Computations and Data in Algebraic Statistics

Carlos Améndola (Technical University of Berlin),
Anthea Monod (Imperial College London),
Elina Robeva (University of British Columbia),
Bernd Sturmfels (Max-Planck Institute for Mathematics in the Sciences)

May 14–19, 2023

This workshop brought together both theoreticians and practitioners from broadly different fields to advance algebraic statistical methodology to data science. In particular, our goal was to advance the computational feasibility of algebraic statistics. Our focus was specifically on direct applicability of algebraic statistical theory to real data sets. Additionally, this workshop explored other statistical settings that are also algebraic in nature and that find practical applications in various different fields, where nonlinear algebra may also be applied.

1 Overview of the Field

Classical statistics heavily relies on vector spaces, with foundational methodologies like principal component analysis and linear regression drawing upon linear algebra theory and numerical linear algebra computations. In contrast, the field of algebraic statistics, a more recent development emerging in the 1990s, presents a fresh perspective on parametric statistical inference. It offers a flexible setting by conceptualizing families of parametric distributions as solution sets of systems of higher-order multivariate polynomial equations, thereby introducing a less restrictive nonlinear polynomial framework.

This approach proves to be a natural fit for addressing crucial statistical questions, including parameter estimation, model description and selection. Examples encompass the estimation of mixture models using the method of moments, the exploration of latent variable models through matrix and tensor decompositions, causality studies in graphical models, phylogenetic tree model reconstruction, and addressing existence and uniqueness concerns in matrix completion problems. The book "Algebraic Statistics" by Sullivant provides an extensive introduction to the subject, offering an algebraic perspective on various statistical questions.

In all the mentioned cases, the algebraic viewpoint directly facilitates the exploration of statistical questions, as these models manifest as solution sets of polynomial equations and inequalities. These solution sets are termed "algebraic varieties" or "semi-algebraic sets," making concepts from algebraic geometry applicable. Beyond algebraic geometry, algebraic statistics extends its connections to fields such as algebraic topology, convex and discrete geometry, as well as tensors and multilinear algebra. The interplay between these mathematical theories, often implemented numerically through efficient software packages in the realm of "nonlinear algebra," underscores the potential for algebraic statistics to evolve. This evolution involves implementing efficient methods for rapid inference under nonlinear models, addressing a fundamental challenge in data applications.

2 Presentation Highlights & Workshop Format

Our workshop consisted of a wide range of presentations in a variety of formats, which created a flexible and dynamic working environment for the research problems discussed.

2.1 Invited Talks

The program consisted of 10 main talks, making use of the hybrid format:

1. Thomas Kahle: *Towards the classification of 5-Gaussoids*
   There exist 60,212,776 gaussoids characterized by 5 variables. This talk discussed various attempts to categorize these gaussoids, such as classification based on their realizability through covariance matrices. The overarching goal of the research is to guarantee that this valuable reservoir of research data aligns with the FAIR principles: These principles encompass ensuring the data's Findability, Accessibility, Interoperability, and Reusability. The aim therefore aligns with the theme of the workshop, since it strives to enhance the overall accessibility and usability of gaussoids and facilitate its broader utilization in practical settings.

2. Anna Seigal: *Linear Causal Disentanglement*
   The challenge of causal disentanglement revolves around discovering a representation of data in which variables are causally connected. This talk studied the particular case of the realm of linear causal disentanglement, where variables exhibit a causal relationship through a linear model encompassing latent and observed variables. The talk presented efforts and strategies to establish both sufficient and, in the most challenging scenarios, necessary conditions for identifying the linear causal disentanglement configuration. This talk was based on joint work with Chandler Squires, Salil Bhate, and Caroline Uhler and contributed to the theme of the workshop by advancing the understanding and criteria for identifying linear causal relationships among variables in complex datasets.

3. Marta Casanellas: *Equations defining phylogenetic varieties: from general Markov to equivariant models*
   Phylogenetics is a field that explores the evolutionary connections among species by analyzing their molecular sequences, typically depicted on a phylogenetic tree or network. A prevalent method in phylogenetic reconstruction involves modeling nucleotide or amino acid substitution along such trees. This is often achieved using a general Markov model or one of its submodels characterized by specific substitution symmetries. The case when symmetries adhere to the actions of a permutation group G on the rows and columns of a transition matrix are known as *G-equivariant models*.

   The concept of G-equivariant models is important because, in comparison to a general Markov model, they involve fewer parameters. This is an important consideration in where computations are involved and in applications to real data, and therefore relevant ot the theme of the workshop. Consequently, their phylogenetic varieties, which describe the potential configurations of evolutionary relationships, are defined by a greater number of equations. The identification of these equations is typically challenging. Over the past decade, algebraic geometry has found application in phylogenetics, aiding in both phylogenetic reconstruction and establishing the identifiability of parameters within complex evolutionary models. The primary focus of this ongoing project lies in unveiling equations for G-equivariant models. The project has the potential to shed light on the identifiability of networks evolving under G-equivariant models, thus offering valuable insights into the intricate dynamics of evolutionary processes within biological systems. Joint work with Jess Fernndez-Snchez from Universitat Politcnica de Catalunya, Spain.

4. Piotr Zwiernik: *Entropic covariance models*
   In the realm of covariance matrix estimation, a pervasive challenge lies in identifying an appropriate model and devising an efficient estimation method. Traditionally, two widely employed strategies involve imposing linear constraints either directly on the covariance matrix or its inverse. Notably, linear constraints on the matrix logarithm of the covariance matrix have also garnered attention. This

presentation introduces a comprehensive framework encompassing linear restrictions on various transformations of the covariance matrix, incorporating the aforementioned approaches. The proposed estimation methodology hinges on solving a convex problem, culminating in an estimator characterized by consistency and asymptotic Gaussian properties, subject to mild conditions on the data-generating distribution.

Following the development of this overarching theoretical foundation, the focus narrows down to instances where linear constraints necessitate specific off-diagonal entries to be zero. In this context, the geometric representation closely mirrors our understanding of Gaussian graphical models. By delving into this specialized scenario, the talk aims to not only provide a generalizable methodology but also to establish connections between covariance matrix estimation and the intricate structure of Gaussian graphical models, offering valuable insights into the interplay between linear constraints and the underlying geometric properties of the data.

5. Sonja Petrovic: *Sampling lattice points on a polytope: Bayesian Updated Lattice Bases Algorithm*

   Fiber sampling commonly employs Markov Chain Monte Carlo (MCMC) methods, navigating edges within a connected fiber graph through moves defined by a Markov basis. However, the efficiency of these methods is constrained by the computational complexity of determining the basis. When dealing with large fibers, computing every move required for connectivity becomes impractical, necessitating dynamic computation. Moreover, chains on graphs with low connectivity may encounter bottlenecks. This talk discussed these challenges and presented a novel biased sampling algorithm designed for extracting a fiber using an easily computable lattice basis. The algorithm's effectiveness is demonstrated through its application to well-known examples with Markov bases fiber graphs featuring identifiable bottlenecks. The presentation also discussed the concept of a fiber discovery rate, which offers background insights into mixing time and Markov chains, and elucidates the underlying principles of this Bayesian biased sampler for points within a fiber. Also addressed were the computational complexities associated with fiber sampling and provide a more accessible and efficient solution for analyzing complex fiber structures. The work was joint with Miles Bakenhus.

6. Serkan Hoşten: *Maximizing the KL-divergence to a toric model.*

   In the context of a discrete statistical model denoted as M, the KL-divergence originating from a point q within the probability simplex to M achieves its minimum at the maximum likelihood estimator. This talk focused on the problem of identifying point(s) q such that the KL-divergence from q to M attains its maximum, specifically when M is characterized as a toric model. Building upon the foundational contributions of Ay, Matus, Montufar, Rauh, and others, this talk delineated the KL-maximizers for a linear model M. Furthermore, a new algorithm to compute these maximizers was presented. This computational approach integrates the combinatorial aspects of the chamber complex associated with the polytope defining M with numerical algebraic geometry techniques, effectively solving multiple systems of equations. The application of the algorithm was demonstrated with computations performed on toric surfaces, independence models, and other toric models with ML degree equal to one. This talk presented results that contribute to the theoretical understanding of KL-divergence in toric models alongside providing a practical and computational methodology for identifying KL-maximizing points in a wide range of statistical settings. This was joint work with Yulia Alexandr.

7. Primož Škraba: Lessons from Computing Persistence

   Persistent homology is the most widely used technique within topological data analysis (TDA) that explores qualitative aspects of data across various scales. Robust to input data perturbations and independent of dimensions and coordinates, PH offers a concise representation of the qualitative features present in the data. The computation of persistent homology poses numerous important and intriguing challenges, as the field is dynamically evolving with the continuous emergence of new algorithms and software implementations. This talk overviews two primary objectives: firstly, to acquaint a broad audience of computational scientists with the theory and computational methods underlying persistent homology, and secondly, to furnish benchmarks for state-of-the-art implementations in persistent homology computation. The talk introduced persistent homology, walked through the computation pipeline with an emphasis on practical applications, and assessed a variety of synthetic and real-world

datasets to appraise the efficacy of currently available open-source implementations for persistent homology computation. This then served as a guide for which algorithms and implementations are most suitable for different types of datasets, offering valuable insights into the landscape of persistent homology computation. Given that TDA, which is another algebraic perspective of statistics that leverages algebraic topology rather than algebraic geometry (as does algebraic statistics), and has had greater success in real data applications and computation, this talk served as an important precursor to considerations for computations and data analysis in algebraic statistics.

8. Mathias Drton: Testing many possibly irregular polynomial constraints.

   In several applications, a statistical hypothesis test can be expressed algebraically through polynomial equality and inequality constraints applied to a statistical parameter that can be readily estimated. However, incorporating these constraints into statistical tests poses challenges, particularly when the number of relevant constraints is comparable to or even exceeds the number of observed samples. Furthermore, standard distributional approximations may be unreliable due to singularities arising from the constraints. To address these issues, this talked proposed a methodology for designing tests that involves estimating the pertinent polynomials using incomplete U-statistics. The approach builds on recent advancements in bootstrap approximation to determine critical values. Specifically, incomplete U-statistics with a computational budget parameter proportional to the sample size were constructed and its effectiveness was demonstrated in handling scenarios where individual U-statistics kernels may exhibit either mixed non-degeneracy or degeneracy.

9. Joe Kileel: Conditionally-independent mixture models

   Is there significance in algebraic statistics when parametric models are absent? How do varieties factor into the equation? This presentation investigated mixtures that are conditionally independent, eliminating the necessity of parameterizing their distributions. Estimation algorithms tailored for high-dimensional scenarios were introduced, employing a method-of-moments framework and streamlined tensor operations. Additionally, the discussion touched upon the algebraic varieties inherent in this problem, providing insights into the existing understanding of these varieties. The results were obtained from joint work with Yifan Zhang, along with joint contributions alongside Yulia Alexandr and Bernd Sturmfels.

10. Elizabeth Gross: Dimensions of phylogenetic networks

    Phylogenetic networks are instrumental in illustrating evolutionary histories of sets of taxa, accommodating instances of horizontal evolution or hybridization. By incorporating a Markov model of evolution into a phylogenetic network, a model emerges that lends itself well to algebraic examination, represented as an algebraic variety. In this presentation, a formula determining the dimension of the variety associated with a triangle-free level-1 phylogenetic network within a group-based evolutionary model was provided. Along the way, a dimension formula for toric fiber products with codimension zero is established. The discussion also presented an application of these findings to identifiability concerns. This was joint work with Robert Krone and Samuel Martin.

2.2 Impromptu Talks

There were also 10 shorter impromptu talks by in-person attendees, where we particularly encouraged presentations by early-career participants:

1. Pardis Semnani - Causal inference in directed, possibly cyclic, graphical models.
2. Pratik Misra - Combinatorial and algebraic perspectives on the marginal independence structure of Bayesian networks.
3. Tobias Boege - More on the status of the classification of 5-Gaussoids.
4. Guido Montfar - Supermodular Rank: Set function decomposition and Optimization .
5. Daniel Bernstein - Maximum likelihood thresholds of linear concentration models.

6. Ernesto lvarez - Reduction Process in Phylogenetics.
7. David Kahle - Variety Distributions and Applications
8. Ruriko Yoshida - Application of tropical geometry to data analysis over phylogenetic space
9. Irem Portakal - Algebraic sparse factor analysis models
10. Luis David Garca Puente - Identifiability of Structural Equation Models

Additionally, there were three moderated discussions that were very interesting. The hybrid format was especially helpful for these, allowing for input from a wide range of researchers in diverse geographic locations.

2.3 Moderated Discussion Sessions

Active group discussions also took place, with each session being moderated by one of the organizers. Some specific open problems were discussed, forming the basis for a set of interesting and exciting future research areas and directions. Examples of such problems are summarized below.

3 Recent Developments & Open Problems

Four notable problems that were discussed in the discussion sessions and free time of the workshop are the following.

1. **The Likelihood Correspondence**:

   A *discrete statistical model* is an algebraic variety $\mathcal{M}$ in the probability simplex

   $$\Delta_n = \{(p_0, p_1, \ldots, p_n) \in \mathbb{R}^{n+1}_{\geq 0} : \sum_{i=0}^{n} p_i = 1\}.$$

   We identify $\mathcal{M}$ with its Zariski closure in the projective space $\mathbb{P}^n$. The data are summarized in a vector of counts $u = (u_0, u_1, \ldots, u_n)$ and these serve as coordinates for a second projective space $\mathbb{P}^n$. The *likelihood correspondence* $\mathcal{L}_M$ is the variety of pairs $(p, u) \in \mathbb{P}^n \times \mathbb{P}^n$ such that $p \in \mathcal{M}$ is a critical point for the log-likelihood function $\sum_{i=0}^{n} u_i \log(p_i)$. This is a variety of dimension n. Its map onto the second factor is k-to-1, where k is the maximum likelihood degree, *the ML degree* of the model $\mathcal{M}$.

   In this problem, we study the likelihood correspondence and its defining ideal for some basic models in algebraic statistics. The guiding open problem is to characterize models $\mathcal{M}$ for with $\mathcal{L}_{\mathcal{M}}$ is a complete intersection in $\mathbb{P}^n \times \mathbb{P}^n$. We begin with linear and their tight connection to *matroids*. Another scenario of interest is rank constraints on matrices, since these models represent *conditional independence*.

2. **Wachspress Models**:

   The barycentric coordinates of a simplex are the probabilities $p_0, p_1, \ldots, p_n$ on our discrete state space. In this project we replace the simplex Δ_n by an arbitrary simple polytope P with $n + 1$ facets. There is a canonical rational map $P \to \Delta_n$, introduced by Wachspress in the context of geometric modeling, which assigns to each point in the polytope a probability distribution of the facets. These serve as the barycentric coordinates of the point. This construction is closely connected to the theory of *positive geometries* in physics.

   Wachspress models pertain to the map $P \to \Delta_n$ as an object of study from the perspective of algebraic statistics. First steps in this direction have already been taken from a Bayesian perspective. However, the topic is still largely unexplored, which motivated its discussion during the unscheduled time slots of the workshop. Building on known commutative algebra results, beginning discussions studied the Wachspress model of a simple polytope P. One aim was to find a formula for its ML degree in terms of P, and to study its maximum likelihood estimation.

3. **Moment Varieties of Classical Distributions**:

   Continuous distributions on $\mathbb{R}^n$ can be represented by their moments of a given order d. These are the entries in a symmetric tensor, namely the moment tensor of d. The sequence of moment tensors, as d varies, characterizes the distribution. To learn a distribution within a given parametric statistical model $\mathcal{M}$, the approach taken was to first consider the empirical moments of the data and then compute their best approximation among the moments arising from $\mathcal{M}$. This approach is known as the *method of moments* and dates back to Pearson in 1894. This in turn led o the study of moment varieties in algebraic statistics.

   This study built on the work by Belkin and Sinha who developed the method of moments within computer science. They argue that the moments of many classical distributions that are used in probability and statistics depend polynomially on model parameters. We studied the moment varieties of these distributions, which was done for uniform on polytopes, Gaussian mixtures, and for Dirac and Pareto distributions, but many other cases remain open. In particular, the question arises around the gamma, uniform, Laplace, exponential, χ^2, inverse Gaussian, Poisson, geometric, and negative binomial distributions, which were viewed as parameterizations of curves and surfaces. When passing to multidimensional versions of these classical distributions, higher-dimensional moment varieties arise.

4. **Positive Grassmannian and Positroids**:

   The positive Grassmannian can be viewed as a natural generalization of the probability simplex. This perspective as been fruitful in the exciting advances on the *amplituhedron* at the interface of algebraic combinatorics and particle physics. In discussing this problem, we examined the positive Grassmannian and its positroid cells through the lends of algebraic statistics.

   The Grassmannian $\mathrm{Gr}(k,n)$ is the image of a surjectivity positive parameterization from $\mathbb{R}^{k(n-k)}$ into the projective space $\mathbb{P}^{\binom{n}{k}-1}$ and it is cut out by the ideal of quadratic Plücker relations. The positive Grassmannian lives in the simplex $\Delta_{\binom{n}{k}-1}$ and we view it as a statistical model whose states are the k-element subsets of $\{1,2,\ldots,n\}$. If we pass to the boundary of the simplex then we obtain the *positroid models*, which also admit a surjectively positive parameterization via *plabic graphs*. In these models, the states are the bases of the positroid and MLE means learning the network parameters from data.

   A first glimpse on Grassmannians as statistical models currently exist, displaying the likelihood correspondence for $\mathrm{Gr}(2,n)$ with $n=4$ and which reports the ML degrees 26 and 156 for $n=5,6$. This was a fruitful topic of discussion in the workshop and inspired future work following the workshop.

5. **Connecting the No-3-Way Interaction Model:**

   Ruriko Yoshida talked about the following open problem. Consider a three-way contingency table $X=(X_{ijk})$ under the no-three-way interaction model. It is well-known that it is very hard to compute a Markov basis for this model.

   Conjecture: Let M be the set of all $2\times 2\times 2$ basic moves. If we allow each cell X_{ijk} to be -1, then M connects all tables in the given fiber.

4 Scientific Progress Made & Outcome of the Meeting

The workshop united theoreticians and practitioners across diverse fields with the aim of advancing algebraic statistical methodology in the context of data science. The primary objective was to enhance the computational viability of algebraic statistics, with a specific emphasis on its direct application to real datasets. Moreover, the workshop delved into other statistical settings characterized by algebraic nature, demonstrating their practical utility across various fields where nonlinear algebra could be effectively employed.

The initiative of the workshop focused on addressing complex mathematical and statistical issues relevant to the evolving landscape of not only academic and industrial issues, but also those related to current events. The primary areas of focus included:

- Developing models for environmental and ecological systems to enhance our understanding of the impact of climate change on these systems.

- Reimagining urban development and economic systems to tackle persistent inequities in daily living activities.
- Establishing theoretical foundations for contemporary statistical learning techniques, aiming to comprehend their implications in widespread use and facilitate their adaptation to novel applications.

While these challenges were formidable, the workshop focused on bringing together experts from diverse fields such as biology, statistics, and mathematics. By collaboratively examining these challenges through the lens of algebraic statistics and nonlinear algebra, a novel perspective was seeded to address these issues collectively.

Combinatorics, algebra, and geometry formed the underlying basis for many statistical challenges in other scientific disciplines represented at the workshop. For instance, modeling and inferring biological and social networks relied on fundamentally algebraic statistical models, often employing combinatorial walks for estimation. Understanding modern statistical learning techniques was rooted in principles derived from algebraic geometry. The historical connection between algebraic and geometric methods and statistics has been a source of inspiration in the field of algebraic statistics.

By tackling specific problems and uncovering the inherent geometry and algebra, this workshop sought to integrate domain-specific expertise with recent developments in algebraic statistics. The overarching objective was not only to make progress in addressing these applications but also to identify the mathematical and computational tools necessary for the future and initiate their development during the workshop, which, since the workshop, has been an active scientific research direction.

References

[1] C. Améndola, J.C. Faugère, B. Sturmfels, *Moment Varieties of Gaussian Mixtures*, Algebraic Statistics 7 (1), 14–28, 2016

[2] S. Aoki, H. Hara, A. Takemura, *Markov Bases in Algebraic Statistics.* Vol. 199, Springer Science & Business Media, 2012.

[3] C. Bocci, L. Chiantini, *An Introduction to Algebraic Statistics with Tensors* Vol. 1. Warsaw, Poland: Springer, 2019.

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

[5] M. Drton, B. Sturmfels, S. Sullivant, *Lectures on Algebraic Statistics* Oberwolfach Lectures Vol. 39, Springer Science & Business Media, 2008.

[6] G. Pistone, E. Riccomagno, H.P. Wynn, *Algebraic Statistics: Computational Commutative Algebra in Statistics*, CRC Press, 2000.

[7] M. Michalek, B. Sturmfels. *Invitation to Nonlinear Algebra*, Graduates Studies in Mathematics Vol. 211, American Mathematical Society, 2021.

[8] L. Pachter, B. Sturmfels, *Algebraic Statistics for Computational Biology*, Cambridge University Press 13, 2005.

[9] S. Sullivant, *Algebraic Statistics*, Graduates Studies in Mathematics Vol. 194, American Mathematical Society, 2018.

[10] P. Zwiernik, *Semialgebraic Statistics and Latent Tree Models*, CRC Press, 2015.