Contributions to Statistics

More information about this series at http://www.springer.com/series/2912

S. Ejaz Ahmed
Editor

Big and Complex Data Analysis

Methodologies and Applications

Springer

Editor
S. Ejaz Ahmed
Department of Mathematics & Statistics
Brock University
St. Catherines, Ontario
Canada

ISSN 1431-1968
Contributions to Statistics
ISBN 978-3-319-41572-7 ISBN 978-3-319-41573-4 (eBook)
DOI 10.1007/978-3-319-41573-4

Library of Congress Control Number: 2017930198

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book comprises a collection of research contributions toward high-dimensional data analysis. In this data-centric world, we are often challenged with data sets containing many predictors in the model at hand. In a host of situations, the number of predictors may very well exceed the sample size. Truly, many modern scientific investigations require the analysis of such data. There are a host of buzzwords in today's data-centric world, especially in digital and print media. We encounter data in every walk of life, and for analytically and objectively minded people, data is everything. However, making sense of the data and extracting meaningful information from it may not be an easy task. Sometimes, we come across buzzwords such as big data, high-dimensional data, data visualization, data science, and open data without a proper definition of such words. The rapid growth in the size and scope of data sets in a host of disciplines has created a need for innovative statistical and computational strategies for analyzing such data. A variety of statistical and computational tools are needed to deal with such type of data and to reveal the data story.

This book focuses on variable selection, parameters estimation, and prediction based on high-dimensional data (HDD). In classical regression context, we define HDD where a number of predictors (d) are larger than the sample size (n). There are situations when the number of predictors is in millions and sample size maybe in hundreds. The modeling of HDD, where the sample size is much smaller than the size of the data element associated with each observation, is an important feature in a host of research fields such as social media, bioinformatics, medical, environmental, engineering, and financial studies, among others. A number of the classical techniques are available when $d < n$ to tell the data story. However, the existing classical strategies are not capable of yielding solutions for HDD. On the other hand, the term "big data" is not very well defined, but its problems are real and statisticians need to play a vital role in this data world. Generally speaking, big data relates when data is very large and may not even be stored at one place. However, the relationship between n and d may not be as crucial when comparing with HDD. Further, in some cases, users are not able to make the distinction between population and sampled data when dealing with big data. In any event, the big data or data science is an emerging field stemming equally from research enterprise and public and private sectors. Undoubtedly, big data is the future of research in a host of research fields, and transdisciplinary programs are required to develop the skills for data scientists. For example, many private and public agencies are using sophisticated number-crunching, data mining, or big data analytics to reveal patterns based on collected information. Clearly, there is an increasing demand for efficient prediction strategies for analyzing such data. Some examples of big data that have prompted demand are gene expression arrays; social network modeling; clinical, genetics, and phenotypic spatiotemporal data; and many others.

In the context of regression models, due to the trade-off between model prediction and model complexity, the model selection is an extremely important and challenging problem in the big data arena. Over the past two decades, many penalized regularization approaches have been developed to perform variable selection and estimation simultaneously. This book makes a seminal contribution in the arena of big data analysis including HDD. For a smooth reading and understanding of the contributions made in this book, it is divided in three parts as follows:

General High-dimensional theory and methods (chapters "Regularization After Marginal Learning for Ultra-High Dimensional Regression Models"–"Bias-Reduced Moment Estimators of Population Spectral Distribution and Their Applications")

Network analysis and big data (chapters "Statistical Process Control Charts as a Tool for Analyzing Big Data"–"Nonparametric Testing for Heterogeneous Correlation")

Statistics learning and applications (chapters "Optimal Shrinkage Estimation in Heteroscedastic Hierarchical Linear Models"–"A Mixture of Variance-Gamma Factor Analyzers")

We anticipate that the chapters published in this book will represent a meaningful contribution to the development of new ideas in big data analysis and will showcase interesting applications. In a sense, each chapter is self-contained. A brief description of the contents of each of the eighteen chapters in this book is provided.

Chapter "Regularization After Marginal Learning for Ultra-High Dimensional Regression Models" (Feng) introduces a general framework for variable selection in ultrahigh-dimensional regression models. By combining the idea of marginal screening and retention, the framework can achieve sign consistency and is extremely fast to implement.

In chapter "Empirical Likelihood Test for High Dimensional Generalized Linear Models" (Zang et al.), the estimation and model selection aspects of high-dimensional data analysis are considered. It focuses on the inference aspect, which can provide complementary insights to the estimation studies, and has at least two notable contributions. The first is the investigation of both full and partial tests, and the second is the utilization of the empirical likelihood technique under high-dimensional settings.

Abstract random projections are frequently used for dimension reduction in many areas of machine learning as they enable us to do computations on a more succinct representation of the data. Random projections can be applied row-

and column-wise to the data, compressing samples and compressing features, respectively. Chapter “Random Projections For Large-Scale Regression” (Thanei et al.) discusses the properties of the latter column-wise compression, which turn out to be very similar to the properties of ridge regression. It is pointed out that further improvements in accuracy can be achieved by averaging over least squares estimates generated by independent random projections.

Testing a hypothesis subsequent to model selection leads to test problems in which nuisance parameters are present. Chapter “Testing in the Presence of Nuisance Parameters: Some Comments on Tests Post-Model-Selection and Random Critical Values” (Leeb and Pötscher) reviews and critically evaluates proposals that have been suggested in the literature to deal with such problems. In particular, the chapter reviews a procedure based on the worst-case critical value, a more sophisticated proposal based on earlier work, and recent proposals from the econometrics literature. It is furthermore discussed why intuitively appealing proposals, for example, a parametric bootstrap procedure, as well as another recently suggested procedure, do not lead to valid tests, not even asymptotically.

As opposed to extensive research of covariate measurement error, error in response has received much less attention. In particular, systematic studies on general clustered/longitudinal data with response error do not seem to be available. Chapter “Analysis of Correlated Data with Error-Prone Response Under Generalized Linear Mixed Models” (Yi et al.) considers this important problem and investigates the asymptotic bias induced by the error in response. Valid inference procedures are developed to account for response error effects under different situations, and asymptotic results are appropriately established.

Statistical inference on large covariance matrices has become a fast growing research area due to the wide availability of high-dimensional data, and spectral distributions of large covariance matrices play an important role. Chapter “Bias-Reduced Moment Estimators of Population Spectral Distribution and Their Applications” (Qin and Li) derives bias-reduced moment estimators for the population spectral distribution of large covariance matrices and presents consistency and asymptotic normality of these estimators.

Big data often take the form of data streams with observations of a related process being collected sequentially over time. Statistical process control (SPC) charts provide a major statistical tool for monitoring the longitudinal performance of the process by online detecting any distributional changes in the sequential process observations. So, SPC charts could be a major statistical tool for analyzing big data. Chapter “Statistical Process Control Charts as a Tool for Analyzing Big Data” (Qiu) introduces some basic SPC concepts and methods and demonstrates the use of SPC charts for analyzing certain real big data sets. This chapter also describes some recent SPC methodologies that have a great potential for handling different big data applications. These methods include disease dynamic screening system and some recent profile monitoring methods for online monitoring of profile/image data that is commonly used in modern manufacturing industries.

Chapter “Fast Community Detection in Complex Networks with a K-Depths Classifier” (Tian and Gel) introduces a notion of data depth for recovery of community structures in large complex networks. The authors propose a new data-driven algorithm, K-depths, for community detection using the L_1 depth in an unsupervised setting. Further, they evaluate finite sample properties of the K-depths method using synthetic networks and illustrate its performance for tracking communities in online social media platform Flickr. The new method significantly outperforms the classical K-means and yields comparable results to the regularized K-means. Being robust to low-degree vertices, the new K-depths method is computationally efficient, requiring up to 400 times less CPU time than the currently adopted regularization procedures based on optimizing the Davis-Kahan bound.

Chapter "How Different are Estimated Genetic Networks of Cancer Subtypes?" (Shojaie and Sedaghat) presents a comprehensive comparison of estimated networks of cancer subtypes. Specifically, the networks estimated using six estimation methods were compared based on various network descriptors characterizing both local network structures, that is, edges, and global properties, such as energy and symmetry. This investigation revealed two particularly interesting properties of estimated gene networks across different cancer subtypes. First, the estimates from the six network reconstruction methods can be grouped into two seemingly unrelated clusters, with clusters that include methods based on linear and nonlinear associations, as well as methods based on marginal and conditional associations. Further, while the local structures of estimated networks are significantly different across cancer subtypes, global properties of estimated networks are less distinct. These findings can guide future research in computational and statistical methods for differential network analysis.

Statistical analysis of big clustered time-to-event data presents daunting statistical challenges as well as exciting opportunities. One of the challenges in working with big biomedical data is detecting the associations between disease outcomes and risk factors that involve complex functional forms. Many existing statistical methods fail in large-scale settings because of lack of computational power, as, for example, the computation and inversion of the Hessian matrix of the log-partial likelihood is very expensive and may exceed computation memory. Chapter "A Computationally Efficient Approach for Modeling Complex and Big Survival Data" (He et al.) handles problems with a large number of parameters and propose a novel algorithm, which combines the strength of quasi-Newton, MM algorithm, and coordinate descent. The proposed algorithm improves upon the traditional semiparametric frailty models in several aspects. For instance, the proposed algorithms avoid calculation of high-dimensional second derivatives of the log-partial likelihood and, hence, are competitive in term of computation speed and memory usage. Simplicity is obtained by separating the variables of the optimization problem. The proposed methods also provide a useful tool for modeling complex data structures such as time-varying effects.

Asymptotic inference for the concentration of directional data has attracted much attention in the past decades. Most of the asymptotic results related to concentration parameters have been obtained in the traditional large sample size and fixed dimension case. Chapter "Tests of Concentration for Low-Dimensional and High-Dimensional Directional Data" (Cutting et al.) considers the extension of existing testing procedures for concentration to the large *n* and large *d* case. In this high-dimensional setup, the authors provide tests that remain valid in the sense that they reach the correct asymptotic level within the class of rotationally symmetric distributions.

"Nonparametric testing for heterogeneous correlation" covers the big data problem of determining whether a weak overall monotone association between two variables persists throughout the population or is driven by a strong association that is limited to a subpopulation. The idea of homogeneous association rests on the underlying copula of the distribution. In chapter "Nonparametric Testing for Heterogeneous Correlation" (Bamattre et al.), two copulas are considered, the Gaussian and the Frank, under which components of two respective ranking measures, Spearman's footrule and Kendall's tau, are shown to have tractable distributions that lead to practical tests.

Shrinkage estimators have profound impacts in statistics and in scientific and engineering applications. Chapter "Optimal Shrinkage Estimation in Heteroscedastic Hierarchical Linear Models" (Kou and Yang) considers shrinkage estimation in the presence of linear predictors. Two heteroscedastic hierarchical regression models are formulated, and the study of optimal shrinkage estimators in each model is thoroughly presented. A class of shrinkage estimators, both parametric and semiparametric, based on unbiased risk estimate is proposed and is shown to be (asymptotically) optimal under mean squared error loss in each model. A simulation study is conducted to compare the performance of the proposed methods with existing shrinkage estimators. The authors also apply the method to real data and obtain encouraging and interesting results.

Chapter "High Dimensional Data Analysis: Integrating Submodels" (Ahmed and Yuzbasi) considers efficient prediction strategies in sparse high-dimensional model. In high-dimensional data settings, many penalized regularization strategies are suggested for simultaneous variable selection and estimation. However, different strategies yield a different submodel with different predictors and number of predictors. Some procedures may select a submodel with a relatively larger number of predictors than others. Due to the trade-off between model complexity and model prediction accuracy, the statistical inference of model selection is extremely important and a challenging problem in high-dimensional data analysis. For this reason, we suggest shrinkage and pretest post estimation strategies to improve the prediction performance of two selected submodels. Such a pretest and shrinkage strategy is constructed by shrinking an overfitted model estimator in the direction of an underfitted model estimator. The numerical studies indicate that post selection pretest and shrinkage strategies improved the prediction performance of selected submodels. This chapter reveals many interesting results and opens doors for further research in a host of research investigations.

Chapter "High-Dimensional Classification for Brain Decoding" (Croteau et al.) discusses high-dimensional classification within the context of brain decoding where spatiotemporal neuroimaging data are used to decode latent cognitive states. The authors discuss several approaches for feature selection including persistent homology, robust functional principal components analysis, and mutual information networks. These features are incorporated into a multinomial logistic classifier, and model estimation is based on penalized likelihood using the elastic net penalty. The approaches are illustrated in an application where the task is to infer, from brain activity measured with magnetoencephalography (MEG), the type of video stimulus shown to a subject.

Principal components analysis is a widely used technique for dimension reduction and characterization of variability in multivariate populations. In chapter “Unsupervised Bump Hunting Using Principal Components” (A. D’ıaz-Pach’on et al.), the authors interest lies in studying when and why the rotation to principal components can be used effectively within a response-predictor set relationship in the context of mode hunting. Specifically focusing on the Patient Rule Induction Method (PRIM), the authors first develop a fast version of this algorithm (fastPRIM) under normality which facilitates the theoretical studies to follow. Using basic geometrical arguments, they then demonstrate how the principal components rotation of the predictor space alone can in fact generate improved mode estimators. Simulation results are used to illustrate findings.

The analysis of high-dimensional data is challenging in multiple aspects. One aspect is interaction analysis, which is critical in biomedical and other studies. Chapter “Identifying Gene-Environment Interactions Associated with Prognosis Using Penalized Quantile Regression” (Wang et al.) studies high-dimensional interactions using a robust approach. The effectiveness demonstrated in this study opens doors for other robust methods under high-dimensional settings. This study will also be practically useful by introducing a new way of analyzing genetic data.

In chapter “A Mixture of Variance-Gamma Factor Analyzers” (McNicholas et al.), a mixture modeling approach for clustering high-dimensional data is developed. This approach is based on a mixture of variance-gamma distributions, which is interesting because the variance-gamma distribution has been underutilized in multivariate statistics—certainly, it has received far less attention than the skew-t distribution, which also parameterizes location, scale, concentration, and skewness. Clustering is carried out using a mixture of variance-gamma factor analyzers (MVGFA) model, which is an extension of the well-known mixture of factor analyzers model that can accommodate clusters that are asymmetric and/or heavy tailed. The formulation of the variance-gamma distribution used can be represented as a normal mean variance mixture, a fact that is exploited in the development of the associated factor analyzers.

In summary, several directions for innovative research in big data analysis were highlighted in this book. I remain confident that this book conveys some of the surprises, puzzles, and success stories in the arena of big data analysis. The research in this arena is ongoing for a foreseeable future.

As an ending thought, I would like to thank all the authors who submitted their papers for possible publication in this book as well as all the reviewers for their valuable input and constructive comments on all submitted manuscripts. I would like to express my special thanks to Veronika Rosteck at Springer for the encouragement and generous support on this project and helping me to arrive at the finishing line.

My special thanks go to Ulrike Stricker-Komba at Springer for outstanding technical support for the production of this book. Last but not least, I am thankful to my family for their support for the completion of this book.

Niagara-On-The-Lake, Ontario, Canada
August 2016

S. Ejaz Ahmed

Contents

Part I General High-Dimensional Theory and Methods

Part II Network Analysis and Big Data

Part III Statistics Learning and Applications