# Preface

Algebraic statistics is a relatively young field based on the observation that many questions in statistics are fundamentally problems of algebraic geometry. This observation is now at least twenty years old and the time seems ripe for a comprehensive book that could be used as a graduate textbook on this topic.

Algebraic statistics represents an unusual intersection of mathematical disciplines, and it is rare that a mathematician or statistician would come to work in this area already knowing both the relevant algebraic geometry and statistics. I have tried to provide sufficient background in both algebraic geometry and statistics so that a newcomer to either area would be able to benefit from using the book to learn algebraic statistics. Of course both statistics and algebraic geometry are huge subjects and the book only scratches the surface on either of these disciplines.

I made the conscious decision to introduce algebraic concepts alongside statistical concepts where they can be applied, rather than having long introductory chapters on algebraic geometry, statistics, combinatorial optimization, etc. that must be waded through first, or flipped back to over and over again, before all the pieces are put together. Besides the three introductory chapters on probability, algebra, and statistics (Chapters 2, 3, and 5, respectively), this perspective is followed throughout the text. While this choice might make the book less useful as a reference book on algebraic statistics, I hope that it will make the book more useful as an actual textbook that students and faculty plan to learn from.

Here is a breakdown of material that appears in each chapter in the book.

**Chapter 1** is an introductory chapter that shows how ideas from algebra begin to arise when considering elementary problems in statistics. These ideas are illustrated with the simple example of a Markov chain. As statistical and algebraic concepts are introduced the chapter makes forward reference to other sections and chapters in the book where those ideas are highlighted in more depth.

**Chapter 2** provides necessary background information in probability theory which is useful throughout the book. This starts from the axioms of probability, works through familiar and important examples of discrete and continuous random variables, and includes limit theorems that are useful for asymptotic results in statistics.

**Chapter 3** provides necessary background information in algebra and algebraic geometry, with an emphasis on computational aspects. This starts from definitions of polynomial rings, their ideals, and the associated varieties. Examples are typically drawn from probability theory to begin to show how tools from algebraic geometry can be applied to study families of probability distributions. Some computational examples using computer software packages are given.

**Chapter 4** is an in-depth treatment of conditional independence, an important property in probability theory that is essential for the construction of multivariate statistical models. To study implications between conditional independence models, we introduce primary decomposition, an algebraic tool for decomposing solutions of polynomial equations into constituent irreducible pieces.

**Chapter 5** provides some necessary background information in statistics. It includes some examples of basic statistical models and hypothesis tests that can be performed in reference to those statistical models. This chapter has significantly fewer theorems than other chapters and is primarily concerned with introducing the philosophy behind various statistical ideas.

**Chapter 6** provides a detailed introduction to exponential families, an important general class of statistical models. Exponential families are related to familiar objects in algebraic geometry like toric varieties. Nearly all models that we study in this book arise by taking semialgebraic subsets of the natural parameter space of some exponential family, making these models extremely important for everything that follows. Such models are called algebraic exponential families.

**Chapter 7** gives an in-depth treatment of maximum likelihood estimation from an algebraic perspective. For many algebraic exponential families maximum likelihood estimation amounts to solving a system of polynomial equations. For a fixed model and generic data, the number of critical points of this system is fixed and gives an intrinsic measure of the complexity of calculating maximum likelihood estimates.

**Chapter 8** concerns the geometry of the cone of sufficient statistics of an exponential family. This geometry is important for maximum likelihood estimation: maximum likelihood estimates exist in an exponential family if and only if the data lies in the interior of the cone of sufficient statistics. This chapter also introduces techniques from polyhedral and general convex geometry which are useful in subsequent chapters.

**Chapter 9** describes Fisher's exact test, a hypothesis test used for discrete exponential families. A fundamental computational problem that arises is that of generating random lattice points from inside of convex polytopes. Various methods are explored including methods that connect the problem to the study of toric ideals. This chapter also introduces the hierarchical models, a special class of discrete exponential family.

**Chapter 10** concerns the computation of upper and lower bounds on cell entries in contingency tables given some lower-dimensional marginal totals. One motivation for the problem comes from the sampling problem of Chapter 9: fast methods for computing bounds on cell entries can be used in sequential importance sampling, an alternate strategy for generating random lattice points in polytopes. A second motivation comes from certain data privacy problems associated with contingency tables. The chapter connects these optimization problems to algebraic methods for integer programming.

**Chapter 11** describes the exponential random graph models, a family of statistical models used in the analysis of social networks. While these models fit in the framework of the exponential families introduced in Chapter 6, they present a particular challenge for various statistical analyses because they have a large number of parameters and the underlying sample size is small. They also present a novel area of study for application of Fisher's exact test and studying the existence of maximum likelihood estimates.

**Chapter 12** concerns the use of algebraic methods for the design of experiments. Specific algebraic tools that are developed include the Gröbner fan of an ideal. Consideration of designs that arise in reliability theory lead to connections with multigraded Hilbert series.

**Chapter 13** introduces the graphical statistical models. In graphical models, complex interactions between large collections of random variables are constructed using graphs to specify interactions between subsets of the random variables. A key feature of these models is that they can be specified either by parametric descriptions or via conditional independence constructions. This chapter compares these two perspectives via the primary decompositions from Chapter 4.

**Chapter 14** provides a general introduction to statistical models with hidden variables. Graphical models with hidden variables are widely used in statistics, but the presence of hidden variables complicates their use. This chapter starts with some basic examples of these constructions, including mixture models. Mixture models are connected to secant varieties in algebraic geometry.

**Chapter 15** concerns the study of phylogenetic models, certain hidden variable statistical models used in computational biology. The chapter highlights various algebraic issues involved with studying these models and their equations. The equations that define a phylogenetic model are known as phylogenetic invariants in the literature.

**Chapter 16** concerns the identifiability problem for parametric statistical models. Identifiability of model parameters is an important structural feature of a statistical model. Identifiability is studied for graphical models with hidden variables and structural equation models. Tools are also developed for addressing identifiability problems for dynamical systems models.

**Chapter 17** concerns the topic of model selection. This is a well-developed topic in statistics and machine learning, but becomes complicated in the presence of model singularities that arise when working with models with hidden variables. The mathematical tools to develop corrections come from studying the asymptotics of Bayesian integrals. The issue of nonstandard asymptotics arises precisely at the points of parameter space where the model parameters are not identifiable.

**Chapter 18** concerns the geometry of *maximum a posteriori* (MAP) estimation of the hidden states in a model. This involves performing computations in the tropical semiring. The related parametric inference problem studies how the MAP estimate changes as underlying model parameters change, and is related to problems in convex and tropical geometry.

**Chapter 19** is a study of the geometry of finite metric spaces. Of special interest are the set of tree metrics and ultrametrics which play an important role in phylogenetics. More generally, the set of cut metrics are closely related to hierarchical models studied earlier in the book.

There are a number of other books that address topics in algebraic statistics. The first book on algebraic statistics was Pistone, Riccomagno, and Wynn's text [**PRW01**] which is focused on applications of computational commutative algebra to the design of experiments. Pachter and Sturmfels' book [**PS05**] is focused on applications of algebraic statistics to computational biology, specifically phylogenetics and sequence alignment. Studený's book [**Stu05**] is specifically focused on the combinatorics of conditional independence structures. Aoki, Hara, and Takemura [**AHT12**] give a detailed study of Markov bases which we discuss in Chapter 9. Zwiernik

[**Zwi16**] gives an introductory treatment of tree models from the perspective of real algebraic geometry. Drton, Sturmfels, and I wrote a short book [**DSS09**] based on a week-long short course we gave at the *Mathematisches Forschungsinstitut Oberwolfach* (MFO). While there are many books in algebraic statistics touching on a variety of topics, this is the first one that gives a broad treatment. I have tried to add sufficient background and provide many examples and exercises so that algebraic statistics might be picked up by a nonexpert.

My first attempt at a book on algebraic statistics was in 2007 with Mathias Drton. That project eventually led to the set of lecture notes [**DSS09**]. As part of Mathias's and my first attempt at writing, we produced two background chapters on probability and algebra which were not used in [**DSS09**] and which Mathias has graciously allowed me to use here.

I am grateful to a number of readers who provided feedback on early drafts of the book. These include Elizabeth Allman, Carlos Améndola Cerón, Daniel Bernstein, Jane Coons, Mathias Drton, Eliana Duarte, Elizabeth Gross, Serkan Hosten, David Kahle, Thomas Kahle, Kaie Kubjas, Christian Lehn, Colby Long, František Matúš, Nicolette Meshkat, John Rhodes, Elina Robeva, Anna Seigal, Rainer Sinn, Milan Studený, Ruriko Yoshida, as well as four anonymous reviewers. Tim Römer at the University Osnabrück also organized a reading course on the material that provided extensive feedback. David Kahle provided extensive help in preparing examples of statistical analyses using R. My research during the writing of this book has been generously supported by the David and Lucille Packard Foundation and the US National Science Foundation under grants DMS 0954865 and DMS 1615660.