

Preface

It would not be a stretch to say that modern statistics is enjoying a youthful bloom in greater China. The Chinese have long been renowned for fundamental contributions to the foundations of science and mathematical thinking. We expect no less from future generations of Chinese statisticians.

At the dawn of the 20th century, statistics had but a small role on the stage of scientific practice; by the close of the century, the use of statistics was ubiquitous in virtually every area of scientific and practical interest, with statisticians having developed many procedures to address substantive problems arising in these various areas. In the last 25 years especially, proceeding in hand with the exponential growth in computing power, new technologies have emerged in various disciplines, leading to data sets that are orders of magnitude larger than those for which the earlier procedures were developed. This prompts questions which these earlier procedures are not well suited to address. Substantial statistical innovations are necessary to take on the task, and while statisticians are responding well to the challenge, much work remains to be done.

This volume, occasioned by the University of Science and Technology (USTC) 50th anniversary celebration, features 15 statistical papers by renowned statisticians and USTC alumni covering 4 topics in high dimensional data analysis and bioinformatics at the frontiers of modern statistical science: (i) statistical genetics, (ii) the analysis of microarray data, (iii) computational biology and statistical learning, and (iv) statistical methods for analyzing high dimensional data.

The three statistical genetics papers cover several important topics in the area. Cui, Zhang, Yang and Li address the problems of bias and power in linkage analysis with mixed affected sibling pair data. They propose three test procedures to address these issues, and show that all three perform satisfactorily. Chen reviews sequential importance sampling algorithms developed in population genetics, as well as a more recently proposed technique developed by Chen that incorporates resampling. Lin proposes a hierarchical Bayesian approach for detecting QTL using model selection techniques. The approach works well, provided the number of markers is not very large.

Four papers address the analysis of microarray data. By incorporating longitudinal information on gene expression, Hong develops a functional hi-

erarchical empirical Bayes approach for detecting TR and TDE genes from MTC gene expression experiments. Using a smoothness assumption on the gene expression trajectories, the gene expression profiles are modeled and approximated by well known basis function expansions. The paper by Lai addresses, in the context of FDR, the problem of estimating the fraction of null hypotheses that are true when a large number of tests are performed; using a nonparametric method, an upper bound on the fraction is obtained. Based on earlier work on normalization methods for microarray data, and information on non-replicating genes, Peng proposes new methods that lead to improved estimation of the intensity functions. In connection with the need for reliable variance estimation for gene expression microarray data, Tong and Wang review several statistical methods for estimating variances in the “large p , small n ” context.

Four papers are addressed to the subject of computational biology and statistical learning. Feng, Xu, Zhang, Li, Xie and Wang study the problem of predicting protein subcellular locations using a machine learning type of approach. Through experiments and comparisons, the authors conclude that using the PSSM generated from PSI-BLAST as input and SVM as classifier leads to better predictive performance. A new learning method called LOCSVMPSI is proposed and recommended based on its even better performance. Chao and Jiang give a general review of methods for extracting information at the biomolecular sequence level, with emphasis on biological sequence alignment. Lin, Simmons, Beecher, Truong and Young apply various classification methods, including RP, SVM, and RF, to identify an important set of metabolites for disease classification. Wang and Xi survey recent developments in the statistical modeling of chromatin Sequences. They argue that chromatin sequences trained by a previously proposed model called DHMM may have larger power in predicting the correct nucleosome positioning.

The final set of papers is concerned with high dimensional data analysis. Chen gives a comprehensive review of recent developments in finite mixture modeling. Guo and Dai propose an iterative procedure to fit a smoothing spline ANOVA model with heterogeneous variances. The method is then applied to a data set with a sample of epileptics. Zeng and Yu address the issue of bias arising in kernel estimation in longitudinal studies. They propose a bias-corrected procedure and derive its large sample properties. Zou derives a computable bound in evaluating the quality of the Gibbs sampler, in the context of estimating the posterior mode of the Lasso distribution.

The bound has direct implications for deriving the Lasso estimator. The papers in this collection illustrate both the types of challenges statisticians face and will continue to face and as well as the opportunities such challenges open up for new statistical work. On one hand, statisticians need to look

inward and develop better statistical methodologies and procedures, and on the other, to look outward and work with researchers in other disciplines to meet the new challenges they bring to the table. There is every reason to believe that statistics in the 21st century will continue to be as exciting an area as it was in the 20th century, and the need to train and develop a talented generation of younger statisticians will be great. We hope that USTC and its alumni will continue its work along these line and we trust that the next 50 years for USTC will be even more fruitful and exciting than the 50 past years that we celebrate today.

From the Editors *Shuangge Ma & Yuedong Wang*

Preface *Shaw-Hwa Lo*
Columbia University

Section I Statistical Genetics

Linkage Analysis with Mixed Samples of Affected Full and Half Sibs in the Presence of Uncertain Relationship

Wenguo Gu, Hong Zhang, Yaning Yang, Zhepei Li

Sequential Importance Sampling with Resampling in Molecular Population Genetics

Yuguo Chen

Multiple-Salt Quantitative Trait Loci Mapping Using Stochastic Search Variable Selection

Nan Lin

Section II Statistical Analysis of Microarray Data

Statistical Analysis of Microarray Time Course Gene Expression Data with Functional Hierarchical Models

Fangxin Hong

A Nonparametric Method for the Conservative Estimation of the Proportion of True Null Hypotheses

Yinglei Lei

Partial Consistent Normalization Methods of Microarray Data

Heng Peng

Variance Estimation for Gene-expression Microarray Data

Tiejun Tong, Yuedong Wang