S. Sullivant: “Algebraic Statistics”. AMS, 2018, xiii + 490 pp

Thomas Kahle, Magdeburg
thomas.kahle@ovgu.de

Algebraic statistics is a young area of mathematics. It started in 1894 with Pearson’s investigation of the crabs in the bay of Naples. Pearson was interested to decide, after measuring body characteristics, if there are two distinct populations of crabs, or just one. He assumed a normal distribution of the ratios of head width to body length with fixed mean and variance inside one population. His goal was to determine whether the data was coming from a mixture of two normal distributions with two distinct means, or just one normal distribution. To do so, he computed the first few moments (equivalent to mean, variance, skewness, . . .) of a mixture of two normal distributions. For example, if μ is the mean of one population and ν the mean of another, and they are mixed with a mixing parameter $\lambda \in [0,1]$, then the mixture’s first moment is $\lambda\mu + (1-\lambda)\nu$. The second moment is $\lambda(\mu^2 + \sigma^2) + (1-\lambda)(\nu^2 + \tau^2)$ where now σ^2 and τ^2 are the variances of the two distributions. Conveniently for the algebraist, the k-th moment of the mixture is a polynomial of total degree $k+1$ in $\lambda, \mu, \nu, \sigma$, and τ. Since there are only five parameters, eventually these formulas must show some algebraic dependence. As it turns out, the sixth moment can be computed from the first five. Deriving and solving a univariate polynomial of degree nine, Pearson could work out the product of the means. Amazingly he found this polynomial by hand and computed its real roots. This allowed him to work backwards and identify the five unknowns $\lambda, \mu, \nu, \sigma, \tau$. Today we would say that he solved an elimination problem. This is akin to Gaussian elimination, just for polynomial systems. In the 20th century these elimination techniques became formalized in the theory of Gröbner bases and in the 21st century we can view Pearson’s work through the lens of the geometry of secant varieties [2].

The method of moments that Pearson had developed fell a little out of fashion when Fisher took over statistics with maximum likelihood estimation. This method is based on optimization over a model. The word “model” is used here for some set of probability distributions, often parametrized. The data is assumed to have been generated by a single true distribution and then the model is a set of possibilities for that true distribution. In the maximum likelihood method one searches for a distribution that leads to the highest probability of observing exactly the data that was actually observed. Very often models are geometric objects and the likelihood can be interpreted as a sort of distance. Then the best guess, the maximum likelihood estimate, is the point in the model that is closest to the data (assuming we can somehow put the data in the same space). As everyone learns in high school, optimization problems are solved by setting the derivative to zero and then solving. For many models that are used in practice, the critical equations are polynomial equations. Clearly algebra, the art of solving equations, ought to contribute to the theory of maximum likelihood estimation. This is the case, for example, if a model is parametrized by rational functions, for then it is cut out by polynomial equations and thus an algebraic variety. Varieties range from smooth to very singular. If the model is smooth enough, then methods of differential geometry have successfully been applied.

This is information geometry [1]. In most cases, however, statistical models are singular and additionally polynomial inequality constraints arise (probabilities are non-negative!). Then (real) algebraic geometry comes into play.

The 19 chapters of “Algebraic Statistics” by Seth Sullivant contain many different mathematical structures and have something to offer for everyone. To name just a few, a combinatorialist can enjoy lattice walks in Chapter 9 or random graphs in Chapter 11. An expert in optimization might consider integer programming approaches to data privacy questions in Chapter 10. A mathematical biologist will find Chapter 15 on phylogenetics interesting. A logician and a commutative algebraist could team up and derive new conditional independence implications using primary decomposition (Chapters 4, 9, 13, and 14). A geometer could discover the tropical Grassmannian while she wanders from the finite metric spaces in phylogenetics using Chapter 19 as a map, and so on.

So what is algebraic statistics? Topology studies shape and algebraic topology uses algebra to study shape. Then algebraic statistics is using algebra to study data and data generating processes. This quick explanation becomes quite accurate if a catholic meaning of algebra is applied, which includes the different mathematical occupations named above. Sullivant’s book is an account of where algebraic statistics stands today and it highlights the many different aspects of the field. This diversity does not mean that algebraic statistics (or the book) are mere assemblies of different things. Maximum likelihood is a connecting theme and conditional independence is another. The second is a central topic in multivariate statistics as conditional independence is deeply rooted in how we think about the world. Scientists aim to understand what causes what and independence is non-causation. This central role is witnessed in that Sullivant dedicates one of his four introductory chapters to it. He places it after the introductions to probability and algebra but before the introduction to statistics.

Conditional independence problems lead to beautiful mathematics. Chapter 4 explains how primary decomposition, the algebraic theory of decomposing the solutions of a polynomial system into irreducible pieces, can be used to find implications for conditional independence. If the random variables under consideration take only finitely many discrete values (for example binary random variables), then a distribution consists just of the finitely many elementary probabilities of outcomes and is thus a vector in some $\mathbb{R}^n$. Conditional independence statements such as “X is independent of Y given Z” translate into polynomial equations in the elementary probabilities. This means that the independence statement holds for a distribution if and only if it satisfies the polynomial equations. If the random variables are normally distributed, the mean $\mu \in \mathbb{R}^n$ and the positive definite covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$ are the right coordinates. In this setting conditional independence is expressed by vanishing of minors of Σ, another mainstay of commutative algebra.

Now suppose that X, Y, Z are binary or normally distributed random variables. If X is independent of Y given Z and Y is independent of Z (given nothing), then it follows that Y is independent of X and Z (given nothing). This is an instance of the “contraction axiom” and Examples 4.3.1/2 in the book show completely automatic proofs using computer algebra. Gröbner bases can take arguments like this to the next level and enable many experimental insights into conditional independence implications. This reveals surprising parallels between the worlds of discrete distributions and (continuous) normal distributions. Thankfully the examples in the book often contain source code snippets that help to get started with relevant computer algebra systems.

Another beautiful connection between statistics and algebraic geometry appears in Chapter 17, which discusses model selection using information criteria. Here one has a partially ordered set or sometimes simply a chain of nested models $M_1 \subseteq M_2 \subseteq \dots$. One tries to decide which model is the smallest that contains the true distribution. Again, maximum likelihood is the method of choice. But now, by the containment of the models, each bigger model will lead to a higher likelihood, so some sort of penalization of the larger models is necessary. The Bayesian information criterion (BIC) evaluates models according to the score $\hat{\ell}_n(M) - \frac{1}{2}\dim(M)\log(n)$ where $\hat{\ell}_n(M)$ denotes the logarithm of the likelihood in model M and n is the size of the sample. Theorem 17.1.3 says that selecting the model with the highest score is asymptotically correct in the sense that this picks the simplest model that contains the true distribution with probability tending to one, as the sample size grows. The penalty term $-\frac{1}{2}\dim(M)\log(n)$ is justified by the validity of the theorem, but there is a deeper connection to the model geometry. If one takes a Bayesian viewpoint on statistics, then one has a prior belief of what the probability of the different models M_i to be the true model might be, and then for each model M_i a prior distribution of what the true distribution might be inside this model. When data comes in, one updates these beliefs using the Bayes rule. This can be used for model selection. Given data, pick the model with the highest updated (posterior) probability. Carrying out this procedure leads one to the evaluation of certain likelihood integrals, but under smoothness assumptions and with Laplace approximation one can recover the BIC. Theorem 17.2.3 says that one ought to pick the model that maximizes $\hat{\ell}_n - \frac{d}{2}\log(n) + O_p(1)$ where $d = \dim(M)$ is the dimension of the model's parameter space and $O_p(1)$ is a term that is bounded in probability. This gives further justification to the BIC as the selection criterion derived from Bayesian analysis equals the BIC plus an error term. At this moment one should pause and think about the smoothness assumptions. Example 17.2.4 shows that if one parametrizes a one-dimensional model using two monomials $t \mapsto (t^3, t^4)$ and puts the true distribution at the singularity $t = 0$, then the asymptotics changes to $\hat{\ell}_n - \frac{1}{6}\log(n) + O_p(1)$. Now the BIC is not really justified anymore. So what is the 1/6? The answer: It is half the real log-canonical threshold, a certain invariant from algebraic geometry.

The key takeaway is that, if the true parameter lies on a singularity of the model, the asymptotics changes and the corrections can be computed from properties of the singularity. This is the content of singular learning theory as developed by Watanabe [4].[1] Specifically, if the parametrization is a polynomial map as in the example above, then the asymptotics is $\hat{\ell}_n - \frac{\lambda}{2}\log(n) + (\mathfrak{m} - 1)\log\log(n) + O_p(1)$ where λ is a positive real number called the learning coefficient and $\mathfrak{m}$ is an integer called the multiplicity. These numbers depend on the model, the prior, and the true parameter. In the smooth one-dimensional

[1] A draft of Watanabe's book was ready just in time for the 2008 Oberwolfach Seminar on Algebraic Statistics. The author of this review received it the night before the workshop by e-mail from the organizers and was charged with printing it in time to be discussed in the hands-on sessions. These discussions influenced [3, Chapter 5].

situation both equal one and one recovers the BIC. In general they can be computed from a real resolution of singularities. This is possible with computer algebra but poses formidable challenges.

Sullivant's Chapter 17 helps tremendously to take the first steps in this area. It lays down the basic principles, highlights the key results, and maps out the literature. Then in Section 17.4 it explains how this method can be put to practice with the 1-factor model, a statistical model where output variables are conditionally independent of each other given one unobservable variable (the factor). Example 17.4.3 gives instructions how to select between the model of total independence and the model with one hidden factor using singular learning theory.

Sullivant's textbook fills a gap in the sense that it is the first to cover algebraic statistics in breadth. It is a book for self-study as much as a reference and guide to the field. The different chapters are samples of the flavors of algebraic statistics, just the right size to create appetite. They also highlight concrete applications and contain pointers to the theoretical buffet. The book also regularly comes back to statistical practice, realistic models, and datasets from the literature. This demonstrates that the author, and more generally the algebraic statistics community, have statistical practice in mind. The much thinner Oberwolfach lecture notes [3] are still not obsolete (not all its open problems are solved!) but Sullivant's new text is clearly the general purpose reference for algebraic statistics now. The breadth of the book also shows the rapid development of algebraic statistics in the 10 year period between it and [3]. Sullivant's text adds an accessible but broad snapshot of where we stand with algebraic statistics. This book is a much anticipated resource and everyone interested in algebraic statistics should be able to find something enticing in it.

References

[1] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of mathematical Monographs*. American Mathematical Society, 2000.

[2] Carlos Améndola, Kristian Ranestad, and Bernd Sturmfels. Algebraic identifiability of gaussian mixtures. *International Mathematics Research Notices*, 2018(21):6556–6580, 2018.

[3] Mathias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on Algebraic Statistics*, volume 39 of *Oberwolfach Seminars*. Springer, Berlin, 2009. A Birkhäuser book.

[4] Sumio Watanabe. *Algebraic geometry and statistical learning theory*. Cambridge University Press, 2009.