

Research Statement

Caroline Uhler, Department of Statistics, UC Berkeley

Algebraic statistics exploits the use of algebraic techniques to study statistical problems. The development of computational algebra software provides a powerful tool to analyze statistical models in the discrete as well as the continuous setting. In my research, I have used methods from combinatorics, computational algebra, and algebraic geometry to study discrete commuting birth-and-death processes and Gaussian graphical models. Algebraic methods have also proven to be useful in the context of applied statistics. I have worked on a range of problems in computational biology and plan to continue research in this direction. My long-term goal is to become a university professor in statistics with a specialization in algebraic statistics and its applications to computational biology.

1 Previous and current research

Birth-and-death processes are among the simplest Markov chains and model a particle that wanders on an interval of the integers by taking unit size steps. These processes arise in fields ranging from queuing theory, where the states represent the number of people waiting in a line, to ecology, where the points represent population size. It is natural to consider Markov chains that have as their state space products of intervals. For instance, the ecology model could be extended to a situation in which one keeps track of the number of females and males. While for birth-and-death processes the finite time behavior is easy to study, higher dimensional models are not even necessarily time reversible anymore. A special case of when the nice one-dimensional theory can be recovered is when the transition matrices in each coordinate direction commute. In collaboration with Steven N. Evans and Bernd Sturmfels, I studied such commuting models and found a unique parameterization. We also characterized the minimal number of constraints on the transition probabilities to ensure commutation. We found that, for example on a grid of size $n \times n$, we only need $3n^2$ constraints on the $4n^2 + 4n$ transition probabilities.

In applications nowadays, we are often faced with problems involving a huge number of random variables, but only a small number of observations. This problem arises for example when studying genetic networks: We seek a model involving a vast number of genes, while we are only given gene expression data of a few individuals. Gaussian graphical models have frequently been used to study gene association networks, and the maximum likelihood estimate (MLE) of the covariance matrix is computed to describe the interaction between different genes. So the following question is extremely important from an applied as well as a theoretical point of view: What is the minimum number of observations in a Gaussian graphical model such that the MLE of the covariance matrix exists? The log likelihood function for Gaussian graphical models is strictly concave and the existence of the MLE is therefore a feasibility problem in convex optimization. Together with my advisor Bernd Sturmfels, I have been studying this problem from the perspective of convex algebraic geometry and we found an algorithm to describe the convex set of sufficient statistics for which the MLE exists. For three-dimensional problems we are also able to give a graphical representation of this set.

The beauty and value of statistics for me stems also from its role as a link between theory and real-world problems. I have been engaged in collaborative cross-disciplinary research, mainly in the context of computational biology. As part of the Neanderthal Genome Analysis Consortium, we analyzed a complete mitochondrial (mt) genome sequence from a 38,000 year-old Neanderthal

individual, establishing that the Neandertal mtDNA falls outside the variation of extant human mtDNAs. Another project was related to the dairy industry. Mastitis is an endemic disease of dairy cows spread all over the world causing a decrease in milk production. Currently, no gold standard test for detecting mastitis is available. I developed an MCMC algorithm to compare the performance of various mastitis tests and to find out which test provides the best results in order to improve the efficiency in milk production. Finally, in my most recent project, I collaborated with a biologist to apply algebraic statistics to disease association studies. Rapid research progress in genotyping techniques have allowed large genome-wide disease association studies. Existing methods often focus on determining associations between single genes and the disease. However, most diseases involve complex relationships between multiple genes and the environment. We developed a method for finding interacting genes and applied this method to a genome-wide dog data set, identifying epistasis associated with canine hair length.

2 Future research

In my previous research on Gaussian graphical models, I have given a geometric characterization of the sufficient statistics for which the MLE exists. However, the connection to the minimal number of observations needed is still not well understood. It is known that the MLE does not exist, if the number of observations is smaller than the maximal clique size of the underlying graph, and it exists, if the number of observations is larger than the treewidth. Steffen Lauritzen posed the following problem: What happens in this gap for non-chordal graphs? I propose to study this problem from two sides: For very small graphs computational algebra techniques and my geometric understanding of the problem should be helpful. For very large graphs I propose to compute asymptotic bounds on the number of observations using the Radon-Nikodym theorem to change to a different probability measure and simplify the problem.

Recently, I participated in a workshop on parameter identification in graphical models, organized by Mathias Drton and Seth Sullivant at the American Institute of Mathematics. I want to gain a better understanding of the structure of graphical models with hidden variables in a Gaussian, discrete, and also non-parametric setting. For small graphs, computational algebra tools can be used to describe constraints imposed by the model on the joint distribution of the observed variables. However, these constraints usually come in the form of complicated polynomial equations and understanding them in terms of the graph structure is challenging. A related question is how to use these constraints to guide the model selection process. Other problems I am interested in are the identifiability of causal effects or parameters and their relation to the underlying graph. Finally, studying the maximum likelihood equations for multivariate Gaussians with arbitrary missing data patterns represents a natural bridge to the previous paragraph. In general, I would like to explore connections between machine learning and computational algebra to study graphical models with hidden variables.

From a more applied point of view, I would like to gain a better understanding of the spatial organization of the chromosomes inside the nucleus. The proximity of chromosomes and genome regions are critical for gene activity, and their radial and relative positions are non-random and similar among similar cell types. Studying the mammalian genome organization under a probabilistic model represents an interesting problem at the interface of computational biology, statistics, convex optimization and geometry. In its simplest form the chromosomes are represented as spheres inside a larger sphere, the nucleus, and the problem can be viewed as a sphere packing problem, where some overlap with other spheres is allowed. However, rotational symmetries will have to be taken into account, which make the problem non-convex. A more realistic model would require the use of ellipsoids to represent the chromosomes. In this model the orientation becomes important, leading to an optimization problem with a rich geometric structure.