] and references given there).
(4) Recently, Blekherman and Sinn characterized all spectrahedral cones for which every extreme ray has rank 1 (see [3]). These are the dual cones to sums of squares on particular real algebraic varieties.

In many applications, one is interested in *projected spectrahedra*, which are the images of spectrahedra under linear projections. This is a far more general class of domains. Scheiderer proved that *every* convex semialgebraic subset of $\mathbb{R}^2$ (more generally, the convex hull of a curve in any ambient space) is a projected spectrahedron [7], while this fails in higher dimensions [8]; specifically, the cone of non-negative polynomials fails to be a projected spectrahedron, except in the well-known cases in which every non-negative polynomial is a sum of squares.

Regarding rank constraints, the possible ranks of matrices representing the boundary points of a *generic* projected spectrahedron have been studied by Sinn and Sturmfels [9], generalizing the bounds for generic spectrahedra given above. In the context of Gaussian graphical models, one is interested in the projection $\pi\colon \mathrm{Sym}_n(\mathbb{R}) \to \mathbb{R}^m$ onto a coordinate subspace specified by the edges of a graph Γ. The general *positive semidefinite completion problem* asks for a simple description of the image $\pi(S)$. This has recently been reexamined in the context of sums-of-squares representations on real algebraic varieties by Blekherman, Sinn and Velasco [2]. The *maximum likelihood treshold* of the graph Γ is the minimal rank r such that for a generic matrix $A \in S$ of rank r, there exists a positive definite matrix B (hence of full rank n) with $\pi(A) = \pi(B)$ (see [10] and [5] and references given there). Sophisticated combinatorial (resp. complex-geometric) arguments in [5] imply for example that the maximum likelihood threshold of any planar graph is at most 4. On the other hand, recent work by Blekherman and Sinn [3] shows that if reality and positivity are fully taken into account, such combinatorial bounds cannot always be sharp.

References

[1] G. Blekherman, D. Plaumann, R. Sinn, and C. Vinzant, *Low-Rank Sum-of-Squares Representations on Varieties of Minimal Degree*, preprint, arXiv:1606.04387.
[2] G. Blekherman, R. Sinn, and M. Velasco, *Do Sums of Squares Dream of Free Resolutions?*, SIAM Journal on Applied Algebra and Geometry, **1**, No. 1, (2017), 175-199.
[3] G. Blekherman and R. Sinn, *Generic Completion Rank and Maximum Likelihood Threshold of Graphs*, preprint, arXiv:1703.07849.
[4] L. Chua, D. Plaumann, R. Sinn, and C. Vinzant, *Gram Spectrahedra*, preprint, arXiv:1608.00234.
[5] E. Gross and S. Sullivant, *The Maximum Likelihood Threshold of a Graph*, to appear in: Bernoulli (2014).

[6] P. Rostalski and B. Sturmfels, *Dualities in Convex Algebraic Geometry*, Rendiconti di Mathematica, Serie VII 30 (2010), 285–327
[7] C. Scheiderer, *Semidefinite representation for convex hulls of real algebraic curves*, preprint, arXiv:1208.3865.
[8] C. Scheiderer, *Semidefinitely representable convex sets*, preprint, arXiv:1612.07048.
[9] B. Sturmfels and R. Sinn, *Generic Spectrahedral Shadows*, SIAM Journal on Optimization, **25**, No. 2 (2015), 1209–1220.
[10] B. Sturmfels and C. Uhler, *Multivariate Gaussians, semidefinite matrix completion, and convex algebraic geometry*, Annals of the Institute of Statistical Mathematics **62** (2010), 603–638.

Moment Varieties of Gaussian Mixtures

Carlos Améndola
(joint work with Jean-Charles Faugère, Kristian Ranestad, Bernd Sturmfels)

In Algebraic Statistics, studying the geometry of maximum likelihood estimation for many commonly used models (such as the discrete exponential family) has been quite successful, with the aid of computational invariants such as the ML degree [1]. For the ubiquitous Gaussian mixture model, this approach does not quite work. The ML estimates are transcendental functions of the data and there is no analogous ML degree bound on the number of critical points of the log-likelihood functions [2].

However, the method of moments provides a suitable algebraic approach. All the moments of a Gaussian distribution are homogeneous polynomials in the mean and covariance parameters. This fact allows us to define algebraic moment varieties $G_{n,d}$ given by vectors of all moments in dimension n up to some order d. Furthermore, moment varieties for a mixture of k Gaussians correspond geometrically to secant varieties $Sec_k(G_{n,d})$.

Problem 5. *Study $Sec_k(G_{n,d})$ for all $k, n, d \geq 1$. dimension? degree? equations?*

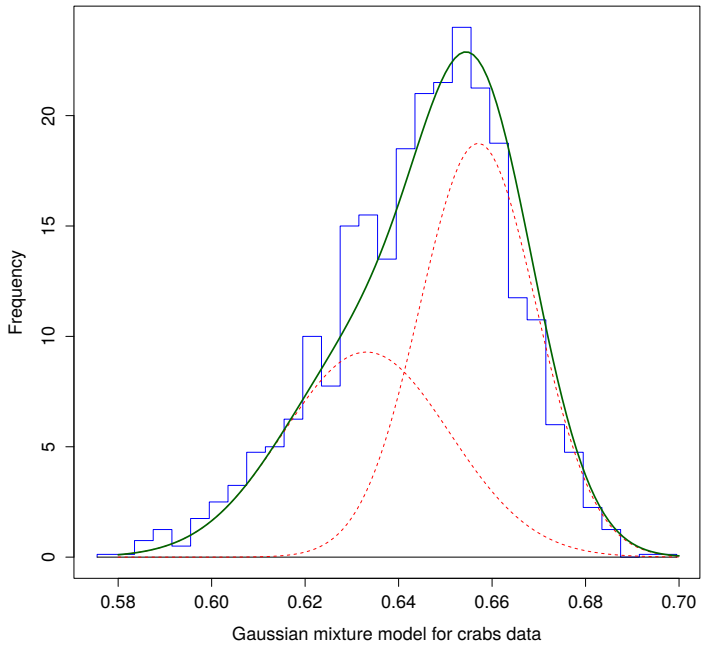
Some progress has been made in [3] and [4]. For example:

Theorem 6. *Let $d \geq 3$. The homogeneous prime ideal of the Gaussian moment surface $\mathcal{G}_{1,d}$ is minimally generated by $\binom{d}{3}$ cubics. These are the 3×3-minors of the $3 \times d$-matrix*

$$H_d = \begin{pmatrix} 0 & m_0 & 2m_1 & 3m_2 & 4m_3 & \cdots & (d-1)m_{d-2} \\ m_0 & m_1 & m_2 & m_3 & m_4 & \cdots & m_{d-1} \\ m_1 & m_2 & m_3 & m_4 & m_5 & \cdots & m_d \end{pmatrix}.$$

The 3×3-minors of the matrix H_d form a Gröbner basis for the prime ideal of the Gaussian moment surface $\mathcal{G}_{1,d} \subset \mathbb{P}^d$ with respect to the reverse lexicographic term order. Thus $\mathcal{G}_{1,d}$ has degree $\binom{d}{2}$ in $\mathbb{P}^d$.

Theorem 7. *The defining homogeneous polynomial of $\mathrm{Sec}_2(\mathcal{G}_{1,6})$ in $\mathbb{P}^6$ is a sum of 31154 monomials of degree 39. This polynomial has degrees $25, 33, 32, 23, 17, 12, 9$ in $m_0, m_1, m_2, m_3, m_4, m_5, m_6$ respectively.*



The statistical motivation is to recover the mixture parameters from observed sample moments up to certain order. This approach was first taken in 1894 by Karl Pearson [5] while trying to fit a mixture of two 1-dimensional Gaussians to measurements of crabs from the Bay of Naples. He showed how to find the unknown means μ_1, μ_2, unknown variances σ_1, σ_2 and mixture proportion α from the first five moments $(n = 1, k = 2, d = 5)$ by solving a polynomial equation of degree 9.

In this way, Pearson's study can be seen as the first paper in Algebraic Statistics!

References

[1] F. Catanese, S. Hoşten, A. Khetan and B. Sturmfels (2006). *The Maximum Likelihood Degree*. American Journal of Mathematics, 128(3), pp. 671-697.

[2] C. Améndola, M. Drton and B. Sturmfels (2016), *Maximum Likelihood Estimates for Gaussian Mixtures are Transcendental*, Mathematical Aspects of Computer and Information Sciences 2015, Berlin, pp. 579-590.

[3] C. Améndola, J.C. Faugère and B. Sturmfels (2016), *Moment Varieties of Gaussian Mixtures*, Journal of Algebraic Statistics 7, pp. 14-28.

[4] C. Améndola, K. Ranestad and B. Sturmfels (2016), *Algebraic Identifiability of Gaussian Mixtures*, arXiv:1612.01129 to appear in International Mathematics Research Notices.

[5] K. Pearson (1894), *Contributions to the Mathematical Theory of Evolution*, Philosophical Transactions of the Royal Society of London, 71-110.

Chordal Networks of Polynomial Ideals

Diego Cifuentes
(joint work with Pablo Parrilo)

Polynomial systems can be used to model many different applications. In most cases the systems arising have a particular sparsity structure, and exploiting such structure can yield significant computational gains. When all polynomials have degree one, we have the special case of systems of linear equations, where it is well known that *chordality* allows for efficient computation [4]. Chordal graphs are also a keystone in constraint satisfaction, graphical models and optimization [3, 5]. We began the study of exploiting chordal structure in polynomial ideals in [1, 2]. In this talk we summarize the results from [2].

Our main contribution is the introduction of a new data structure to represent structured polynomial ideals, that we call *chordal networks*. Chordal networks attempt to fix an intrinsic issue of Gröbner bases: they destroy the underlying graphical structure of the system [1, Ex 1.2]. As a consequence, polynomial systems with simple structure (e.g., the chromatic ideal of a cycle [2, Ex 1.1]) may have overly complicated Gröbner bases. In contrast, chordal networks will always preserve the underlying chordal graph.

Chordal networks describe a decomposition of the (potentially complicated) polynomial ideal into simpler (triangular) polynomial sets. This decomposition gives quite a rich description of the underlying variety. In particular, chordal networks can be efficiently used to compute dimension, cardinality, equidimensional components and also to test radical ideal membership. Remarkably, several families of polynomial ideals (with exponentially large Gröbner bases) admit a compact chordal network representation, of size proportional to the number of variables.

Chordal structure arises in many different problems and we believe that algebraic geometry algorithms should take advantage of it. Preliminary implementation of our methods showed *orders of magnitude reduction* against state-of-the-art algorithms. We showed that chordality helps solve polynomial systems coming from graph colorings, cryptography, sensor networks and differential equations [1]. We also applied our methods to compute irreducible decompositions and radical ideal membership in cases from algebraic statistics [2].

References

[1] D. Cifuentes, P. A. Parrilo, *Exploiting chordal structure in polynomials ideals: A Gröbner bases approach*, SIAM Journal on Discrete Mathematics **30**, no. 3 (2016), 1534–1570.

[2] D. Cifuentes, P. A. Parrilo, *Chordal networks of polynomial ideals*, SIAM Journal on Applied Algebra and Geometry **1**, no. 1 (2017), 73–110.

[3] S. L. Lauritzen, D. J. Spiegelhalter, *Local computations with probabilities on graphical structures and their application to expert systems*, Journal of the Royal Statistical Society, Series B (1988), 157–224.

[4] D. J. Rose, R. E. Tarjan, G. S. Lueker. *Algorithmic aspects of vertex elimination on graphs*, SIAM Journal on Computing **5**, no. 2 (1976), 266–283.

[5] L. Vandenberghe, M. S. Andersen, *Chordal graphs and semidefinite optimization*, Foundations and Trends in Optimization **1**, no. 4 (2015), 241–433.

Higher Order Singular Values: Gram Determinants of Real Binary Tensors

Anna Seigal

The Gram determinants are a tuple of quadratic invariants of a tensor. We introduce the *Gram locus*, the possible Gram determinants of real binary tensors of fixed size. It is the "set of feasible higher-order singular values", from [3], under change of coordinates. We propose a semi-algebraic characterization of the Gram Locus. This answers a question raised by Hackbusch and Uschmajew concerning the higher-order singular values of tensors.

A binary tensor consists of 2^n entries arranged in hypercube format $2\times 2\times\cdots\times 2$. There are n ways to flatten such a tensor into a matrix of size $2 \times 2^{n-1}$. For each flattening, M, the Gram determinant is $\det(MM^T)$. We map a real tensor to its tuple of Gram determinants:

$$\mathcal{G} : \mathbb{R}^2 \otimes \cdots \otimes \mathbb{R}^2 \to \mathbb{R}^n$$

$$(a_{ij\ldots k}) \mapsto (d_1, \ldots, d_n).$$

The *Gram locus* is the image $\mathcal{G}(\mathcal{B})$, where $\mathcal{B}$ is the Frobenius unit ball of tensors.

Theorem 8. *The convex hull of the Gram locus $\mathcal{G}(\mathcal{B})$ is described by the following linear inequalities in the determinants d_i:*

$$d_i \leq \sum_{j\neq i} d_j, \qquad 0 \leq d_i \leq \frac{1}{4}, \qquad 1 \leq i \leq n.$$

In words, each Gram determinant is bounded by the sum of the others.

The true Gram locus is a non-convex semi-algebraic set. We give its description for $2 \times 2 \times 2$ tensors, depicted in Figure 1.

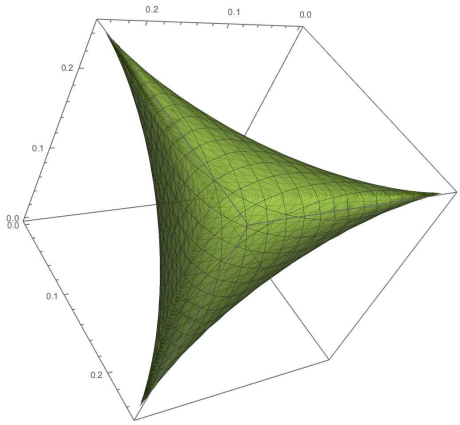


Figure 1. The boundary of the Gram locus for $2 \times 2 \times 2$ tensors.

We propose the general form for the Gram locus, see [5, Conjecture 1.5]. It is concisely expressed as the non-negativity of a single polynomial.

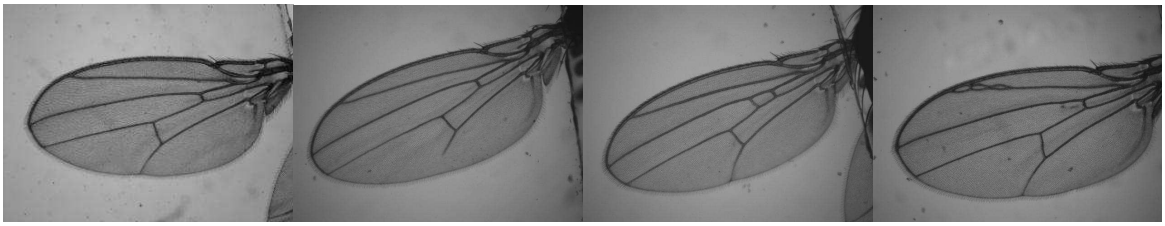
References

[1] L. De Lathauwer, B. De Moor, J. Vandewalle: *A Multilinear Singular Value Decomposition*, SIAM J. Matrix Anal. Appl. (2000) Vol 21 no. 4 1253-1278.
[2] W. Hackbusch, D. Kressner, A. Uschmajew: *Perturbations of Higher-Order Singular Values*, Preprint (2016).
[3] W. Hackbusch, A. Uschmajew: *On the Interconnection between the Higher-Order Singular Values of real tensors*, A. Numer. Math. (2016).
[4] K. Kubjas, P.A. Parrilo, B. Sturmfels: *How to Flatten a Soccer Ball*, Preprint arXiv:1606.02253 (2016).
[5] A. Seigal: *Gram Determinants of Real Binary Tensors*, Preprint arXiv:1612.04420 (2016).

Non-noetherian Modules from Persistent Homology

Ashleigh Thomas
(joint work with Justin Curry, Ezra Miller)

This work recasts and greatly extends the algebraic foundations of persistent homology for the purpose of studying an important question concerning the evolution of changes to discrete morphological features. The model organism for this study is the fruit fly *Drosophila melanogaster*, specifically the wing veination pattern (the first image is normal; the others are topologically abnormal):



The project is joint with biologist David Houle (Florida State). The question is how topologically new features arise in a population with high enough frequency for selection to act, given that the normal topological type is highly canalized—that is, the probability of topological variation away from normal is small. The hypothesis is that directional selection pushes continuous variation in the wing developmental program beyond a threshold, thereby resulting in novel wing vein topologies.

Our mathematical analysis summarizes the metrically embedded planar graphs (wing veins) using multiparameter persistent homology to allow graphs with different topologies to be compared in a single statistical analysis. This data structure is natural in this setting for a number of reasons, including biological interpretability, but it outputs infinitely generated $\mathbb{R}^2$-graded modules over the (non-noetherian) ring of polynomials in two variables with real exponents. Although the generating sets for our modules are uncountable, they occur along semialgebraic varieties and are *finitely encoded*. Consequently, we produce finite data structures for the modules and describe how to reduce algebraic questions about these multiparameter persistence modules to questions about finitely generated $\mathbb{Z}^n$-graded modules over ordinary polynomial rings, which is standard combinatorial commutative algebra.

Statistics on collections of multiparameter persistence modules a priori requires sampling from moduli spaces of modules, which can be nontrivial [1]. But in this work in progress, we aim to prove that for the fly wing application, as with any other application where persistent homology is constructed geometrically from semialgebraic data in $\mathbb{R}^n$, the isomorphism class of each persistence module is determined by its *rank function*, which records the ranks of the homomorphisms between the multigraded pieces of the module.

References

[1] G. Carlsson, A. Zomorodian, *The Theory of Multidimensional Persistence*, Discrete Comput Geom (2009) 42: 71. doi:10.1007/s00454-009-9176-0

Generic Parameter Identifiability in Linear Structural Equation Models with Latent Factors

Luca Weihs

(joint work with Mathias Drton, Rina Foygel Barber)

Linear structural equation models (L-SEMs) are a popular modeling strategy for representing linear causal relationships between random variables. In particular, the joint distribution of a random vector $X = (X_1, ..., X_n)^T$ is distributed as an L-SEM if it can be expressed in matrix form as

$$X = \lambda_0 + \Lambda^T X + \epsilon$$

where $\Lambda = (\lambda_{vw}) \in \mathbb{R}^{n \times n}$ and $\lambda_0 \in \mathbb{R}^n$ are unknown parameters, and $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$ is distributed as a multivariate normal distribution with mean zero and covariance matrix Ω. L-SEMs can be naturally identified with directed graphs with observed and latent (hidden) nodes. Here the observed vertices correspond the components of the random vector X and the lack of a directed edge from X_i to X_j implies that $\lambda_{ij} = 0$. Latent vertices in the graph, used to model confounding, then determine the structure of Ω. A key question of interest for these models is that of generic parameter identifiability; that is, whether for a generic choice of (Λ, Ω) respecting the graph of an L-SEM, one can recover entries of (Λ, Ω) from the covariance matrix of X,

$$\mathrm{Cov}(X) = \phi(\Lambda, \Omega) := (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}.$$

Prior work on this topic has focused primarily upon the case where the latent vertices are required to have exactly two children, in this case one summarizes the effects of these latent vertices using bidirected edges. In particular, the half-trek criterion of Foygel et al. [1] presents a necessary graphical criterion for determining generic identifiability which can be checked in polynomial time. We extend the half-trek criterion to the setting in which latent variables are allowed to have arbitrarily many children, this allows for the generic identifiability of many parameters which would otherwise be unidentifiable in the simpler setting. Surprisingly, we show that verifying our new latent-factor half-trek criterion is NP-complete

and thus we cannot expect to discover a polynomial time algorithm to check its conditions.

References

[1] R. Foygel, J. Draisma, and M. Drton, *Half-trek criterion for generic identifiability of linear structural equation models*, Ann. Statist. **40**(3):1682-1713 (2012).

Tell Me: How Many Modes does the Gaußian Mixture Have ...

Christian Haase
(joint work with Carlos Améndola, Alexander Engström)

It is not known whether or not the probability density function of a mixture of k Gaußians in dimension d can have infinitely many local maxima/modes.

This is a scandal!

It has been conjectured [6] that the maximal number of modes equals $\binom{d+k-1}{d}$. We construct examples of mixtures with $\binom{k}{d}+k$ modes, thus confirming the lower bound in $d=2$. We also prove the first upper bound of $2^{d+\binom{k}{2}}(5+3d)^k$, ...

... provided the number is finite.

There has been a lively discussion after the talk and during the following days. Here are some questions and interesting directions to pursue further, suggested by workshop participants.

- Given some natural prior on the parameters, what is the expected number of modes? What is the distribution of the number of modes?
- In the isotropic case, the best lower bound is $k^{1.261}$ in dimension $d = 2\log_3 k$ [3]. Is the number of modes of isotropic mixtures in fixed dimension bounded linearly in k?
- Are there natural/easily verifiable sufficient conditions to ensure that there are no more than k modes?
- For discrete variables, there is the notion of "strong modes", and their number obeys stricter upper bounds. Is there a notion of strong modes, maybe in terms of Eigenvalues of the Hessian, in the Gaußian case?
- Compare to the body of work on the number of modes of a gravitational potential of finitely many point masses.

References

[1] G. Alexandrovich, H. Holzmann and S. Ray, *On the number of modes of finite mixtures of elliptical distributions*, Algorithms from and for Nature and Life, Springer International Publishing (2013) 49–57.

[2] M. Carreira-Perpiñán and C. Williams, *On the number of modes of a Gaussian mixture*, Scale-Space Methods in Computer Vision, Lecture Notes in Computer Science **2695** (2003), 625–640.

[3] H. Edelsbrunner, B. Fasy and G. Rote, *Add isotropic Gaussian mixtures at own risk: more and more resilient modes in higher dimensions*, Proc. of 27th Annual Symposium of Computational Geometry (2012).
[4] S. Ray and B. Lindsay, *The topography of multivariate normal mixtures*, Annals of Statistics **33** (2005), 2042–2065.
[5] S. Ray and D. Ren, *On the upper bound of the number of modes of a multivariate normal mixture*, Journal of Multivariate Analysis **108** (2012), 41–52.
[6] R. Steele, B. Sturmfels and S. Watanabe, *Singular learning theory: connecting algebraic geometry and model selection in statistics*, American Institute of Mathematics Workshop Summary, `http://aimath.org/pastworkshops/modelselectionrep.pdf` (2011)

Quantum Non-Locality and Latent Causal Structures

David Gross

In the past few years, there has been an increasing interest among researchers from quantum information theory in describing the set of marginal distributions that are compatible with a given Bayesian network.

To understand the reason, we briefly sketch sketch the fundamental argument leading to *Bell's inequalities*. A Bell experiment is a physical procedure that follows a causal structure [1] described by a simple Bayesian network (Fig. 1). The simplest case that yields non-trivial results involves two observers – traditionally referred to as Alice and Bob – that operate experimental equipment in two distance places. A source emitting physical systems at regular time intervals to Alice and Bob is placed in the middle. The two experimenters perform measurements on the incoming particles and record their results. Due to the large distance between Alice and Bob, any dependencies between their respective records has to result from dependencies between the particles emanating from the central source. Thus, the joint distribution of the involved random variables should factorize with respect to the Bayesian network displayed in Fig. 1. A few lines of algebra (presented in the talk and produced in many textbooks, e.g. [2]) shows that the set of marginal distributions of Alice's and Bob's variables that are compatible with the given Bayesian network forms a convex polytope, which is a proper subset of the entire probability simplex. This is the *Bell polytope* and the linear inequalities defining its facets are *Bell's inequalities*. These inequalities have to be satisfied by any process whose causal structure conforms with the direct acyclic graph defining the Bayesian network. Yet, for some such processes, quantum mechanics predicts that the inequalities be violated – and experimental results indeed confirm this prediction.

One is forced to conclude that either the graph does not represent the actual causal structure of the process, or that something more fundamental is happening. Physicists have gone to extreme lengths in order to ensure that the causal constraints are actually respected. For example, many setups are now such that Alice and Bob respective measurements on a particle pair are space-like separated. By the special theory of relativity, this means that information can flow directly

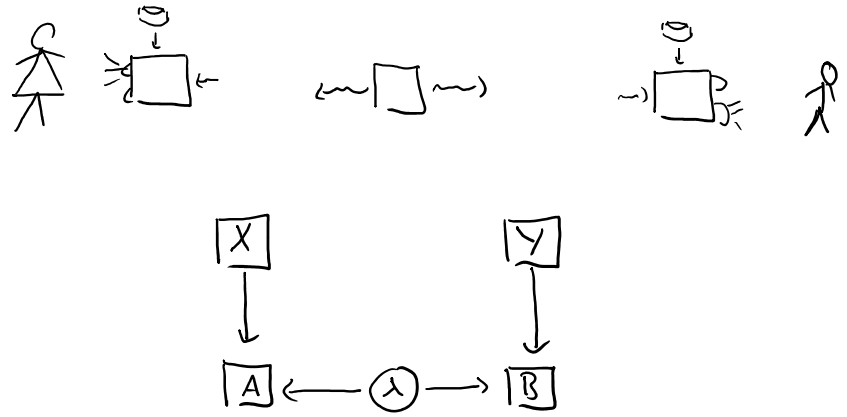


FIGURE 1. Sketch of a Bell experiment (top) and a directed acyclic graph (bottom) describing its causal structure. Top: Alice and Bob receive a pair of particles from a source placed between them. They then each perform a measurement on their respective particle. The measurement devices are depicted as boxes that take a particle and a random bit (the coins) as input. They process these inputs in an unspecified way and indicate the result to the experimenter (as represented by the two light bulbs on each box in the top figure). We emphasize that this extremely vague language is a strength rather than a weakness of the argument that applies in great generality. Bottom: There are two random variables on Alice's side: her coin X and the measurement outcome A. The situation on Bob's side is analogous. Any dependencies between (X, A) and (Y, B) result from their interaction with the particle pair. The distribution of the particles is described by the variable λ. This variable is not assumed to be directly accessible. We are thus interested in properties of the marginal distribution of (X, A, Y, B). Here and in the following, observed random variables are denoted by squared boxes, whereas latent variables appear in circles.

from Alice to Bob only if our understanding of the structure of space-time is fundamentally flawed. Thus, reluctantly, the mainstream of modern physics has come to the conclusion that one has to reject the possibility that all physical processes can be described using the framework of classical probability theory. In particular, the most-commonly held view is that Bell inequality violations mean that *one cannot consistently assign a value to unmeasured physical properties.* While this short introduction cannot possibly give justice to this insight, it is every bit as foundational, surprising, and revolutionary as it sounds. It's first consequence was to bring to a halt all attempts (advocated e.g. by Einstein) to find a successor to quantum mechanics that would reproduce the experimental results while also making predictions independently of any observer's choice of what to measure. Today,

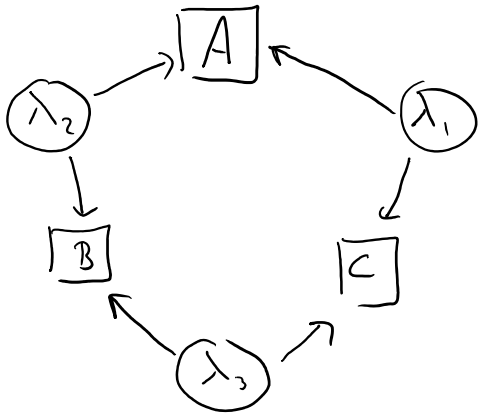


FIGURE 2. The *triangle scenario*. Three observed random variables A, B, C each depend on two latent factors λ_i.

Bell's result forms the basis not just of an improved understanding of quantum physics, but also of early technological applications — most prominently *quantum key expansion*.

The important aspect to notice is that Bell's argument does not presume validity of quantum mechanics. At its heart, it is a statement about marginals of completely classical Bayesian networks (with quantum mechanics only used as a hint for where to look for violations). Thus, our ability to generalize these fundamental findings to more complex causal structures hinges on the availability of constraints on marginals of Bayesian networks. Unfortunately from the point of view of quantum physics, this algebraic statistics problem seems to be little understood.

Figure 2 shows the *triangle scenario*—one of the DAGs which quantum physicists have recently looked at, searching for more general forms of the effect identified by Bell (a non-exhaustive list of references is [3, 4, 5, 6, 7, 8, 9, 10]). However, even in this extremely simple and natural setting, a concise description of the marginals seems to be unavailable. Here, I list some seemingly elementary questions about this Bayesian network. To the best of my knowledge (and to my great embarrassment), all of them remain open despite repeated attempts by otherwise accomplished researchers. It is my hope that this teaser will trigger the attention of the algebraic statistics community, whose help quantum physics clearly needs here.

In Figure 2, assume for simplicity that the three observed random variables are binary. Initially, we also limit the alphabet size of the three hidden variables to some fixed number, say d. Then the set of observable distributions compatible with this structure is a subset C_d of the 7-dimensional probability simplex of three binary random variables. In principle, quantifier elimination should be able to decide whether any given distribution is an element of C_d. In practice, however, even this scenario seems already to be out of reach for quantifier elimination on standard hardware. Problem: Give a practical algorithm that decides membership

with C_d. If that's too hard, do it for C_2. Next, it is clear that if $d < d'$, then $C_d \subset C_{d'}$. Does there exist a finite d such that $C_d = C_{d'}$ for all $d' \geq d$? If so, what is its value? Related: is the union C_∞ of all C_d's a closed set? Find an algorithm that tests membership with C_∞. (Even though it seems unlikely, for all we know, the membership problem for C_∞ could be Turing-undecidable.) More quantitatively: Upper-bound the difference between C_d and C_∞ in ℓ_1-norm. Does the problem simplify if the observed distribution is invariant under cyclic permutations or under all permutations of the three variables? In particular, can one use a symmetric model for symmetric distributions, without having to increase the hidden alphabet size d?

References

[1] J. Pearl, *Causality*, Cambridge University Press 2009.
[2] A. Peres, *Quantum Theory: Concepts and Methods*, Springer 2006.
[3] R. Chaves, T. Fritz. *Entropic approach to local realism and noncontextuality*, Physical Review A 85 (2012), 032113.
[4] T. Fritz, *Beyond Bell's theorem: correlation scenarios*, New Journal of Physics 14 (2012), 103001.
[5] R. Chaves, L. Luft, D. Gross *Causal structures from entropic information: Geometry and novel scenarios*, New J. Phys. 16, 043001 (2014).
[6] R. Chaves, C. Majenz, D. Gross *Information-Theoretic Implications of Quantum Causal Structures*, Nature Communications 6, 5766 (2015).
[7] E. Wolfe, R.W. Spekkens, T. Fritz *The Inflation Technique for Causal Inference with Latent Variables* arXiv:1609.00672 (2016).
[8] R. Evans, *Margins of discrete Bayesian networks*, arXiv:1501.02103 (2016).
[9] M. Weilenmann, R. Colbeck, *Non-Shannon inequalities in the entropy vector approach to causal structures*, arXiv:1605.02078 (2016).
[10] A. Kela, K. von Prillwitz, J. Aberg, R. Chaves, D. Gross *Semidefinite tests for latent causal structures*, arXiv:1701.00652 (2017).

Phylogenetic Mixtures and Linear Invariants for Evolutionary Models with Non-uniform Stationary Distribution

Marta Casanellas

(joint work with Mike Steel)

The discovery of the first linear *topology invariants* for Markov models of nucleotide substitution by James Lake in 1987 (see [5]) attracted great attention. Topology invariants for a phylogenetic tree T and a Markov model $\mathcal{M}$ are polynomials that vanish on any distribution arising from a Markov process on T whose transition matrices belong to $\mathcal{M}$ (and do not satisfy this property for another tree). They are useful because they allow (theoretically, and if one knows enough topology invariants) the identification of the tree topology from which a distribution arises.

On one hand, the interesting property of Lake's invariants was that they were *linear*. Although this may seem too simple for mathematicians, a linear topology invariant for a tree T has the property that it also vanishes on any *mixture* of distributions on the same tree T. Therefore, linear topology invariants do not

only allow to detect the tree topology from a distribution arising on a tree but also from a mixture of distributions on the same tree. Mixtures of distributions on the same tree appear often in biology, for example when one considers alignments containing coding and non-coding sites, different genes, or also different codon positions.

On the other hand, the drawback of Lake's invariants was that they were only valid for some simple models that were restricted to uniform stationary distribution (namely, Kimura 2-parameter and the Jukes-Cantor models). These models are too simple to account for certain biological processes. Although it would be desirable to have linear topology invariants for more complex models, it is well known that there are models for which they do not exist. This is the case of the Kimura 3-parameter model, the simplest model that encompasses Kimura 2-parameter and Jukes-Cantor.

We study linear invariants for the *equal-input model*, which can be thought of as the simplest Markov process that allows different states to have different stationary distribution. This is done for any number κ of states and any number of taxa and by assuming that the stationary distribution π is fixed. We describe the set of all linear invariants, distinguishing between *model invariants* (linear invariants that vanish on any distribution arising form the equal-input Markov process on any tree) and topology invariants. Whereas topology invariants can be used for phylogenetic topology reconstruction, model invariants are appropriate for model selection (see [4]).

The space of linear model invariants for the equal-input model is dual to the affine space $\mathcal{D}^\pi$ of mixtures of distributions that arise from the equal-input model on any tree topology on a set of n taxa. We provide a set of linearly independent points that span $\mathcal{D}^\pi$ for any n and any distribution π (see Theorem 1 in [2]) and prove the following result:

Theorem 9. *The affine space $\mathcal{D}^\pi$ has dimension $|\Sigma_\kappa| - 1$, where Σ_κ is the set of partitions of $[n] = \{1, \ldots, n\}$ of size at most κ.*

This generalizes previous results of Matsen, Mossen and Steel in [6] for binary states and of Casanellas, Kedzierska and Fernández-Sánchez (see [1]).

The space of linear topology invariants on a tree T is dual to the space $\mathcal{D}^\pi_T$ of mixtures of distributions that arise from the equal-input model on T. We provide a set of linearly independent points that span this space (this is equivalent to giving a set of linearly independent linear invariants) and prove the following result (see Theorem 2 in [2]):

Theorem 10. *The dimension of the affine space $\mathcal{D}^\pi_T$ is equal to the number of full subforests of the tree T. When T is a trivalent tree, $\mathcal{D}^\pi_T$ has dimension equal to the Fibonacci number F_{2n-1}.*

These results generalize previous results for the Jukes-Cantor model obtained by Steel and Fu (see [3], [7]). This study of linear topology invariants has allowed us to generalize Lake's invariants to the equal-input model (and to certain more general models, see Proposition 3 in [2]). While this type of invariants is sufficient

to describe the space of linear topology invariants for quartets and for $\kappa = 4$, we prove that it might not sufficient for different number of leaves or different number of states.

References

[1] Casanellas, M., Fernández-Sánchez, J., Kedzierska, A.M., *The space of phylogenetic mixtures for equivariant models*, Alg. Mol. Biol. **7** (2012), 33.

[2] Casanellas, M., Steel, M., *Phylogenetic mixtures and linear invariants for equal input models*, J. Math. Bio. **74** (2017), 1107–1138.

[3] Fu, Y.X., *Linear invariants under Jukes' and Cantor's one-parameter model*, J. Theor. Biol. **173** (1995), 339–352.

[4] Kedzierska, A., Drton, M., Guigó, R., Casanellas, M., *SPIn: model selection for phylogenetic mixtures via linear invariants*, Mol. Biol. Evol. **29** (2012), 929–937.

[5] Lake, J., *A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony*, Molec. Biol. Evol. **4** (1987), 167–191.

[6] Matsen, F.A., Mossel, E., Steel, M., *Mixed-up trees: The structure of phylogenetic mixtures*, Bull. Math. Biol. **70** (2008), 1115–1139.

[7] Steel, M.A., Fu, Y.X.,
Classifying and counting linear phylogenetic invariants for the Jukes–Cantor model, J. Comput. Biol. **2** (1995), 39–47.

Exact Tests for Stochastic Block Models and Extensions to Latent-variable Log-linear Models

Petrović, Sonja

(joint work with Vishesh Karwa, Debdeep Pati, Liam Solus, Nikita Alexeev, Mateja Raič, Dane Wilburne, Robert Williams, Bowei Yan)

Problem background. Analysis of networks – relational data on a fixed set of nodes/actors – has become increasingly popular with wide-ranging applications at the intersection of applied mathematics, statistics, computer science and machine learning. Exciting theoretical and algorithmic developments have been motivated by the ever-increasing availability of network data in diverse fields such as social sciences, web recommender systems, protein networks, genomics and neuroscience, to name a few. There is a rich literature on probabilistic modeling of network data; [9] contains a detailed review. In particular, the stochastic block model (or SBM), originally proposed in the social sciences ([6], [11]), has since attained centerstage in computer science, statistics and machine learning as one of the more popular approaches that induce community structures in networks. In an SBM, each of the n nodes in the network belongs to one of the k pre-defined blocks (or groups), $k \leq n$. As a generalization of the Erdös-Rényi random graph model, an SBM allows the probabilities of occurrences of the edges between different pairs of nodes to be distinct and – crucially – depend on the block membership of the two nodes in the pair. SBMs have been further extended to allow for various heterogeneities in the model that fit a myriad of application scenarios and block/node parameter options, incorporating node degrees, etc.

In the growing literature on probabilistic network models, the question of whether

these models provide an adequate fit to the data has received relatively little attention. However, this is an important practical question which is not only relevant for the adequacy of a single model, but can be generalized to ask questions whether we should fit a block model, or its variants, to the data. Even so, a large part of the literature on computation and modeling does not address model adequacy issues beyond heuristic algorithms [12], [2]. This is largely due to inherent model complexity or degeneracy and the lack of tools that can handle network models and sparse small-sample data. Unfortunately, most goodness-of-fit tests are based on *large sample approximations* that are not applicable in many settings. In contrast, various problems relating to *exact* goodness-of-fit tests are well-studied for contingency tables and some examples of exponential random graph models - one can view these developments as a cornerstone of algebraic statistics for contingency tables (see, for example, [1], [3], [10], and the literature cited therein). There is, however, a dearth of such tests for stochastic block models and its variants when the block assignments of nodes, as well as the number of blocks, are unknown.

Results and methodology. Since stochastic block models with unknown block assignment are widely used in analysis of real-world network data, in [13] we focus on finite-sample goodness-of-fit tests for three different variants of SBMs. Specifically, we describe the non-asymptotic test for both known and unknown block assignment of nodes; the latter is built from the former. An exact testing framework that is applicable to any variant of the SBM when the block assignment for each node is *known* uses by now a well-known approach in algebraic statistics: conditioning on the sufficient statistic of the model and sampling from the conditional distribution on the model fiber using Markov bases. In this case, there are two main ingredients: 1) a valid choice of a goodness-of-fit statistic and 2) a good way to sample from the conditional distribution given the sufficient statistics. Commonly used statistics choices include the chi-square, however, for some simplistic SBMs it collapses to a constant on the fiber and renders the test meaningless; to that end, we develop a block-corrected version of χ^2 and show it provides a valid test. Unsurprisingly, sampling methods rely on Markov bases, which we derive for the three variants of the SBM. We achieve good performance using dynamic Markov bases through an interpretation of the models as log-linear models on 0/1 contingency tables, as suggested in [5]. The algebraic and geometric model structure also inspires some theoretical results on the geometry of the three model polytopes with direct implications on the existence of MLE.

The usually more interesting scenario in applications is when the block assignments are *latent*; this is a new scenario for a finite-sample test in algebraic statistics. Specifically, when the block assignment and the number of blocks are not known, the entire network is a minimal sufficient statistic, so conditioning as above achieves nothing. Instead, we propose a novel way to exploit the exact test for the non-latent model in combination with a consistent, asymptotically valid method for estimating the block assignment. When the number of blocks is known, the Bayesian approach [14] demonstrates good empirical evidence for consistent estimation of block assignments. When the number of blocks is unknown, the

mixture-of-finite-mixtures method for SBM from [8] is one such method that is provably consistent. We take such methods as a black box, essentially, and, not surprisingly, we assume that they offer a good estimator of block assignments.

The interpretation of our goodness-of-fit test varies in a few important ways depending on the method used to estimate the latent block assignments. If a classical method is used to estimate the block assignments, what we propose is an exact conditional test [7] by sampling from the fiber conditional on the estimated block assignments. If a Bayesian approach is used to provide a posterior distribution of the block assignments, one can sample from the posterior predictive distribution by first drawing samples from posterior distribution of the block assignments and then sampling from the fiber conditional on the block assignments. Posterior predictive checks for model validation are popular Bayesian counterparts of p-values and are suited to latent variables models or models with abundance of nuisance parameters. Sampling from the posterior predictive distribution is challenging in general and this intermediate step of sampling from the *conditional fiber* provides an efficient way to gather the samples. The posterior quantity analogous to p-value here is the posterior predictive-p-value, which can be viewed as the posterior mean of the classical p-value. Despite its controversy regarding issues with calibration, we obtained promising results in delivering accurate Type I and II errors. We test our methods on synthetic and real data sets: the Karate network and the brain connectome data.

Extensions. The main point to take away is that latent-block SBMs are mixtures of known-block ERGMs. While we focused here on three variants of the SBM, the proposed methodology *extends to any mixture of log-linear models* on discrete data. The geometry of mixtures is explained in [4], who explore the link between the geometric and statistical model properties and the implications on parameter estimation.

I would be extremely interested in using this methodology on another type of data in a non-network application. While we have a general approach for constructing Markov moves on the fly (i.e., dynamically, so that they are applicable and data-dependent), I am not aware of a general method to estimate the mixture parameters with provably good properties.

References

[1] Satoshi Aoki, Hisayuki Hara, and Akimichi Takemura, Markov bases in algebraic statistics, Springer Series in Statistics, Springer New York, 2012.

[2] Nicole B. Carnegie, Pavel N. Krivitsky, David R. Hunter, and Steven M. Goodreau, An approximation method for improving dynamic network model fitting, Journal of Computational and Graphical Statistics 24 (2015), no. 2, 502–519.

[3] Persi Diaconis and Bernd Sturmfels, Algebraic algorithms for sampling from conditional distributions, Annals of Statistics 26 (1998), no. 1, 363–397.

[4] Stephen E. Fienberg, Patricia Hersh, Alessandro Rinaldo, and Yi Zhou, Maxi- mum likelihood estimation in latent class models for contingency table data, vol. Algebraic and Geometric Methods in Statistics, Cambridge University Press, 2007.

[5] Stephen E. Fienberg, Michael M. Meyer, and Stanley S. Wasserman, Statistical analysis of multiple sociometric relations, Journal of the American Statistical Association 80 (1985), no. 389, 51–67.
[6] Stephen E. Fienberg and Stanley S. Wasserman, Categorical data analysis of single sociometric relations, Sociological methodology 12 (1981), 156–192.
[7] Ronald Aylmer Fisher, Statistical methods for research workers, Genesis Publishing Pvt Ltd, 1925.
[8] Junxian Geng, Anirban Bhattacharya, and Debdeep Pati, Probabilistic community detection with unknown number of communities, arXiv preprint arXiv:1602.08062 (2016).
[9] Anna Goldenberg, Alice X. Zheng, Stephen E. Fienberg, and Edoardo M. Airoldi, A survey of statistical network models, Foundations and Trends in Machine Learning 2 (2010), no. 2, 129–233.
[10] Elizabeth Gross, Sonja Petrović, and Despina Stasi, Goodness-of-fit for log-linear network models: Dynamic Markov bases using hypergraphs, Annals of the Institute of Statistical Mathematics (2016), DOI: 10.1007/s10463-016-0560-2.
[11] Paul W. Holland, Kathryn B. Laskey, and Samuel Leinhardt, Stochastic blockmodels: First steps, Social networks 5 (1983), no. 2, 109–137.
[12] David R. Hunter, Steven M. Goodreau, and Mark S. Handcock, Goodness of fit of social network models, Journal of the American Statistical Association 103 (2008), no. 481, 248–258.
[13] Vishesh Karwa, Debdeep Pati, Sonja Petrović, Liam Solus, Nikita Alexeev, Mateja Raič, Dane Wilburne, Robert Williams, Bowei Yan, Exact tests for stochastic block models, Submitted. Preprint available at arXiv:1612.06040.
[14] Debdeep Pati and Anirban Bhattacharya, Optimal bayesian estimation in stochastic block models, arXiv preprint arXiv:1505.06794 (2015).

Kullback Information of Gaussian Mixtures

Shaowei Lin
(joint work with Carlos Améndola, Mathias Drton)

Watanabe et al. proved in 2004 that the Kullback information of Gaussian mixtures is not an analytic function at points on the boundary of the parameter space where one of the mixing parameters is zero [3]. We present a partial translation of their result, which was published in Japanese. In particular, we show that the Kullback information is equivalent to the squared distance between the true distribution and the model distribution, using a resolution of singularities. We also show that the Kullback information is in fact equivalent to a polynomial – the sum of squares of the moments of a Gaussian mixture.

Given a model $p(x|\omega)$ with states $x \in \mathbb{R}^n$ and parameters $\omega \in \Omega$, the *Kullback information* $K : \Omega \to \mathbb{R}$ at the true distribution $p(x|\omega^*)$ is defined by

$$K(\omega) = \int_{\mathbb{R}^n} p(x|\omega^*) \log \frac{p(x|\omega^*)}{p(x|\omega)} dx = \int_{\mathbb{R}^n} f(x,\omega)^2 \, S(x,\omega) dx,$$

where $f(x,\omega) = (\, p(x|\omega)/p(x|\omega^*) - 1)^2$,

$$S(x,\omega) = S\left(\frac{p(x|\omega)}{p(x|\omega^*)}\right) p(x|\omega^*), \quad \text{and} \quad S(t) = \frac{-\log t + t - 1}{(t-1)^2}.$$

The Kullback information tells us a lot about the behavior of learning algorithms. For instance, the marginal likelihood integral

$$Z_N = \int_\Omega \prod_{j=1}^N p(x_j|\omega)\varphi(\omega)d\omega$$

where $x_1, \ldots, x_N \in \mathbb{R}^n$ are the observed data points, is asymptotically

$$\log Z_N = \sum_{i=1}^N \log p(x_i|\omega^*) - \lambda \log N + (\theta - 1) \log\log N + \eta_N,$$

where (λ, θ) is the *real log canonical threshold* (RLCT) of $K(\omega)$, and η_N is a random variable whose expectation tends to a constant as $N \to \infty$. Watanabe showed that this asymptotic result holds for Gaussian mixtures, even though the Kullback information is non-analytic [3, Theorem 2].

Finding the RLCT of non-analytic functions can be computationally challenging. To overcome this problem, it is often useful to find simpler functions or even polynomials that have the same RLCT. Given $f, g : \Omega \to \mathbb{R}_{\geq 0}$, f and g are *equivalent* over Ω if there exists constants $c_1, c_2 > 0$ such that $c_1 f(\omega) \leq g(\omega) \leq c_2 f(\omega)$ for all $\omega \in \Omega$. If two functions are equivalent, then their RLCTs are equal.

Our goal to show that $K(\omega)$ is equivalent to the density distance

$$L(\omega) = \int_{\mathbb{R}^n} \left(p(x|\omega) - p(x|\omega^*)\right)^2 dx = \int_{\mathbb{R}^n} f(x,\omega)^2\, p(x|\omega^*)^2 dx.$$

We need two assumptions. First, we assume that the prior $\varphi(\omega)$ is supported in a compact semi-analytic set Ω. Moreover, for all $\omega \in \Omega$, $\varphi(\omega) = \varphi_0(\omega)\varphi_1(\omega)$ where $\varphi_0(\omega) > 0$ is C^∞-smooth and $\varphi_1(\omega) \geq 0$ is analytic. Second, we assume that there exists a real-analytic function $\bar{S} : \mathbb{R}^n \to \mathbb{R}$ such that for all $\omega \in \Omega$,

$$S(x,\omega) < \bar{S}(x) \quad \text{and} \quad p(x|\omega^*)^2 < \bar{S}(x),$$

and the integral $\bar{K}(\omega) = \int_{\mathbb{R}^n} f(x,\omega)^2\, \bar{S}(x)dx$ is finite and real-analytic over Ω. For Gaussian mixtures, Watanabe verified that these two assumptions are satisfied [3].

From the assumptions, both $K(\omega)$ and $L(\omega)$ are bounded above by $\bar{K}(\omega)$. Since $\bar{K}(\omega)$ is real-analytic, there exists a resolution of singularities $\rho : \mathcal{M} \to \Omega$ where in each chart with local coordinates $\mu = (\mu_1, \ldots, \mu_m)$, we have

$$\bar{K}(\rho(\mu)) = \mu^{2\kappa}$$

for some non-negative integer vector $\kappa = (\kappa_1, \ldots, \kappa_m)$. Locally in each chart, we have the series expansion of the real-analytic function

$$f(x, \rho(\mu)) = \sum_{\tau \geq 0} a_\tau(x)\mu^\tau.$$

By considering the resulting expansion

$$\mu^{2\kappa} = \int_{\mathbb{R}^n} \Big(\sum_{\tau \geq 0} a_\tau(x)\mu^\tau\Big)^2 \bar{S}(x)dx$$

of $\bar{K}(\omega)$, we see that $a_\tau(x) \neq 0$ only if $\tau \geq \kappa$. Therefore,

$$f(x, \rho(\mu)) = a(x, \mu)\mu^\kappa \quad \text{and} \quad \int_{\mathbb{R}^n} a(x, \mu)^2 \, \bar{S}(x)dx = 1$$

for some non-vanishing real-analytic function $a(x, \mu)$. Now, by choosing sufficiently large compact subsets $\mathcal{C}_1, \mathcal{C}_2 \subset \mathbb{R}^n$, we have lower bounds

$$\mu^{2\kappa} \geq K(\rho(\mu)) \geq \int_{\mathcal{C}_1} f(x, \omega)^2 \, S(x, \omega)dx = \mu^{2\kappa} \int_{\mathcal{C}_1} a(x, \mu)^2 \, S(x, \omega)dx,$$
$$\mu^{2\kappa} \geq L(\rho(\mu)) \geq \int_{\mathcal{C}_2} f(x, \omega)^2 \, p(x|\omega^*)^2 dx = \mu^{2\kappa} \int_{\mathcal{C}_2} a(x, \mu)^2 \, p(x|\omega^*)^2 dx,$$

where the integral coefficients of μ^κ are positive functions of μ. Consequently, both $K(\rho(\mu))$ and $L(\rho(\mu))$ are locally equivalent to $\mu^{2\kappa}$. Because there are finitely many charts, it follows that $K(\omega)$ is equivalent to $L(\omega)$.

Finally, we remark that the density distance $L(\omega)$ is equivalent to a polynomial. Let $\alpha = (\alpha_1, \ldots, \alpha_k) \geq 0$, $|a| := \sum \alpha_i = 1$, and $\mu = (\mu_1, \ldots, \mu_k) \in \mathbb{R}^{n \times k}$. Given $\omega = (\alpha, \mu)$, we define the Gaussian mixture model (GMM) with distribution

$$p(x|\omega) = (2\pi)^{-n/2} \sum_{j=1}^{k} \alpha_j \exp(-\frac{1}{2}\|x - \mu_j\|^2), \quad x \in \mathbb{R}^n.$$

If the true distribution has $k_0 \leq k$ components, then $L(\omega)$ is equivalent to

$$P(\omega) = \sum_{1 \leq |r| \leq k + k_0} \left(P_{r,k}(\omega) - P_{r,k_0}(\omega^*)\right)^2$$

where each $P_{r,k}(\omega)$ is the convex sum of monomials

$$P_{r,k}(\omega) = \sum_{k=1}^{k} \alpha_h \mu_h^r.$$

Hence, the ML variety is a fiber over a secant map of Veronese embeddings. A key step in the proof of this result involves observing that due to Parseval's Theorem, the density distance is equal to the *characteristic distance*

$$\chi(\omega) = \int_{\mathbb{R}^n} |\phi(t|\omega) - \phi(t|\omega^*)|^2 \, dt$$

where the complex-valued function

$$\phi(t|\omega) = \int_{\mathbb{R}^n} e^{itx} p(x|\omega)dx$$

is the characteristic function or Fourier transform of $p(\,\cdot\,|\omega)$.

References

[1] S. Lin: Ideal-Theoretic Strategies for Asymptotic Approximation of Marginal Likelihood Integrals, *Journal of Algebraic Statistics* **8**:1 (2017).

[2] S. Watanabe: *Algebraic Geometry and Statistical Learning Theory*, Cambridge Monographs on Applied and Computational Mathematics **25**, Cambridge University Press (2009).

[3] S. Watanabe, K. Yamazaki, and M. Aoyagi: Kullback information of normal mixture is not an analytic function, *Technical Report of IEICE* (in Japanese) NC2004-50 (2004) 41-46.

Restricted Boltzmann Machines

Guido Montúfar

(joint work with Jason Morton, Johannes Rauh)

The restricted Boltzmann machine with n visible binary variables and m hidden binary variables is the set $\mathcal{M}_{n,m}$ of probability distributions of the form

$$p_\theta(x) = \frac{1}{Z(\theta)} \sum_{y\in\{0,1\}^m} \exp\Big(\sum_{i,j} \theta^I_{i,j} x_i y_j + \sum_i \theta^V_i x_i + \sum_j \theta^H_j y_j\Big), \quad x \in \{0,1\}^n,$$

parametrized by $\theta = (\theta^I, \theta^V, \theta^H) \in \mathbb{R}^{nm+n+m}$. The partition function $Z(\theta)$ ensures that $\sum_{x\in\{0,1\}^n} p_\theta(x) = 1$. The restricted Boltzmann machine is a prominent generative model in machine learning and building block of deep neural networks. In mathematical terms, $\mathcal{M}_{n,m}$ is a semialgebraic subset of the simplex Δ_{2^n-1} of probability distributions on $\{0,1\}^n$. It can be regarded as the set of $2\times\cdots\times 2$ (n times) probability tables expressible as normalized entrywise Hadamard products of m tables of non-negative tensor rank at most two. In particular, $\mathcal{M}_{n,1}$ is the set of mixtures of pairs of binary product distributions.

In this talk I present recent advances on two central questions about the geometry of the restricted Boltzmann machine:

- What is the dimension of the set $\mathcal{M}_{n,m}$?
- Given n, what is the smallest m for which $\overline{\mathcal{M}_{n,m}} = \Delta_{2^n-1}$?

Regarding the first point, in [5] Jason Morton and I completed the dimension characterization started by Cueto, Morton, and Sturmfels [1], and proved their conjecture stating that the restricted Boltzmann machine always has the expected dimension.

Theorem 11. *For any non-negative integers n and m, the restricted Boltzmann machine with n visible and m hidden binary units has the dimension expected from parameter counting, that is,* $\dim(\mathcal{M}_{n,m}) = \min\{2^n - 1, (n+1)(m+1) - 1\}$.

Regarding the second point, in [6] Johannes Rauh and I obtained a new result on the minimal number of hidden units m that suffices to approximate any probability distribution on $\{0,1\}^n$ to within any desired degree of accuracy, thereby improving a series of previous results [2, 3, 7, 4].

Theorem 12. *Every probability distribution on $\{0,1\}^n$ can be approximated arbitrarily well by probability distributions from the restricted Boltzmann machine with n visible and m hidden binary units whenever $m \geq \frac{2(\ln(n-1)+1)}{n+1}(2^n-(n+1)-1)+1$.*

A simple lower bound on the minimal sufficient number of hidden units is $m \geq \frac{2^n}{n+1} - 1$. Thus we now know that the behavior of the exact number is between $\frac{2^n}{n}$ and $\frac{\log n}{n}2^n$. We would love to see more advances closing this gap.

References

[1] M. A. Cueto, J. Morton, and B. Sturmfels. Geometry of the restricted Boltzmann machine. In M. A. G. Viana and H. P. Wynn, editors, *Algebraic methods in statistics and probability II*, pages 135–153. AMS, 2010.

[2] Y. Freund and D. Haussler. Unsupervised learning of distributions of binary vectors using 2-layer networks. In J. E. Moody, S. J. Hanson, and R. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 912–919. Morgan Kaufmann, 1991.

[3] N. Le Roux and Y. Bengio. Representational power of restricted Boltzmann machines and deep belief networks. *Neural Computation*, 20(6):1631–1649, 2008.

[4] G. Montúfar and N. Ay. Refinements of universal approximation results for deep belief networks and restricted Boltzmann machines. *Neural Computation*, 23(5):1306–1319, 2011.

[5] G. Montúfar and J. Morton. Dimension of marginals of Kronecker product models. *SIAM Journal on Applied Algebra and Geometry*, 1(1):126–151, 2017.

[6] G. Montúfar and J. Rauh. Hierarchical models as marginals of hierarchical models. *International Journal of Approximate Reasoning, in press*, 2016. Available online at `http://dx.doi.org/10.1016/j.ijar.2016.09.003`.

[7] L. Younes. Synchronous Boltzmann machines can be universal approximators. *Applied Mathematics Letters*, 9(3):109–113, 1996.

An Algebraic Theory of Negative Dependence

Petter Brändén

In this talk we give a short introduction to an algebraic approach to negative dependence in discrete probability theory developed by Borcea, Liggett and the author in [1]. Let $X = (X_1, \ldots, X_n)$ be a vector of random variables taking values 0 or 1. The *multivariate partition function* of X is the polynomial

$$Z_X(\mathbf{z}) = \mathbb{E}(\mathbf{z}^X) = \sum_{\alpha \in \{0,1\}^n} \mathbb{P}(X = \alpha)\mathbf{z}^\alpha,$$

where $\mathbf{z} = (z_1, \ldots, z_n)$ and $\mathbf{z}^\alpha = z_1^{\alpha_1} \cdots z_n^{\alpha_n}$. Recall that X is *negatively associated* if

$$\mathbb{E}(f(X)g(X)) \leq \mathbb{E}(f(X)) \cdot \mathbb{E}(g(X)),$$

for all functions $f, g : \{0,1\}^n \to \mathbb{R}$ depending on disjoint sets of variables. A polynomial $P(z_1, \ldots, z_n) \in \mathbb{C}[z_1, \ldots, z_n]$ is *stable* if $P(z_1, \ldots, z_n) \neq 0$ whenever all variables lie in the half-plane $\{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$.

One of the main theorem in [1] from which many other results may be deduced is

Theorem 13. *If $Z_X(\mathbf{z})$ is stable, then X is negatively associated.*

References

[1] J. Borcea, P. Brändén, T. M. Liggett, Negative dependence and the geometry of polynomials, J. Amer. Math. Soc. **22** (2009), no. 2, 521–567.

Statistics on Interesting Spaces and Interesting Statistics on Spaces

Stephan F. Huckemann

We set objectives of *Non-Euclidean Statistics*,

(1) design of data descriptors as minimizers of generalized Fréchet functions

$$E_n = \operatorname*{argmin}_{p\in\hat{P}} \sum_{j=1}^{n} \rho(X_j, p)^2$$

where $\hat{P}$ is a possibly random descriptor space approximating the data $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X \in Q$ best – for example the set of all points (*nested means*) on a random geodesic, cf. [12, 10], or the set of all lower dimensional small spheres (*principal nested spheres*) within a random higher dimensional small subsphere, cf. [13];

(2) obtain their asymptotic distribution ([4, 1, 5]) building on *Ziezold's strong consistency* ([18, 6]),

$$\bigcap_{n=0}^{\infty} \overline{\bigcup_{k=n}^{\infty} E_k} \subset E(X) \text{ a. s.} \tag{1}$$

with $E(X) = \operatorname{argmin}_{p\in P} \mathbb{E}\big[\rho(X,p)^2\big]$, where P is a suitable descriptor space;

(3) in order to obtain inferential results for driving applications such as RNA residue geometries, cf. [3], or filament orientations in early human mensenchymal stem cell differentiation, cf. [9];

in relation to some objectives of *Algebraic Statistics* with its focus

(4) on algebraic equations.

The hunt for analogs of PCA to non-Euclidean spaces has recently led to *barycentric subspaces* by [16, 15] which naturally form a flag

$$\{\mu\} = p_0 \subset p_1 \subset \ldots p_{m-1} \subset p_m = Q$$

of nested subspaces beginning from a Fréchet mean $\mu \in Q$. The potential, describing these and other flags by sequences of equations has yet been neither deeply investigated nor throroughly exploited for statistical applications.

We also discuss some of the challenges arising when investigating assymptotic distributions of entire flags and single elements thereof. Among those is *manifold stability*, which asserts that a descriptor is assumed on the manifold part of a possibly non-trivially stratified space (cf. [7]). For example, on a non-negative curvature cone, the intrinsic mean is never a singular cone point, unless all mass is concentrated there. This is no longer true for a non-positive curvature cone,

where the cone point may be *sticky*, i.e. there are non-trivial distributions, that with every small perturbation thereof, have a unique intrinsic mean at the cone point, cf. [11]. It seems that a counterpart of stickiness, namely *smeariness*, can only occur in non-negative curvature scenarios, and cases where $E(X)$ is not discrete may be linked to limiting cases of infinite smeariness.

Often, geometries which seem very benign, for instance the canonical flat geometry of the torus, are statistically not at all benign, because every data set is arbitrarly well approximated by almost any geodesic. Similarly, the phenomenon of stickiness yielding degenerate limiting fluctuation, may not allow for any asymptotic statistic, as can be the case in BHV tree space ([17]) introduced by [2]. To this end one may slightly change geometries as in [3] or consider new geometries, cf. [14]. These and other, still open research problems in this context, some of which listed in [8] are expanded.

References

[1] Bhattacharya, R. N. and V. Patrangenaru (2005). *Large sample theory of intrinsic and extrinsic sample means on manifolds II*, The Annals of Statistics, **33**(3), 1225–1259. 2015 proceedings, 22 - 29.

[2] Billera, L., S. Holmes and K. Vogtman (2001). *Geometry of the space of phylogenetic trees*, Advances in Applied Mathematics, **27**, 733–767.

[3] Eltzner, B., S. F Huckemann and K. V. Mardia (2015). *Torus Principal Component Analysis with an Application to RNA Structures*, arXiv:1511.04993

[4] Hendriks, H. and Z. Landsman (1998). *Mean location and sample mean location on manifolds: asymptotics, tests, confidence regions*, Journal of Multivariate Analysis, **67**, 227–243.

[5] Huckemann, S. (2011a). *Inference on 3D Procrustes means: Tree boles growth, rank-deficient diffusion tensors and perturbation models*, Scandinavian Journal of Statistics **38**(3), 424–446.

[6] Huckemann, S. (2011b). *Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth*, The Annals of Statistics, **39**(2), 1098–1124.

[7] Huckemann, S. (2012). *On the meaning of mean shape: Manifold stability, locus and the two sample test*, Annals of the Institute of Statistical Mathematics, **6**(6), 1227–1259.

[8] Huckemann, S. F. and B. Eltzner (2014). *Stickiness and Smeariness*, in Mini-Workshop: Asymptotic Statistics on Stratified Spaces, Oberwolfach Report No. 44/2014, 2520–2521.

[9] Huckemann, S. F. and B. Eltzner (2016). *Backward nested descriptors asymptotics with inference on stem cell differentiation*, arXiv:1609.00814.

[10] Huckemann, S., T. Hotz, and A. Munk (2010). *Intrinsic shape analysis: Geodesic principal component analysis for Riemannian manifolds modulo Lie group actions (with discussion)*, Statistica Sinica **20**(1), 1–100.

[11] Huckemann, S., J. C. Mattingly, E. Miller, and J. Nolen (2015). *Sticky central limit theorems at isolated hyperbolic planar singularities*, Electronic Journal of Probability **20**(78), 1–34.

[12] Huckemann, S. and H. Ziezold (2006). *Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces*, Advances in Applied Probability **38**(02), 299–319

[13] Jung, S., I. L. Dryden, and J. S. Marron (2012). *Analysis of principal nested spheres*, Biometrika **99**(3), 551–568.

[14] Lin B. and R. Yoshida (2016). *Tropical Fermat-Weber points*, arXiv:1604.04674.

[15] Nye, T. MW, X. Tang, G. Weyenberg and R. Yoshida (2016). *Principal component analysis and the locus of the Frechet mean in the space of phylogenetic trees*, arXiv preprint arXiv:1609.03045

[16] Pennec, X. (2015). *Barycentric subspaces and affine spans in manifolds*, in International Conference on Networked Geometric Science of Information (Springer), 12–21.
[17] Skwerer, S., E. Bullitt, S. Huckemann, E. Miller, I. Oguz, M. Owen, V. Patrangenaru, S. Provan, J.S Marron (2014). *Tree-oriented analysis of brain artery structure*, Journal of Mathematical Imaging and Vision, **50**, 126–143.
[18] Ziezold, H. (1977). *Expected figures and a strong law of large numbers for random elements in quasi-metric spaces*, Transaction of the 7th Prague Conference on Information Theory, Statistical Decision Function and Random Processes A, 591–602.

Tropical Principal Component Analysis

Ruriko Yoshida
(joint work with Leon Zhang and Xu Zhang)

A dimensionality reduction is applied to high-dimensional data sets in order to solve the problem called the "curse of dimensionality". This term refers to the problems associated with multivariate data analysis as the dimensionality increases. This problem of multidimensionality is acute in the rapidly growing area of phylogenomics, which can provide insight into relationships and evolutionary patterns of a diversity of organisms, from humans, plants and animals, to microbes and viruses. In this project, we are interested in applying the tropical metric in the max-plus algebra to computation of the principal component analysis over the space of rooted equidistant phylogenetic trees on m leaves, that is realized as the set of all ultrametrics. In this project, the proposed process of reducing the dimension of the multidimensional data sets on the "treespace" is to take data points in the space into a lower dimensional plane which minimizes the sum of distance between each point in the data set and their orthogonal projection onto the plane, that is, an optimization problem such that minimizing projection residuals between data points and their projections on the plane via the tropical metric in the max-plus algebra.

In this project, we assume that phylogenetic trees of m leaves are equidistant trees. Let D be a distance matrix computed from a phylogenetic tree, that is, a nonnegative symmetric $m \times m$-matrix $D = (d_{ij})$ with zero entries on the diagonal such that all triangle inequalities are satisfied:

$$d_{ik} \leq d_{ij} + d_{jk} \quad \text{for all } i, j, k \text{ in } [m] := \{1, 2, \ldots, m\}.$$

If a distance matrix D is computed from an equidistance tree, it is well-known that elements in D satisfy the following strengthening of the triangle inequalities [CITE]:

$$d_{ik} \leq \max(d_{ij}, d_{jk}) \quad \text{for all } i, j, k \in [m]. \tag{1}$$

If (1) holds then the metric D is called an *ultrametric*. The set of all ultrametrics contains the ray $\mathbb{R}_{\geq 0}\mathbf{1}$ spanned by the all-one metric $\mathbf{1}$, which is defined by $d_{ij} = 1$ for $1 \leq i < j \leq m$. The image of the set of ultrametrics in the quotient space $\mathbb{R}^{\binom{m}{2}}/\mathbb{R}\mathbf{1}$ is denoted $\mathcal{U}_m$ and called the *space of ultrametrics*. Therefore, we can consider the space of ultrametrics as a treespace for all possible equidistant

phylogenetic trees with m leaves. Let $e = \binom{m}{2}$. Tropical geometry gives an alternative geometric structure on $\mathcal{U}_m$, via the graphic matroid of the complete graph [3, Example 4.2.14], i.e., $\mathcal{U}_m$ can be written as a tropical linear space under the max-plus algebra. Under the max-plus algebra, we define $a \oplus b = \max\{a, b\}$ and $a \odot b = a + b$ where $a, b \in \mathbb{R}$.

In order to compute the distance between two points v, w on $\mathcal{U}_m$, we will use the *tropical distance* which can be computed as follows:

$$d_{\mathrm{tr}}(v, w) \;=\; \max\big\{\, |v_i - w_i - v_j + w_j| \;:\; 1 \le i < j \le e \,\big\}, \tag{2}$$

where $v = (v_1, \ldots, v_e)$ and $w = (w_1, \ldots, w_e)$. This metric is also known as the *generalized Hilbert projective metric* [1, §2.2], [2, §3.3].

Our problem is to mimic the $(s-1)$th principal component so that we can find the convex hull which minimizes the distances between each point in the sample to its projection onto the convex hull. Note that wee can re-write this problem the following problem.

Problem 14. *We want to find the solution for the following optimization problem:*

$$\min_{D^{(1)}, D^{(2)}, D^{(3)} \in \mathcal{U}_m} \sum_{i=1}^{n} d_{\mathrm{tr}}(d_i, d_i')$$

where

$$d_i' = \lambda_1^i \odot D^{(1)} \;\oplus\; \lambda_2^i \odot D^{(2)} \;\oplus\; \lambda_3^i \odot D^{(3)}, \quad \text{where } \lambda_k^i = \min(d_i - D^{(k)}),$$

and

$$d_{\mathrm{tr}}(d_i, d_i') = \max\{|d_i(k) - d_i'(k) - d_i(l) + d_i'(l)| : 1 \le k < l \le e\}$$

with $d_i = (d_i(1), \ldots, d_i(e))$ *and* $d_i' = (d_i'(1), \ldots, d_i'(e))$.

Proposition 15. *Problem 14 can be formulated as the following optimization problem:*

$$\min_{\Delta_1, \ldots, \Delta_n \in \mathbb{R}; d_1', \ldots, d_n' \in \mathcal{S}'} \sum_{i=1}^{n} \Delta_i \tag{3}$$

$$\text{subject to } \Delta_i \ge d_i(k) - d_i'(k) - d_i(l) + d_i'(l), \forall 1 \le k < l \le e, i = 1, 2, \ldots, n$$
$$\Delta_i \ge -[d_i(k) - d_i'(k) - d_i(l) + d_i'(l)], \forall 1 \le k < l \le e, i = 1, 2, \ldots, n$$

where

$$(d_i')(k) = \max_{\lambda_1^i, \lambda_2^i, \lambda_3^i \in \mathbb{R}; D^{(1)}, D^{(2)}, D^{(3)} \in \mathcal{U}_m} (\lambda_1^i + D^{(1)}(k), \lambda_2^i + D^{(2)}(k), \lambda_3^i + D^{(3)}(k))$$

$$\text{subject to } \lambda_1^i + D^{(1)}(t) \le d_i(t), \forall t = 1, 2, 3, \ldots, e$$
$$\lambda_2^i + D^{(2)}(t) \le d_i(t), \forall t = 1, 2, 3, \ldots, e$$
$$\lambda_3^i + D^{(3)}(t) \le d_i(t), \forall t = 1, 2, 3, \ldots, e.$$

References

[1] M. Akian, S. Gaubert, N. Viorel and I. Singer: *Best approximation in max-plus semimodules*, Linear Algebra Appl. **435** (2011) 3261–3296.

[2] G. Cohen, S. Gaubert and J.P. Quadrat: *Duality and separation theorems in idempotent semimodules*, Linear Algebra Appl. **379** (2004) 395–422.

[3] D. Maclagan and B. Sturmfels: *Introduction to Tropical Geometry*, Graduate Studies in Mathematics, 161, American Mathematical Society, Providence, RI, 2015.

Distinguishing Phylogenetic Networks

Elizabeth Gross

(joint work with Colby Long)

Phylogenetic trees are graphical summaries of the evolutionary history of a set of species. In a phylogenetic tree, the interior nodes represent extinct species, while the leaves represent extant, or living, species. While trees are a natural choice for representing evolution visually, by restricting to the class of trees, it is possible to miss more complicated events such as hybridization and horizontal gene transfer. For more complete descriptions, phylogenetic networks, directed acyclic graphs, are increasingly becoming more common in evolutionary biology. Here we focus on phylogenetic networks and the algebraic problems associated to their inference.

Using aligned DNA sequences, several methods exist for reconstructing phylogenetic trees, with most of these methods falling into two main classes, distance-based methods and model-based methods. Model-based methods are amenable to analysis using algebraic geometry and have been a recurring theme in algebraic statistics [1]. In a model-based method, it is assumed that evolution proceeds according to a Markov process along a tree $\mathcal{T}$. Each edge e of $\mathcal{T}$ has an associated transition probability matrix M_e whose (i,j)th entry is the probability of transitioning from state j into state j. For applications to phylogenetics, it is commonly assumed that each vertex can take values in one of four possible states, A, C, G, T and that the transition matrices all adher to a specific pattern. For example, under the Jukes-Cantor model, each transition matrix is assumed to have the following form

$$\begin{pmatrix} \alpha & \beta & \beta & \beta \\ \beta & \alpha & \beta & \beta \\ \beta & \beta & \alpha & \beta \\ \beta & \beta & \beta & \alpha \end{pmatrix}.$$

In a tree-based Markov model, the joint probability distribution of the leaves can be described with a polynomial map $\phi_{\mathcal{T}}$ from the parameter space (the entries of the transition matrices) to probability space. We define $\mathcal{V}_{\mathcal{T}}$ to be the Zariski closure of the image of $\phi_{\mathcal{T}}$.

It is known that under the general Markov model and the popular group-based models, e.g. Jukes-Cantor, Kimura 2-parameter (K2P), and Kimura 3-parameter (K3P), two distinct unrooted n-leaf trees $\mathcal{T}_1$ and $\mathcal{T}_2$ are *distinguishable*, that is $\mathcal{V}_{\mathcal{T}_1} \cap \mathcal{V}_{\mathcal{T}_2}$ is a proper subvariety of $\mathcal{V}_{\mathcal{T}_1}$ and of $\mathcal{V}_{\mathcal{T}_2}$. These distinguishability results

imply generic identifiability of the tree parameter for these models. Our study focuses on a similar question for networks: *Given two distinct n-leaf networks $\mathcal{N}_1$ and $\mathcal{N}_2$, under what conditions is it true that $\mathcal{V}_{\mathcal{N}_1} \cap \mathcal{V}_{\mathcal{N}_2}$ is a proper subvariety of $\mathcal{V}_{\mathcal{N}_1}$ and of $\mathcal{V}_{\mathcal{N}_2}$?*

We first begin by defining phylogenetic networks. A *phylogenetic network* $\mathcal{N}$ [2] on a set of leaves X is a rooted acyclic digraph with no edges in parallel such that the root has out-degree two, a vertex with out-degree zero has in-degree one, the set of vertices with out-degree zero is X, and all other vertices either have in-degree one and out-degree two, or in-degree two and out-degree one. A vertex with indegree two and outdegree one is called a *reticulation vertex* and edges directed into a reticulation edge are called *reticulation edges*. The main result of our work focuses on *k-cycle networks*, semi-directed networks with one reticulation vertex that contains a k-cycle.

In phylogenetic network models, as in phylogenetic tree models, a transition matrix is assigned to each edge and evolution is assumed to proceed according to a Markov process. In this setting, network models can be viewed as the mixture of two tree models where the transition matrices for the edges in each tree are inherited from the network. Then the model of the network $\mathcal{M}_{\mathcal{N}}$ is the image of a polynomial map from the parameter space of the network to the probability simplex,

$$\phi_{\mathcal{N}} : \Theta_{\mathcal{N}} \times [0,1] \to \Delta^{4^n-1}, \ (\theta, \delta) \mapsto \delta\phi_{\mathcal{T}_1}(\theta) + (1-\delta)\phi_{\mathcal{T}_2}(\theta).$$

We define $\mathcal{V}_{\mathcal{N}}$ to be the Zariski closure of $\mathcal{M}_{\mathcal{N}}$ and $I_{\mathcal{N}} := I(\mathcal{V}_{\mathcal{N}})$.

Working under the Jukes-Cantor model and in Fourier coordinates (see[4]), we can classify all possible ideals for 4-leaf k-cycle networks:

Proposition (G.-Long): For 4-leaf k-cycle Jukes-Cantor networks, there are

- 3 unique 6-dimensional ideals corresponding to 2-cycle networks (trees).
- 6 unique 7-dimensional ideals corresponding to 3-cycle networks.
- 12 unique 8-dimensional ideals corresponding to 4-cycle networks.

Although there are 4-leaf 3-cycle networks whose varieties are subvarieties of varieties corresponding to 4-leaf 4-cycle networks, we are able to distinguish networks once the size of the cycle is large enough.

Theorem (G.-Long): For a fixed number of leaves, all pairs of k-cycle Jukes-Cantor networks with $k \geq 4$ are distinguishable.

While this handles most pairs of k-cycle networks in the Jukes-Cantor setting, this is only the beginning of the phylogenetics networks story from the algebraic and geometric viewpoint. There are several directions that can be pursued in algebraic statistics for phylogenetic networks. To this end, we list several open problems here.

Open Problem 1: For a fixed n, which pairs of k-cycle networks are distinguishable under other group-based models such as Cavender-Farris-Neyman (CFN), K2P, and K3P?

Since the CFN model has less parameters than the Jukes-Cantor model, it may be less likely to get as strong distinguishability results as the main theorem above, although an investigation would be interesting algebraically. To start, we have the following classification result for 4-leaf k-cycle CFN networks:
Proposition (G.-Long): For 4-leaf k-cycle CFN networks, there are

- 3 unique 6-dimension ideals corresponding to 2-cycle (trees) and 3-cycle networks.
- 3 unique 7-dimensional ideals corresponding to 4-cycle networks.

Open Problem 2: A larger class of networks that includes k-cycle networks are *level-1* networks, networks in which every edge belongs to at most one cycle. For a fixed n and fixed group-based model, which level-1 networks are distinguishable?

After level-1 networks, the next step might be to examine *tree-child* networks. Tree-child networks are identifiable from their *trinets*, which are induced subnetworks on 3-leaves [5].

Open Problem 3: For CFN, Jukes-Cantor, K2P, and K3P, given a network $\mathcal{N}$, what are the defining polynomials of $\mathcal{V}_\mathcal{N}$?

(A) (B)

FIGURE 1. Two different 4-leaf network k-cycle topologies.

For example, under the CFN model with the two states denoted 0 and 1, and using Fourier coordinates, the defining ideal for the 4-leaf 3-cycle network $\mathcal{N}_1$ pictured in Figure 1A is $I_{\mathcal{N}_1} = \langle p_{0110}p_{1001} - p_{0101}p_{1010},\ \ p_{0011}p_{1100} - p_{0000}p_{1111}\rangle$, while the defining ideal for the 4-leaf 4-cycle network $\mathcal{N}_2$ pictured in Figure 1B is $I_{\mathcal{N}_2} = \langle p_{0110}p_{1001} - p_{0101}p_{1010} + p_{0011}p_{1100} - p_{0000}p_{1111}\rangle$. Knowing the defining polynomials would be useful for model selection, which is the next step after establishing identifiability.

References

[1] M. Drton, B. Sturmfels and S. Sullivant, *Lectures on algebraic statistics*, Oberwolfach Seminars **39**, Birkhäuser (2009)
[2] A. R. Francis and M. Steel, Which phylogenetic networks are merely trees with additional arcs? *Systematic biology*, 64 no. 5 (2015) 768–777.
[3] E. Gross and C. Long, *Distinguishing phylogenetic networks*, In preparation.
[4] B. Sturmfels and S. Sullivant. Toric ideals of phylogenetic invariants, *Journal of Computational Biology*, 12 (2005) 204–228.
[5] L. Van Iersel and V. Moulton, Trinets encode tree-child and level-2 phylogenetic networks, *Journal of mathematical biology* 68 no. 7 (2014) 1707–1729.

The Correlation Space of Gaussian Latent Tree Models and Model Selection without Fitting

Piotr Zwiernik

(joint work with John Aston, Nat Shiers, and Jim Smith)

In phylogenetics and lynguistics latent tree models are used to model evolutionary processes. Model selection procedures are employed to choose the best tree fitting the data. However, deciding if the tree hypothesis is consistent with the data is typically hard. We provide the full semialgebraic description of Gaussian latent tree models and link them to phylogenetic oranges. We then use this geometric description to propose a quick and robust way of choosing the best tree, or, of testing the tree hypothesis.

References

[1] N. Shiers, P. Zwiernik, J. Aston, J.Q. Smith *The correlation space of Gaussian latent tree models and model selection without fitting*, Biometrika (2016) 103 (3): 531–545.
[2] V. Moulton, and M. Steel. *Peeling phylogenetic 'oranges'.* Advances in Applied Mathematics 33, no. 4 (2004): 710-727.
[3] P. Zwiernik. *Semialgebraic statistics and latent tree models.* CRC Press, 2016.

The Maximum Likelihood Data Singular Locus

Emil Horobeţ

(joint work with Jose I. Rodriguez)

For general data, the number of complex solutions to the likelihood equations is constant and this number is called the (maximum likelihood) ML-degree of the model. In this talk, we describe the special locus of data for which the likelihood equations have a solution in the model's singular locus. The talk is based on the recent results of the authors [4, 5].

Maximum likelihood estimation is an important problem in statistics. On a statistical model, one wishes to maximize the likelihood function for given data. The algebraic approach to this problem determines every critical point of the likelihood function on the the model's closure. For general data there will be

finitely many regular complex critical points. Moreover, this number remains constant and is called the maximum likelihood degree of the statistical model.

The (ML) maximum likelihood degree was introduced in [2] and [6]. In [7] Huh relates the ML degree of a smooth model to a topological Euler characteristic and used topological methods to classify varieties with ML degree one [8]. Recently, Euler characteristics and Gaussian degrees have been used to answer questions about the ML degree of a singular variety [1, 10, 11].

One reason to study the ML degree is because continuous deformations of the data induce continuous deformations of the critical points. So by deforming generic data to specific data, we are able to determine the critical points of the likelihood function as seen in [3] for example. For most choices of specific data, the critical points deform to distinct and regular critical points. However, for special choices of specific data, special behavior may occur. One type of special behavior is when the deformed critical points are no longer distinct. This was discussed from a computational view in [9] for the likelihood equations.

In this talk we discuss a different type of special behavior. We are interested in deformations of data leading a critical point into the singular locus. We call the closure of this type of special data the **(ML) maximum likelihood data singular locus**.

Our main theorem bounds the data singular locus. We give an algebraic variety contained in the data singular locus and an algebraic variety containing the data singular locus. These bounds connect dual varieties and Hadamard geometry to the ML data singular locus. We will give examples to show these bounds are strict.

Theorem 16. *Let X be an algebraic statistical model in $\mathbb{P}^{n+1}$. Then, the following two inclusions hold*

$$(X_{sing} \setminus \mathcal{H}) * [1 : \ldots : 1 : -1] \subseteq_{(1)} \mathrm{DS}(X) \subseteq_{(2)} (X_{sing} \setminus \mathcal{H}) * X^*,$$

where $\mathrm{DS}(X)$ is the data singular locus, X^* is the dual variety, $X_{sing} \setminus \mathcal{H}$ is the open part of the singular locus where none of the coordinates are zero and the Hadamard product $*$ is considered as in [5].

References

[1] Budur, N., and Wang, B. The Signed Euler Characteristic of Very Affine Varieties. *Int. Math. Res. Not. IMRN*, 14 (2015), 5710–5714.

[2] Catanese, F., Hoşten, S., Khetan, A., and Sturmfels, B. The maximum likelihood degree. *American Journal of Mathematics 128*, 3 (2006), 671–697.

[3] Hauenstein, J., Rodriguez, J. I., and Sturmfels, B. Maximum likelihood for matrices with rank constraints. *J. Algebr. Stat. 5*, 1 (2014), 18–38.

[4] Horobeţ, E. The Data Singular and the Data Isotropic Loci for Affine Cones. *Comm. Algebra, Volume 45 (2017), Issue 3, 1177 – 1186.*

[5] Horobeţ, E., and Rodriguez, J.I. The Maximum Likelihood Data Singular Locus. *J. Symbolic Comput., Volume 79 (2017), Part 1, 99–107.*

[6] Hoşten, S., Khetan, A., and Sturmfels, B. Solving the likelihood equations. *Found. Comput. Math. 5*, 4 (2005), 389–407.

[7] Huh, J. The maximum likelihood degree of a very affine variety. *Compos. Math. 149*, 8 (2013), 1245–1266.

[8] Huh, J. Varieties with maximum likelihood degree one. *J. Algebr. Stat. 5*, 1 (2014), 1–17.
[9] Rodriguez, J. I., and Tang, X. Data-discriminants of likelihood equations. In *Proceedings of the 2015 ACM on International Symposium on Symbolic and Algebraic Computation* (New York, NY, USA, 2015), ISSAC '15, ACM, pp. 307–314.
[10] Rodriguez, J. I., and Wang, B. The maximum likelihood degree of mixtures of independence models. *preprint arXiv:1505.06536* (2015).
[11] Wang, B. Maximum likelihood degree of Fermat hypersurfaces via Euler characteristics. *Proceedings of the American Mathematical Society 144.9 (2016): 3649-3655*

Testing Membership of the Likelihood Correspondence

Jose Israel Rodriguez

Maximum likelihood estimation is a fundamental computational task in statistics. A typical problem encountered in its applications is the occurrence of multiple local maxima. To be certain that a global maximum of the likelihood function has been achieved, one can locate all solutions to a system of polynomial equations called likelihood equations. The number of solutions to these equations is called the maximum likelihood degree (ML degree) and gives a measure of complexity to the global optimization problem. When the data yields finitely many solutions we can find them all. In my talk, I discussed a trace test to verify the computation of a multiprojective witness set which can be used to to test membership. In this extended abstract, the likelihood correspondence witness set will be defined and with it we will test membership.

We consider a statistical model, denoted by $\mathcal{M}$, contained in the $(n-1)$-dimensional probability simplex $\Delta_{n-1} := \{(p_1, p_2, \ldots, p_n) \in \mathbb{R}^n_{\geq 0} : \sum p_i = 1\}$. The Zariski closure of this model in $\mathbb{P}^{n-1}_{\mathbb{C}}$ is denoted by $X := \bar{\mathcal{M}}$, where $\mathcal{M}$ is a subset of the real points of X restricted to the affine chart defined by $\sum p_i = 1$. In other words, X is a projective variety, i.e., an algebraic statistical model. For fixed data $u = (u_1, u_2, \ldots, u_n) \in \mathbb{N}^n$, we would like to maximize the *likelihood function*

$$\ell_u(p) := p_1^{u_1} p_2^{u_2} \cdots p_n^{u_n}$$

on the statistical model $\mathcal{M}$. The maximizer in the model is a probability distribution $\hat{p}$ called the *maximum likelihood estimate*. When this maximizer is a local maxima, it can be found by computing the critical points of the function restricted to the regular locus of X. These are determined by solving a system of polynomials called *likelihood equations*. The *likelihood correspondence of* X is the Zariski closure of all pairs of critical points with data, i.e.,

$$\mathfrak{L}_X := \overline{\{(p,u) : p \in X_{reg} \text{ and a critical point of } \ell_u(p)\}} \subset \mathbb{P}^{n-1} \times \mathbb{P}^{n-1}.$$

This correspondence is $(n-1)$-dimensional and its projection to $\mathbb{P}^{n-1}_u$ is finite to one. The fiber of the projection are critical points of the likelihood function for that choice of data. A generic fiber has finitely many points and the its cardinality is known as the *maximum likelihood degree* (ML degree).

Example 17. *Consider the algebraic statistical model defined by* $f := (p_1 + p_2 + p_3)^3 - 30p_1p_2p_3 = 0$ *in* $\mathbb{P}^2$. *The likelihood correspondence* $\mathfrak{L}_X \subset \mathbb{P}^2 \times \mathbb{P}^2$ *is defined*

by $f = 0$ and the irreducible polynomial (1)*. For general u, we find three solutions to the equations above and conclude the maximum likelihood degree is three.*

$$\det \begin{bmatrix} \frac{u_1}{p_1} & \frac{u_2}{p_2} & \frac{u_3}{p_{13}} \\ 1 & 1 & 1 \\ \frac{\partial f}{\partial p_1} & \frac{\partial f}{\partial p_2} & \frac{\partial f}{\partial p_3} \end{bmatrix}. \tag{1}$$

Question 6. *Is the point $z := (p, u)$ in the likelihood correspondence $\mathfrak{L}_X$?*

To answer this question we define the *witness set* of the likelihood correspondence. Let $\mathcal{L}^{(0)}, \mathcal{L}^{(1)}, \ldots, \mathcal{L}^{(\dim X)}$ denote a general linear spaces where $\mathcal{L}^{(i)}$ is the intersection of i hyperplanes defined by linear forms in the indeterminants $p_1, \ldots, p_n$ with the intersection of $(n-1-i)$ hyperplanes defined by linear forms in the indeterminants $u_1, \ldots, u_n$. When $i = 0$, this is equivalent to fixing $u \in \mathbb{P}^{n-1}$. When $\mathcal{L}^{(i)}$ contains the point $z \in \mathbb{P}^{n-1} \times \mathbb{P}^{n-1}$, we denote this linear space by $\mathcal{L}_z^{(i)}$.

Definition 18. *The witness set for $\mathfrak{L}_X$ is the following formal union of sets of points:*

$$\mathfrak{W}(\mathfrak{L}_X) := \bigsqcup_{i=0}^{\dim X} \mathcal{L}^{(i)} \cap \mathfrak{L}_X.$$

The ith element of $\mathfrak{W}(\mathfrak{L}_X)$ is a set of finitely many points. When $i = \dim X$ this is the degree of X. When $i = 0$ the number of points is the ML degree.

Example 19. *The witness set for the likelihood correspondence $\mathfrak{L}_X$ from Ex. 17 consists of 6=3+3 points. The linear spaces $\mathcal{L}^{(0)}$ and $\mathcal{L}^{(1)}$ are each defined by two equations:*

$$\mathcal{L}^{(0)} : \begin{cases} 13u_1 - 3u_2 + 0u_3 = 0 \\ 0u_1 + 14u_2 - 13u_3 = 0, \end{cases} \qquad \mathcal{L}^{(0)} : \begin{cases} 1u_1 + 2u_2 - 2u_3 = 0 \\ 3p_1 - 2p_2 - p_3 = 0. \end{cases}$$

An example of such a witness set is given by $\mathfrak{W}(\mathfrak{L}_X) =$

$$\begin{aligned} &= && \mathfrak{L}_X \cap \mathcal{L}^{(0)} && \sqcup && \mathfrak{L}_X \cap \mathcal{L}^{(1)} \\ &= \left\{ \begin{array}{c} ([.46 : .28 : .26], [3 : 13 : 14]) \\ ([.22, .38, .40], [3 : 13 : 14]) \\ ([-1.35 : .10 : .25], [3 : 13 : 14]) \end{array} \right\} && \sqcup && \left\{ \begin{array}{c} ([.36 : .42 : .22], [8 : 10 : 4]) \\ ([.31 : .24 : .45], [8 : 10 : 4]) \\ ([.02 : 1.06 : -2.08], [8 : 10 : 4]) \end{array} \right\}. \end{aligned}$$

We use the witness set for $\mathfrak{L}_X$ to test membership by deforming $\mathcal{L}^{(i)}$ to $\mathcal{L}_x^{(i)}$.

Theorem 7. *Fix a point $z \in \mathbb{P}^{n-1} \times \mathbb{P}^{n-1}$. The point z is in $\mathfrak{L}_X$ if only if there exist $\mathcal{L}_z^{(i)}$ such that $\mathfrak{L}_X \cap \mathcal{L}_z^{(i)}$ is locally zero dimensional at the point z.*

This theorem (the content of Prop. 3.3 of [2]) allows us to construct a membership test for $\mathfrak{L}_X$ by taking advantage of coefficient homotopy theory.

Example 20. *We can test membership for the point $z = ([.22 : .38 : .40], [10 : 10 : 10])$ of $\mathfrak{L}_X$ from Ex. 19. Deforming $\mathcal{L}^{(0)}$ to $\mathcal{L}_z^{(0)}$ can take the real points to three complex point, none of which are z. However, if we take $\mathcal{L}^{(1)}$ to $\mathcal{L}_z^{(1)}$ then one of the three points are taken to z. Thus, we would conclude z is in $\mathfrak{L}_X$.*

References

[1] F. Catanese, S. Hoşten, A. Khetan, and B. Sturmfels. The maximum likelihood degree. *Amer. J. Math.*, 128(3):671–697, 2006.

[2] J. D. Hauenstein and J. I. Rodriguez. Multiprojective witness sets and a trace test *Arxiv.*, 1507.07069.

[3] S. Hoşten, A. Khetan, and B. Sturmfels. Solving the likelihood equations. *Found. Comput. Math.*, 5(4):389–407, 2005.

[4] J. Huh and B. Sturmfels. Likelihood geometry. In *Combinatorial algebraic geometry*, volume 2108 of *Lecture Notes in Math.*, pages 63–117. Springer, Cham, 2014.

[5] A. Leykin, J. I. Rodriguez, and F. Sottile, *Trace test*, arXiv:1608.00540.

Learning Bayesian Networks via Edge Walks on DAG Associahedra

Liam Solus

(joint work with Yuhao Wang, Caroline Uhler, and Lenka Matejovicova)

Discovering causal relations is a fundamental problem across a wide variety of research areas, including computational biology, sociology, economics, and many others [3]. A *Bayesian network*, or DAG model, is a type of graphical model based on a *directed acyclic graph (DAG)* $\mathcal{G} = ([p], A)$ with node set $[p]$ and collection of arrows A. The model associates to each node i a random variable X_i, and uses the arrows to $\mathcal{G}$ to encode the causal influences amongst the random variables $(X_1, \ldots, X_p)$. Namely, we say a joint distribution $\mathbb{P}$ satisfies the *Markov assumption* with respect to the DAG $\mathcal{G}$ if $\mathbb{P}$ satisfies the conditional independence (CI) relations $X_i \perp\!\!\!\perp X_{\mathrm{Nd}(i)\setminus\mathrm{Pa}(i)} \mid X_{\mathrm{Pa}(i)}$, where $\mathrm{Pa}(i)$ denotes the collection of *parents* of node i in $\mathcal{G}$ and $\mathrm{Nd}(i)$ denotes it collection of *nondescendants*. When a distribution $\mathbb{P}$ satisfies the Markov assumption with respect to a DAG $\mathcal{G}$ it also satisfies a (potentially) larger family of CI relations which are captured via the combinatorics of $\mathcal{G}$ by a notion of *directed separation*, known as *d-separation* [3]. A fundamental problem in causal inference is to recover a DAG $\mathcal{G}$ whose *Markov properties* (i.e. those CI relations implied by the Markov assumption) encode a collection of observed CI relations $\mathcal{C}$. Unfortunately, this problem is not well-defined since multiple DAGs can have the exact same set of d-separation statements (and therefore the exact same set of CI relations implied by the Markov assumption). If two DAGs have the same set of d-separation statements they are called *Markov equivalent* and are said to belong to the same *Markov equivalence class (MEC)*. Our fundamental problem then becomes: Given a collection of observed CI relations $\mathcal{C}$ drawn from some distribution $\mathbb{P}$ that satisfies the Markov assumption with respect to an unknown DAG $\mathcal{G}$, can we recover the MEC of $\mathcal{G}$?

An immense amount of work has been done in regards to this problem. Proposed solutions come in the form of algorithms that take in the collection of observed CI relations $\mathcal{C}$ (and possibily some other parameters) and return a DAG $\mathcal{G}$. The success of an algorithm can be measured in terms of its efficiency as well as its associated *consistency gaurantees*, i.e. those assumptions on the data-generating probability distribution that guarantee the algorithm will recover the true MEC.

Typcially, model selection algorithms are grouped into two categories: *constraint-based* and *score-based* algorithms. Constraint-based algorithms use conditional independence tests to recover a DAG $\mathcal{G}$. Such algorithms tend to perform well under the *faithfulness assumption*, i.e. the assumption that the only CI relations satisfied by the data-generating distribution are precisely those encoded via d-separation statements in the true DAG $\mathcal{G}$. Popular constraint-based algorithms such as the *PC algorithm* [3] are often used since they return only a single, clear candidate DAG. However, such methods are subject to the propagation of errors in statistical independence test results and lack any measure of confidence for the output DAG. Score-based algorithms are optimization hueristics in which each potential DAG is assigned a score and then the algorithm then attempts to select the DAG with the optimal score. One of the most popular score-based algorithms is the *Greedy Equivalence Search (GES)* which was shown to be consistent with respect to the *Bayesian Information Criterion (BIC)* in [1]. While these methods come with a natural measure of confidence in the output DAGs, the choice of maximal DAG may be ambiguous.

Many algorithms also blend features of constraint-based and score-based algorithms. In this talk, we considered a *hybrid algorithm* which we refer to as the *Greedy SP* algorithm. This algorithm is hybrid in the sense that it utilizes conditional independence tests while working to optimize a score function. Given a collection of observed CI relations $\mathcal{C}$ and a permutation $\pi = \pi_1 \dots \pi_p$ the Greedy SP algorithm assigns a DAG $\mathcal{G}_\pi := ([p], A_\pi)$ to π by asserting

$$\pi_i \to \pi_j \in A_\pi \quad \iff \quad i < j \text{ and } \pi_i \not\perp\!\!\!\perp \pi_j \mid \{\pi_1, \dots, \pi_{\max(i,j)}\} \backslash \{\pi_i, \pi_j\},$$

for all $1 \leq i < j \leq p$. The algorithm initializes at some permutation π and it considers the *covered arrows* within $\mathcal{G}_\pi$; i.e., those arrows $i \to j$ for which the collection of parent nodes of j other than i are the parents of i. Using a depth-first-search approach, Greedy SP transposes the letters of π that are the endpoints of covered arrows in $\mathcal{G}_\pi$ to produce new permutations τ. In one of these new permutations is such that $\mathcal{G}_\tau$ has strictly fewer edges that $\mathcal{G}_\pi$, then Greedy SP moves to this permutation and repeats the process. This algorithm has a very natural interpretation as an edge-walk along a subset of the edges of a certain convex polytope known as a *DAG Associahedron* [2].

In this talk, we observed that Greedy SP is consistent under the faithfulness assumption and compared it via simulated data to the PC algorithm. It turns out that Greedy SP is the first *permutation-based* DAG model selection algorithm in the literature with a consistency guarantee. This guarantee follows from the fact that if the data-generating distribution is faithful to the true (sparsest) permutation DAG $\mathcal{G}_{\pi^*}$ then every other permutation DAG $\mathcal{G}_\pi$ is an *independence map* of $\mathcal{G}_{\pi^*}$ [4]. A DAG $\mathcal{H}$ is an independence map of $\mathcal{G}$, denoted $\mathcal{G} \leq \mathcal{H}$, if every CI relation entailed by $\mathcal{H}$ is also entailed by $\mathcal{G}$. It turns out that $\mathcal{H}$ can be transformed into $\mathcal{G}$ using a sequence of covered arrow reversals and arrow deletions [1], and this fact is key to both the proof of consistency for GES with BIC and Greedy SP under faithfulness. Synthetic data studies suggest that with relatively few runs

and reasonable search depth bounds, Greedy SP generally outperforms the PC algorithm.

References

[1] D. M. Chickering. *Optimal structure identification with greedy search.* Journal of Machine Learning Research (2002): 507-554.

[2] F. Mohammadi, C. Uhler, C. Wang, and J. Yu. *Generalized permutohedra from probabilistic graphical models.* ArXiv preprint: http://arxiv.org/abs/1606.01814v1 (2016).

[3] P. Spirtes, C. N. Glymour and R. Scheines. *Causation, Prediction, and Search.* MIT Press, Cambridge, 2001.

[4] L. Solus, Y. Wang, C. Uhler, and L. Matejovicova. *Consistency guarantees for permutation-based causal inference algorithms.* ArXiv preprint: https://arxiv.org/abs/1702.03530 (2017).

Matroid Representations: Algebra and Entropy

František Matúš

A matroid (N, r) consists of a finite ground set N and rank function r, see [9].

Over a field $\mathbb{F}$, the matroid is *multilinear* of degree $\delta \geq 1$ if there exist subspaces E_i, $i \in N$, of a linear space over $\mathbb{F}$ such that $\delta \cdot r(I) = \dim E_I$, $I \subseteq N$. Here, E_I denotes the inner sum of E_i over $i \in I$. In the special case $\delta = 1$, the *linear* matroids over $\mathbb{F}$ arise.

The matroid is *algebraic* over $\mathbb{F}$ if there exist not necessarily different elements e_i, $i \in N$, of an extension field of $\mathbb{F}$ such that $r(I) = \dim_{\mathrm{tr}} \mathbb{F}(I)$ for $I \subseteq N$. Here, $\dim_{\mathrm{tr}}$ denotes the transcendence dimension over $\mathbb{F}$ and $\mathbb{F}(I)$ the smallest subfield of the extension field that contains $\mathbb{F}$ and $\{e_i \colon i \in I\}$.

The matroid is *partition representable* of the degree $d \geq 2$ if a $d^{r(N)}$-element set admits partitions π_i, $i \in N$, such that the meet-partition $\pi_I = \bigwedge_{i \in I} \pi_i$ has $d^{r(I)}$ blocks of the same size, $I \subseteq N$. A parallel language for this kind of representations is that of ideal secret sharing schemes from cryptography. The definition can be reformulated also in terms of generalized quasigroup equations [3].

A polymatroid (N, h) has a real-valued rank function h.

For random variables ξ_i, $i \in N$, that take only finitely many values, the mapping that sends $I \subseteq N$ to the Shannon entropy of $(\xi_i \colon i \in I)$ is a polymatroidal rank function. The corresponding polymatroids are *entropic*. Their rank functions exhaust the entropy region [5]. A matroid is partition representable if and only if a positive multiple of its rank function is entropic.

A polymatroid (N, g) is *almost entropic* if there exists a sequence of entropic polymatroids (N, h_n) such that $h_n \to g$ pointwise, thus if g belongs to the closure of the entropy region. This defines in particular the almost entropic matroids and their representations by infinite sequences of probability the distributions. The class of these matroids provides a description of the entropy regions [4, Theorem 5].

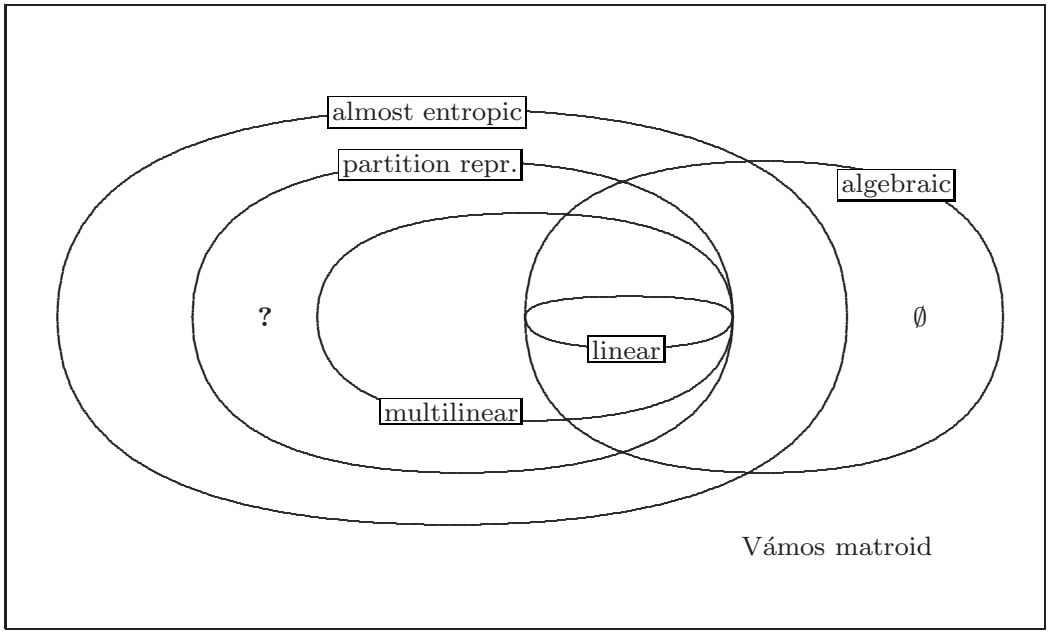


FIGURE 1. Classes of matroids.

Open problems and conjectures

1. Given a matroid, classify its partition representations of a given degree. For the graphical matroid of K_4 see [3]. Dowling geometries are studied in [8].

2. The class of partition representable matroids of a given degree has finitely many exluded minors. This is analogous to Rota's conjecture, see [1].

3. Existence of a partion representable matroid that is not multilinear, see the question mark in the figure.

4. Existence of an algebraic matroid whose dual is not algebraic.

The algebraic matroids are almost entropic by [7], which means the empty set in the figure. The class of almost entropic matroids is not closed under the duality [2]. This may give new insights to the last conjecture.

A more detailed discussion of Figure 1 is in [6].

Acknowledgement

This work was supported by Grant Agency of the Czech Republic under Grant 16-12010S

REFERENCES

[1] J. Geelen, B. Gerards and G. Whittle, *Solving Rota's conjecture.* Notices AMS **61** (2014), 736–743.

[2] T. Kaced, *The entropy region is not closed under duality.* arXiv:1611.04109 [cs.IT] (2016).

[3] F. Matúš, *Matroid representations by partitions*, Discrete Mathematics **203** (1999), 169–194.

[4] F. Matúš, *Two constructions on limits of entropy functions*, IEEE Trans. Information Theory **53** (2007), 320–330.

[5] F. Matúš and L. Csirmaz, *Entropy region and convolution.* IEEE Trans. Information Theory **62** (2016), 6007–6018.

[6] F. Matúš, *Classes of matroids closed under the minors and principal extensions*, (2017) (to appear in Combinatorica)
[7] F. Matúš, *Algebraic matroids are almost entropic.* (2017) (in preparation)
[8] F. Matúš and A. Ben-Efraim *Dowling geometries represented by partitions.* (2017) (in preparation)
[9] J.G. Oxley, *Matroid Theory.* Oxford Graduate Texts in Mathematics **21** (2011), Oxford University Press, Oxford. (Second Edition)

František Matúš is with Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Prague, Czech Republic (matus@utia.cas.cz).

Propagating Polynomial Equations

Jan Draisma

My talk concerned the following central question:

> *Given a sequence $X_1, X_2, \ldots, X_n, \ldots$ of algebraic varieties, do their equations look alike for $n \gg 0$?*

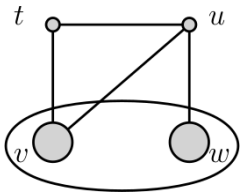
Here, an *algebraic variety* is the solution set X to a system of polynomial equations defined on a finite-dimensional vector space A.

Running example. If X_n is the variety of $n \times n$-matrices of rank at most 1 inside the space A_n of $n \times n$-matrices, then X_2 is the solution set of $x_{11}x_{22} - x_{12}x_{21} = 0$. Nonnegative, real solutions whose entries add up to 1 form the statistical model of independence. For $n \geq 2$, X_n is defined by the equations obtained from the one above by permuting rows and columns. In this manner, all equations are captured by the single equation for X_2.

Propagating. The word *propagating* in the title refers to the following easy observation: if $\pi : A_m \to A_n$ is a linear map with $\pi(X_m) \subseteq X_n$, then each equation f for X_n of degree d yields an equation $f \circ \pi$ of degree (at most) d for X_m. My talk discussed several instances where equations of finitely many of the X_n actually yield equations of *all* of them by propagation along such linear maps.

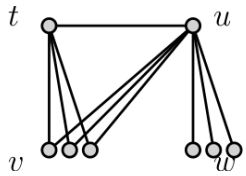
Markov random fields. Consider *Markov random fields*, or equivalently (parameterised) *undirected graphical models* for discrete random variables. The running example is the special case where the graph consists of two isolated vertices.

There are at least two different ways in which such a model is part of a sequence. First, one may increase the state space of some of the variables:



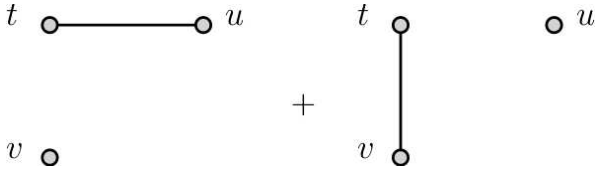
Then the associated (toric) algebraic variety is in a natural manner a *contravariant* variety over the category **FI** of finite sets with injective maps, and propagating polynomial equations along linear maps coming from the morphisms in **FI** yields equations for all varieties in the sequence under the (somewhat restrictive) condition that the variables whose state spaces are being increased form an independent set in the graph (see [HS12] and [DEKL16] for generalisations of this statement).

Second, one may increase the graph by glueing copies of a fixed graph along a fixed subgraph:



In this case the associated variety turns out to be a *covariant* variety over the category **FS** of finite sets with surjective maps, and the corresponding stabilisation result was conjectured in [RS16], verified in the case of the complete bipartite graph $K_{3,n}$ with binary states in [RS14], and proved in [DO16]. The theorem says that iterated toric fibre products [Sul07] of a finite list of Hadamard-stable varieties have ideals generated in bounded degree. The proof uses a beautiful combinatorial result from [Mac01] that monomial ideals in a finite-dimensional polynomial ring are well-partially ordered by reversed inclusion.

Mixtures. In another direction, one can take mixtures of models like the one above, and let the state spaces vary. For instance, the mixture of the following two models:



where on the left v is independent of (t, u) and on the right u is independent of (t, v), corresponds to certain tensors of *slice rank* [TS16] at most 2 in a tensor product of three vector spaces. A theorem in a forthcoming joint paper with Oosterhof is that such tensors are characterised by polynomials of degree at most 6 regardless of the sizes of the state spaces of the three variables; the degree-6 equations are certain 2×2-determinants of matrices whose entries are 3×3-determinants. Finding the equations for more complicated mixtures is a real challenge!

This model has more symmetry than just the combinatorial categories **FI** and **FS** above, and should be seen as a variety over the category **Vec** of finite-dimensional vector spaces. In the recent preprint [Dra17] a general (topological, i.e., not ring-theoretic) finiteness result is proved for such varieties.

Propagating inequalities? Results on tensors of nonnegative rank 2 [ARSZ15] and matrices of nonnegative rank 3 [KRS15] as well as on the Gaussian two-factor model [DX10] lead to the following natural question:

Do recent finiteness results for propagating polynomial equations have analogues for propagating polynomial inequalities?

This question is widely open. A positive answer might have much impact on algebraic statistics and optimisation, where models are intrinsically semi-algebraic sets over the reals rather than complex algebraic varieties.

References

[ARSZ15] Elizabeth S. Allman, John A. Rhodes, Bernd Sturmfels, and Piotr Zwiernik. Tensors of nonnegative rank two. *Linear Algebra Appl.*, 473:37–53, 2015.

[DEKL16] Jan Draisma, Rob H. Eggermont, Robert Krone, and Anton Leykin. Noetherianity for infinite-dimensional toric varieties. *Algebra & Number Theory*, 9(8):1857–1880, 2016.

[DO16] Jan Draisma and Florian M. Oosterhof. Markov random fields and iterated toric fibre products. 2016. Preprint, `arXiv:1612.06737`.

[Dra17] Jan Draisma. Topological noetherianity of polynomial functors. 2017. Preprint, `arXiv:1705.01419`.

[DX10] Mathias Drton and Han Xiao. Finiteness of small factor analysis models. *Annals of the Institute of Statistical Mathematics*, 62(4):775–783, 2010.

[HS12] Christopher J. Hillar and Seth Sullivant. Finite Gröbner bases in infinite dimensional polynomial rings and applications. *Adv. Math.*, 221:1–25, 2012.

[KRS15] Kaie Kubjas, Elina Robeva, and Bernd Sturmfels. Fixed points of the EM algorithm and nonnegative rank boundaries. *Ann. Stat.*, 43(1):422–461, 2015.

[Mac01] Diane MacLagan. Antichains of monomial ideals are finite. *Proc. Am. Math. Soc.*, 129(6):1609–1615, 2001.

[RS14] Johannes Rauh and Seth Sullivant. The Markov basis of $k_{3,N}$. 2014. Preprint, `1406.5936`.

[RS16] Johannes Rauh and Seth Sullivant. Lifting Markov bases and higher codimension toric fiber products. *J. Symb. Comp*, 74:276–307, 2016.

[Sul07] Seth Sullivant. Toric fiber products. *J. Algebra*, 316(2):560–577, 2007.

[TS16] Terence Tao and Will Sawin. Notes on the "slice rank" of tensors. 2016. `https://terrytao.wordpress.com/2016/08/24/notes-on-the-slice-rank-of-tensors/`.

Studying the Posterior Distribution of Overfitted Hidden Markov Models

Judith Rousseau

1. Introduction: asymptotic posterior distribution in hidden Markov models

Finite state space hidden Markov models are dynamical extensions of mixture models and can be represented as: conditionnally on hidden (i.e. latent or unobserved) states $x_{1:n} = (x_1, \cdots, x_n)$ the observations y_t' s are independent with distribution $[y_t|x_t = s] \sim g_{\gamma_s}$, $s \leq K$ and $\gamma_s \in \Gamma \subset \mathbb{R}^d$ and $x_{1:n}$ form a finite state space Markov chain on $\{1, \cdots, K\}$ with transition matrix Q and initial distribution μ. We write $Q = (q_{i,j}, i, j \leq K)$, the parameter is then defined as $\theta = (q_{i,j}, i \leq K, j \leq K-1, \gamma_1, \cdots, \gamma_K)$ and by Θ_K the associated parameter space. We also denote by $\mathbb{P}^n_{\theta,\mu}$ the distribution of $y_{1:n}$ associated to the parameter

θ and initial distribution μ of $x_{1:n}$ and by $\mathbb{P}_\theta$ the stationary distribution of $y_{1:n}$ associated to the parameter θ when it exists.

Under regularity conditions, if the true distribution of the observations $y_{1:n}$ is that of a hidden Markov model as described above with transition matrix Q^* on $\{1, \cdots, K\}$ which is ergodic and irreducible and if the true emission parameters $\gamma_1^*, \cdots, \gamma_K^*$ are all distinct, then the maximum likelihood estimator $\hat{\theta}$, is consistent and asymptotically Gaussian and efficient. Similarly for any Bayesian approach based on a prior distribution on $(Q, \gamma_1, \cdots, \gamma_K)$ absolutely continuous with respect to the Lebesgue measure on $\mathcal{S}_K^K \times \Gamma^K$, where $\mathcal{S}_K$ is the K dimensional simplex, with positive and continuous density at $Q^*, \gamma_1^*, \cdots, \gamma_K^*$, then the so-called Bernstein von Mises theorem holds. This means that the posterior distribution of $\sqrt{n}(\theta - \hat{\theta})$ converges to a Gaussian distribution with mean 0 and variance $I^{-1}(\theta^*)$ under $\mathbb{P}_{\theta^*}$, where $I^{-1}(\theta^*)$ is the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta^*)$ under $\mathbb{P}_{\theta^*}^n$.

In this talk we are interested in the behaviour of the posterior distribution when the true parameter θ^* corresponds to a Markov chain living on a subset of $\{1, \cdots, K\}$, in other words when there exists $K^* < K$ and an ergodic and irreducible transition matrix Q^* on $\{1, \cdots, K^*\}$ and $\gamma_1^*, \cdots, \gamma_{K^*}^*$ all distinct such that $[y_t | x_t = s] \sim g_{\gamma_s^*}$ and $x_{1:n}$ form a Markov chain on $\{1, \cdots, K^*\}$ with transition matrix Q^*.

In this case the model defined on Θ_K has a singularity at θ^* and is not identifiable. The non-identifiability can be observed by noting that for all n, $P_{\theta^*}^n$ can be represented on Θ_K by either merging extra components or by emptying extra components. As an illustration of the merging configuration, consider $(Q, \gamma_1, \cdots, \gamma_K)$ with $\gamma_{K^*}^* = \gamma_{K^*+1} = \cdots = \gamma_K$ and $Q = (q_{i,j}, i, j \leq K)$ satisfying for all $i \leq K^*$,

$$q_{i,j} = q_{i,j}^*, \quad j \leq K^* - 1; \quad \sum_{j=K^*}^{K} q_{i,j} = q_{i,K^*}^*$$

and for all $i \geq K^*$,

$$q_{i,j} = q^* K^*, j, \quad j \leq K^* - 1; \quad \sum_{j=K^*}^{K} q_{i,j} = q_{K^*,K^*}^*.$$

The emptying of the extra component is verified for instance by parameters defined as $\gamma_j = \gamma_j^*$ for all $j \leq K^*$, and Q defined by

$$q_{i,j} = q_{i,j}^*, i, j \leq K^*, \quad q_{i,j} = 0 \, \forall i \leq K, j > K^*, \quad q_{i,j} = q_{K^*,j} \quad \forall j \leq K^*.$$

In this case, if it can be proved that the maximum likelihood estimator $\hat{\theta}$ can converge to the set $\tilde{\Theta}^*$ of parameter values $\theta \in \Theta_K$ such that $\mathbb{P}_\theta^n = \mathbb{P}_{\theta^*}^n$ for all n, no other more precise statements has been obtained for such an estimator.

On the other hand, the same kind of problems has been studied in the context of static mixtures, i.e. when the unobserved states $x_{i:n}$ are independent and identically distributed with distribution $\mathbf{p} = (p_1, \cdots, p_K)$. Similarly, the model is non indentifiable and singular at parameter values corresponding to $K^* < K$ states latent variables, with merging of extra states represented by, for instance $\gamma_j = \gamma_j^*$,

$j \leq K^*$ and $\gamma_{K^*+1} = \cdots = \gamma_K = \gamma^*_{K^*}$ and $p_j = p^*_j$ for $j \leq K^* - 1$ and emptying the extra states by having $p_{K^*+1} = \cdots = p_K = 0$ while $\gamma_j = \gamma^*_j$ for $j \leq K^*$. Here again the maximum likelihood converges to $\tilde{\Theta}^*$ and not much else can be said on this estimator. However in [3], it was proved that, by choosing correctly the prior distribution on the parameter $\mathbf{p}$, the posterior distribution has asymptotically a more stable behaviour.

More precisely if $\mathbf{p} \sim \mathcal{D}(\alpha_1, \cdots, \alpha_K)$ (a Dirichlet distribution) and if the γ_j's are independent and identically distributed with distribution π_γ which has positive and continuous density at γ^*_j, $j \leq K^*$, then

- If $\max(\alpha_1, \cdots, \alpha_K) < d/2$ then

$$\Pi\left(\sum_{j>K^*} p_j > M(\log n)^q/\sqrt{n} | y_{1:n}\right) = o_{p_{\theta^*}}(1)$$

- If $\min(\alpha_1, \cdots, \alpha_K) > d/2$ then

$$\Pi\left(\sum_{j>K^*} p_j < (\log n)^{-q} | y_{1:n}\right) = o_{p_{\theta^*}}(1)$$

where in the above equalities $\Pi(\cdot|x_{1:n})$ denotes the posterior distribution and to simplify notation $\sum_{j>K^*} p_j$ is associated to the permutation σ^* of the labels which minimizes $\sum_{j>K^*} p_{\sigma(j)}$ over all permutations. The above results thus state that if the hyperparameters of the Dirichlet prior satisfy $\max \alpha_j < d/2$ then the extra components have asymptotically probability going to 0 (emptying of the extra states) under the posterior distribution while in the second case the posterior distribution concentrates to configurations of $\tilde{\Theta}^*$ corresponding to the merging of the extra states.

One of the consequences of the above result is that the log marginal likelihood is bounded from above by the following singular BIC approximation

$$\log m_n(y_{1:n}) = \log\left(\int_{\Theta_K} f^n_\theta(y_{1:n}) d\pi(\theta)\right) = \ell_n(\hat{\theta}_n) - D(K^*, K)/2 + O_P(\log\log n)$$

when $\alpha = \alpha_1 = \cdots = \alpha_K < d/2$, where $\ell_n(\theta)$ is the log - likelihood and $D(K^*, K) = (K^* d + K^* - 1 + \alpha(K - K^*))$. Interestingly this result is obtained via a technic quite different from the algebraic statistical approach of [5], see also [1].

The question is : can we extend this phase - transition result to the case of hidden Markov models ?

2. Case of hidden Markov models

The case of hidden Markov models is more complex and so far has not been treated using algebraic methods. Using the technic of proof as in [3], [2] and [4] study the asymptotic behaviour of the posterior distribution in hidden Markov models, when

the true number of states K^* is smaller than K the assumed number of states on model Θ_K. Their aim is to understand if

$$\Pi\left(\sum_{j>K^*}\mu_Q(j) > u_n|y_{1:n}\right) = o_p(1) \tag{1}$$

for some sequence $u_n = o(1)$, where μ_Q is the stationary distribution associated to the transition matrix Q. In [2] and [4] only a partial caracterisation of the behaviour of the posterior distribution has been derived, which we now briefly recall.

Consider a prior distribution on θ in the form

$$\forall i \leq K, \quad q_{i.} = (q_{i,j}, j \leq K) \overset{ind}{\sim} \mathcal{D}(\alpha_{i1}, \cdots, \alpha_{i,K}), \quad \gamma_j \overset{iid}{\sim} \pi_\gamma.$$

In [4] the authors show that if $\alpha_{i1} = \cdots = \alpha_{ip} = \bar{\alpha}$ and $\alpha_{ip+1} = \cdots = \alpha_{iK} = \underline{\alpha}$ with $\bar{\alpha} > A(K, K^*, d)$ and $\underline{\alpha} < a(K, K^*, d)$ where $A(K, K^*, d)$ is a constant rapidly growing with K and $a(K, K^*, d)$ is a small constant decreasing with K, then (1) is valid. In the case where $K = 2$ and K^*, [2] provide a sharper result where under some condition (1) is not valid.

The difficulty in handling the hidden Markov models, as it is done in [2] and [4] is that first a concentration result in terms of the L_1 norm, for the stationary density of two consecutive observations

$$f_\theta(y_1, y_2) = \sum_{i_1,i_2=1}^{K} \mu_Q(i_1) q_{i_1,i_2} g_{\gamma_{i_1}}(y_1) g_{\gamma_{i_2}}(y_2)$$

must be obtained. However in any neighbourhood of Q^* there exist transition matrices Q which are non ergodic or reducible and the likelihood does not have a well understood behaviour near such values of the parameter. This first induces a posterior concentration rate on f_{θ^*} which is much larger than the usual $1/\sqrt{n}$ and second a constraint on the hyper parameters α_{ij} which ensures that the prior penalizes enough neighbourhoods of non ergodic Markov chains.

As a consequence $A(K, K^*, d)$ and $a(K, K^*, d)$ are probably not sharp cutting points to descriminate between an emptying of the extra components of the Markov chain and a merging of these components. A simulation study we have run indicates that (1) seems to hold for $\bar{\alpha}$ much smaller than $A(K, K^*, d)$.

My question is then : can the technics of algebraic statistics tackle this problem?

References

[1] M. Drton and M. Plummer (2017). A Bayesian information criterion for singular models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, to appear.

[2] E. Gassiat and J. Rousseau (2014) : About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli*, **20**, 2039–2075.

[3] J. Rousseau and K. Mengersen (2011) Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **73**, 689–710.

[4] Z. van Havre, J. Rousseau, N. White and K. Mengersen (2015) Overfitting hidden Markov models with an unknown number of states. *Preprint*
[5] S. Watanabe *Algebraic geometry and statistical learning theory*, **25** , Cambridge Monographs on Applied and Computational Mathematics. (2009). Cambridge University Press.

Optimal Experimental Design that Minimizes the Width of Simultaneous Confidence Bands

Satoshi Kuriki
(joint work with Henry P. Wynn)

We propose an optimal experimental design for a curvilinear regression model that minimizes the band-width of simultaneous confidence bands. Simultaneous confidence bands for nonlinear regression are constructed by evaluating the volume of a tube about a curve that is defined as a trajectory of a regression basis vector [3, 6, 8]. This methodology is referred to as the volume-of-tube method, which has been developed for the approximation of the tail probability of the maximum of a smooth Gaussian random field [1, 4, 9, 10]. The proposed experimental design criterion is constructed based on the volume of a tube, and the corresponding optimal experimental design is referred to as the minimum-volume optimal design.

For Fourier and weighted polynomial regressions with polynomial variance function, the problem is formalized as one of minimization over the cone of Hankel positive definite matrices, and the criterion to minimize is expressed as an elliptic integral. We show that the Möbius group keeps our problem invariant, and hence, minimization can be conducted over cross-sections of orbits. The Möbius group action on polynomial variance function defined here is the same group action on the Cauchy distribution introduced by [7].

We demonstrate that for the weighted polynomial regression and the Fourier regression with three bases, the minimum-volume optimal design forms an orbit of the Möbius group containing D-optimal designs as representative elements. We also characterize the D-optimal design for weighted polynomial design in terms of group action (cf. [2]).

The case where the number of explanatory variables is more than one is remaining as a future research topic. The group invariance under the multivariate Möbius group is proved again, and the same properties as the univariate case are expected to hold.

This talk is based on [5].

References

[1] R. J. Adler and J. E. Taylor, *Random Fields and Geometry*, Springer, 2007.
[2] H. Dette, L. M. Haines, and L. Imhof, *Optimal designs for rational models and weighted polynomial regression*, The Annals of Statistics, **27** (4) (1999), 1272–1293.
[3] S. Johansen and I. M. Johnstone, *Hotelling's theorem on the volume of tubes: Some illustrations in simultaneous inference and data analysis*, The Annals of Statistics, **18** (2) (1990), 652–684.

[4] S. Kuriki and A. Takemura, *Tail probabilities of the maxima of multilinear forms and their applications*, The Annals of Statistics, **29** (2) (2001), 328–371.
[5] S. Kuriki and H. P. Wynn, Optimal experimental design that minimizes the width of simultaneous confidence bands, arXiv:1704.03995 [math.ST], 2017.
[6] X. Lu and S. Kuriki, *Simultaneous confidence bands for contrasts between several nonlinear regression curves*, Journal of Multivariate Analysis, **155** (2017), 83–104.
[7] P. McCullagh, *Möbius transformation and Cauchy parameter estimation*, The Annals of Statistics, **24** (2) (1996), 787–808.
[8] D. Q. Naiman, *Conservative confidence bands in curvilinear regression*, The Annals of Statistics, **14** (3) (1986), 896–906.
[9] J. Sun, *Tail probabilities of the maxima of Gaussian random fields*, The Annals of Probability, **21** (1) (1993), 34–71.
[10] A. Takemura and S. Kuriki, *On the equivalence of the tube and Euler characteristic methods for the distribution of the maximum of Gaussian fields over piecewise smooth domains*, The Annals of Applied Probability, **12** (2) (2002), 768–796.

Introduction to Normaliz

Tim Römer

Normaliz [4] is an open source tool developed in Osnabrück by the Normaliz Team. It implements algorithms especially for the computation of lattice points in rational polyhedra. Seen from the point of view of algebra it develops algorithms to solve linear diophantine systems. Here a polyhedron and a lattice can either be defined by generators (extreme rays of cones, vertices of polyhedra, generators of the lattice), or by constraints (inequalities, equations, congruences). Note that the conversion between generators and constraints is already an important part of Normaliz.

The current version of Normaliz is 3.2.0. It is implemented in C++. It uses GMP for infinite precision arithmetic and OpenMP for parallelization. For the subdivision of "large" simplicial cones it can use (optional) the IP solver SCIP [7]. Normaliz offers the API *libnormaliz*. Normaliz has interfaces to CoCoA, GAP, Macaulay 2, Python and Singular. It is used, e.g., by polymake [8], Regina [6] or SecDec-3.0 [3].

For the mathematical background of Normaliz we refer to its documentation which is available at its homepage; see [4]. Its main computation goals are:

- convex hulls and dual cones,
- conversion from generators to constraints and vice versa,
- triangulations, disjoint decompositions and Stanley decompositions,
- Hilbert basis of rational (not necessarily pointed) cones,
- normalization of affine monoids,
- lattice points of rational polytopes and (unbounded) polyhedra,
- Hilbert series and (quasi) polynomials under $\mathbb{Z}$-gradings,
- generalized Ehrhart series and Lebesgue integrals of polynomials over rational polytopes via NmzIntegrate.

These goals can be selected via command line options or in the input file. Some of the tools of Normaliz are:

- linear algebra over $\mathbb{Z}$,
- Fourier-Motzkin elimination,
- pyramid decomposition and triangulation,
- evaluation of simplicial cones,
- reduction of a system of generators to the Hilbert basis,
- Stanley decomposition (for Hilbert series),
- a variant of Pottier's "dual" algorithm for Hilbert bases.

Also the algorithmic variants can be selected via command line options.

For some benchmarks we refer to [2], [5] and [9]. Normaliz was cited more than one hundred times in the literature. For example, it was used in algebraic statistics by Sturmfels–Welker [11] to study linear ordering polytopes, which can be seen as the model polytopes of the associated toric statistical models. Another example of a very interesting application is its use in [10] to compute the set of "holes" of a semigroup.

References

[1] J. Abbott, A. M. Bigatti and G. Lagorio, *CoCoA-5: a system for doing Computations in Commutative Algebra*. Available at `http://cocoa.dima.unige.it`.

[2] B. Assarf et al., *Computing convex hulls and counting integer points with polymake*. Math. Program. Comput. **9** (2017), no. 1, 1–38.

[3] S. Borowka et al., *SecDec – A program to evaluate dimensionally regulated parameter integrals numerically*. Available at `https://secdec.hepforge.org`.

[4] W. Bruns, B. Ichim, T. Römer, R. Sieg and C. Söger, *Normaliz. Algorithms for rational cones and affine monoids*. Available at `https://www.normaliz.uni-osnabrueck.de`.

[5] W. Bruns, B. Ichim and C. Sger, *The power of pyramid decomposition in Normaliz*. J. Symbolic Comput. **74** (2016), 513–536.

[6] B. A. Burton, *Regina: Software for low-dimensional topology*. Available at `http://regina.sourceforge.net`.

[7] Gamrath et al., *The SCIP Optimization Suite 3.2*. ZIB-Report **15–60** (2016). Available at `http://scip.zib.de`.

[8] M. Joswig, B. Mller and A. Paffenholz, `polymake` *and lattice polytopes*. In: 21st International Conference on Formal Power Series and Algebraic Combinatorics (FPSAC 2009), 491–502, Discrete Math. Theor. Comput. Sci. Proc., AK, Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2009.

[9] M. Kppe and Y. Zhou, *New computer-based search strategies for extreme functions of the Gomory–Johnson infinite group problem*. See arXiv:1506.00017.

[10] F. Kohl, Y. Li, J. Rauh and R. Yoshida, *Semigroups – A Computational Approach*. See arXiv:1608.03297.

[11] B. Sturmfels and V. Welker, *Commutative algebra of statistical ranking*. J. Algebra **361** (2012), 264–286.

Identification of Linear Dynamic Systems: Structure Theory and its Relation to Estimation

Manfred Deistler

We are concerned with the problem of finding a linear dynamic model from (discrete, equidistant) time series data. The linear dynamic models considered are either ARMA models $(a(z), b(z))$, $a(z) = \sum_{j=0}^{p} a_j z^j$, $b(z) = \sum_{j=0}^{q} b_j z^j$; $a_j, b_j \in$

$\mathbb{R}^{s\times s}$, $z \in \mathbb{C}$ or state space models (A, B, C), $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times s}$, $C \in \mathbb{R}^{s\times n}$ driven by unobserved white noise (ε_t) with $\Sigma = \mathbb{E}\varepsilon_t\varepsilon_t' > 0$.

In both cases the transferfunctions

$$k(z) = a^{-1}(z)b(z) = I + C(I - Az)^{-1}zB$$

are rational. A standard assumption is that $k(z)$ has no poles or zeros for $|z| \leq 1$. We then say that $k(z)$ is causal and miniphase. Then the corresponding solution

$$y_t = k(z)\varepsilon_t = \sum_{j=0}^{\infty} k_j\varepsilon_{t-j}$$

gives a stationary output process and corresponds to the Wold decomposition.

The spectral density of (y_t) is of the form

$$f(z) = (2\pi)^{-1}k(z)\,\Sigma\,k(\overline{z})'.$$

If we assume $k(0) = I$, then $k(z)$ is uniquely determined from $f(z)$ and thus from the population second moments of (y_t). The second moments $f(z)$ (and thus, in a certain sense the transfer functions $k(z)$) represent the external behaviour of the system, whereas the ARMA and state space parameters describe the internal characteristics. Every rational, causal and miniphase transfer function can be realized by an ARMA or state space system. For simplicity, from now on, we restrict ourselves to ARMA systems here.

If we have no "structural" a-priori restrictions, we have two ("first") model classes, namely:

- U_A: the set of all rational, causal and miniphase $s \times s$ transfer functions with $k(0) = I$
- T_A: the set of all ARMA systems (a, b) where a, b are $s \times s$ polynomial matrices satisfying additional conditions, such as left coprimeness, stability and the miniphase assumption

and the mapping

$$\pi : T_A \to U_A : \quad k(z) = a^{-1}(z)b(z).$$

The problems here are as follows:

- T_A is not finite-dimensional
- π is (surjective but) not injective, i.e. we lack identifiability
- There exists no continuous selection $U_A \to T_A$, when U_A is endowed with the so called pointwise topology T_{pt}.

For this reason a "semi non-parametric approach" is taken: In a first step, the model classes are broken into pieces to obtain the final model classes:

This is done as follows:

- First we define $M(n) \subset U_A$, the set of all transfer functions of order n. $M(n)$ is a real analytic manifold of dimension $2ns$, which, in the multivariable case, cannot be described by one chart.

- In the next step, suitable charts U_α, described by a multiindex α, obtained from the block Hankel matrix of the transfer function are chosen. This leads to a parametrization

$$\psi_\alpha : U_\alpha \to T_\alpha : \quad \psi_\alpha \left(\pi/_{T_\alpha}(\tau)\right) = \tau$$

  where τ are the ARMA parameters.

We have $\overline{U}_\alpha = \overline{M(n)} = \bigcup_{i \le n} M(i)$, where the bar denotes the closure, and ψ_α is a homeomorphism.

This leads to a model selection problem, n and α have to be estimated from data, n e.g. by AIC or BIC.

Given the selected n, α, in a second step, we use ML-type procedures to estimate $\theta = \begin{pmatrix} \tau \\ \text{vech}\Sigma \end{pmatrix}$: $-2T^{-1}$ log of the Gaussian Likelihood is of the form

$$L_T(\theta) = T^{-1} \log\det \Gamma_T(\theta) + \begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix}' \Gamma_T(\theta)^{-1} \begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix},$$

where T is sample size and $\Gamma_T(\theta)$ is the $Ts \times Ts$ covariances matrix corresponding to the ARMA parameters θ.

A coordinate free consistency result has been given by Hannan (see [1]): We have

$$\hat{k}_T(z) \to k_0(z) \qquad (\text{in } T_{pt} \text{ a.s.})$$
$$\hat{\Sigma}_T \to \Sigma_0 \qquad \text{a.s.}$$

where $\hat{k}_T$, $\hat{\Sigma}_T$ denote the maximum likelihood estimators and k_0, Σ_0 the true values. Then, the continuity of ψ_α implies

$$\hat{\theta}_T = \psi_\alpha \left(\hat{k}_T(z)\right) \to \theta_0 = \psi_\alpha(k_0) \quad \text{a.s.}$$

Note that:

- $\overline{M(n)} \supseteq \pi(\overline{T}_\alpha)$
- U_α is open in $\overline{M(n)}$
- $\overline{T}_\alpha - T_\alpha$ contains equivalence classes corresponding to lower dimensional systems
- $\overline{M(n)} - \pi(\overline{T}_\alpha)$ corresponds to the point of infinity in the parameter space

References

[1] E.J. Hannan and M. Deistler, "The Statistical Theory of Linear Systems". Wiley, New York, 1988, Reprint in SIAM Classics in Applied Mathematics, Philadelphia, 2012.

[2] M. Deistler, "System Identification - General Aspects and Structure". In: G. Goodwin (ed) "System Identification and Adaptive Control" (Festschrift for B.D.O. Anderson), Springer, London, 3 – 26, 2001.

[3] M. Deistler, "A Birds Eye View on System Identification". In: A. Chiuso, A. Ferrante and S. Pinzoni (eds): Festschrift for G. Picci, Lecture Notes in Information and Control Sciences, Vol 364, Springer, 59-71, 2007.

Species Tree Identifiability from Split Probabilities

Elizabeth S. Allman
(joint work with James H. Degnan, John A. Rhodes)

It is well known that gene trees and species trees may differ. A gene tree constructed from a molecular alignment describes the relationship between those particular sampled gene lineages. In contrast, a species tree describes the relationships between the species from which individuals might be sampled to construct a gene tree. Typically, the species tree is the parameter of interest, but one obtains information about it only indirectly from gene trees.

There are many reasons for discordance between gene trees and species trees, including lateral gene transfer, hybridization, and *incomplete lineage sorting*. The phenomenon of incomplete lineage sorting occurs because the merging of two sampled gene lineages back in the past can predate the divergence of species. For example, in the figure below we see that the gene tree $((((C,D),A),B),E)$ can form in the species tree $((((a,b),c),d),e)$. Here, the species tree is denoted with thick lines as a 'fat' tree; the gene tree within the species tree is shown as a stick tree.

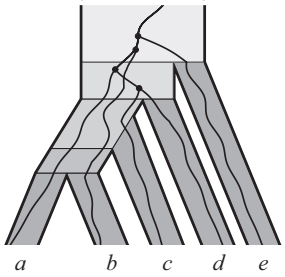


Figure: A gene tree can have a topology different from the species tree because of the phenomenon of Incomplete Lineage Sorting.

The *multispecies coalescent* models the population genetic effect of incomplete lineage sorting. The parameters of this model are the species tree topology, and internal branch lengths. Probabilities (more accurately, densities) of rooted, metric gene trees can be computed under this model for all $(2n-3)!!$ factorial possible rooted topologies, where n is the number of taxa. In work of Degnan and Salter [2], probabilities for the rooted, *topological* gene tree distribution were computed for the general case of n taxa.

Inference of species trees from many gene alignments or large multi-locus datasets is a challenging, but important, problem. Because the topological gene tree distribution is so large, with $(2n-3)!!$ non-zero entries under the multispecies coalescent model, one might investigate summaries of the distribution that retain enough information for accurate and fast inference. We investigate the collection of *split probabilities*. If $\mathcal{X}$ is the collection of taxa at the tips of a species tree σ, the leaf labels, then for a subset $\emptyset \neq \mathcal{A} \subset \mathcal{X}$ the split $\mathcal{A} \mid \mathcal{X} \setminus \mathcal{A}$ has a positive probability under the multispecies coalescent of occurring on any topological gene tree. This probability, $\mathbb{P}_\sigma(Sp(\mathcal{A} \mid \mathcal{X} \setminus \mathcal{A})$, can be computed by summing the

probabilities of all rooted gene tree topologies which display the split $\mathcal{A} \mid \mathcal{X} \setminus \mathcal{A}$. Such a summary has appeal to practitioners, since estimating split probabilities from a collection of estimated gene trees is fast and easy, and there are likely to be many fewer sampling zeros.

We prove a number of results about the collection of split probabilities [1]. First, we note that the collection of split probabilities is not a distribution; instead the sum of all such non-trivial split probabilities is $n-3$, the number of internal edges in the species tree. We then investigate the question of species tree topology *identifiability* under the multispecies coalescent model, showing that the unrooted species tree topology is indeed identifiable from split probabilities. This is accomplished by computing *split invariants*, polynomial equations that must hold for any collection of split probabilities generated under the multispecies coalescent model. As an interesting extension we prove that using linear inequalities —that is, ideas from semi-algebraic geometry— the root of the species tree is also identifiable from split probabilities for all trees except an exceptional one, the fully balanced 6-taxon tree.

As a final result, we show that the heuristic distance-based method of species tree estimation introduced by Liu [3] called NJ_{st} is in fact a model-based method. More precisely, the NJ_{st} estimate is a consistent estimator of the species tree topology under the multispecies coalescent model, and in fact only uses the summary split probabilities for its estimation.

There are many, many open questions for interested researchers to investigate. Can branch lengths in a species tree σ be identified from split probabilities? Is there an improved NJ_{st} algorithm that can estimate the root on the species tree parameter? Are there higher degree invariants?

References

[1] E. Allman, J. Degnan, and J. Rhodes, *Split probabilities and species tree inference under the multispecies coalescent model*, (2017), preprint.
[2] J. H. Degnan and L. A. Salter. *Gene tree distributions under the coalescent process*, Evolution, **59**(1) (2005), 24–37.
[3] L. Liu and L. Yu, *Estimating species trees from unrooted gene trees*, Syst. Biol, **60** (2011), 661–667.

Rank One Tensor Completion

Mario Kummer
(joint work with T. Kahle, K. Kubjas, Z. Rosen)

The talk is based on the article [KKKR17].

Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$ denote either the field of real or complex numbers. We consider a tensor $T \in \mathbb{K}^{d_1} \otimes \cdots \otimes \mathbb{K}^{d_n}$ of which we only know some entries, indexed by a subset $E \subseteq D = [d_1] \times \cdots \times [d_n]$ where $[d_i] = \{1, \ldots, d_i\}$. Based on this limited knowledge we want to decide whether T could possible be of rank one, i.e. expressable as $T = \theta_1 \otimes \cdots \otimes \theta_n$ for some $\theta_i \in \mathbb{K}^{d_i}$. This question is known as the problem of

rank one tensor completion. Wether or not such a partial tensor is completable depends on $\mathbb{K}$ as the following example that goes back to Kruskal [Kru89] shows.

Example. *Let* $n = 3$, $d_1 = d_2 = d_3 = 2$ *and*

$$E = \{(1,1,2),(1,2,1),(2,1,1),(2,2,2)\}.$$

The partial tensor given by

$$x_{112} = 1, x_{121} = 1, x_{211} = 1, x_{222} = -1$$

is completable over $\mathbb{C}$ *but not over* $\mathbb{R}$.

Since the set of completable tensors is parametrized by monomials, checking whether a partial tensor is completable over the complex numbers can be done using the theory of toric varieties. For example in [ES96, Prop. 8.7] an explicit set of equations is given.

For a partial tensor to be completable over the real numbers it is of course necessary that it is completable over the complex numbers. We give a criterion on E which allows us to decide whether every partial tensor that is completable over $\mathbb{C}$ is also completable over $\mathbb{R}$. Moreover, we show that whether a partial tensor is completable over $\mathbb{R}$ only depends on the signs of the observed entries — given that it is completable over $\mathbb{C}$.

Finally, we consider the problem of deciding whether T is the joint distribution of independent discrete random variables. This is equivalent to T being expressable as $T = \theta_1 \otimes \cdots \otimes \theta_n$ where the entries of each θ_i are nonnegative and sum up to one. In this case the set of completable partial tensors can not be described by equations and sign distributions anymore. We focus on the case where this set is full-dimensional and where $|E| = \sum_{i=1}^{n}(d_i - 1)$. We describe the algebraic boundary of the completable region, i.e. the Zariski closure of its Euclidean boundary. Apart from the coordinate hyperplanes this consists of a single irreducible hypersurface H. We describe this hypersurface as the image of a linear subspace of $\mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_n}$ under the monomial map parametrizing the completable tensors. The equations of this linear subspace can be explicitly determined from E.

We end with two questions which we consider to be decidable within a reasonable amount of effort. The first one should be approachable using toric or tropical geometry.

Problem. *What is the degree of the hypersurface* H*?*

The second question turned out to be true in all examples that we computed. It is also true in the case when $d_1 = \cdots = d_n = d$ and E is the set of all diagonal indices, i.e. those of the form $(i, \ldots, i)$ for $i = 1, \ldots, d$ [KR16].

Problem. *Is the set of all partial tensors with nonnegative entries that are not completable to a tensor corresponding to the joint distribution of independent random variables always a convex set?*

References

[ES96] David Eisenbud and Bernd Sturmfels. Binomial ideals. *Duke Math. J.*, 84(1):1–45, 1996.

[KKKR17] Thomas Kahle, Kaie Kubjas, Mario Kummer, and Zvi Rosen. The Geometry of Rank-One Tensor Completion. *SIAM J. Appl. Algebra Geom.*, 1(1):200–221, 2017.

[KR16] Kaie Kubjas and Zvi Rosen. Matrix completion for the independence model. *Journal of Algebraic Statistics (to appear), arXiv:1407.3254*, 2016.

[Kru89] J. B. Kruskal. Rank, decomposition, and uniqueness for 3-way and N-way arrays. In *Multiway data analysis (Rome, 1988)*, pages 7–18. North-Holland, Amsterdam, 1989.

Attempts to Characterize Extreme Supermodular Functions

Milan Studený

Supermodular functions have been investigated in various branches of mathematics, in particular in connection with cooperative games, theory of impresise probabilities and conditional independence structures. Submodular functions, their mirror images, were studied in matroid theory and combinatorial optimization.

Definition A set function $m : \mathcal{P}(N) \to \mathbb{R}$, where $\mathcal{P}(N) := \{A : \ A \subseteq N\}$ denotes the power set of a finite non-empty set N, is *supermodular* if it satisfies inequalities

$$\forall\, C, D \subseteq N \qquad m(C) + m(D) \ \leq \ m(C \cup D) + m(C \cap D)\,.$$

A function m is called *ℓ-standardized* if $m(S) = 0$ for any $S \subseteq N$, $|S| \leq 1$. An ℓ-standardized supermodular function will be called *extreme* if it generates an extreme ray of the cone of ℓ-standardized supermodular functions. □

Extreme supermodular functions play an important role in the context of testing *conditional independence implications* [1]. This motivated a long-term plan to get a *complete characterization* of extreme supermodular functions, by which is meant an enumeration procedure generating, for any given $n = |N|$, all extreme rays of the supermodular cone. This seems to be a quite ambitious plan because the number of extreme rays grows rapidly with $n = |N|$; their number for $n = 5$ is 117978 and decompose into 1319 permutation types [7]. Nevertheless, some partial results have been achieved. One of them is a simple linear-algebraic criterion for testing the extremity [9] based on a game-theoretical concept of a core polytope.

Definition Given a supermodular function m with $m(\emptyset) = 0$, its *core* is a bounded polyhedron (= polytope) in $\mathbb{R}^N$ defined by

$$C(m) \ := \ \{\, x \in \mathbb{R}^N \ : \ \sum_{i \in N} x_i = m(N) \ \&\ \forall\, S \subseteq N \quad \sum_{i \in S} x_i \geq m(S) \,\}\,.$$

Every vertex of the core $v = [v_i]_{i \in N} \in \mathsf{ext}\, C(m)$ can be assigned the respective *tightness class* $\mathcal{T}_v^m = \left\{S \subseteq N : \ m(S) = \sum_{i \in S} v_i\right\}$ of sets. The (combinatorial) *core structure* of m is then the collection of classes of subsets of N, namely $\{\, \mathcal{T}_v^m \ : \ v \in \mathsf{ext}\, C(m) \,\}$; see [3, § 2] where this concept was introduced. □

The criterion from [9] ascribes a simple linear equation system to the core structure of (an ℓ-standardized supermodular function) m and the function m is

shown to be extreme iff the solution to the equation system is unique up to a real multiple. The criterion is easy to implement on a computer.

The cores of supermodular functions coincide with the polytopes known as *generalized permutohedra* [6]. These polytopes also occurred recently in the context of algebraic statistics because they can be used to describe alternatively some of probabilistic graphical models [4].

The characterization problem mentioned above leads to the study of

(i): the face-lattice of the cone of supermodular functions (with usual inclusion ordering)

because extreme supermodular functions corresponds to its atoms. This lattice is, moreover, isomorphic to two other lattices, namely,

(ii): the lattice of equivalence classes of *generalized permutohedra*, pre-ordered by the relation $P \preceq Q$ iff P is a *Minkowski summand* of Q,

(iii): the lattice of *normal fans* coarsening the permutation fan (= braid arrangement fan) ordered by the relation $\mathcal{N} \preceq \mathcal{M}$ iff a fan $\mathcal{N}$ coarsens a fan $\mathcal{M}$. These fans correspond to *submodular rank tests* [5].

On the top of that, the above face-lattice is also anti-isomorphic to

(iv): the lattice of structural independence models (with independence-inclusion ordering) [8], and to

(v): the lattice of combinatorial *core structures* induced by supermodular functions, ordered by refinement relation: a core structure $\mathfrak{C}$ *refines* a core structure $\mathfrak{D}$ iff $\forall \mathcal{C} \in \mathfrak{C}\ \exists \mathcal{D} \in \mathfrak{D}$ with $\mathcal{C} \subseteq \mathcal{D}$.

Note that the refinement relation for combinatorial core structures is analogous to the concept of a refinement from [2, Definition 2.3.8], used to describe relations between different polyhedral subdivisions of a polytope.

The open questions/tasks I plan to deal with in near future concern the concept of a (combinatorial) core structure. The nearest goal is to try to characterize those collections of classes of subsets of N which are core structures for supermodular functions. Further research theme/topic is developing methods for easy generating extreme supermodular functions.

Acknowledgements. The work on this topic has been supported from the GAČR project n. 16-12010S.

References

[1] R. Bouckaert, R. Hemmecke, S. Lindner, M. Studený. Efficient algorithms for conditional independence inference. *Journal of Machine Learning Research* 11 (2010) 3453–3479.

[2] J.A. De Loera, J. Rambau, F. Santos. *Triangulations.* Algorithms and Computation in Mathematics 25, Springer, Berlin 2010.

[3] J. Kuipers, D. Vermeulen, M. Voorneveld. A generalization of the Shapley-Ichiishi result. *International Journal of Game Theory* 39 (2010) 585–602.

[4] F. Mohammadi, C. Uhler, C. Wang, J. Yu. Generalized permutohedra from probabilistic graphical models. A manuscript from June 2016, available at arXiv:1606.0814 .

[5] J. Morton, L. Pachter, A. Shiu, B. Sturmfels, O. Wienand. Convex rank tests and semigraphoids. *SIAM Journal on Discrete Mathematics* 23 (2009) 1117–1134.

[6] A. Postnikov, V. Reiner, L. Williams. Faces of generalized permutohedra. *Documenta Mathematica* 13 (2008) 207–273.
[7] M. Studený, R.R. Bouckaert, T. Kočka. Extreme supermodular set functions over five variables. Research report n. 1977, Institute of Information Theory and Automation, Prague, January 2000.
[8] M. Studený. *Probabilistic Conditional Independence Structures*. Springer, 2005.
[9] M. Studený, T. Kroupa. Core-based criterion for extreme supermodular functions. *Discrete Applied Mathematics* 206 (2016) 122–151.

Reporter: Carlos Enrique Améndola Cerón