Preface

Statistical genetics results from the merger of genetics and statistics into a coherent quantitative theory for predicting and interpreting genetic data. Based on this theory, statistical geneticists developed techniques that made notable contributions to animal and plant breeding practices in the second half of the last century as well as to advances in human genetics.

There has been an enormous research impetus in statistical genetics over the last 10 years. Arguably, this was stimulated by major breakthroughs in molecular genetics, by the advent of automatic data-recording devices, and by the possibility of applying computer-intensive statistical methods to large bodies of data with relative ease. Data from molecular biology and biosensors are characterized by their massive volume. Often, intricate distributions need to be invoked for appropriate modeling. Data-reduction techniques are needed for accounting for the involved nature of these data and for extracting meaningful information from the observations. Statistical genetics plays a major role in this process through the development, implementation, and validation of probability models for inference. Many of these models can be daunting and, often, cannot be fitted via standard methods. Fortunately, advances in computing power and computer-based inference methods are making the task increasingly feasible, especially in connection with likelihood and Bayesian inference.

Two important breakthroughs in computational statistics have been the bootstrap and Markov chain Monte Carlo (MCMC) methods. In this book we focus on the latter. MCMC was introduced into the statistical literature in the late 1980s and early 1990s, and incorporation and adaptation of the methods to the needs of quantitative genetic analysis was relatively rapid,

particularly in animal breeding. Also, MCMC is having a major impact in applied statistics (especially from a Bayesian perspective), opening the way for posing models with an enormous amount of flexibility. With MCMC, it is possible to arrive at better descriptions of the perceived underlying structures of the data at hand, free from the strictures of standard methods of statistical analysis.

The objective of this book is to present the main ideas underlying likelihood and Bayesian inference and MCMC methods in a manner that is accessible to numerate biologists, giving step-by-step derivations and fully worked-out examples. Most of these examples are from quantitative genetics and, although not exclusively, we focus on normal or generalized linear models.

Most students and researchers in agriculture, biology, and medicine lack the background needed for understanding the foundations of modern biometrical techniques. This book has been written with this particular readership in mind. A number of excellent books describing MCMC methods have become available in recent years. However, the main ideas are presented typically in a technically demanding style, as these books have been written by and addressed to statisticians. The statistician often has the mathematical background needed to "fill in the blanks". What is tedious detail to a statistician, so that it can be omitted from a derivation, can cause considerable consternation to a reader with a different background. In particular, biologists need a careful motivation of each model from a subject matter perspective, plus a detailed treatment of all the algebraic steps needed to carry out the analysis. Cavalier statements such as "it follows immediately", or "it is easy to show", are encountered frequently in the statistical literature and cause much frustration to biological scientists, even to numerate ones. For this reason, we offer considerably more detail in the developments than what may be warranted for a more mathematically apt audience. We do not apologize for this, and hope that this approach will be viewed sympathetically by the scientific community to which we belong. Nevertheless, some mathematical and statistical prerequisites are needed in order to be able to extract maximum benefit from the material presented in this book. These include a beginning course in differential and integral calculus, an exposure to elementary linear algebra (preferably with a statistical bent), an understanding of probability theory and of the concepts of statistical inference, and a solid grounding in the applications of mixed effects linear models. Most students of quantitative genetics and animal breeding acquire this preparation during the first two years of their graduate education, so we do not feel that the requirements are especially stringent. Some applied statisticians reading this book may be caught by the quantitative genetics jargon. However, we attempt to relate biological to statistical parameters and we trust that the meaning will become clear from the context.

The book is organized into four parts. Part I (Chapters 1 and 2) presents a review of probability and distribution theory. Random variables and their distributions are introduced and illustrated. This is followed by a discussion on functions of random variables. Applied and theoretical statisticians can skip this part of the book safely, although they may find some of the examples interesting.

The first part lays the background needed for introducing methods of inference, which is the subject of the seven chapters in Part II. Chapters 3 and 4 cover the classical theory of likelihood inference. Properties of the maximum likelihood estimator and tests of hypotheses based on the Neyman–Pearson theory are discussed. An effort has been made to derive, in considerable detail, many of the important asymptotic results and several examples are given. The problems encountered in likelihood inference under the presence of nuisance parameters are discussed and illustrated. Chapter 4 ends with a presentation of models for which the likelihood does not have a closed form. Bayesian inference is the subject of chapters 5-8 in Part II . Chapter 5 provides the essential ingredients of the Bayesian approach. This is followed by a chapter covering in fair detail the analysis of the linear model. Chapter 7 discusses the role of the prior distribution in Bayesian analysis. After a short tour of Bayesian asymptotics, the concepts of statistical information and entropy are introduced. This is followed by a presentation of Bayesian analysis using prior distributions conveying vague prior knowledge, perhaps the most contentious topic of the Bayesian paradigm. The chapter ends with an overview of a technically difficult topic called reference analysis. Chapter 8 deals briefly with hypothesis testing from a Bayesian perspective. Chapter 9, the final one of this second part, provides an introduction to the expectation–maximization (EM) algorithm, a topic which has had far-reaching influences in the statistical genetics literature. This algorithm is extremely versatile, and is so inextricable from the statistical structure of a likelihood or Bayesian problem that we opted to include it in this part of the book.

The first two parts of the book described above provide the basis for positing probability models. The implementation and validation of models via MCMC requires some insight on the subtleties on which this technique is based. This is presented in Part III, whose intent is to explain this remarkable computational tool, within the constraints imposed by the authors' limited mathematics. After an introduction to discrete Markov chains in Chapter 10, the MCMC procedures are discussed in a detailed manner in Chapter 11. An inquisitive reader should be able to follow the derivation of the acceptance probability of various versions of the celebrated Metropolis–Hastings algorithm, including reversible jump. An overview of methods for analyzing MCMC output is the subject of Chapter 12.

Part IV gives a presentation of some of the models that are being used in quantitative genetics at present. The treatment is mostly Bayesian and the models are implemented via MCMC. The classical Gaussian mixed

model for single- and multiple-trait analyses is described in Chapter 13. Extensions are given for robust analyses using t distributions. The Bayesian MCMC implementation of this robust analysis requires minor changes in a code previously developed for analyzing Gaussian models, illustrating the remarkable versatility of the MCMC techniques. Chapter 14 discusses analyses involving ordered categorical traits based on the threshold model of Sewall Wright. This chapter also includes a Bayesian MCMC description of a model for joint analysis of categorical and Gaussian responses. Chapter 15 deals with models for the analysis of longitudinal data, and the book concludes with Chapter 16, which introduces segregation analysis and models for the detection of quantitative trait loci.

Although this book can be used as a text, it cannot claim such status fully. A textbook requires carefully chosen exercises, and probably a more linear development than the one presented here. Hence, these elements will need to be provided by the instructor, should this book be considered for classroom use. We have decided not to discuss software issues, although some reasonably powerful public domain programs are already available. The picture in this area is changing too rapidly, and we felt that many of our views or recommendations in this respect would probably be rendered obsolete at the time of publication.

The book evolved from cooperation between the two authors with colleagues from Denmark and Wisconsin leading to a series of papers in which the first applications in animal breeding of Bayesian hierarchical models computed via MCMC methods were reported. Subsequently, we were invited to teach or coteach courses in Likelihood and Bayesian MCMC analysis at Ames (USA), Armidale (Australia), Buenos Aires (Argentina), Edinburgh (Scotland), Guelph (Canada), Jokioinen (Finland), Liège (Belgium), Lleida (Spain), Madison (USA), Madrid (Spain), Milan (Italy), Montecillo (Mexico), Piracicaba (Brazil), Ribeirao Preto (Brazil), Toulouse (France), Uppsala (Sweden), Valencia (Spain), and Viçosa (Brazil). While in the course of these teaching experiences, we thought it would be useful to amalgamate some of our ideas in book form. What we hope you will read is the result of several iterations, starting from a monograph written by Daniel Sorensen and entitled "Gibbs Sampling in Quantitative Genetics". This was published first in 1996 as Internal Report No. 82 by the Danish Institute of Agricultural Sciences (DIAS).

Colleagues, friends, and loved ones have contributed in a variety of ways toward the making of this book. Carlos Becerril, Agustín Blasco, Rohan Fernando, Bernt Guldbrandtsen (who also made endless contributions with LaTeX related problems), Larry Schaeffer, and Bruce Walsh worked through a large part of the manuscript. Specific chapters were read by Anders Holst Andersen, José Miguel Bernardo, Yu-mei Chang, Miguel Pérez Enciso, Davorka Gulisija, Shyh-Forng Guo, Mark Henryon, Bjorg Heringstad, Just Jensen, Inge Riis Korsgaard, Mogens Sandø Lund, Nuala

Sheehan, Mikko Sillanpää, Miguel Angel Toro, and Rasmus Waagepetersen. We acknowledge their valuable suggestions and corrections. However, we are solely responsible for the mistakes that evaded scrutiny, as no book is entirely free of errors. Some of the mistakes find a place in the book by what one may mercifully call random accidents. Other mistakes may reflect incomplete knowledge of the topic on our side. We would be grateful if we could be made aware of these errors.

We wish to thank colleagues at the Department of Animal Breeding and Genetics, DIAS, and at the Departments of Animal Sciences and of Dairy Science of the University of Wisconsin-Madison for providing an intellectually stimulating and socially pleasant atmosphere. We are in special debt to Bernt Bech Andersen for much support and encouragement, and for providing a rare commodity: intellectual space.

We acknowledge John Kimmel from Springer-Verlag for encouragement and patience. Tony Orrantia, also from Springer-Verlag, is thanked for his sharp professional editing.

DG wishes to thank Arthur B. Chapman for his influential mentoring and for his views on the ethics of science, the late Charles R. Henderson for his pioneering work in linear models in animal breeding, and my colleagues and friends Jean-Louis Foulley, Rohan Fernando, and Sotan Im, from whom I learned much. DG had to fit the book into a rather hectic schedule of lecturing and research, both at home and overseas. This took much time away from Graciela, Magdalena, and Daniel Santiago, but they always gave me love, support and encouragement. I also wish to thank Gorgias and Alondra (my parents), the late Tatu, Morocha, and Héctor, and Chiquita, Mumu, Arturo, and Cecilia for their love.

This book was written "at work", at home, in airports and in hotels, on week-ends and on vacation. Irrespective of place, DS received consistent support from Maiken, Jon, and Elsebeth. They accepted that I was unavailable, and put up with moments of frustration (often in good spirit) when things did not work out. I was influenced by and am in debt to my early teachers in Reading and Edinburgh, especially Robert Curnow, Bill Hill, and Alan Robertson. Brian Kennedy introduced me to mixed linear model theory while I was a post-doc in Davis, California, and later in Guelph. I have learned much from him. To my parents I owe unveiling for me at an early age, that part of life that thrives on the top of trees, in worlds of reason and poetry, where it finds its space and achieves its splendor.