TAPAS OF ALGEBRAIC STATISTICS

CARLOS AMÉNDOLA, MARTA CASANELLAS, AND LUIS DAVID GARCÍA PUENTE

1. What is Algebraic Statistics?

Algebraic statistics is an interdisciplinary field that uses tools from computational algebra, algebraic geometry, and combinatorics to address problems in statistics and its applications. A guiding principle in this field is that many statistical models of interest are *semialgebraic sets*—a set of points defined by polynomial equalities and inequalities. Algebraic statistics is not only concerned with understanding the geometry and algebra of the underlying statistical model, but also with applying this knowledge to improve the analysis of statistical procedures, and to devise new methods for analyzing data.

A well-known example of this principle is the *model of independence* of two discrete random variables. Two discrete random variables X, Y are independent if their joint probability factors into the product of the marginal probabilities. Equivalently, X and Y are independent if and only if every 2×2-minor of the matrix of their joint probabilities is zero. These quadratic equations, together with the conditions that the probabilities are nonnegative and sum to one, define a semialgebraic set.

In 1998, Persi Diaconis and Bernd Sturmfels showed how one can use algorithms from computational algebraic geometry to sample from conditional distributions. This work is generally regarded as one of the seminal works of what is now referred to as algebraic statistics. However, algebraic methods can be traced back to R. A. Fisher, who used Abelian groups in the study of factorial designs, and Karl Pearson, who used polynomial algebra to study Gaussian mixture models.

Algebraic statistics is a broad field actively expanding from discrete statistical models, contingency table analysis, and experimental design to Gaussian models, singular learning theory, and applications to phylogenetics, machine learning, and biochemical reaction networks. In this note, we will address two recent contributions to this field: an extension of Pearson's work on Gaussian mixtures and some recent results in phylogenetics.

Carlos Améndola is postdoctoral researcher in the Department of Mathematics at Technische Universität München. His email address is carlos.amendola@tum.de. Marta Casanellas is associate professor and vice director of research in the Department of Mathematics at Universitat Politècnica de Catalunya. Her email address is marta.casanellas@upc.edu. Luis David García Puente is associate professor and assistant department chair in the Department of Mathematics and Statistics at Sam Houston State University. His email address is lgarcia@shsu.edu.

2. Algebraic Statistics of Gaussian Mixtures

In 1894 the famous statistician Karl Pearson [5] wanted to explain the asymmetry observed in data measured from a population of Naples' crabs, believing it was possible that two subpopulations of crabs were present in the sample. The corresponding statistical model is known as a Gaussian mixture; in this case a mixture of two univariate Gaussian distributions, each with its own mean and variance. In order to recover the parameters from the sample, Pearson introduced the *method of moments*, matching the density moments to the sample moments. He obtained the following system of polynomial equations in the means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and mixture proportions α_1 and α_2:

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

After considerable effort and cleverness, Pearson managed to eliminate variables to obtain a ninth degree polynomial relation in the single unknown $x = \mu_1\mu_2$,

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

where $\lambda_4 = 9m_2^2 - 3m_4$ and $\lambda_5 = 30m_2m_3 - 3m_5$. After substituting his numerical moment estimates m_i, he found the real roots of this nonic and determined if they could correspond to a solution for the mixture model. We see his approach as one of the first instances of algebraic statistics. Pearson's work leads to natural questions:

Problem 1. *Can Pearson's method be generalized for a mixture of k Gaussians? How many moments are needed to recover the parameters? Is there an analogous polynomial to* (2)*? What is its degree? What about Gaussians in higher dimensions?*

Recovering the parameters from data drawn from a Gaussian mixture is an important problem in statistics, computer science, and machine learning. Answers to the above questions shed light on the computational complexity and the effectiveness of several algorithms proposed in these areas. The key point is that all the moments of a mixture of Gaussians are **polynomials** in the parameters, so they define *moment varieties* that can be studied algebraically.

Recent progress with this approach has been made by Améndola et al. [2, 3], with partial answers. For example, it was shown [3] that considering all the moments up to order $3k-1$ will yield generically a finite number of Gaussian mixture densities with the same matching moments. In other words, the polynomial moment system generalizing (1) will generically have a finite number of solutions for the $3k$ unknown parameters $\mu_i, \sigma_i, \alpha_i$ for $1 \leq i \leq k$. For $k = 2$ this is Pearson's number 9. For $k = 3$ it was found [2] that the corresponding

degree is 225. In contrast, perhaps shockingly, the system of 20 polynomial equations in 20 unknowns corresponding to the moments up to order three of mixtures of two Gaussians in 3-dimensional space $\mathbb{R}^3$ will have generically *infinitely* many solutions. This means that one needs to consider higher order moments in order to recover the parameters. A complete classification of such defective cases is still open.

3. Algebraic Statistical Phylogenetics

Algebraic statistics has been also used in phylogenetics. Phylogenetics seeks to explain the ancestral relationships among a group of living species. These relationships are usually represented in a *phylogenetic tree* as in Figure 1, where the leaves are in bijection with the living species, the interior nodes represent ancestral species, the root is the common ancestor to all the species in the tree, and the edges represent an evolutionary process that led from one ancestral species to the next. Figure 1 shows three possible phylogenetic trees that could explain the evolution of human, gorilla, lemur, and macaque.

In order to infer the phylogenetic tree that best explains the evolution of the species, one uses the genome of the living species and models the substitution of nucleotides using a Markov process on trees, assuming that each position in the genome evolves in the same way and independently of the others. We denote the set of four nucleotides by $\{A, C, G, T\}$. A discrete random variable taking values in this set is assigned to each node of the tree. For each edge, the probabilities of substitution of nucleotides between the two species at the ends of the edge are recorded in a transition matrix (a Markov matrix). The entries of these matrices, together with the distribution of nucleotides at the root of the tree, form the parameters of the model. Then, the probability of observing a certain pattern of nucleotides at the leaves of the tree can be written in terms of these parameters, by assuming that the evolutionary processes of two edges incident at a node v is independent given the observations at v.

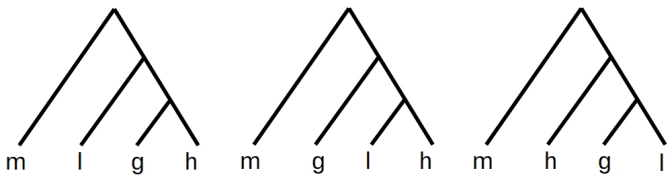


Figure 1. Three phylogenetic trees representing three possible evolutionary histories of human (h), gorilla (g), macaque (m), and lemur (l).

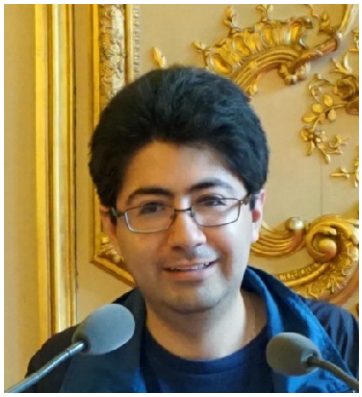
Again, the key point is that for each phylogenetic tree τ, the map φ_τ that sends each set of parameters to the vector of probabilities of patterns $AA\ldots A$, $AA\ldots C$, $\ldots$, $TT\ldots T$ at the leaves is a **polynomial map**, and hence its image is (almost) an algebraic variety V_τ. Different trees (as the ones in Figure 1) lead to different algebraic varieties, and the goal is to use the equations that define these algebraic varieties in order to decide, given a data point (that is, a sequence of nucleotides for each species at the leaves), to which variety is closest (in some sense).

The idea of using polynomial equations in phylogenetics is not due to mathematicians but to biologists. Indeed, in the late 1980s, biologists James A. Cavender, Joseph Felsenstein, and James A. Lake already realized that the equations satisfied by the pattern probabilities on a phylogenetic tree could help in inferring the tree without having to estimate the parameters of the model. It is precisely this, the fact of not having to estimate the parameters, that makes algebraic statistics potentially useful in phylogenetics. However, selecting a set of equations that define the algebraic variety cannot be done in a canonical way. Moreover, the codimension of these varieties grows exponentially in the number of leaves, so using them directly may not be a practical choice.

A recent approach to indirectly using these algebraic varieties is based on the following result due to Elizabeth Allman and John Rhodes: Assume the vector of probabilities $p = (p_{AA\ldots A}, p_{AA\ldots C}, \ldots, p_{TT\ldots T})$ belongs to the image of φ_τ. Any edge of τ splits the set of leaves into two subsets a and b, giving rise to a matrix $M_{a,b}$ whose rows (resp. columns) are labeled by the states at the leaves in a (resp. b) and whose entries are the corresponding probabilities in p. Then the matrix $M_{a,b}$ has rank at most 4. This result leads to equations satisfied by the points of the variety (the 5×5 minors must vanish), and it also gives the possibility to test candidate phylogenetic trees by checking how far certain matrices are from the set of rank 4 matrices. This distance can be easily computed using singular value decomposition. This approach has been recently exploited [1, 4] with great success on both simulated and real data. As a consequence, algebraic tools have finally attracted the attention of biologists and have been implemented in some widely used packages of phylogenetic inference.

There are several books and even a journal dedicated to algebraic statistics. The R package **algstat** contains many computational algebraic statistics tools including the state of the art implementation of the Diaconis-Sturmfels sampling method. The upcoming book *Algebraic Statistics* by Seth Sullivant is a great resource for graduate students and researchers interested in learning more about this exciting field.

4. About the Authors



Carlos Enrique Améndola Cerón wants everyone to know that algebraic statistics is a very cool subject. Inspired by his PhD advisor Bernd Sturmfels, he believes applied algebraic geometry topics have enormous research potential. Additionally, he enjoys playing board games, traveling around the world, and dreaming about far, far away galaxies.

Marta Casanellas did her PhD in algebraic geometry. After a postdoc at UC Berkeley, she shifted her interest towards the applications of algebraic techniques in phylogenetics.

Luis David García Puente works in applied and computational algebraic geometry, algebraic statistics, and algebraic combinatorics. He has devoted much of his professional energy to broadening participation in the mathematical sciences through directing undergraduate research. He also enjoys salsa dancing with his daughters, coaching soccer, and playing racquetball.

References

[1] Elizabeth S. Allman, Laura S. Kubatko, and John A. Rhodes. Split scores: A tool to quantify phylogenetic signal in genome-scale data. *Systematic Biology*, 66:620–636, 2016.

[2] Carlos Améndola, Jean-Charles Faugère, and Bernd Sturmfels. Moment varieties of Gaussian mixtures. *J. Algebr. Stat.*, 7(1):14–28, 2016.

[3] Carlos Améndola, Kristian Ranestad, and Bernd Sturmfels. Algebraic identifiability of Gaussian mixtures. *International mathematics research notices*, 2017.

[4] Jesús Fernández-Sánchez and Marta Casanellas. Invariant versus classical quartet inference when evolution is heterogeneous across sites and lineages. *Molecular Biology and Evolution*, 65:280–291, 2016.

[5] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 185:71–110, 1894.