# ALGEBRAIC STATISTICS 2015

## 8-11 JUNE, GENOA, ITALY

INVITED SPEAKERS
**Elvira Di Nardo** - University of Basilicata
**Thomas Kahle** - OvGU Magdeburg
**Sonja Petrović** - Illinois Institute of Technology
**Jim Q. Smith** - University of Warwick
**Bernd Sturmfels** - University of California Berkeley

website: **http://www.dima.unige.it/~rogantin/AS2015/**

TUTORIALS
**Luis García-Puente** - Sam Houston State University
**Luigi Malagò** - Shinshu University and INRIA Saclay
**Giovanni Pistone** - Collegio Carlo Alberto
**Piotr Zwiernik** - University of Genoa

IMPORTANT DATES
30 April - deadline for abstract submissions
31 May - registration closed

UNIVERSITÀ DEGLI STUDI DI GENOVA

UPO UNIVERSITÀ DEL PIEMONTE ORIENTALE

Bernoulli Society for Mathematical Statistics and Probability

DE CASTRO STATISTICS

**Monday, June 8, 2015**



**Tuesday, June 9, 2015**

**Posters**:

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Exponential Varieties

Bernd Sturmfels[1]

[1] *University of California Berkeley, USA,* bernd@math.berkeley.edu

Exponential varieties arise from exponential families in statistics. These real algebraic varieties have strong positivity and convexity properties, generalizing those of toric varieties and their moment maps. Another special class, including Gaussian graphical models, are varieties of inverses of symmetric matrices satisfying linear constraints. We present a general theory of exponential varieties, with focus on those defined by hyperbolic polynomials. This is joint work with Mateusz Michałek, Caroline Uhler, and Piotr Zwiernik.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Tying Up Loose Strands: Defining Equations of the Strand Symmetric Model

Colby Long[1], Seth Sullivant[1]

[1] *North Carolina State University, Raleigh, NC, USA,* {celong2, smsulli2 }@ncsu.edu

The strand symmetric model is a phylogenetic model designed to reflect the symmetry inherent in the double-stranded structure of DNA. As the four DNA bases are always paired across the two strands (A with T and C with G), a change of base in one strand will induce a complementary change in the other. Thus, we impose the following restrictions on the entries of the transition matrices: $\theta_{AA} = \theta_{TT}, \theta_{AC} = \theta_{TG}, \theta_{AG} = \theta_{TC}, \theta_{AT} = \theta_{TA}, \theta_{CC} = \theta_{GG}, \theta_{CG} = \theta_{GC}, \theta_{CT} = \theta_{GA}, \theta_{GT} = \theta_{CA}$. Imposing only these restrictions gives the general strand symmetric model (SSM). The phylogenetic invariants of a model are algebraic relationships that must be satisfied by the probability distributions arising from the model. Their study was originally proposed as a method for reconstructing phylogenetic trees [3, 9], but they have also been useful theoretical tools in proving identifiability results (see e.g. [1]). Results in [5] imply that to determine generators of the ideal of phylogenetic invariants for the SSM for any trivalent tree, it suffices to determine them for the claw tree, $K_{1,3}$.

Though the general strand symmetric model itself is not group-based, Casanellas and the second author [2] showed that it is still amenable to the Fourier/Hadamard transform technique of [6, 11]. In the Fourier coordinates, it becomes evident that the parameterization of the model for $K_{1,3}$ is a coordinate projection of the secant variety of the Segre embedding of $\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3$. From this observation, the same authors were able to find 32 degree three and 18 degree four invariants of the homogenous ideal for $K_{1,3}$ and to show that these invariants generate the ideal up to degree four. Whether or not these equations generate the entire ideal was heretofore unknown. Our recent finding is the following theorem.

**Theorem 1** *[10] The vanishing ideal of the strand symmetric model for the graph $K_{1,3}$ is minimally generated by* 32 *cubics and* 18 *quartics. The ideal has dimension* 20*, degree* 9024*, and Hilbert series*

$$\frac{1+12t+78t^2+332t^3+984t^4+1908t^5+2394t^7+1908t^8+984t^9+332t^{10}+78t^{11}+12t^{12}+t^{13}}{(1-t)^{20}}.$$

In this talk, we will discuss the procedure that we used to obtain this result and how the same methods might be applied to other problems arising in algebraic statistics. Our two major tools are the tropical secant dimension approach of Draisma [4] and the following lemma.

**Lemma 1** *[7, Proposition 23] Let k be a field and $J \subset k[x_1, \ldots, x_n]$ be an ideal containing a polynomial $f = gx_1 + h$ with $g, h$ not involving $x_1$ and $g$ a non-zero divisor modulo $J$. Let $J_1 = J \cap k[x_2, \ldots, x_n]$ be the elimination ideal. Then $J$ is prime if and only if $J_1$ is prime.*

First, we use the tropical secant dimension approach to determine the dimension of the variety of probability distributions arising from the model. Then, using Macaulay2 [8], we show that the ideal generated by these fifty equations defines a variety of the same dimension. Finally, with the aid of symbolic computation (again using Macaulay2), we generate a decreasing sequence of elimination ideals, and apply Lemma 1 to demonstrate that the ideal in question is prime. Thus, the variety defined by these equations is irreducible, contains the parameterization, and is of the same dimension as the parameterization, from which the result follows.

We have since used this same method to determine generators for the secant ideals of the binary Jukes-Cantor model for trees with six or fewer leaves. In general, the same procedure may prove useful in instances where one can compute low degree equations of an ideal and wishes to determine if these equations generate the entire ideal under consideration.

# References


[1] E.S. Allman and J.A. Rhodes (2006), The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *J. Comp. Biol.*, **13(5)**, 1101–1113.

[2] M. Casanellas and S. Sullivant (2005) The Strand Symmetric Model. In *Algebraic Statistics for Computational Biology, chapter 16* (B. Sturmfels and L. Pachter), Cambridge University Press, Cambridge, United Kingdom.

[3] J.A. Cavender and J. Felsenstein (1987) Invariants of phylogenies in a simple case with discrete states. *J. of Class.*, **4**, 57–71.

[4] J. Draisma (2005), A tropical approach to secant dimensions. *J. Pure Appl. Algebra*, **212(2)**, 349–363.

[5] J. Draisma and J. Kuttler (2009), On the ideals of equivariant tree models. *Math. Ann.*, **344(3)**,619–644.

[6] S.N. Evans and T.P. Speed (1993), Invariants of some probability models used in phylogenetic inference. *Ann. Statist*, **21(1)**,355–377.

[7] L.D. Garcia, M. Stillman, and B. Sturmfels (2005), Algebraic geometry of bayesian networks. *Journal of Symbolic Computation*, **39(3-4)**, 331–355.

[8] D.R. Grayson and M.E. Stillman. Macaulay2, a software system for research in algebraic geoemetry. Available at http://www.math.uiuc.edu/Macaulay2/, 2002.

[9] J. A. Lake (1987), A rate-independent technique for analysis of nucleaic acid sequences:evolutionary parsimony. *Molecular Biology and Evolution*, **4**, 167–191.

[10] C.Long and S. Sullivant (2015), Tying up lose strands: Defining equations of the strand symmetric model. *Journal of Algebraic Statistics*, (to appear).

[11] L. Székely, P.L. Erdös, M.A. Steel, and D. Penny (1993), A fourier inversion formula for evolutionary trees. *Applied Mathematics Letters*, **6(2)**,13–17.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# From Factorial Designs to Hilbert Schemes

Lorenzo Robbiano[1]

[1] *University of Genova, Italy,* robbiano@dima.unige.it

This talk is meant to explain the evolution of research which originated a few years ago from some problems in statistics. In particular, the inverse problem for factorial designs gave birth to new ideas for the study of special schemes, called Border Basis Schemes. They parametrize zero-dimensional ideals which share a common quotient basis, and turn out to be open sets in the corresponding Hilbert Schemes.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Decomposing Tensors into Frames

Elina Robeva[1]

[1] *University of California, Berkeley, USA,* erobeva@gmail.com

Recent machine learning studies [1, 2] have shown that one can learn latent variable model parameters by finding the decomposition of a given tensor. This is done by applying a transformation to the tensor and then using the tensor power method. In [1], the authors study tensors $T$ of rank at most $n$ of size $n \times n \times \cdots \times n$ ($d$ times), which represent the observed data from a latent variable model. They show that by transforming $T$ to an orthogonally decomposable tensor $T_{od}$, the power method recovers the decomposition of $T_{od}$ and that also gives the decomposition of $T$. It turns out that the elements in the decomposition of $T_{od}$ are robust eigenvectors. In [4] we study the algebraic geometry of orthogonally decomposable tensors. We give a formula for all of their eigenvectors in terms of the robust ones. Furthermore, we propose equations that define the variety of all orthogonally decomposable tensors.

In [2], the authors consider $n \times n \times \cdots \times n$ tensors of rank higher than $n$ and show that under certain conditions, the tensor power method can still recover the decomposition. Motivated by this study, we take [4] one step further. In join work with Luke Oeding and Bernd Sturmfels [3] we study the algebraic geometry of symmetric tensors which can be decomposed as $T = \sum_{i=1}^{r} \lambda_i v_i^{\otimes d}$ where $v_1, \ldots, v_r$ form a unit norm tight frame. We explain for which types of frames one can use the tensor power method to recover the decomposition. In the case $n = 2$, the variety of frame decomposable tensors is given by the vanishing of the maximal minors of a certain matrix whose entries are linear in the entries of $T$. Using this representation, we can recover the decomposition of such tensors efficiently.

## References

[1] A. Anandkumar, R. Ge, D. Hsu, S. Kakade and M. Telgarsky: *Tensor decompositions for learning latent variable models*, Journal of Machine Learning Research **15** (2014) 2773–2832.

[2] A. Anandkumar, D. Hsu, M. Janzamin and S. Kakade: *When are Overcomplete Topic Models Identifiable? Uniqueness of Tensor Tucker Decompositions with Structured Sparsity*. Advances in Neural Information Processing Systems 26, pp.1986-1994, 2013.

[3] L. Oeding, E. Robeva and B. Sturmfels: *Decomposing Tensors into Frames*. arXiv:1243678.

[4] E. Robeva: *Orthogonal decomposition of symmetric tensors*, arXiv:1409.6685.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# R package for algebraic statistics

Luis García-Puente[1]

[1] *Sam Houston State University, USA,* lgarcia@shsu.edu

In this tutorial we will introduce the R package ?algstat?. R is a free software environment for statistical computing and graphics. The package algstat provides functionality for algebraic statistics in R. We will discuss some of its features such as exact inference in log-linear models for contingency table data, analysis of ranked and partially ranked data, and basic multivariate polynomial manipulation through its interface with computer algebra systems such as Macaulay2 and Bertini. The tutorial will include a large practical/hands on component. No previous experience with R or other computer algebra systems is required.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Algebraic geometry of Poisson regression

Thomas Kahle [1]

[1] *OvGU Magdeburg, Germany,* thomas.kahle@ovgu.de

Designing experiments for generalized linear models is tricky because the optimal design depends on unknown parameters. Here we investigate local optimality. We try to understand, for each design, its region of optimality in parameter space. In some cases these regions are semi-algebraic and feature interesting symmetries. We demonstrate this with the Rasch Poisson counts model. This is joint work with Rainer Schwabe.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# What are shell structures of random networks telling us?

Sonja Petrović[1]

[1] *Illinois Institute of Technology, USA,* Sonja.Petrovic@iit.edu

In the network (random graphs) literature, network analyses are often concerned - either directly or indirectly - with the degrees of the nodes in the network. Familiar statistical frameworks, such as the beta or p1 models, associate probabilities to networks in terms of their degree distributions. However, this approach may fail to capture certain vital connectivity information about the network. Often, it matters not just to how many other nodes a particular node in the network is connected, but also to which other nodes it is connected. Degree-centric analyses are not well-suited to model such situations. This talk introduces a model family for one such connectivity structure motivated by examples of social networks, and discusses the relevant algebraic/geometric problems, simulations and sampling algorithms. (Joint work with Karwa, Pelsmajer, Stasi, Wilburne)

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# The geometry of the statistical model for range–based localization

M. Compagnoni[1], R. Notari[1], A.A. Ruggiu[2], F. Antonacci[3], A. Sarti[3]

[1] *Politecnico di Milano - Dipartimento di Matematica, Milano, Italy,* {marco.compagnoni,roberto.notari}@polimi.it
[2] *Linköping University - MAI, Linköping, Sweden,* andrea.ruggiu@liu.se
[3] *Politecnico di Milano - Dipartimento di Elettronica, Informatica e Bioingegneria, Milano, Italy,*
{fabio.antonacci,augusto.sarti}@polimi.it

A plenty of problems in science and engineering are formulated in terms of distances (or ranges) between couples of points in a given set. Without any claim to exhaustiveness, we can list a number of research fields where the above issue plays a key role: radar and sonar technology, wireless sensor networks, statics, robotics, molecular conformation and dimensionality reduction in statistics and machine learning (see [1]).

In this talk, we focus on the problem of locate a radiant source using range measurements. This is the prototype problem in active localization technologies such as radar and active sonar. In this situations, the measurements are the time delays between the transmission of a pulse signal and the reception of its echo. Assuming known and constant the speed of propagation, the Time Of Arrival (TOA) of the signal is directly related to the range between the source and the corresponding emitter/receiver. The goal of the localization is to find the source position from the TOAs.

The localization problem is a fundamental issue also for wireless sensor networks. Indeed, the network routers must be updated of the positions of the sensors (e.g. smartphones), in order to adapt routes, frequencies, and network ID data accordingly. It is well known that the distance between any pair of sufficiently close sensors is strongly correlated to the battery charge used in their communications. Furthermore, by the fact that the positions of the fixed elements of the network (e.g routers and repeaters) are known, it follows that wireless sensor networks localization has many analogies to a multi–source localization problem based on TOA measurements.

In mathematical literature, the problems involving ranges measurements have been intensively studied in the context of Euclidean Distance Geometry (DG) [1]. The fundamental problem in DG can be formulated in terms of the embeddability of a weighted graph $G = (V,E)$ into a suitable $k$–dimensional Euclidean space. Roughly speaking, one has to understand when the set $V$ of vertices $p_i,\ i = 1,\dots,n$ actually corresponds to a set of points $\phi(p_i),\ i = 1,\dots,n$ in $R^k$, where the Euclidean distance $\|\phi(p_i) - \phi(p_j)\|$ is equal to the weight of the edge $e_{ij} \in E$. In localization problems the vertices $p_i$ of $G$ correspond to the sensors and sources, while the weighted edges $e_{ij}$ are the available range measurements.

In real world applications the range measurements are affected by noise. The source estimation, in particular the Maximum Likelihood Estimation (MLE), is a non linear and non convex problem, therefore it is difficult to globally solve it. This is why researchers have studied many different approaches and algorithms that give rise to robust estimations, but that are suboptimal from a statistical point of view. The wish to give different perspective on this important problem has been one of the main motivation of our work.

In this talk we give an overview of the range–based localization from the point of view of differential and algebraic geometry. We focus on the range–based source localization with two and three calibrated and synchronous sensors [2]. Firstly, we define the stochastic model for range measurements. It encodes the range–based localization into a map from the Euclidean space containing source and receivers to the space of range measurements. Then, we offer a complete characterization of such a map. We address the identifiability problem and we describe in great details the geometry of the sets of feasible range measurements.

The most interesting results concern the case of three non collinear receivers. If also the source belongs to the receivers plane, then the set of feasible measurements is contained in a Kummer's quartic surface. We give a detailed description of its geometric properties, both from an algebraic and a differential point of view. In particular, we investigate the link between the properties of the Kummer's and some distinguished sets in the physical Euclidean plane. On the contrary, if the source stays outside the receivers plane, the set of feasible measurements is the domain bounded by the previous surface. Thanks to classical results on Kummer's surfaces, we have been able to compute the convex hull of the set of measurements and a linear approximation of it.

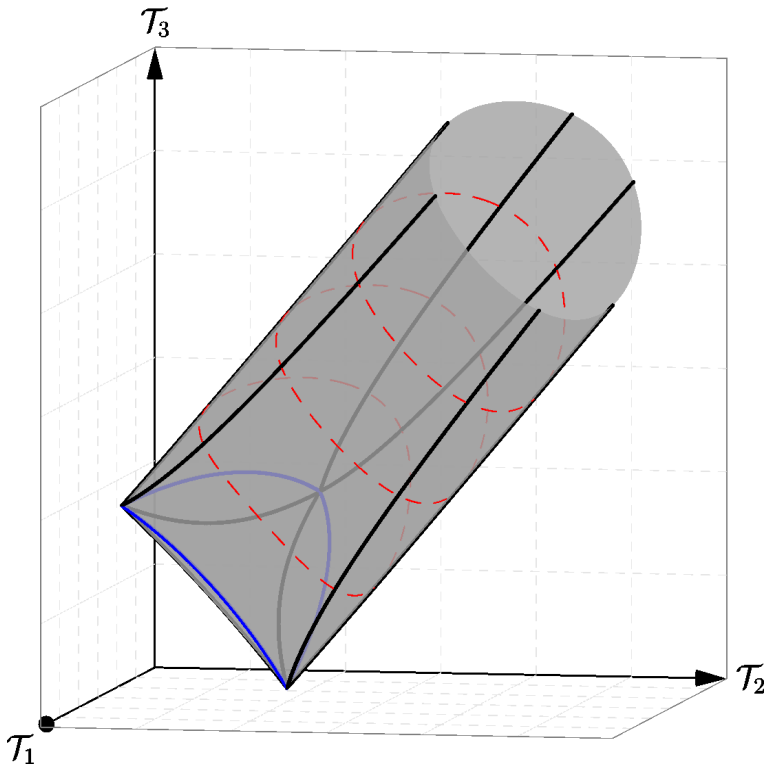



Figure 1: An example of the set of feasible range measurements for a non collinear configuration of receivers. The grey surface is the real part of a Kummer's surface contained in the first octant of the space of the ranges. The bold curves are arcs of conics and they are the asymptotic curves of the surface. They meet at the singular points of the surface, which are the images of the receivers.

We performed numerical tests on Euclidean Distance degree of the surface. Currently, we are studying the Euclidean Distance discriminant and the extension to a larger number of receivers and sources. Our analysis is helpful also for the study of pseudorange–based localization (e.g. Global Positioning System). Indeed, from a mathematical point of view, the two models are related through a linear projection.

## References


[1] L. Liberti and C. Lavor and N. Maculan and A. Mucherino (2014), Euclidean Distance Geometry and Applications, *SIAM REVIEW*, **56-1**, 3-69.

[2] M. Compagnoni, R. Notari, A.A. Ruggiu, F. Antonacci and A. Sarti, (2015), The Algebro-Geometric Study of Range Maps, preprint.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Matrices of nonnegative rank at most three

Rob H. Eggermont[1], Emil Horobeţ[2] and Kaie Kubjas[3]

[1] *Eindhoven University of Technology, The Netherlands, r.h.eggermont@tue.nl*
[2] *Eindhoven University of Technology, The Netherlands, e.horobet@tue.nl*
[3] *Aalto University, Finland, kaie.kubjas@gmail.com*

The *nonnegative rank* of a matrix $M \in \mathbb{R}^{m\times n}_{\geq 0}$ is the smallest $r \in \mathbb{N}$ such that there exist matrices $A \in \mathbb{R}^{m\times r}_{\geq 0}$ and $B \in \mathbb{R}^{r\times n}_{\geq 0}$ with $M = AB$. Matrices of nonnegative rank at most $r$ form a semialgebraic set, i.e. they are defined by Boolean combinations of polynomial equations and inequalities. We denote this semialgebraic set by $\mathscr{M}^r_{m\times n}$. If a nonnegative matrix has rank 1 or 2, then its nonnegative rank equals its rank. In these cases, the semialgebraic set $\mathscr{M}^r_{m\times n}$ is defined by $2\times 2$ or $3\times 3$-minors respectively together with the nonnegativity constraints. In the first interesting case when $r = 3$, a semialgebraic description is given by Robeva, Sturmfels and Kubjas [4, Theorem 4.1].

This description is in the parameter variables of $A$ and $B$, where $M = AB$ is any size 3 factorization of $M$, so it is not clear from the description what (the Zariski closure of) the boundary is. Some boundary components are defined by the ideals $\langle x_{ij}\rangle$, where $1 \leq i \leq m, 1 \leq j \leq n$ and $x_{ij}$ denote the coordinates on $M_{m\times n}$. We call them the trivial boundary components.

In this talk we present the proof and some consequences of the main result from [2], previously conjectured in [4, Conjecture 6.4]:

**Theorem 2 ([2], Theorem 1.1)** *Let $m \geq 4, n \geq 3$ and consider a nontrivial irreducible component of $\overline{\partial \mathscr{M}^3_{m\times n}}$. The prime ideal of this component is minimally generated by $\binom{m}{4}\binom{n}{4}$ quartics, namely the $4\times 4$-minors, and either by $\binom{m}{3}$ sextics that are indexed by subsets $\{i,j,k\}$ of $\{1,2,\ldots,m\}$ or $\binom{n}{3}$ sextics that are indexed by subsets $\{i,j,k\}$ of $\{1,2,\ldots,n\}$. These form a Gröbner basis with respect to graded reverse lexicographic order.*

One motivation for studying the nonnegative matrix rank comes from statistics. A probability matrix of nonnegative rank $r$ records joint probabilities $\mathrm{Prob}(X = i, Y = j)$ of two discrete random variables $X$ and $Y$ with $m$ and $n$ states respectively that are conditionally independent given a third discrete random variable $Z$ with $r$ states. The intersection of $\mathscr{M}^r_{m\times n}$ with the probability simplex $\Delta_{mn-1}$ is called the $r$-th mixture model, see [5, Section 4.1] for details. Nonnegative matrix factorizations appear also in audio processing [3], image compression and document analysis [5].

Understanding the Zariski closure of the boundary is necessary for solving optimization problems on $\mathscr{M}^r_{m\times n}$ with the certificate that we have found a global maxima. One example of such an optimization problem is the maximum likelihood estimation, i.e. given data from observations one would like to find a point in the $r$-th mixture model that maximizes the value of the likelihood function. To find the global optima, one would have to use the method of Lagrange multipliers on the Zariski closure of the semialgebraic set, its boundaries and intersections of boundaries.

# References


[1] Mathias Drton, Bernd Sturmfels and Seth Sullivant, *Lectures on Algebraic Statistics*, Oberwolfach Seminars 39. Birkhäuser Verlag, 2009.

[2] Rob H. Eggermont, Emil Horobeţ and Kaie Kubjas, *Algebraic boundary of matrices of nonnegative rank at most three*. Available at arXiv:1412.1654.

[3] Sebastian Ewert and Meinard Müller, *Score-Informed Source Separation for Music Signals*, in Multimodal Music Processing, Dagstuhl Follow-Ups 3 (2012), 73–94.

[4] Kaie Kubjas, Elina Robeva and Bernd Sturmfels, *Fixed Points of the EM Algorithm and Nonnegative Rank Boundaries*, Annals of Statistics, 43 (2015), 422–461.

[5] Daniel D. Lee and H. Sebastian Seung, *Learning the parts of objects by non-negative matrix factorization*, Nature 401 (1999), 788–791.

# Algebraic Graph Limits

A. Engström[1], P. Norén[2]

[1] *Aalto University, Finland,* alexander.engstrom@aalto.fi
[2] *North Carolina State University, USA,* pgnoren2@ncsu.edu

The theory of graph limits associates random graph models to symmetric measurable functions on the unit square [1, 2, 3]. We investigate what happens when these functions are polynomials. For low degree polynomials the models appearing are familiar and important, for example preferential attachment and Erdös-Rényi. The higher degree polynomials are also useful as any graph limit can be arbitrarily well approximated by a polynomial. We show that this setup could be useful in applications: To determine the parameters of an algebraic graph limit that fits observed data best one can use numerical algebraic geometry efficiently.

## References

[1] P. Diaconis and S. Janson: *Graph limits and exchangeable random graphs.* Rend. Mat. Appl. (7)28 (2008),no.1, 33 – 61.

[2] A. Engström and P. Norén: *, Polytopes from subgraph statistics*.Preprint at arXiv:1011.3552.

[3] L. Lovász and B. Szegedy: *Limits of dense graph sequences.* J. Combin. Theory Ser. B96 (2006), no.6, 933–957.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Hypergraph Decompositions and Toric Ideals

E. Gross[1], K. Kubjas[2]

[1] *San José State University, USA,* elizabeth.gross@sjsu.edu
[2] *Aalto University, Helsinki, Finland,* kaie.kubjas@gmail.com

Let $H = (E, V)$ be a hypergraph where $V$ is the set of vertices and $E \subseteq 2^V \setminus \{\emptyset\}$ is the set of hyperedges. Let $k[p_e : e \in E]$ and $k[q_v : v \in V]$ be polynomial rings over a field $k$. The toric ideal $I_H$ of a hypergraph $H$ is a binomial ideal defined as the kernel of the ring homomorphism

$$\begin{aligned} k[p_e : e \in E] \to &\quad k[q_v : v \in V], \\ p_e \mapsto &\quad \prod_{v \in e} q_v. \end{aligned}$$

Any toric ideal arises as the kernel of a monomial map encoded by an integer matrix $A$. In the case of the toric ideal $I_H$ of a hypergaph $H$, this is the incidence matrix of the hypergraph $H$. In fact all, all toric ideals defined by 0-1 matrices can be regarded as toric ideals of hypergraphs.

In this ongoing work we study generating sets of toric ideals $I_H$ of hypergraphs or equivalently, by the Fundamental Theorem of Markov Bases [1], Markov bases of log-linear models with square-free parameterizations. In particular, we address a modification of [2, Problem 6.3], which was also asked by Sonja Petrović at the open problem session of Algebraic Statistics 2014 at IIT:

**Problem** Given a hypergraph $H$ that is obtained by identifying vertices from two smaller hypergraphs $H_1$ and $H_2$, is it possible to obtain generating set of $I_H$ from the generating set of $I_{H_1}$ and $I_{H_2}$?

We give an affirmative answer to this question, and give explicit constructions for a Markov basis and the Graver basis. As an example, consider the monomial sunflower [2] in Figure 2. We can construct larger monomial sunflowers by taking an even number of copies of $H$ and identifying all copies of the vertex $v_1$. We consider 128 copies of the sunflower. If we split it into two and apply our construction for computing a Markov basis, then we get a 10 times speed up compared to computing a Markov basis directly using `Macaulay2` interface for `4ti2`.

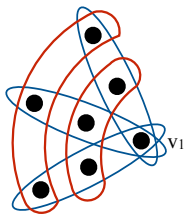



Figure 2: Monomial sunflower.

## References

[1] P. Diaconis and B. Sturmfels (1998), Algebraic algorithms for sampling from conditional distributions, *Ann. Statist.*, **26**, 363-397.
[2] S. Petrović and D. Stasi (2014), Toric algebra of hypergraphs, *J. Algebraic Combin.*, **39**, 187-208.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# The maximum likelihood degree of rank 2 matrices via Euler characteristics

Jose Isreael Rodriguez[1], Botong Wang[2]

[1] *University of Notre Dame, Indiana, USA,* jo.ro@ND.edu

[2] *University of Notre Dame, Indiana, USA,* bwang3@ND.edu

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 [3, 5]. In this talk, we provide closed form expressions for ML degrees of rank at most 2 matrices, which are the Zariski closure of mixtures of independence models. This answers an outstanding conjecture of [2].

We consider the case for two discrete random variables, having $m$ and $n$ states respectively. A joint probability distribution for two such random variables is written as an $m \times n$-matrix:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{bmatrix}. \tag{1}$$

The $(i,j)$th entry $p_{ij}$ represents the probability that the first variable is in state $i$ and the second variable is in state $j$. By a statistical model, we mean a subset $\mathscr{M}$ of the probability simplex $\triangle_{mn-1}$ of all such matrices $P$. The models we consider in this talk are $\mathscr{M}_{mn}$ the set of rank at most 2 matrices.

If i.i.d. samples are drawn from some distribution $P$, then we summarize the data also in a matrix:

$$u = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{m1} & u_{m2} & \cdots & u_{mn} \end{bmatrix}. \tag{2}$$

The entries of $u$ are non-negative integers where $u_{ij}$ is the number of samples drawn with state $(i,j)$.

The likelihood function corresponding to the data matrix $u$ is given by

$$\ell_u(p) := p_{11}^{u_{11}} p_{12}^{u_{12}} \cdots p_{mn}^{u_{mn}} \tag{3}$$

Maximum likelihood estimation is an optimization problem. This problem consists of determining, for fixed $u$, the argmax of $\ell_u(p)$ on a statistical model $\mathscr{M}$. The optimal solution is called the maximum likelihood estimate (mle). For the models we consider, the mle is a solution to the likelihood equations. So by solving the likelihood equations, we solve the maximum likelihood estimation problem. Since the ML degree is the number of solutions to the likelihood equations, it gives a measure on the difficulty of the problem.

In [2], the following table of ML degrees of $\mathscr{M}_{mn}$ were computed for various $m$ and $n$:

| $n=$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| $m=3:$ | 10 | 26 | 58 | 122 | 250 | 506 | 1018 | 2042 |
| $m=4:$ | 26 | 191 | 843 | 3119 | 6776 | | | |

The first main result of this talk and of [7] proves a formula for the first row of the table:

$$\texttt{MLdegree}\mathscr{M}_{3n} = 2^{n+1} - 6 \texttt{ for } n \geq 3.$$

Our techniques relate ML degrees to Euler characteristics. Work by Huh [4] has shown that the ML degree of smooth algebraic statistical models $\mathscr{M}$ with Zariski closure $X$ equals the signed topological characteristic of an open subvariety $X^o$ where $X^o$ is the set of points of $X$ with nonzero coordinates and non-zero coordinate sums. More recent work of Budur and Wang [1] show that the ML degree of a singular model is a stratified topological invariant. They show that the Euler characteristic of $X^o$ is a sum of ML degrees weighted by Euler obstructions, which can be thought of measuring the multiplicity of the singular locus.

We further develop these techniques and apply them to the mixture model $\mathscr{M}_{mn}$. Doing so, we provide a recursion to compute the ML degree of $\mathscr{M}_{mn}$ for any fixed $m$ and $n$, see [7]. This talk will present an example based on DiaNA's dice [6] to bridge statistics, Euler characteristics, and applied algebraic geometry, concluding with the recursion.

## References


[1] N. Budur and B. Wang. Bounding the maximum likelihood degree. *Preprint* arXiv:1411.3486.

[2] J. Hauenstein, J. I. Rodriguez, and B. Sturmfels. Maximum likelihood for matrices with rank constraints. *Journal of Algebraic Statistics*, 2012.

[3] S. Hoşten, A. Khetan, and B. Sturmfels. Solving the likelihood equations. *Foundations of Computational Mathematics*, 5(4):389–407, 2005.

[4] J. Huh, *The maximum likelihood degree of a very affine variety*, Compos. Math. **149** (2013), no. 8, 1245–1266.

[5] J. Huh and B. Sturmfels. *Likelihood geometry*, pages 63–117. Springer International Publishing, 2014.

[6] L. Pachter and B. Sturmfels. *Algebraic statistics for computational biology*, pages 3–42. Cambridge Univ. Press, New York, 2005.

[7] J. Rodriguez and B. Wang. The maximum likelihood degree of rank 2 matrices via Euler characteristics. *Preprint.*

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# A Linear-Algebraic Criterion for Indecomposable Generalized Permutohedra

Milan Studený[1], Tomáš Kroupa[2]

[1] *Institute of Information Theory and Automation of the CAS, Czech Republic,* studeny@utia.cas.cz

[2] *Department of Mathematics, University of Milan, Italy,* Tomas.Kroupa@unimi.it

The contribution concerns the geometry of *conditional independence* (CI). Morton [4, Thm 2.4.3] in his thesis established a one-to-one correspondence between (the class of) structural CI models [8] and (the class of) certain polytopes, namely Minkowski summands of the permutohedron. These polytopes are known in the recent literature as *generalized permutohedra*.

The generalized permutohedra were introduced by Postnikov and his co-workers [5, 6] as the polytopes obtainable by moving vertices of the usual permutohedron while the directions of edges are preserved. Their connection to supermodular and submodular functions has been indicated by Doker in his thesis [1].

In our recent manuscript [9, Cor 11] we have observed that generalized permutohedra, in fact, coincide with (the class of) polytopes which were formerly studied in the context of the cooperative game theory, namely with the *cores* of supermodular games, called *convex games* in that context [7]. We have been interested in the description (and later possible characterization) of those supermodular games that are *extreme* (= generating the extreme rays of the cone of standardized supermodular games). It turns out that the core polytopes for these extreme supermodular games are just those generalized permutohedra $P$ that are *indecomposable* in sense of Meyer [3], which means that every Minkowski summand of $P \subseteq \mathbb{R}^N$ has the form $\alpha \cdot P \oplus \{v\}$, where $\alpha \geq 0$ and $v \in \mathbb{R}^N$.

Motivated by the game-theoretical point of view, we have offered in [9] a simple linear-algebraic criterion to recognize whether a (standardized) supermodular game is extreme. The criterion is based on a vertex-description of the corresponding core polytope, which is easy to find owing to a classic result by Shapley [7]. Our criterion leads to solving a linear equation system determined by the combinatorial *core structure*, which is a concept recently pinpointed in the context of game theory [2].

Thus, our result gives, as a by-product, a criterion to recognize whether a given generalized permutohedron is indecomposable. Note that the criterion is different (and simpler than) Meyer's general criterion to recognize indecomposable polytopes based on their facet-description [3].

In the first part of the presentation, the first author plans to recall the motivation and explain the wider context as indicated in this abstract. In the second part of the presentation, the second author plans formulate the criterion from [9] and illustrate it by a few simple examples.

*Acknowledgements.* Milan Studený is supported from the GAČR project n. 13-20012S. Tomáš Kroupa gratefully acknowledges the support from Marie Curie Intra-European Fellowship OASIG (PIEF-GA-2013-622645).

# References


[1] J.S. Doker. Geometry of generalized permutohedra. PhD thesis, University of California Berkeley, 2011.

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

[3] W.J. Meyer. Indecomposable polytopes. *Transaction of the American Mathematical Society* 190 (1974) 77–86.

[4] J.R. Morton. Geometry of conditional independence. PhD thesis, University of California Berkeley, 2007.

[5] A. Postnikov. Permutohedra, associahedra, and beyond. *International Mathematics Research Notices* 6 (2009) 1026–1106; see also the previous manuscript from July 2005 available at `arxiv.org/abs/math/0507163`.

[6] A. Postnikov, V. Reiner, L. Williams. Faces of generalized permutohedra. *Documenta Mathematica* 13 (2008) 207–273.

[7] L.S. Shapley. Cores of convex games. *International Journal of Game Theory* 1 (1971/72) 11–26.

[8] M. Studený. *Probabilistic Conditional Independence Structures*. Springer, 2005.

[9] M. Studený, T. Kroupa. Core-based criterion for extreme supermodular functions. A manuscript submitted to Discrete Applied Mathematics (October 2014), available at `arxiv.org/abs/1410.8395`.

# Symbolic methods in statistics: elegance towards efficiency

Elvira Di Nardo[1]

[1] *University of Basilicata, Italy,* elvira.dinardo@unibas.it

In the last ten years, the employment of symbolic methods has substantially extended both the theory and the applications of statistics. By symbolic methods we refer to the set of manipulation techniques aiming to perform algebraic calculations through an algorithmic approach. The goal is to find efficient mechanical processes to pass to a computer. Typically these algebraic expressions are encountered within statistical inference or parameter estimation. Recent connections with free probability and its applications, within random matrices and other satellite area, have extended its boundaries of applicability. To find efficient symbolic algorithms challenges with new problems involving both computational and conceptual issues. There are many packages devoted to numerical/graphical statistical tool sets but not doing algebraic/symbolic computations. The packages filling this gap are not open source. R is a much stronger numeric programming environment and the procedures including symbolic software are not yet specifically oriented for statistical calculations. So the availability of a widely spread open source symbolic platform will be of great interest, especially if there are interface capabilities to external programs. The conceptual aspects related to symbolic methods involve more strictly mathematical issues. In this picture, the combinatorics has no doubt a preeminent role. But, what we regard as symbolic computation is now evolving towards an universal algebraic language which aims to combine syntactic elegance and computational efficiency. Experience have shown that syntactic elegance often requires the acquisition of innovative techniques and to climb this steep learning curve can be a deterrent to pursue the goal. But, having got a different and deeper viewpoint, the efficiency is obtained as by product and the result can be surprisingly better of what you expected. Working examples will be polykays for random vectors or random matrices, with special reference to non-central Wishart distributions.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# The Geometry of Chain Event Graphs

Jim Q. Smith[1]

[1] *The University of Warwick, UK,* j.q.smith@warwick.ac.uk

The class of chain event graphs (CEGs) - which contains the class of discrete Bayes Nets as a special case - has now been established as a widely applicable modeling tool. But the family also enjoys some interesting associated mathematical structure. A CEG is specifed through an event tree with some of its edge probabilities being equated. So in particular each of its atoms - its root to leaf paths - has a monomial associated to it corresponding to a product of edge probabilities. It therefore follows that, in particular, the class of probability measures associated with each given CEG can be mapped on to a family of polynomials. This gives a new area of statistics where techniques of algebraic geometry can be usefully applied. In this talk I will illustrate how we have recently used this algebraic description to come to a better understanding of the statistical equivalence classes of CEGs. The potential uses of this classification for causal discovery will then be explored. This is joint work with one of my PhD students: Christiane Görgen.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# On patterns of conjunctive forks

Vašek Chvátal[1], František Matúš[2] and Yori Zwols[3]

[1] *Department of Computer Science and Software Engineering, Concordia University, Montreal, Quebec H3G 1M8, Canada,* chvatal@cse.concordia.ca
[2] *Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Prague 8, Czech Republic,* matus@utia.cas.cz
[3] *Google, London, UK,* yzwols@gmail.com

H. Reichenbach [13, Chapter 19] defined a *conjunctive fork* as an ordered triple $(A,B,C)$ of events in a probability space $(\Omega,\mathscr{F},\mathrm{P})$ such that

$$\mathrm{P}(A\cap C|B)=\mathrm{P}(A|B)\,\mathrm{P}(C|B)\,,\qquad \mathrm{P}(A|B)>\mathrm{P}(A|\Omega\setminus B)\,,$$
$$\mathrm{P}(A\cap C|\Omega\setminus B)=\mathrm{P}(A|\Omega\setminus B)\,\mathrm{P}(C|\Omega\setminus B)\,,\qquad \mathrm{P}(C|B)>\mathrm{P}(C|\Omega\setminus B)\,.$$

Implicit in this definition is the assumption $0<\mathrm{P}(B)<1$, which is necessary to make the conditional probabilities well-defined. The context of discourse was physics: conjunctive forks play a central role in Reichenbach's causal theory of time. In this role, they have attracted considerable attention: over one hundred publications, such as [1], [14], [15], [6], [2], [4], [16], [7], [10], refer to them. A similar notion was introduced earlier by P. Kendall and P. Lazarsfeld [8, Part I, Section 2] in the context of sociology.

As recognized by W. Spohn [16, page 2], Reichenbach's definition can be interpreted in terms of *indicator functions* $\mathbb{1}_A$ of events $A$. The above equalities mean that the random variables $\mathbb{1}_A$ and $\mathbb{1}_C$ are conditionally independent given $\mathbb{1}_B$. The strict inequalities are equivalent to $\mathrm{P}(AB)>\mathrm{P}(A)\mathrm{P}(B)$ and $\mathrm{P}(BC)>\mathrm{P}(B)\mathrm{P}(C)$. This means that the covariance of $\mathbb{1}_A$ and $\mathbb{1}_B$ is positive and so is that of $\mathbb{1}_B$ and $\mathbb{1}_C$.

Given (not necessarily distinct) events $A,B,C\in\mathscr{F}$, let $(A,B,C)_{\mathrm{P}}$ mean that the triple of events is a conjunctive fork. Events $A_i$ indexed by $i$ in finite set $N$ will be said to *fork represent* a ternary relation $\mathscr{R}$ on a ground set $N$ if and only if

$$(i,j,k)\in\mathscr{R}\;\;\Leftrightarrow\;\;(A_i,A_j,A_k)_{\mathrm{P}}\,.$$

A ternary relation will be called *fork representable* if and only if it admits a fork representation. In this contribution the finite fork representable relations are characterized in a way which implies that fork representability of finite ternary relations can be tested in polynomial time.

Following E. Pitcher and M. F. Smiley [12] a ternary relation $\mathscr{R}$ is called a *betweenness* if and only if it satisfies, for all choices of elements $i,j,k$ of its ground set,

$$(i,j,k)\in\mathscr{R}\;\;\Rightarrow\;\;(k,j,i)\in\mathscr{R}\,,$$
$$(i,j,k)\in\mathscr{R}\;\text{ and }\;(i,k,j)\in\mathscr{R}\;\;\Rightarrow\;\;j=k\,,$$

and $(i,j,j)\in\mathscr{R}$. It is called *weak betweenness* if it satisfies the above two implications,

$$(i,k,j)\in\mathscr{R}\;\;\Rightarrow\;\;(i,j,j)\in\mathscr{R},(j,k,k)\in\mathscr{R}\;\text{ and }\;(k,i,i)\in\mathscr{R}$$

and $(i,i,i)\in\mathscr{R}$. Given a ternary relation $\mathscr{R}$, let $V_{\mathscr{R}}=\{i\in N\colon (i,i,i)\in\mathscr{R}\}$ and $\overset{\mathscr{R}}{\sim}$ be the binary relation on $V_{\mathscr{R}}$ defined by

$$i\overset{\mathscr{R}}{\sim}j\;\;\text{ if and only if }\;\;(i,j,i)\in\mathscr{R}\text{ and }(j,i,j)\in\mathscr{R}.$$

The relation $\mathscr{R}$ will be called *regular* if and only if $\mathscr{R} \subseteq V_{\mathscr{R}}^3$, $\overset{\mathscr{R}}{\sim}$ is an equivalence relation and

$$(i,j,k) \in \mathscr{R},\ i \overset{\mathscr{R}}{\sim} i',\ j \overset{\mathscr{R}}{\sim} j',\ k \overset{\mathscr{R}}{\sim} k' \ \Rightarrow\ (i',j',k') \in \mathscr{R}.$$

Every regular ternary relation $\mathscr{R}$ generates its *quotient relation* which is the ternary relation $\mathscr{Q}$ on the set of equivalence classes of $\overset{\mathscr{R}}{\sim}$ defined by $(I,J,K) \in \mathscr{Q}$ if and only if $(i,j,k) \in \mathscr{R}$ for at least one $i$ in $I$, at least one $j$ in $J$, and at least one $k$ in $K$. For the quotient $\mathscr{Q}$, write

$$E_{\mathscr{Q}} = \{\{I,J\} : I \neq J,\ (I,J,J) \in \mathscr{Q},\ (J,I,I) \in \mathscr{Q}\}$$

and note that if $(I,J,K)$ belongs to $\mathscr{Q}$ and $I,J,K$ are pairwise distinct, then all three $\{I,J\}$, $\{J,K\}$, $\{K,I\}$ belong to $E_{\mathscr{Q}}$. The quotient $\mathscr{Q}$ will be called *solvable* if and only if the system

$$\begin{aligned} &x_{\{I,K\}} = x_{\{I,J\}} + x_{\{J,K\}} && \text{for } (I,J,K) \in \mathscr{Q} \text{ pairwise distinct,} \\ &x_{\{I,J\}} > 0 && \text{for } \{I,J\} \in E_{\mathscr{Q}} \end{aligned}$$

has a solution.

The main result to be presented can be now formulated as follows. *A ternary relation on a finite set is fork representable if and only if it is regular and its quotient relation is a solvable weak betweenness.* Hence, the fork representability of a ternary relation $\mathscr{R}$ can be verified in time polynomial in $|\mathscr{R}|$, by the epoch-making result of L.G. Khachiyan [9].

Constraints of the fork type define new classes of semialgebraic varieties similar to the conditional independence varieties [5] arising from conditional independence structures [11, 17]. Open problems related to the varieties will be discussed.

# References


[1] P. von Bretzel, Concerning a probabilistic theory of causation adequate for the causal theory of time. *Synthese* **35** (1977) 173–190.

[2] N. Cartwright and M. Jones, How to hunt quantum causes. *Erkenntnis* **35** (1991) 205–231.

[3] V. Chvátal and Baoyindureng Wu, On Reichenbach's causal betweenness, *Erkenntnis* **76** (2012) 41–48.

[4] P. Dowe, Process causality and asymmetry. *Erkenntnis* **37** (1992) 179–196.

[5] M. Drton, B. Sturmfels and S. Sullivant, *Lectures on Algebraic Statistics.* Oberwolfach Seminars, 2009, Birkhäuser.

[6] F.S. Ellett, Jr. and D.P. Ericson, Correlation, partial correlation, and causation. *Synthese* **67** (1986) 157–173.

[7] G. Hofer-Szabó, M. Rédei and L.E. Szabó, On Reichenbach's common cause principle and Reichenbach's notion of common cause. *The British Journal for the Philosophy of Science* **50** (1999) 377–399.

[8] P.L. Kendall and P.F. Lazarsfeld, Problems of survey analysis. In: *Continuities in Social Research: Studies in the Scope and Method of "The American Soldier"* (R.K. Merton and P.F. Lazarsfeld, eds.), The Free Press, Glencoe, IL, 1950, pp. 133–196.

[9] L.G. Khachiyan, A polynomial algorithm in linear programming. *Doklady Akademii Nauk SSSR* **244** (1979) 1093–1096. (Russian)

[10] K.B. Korb, Probabilistic causal structure. In: *Causation and Laws of Nature* (H. Sankey, ed.), Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 265–311.

[11] J. Pearl, *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufman, 1988, San Mateo, California.

[12] E. Pitcher and M.F. Smiley, Transitivities of betweenness. *Transactions of the American Mathematical Society* **52** (1942) 95–114.

[13] H. Reichenbach, *The Direction of Time.* University of California Press, 1956, Berkeley and Los Angeles.

[14] W.C. Salmon, Probabilistic causality. *Pacific Philosophical Quarterly* **61** (1980) 50–74.

[15] W.C. Salmon, *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, 1984, Princeton.

[16] W. Spohn, On Reichenbach's principle of the common cause. In: *Logic, Language, and the Structure of Scientific Theories* (W. Salmon and G. Wolters, eds.), University of Pittsburgh Press, Pittsburgh, 1994, pp. 211–235.

[17] M. Studený, *Probabilistic Conditional Independence Structures.* Springer 2005, New York.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Maximum likelihood threshold of a graph

Elizabeth Gross[1], Seth Sullivant[2]

[1] *San José State University, San José, CA, USA,* elizabeth.gross@sjsu.edu

[2] *North Carolina State University, Raleigh, NC, USA,* smsulli2@ncsu.edu

Let $X = (X_1, \ldots, X_m)$ be a $m$-dimensional random vector distributed according to a multivariate normal distribution, i.e. $X \sim \mathcal{N}(\mu, \Sigma)$. In a Gaussian graphical model, an undirected graph $G = (\{1, \ldots, m\}, E)$ encodes the conditional independence structure of the distribution: the edge $(i, j) \notin E$ if and only if $X_i$ and $X_j$ are conditionally independent given the remaining variables.

For Gaussian graphical models, when the number of observations $n$ is larger than the number of random variables $m$, the maximum likelihood estimator (MLE) is known to exist with probability one. But it is often the case, especially in biological applications, that $m \gg n$. In this setting, it is still possible for the MLE to exist with probability one, which invites the question: *For a given graph G, what is the smallest n such that the maximum likelihood estimator of* $\Sigma$ *exists almost surely?* We call the resulting graph invariant the *maximum likelihood threshold* and denote it $\mathrm{mlt}(G)$.

We show that this graph parameter, the maximum likelihood threshold, is connected to the theory of combinatorial rigidity. In particular, if the edge set of a graph $G$ is an independent set in the $(n-1)$-dimensional generic rigidity matroid, then the maximum likelihood threshold of $G$ is less than or equal to $n$. This connection implies many results about the maximum likelihood threshold for large classes of graphs. For example, if $G$ has an empty $n$-core then $\mathrm{mlt}(G) \leqslant n$. Or, as a corollary, let $Gr_{k_1,k_2}$ denote the $k_1 \times k_2$ grid graph with $k_1, k_2 \geqslant 2$, then $\mathrm{mlt}(Gr_{k_1,k_2}) = 3$. This extends the result in [2] that shows that $\mathrm{mlt}(Gr_{3,3}) = 3$.

## References

[1] S. Buhl. On the existence of maximum likelihood estimators for graphical Gaussian models. *Scand. J. Statist.* **20** (1993), no. 3, 263–270.

[2] C. Uhler. Geometry of maximum likelihood estimation in Gaussian graphical models. *Ann. Statist.* **40** (2012), no. 1, 238–261.

[3] E. Gross and S. Sullivant. The maximum likelihood threshold of a graph. arXiv:1404.6989.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# The Algebra of Integrated Partial Belief Systems

M. Leonelli[1], J. Q. Smith[1], E. Riccomagno[2]

[1] *Department of Statistics, The University of Warwick, UK,* {m.leonelli,j.q.smith}@warwick.ac.uk
[2] *Dipartimento di Matematica, Universitá degli Studi di Genova, Italy,* riccomagno@dima.unige.it

Probabilistic decision support tools for single agents are now, although still being refined, well developed and used in practice in a variety of domains. However, the size of current applications requires expert judgements coming from different panels of experts with diverse expertise. For example in nuclear emergency management judgements concerning the safety of the source term, the atmospheric dispersion of a cloud of contamination and the effects on human health deriving from radioactive intake, among the others, need to be taken into account in the decision making process [4, 6].

Integrating Decision Support Systems (IDSSs) [5, 8] have been recently defined to generalize the coherence of Bayesian decision support for single agents to the more realistic multi-expert setting. These take only a few selected probabilistic outputs from each model used by the panels and then paste them together to provide a unique coherent evaluation of the overall problem. A variety of different methodologies can be now employed by the panels to model the domain under their jurisdiction, as for example large scale hierarchical Bayesian spatio-temporal models based on advanced computational algorithms [1] or probabilistic emulators over massive deterministic simulators [3].

Under conditions formally and extensively discussed in [8], a variety of both dynamic and non-dynamic graphical models can be used as an overarching integrating tool to provide a unique coherent picture of the whole problem, in such a way that the judgements of the different panels do not contradict each other. Importantly, current technology allows each of the components of the IDSS to be designed to be fast. Thus the *distributive* nature of IDSSs guarantees that overall estimates can be produced in real time.

The theory of IDSSs has mostly focused on the inferential full-distributional difficulties associated to this integration. However a formal Bayesian decision analysis is based on the maximization of an expected utility function that often only depends on some simple summaries of key output variables, as for example some low order moments. This is the case for example when the utility function is a low degree polynomial. By requesting only this information, the implementation of an IDSS can become orders of magnitude more manageable. Then panels just need to communicate a few summaries from their sample: a trivial and fast task to perform. Surprisingly, it is common to be able to partially define a coherent and distributed system with this property.

In this framework expected utilities are polynomials whose indeterminates are functions of the panels' delivered summaries. A study of the polynomial structure of expected utilities enables us to identify in a variety of examples the required summaries needed by different panels of experts that allows the decision center to compute beliefs for the computation of the expected utility scores it needs to rank its alternatives. These summaries are associated both with the shape of the utility function and the form of the probability density. This formal analysis also enables us to identify a minimal set of independence assumptions that guarantee coherence in partially defined systems. We will show that it is often only necessary that some of the moments satisfy certain polynomial relationships and explore some of these properties that we will term *partial* and *moment independence* relationships.

We will present the methodology in a variety of examples, including standard Gaussian Bayesian networks, a class of asymmetric models called staged trees [7] and Bayes linear directed graphs [2]. In addition we will look at situations where the shape of the expected utility function informs the (IDSS) that panels will rarely come to an agreement on the course of action, since the optimal decision would only favour one group. We will term such situations *conflict models*.

# References


[1] S. Banerjee, B.P. Carlin and A.E. Gelfand (2014). Hierarchical modeling and analysis for spatial data. Crc Press.

[2] M. Goldstein and D. Wooff (2007). Bayes Linear Statistics, Theory & Methods. John Wiley & Sons.

[3] M.C. Kennedy and A. O'Hagan (2001). Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(3), 425-464.

[4] M. Leonelli and J.Q. Smith (2013). Using graphical models and multi-attribute utility theory for probabilistic uncertainty handling in large systems, with application to the nuclear emergency management. In Data Engineering Workshops (ICDEW), (pp. 181-192). IEEE.

[5] M. Leonelli and J.Q. Smith (2015). Bayesian decision support for complex systems with many distributed experts. (invited revision to AoOR)

[6] K.N. Papamichail and S. French (2013). 25 Years of MCDA in nuclear emergency management. IMA Journal of Management Mathematics, 24(4), 481-503.

[7] J.Q. Smith (2010). Bayesian decision analysis: principles and practice. Cambridge University Press.

[8] J.Q. Smith, M.J. Barons and M. Leonelli (2015). Coherent inference for integrated decision support systems. (in preparation)

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Conditional Independence Ideals With Hidden Variables

Johannes Rauh[1], Fatemeh Mohammadi[2]

[1] *York University, Toronto, Canada,* rauh@math.uni-hannover.de

[2] *Institute of Science and Technology, Austria,* fatemeh.mohammadi@ist.ac.at

Conditional independence (CI) constraints among a collection of discrete random variables can be formulated algebraically in terms of the determinants of certain $2 \times 2$-submatrices of the probability tensor. This allows to associate a corresponding CI ideal with each collection of CI statements, and implications among CI statements can be analyzed by studying these CI ideals. A primary decomposition often allows to understand the corresponding CI variety, which contains all probability distributions satisfying the CI constraints, and to decompose this set into irreducible subsets with a good statistical interpretation. Algebraically, the analysis is complicated by the fact that CI ideals need not be radical. Furthermore, not all irreducible components need to contain a probability distribution.

In the case of *saturated* CI statements, when all statements involve the same variables, the CI ideal is a binomial ideal, which greatly simplifies the analysis. A special class of such statements leads to *binomial edge ideals*, which are radical and which have a nice primary decomposition with a good statistical interpretation. The name of these ideals comes from the fact that they can be described by a graph, where each node corresponds to a column of a matrix of probabilities and where the edges describe the corresponding determinants.

Our goal is to study what happens in the presence of hidden variables; i.e. when taking the marginal distribution over the hidden variables. When some of the variables are considered as hidden, CI statements lead to polynomial equations that correspond to higher minors (i.e. determinants of larger submatrices) of the probability tensor. In this case, the polynomial equations do not give a full characterization of the marginal model (i.e. to characterize the marginal model, polynomial inequalities are needed), but nevertheless, information about the corresponding CI ideal helps to understand the marginal model.

In special cases, this leads to *determinantal hypergraph ideals*, in which a hypergraph determines which minors of a given matrix of probabilities are contained in the ideal. Such ideals have already been studied abstractly, without observing the relation to statistics. Unlike binomial edge ideals, determinantal hypergraph ideals are not necessarily radical, and so far, no satisfying description of the primary decomposition is known in the general case. Only under strong conditions on the hypergraph is it possible to generalize the methods used for binomial edge ideals. Instead of trying to generalize the old methods, another idea is to start with examples that are natural from the point of view of statistics. These examples correspond to large hypergraphs with many symmetries.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# The facets of the cut polytope and the extreme rays of cone of concentration matrices of series-parallel graphs

Liam Solus[1], Caroline Uhler[2], Ruriko Yoshida[1]

[1] *University of Kentucky, USA,* {liam.solus,ruriko.yoshida}@uky.edu

[2] *Institute of Science and Technology Austria, Austria,* caroline.uhler@ist.ac.at

For a graph $G$ with $p$ vertices the cone of concentration matrices consists of all real positive semidefinite $p \times p$ matrices with zeros in entries corresponding to missing edges of $G$. The extremal rays of this cone and their associated ranks have applications to matrix completion problems, maximum likelihood estimation in Gaussian graphical models in statistics, and Gauss elimination for sparse matrices. It is well-known that the extremal rays of this cone in the case of the cycle are either rank 1 or rank $p-2$. Similarly, the cut polytope of the cycle has facets of two distinct shapes. Using hyperplane translations and general duality theory of spectrahedra, we demonstrate that a facet of a fixed shape corresponds to an extremal ray of a fixed rank. This shows that, in the case of the cycle, the different facet shapes in the cut polytope identify the ranks of extremal rays in the cone of concentration matrices, and this correspondence arises from the cutsets defining the facets. More generally, we show that any series-parallel graph $G$ has the facet-ray identification property, that is, the normal vectors to the facets of the cut polytope identify extremal rays in the cone of concentration matrices of G.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Information Geometry and Algebraic Statistics on a finite state space and on Gaussian models

Giovanni Pistone[1], Luigi Malagò[2]

[1] *de Castro Statistics, Collegio Carlo Alberto, Moncalieri, Italy,* giovanni.pistone@carloalberto.org

[2] *Shinshu University, Japan, and INRIA Saclay - Ile-de-France, France* malago@shinshu-u.ac.jp

It was shown by C. R. Rao in a paper published 1945 that the set of positive probabilities on a finite state space $\{0,1,\dots,n\}$ is a Riemannian manifold in a way which is of interest for Statistics. It was later pointed out by Sun-Ichi Amari, that it is actually possible to define two other affine geometries of Hessian type on top of the classical Riemannian geometry. Amari gave to this new topic the name of Information Geometry. Information Geometry and Algebraic statistics are deeply connected because of the central place occupied by exponential families in both fields. The present course is focused mainly on Differential Geometry, but arguments from the theory of Toric Models will be important.

- Lecture 1 (Pistone) The Differential Geometry of the Simplex.
- Lecture 2 (Pistone) The differential Geometry of statistical models.
- Lecture 3 (Malagò) Applications to Optimization and Machine Learning.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Latent tree graphical models

Piotr Zwiernik[1]

[1] *University of Genova, Italy,* piotr.zwiernik@gmail.com

The aim of this short lecture course is to introduce various mathematical and statistical aspects of latent tree graphical models. The latent tree graphical model is a special type of a statistical graphical model. The associated graph is a tree, which gives a tractable model with a rich combinatorial structure. What makes this model more complicated and also more interesting is that some variables in the system are assumed to be latent (not observed). This adds modeling power but also leads to various statistical issues. For example the associated likelihood function is multimodal and its maxima often lie on the boundary of the parameter space (and hence they are not critical points of the likelihood function). Another important statistical problem is that these models may be not identifiable.

I will discuss the following related topics:

1. Trees, tree metrics and spaces of trees: basic graph-theoretic tree concepts, tree metrics and other tree spaces that arise naturally in the study of latent tree graphical models.
2. Latent tree graphical models: model definition, links to Bayesian networks and undirected graphical models on trees; identifiability and moment structure.
3. Tree inference and parameter estimation: overview of methods for learning the underlying tree structure which is of interest in many applications; the structural EM algorithm for the MLE estimation and other approximate methods.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Pearson's Crabs: Algebraic Statistics in 1894

C. Améndola Cerón

*Technische Universität, Berlin, Germany,* carlos.amendola@cims.nyu.edu

This is a work in progress that begins by revisiting Karl Pearson's 1894 seminal paper [1] where he took on the problem of estimating the parameters of a mixture model of two univariate Gaussians. His motivation was to explain the asymmetry observed in data measured from a population of Naples' crabs, believing it might have been due to the fact that two subpopulations were present in the sample. In order to solve this problem, Pearson uses the method of moments to obtain the following system of polynomial equations in the means $\mu_1$, $\mu_2$, the variances $\sigma_1^2$, $\sigma_2^2$ and the mixture proportions $\alpha_1$, $\alpha_2$:

$$\begin{aligned}
\alpha_1+\alpha_2 &= 1 && (4)\\
\alpha_1\mu_1+\alpha_2\mu_2 &= 0 && (5)\\
\alpha_1(\mu_1^2+\sigma_1^2)+\alpha_2(\mu_2^2+\sigma_2^2) &= m_2 && (6)\\
\alpha_1(\mu_1^3+3\mu_1\sigma_1^2)+\alpha_2(\mu_2^3+3\mu_2\sigma_2^2) &= m_3 && (7)\\
\alpha_1(\mu_1^4+6\mu_1^2\sigma_1^2+3\sigma_1^4)+\alpha_2(\mu_2^4+6\mu_2^2\sigma_2^2+3\sigma_2^4) &= m_4 && (8)\\
\alpha_1(\mu_1^5+10\mu_1^3\sigma_1^2+15\mu_1\sigma_1^4)+\alpha_2(\mu_2^5+10\mu_2^3\sigma_2^2+15\mu_2\sigma_2^4) &= m_5. && (9)
\end{aligned}$$

After considerable effort and cleverness, Pearson manages to eliminate variables from (4)-(9) to obtain a ninth degree polynomial in the single variable $x=\mu_1\mu_2$,

$$\begin{aligned}
&24x^9-28\lambda_4x^7+36m_3^2x^6-(24m_3\lambda_5-10\lambda_4^2)x^5-(148m_3^2\lambda_4+2\lambda_5^2)x^4+(288m_3^4-12\lambda_4\lambda_5m_3-\lambda_4^3)x^3\\
&\qquad+(24m_3^3\lambda_5-7m_3^2\lambda_4^2)x^2+32m_3^4\lambda_4x-24m_3^6=0, \qquad (10)
\end{aligned}$$

where $\lambda_4=9m_2^2-3m_4$ and $\lambda_5=30m_2m_3-3m_5$. After substituting his numerical data, he finds the real roots of this nonic in more than one example (a "heroic" task in his time), and determines if they can correspond to a solution for the mixture model.

Nevertheless, even though his modeling ideas would greatly impact scientific research, the amount of computational effort needed in Pearson's method led to various attempts to find alternatives that culminated in maximum likelihood methods such as the popular EM (Expectation Maximization) algorithm. It has not been until recently ([2],[3]) that Pearson's method of moments approach has been adapted to higher dimensions in a computationally efficient way and even proven to be optimal in a particular sense ([4]).

It is in this framework that we believe the present research direction becomes significant. Concretely, we can translate Pearson's original problem into a study of a special algebraic variety, and bring Pearson's work under an Algebraic Statistics light. Indeed, a first result is that computing a Gröbner basis for the suitable elimination ideal associated to (4)-(9) produces an irreducible 9th degree polynomial relation that coincides with Pearson's polynomial (10) in its full 30 term expansion in the $m_i$'s. It is also natural in this formulation to deal with the presence of multiple solutions (non-identifiability) or of none at all; two problems that Pearson faced. For the first, he proposes to discriminate by looking at the proximity to the sixth moment $m_6$; we can translate this to the study of the irreducible polynomial relation among the first six moments $m_1,\ldots,m_6$. For the second problem, Pearson gives an interesting interpretation in terms of homogeneity of the population and evolution, while we take it that he encountered a distribution that belongs to the secant statistical model but not to the mixture one.

In summary, unbeknownst to him, Pearson was working in an Algebraic Statistics framework, obtaining results and running into problems that may be expressed in modern language and techniques. Our main motivation is to do so and find what the new perspective offers in this context.

# References


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

[2] A. Kalai, A. Moitra, and G. Valiant (2010), Efficiently learning mixtures of two Gaussians, *STOC*, ACM, 553-562.

[3] M. Belkin and K. Sinha (2010), Polynomial learning of distribution families, *FOCS*, IEEE Computer Society, 103-112.

[4] M. Hardt and E. Price (2014), Tight bounds for learning a mixture of two Gaussians. Technical Report, *ArXiv* 1404.4997v2.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# The algebraic method in experimental designs

Y. Berstein[1], H. Maruri-Aguilar[2], S. Onn[3], E. Riccomagno[4], E. Sáenz de Cabezón[5], H. Wynn[6]

[1] *McCombie Lab, Cold Spring Harbor Laboratory NY, USA,* yberstei@cshl.edu
[2] *School of Mathematics, Queen Mary University of London, UK,* H.Maruri-Aguilar@qmul.ac.uk
[3] *Technion - Israel Institute of Technology, Haifa,* onn@technion.ac.il
[4] *Dipartimento di Matematica, Università degli Studi di Genova, Italy,* riccomag@dima.unige.it
[5] *Departamento de Matemáticas y Computación, Universidad de La Rioja, Spain, eduardo.saenz-de-cabezon@unirioja.es*
[6] *CATS, London School of Economics, UK,* h.wynn@lse.ac.uk

The algebraic approach to identify models in experimental designs represents a set of points $\mathscr{D} \subset R^k$ (design) by the polynomial ideal $I(\mathscr{D}) \subset R[x]$ generated by it. The monomial basis for the quotient $R[x]/I(\mathscr{D})$ corresponds to a statistical model that satisfies desirable properties such as hierarchy and identifiability [5].

This poster describes recent developments of our group on the relation between designs and identified models. Three topics form the core of it, they are descriptions of the designs based on

1. The graded degree of models associated with it. This degree of algebraic models is termed aberration and we give bounds for it [1].

2. The complexity of identifiable monomial models in terms of Betti numbers [3]. Our results coincide in some specific cases with lex-segment ideals that maximize Betti numbers [2].

3. The identifiability of models when the design $\mathscr{D}$ is considered as a subset of a bigger grid design $\mathscr{F}$ and thus it has a complementary (disjoint) fraction $\mathscr{D}'$ such that $\mathscr{D} \cup \mathscr{D}' = \mathscr{F}$. Models for complementary designs satisfy a general version of Alexander duality [4].

## References

[1] Y. Berstein, H. Maruri-Aguilar, S. Onn, E. Riccomagno and H. Wynn. (2010). Minimal average degree aberration and the state polytope for experimental designs. *Annals of the Institute of Statistical Mathematics*, **62**(4), 673-698.

[2] A. Bigatti (1993), Upper bounds for the Betti numbers of a given Hilbert function. *Communications in Algebra*, **21**(7), 2317-2334.

[3] H. Maruri-Aguilar, E. Sáenz de Cabezón and H. Wynn. (2012). Betti numbers of polynomial hierarchical models for experimental designs. *Annals of Mathematics and Artificial Intelligence*. **64**(4), 411-426.

[4] H. Maruri-Aguilar, E. Sáenz de Cabezón and H. Wynn (2013), Alexander duality in experimental designs, *Annals of the Institute of Statistical Mathematics* **65**, 667-686.

[5] G. Pistone, E. Riccomagno and H. Wynn (2001). *Algebraic Statistics*, Monographs on Statistics and Applied Probability, vol 89. Chapman & Hall/CRC, Boca Raton.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# A Statistical Package in CoCoA-5

A. M. Bigatti[1], M. Caboara[2]

[1] *Università degli Studi di Genova, Italy,* bigatti@dima.unige.it

[2] *Università degli Studi di Pisa, Italy,* caboara@dm.unipi.it

In this poster we present a software package for CoCoA (a system for Computations in Commutative Algebra) using Gröbner basis theory and some methods of Algebraic Geometry to solve a relevant problem in Statistics, more specifically in the Design of Experiments. Namely suppose we are given a Full Factorial Design $D$ and a complete polynomial model $P$, whose support is contained in the order ideal of monomials defined by $D$. We show how to construct families of ideals defining Fractions $F$ of $D$ which are minimally identified by $P$.

## References

[1] J. Abbott, A.M. Bigatti, *CoCoALib: a C++ library for doing Computations in Commutative Algebra*, Available at http://cocoa.dima.unige.it/cocoalib,

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

[3] M. Caboara, L. Robbiano (1997) *Families of Ideals in Statistics* Proceedings of the ISSAC97 Conference (Maui, Hawaii, July 1997) Kuchlin ed., New York, pp. 404-409

[4] L. Robbiano (1998) *Gröbner Bases and Statistic* in *Gröbner Bases and Applications. Proc. of the Conf. 33 Years of Gröbner Bases* (B.Buchberger and F.Winkler eds.) Cambridge University Press, London Mathematical Society Lecture Notes Series, 251.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Exploiting Symmetry in Characterizing Bases of Toric Ideals

I. Z. Burke[1]

[1] *National University of Ireland (NUI) Galway, Ireland,* i.burke1@nuigalway.ie

Every ideal $I \subseteq k[x_1,...,x_m]$ is invariant under all permutations of some subgroup $G \subseteq S_m$. Many toric ideals arising from statistical models possess rich symmetric structures (i.e. the corresponding $G$ is non-trivial and interesting). Here we take advantage of this structure in an attempt to fully characterize the various bases (Markov, Universal Gröbner, Graver) of a well-known family of toric ideals arising from independence models.

We denote by $I_{[2^n]}$ the ideal associated with the independence model of a $2 \times 2 \times \cdots \times 2$ ($n$ times) contingency table, see e.g. [1]. We show that this ideal is invariant under $G \cong BC_n$, where $BC_n$ is the hyperoctahedral group of dimension $n$. Of particular interest is the *Universal Gröbner Basis* of $I_{[2^n]}$, denoted $\mathscr{U}_{[2^n]}$, which we define to be the union of all reduced Gröbner Bases of $I_{[2^n]}$.

We say that $u = (u_1,...,u_{2^n}), v = (v_1,...,v_{2^n}) \in I_{[2^n]}$ have the same *combinatorial type*, or $u \sim v$, if $u \in \{\pm(s \cdot v)\}$ for some $s \in S_m$, where $m = 2^n$. It is shown that in general, $u \sim v$ does not imply $u \in G \cdot v$, where $G \cdot v$ denotes the orbit of $v \in I_{[2^n]}$ under the action of $G$. Using a geometric approach, we develop formulae for the number of distinct combinatorial types in a minimal Markov basis of $I_{[2^n]}$ and the number of distinct minimal Markov bases of $I_{[2^n]}$ in general.

Motivated by a theorem of Sturmfels [2], we use the programs `4ti2` and `polymake` to examine the fiber polytopes of the $n$-way independence model. Provisional results are included and we make a number of conjectures about the general structure of $\mathscr{U}_{[2^n]}$.

## References

[1] S. Aoki, H.Hara and A. Takemura (2012), *Markov Bases in Algebraic Statistics*, Springer Verlag, New York.

[2] B. Sturmfels (1996), *Gröbner Bases and Convex Polytopes*, AMS University Lecture Series (Vol. 8), Providence, Rhode Island.

# Bayesian computation for exponential random graph models

A. Caimo

*University of Lugano, Switzerland,* alberto.caimo@usi.ch

Recent research in statistical social network analysis has demonstrated the advantages and effectiveness of probabilistic approaches to network data. In fact, Bayesian methods are becoming increasingly popular as techniques for modeling social networks [1, 2, 3].

The likelihood of exponential random graph models [4] represents the probability distribution of a network graph $y$ and can be expressed as:

$$p(y|\theta) = \frac{\exp\{s(y)^t\theta\}}{z(\theta)} \tag{11}$$

This equation states that the probability of observing a given network graph $y$ is equal to the exponent of the observed graph statistics $s(y)$ multiplied by parameter vector $\theta$ divided by a normalising constant term $z(\theta)$. The latter is calculated over the sum of all possible graphs on $n$ nodes and it is therefore extremely difficult to evaluate for all but trivially small graphs.

Following the Bayesian paradigm, prior distribution is assigned to $\theta$. The posterior distribution of $\theta$ given the data $y$ is:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}. \tag{12}$$

Direct evaluation of $p(\theta|y)$ requires the calculation of both the likelihood $p(y|\theta)$, which is computationally demanding if not intractable, and the marginal likelihood $p(y)$ which is typically intractable.

In this poster we present some Monte Carlo strategies for doubly intractable distributions which improve the efficiency of Bayesian methods for exponential random graph models and increase their scalability to large network graphs. The analysis is carried out using the `Bergm` package for `R` [5].

## References

[1] J. Koskinen, G. Robins and P. Pattison (2010), Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. *Statistical Methodology*, **7**, 366 – 384.

[2] A. Caimo and N. Friel (2011), Bayesian Inference for Exponential Random Graph Models, *Social Networks*,**33**, 41 – 55.

[3] A. Caimo and A. Mira (2015), Efficient Computational Strategies for Doubly Intractable Problems with Applications to Bayesian Social Networks, *Statistics and Computing*, **25**, 113 – 125.

[4] D. Lusher, J. Koskinen, and G. Robins (2012), *Exponential random graph models for social networks: Theory, methods, and applications.* Cambridge University Press.

[5] A. Caimo and N. Friel (2014), Bergm: Bayesian Exponential Random Graphs in R, *Journal of Statistical Software*, **61**(2), 1 – 25.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Degree bounds on tree models

J. Draisma[1], R.H. Eggermont[1]

[1] *Eindhoven University of Technology, The Netherlands,* {j.draisma,r.h.eggermont}@tue.nl

Tree models are families of probability distributions used in modelling the evolution of a number of extant species from a common ancestor. Such a model can be constructed as follows. Let $G$ be a finite group. Let $T$ be a finite rooted tree, and attach to each of its vertices $q$ an alphabet $B$, acted upon by $G$. To the root $r$ of $T$, attach a probability distribution $\pi$ on $B$, and to each edge $q \to q'$ of $T$, attach a $B \times B$-matrix $A_{qq'}$ for which the $(b,c)$-th entry should be interpreted as the probability that letter $b$ mutates into letter $c$. This gives rise to a probability distribution on $B^{\mathrm{leaf}(T)}$ by

$$P(\mathbf{b}) = \sum_{\mathbf{b}' \in B^{\mathrm{vert}(T)} \text{ extending } \mathbf{b}} \pi(b'_r) \cdot \prod_{q \to q' \in \mathrm{edge}(T)} A_{qq'}(b'_q, b'_{q'}). \tag{13}$$

The set of probabilities thus obtained in $\mathbb{R}^{\mathrm{leaf}(T)}$ (or even $\mathbb{C}^{\mathrm{leaf}(T)}$) is called the equivariant model for the triple $(T,B,G)$. Its Zariski closure is an object of interest in algebraic statistics, and finding the ideal of this variety, or even just a set of defining equations, can be quite difficult in general.

Sturmfels and Sullivant [SS05] conjecture that if $G$ is Abelian and $G$ acts transitively on $B$, then the ideals of the equivariant model assigned to the star with $n$ leaves are defined in degree $|G|$, independent of $n$. In this situation, Michałek [Mic13] has shown that for general trees, there exists a degree bound for the associated projective schemes, though he still requires that $G$ acts transitively on $B$. We will show the following theorem, which is weaker than both the conjecture and Michałek's result in the sense that it is a set-theoretical result and that the degree bound is unknown, but which does allow for more freedom in the trees and the alphabets.

**Theorem 3** *For a fixed finite alphabet $B$ and a fixed Abelian group $G$ with a fixed action on $B$, there exists a uniform bound $D = D(B,G)$ such that for any finite tree $T$ the Zariski closure of the equivariant tree model for $(T,B,G)$ is defined by equations of degree at most $D$.*

We show this theorem by working in a slightly more general setting, in which each vertex can be associated to a different alphabet. We do this by first reducing to the case of star trees with the same alphabet attached to each leaf (following along the lines of Draisma and Kuttler [DK09]), and then constructing a projective limit of these trees. In this infinite-dimensional setting, we will prove Noetherianity up to symmetry, and we will use this to show that up to symmetry, finitely many equations are needed to cut out the variety in finite dimension.

This talk is based on [DE14].

# References

[DE14] Jan Draisma and Rob H. Eggermont (2014), Finiteness results for Abelian tree models, J. Eur. Math. Soc.. To appear; preprint available from http://arxiv.org/abs/1207.1282.

[DK09] Jan Draisma and Jochen Kuttler (2009), On the ideals of equivariant tree models, *Math. Ann.*, **344**, 619-644.

[Mic13] Mateusz Michałek (2013), Constructive degree bounds for group-based models, *J. Comb. Theory, Ser. A*, **120**,1672-1694.

[SS05] Bernd Sturmfels and Seth Sullivant (2005), Toric ideals of phylogenetic invariants, *Journal of Computational Biology*, **12**, 204-228.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Optimality criteria and geometry of fractional factorial designs

Roberto Fontana[1], Fabio Rapallo[2], Maria Piera Rogantin[3]

[1] *Politecnico di Torino, Italy,* roberto.fontana@polito.it
[2] *Università del Piemonte Orientale, Alessandria, Italy,* fabio.rapallo@unipmn.it
[3] *Università di Genova, Italy,* rogantin@dima.unige.it

The geometric structure of the fraction plays a prominent role for classifying fractional factorial designs. The characterization of the geometric structure of a fraction can be made through combinatorial invariants associated to the model matrix of the full factorial design.

A first result in this direction concerns the case of saturated fractions, i.e., fractions with as many points as the number of parameters, and such that all the parameters are estimable and no degrees of freedom remain to estimate the error term. Consider a model with $p$ parameters and with model matrix $X$, and compute the circuits $f_1, \ldots, f_L$ of $X^t$. In [1] and [2] the author proved that a fraction with exactly $p$ points is saturated is and only if it does not contain any of the supports of the circuits $f_1, \ldots, f_L$.

Further investigations have been performed in order to check the connections between the geometric structure of a fraction and some optimality criteria. In small cases, where the enumeration of all saturated fraction is possible, we have compared the $D$-efficiency of the fractions with its geometric structure. In particular, for a given fraction $\mathscr{F}$ we have computed the quantities

$$b_i = \#\mathrm{supp}(f_i)\,, \qquad\qquad b_{i,\mathscr{F}} = \#(\mathrm{supp}(f_i) \cap \mathscr{F})$$

for $i = 1, \ldots, L$, and we have summarized the results through the following indices:

$$g_2(\mathscr{F}) = \sum_{i=1}^{L} (b_i - b_{i,\mathscr{F}})^2\,, \qquad\qquad g_3(\mathscr{F}) = \max_i (b_{i,\mathscr{F}})\,.$$

As illustrated in [3], experiments and simulations show that $D$-optimal fractions present the highest values of the indices $g_2$ and $g_3$.

Finally, under the point of view of model-free optimality criteria, the complex indicator function of a fraction is a major tool for the computation of the Generalized Word-Length Pattern. The complex coding and the corresponding indicator function are powerful ingredients to define properly Orthogonal Arrays in the multi-level case. Indeed, the expression of the aberration depends only on the level counts and does not involve explicit computations with complex numbers.

## References

[1] R. Fontana, F. Rapallo and M.P. Rogantin (2014), A Characterization of Saturated Designs for Factorial Experiments. *J. Statist. Plann. Inference*, **147**, 204-211.

[2] R. Fontana, F. Rapallo and M.P. Rogantin (2014), Two factor saturated designs: Cycles, Gini index and state polytopes. *J. Stat. Theory Pract.*, **8**(1), 66-82.

[3] R. Fontana, F. Rapallo and M.P. Rogantin (2014). $D$-optimal saturated designs: A simulation study. In *Topics in statistical simulation* (V.B. Melas, S. Mignani, P. Monari, L. Salmaso eds.), Springer Verlag, New York, 183-190.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Symbolic Computations for Defining Diffusion Processes on Torus Using Dihedral Angles Coordinates

M. Golalizadeh[1], M. Rahimi[2]

[1] *Department of Statistics, Tarbiat Modares University, Tehran, Iran,* golalizadeh@modares.ac.ir

[2] *Department of Statistics, Tarbiat Modares University, Tehran, Iran,* milad-tmu@yahoo.com

There are two important angles to determine the structure of protein. They are usually written as $\phi$ and $\psi$, where both range in $[0,2\pi]$, and the loci of them are points on torus. They are called dihedral angles and have already been used to plot the structure of proteins on the plane, well known as the Ramachandran plot (see, e.g. [2]). Along with this, a recent application of statistics in biological sciences is concerned about studying bivariate distributions to describe the joint variability of the dihedral angles. Among many densities the bivariate von Mises distributions play a key role. The idea of these densities have been appeared first in [7]. But, two well known distributions on torus, sine and cosine models and their statistical properties were comprehensively studied in [2]. Thereafter, these models have been extended in other directions including multivariate cases. Particularly, the multivariate sine model has been studied in [1]. Recently, generalized multivariate sin model distribution has also been investigated in [3].

Mathematically, a torus with big and small radii $R$, and $r$, respectively, can be described with the parametric from,

$$x=(R+r\cos\phi)\cos\psi,\quad y=(R+r\cos\phi)\cos\psi,\quad z=r\sin\psi. \tag{14}$$

Our aim in this paper is to derive the structure of standard Brownian motion on torus. There is a method for obtaining the infinitesimal generator of a Brownian motion on a manifold by using the metric tensor. In fact, it uses a second order differential operator, called the Laplace-Beltrami operator. It is usually defined as " div grad " in textbooks (see, e.g. [4], §4.3 and [5], pp. 256-270) and can be used to obtain the infinitesimal parameters of a diffusion process such as Brownian motion on a Riemannian manifold. Generally, the Laplace-Beltrami operator on a manifold $\mathbf{M}$ with metric tensor $G=[g_{ij}]$, for any $i,j=1,\ldots,n$, and the coordinates $u=(u_1,\ldots,u_n)^T$, is defined by

$$\triangle f=\frac{1}{\sqrt{\det(G)}}\sum_{i=1}^{n}\frac{\partial}{\partial u_i}\Big(\sqrt{\det(G)}\sum_{j=1}^{n}g^{ij}\frac{\partial f}{\partial u_j}\Big), \tag{15}$$

where $g^{ij}$ is the $(i,j)$ entry of $G^{-1}$, the inverse matrix of $G$, and $f$ is a function satisfying some conditions. Interestingly, there is a closed relationship between this operator and the elements of a standard Brownian motion on manifold. Following [6], the infinitesimal generator of standard Brownian motion on a manifold $\mathbf{M}$ with metric tensor $G=[g_{ij}]$ is given by one half of the Laplace-Belterami operator on $\mathbf{M}$. Hence, we can derive the infinitesimal drift and diffusion coefficients of Brownian motion on a manifold $\mathbf{M}$ with coordinates $u=(u_1,\ldots,u_n)^T$, for any $i,j=1,\ldots,n$, as, respectively,

$$\mathrm{drift}(\mathrm{d}(u_i)_t)=\mu_i(u)\mathrm{d}t=\frac{1}{2\sqrt{\det(G)}}\sum_{j=1}^{n}\frac{\partial}{\partial u_j}\Big(\sqrt{\det(G)}g^{ij}\Big)\mathrm{d}t \tag{16}$$

and

$$\mathrm{d}(u_i)_t\mathrm{d}(u_j)_t=(\sigma\sigma^T)_{ij}\mathrm{d}t=g^{ij}\mathrm{d}t. \tag{17}$$

Recalling Eq. (14) and using simple computations, particularly algebraic calculus, it can be shown that the tensor metric for the points on this torus is given as

$$G = \begin{pmatrix} (R + r\cos\phi)^2 & 0 \\ 0 & r^2 \end{pmatrix}.$$

Now, we can derive the infinitesimal drifts for $\phi_t$ and $\psi_t$ using $G^{-1}$ and Eq. (16). It is a bit of calculus to show that

$$\text{drift}(\mathrm{d}\phi_t) = 0, \qquad \text{drift}(\mathrm{d}\psi_t) = \frac{-\sin\psi_t}{2r(R + r\cos\phi_t)}\mathrm{d}t. \tag{18}$$

Similarly, those infinitesimal coefficients are obtained using $G^{-1}$ and Eq. (17). Particularly, we have

$$\mathrm{d}\phi_t \mathrm{d}\psi_t = \begin{pmatrix} \frac{1}{(R+r\cos\phi_t)^2} & 0 \\ 0 & \frac{1}{r^2} \end{pmatrix} \mathrm{d}t. \tag{19}$$

So, a general form of an Stochastic Differential Equation (SDE) of the standard Brownian motion on torus using the dihedral angles, as the local coordinates, can be written as

$$\begin{pmatrix} \mathrm{d}\phi_t \\ \mathrm{d}\psi_t \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{-\sin\phi_t}{2r(R+r\cos\phi_t)} \end{pmatrix} \mathrm{d}t + \begin{pmatrix} \frac{1}{R+r\cos\phi_t} & 0 \\ 0 & \frac{1}{r} \end{pmatrix} \begin{pmatrix} \mathrm{d}B_1(t) \\ \mathrm{d}B_2(t) \end{pmatrix}, \tag{20}$$

where $B_t = (B_1(t), B_2(t))^T$ represents the two dimensional Brownian motion on the Euclidean plane. To simulate the data following the standard Brownian motion on torus, identified by the dihedral angles, one can utilize the Euler discretized version of the SDE in Eq. (20).

To derive our manipulations in the above representation, we benefited from symbolic computations in both Maple and Mathematica. Particularly, a combination of algebraic statistics and differential geometry along with visual representations were employed to validate our results.

It is known that any standard Brownian motion defined on a manifold will eventually reach to its equilibrium provided the corresponding diffusion process are satisfied on some criteria. So, our objective is to derive the equilibrium density of the SDE given by Eq. (20). To obtain a density from multidimensional SDE is a tough job, unless it is believed that the components describing the diffusion processes are mutually independent. This is not the case for our SDE given by Eq. (20) since the drift of $\psi_t$ depends on the stochastic process $\phi_t$. Another common procedure is to evaluate the densities, already defined on the corresponding manifold, to be satisfied in the Kolmogorov forward equation. Because our motivating example comes from biological sciences, we were conducted to review those popular and relevant densities in these contexts.

We have checked the sine and cosine models using the SDE given by Eq. (20) and Kolmogorov forward equation. However, the results were not promising. So, it sounds that we should seek the other densities on torus which have not been derived yet! This is our ongoing research.

# References


[1] K. V. Mardia, G. Huges, C. C. Taylor and H. Sing (2008), A multivariate von Mises Distribution with Applications to bioinformatics, *Canadian Journal of Statistics*, **36**, 99-109.

[2] K. V. Mardia, C. C. Taylor and G. K. Subramaniam (2007), Protein Bioinformatics and Mixtures of Bivariate von Mises Distributions for Angular Data, *Biometrics*, **63**, 505-512.

[3] K. V. Mardia and J. Voss (2014), Some Fundamental Properties of Multivariate von Mises Distribution, *Communication in Statistics -Theory and Methods*, **43**, 1132-1144.

[4] H. P. McKean (1969), *Stochastic Integrals*, Academic Press, New York.

[5] P. A. Meyer (1966), *Probability and Potentials*, Blaisdell, Waltham, Mass.

[6] B. Øksendal (1998), *Stochastic Differential Equation, An Introduction with Applications*, Springer-Verlag, Berlin.

[7] H. Singh, V. Hnizdo and E. Demchuk (2002), Probabilistic Model for Two Dependent Circular Variables, *Bioinformatics*, **89**, 719-723.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Equivalence Classes of Chain Event Graph Models

Ch. Görgen[1], Jim Q. Smith[1]

[1] *University of Warwick, UK,* {c.gorgen, j.q.smith}@warwick.ac.uk

The Chain Event Graph (CEG) is a recently developed graphical statistical model based on an event tree [1]. It provides a very compact graphical description of the unfolding of events in a discrete setting.

CEGs have been successfully applied in a whole range of areas of statistical inference. They are more expressive and less redundant than tree representations. For instance, [2] list separation theorems to read all sorts of conditional independence statements of functions of variables from a given CEG, [3] and [4] found advantages in model selection and learning, [5] states propagation algorithms and [6] as well as [7] examined causal interpretations.

We are currently working on a characterisation of *statistically equivalent* CEGs. This is an important concept for three reasons. The first is computational: CEGs constitute a massive model space. By identifying a single representative within an equivalence class we significantly reduce the search effort across that space. The second is coherence: for Bayesian model selection, [8] and others have argued that two statistically equivalent models (always giving the same likelihood) should be given the same prior. Otherwise in Bayes factor model selection, one model will be given spuriously preference over another. Thus, it is critical to know when two CEGs make the same distributional assertions—that is, are statistically equivalent. The third reason is inference: causal discovery algorithms can be applied to CEGs just as well as BNs to discover putative causal ordering [7]. However to do this we need to know there is an unambiguous causal ordering inferred from a given dataset. Just as for Bayesian networks (BN) to confirm this we need to know all elements in a selected model class have the same directionality.

Notably, the class of CEG models contains BNs as a subclass. Now, different BNs which are in the same equivalence class (in the sense of [8]) always share some of their topological structure. Namely, they have the property that they can be characterised as the ones that share the same *pattern*. Consequently, a mixed graph can act as a representation for a class of equivalent BNs. However, sadly no such elegant common representation is available for the much larger class of CEGs. In fact, we know that even two equivalent CEGs which are also equivalent to some BN representation do not necessarily have distinguishable topological characteristics in common and their graphs can look quite different.

Nevertheless, [2] examine the implied conditional independence properties in various subclasses of the CEG class and link these to the graphical description of a model within such a subclass. However, even in these restricted settings, the graph is not sufficiently informative to be able to function as a representative of an equivalence class.

Our aim in this presentation is to algebraically—as opposed to graphically—characterise statistical equivalence in the context of CEG models.

We note that in the proofs in [5], it can be seen that in fact it was sets of *polynomial equations* associated with atomic probabilities—which are monomials in a certain polynomial ring determined by the model—that yielded powerful propagation theorems and it was only subsequently that these were translated into a graphical syntax. Similarly, in [6] and [9], it was again the algebraic properties of certain polynomials implicit in a CEG which led to various causal implications to be proved. This is consistent: In most of the literature surrounding CEG models, we notice that proofs always fall back on an algebraic description of the model class.

We will therefore investigate the efficacy of applying algebraic methods directly to characterise equivalent CEGs. We note that a characterisation in algebraic terms has already been extremely successful in the study of properties of Bayesian networks even when this class enjoys the pattern property mentioned above. For the class of CEGs the reasoning centres mainly around the construction of a polynomial which equals the sum of atomic probabilities in a model, which we call the *interpolating polynomial*, after [10]. We will see in the following that this function not only captures the graphical structure it is defined on but also yields information about conditional independence readable from a CEG and about the class of probability models containing a given event tree.

Our ansatz is to start from well-understood equivalence classes, in particular those of discrete decomposable Bayesian networks, and use their structure to understand CEG equivalence classes within this new framework. In this context, event trees will be used as a vehicle to switch between different representations. We report on some recent results.

# References


[1] J. Q. Smith and P. E. Anderson (2008), *Conditional independence and Chain Event Graphs*, Journal of Artifical Intelligence, vol. 172, pp. 42–68

[2] P. A. Thwaites and J. Q. Smith, (2011)*Separation theorems for Chain Event Graphs*, CRiSM, vol. 11-09.

[3] G. Freeman and J. Q. Smith, *Bayesian MAP model selection of Chain Event Graphs*, CRiSM, vol. 09-06.

[4] L. M. Barclay, J. L. Hutton, and J. Q. Smith (2013), *Refining a Bayesian Network using a Chain Event Graph*, International Journal of Approximate Reasoning, vol. 54, pp. 1300–1309.

[5] P. A. Thwaites, J. Q. Smith, and R. G. Cowell (2008), *Propagation using Chain Event Graphs*, in Proceedings of the 24th Conference on Uncertainty in Artifical Intelligence, (Helsinki), pp. 546–553.

[6] P. A. Thwaites, J. Q. Smith, and E. Riccomagno (2010), *Causal Analysis with Chain Event Graphs*, Journal of Artificial Intelligence, vol. 174, pp. 889–909.

[7] R. G. Cowell and J. Q. Smith (2014), *Causal discovery through MAP selection of stratified Chain Event Graphs*, Electronic Journal of Statistics, vol. 8, pp. 965–997.

[8] D. Heckerman (1998), *A Tutorial on Learning with Bayesian Networks*, pp. 301–354. The MIT Press.

[9] E. Riccomagno and J. Q. Smith (2009), *The Geometry of Causal Probability Trees that are Algebraically Con- strained*, Optimal Design and Related Areas in Optimization and Statistics, vol. 28, pp. 133–154.

[10] G. Pistone, E. Riccomagno (2001), and H. P. Wynn, *Gröbner bases and factorisation in discrete probability and Bayes*, Statistics and Computing, vol. 11, pp. 37–46.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Intrinsic and extrinsic means and curvature of metric cones

Kei Kobayashi[1,2], Henry P. Wynn[3]

[1] *The Institute of Statistical Mathematics, Tokyo, Japan,* kei@ism.ac.jp
[2] *JST PRESTO*
[3] *London School of Economics, UK,* h.wynn@lse.ac.uk

When the support of a data distribution is restricted to a path-connected subset $\mathscr{M}$, called "data space", of the whole Euclidean space $\mathbb{E}^d$, neither the population mean nor sample mean is necessarily in the data space. If some "mean" in the data space is required, the intrinsic mean can be the first option and that is defined by

$$\hat{\mu} = \arg\min_{m \in \mathscr{M}} \sum_{i=1}^{n} d(X_i, m)^2$$

for samples $X_1, \ldots, X_n \in \mathscr{M}$. Here $d(\cdot,\cdot)$ is the geodesic distance in $\mathscr{M}$. The intrinsic mean is not necessarily unique but some sufficient conditions for the uniqueness have been studied by using the curvature or CAT(k) property of the data space.

To compute the intrinsic mean of a given data set, we need to calculate the geodesic distances but that is difficult for complicated data spaces. For such cases, the extrinsic mean using the embedding Euclidean metric can be the second option:

$$\hat{\mu} = \arg\min_{m \in \mathscr{M}} \sum_{i=1}^{n} \|X_i - m\|^2.$$

The extrinsic mean is again not necessarily unique. In this presentation, we consider a variation of the intrinsic mean

$$\hat{\mu} = \arg\min_{m \in \mathscr{M}} \sum_{i=1}^{n} g(d(X_i, m))^2$$

with a non-decreasing concave function $g$ such that $g(0) = 0$. For such a function $g$, $\tilde{d}(\cdot,\cdot) := g(d(\cdot,\cdot))$ becomes a distance function but not necessarily a geodesic distance function. Since the curvature and CAT(k) properties are usually defined only for geodesic metric spaces, change of the metric $d$ by $g$ prevents arguments of the uniqueness of the mean in the aspect of the curvature of the data space. In order to solve this problem, we propose the following concave function $g_\beta$ with a parameter $\beta > 0$:

$$g_\beta(z) = \begin{cases} \sin(\frac{\pi z}{2\beta}), \text{ for } 0 \leqslant z \leqslant \beta, \\ 1, \text{ for } z > \beta. \end{cases}$$

With this specific concave function, the intrinsic mean becomes equivalent to a variety of extrinsic mean for which the data space is embedded into a metric cone but not into the Euclidean space. The parameter $\beta$ corresponds to the distance from the origin of the cone to the embedded data space and the CAT(k) property of the metric cone can be controlled by $\beta$. In application, data spaces are often embedded into the Euclidean space with a much higher dimension, but the embedding metric cone is only one-dimensional higher than the data space.

We propose the intrinsic means and the corresponding variances by using the new distance defined by $g_\beta$ above and show examples of their applications to some real data. We also present some relations between the metric-cone embedding and the correlation function. This presentation is mainly based on our preprint arXiv:1401.3020 [math.ST] and some more resent results.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Numerical Algebraic Fan of a Design for Statistical Model Building

Nikolaus Rudak[1], Sonja Kuhnt[1], Eva Riccomagno[2]

[1] *Department of Computer Science, Dortmund University of Applied Sciences and Arts, Germany,* sonja.kuhnt@fh-dortmund.de, rudak@statistik.tu-dortmund.de

[2] *Department of Mathematics, University of Genova, Italy,* riccomag@dima.unige.it

The identification of experimental designs by a polynomial ideal has long been explored in algebraic statistics. A key result being that features and properties of the ideal provide insight into the structure of models identifiable by the design [3, 5]. Holliday et al. [3] for example apply this approach to an incomplete standard factorial design in the automotive industry, Bates et al. [1] search for good polynomial meta-models for computer experiments.

Here, we treat a problem of model identifiability in a two-stage process, where observations or predictions from a well-chosen experimental design are themselves input variables for an eventual output of interest. Our work is motivated throughout by a thermal spraying process for which different modeling strategies are compared. Figure 3 depicts the two-stage process.

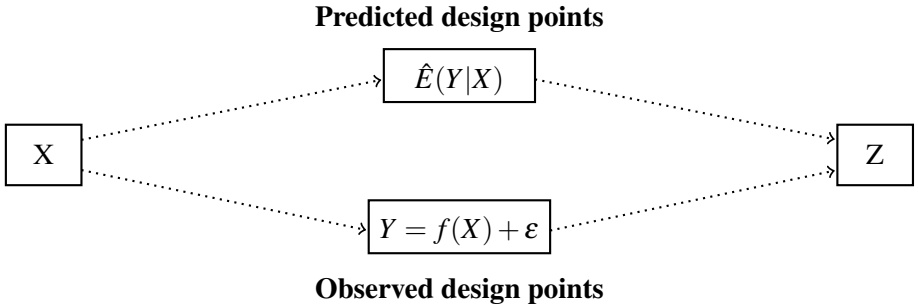



Figure 3: Occurence of noisy design in two-stage process

In a thermal spraying process spray material is melted up in a spraying gun and a gas stream accelerates the heated particles towards a surface which is to be coated [7]. A number of process parameters, such as the amount of used kerosene, can be set to chosen values. These parameters are summarised in the vector $X$, for which a full factorial design with a center point is used as experimental design $D_X$.

Properties of the particles in flight, such as their velocity and temperature, are measured as intermediate output, denoted by the random vector $Y$. Either these observations themselves ($D_Y$) or predicted values from models built between $X$ and $Y$ ($D_{\hat{Y}}$) then present the design for the eventually observed responses $Z$. Note, that our current interest lies in identifiable models from $Y$ to $Z$. The models treated in both parts are from the class of linear models.

The design of interest, either $D_Y$ or $D_{\hat{Y}}$, may consist of noisy points, such that the class of identifiable models from $Y$ to $Z$ determined by the standard approach might be unstable. More specifically, a model might only be identifiable due to a small deviation from a more regular design. We solve this problem by switching from symbolic, exact computation to numerical computations in the calculation of the design ideal and of its fan. We employ an algorithm from Fassino [2], whose use in statistics is new. The design is identified by a set of polynomials which "almost vanish" at the design points. Thereby we derive a procedure to construct (numerical) statistical fans [6].

For the thermal spraying application algebraic fans based on three standard term orders (lexicographical, degree lexicographical and reverse degree lexicographical ordering) are calculated for different predicted designs $D_{\hat{Y}}$ as well as the observed design $D_y$. Furthermore, the order of the main factors is permuted, such that to each design we get a (subset of its) fan. We compare the leaves within each subfan by looking at the most frequent monomials.

Coming back to statistical modelling the leafs are used as saturated models for a forward and backward selection based on the AIC criterion. The resulting models are compared to those derive from a models selection based on a second-order polynomial model in terms of AIC and $R^2$ values. We find, that we achieve a much improved model selection due to the enhanced knowledge of the space of identifiable models.

# References


[1] R.A. Bates, B. Giglio and H.P. Wynn (2003), A global selection procedure for polynomial interpolators, *Technometrics*, **45**, 246-255.

[2] C. Fassino (2010), Almost vanishing polynomials for sets of limited precision points, *J. Symbolic Comput.* **44**, 19-37.

[3] T. Holliday, G. Pistone, E. Riccomagno and H.P. Wynn (1999), The application of computational algebraic geometry to the analysis of designed eyperiments: a case study, *Comput. Statisti.*, **14**, 213-231.

[4] G. Pistone, E. Riccomagno and H.P. Wynn (2000), *Algebraic Statistics: Computational Commutative Algebra in Statistics*, Chapman & Hall/CRC, Bota Raton.

[5] E. Riccomagno (2009). A short history of algebraic statistics. *Metrika*, **69**, 397-418.

[6] N. Rudak, S. Kuhnt, W. Riccomagno (2013). Numerical algebraic fan of a design for statistical model building. *SFB 823 Discussion Paper 4/13, TU Dortmund University, Dortmund, Germany.*

[7] W. Tillmann, E. Vogli, B. Hussong, S. Kuhnt and N. Rudak (2010) *Relations between in flight particle characteristics and coating properties by HVOF spraying*, Proceedings of ITSC 2010 Conference, 2010.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Persistence of terms in lasso

Hugo Maruri-Aguilar[1], Simon Lunagómez[2]

[1] *School of Mathematics, Queen Mary University of London, UK,* H.Maruri-Aguilar@qmul.ac.uk
[2] *Department of Statistical Science, University College London, UK,* lunagomez@fas.harvard.edu

The lasso methodology has recently atracted attention in the context of models with hierarchy restrictions. In these models, an interaction term is allowed only if both main effects are active (strong hierarchy) or if at least one main effect is active (weak hierarchy). For example, under strong hierarchy appearance of the term $x_1x_2$ in a model requires both $x_1$ and $x_2$, while under weak hierarchy at least one of $x_1, x_2$ is needed. Recent approaches to hierarchical lasso include convex relaxation of the problem [1, 2] and nonlinear parameterization and Bayesian proposals [4, 5].

Our poster aims to describe the evolution of model terms as the parameter of regularization takes different values. As model terms become active or inactive, the structure of interactions changes. We describe the model in terms of components and cycles, borrowing from recent developments in computational topology in the area of persistent homology [3]. In our setting, components are groups of variables that interact, whereas cycles describe higher order interactions that are not currently included in the model. In the poster we briefly describe the elements from statistics and computational topology involved and present and discuss preliminary results.

# References

[1] J. Bien, N. Simon and R. Tibshirani (2013), Convex hierarchical testing of interactions, accepted in *Annals of Statistics*.

[2] J. Bien, J. Taylor and R. Tibshirani (2013), A lasso for hierarchical interactions, *Annals of Statistics*, **41**(3), 1111-1141.

[3] G. Carlsson (2009), Topology and data, *Bulletin of the American Mathematical Society*, **46**(2), 255-308.

[4] N. H. Choi, W. Li and J. Zhu (2010), Variable selection with the strong hieredity constraint and its oracle property, *Journal of the American Statistical Association*, **105**(489), 354-364.

[5] H. Noguchi, Y. Ojima and S. Yasui (2015), Bayesian lasso with effect heredity principle, In *Frontiers in Statistical Quality Control* (S. Knoth and W. Schmid eds.), Springer-Verlag, Berlin.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Algebraic Representation of Gaussian Model Combinations

M.S. Massa[1], E. Riccomagno[1],

[1] *Department of Statistics, University of Oxford, United Kingdom,* massa@stats.ox.ac.uk
[2] *Universita di Genova, Italy,* riccomag@dima.unige.it

Markov combinations of Gaussian models build a new model incorporating the constraints imposed by each initial model and imposing conditional independence between variables not jointly observed. Such problems correspond to finding a positive definite completion of the covariance/concentration matrix of all the variables of interest. The missing component is the sub-matrix corresponding to the random variables not jointly observed. We will firstly represent the covariance and concentration matrix of the combinations in term of matrix operations. Then we will translate the matrix description into an algebraic geometry setting and derive explicit forms of the combinations within this framework. In particular, a representation of each combination will be given in terms of polynomial ideals. We will show the usefulness of this approach by looking at the combination of Gaussian graphical models with equality and inequality constraints imposed by conditional independence relations, stationary constraints or positive definiteness requirements, for example. We will conclude with an illustrative example based on real data.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Conditional independence relations in Gaussian DAG models

Fatemeh Mohammadi[1]

[1] *IST Austria, Austria ,* fatemeh.mohammadi@ist.ac.at

Let $G$ be a directed acyclic graph (DAG) on the vertex set $[m]$. We associate to each vertex $i$ of $G$ a Gaussian random variable $X_i$, and let $\Sigma$ be the corresponding covariance matrix. A conditional independence statement $X_A \perp\!\!\!\perp X_B | X_C$ holds if and only if the set $C$ $d$-separates the sets $A$ and $B$ in $G$, or equivalently, the rank of the submatrix $\Sigma_{A\cup C, B\cup C}$ is equal to the cardinality of $C$. However, there exist additional subdeterminants of $\Sigma$ that are zero, but do not correspond to conditional independence relations associated to $G$.

In [2], Sullivant-Talaska-Draisma present a combinatorial characterization (in terms of treks) of subdeterminants of the covariance matrix $\Sigma$ corresponding to $d$-separation. In [3], Uhler-Raskutti-Bühlmann-Yu obtain a combinatorial characterization for *all* subdeterminants of the concentration matrix $\Sigma^{-1}$ in terms of the edge parameters of $G$.
Now take a collection $S = \{C_1, \ldots, C_r\}$ of conditional independence relations on the random variables $\{X_\ell : \ell \in [m]\}$ that are not necessarily coming from $G$. We are interested in finding maximal sets $S$ of conditional independence relations that do not force any of the edge parameters to be equal to zero. For each $C_i$ we denote by $p_{C_i}$ the corresponding subdeterminant of $\Sigma^{-1}$. Now we study a similar question as in [2], [3]: Does the saturation of the ideal generated by $p_{C_i}$ (with respect to the product of the edge parameters) contain any additional subdeterminant? In particular, we are interested in finding a combinatorial way to interpret these subdeterminants in terms of paths in $G$. For example, using [1], [3] we can show that for any DAG model on a cactus graph (i.e. a connected graph in which any two cycles have at most one vertex in common), the cardinality of $S$ cannot be greater than one.
This is based on ongoing joint work with Caroline Uhler.

## References

[1] Jacob Ponstein, *Self-avoiding paths and the adjacency matrix of a graph*, SIAM Journal on Applied Mathematics 14.3 (1966): 600 – 609.

[2] Seth Sullivant, Kelli Talaska, and Jan Draisma, *Trek separation for Gaussian graphical models.* The Annals of Statistics (2010): 1665–1685.

[3] Caroline Uhler, Garvesh Raskutti, Peter Bu?hlmann, and Bin Yu, *Geometry of the faithfulness assumption in causal inference.* The Annals of Statistics 41.2 (2013): 436–463.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Mode Poset Probability Polytopes

Guido Montúfar[1], Johannes Rauh[2]

[1] *Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany,* montufar@mis.mpg.de

[2] *Leibniz Universität Hannover, Germany,* rauh@math.uni-hannover.de

A mode of a probability vector is a local maximum with respect to some vicinity structure on the set of elementary events. The mode inequalities cut out a polytope from the simplex of probability vectors. Related to this, a strong mode is an elementary event that has more probability mass than all its direct neighbors together. The set of probability distributions with a given set of strong modes is again a polytope.

The patterns of modes that are possible in a given vicinity structure define special types of partial orders in the coordinates of the probability vectors. Mode poset probability polytopes are special types of the well known order and poset polytopes [1].

Statistical models are usually constrained in the patterns of modes and strong modes that they can realize. The set of patterns of modes or strong modes realizable by a statistical model gives a combinatorial characteristic of the model. In the case of exponential families, this is closely related to the combinatorics of convex supports and their normal fans, whereby it also incorporates the extrinsic vicinity structure on the set of elementary events.

The mode characteristic of a statistical model can be used to describe a portion of its complement and represents a coarse form of implicit semialgebraic description. This idea can be used to draw sensible distinctions between certain types of statistical models with hidden variables, such as softmax naïve Bayes models and restricted Boltzmann machines [2].

We study the vertices, the facets, and the volume of mode poset probability polytopes, depending on the sets of (strong) modes and the vicinity structures. We use the obtained results on a few examples describing statistical models with hidden variables.

## References

[1] R. Stanley. Two poset polytopes. *Discrete Comput. Geom.*, 1:9–23, 1986.

[2] G. Montúfar and J. Morton. When does a mixture of products contain a product of mixtures? *SIAM Journal on Discrete Mathematics*, 29:321–347, 2015.

# Computing simplicial complexes with CoCoA

E. Palezzato[1]

[1] *University of Genoa, Italy,* palezzato@dima.unige.it

In this poster we present the package for the algebraic software CoCoA that computes the simplicial homology with explicit description of the generators of a given simplicial complex.

We also compute the Alexander Dual ideal and the Stanley-Reisner ideal associated to the simplicial complex and other invariants, for example the f-vector.

Simplicial complexes are a natural choice to represent geometric objects computationally because they can be easily implemented in a computer by an enumeration of the top simplices. In fact, the knowledge of the top simplices of a complex uniquely determines the whole complex: given the collection of top simplices, it is sufficient to take all their possible faces to construct the complex.

Notice that the simplicial complex approach can be used in the study of clustering data for example for the case of datasets of large scale networks of neurons.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Finding the Statistical Fan of an Experimental Design

D. Pavlov[1]

[1] *Institute of Applied Astronomy, St. Petersburg, Russia,* dpavlov@ipa.nw.ru

This work deals with a problem that was stated in [1]: given an experimental design $\mathscr{F}$, find polynomial models identifiable by $\mathscr{F}$. The set of all such models was named *statistical fan* in [2]. A known solution for the problem is based on the connection between experimental designs and zero-dimensional polynomial ideals. Let us denote $I(\mathscr{F})$ the ideal of polynomials who turn to zero on every point from $\mathscr{F}$ — the *ideal of points*. From each Gröbner basis of $I(\mathscr{F})$ one can trivially obtain the monomial basis of its quotient ring, and the polynomial model built on these monomials will be identifiable by $\mathscr{F}$. The collection of polynomial models obtained from all possible Gröbner bases is called an *algebraic fan*. Algorithms are known [3, 4] for finding the (always finite) set of all possible Gröbner bases (Gröbner fan) given any single Gröbner basis. A single Gröbner basis can be found using the Buchberger-Möller algorithm [5].

The algebraic fan of $\mathscr{F}$ is a subset of the statistical fan, and as it is shown in [2], the inclusion can be strong. This work provides some more examples of differences between algebraic and statistical fans. GFan [4] was used to obtain the algebraic fans for these examples, while a dedicated algorithm was designed and implemented for finding the full statistical fan of an arbitrary experimental design.

The idea of the algorithm is based on the concept of *design matrix* [6]. Let us denote the polynomial model of interest as

$$y = f(x_1, \ldots, x_d) = \sum_{i=1}^{n} c_i t_i, \tag{21}$$

where each of $\{c_i\}$ is a coefficient from some field $K$, and each of $\{t_i\}$ is a monomial: $t_i = \prod_{j=1}^{d} x_j^{\alpha_{ij}}$. The set of monomials $\{t_i\}$ is called the *support* of $f$, and it will be denoted $\mathrm{Supp}(f)$ hereafter. At this point, we put a note that the definitions of both algebraic and statistical fans imply that the used polynomials models are *complete*, i.e. if $\mathrm{Supp}(f)$ has some monomial $t$, then it has all $t$'s divisors as well.

Let the experimental design $\mathscr{F}$ have $N$ $d$-dimensional points: $\mathscr{F} = \{p_1, \ldots, p_N\}$. Then the design matrix $X$ is defined in the following way:

$$X_{ij}(\mathscr{F}, \mathrm{Supp}(f)) = t_j(p_i) = \prod_{k=1}^{d} {p_i}_k^{\alpha_{jk}}, \quad i = 1..N, j = 1..n \tag{22}$$

Model $y = f(x_1, \ldots, x_d)$ is identifiable by design $\mathscr{F}$ iff $\mathrm{rank}(X(\mathscr{F}, \mathrm{Supp}(f))) = n$. That obviously requires $n \leqslant N$. Without the loss of generality, the search can be restricted to polynomial models that are identified unambiguously, this requiring $n = N$. So the problem of finding the statistical fan of $\mathscr{F}$ can be restated as the following: find every complete support $S = \{t_1, \ldots, t_n\}$ of size $n$ whose design matrix $X(\mathscr{F}, S)$ is nonsingular (has rank $n$).

The condition $\mathrm{rank}(X(\mathscr{F}, S)) = n = |S|$ means linear independence of monomials from $S$ in $I(\mathscr{F})$. That necessitates linear independence of any subset of $S$ in $I(\mathscr{F})$, which means that $\mathrm{rank}(X(\mathscr{F}, S')) = |S'|$ for any $S' \subset S$. That, plus the condition that $S'$ is complete, gives an invariant for a combinatorial procedure. The procedure starts with $S_0 = \{\mathbf{1}\}$ and recursively enumerates all possible sequences of monomials, adding a single monomial to the sequence at each step: $S_{i+1} = S_i \cup t_{i+1}$. Since there is generally more that one variable ($d > 1$), there is more than

one possible $S_{i+1}$ for a given $S_i$. The procedure goes all the paths sequentially, and recursively all the paths that branch from them. Some conditions must be hold to keep the invariant:

1. The completeness condition. For a given $S_i$, $t_{i+1}$ should be in "dimple" position regarding to $S_i$. That is, for every $k \in [1..d]$, either $t_{i+1}$ must have no $x_k$ multiple, or there must be $t' \in S_i$ such that $t_{i+1} = t'x_k$.

2. The rank condition. If $\mathrm{rank}(X(\mathscr{F}, S_i \cup t_{i+1})) = i$ instead of $i+1$, then the branch starting from $t_{i+1}$ is not run, and $t_{i+1}$ (along with all its multiples) is excluded from the sequence in the current branch and all of its sub-branches. That is, $t_{i+1}$ is seen as a linear combination of monomials from $S_i$ in $I(\mathscr{F})$.

The procedure essentially is the enumeration of Young tableaux (condition 1) with some additional restriction (condition 2). The technique for keeping the set of "dimple" monomials for condition 1 was presented in [7]. Now, if $i = n$, the branch is terminated and the current $S_n$ is yielded as one of the results. Also, the implementation of the algorithm uses some techniques to avoid repeated calculations:

1. The values of $t_j(p_i)$, used in the calculation of the design matrix, are cached. For each encountered monomial, its values in all of $\{p_i\}$ are stored in a hash table and are accessed from it in the future steps.

2. Each encountered sequence is stored in a hash table as a set (a Young diagram). Future branches that have their sequences resulting in the same set are not run. For example, if the sequence $\{\mathbf{1}, x, y\}$ has been processed, the sequence $\{\mathbf{1}, y, x\}$ should not be, since it will not give any new results.

3. The rank of the matrix $X(\mathscr{F}, S_i \cup t_{i+1})$ is checked in an incremental way using intermediate results obtained on the previous step with matrix $X(\mathscr{F}, S_i)$. That required an implementation of some modified Gauss-Jordan elimination scheme.

**Example 1.** Box-Behnken design with $d = 3$: $\mathscr{F} = \{(0,\pm1,\pm1),(\pm1,0,\pm1)(\pm1,\pm1,0)\}$. The algebraic fan of $\mathscr{F}$ contains 12 polynomial models. The statistical fan of $\mathscr{F}$ contains 14, the two extra being (listing just the monomials): $\{1, z, z^2, y, yz, y^2, y^2z, x, xz, xz^2, xy, x^2, x^2y\}$ and $\{1, z, z^2, y, yz, yz^2, y^2, x, xz, xy, xy^2, x^2, x^2z\}$.

**Example 2.** Box-Behnken design with $d = 4$: the algebraic fan has 48 models, while the statistical fan has 96.

**Example 3.** Box-Wilson design with $d = 3$: $\mathscr{F} = \{(0,\pm1,\pm1),(\pm1,0,\pm1)(\pm1,\pm1,0),(\pm1,0,0),(0,\pm1,0),$ $(0,0,\pm1),(0,0,0)\}$. The algebraic fan is equal to the statistical fan and has 6 models.

**Example 4.** Box-Wilson design with $d = 4$: the algebraic fan has 12 models, while the statistical fan has 54.

**Example 5.** $\mathscr{F} = \{(1,3),(1,1),(0,1),(0,2),(0,0),(2,0)\}$. The algebraic fan consists of two models: $\{1, x, y, xy,$ $y^2, y^3\}$ and $\{1, x, y, xy, x^2, y^2\}$. The statistical fan has two more models: $\{1, x, y, x^2, y^2, y^3\}$ and $\{1, x, y, xy, y^2, xy^2\}$. The term $xy^2$ is not present in any model from the algebraic fan.

**Example 6.** Considering a Boolean field, $d = 4$, and $\mathscr{F} = \{(0,0,0,0),(0,1,1,1),(0,0,1,0),(1,0,0,0),(0,0,1,1),$ $(1,0,1,1),(0,1,0,0),(1,1,0,0)\}$. The albegraic fan has two models: $\{1, x, y, z, w, xy, xw, yw\}$ and $\{1, x, y, z, w, xy, xz, yz\}$, while the statistical fan has two more: $\{1, x, y, z, w, xy, xz, yw\}$ and $\{1, x, y, z, w, xy, yz, xw\}$.

Author would like to thank Nikolay Vasiliev and Konstantin Usevich for useful insights that and discussions that preceded the appearance of this work.

# References


[1] G. Pistone and H. Wynn (1996). Generalised confounding with Gröbner bases, *Biometrica*, **83**, 653–666.

[2] G. Pistone, E. Riccomagno, and H. Wynn (2001). Algebraic statistics: computational commutative algebra in statistics, Chapman & Hall/CRC.

[3] H. Maruri-Aguilar (2005). Universal Gröbner Bases for Designs of Experiments, *Rend. Istit. Mat. Univ. Trieste*, textbf37, 95–119.

[4] A. Jensen (2011). Gfan, a software system for Gröbner fans and tropical varieties, `http://www.math.tu-berlin.de/~jensen/software/gfan/gfan.html`

[5] H. Möller and B. Buchberger (1982). The Construction of Multivariate Polynomials with Preassigned Zeros, *Proceesings of European Conference on Computer Algebra*, 24–31.

[6] E. Riccomagno, M. Caboara, G. Pistone, and H. P. Wynn (1997), The fan of an experimental design, *SCU Research Report*, **10**.

[7] N. Vasiliev and D. Pavlov (2009). Enumeration of finite monomial orderings and combinatorics of universal Gröbner bases. Programming and Computer Software, **35**, 79–89.

# Generalized Fréchet bounds: from contingency tables to discrete copulas

Elisa Perrone[1]

[1] *Johannes Kepler University, Linz, Austria,* elisa.perrone@jku.at

Copula functions are largely employed in applied statistics as a flexible tool to describe the behavior of the dependence between random variables. As a matter of fact, according to the Sklar's theorem (see refs. [7, 6]), the joint bivariate distribution function $F_{XY}$ of two random variables $X$ and $Y$ with univariate marginal distribution functions $F$ and $G$, respectively, can be written as

$$F_{XY}(x,y) = C(F_X(x), F_Y(y)) \qquad (x,y \in R) \tag{23}$$

where $C$ is a bivariate copula uniquely determined on the set $RanF_X \times RanF_Y$. Copulas are a very strong tool in modeling in a continuous setting. Nevertheless, the use of such functions into a discrete one needs to be treated differently and more carefully. The limitations and dangers of an undiscriminating transposition of modeling by copulas in a discrete framework have been analyzed and discussed by several authors in the literature (e.g., see ref. [2]). Due to the complexity and difficulties in extending the results from the continuous framework to the discrete one, the concept of *discrete copulas* has been introduced. Such discrete copula functions can be regarded as the restriction of a discrete bivariate distribution function with uniform discrete univariate marginals (see refs. [4, 3]). The statistical meaning of these families has been discussed in [5, 2], where constructions of discrete copulas given a contingency table have been shown.

In this work we study the connection between discrete copulas and contingency tables. To this end, we start from the context considered in [1, 8] with respect to the problem of finding upper and lower bounds on cell counts in cross classifications of nonnegative counts. In particular, we investigate the meaning of the generalized Fréchet bounds obtained in [1] in the context of discrete copulas with the goal of better understanding the shape of the set of all possible discrete copulas. We also present the relationship between the generalization of the Fréchet bounds in [1] and the classical ones. Finally, we discuss some potential extensions to the multidimensional case.

## References

[1] A. Dobra and S. Fienberg (2000), *Bounds for cell entries in contingency tables given marginal totals and decomposable graphs*, Proceedings of the National Academy of Sciences, 97 (22), pp.11885-11892.

[2] C. Genest and J. Nešlehová (2007), *A Primer on Copulas for Count Data*, Astin Bulletin, 37 (2), pp. 475-515.

[3] A. Kolesárová, R. Mesiar, J. Mordelová, and C. Sempi (2006), *Discrete Copulas*, IEEE Transactions on Fuzzy System, 14 (5), pp. 698-705.

[4] G. Mayor, J. Suñer, and J. Torrens (2005), *Copula-like operations on finite settings*, IEEE Trans. Fuzzy Systems, 13 (4), pp.468-477.

[5] R. Mesiar (2005), *Discrete Copulas - what they are*, Proceedings of the Joint 4th Conference of the European Society for Fuzzy Logic and Technology and the 11th Rencontres Francophones sur la Logique Floue et ses Applications, Barcelona, Spain, September 7-9, 2005, pp. 927-930.

[6] R.B. Nelsen (2006), *An Introduction to Copulas*, Springer, Ney York, second edition.

[7] A. Sklar (1959), *Fonctions de répartition à n dimensions et leurs marges*, Publ. Inst.Statist. Univ. Paris, 8, pp. 229-231.

[8] A.B. Slavkovic and S.E. Fienberg (2004), *Bounds for cell entries in Two-Way tables given conditional relative frequencies*, PSD 2004, LNCS 3050, , pp. 30-43, Springer-Verlag Berlin Heidelberg.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Algebraic-statistics tools for network dynamics and connectivity in *in vitro* cortical cultures

Virginia Pirino[1], Eva Riccomagno[2], Sergio Martinoia[1] and Paolo Massobrio[1]

[1] *Department of Informatics, Bioengineering, Robotics and System Engineering, University of Genova, Italy*
virginia.pirino@edu.unige.it, sergio.martinoia@unige.it, paolo.massobrio@unige.it
[2] *Department of Mathematics, University of Genova, Italy,* riccomag@dima.unige.it

To address the issue of extracting useful information from large data-set of large scale networks of neurons, I propose an approach that involves both algebraic-statistical and topological tools. The first part of my research project is devoted to the investigation of the electrophysiological behavior of *in vitro* cortical assemblies both during spontaneous and stimulus-evoked activity coupled to Micro-Electrode Arrays (MEAs). The goal is to identify core sub-networks of repetitive and synchronous patterns of activity and to characterize them. The analysis is performed at different resolution levels using a clustering algorithm that reduces the network dimensionality. To better visualize the results, I provide a graphical representation of the detected sub-networks and characterize them with a topological invariant, i.e. the sequence of Betti numbers computed on the associated simplicial complexes. The results show that the extracted sub-populations of neurons have a more heterogeneous firing rate with respect to the entire network. Furthermore, the comparison of spontaneous and stimulus-evoked behavior reveals similarities in the identified clusters of neurons, indicating that in both conditions similar activation patterns drive the global network activity. The second part of my research project treats the issue of parameters sensitivity's analysis. In order to validate the developed method, each parameter has been studied in all its variation range and the corresponding results have been compared. Finally, the third part is devoted to the comparison of the proposed methodology with gold standard methods used in this field like Cross-Correlation function and Transfer Entropy algorithm.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Use of Algebraic Statistics in Epidemiological context

Fulvio Ricceri[1,2]

[1] *Unit of Epidemiology, Regional Health Service ASL TO3, Grugliasco (TO), Italy,* fulvio.ricceri@unito.it

[2] *Unit of Cancer Epidemiology,Department of Medical Sciences, University of Turin, Italy*

**Introduction**

In the last years, there was a huge theoretical development of Algebraic Statistics, covering different fields of interest. However, still not many applications using real data have been provided. On the other hand, biostatisticians faced the problem of the analysis of the so-called "big data" coming, for example, from genetic investigations. So, they have a need of new and rigorous methods able to solve arising theoretical and computational problems.

An interesting and useful application of Algebraic Statistics may be the study of interactions between epidemiological variables. In fact, traditional methods of studying interactions failed when working with a large number of variables compared to the number of subjects.

Together with other colleagues, I studied the application of the independence model of two variables $(X_1, X_2)$ from a third $X_3$ applied to the dependence of two genetic variables from the occurrence of cancer ("gene-gene" interaction) and I propose here one of the possible future directions of Algebraic-Biostatistic research.

**First example: EPIC-Genair Study**

The aim of this first example (already published in [1]) is to develop a model that use computational algebraic methods (namely, the Diaconis-Sturmfels algorithm [2]) in order to test the effect of the combination of two variables (here, polymorphisms) on the risk of a disease (here, cancer). This analysis was performed in the Gen-Air study. Gen-Air is a case-control study nested in the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort, that is a multicenter European study, in which more than 500 000 healthy volunteers were recruited in 10 European countries [3]. The aim of Gen-Air was to study the relationship between some types of cancer (bladder cancer, lung cancer, and leukaemia), air pollution, environmental tobacco smoke, and genetic polymorphisms [4].

The most suitable model to apply is the model of independence of two random variables $X_1, X_2$ (the polymorphisms), from a third one $X_3$ (presence or absence of cancer); in symbols:

$$P(X_1 = i, X_2 = j, X_3 = k) = P(X_1 = i, X_2 = j)P(X_3 = k) \tag{24}$$

The parametric equations of model (24) are of the form: $p_{ijk} = \theta_{ij}\mu_k$. We associate to this model the ideal $I_{12,3}$ that contains all polynomials in the variables $p_{ijk}$ that identically vanish on the points given by the parametrization. It is a toric ideal and a Gröbner basis for it is given by binomials of the form:

$$p_{i_1 j_1 k_1} p_{i_2 j_2 k_2} - p_{i_1 j_1 k_2} p_{i_2 j_2 k_1} \tag{25}$$

Then, using the Gröbner basis of the ideal associated to the model, we can apply the Diaconis-Sturmfels algorithm, obtaining in a simpler way a Monte Carlo sampling.

Using this method we were able to identify several interactions that influence the risk of cancer. We compared them with the results from the traditional logistic model and they were consistent.

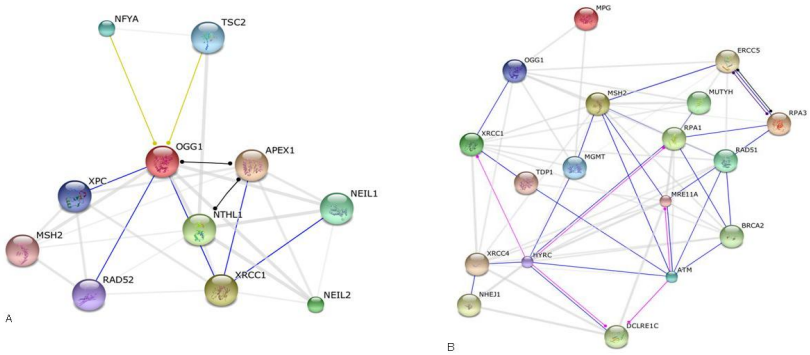
**Second example: DRGP Study**

The same method was applyed to the genotype-phenotype correlation study, a study that investigate the relation between DNA repair capacity and DNA repair genes polymorphisms on healthy subjects [5]. Results are still unpublished.

It is well known that the variant allele in the polymorphism rs1052133 in *OGG1* gene (Ser326Cys) reduces DNA glycosylase activity. For this reason, we decided to explore the interaction between DNA glycosylase activity, *OGG1* rs1052133 polymorphism and other genetic variants.

Among the results, it is interesting to observe that *OGG1* rs1052133 interacts with other functional polymorphisms belonging to the same repair system (Base Excision Repair): *XRCC1* rs17655484 and *MUTYH* rs3219474 ($p-$value$= 0.00001$ and $p-$value$= 0.00003$, respectively).

In the figure we showed the predicted (A) and observed (B) protein-protein interaction for OGG1, estimated using STRING 9.0 (http://string-db.org)




**Further prospective**

Several other independence models using Algebraic Statistics have been studied [6, 7]. These models could be used in order to explore pathways among variables.

For example, in the EPIC study, a full set of nutritional variables are available for all the over 500.000 subjects. Our idea is to test several independence models in this database in order to identify food patterns.

Another example is the study of clusters of carcinogenic exposures in several occupational epidemiology studies.

**Acknowledgement**

I would like to thank Lea Terracini, Margherita Roggero, Claudia Fassino, and Maria Laura Torrente for their help in the mathematical work. I would also thank Paolo Vineis and Giuseppe Matullo for the availability of data. Thanks also to Fabio Rapallo, Angelo d'Errico and Carlotta Sacerdote for the sharing of further prospectives.

# References

[1] F. Ricceri, *et al.* (2012), Algebraic Methods for Studying Interactions Between Epidemiological Variables, *Math. Model. Nat. Phenom*, **7**, 227-252.

[2] P. Diaconis and B. Sturmfels (1998), Algebraic algorithms for sampling from conditional distributions, *Ann. Statist.*, **26**, 363-397.

[3] E. Riboli (2001), The European Prospective Investigation into Cancer and Nutrition (EPIC): plans and progress, *J. Nutr.*, **131**, 170-175.

[4] M. Peluso, *et al.* (2005) Methodology of laboratory measurements in prospective studies on gene-environment interactions: the experience of GenAir, *Mutat Res*, **574**, 92-104.

[5] F. Ricceri, *et al.*, (2011) Involvement of MRE11A and XPA gene polymorphisms in the modulation of DNA double-strand break repair activity: a genotype-phenotype correlation study., *DNA Repair*, **10**, 1044-50.

[6] B. Sturmfels, (2003) Algebra and geometry of statistical models, *Tech. report, John von Neumann Lectures*, TU München

[7] L. Patcher and B. Sturmfels (2005), *Algebraic Statistics for Computational Biology*, Cambridge University Press, Cambridge.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# A new crossing criterion to assess path-following performance for Unmanned Marine Vehicles

Eleonora Saggini[1], Maria-Laura Torrente[2]

[1] *Eleonora Saggini, CNR-ISSIA, Genova, Italy,* saggini@dima.unige.it

[2] *Maria-Laura Torrente, Università di Genova, Italy,* torrente@dima.unige.it

A current trend in marine robotics consists of performance evaluation of Unmanned Marine Vehicles (UMVs) guidance systems. Among the recent literature on the topic (see for instance [2, 3, 4, 5, 8, 9, 12]), we follow the work of [8]. In this paper the problem of evaluating robotic systems performances is addressed by means of three steps: a) define a performance metrics $M$, based on reference, measurements, state signals, control actions, if accessible (it could be a combination of different performance indices and should be computed for each experiment); b) define a system performance $P$ as a function of $M$ computable (at least theoretically) with respect to all possible reference signals and initial conditions; c) design a limited suitable set of experiments.

We concentrate on step a). To this aim, we derive from [8] an innovative criterion to evaluate system performance which gives an important contribution to the definition of $M$. The new criterion can be described as follows. Let $\mathbb{X} = \{(x_{R,i}, y_{R,i}),\ i = 1,\dots,s\} \subset \mathbb{R}^2$ and $\mathbb{Y} = \{(x_{V,i}, y_{V,i}),\ i = 1,\dots,s'\} \subset \mathbb{R}^2$ be two sets identifying the reference and vehicle paths, respectively. The method consists of the following two steps:

1. *Approximation of the target path through a polynomial curve:* compute an algebraic curve $f = 0$ that approximates the points of $\mathbb{X}$ by less than a tolerance $\varepsilon_1$;

2. *Identification of the robot well-approximated positions:* select those points of $\mathbb{Y}$ far from the reference path $f = 0$ for more than a tolerance $\varepsilon_2$.

There are several methods in the literature to address step 1, among which an interesting class is formed by recently developed algorithms relying on tools from Numerical Commutative Algebra (among the others we recall [1, 6, 7, 10, 11]). In general, the input is a set of points in $\mathbb{R}^n$ and the output is a real multivariate polynomial $f$ in $n$ variables whose zero-locus defines an algebraic hypersurface of $\mathbb{R}^n$ representing a good approximation (in some sense) of the geometrical arrangement of the input points. For the evaluation of UMVs, obviously we have $n = 2$. In this paper, we concentrate on the problem outlined in step 2, where a rule is required to classify whether or not each point of $\mathbb{Y}$ lies close to the curve $f = 0$ by less than $\varepsilon_2$. To this aim, we follow the general approach illustrated in [13] and extend it to the Euclidean case.

Let $f = 0$ be an algebraic plane curve (for instance, the curve found in step 1 applying algorithms form Numerical Commutative Algebra) and let $\varepsilon_2$ be a fixed tolerance for the closeness of a generic point $(x_V, y_V) \in \mathbb{Y}$ to $f = 0$. In [13] a crossing algorithm is presented based on a suitable criterion which, depending on the evaluation $f(x_V, y_V)$ and on the local differential geometry of $f = 0$, states whether or not the curve $f = 0$ crosses the $\infty$-neighbourhood of radius $\varepsilon_2$ of $(x_V, y_V)$. In this paper, we propose new analytic bounds on which we plan to produce a rule to identify the points of $\mathbb{Y}$ well-approximated by $f = 0$. The bounds are provided for the general case of hypersurfaces. Following standard notation, we denote by $B_R(p) = \{q \in \mathbb{R}^n \mid \|q - p\|_2 < R\}$ the 2-ball centered at $p \in \mathbb{R}^n$ and given radius $R$, by $\mathrm{Jac}_f(x_1,\dots,x_n) := \left(\frac{\partial f}{\partial x_1},\dots,\frac{\partial f}{\partial x_n}\right)$ the *Jacobian* (or *gradient*) of $f$, and by $H_f(x_1,\dots,x_n) := \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{i,j=1,\dots,n}$ the $n \times n$ symmetric *Hessian matrix* of $f$. For length reasons we omit proofs and deeper details.

**Proposition 4** *Let $\varepsilon > 0$, let $f$ be a non-costant polynomial of $\mathbb{R}[x_1,\ldots,x_n]$ and let $p$ be a point of $\mathbb{R}^n$.*

*a) Let $\mathsf{H} := \max_{q\in B(p)} \|H_f(q)\|_2$; if*

$$|f(p)| > \|\mathrm{Jac}_f(p)\|_2\varepsilon + \frac{\mathsf{H}}{2}\varepsilon^2 =: B_1, \tag{26}$$

*then the hypersurface of equation $f = 0$ does not cross $B_\varepsilon(p)$.*

*b) Suppose that $f$ has degree $\geqslant 2$; if*

$$|f(p)| > \|\mathrm{Jac}_f(p)\|_2\varepsilon + \frac{1}{2}\sigma_{\max}(H_f(p))\varepsilon^2 := B_1', \tag{27}$$

*then the hypersurface of equation $f = 0$ does not cross $B_\varepsilon(p)$ neglecting contributions of order $\mathrm{O}(\varepsilon^3)$.*

**Proposition 5** *Let $\varepsilon > 0$, let $f$ be a degree $\geqslant 2$ polynomial of $\mathbb{R}[x_1,\ldots,x_n]$, and let $p$ be a point of $\mathbb{R}^n$ such that $\mathrm{Jac}_f(p)$ is not the zero vector.*

*a) Let $\mathsf{H} := \max_{q\in B_\varepsilon(p)} \|H_f(q)\|_2$, let $R$ be a positive real number such that $R < \min\left\{\varepsilon, \frac{\|\mathrm{Jac}_f(p)\|_2}{\mathsf{H}}\right\}$ and let $J = \frac{1}{\inf_{q\in B_R(p)} \|\mathrm{Jac_f(q)}\|_2}$. If*

$$|f(p)| < \frac{2R}{\mathsf{J}(2+\mathsf{H}\mathsf{J}R)} =: B_2, \tag{28}$$

*then the hypersurface of equation $f = 0$ crosses $B_\varepsilon(p)$.*

*b) Suppose that the Hessian matrix $H_f(p)$ is nontrivial, let $R$ be a positive real number s.t. $R < \min\left\{\varepsilon, \frac{\|\mathrm{Jac}_f(p)\|_2}{\|H_f(p)\|_2}\right\}$ and set $\Theta := \|\mathrm{Jac}_f^\dagger(p)\|_2 + 3\sqrt{r}\frac{\|H_f(p)\|_2}{\|\mathrm{Jac}_f(p)\|_2^2}R$ where $r = \mathrm{rank}(D\mathrm{Jac}_f^\dagger(p))$. If*

$$|f(p)| < \frac{2R}{\Theta(2+\sqrt{r}\|H_f(p)\|_2\Theta R)} =: B_2', \tag{29}$$

*then the hypersurface of equation $f = 0$ crosses $B_\varepsilon(p)$ neglecting order $\mathrm{O}(R^2)$ contributions.*

# References


[1] J. Abbott, C. Fassino and M. Torrente (2008), Stable border bases for ideals of points, *J. Symbolic Comput.*, 43(12), pp. 883–894.

[2] M. R. Benjamin, J. A. Curcio, J. J. Leonard and P. M. Newman (2006), Navigation of unmanned marine vehicles in accordance with the rules of the road, in *Robotics and Automation*, ICRA 2006. Proceedings IEEE International Conference, pp. 3581–3587.

[3] M. Bibuli, G. Bruzzone, M. Caccia and L. Lapierre, Path-following algorithms and experiments for an unmanned surface vehicle, *Journal of Field Robotics*, vol. 26, no. 8, pp. 669–688, 2009.

[4] M.Caccia, M.Bibuli, G.Bruzzone and L.Lapierre (2012), Further advances in Unmanned Marine Vehicles. IET, Control Engineering Series 77, ch. Vehicle-following for unmanned surface vehicles, pp. 201–230.

[5] M. Caccia, E. Saggini, M. Bibuli, G. Bruzzone, E. Zereik and E. Riccomagno (2013), Towards good experimental methodologies for unmanned marine vehicles in *Lecture Notes in Computer Science*, Springer, pp. 365–372.

[6] C.Fassino (2010), Almost vanishing polynomials for sets of limited precision points, *J. Symbolic Comput.*, vol. 45, no. 1, pp. 19–37.

[7] C. Fassino and M. Torrente, Simple Varieties for Limited Precision Points (2013), *Theoret. Comput. Sci.* 479, pp. 174–186.

[8] E. Saggini, M. Torrente, E. Riccomagno, M. Bibuli, G. Bruzzone, M. Caccia and E. Zereik (2014), Assessing path-following performance for Unmanned Marine Vehicles with algorithms from Numerical Commutative Algebra, *IEEE Control and Automation (MED)* 22nd Mediterranean Conference of, University of Palermo, Palermo, Italy, June 16-19, 2014, pp. 752–757.

[9] E. Saggini, E. Zereik, M. Bibuli, G. Bruzzone, M. Caccia and E. Riccomagno (2014), Performance indices for evaluation and comparison of unmanned marine vehicles? guidance systems, in *19th IFAC World Congress*, Cape Town, South Africa.

[10] T. Sauer (2007), Approximate varieties, approximate ideals and dimension reduction, *Numer. Algorithms*, 45, (1-4), pp. 295–313.

[11] H. Stetter (1999), Polynomials with coefficients of limited accuracy, in *Proceedings of Computer Algebra in Scientific Computing*, vol. 51. Berlin, Springer, pp. 409–430.

[12] S. Tadokoro and A. Jacoff (2011), Performance metrics for response robots [industrial activities], *IEEE Robotics & Automation Magazine*, vol. 18, no. 3, pp. 12–14.

[13] M. Torrente and M. C. Beltrametti (2014), Almost vanishing polynomials and an application to the Hough transform, *J. Algebra Appl.* Vol. 13, No. 8.

**Algebraic Statistics 2015**
**Genova, Italy**
**June 8–11, 2015**

# Frame Permutation Quantization and $\sum\Delta$ in the coding Theory

Zahra Shakerpour[1], Amir Khosravi[2]

[1] *Payame Noor University, Department of Engineering, Chadegan, Isfahan, Iran,* zhm.sheida@gmail.com

[2] *Department of Mathematics, Kharazmi University, Tehran, Iran,* Khosravi@khu.ac.ir

The theory of frame for a Hilbert space plays a fundamental role in signal processing, image processing, data compression, sampling theory and more. In this paper, at first we show that Sigma-Delta quantization is a method of representing band limited signals by $0, 1$ sequences that are from regularly spaced samples of this signals and we peruse the performance of finite frames for encoding and decoding of vectors by applying first-order Sigma-Delta quantization to the frame coefficients.

Furthermore, it is shown that for piecewise continuous frames with compact support, the associated regular frame systems can be decomposed into a finite number of linearly independent sets. Starting the alternating projection scheme from the estimate provided by linear decoding is a way to find a consistent estimate that automatically improves this decoding scheme.

The question is new to have an analytical evaluation of this improvement. Our approach is to find an upper bound to any consistent estimate of X and compare it with the expected. MSE of a classical estimate and this is to be compared with the classical decoding MSE, which is proportional to $(1/R)$.

This means that under the special condition on the quantization threshold crossing of one of theorem in this paper. In addition to redundant representations obtained with frame are playing an ever-expanding role in signal processing data to design flexibility and other desirable properties. One such favorable property is robustness to additive noise. This robustness, carried over to quantization noise (without regard to whether it is random or signal-independent), explains the success of both ordinary over sampled analog-to-digital conversion (ADC) and Sigma-Delta ADC with the canonical linear reconstruction. But the combination of frame expansions with scalar quantization is considerably more interesting and intricate because boundedness of quantization noise can be exploited in reconstruction and frames and quantizers can be designed jointly to obtain favorable performance. This paper introduces a new use of finite frames in vector quantization: Frame Permutation Quantization (FPQ). In FPQ permutation source decoding (PSC) is applied to a frame expansion of a vector. This means that the vector is represented by a partial ordering of the frame coefficients or by signs of the frame coefficients. FPQ provides a space partitioning that can be combined with additional signal constraints a prior knowledge to generate a variety of vector quantizers.