

## Preface

The first eight years of the twenty-first century has witnessed the explosion of data collection, with relatively low costs. Data with curves, images and movies are frequently collected in molecular biology, health science, engineering, geology, climatology, economics, finance, and humanities. For example, in biomedical research, MRI, fMRI, microarray, and proteomics data are frequently collected for each subject, involving hundreds of subjects; in molecular biology, massive sequencing data are becoming rapidly available; in natural resource discovery and agriculture, thousands of high-resolution images are collected; in business and finance, millions of transactions are recorded every day. Frontiers of science, engineering, and humanities differ in the problems of their concerns, but nevertheless share a common theme: massive or complex data have been collected and new knowledge needs to be discovered. Massive data collection and new scientific research have strong impact on statistical thinking, methodological development, and theoretical studies. They have also challenged traditional statistical theory, methods, and computation. Many new insights and phenomena need to be discovered and new statistical tools need to be developed.

With this background, the Center for Statistical Research at the Chinese Academy of Science initiated the conference series “International Conference on the Frontiers of Statistics” in 2005. The aim is to provide a focal venue for researchers to gather, interact, and present their new research findings, to discuss and outline emerging problems in their fields, to lay the groundwork for future collaborations, and to engage more statistical scientists in China to conduct research in the frontiers of statistics. After the general conference in 2005, the 2006 International Conference on the Frontiers of Statistics, held in Changchun, focused on the topic “Biostatistics and Bioinformatics”. The conference attracted many top researchers in the area and was a great success. However, there are still a lot of Chinese scholars, particularly young researchers and graduate students, who were not able to attend the conference. This hampers one of the purposes of the conference series. However, an alternative idea was born: inviting active researchers to provide a bird-eye view on the new developments in the frontiers of statistics, on the theme topics of the conference series. This will broaden significantly the benefits of statistical research, both in China and worldwide. The edited books in this series aim at promoting statistical research that has high societal impacts and provide not only a concise overview on the recent developments in the frontiers of statistics, but also useful references to the literature at large, leading readers truly to the frontiers of statistics.

This book gives an overview on recent development on biostatistics and bioinformatics. It is written by active researchers in these emerging areas. It is intended

to give graduate students and new researchers an idea where the frontiers of biostatistics and bioinformatics are, to learn common techniques in use so that they can advance the fields via developing new techniques and new results. It is also intended to provide extensive references so that researchers can follow the threads to learn more comprehensively what the literature is and to conduct their own research. It covers three important topics in biostatistics: Analysis of Survival and Longitudinal Data, Statistical Methods for Epidemiology, and Bioinformatics, where statistics is still advancing rapidly today.

Ever since the invention of nonparametric and semiparametric techniques in statistics, they have been widely applied to the analysis of survival data and longitudinal data. In Chapter 1, Jianqing Fan and Jiancheng Jiang give a concise overview on this subject under the framework of the proportional hazards model. Nonparametric and semiparametric modeling and inference are stressed. Dongling Zeng and Jianwen Cai introduce an additive-accelerated rate regression model for analyzing recurrent event in Chapter 2. This is a flexible class of models that includes both additive rate model and accelerated rate models, and allows simple statistical inference. Longitudinal data arise frequently from biomedical studies and quadratic inference function provides important approaches to the analysis of longitudinal data. An overview is given in Chapter 3 on this topic by John Dziak, Runze Li, and Annie Qiu. In Chapter 4, Yi Li gives an overview on modeling and analysis of spatially correlated data with emphasis on mixed models.

The next two chapters are on statistical methods for epidemiology. Amy Laird and Xiao-Hua Zhou address the issues on study designs for biomarker-based treatment selection in Chapter 5. Several trial designs are introduced and evaluated. In Chapter 6, Jinbo Chen reviews recent statistical models for analyzing two-phase epidemiology studies, with emphasis on the approaches based on estimating-equation, pseudo-likelihood, and maximum likelihood.

The last four chapters are devoted to the analysis of genomic data. Chapter 7 features protein interaction predictions using diverse data sources, contributed by Yin Liu, Inyoung Kim, and Hongyu Zhao. The diverse data sources information for protein-protein interactions is elucidated and computational methods are introduced for aggregating these data sources to better predict protein interactions. Regulatory motif discovery is handled by Qing Zhou and Mayetri Gupta using Bayesian approaches in Chapter 8. The chapter begins with a basic statistical framework for motif finding, extends it to the identification of *cis*-regulatory modules, and then introduces methods that combine motif finding with phylogenetic footprint, gene expression or ChIP-chip data, and nucleosome positioning information. Cheng Li and Samir Amin use single nucleotide polymorphism (SNP) microarrays to analyze cancer genome alterations in Chapter 9. Various methods are introduced, including paired and non-paired loss of heterozygosity analysis, copy number analysis, finding significant altered regions across multiple samples, and hierarchical clustering methods. In Chapter 10, Evan Johnson, Jun Liu and Shirley Liu give a comprehensive overview on the design and analysis of ChIP-chip data on genome tiling microarrays. It spans from biological background and ChIP-chip experiments to statistical methods and computing.

The frontiers of statistics are always dynamic and vibrant. Young researchers

are encouraged to jump into the research wagons and cruise with tidal waves of the frontiers. It is never too late to get into the frontiers of scientific research. As long as your mind is evolving with the frontiers, you always have a chance to catch and to lead next tidal waves. We hope this volume helps you getting into the frontiers of statistical endeavors and cruise on them thorough your career.

Jianqing Fan, Princeton

Xihong Lin, Cambridge

Jun Liu, Cambridge

August 8, 2008