CONTEMPORARY MATHEMATICS

622

Centre de Recherches Mathématiques Proceedings

Perspectives on Big Data Analysis:

Methodologies and Applications

International Workshop on Perspectives
on High-dimensional Data Analysis II
May 30–June 1, 2012
Centre de Recherches Mathématiques,
Université de Montréal, Montréal

S. Ejaz Ahmed
Editor

American Mathematical Society
Providence, Rhode Island

Centre de Recherches Mathématiques
Montréal, Québec, Canada

Perspectives on Big Data Analysis:

Methodologies and Applications

CONTEMPORARY
MATHEMATICS

622

Centre de Recherches Mathématiques Proceedings

Perspectives on Big Data Analysis: Methodologies and Applications

International Workshop on Perspectives
on High-dimensional Data Analysis II
May 30–June 1, 2012
Centre de Recherches Mathématiques,
Université de Montréal, Montréal

S. Ejaz Ahmed
Editor

American Mathematical Society
Providence, Rhode Island

Centre de Recherches Mathématiques
Montréal, Québec, Canada

2010 *Mathematics Subject Classification.* Primary 68T99, 62H25, 62J07, 62F05, 62G05, 62M05, 62G08, 60-XX, 62G99, 62H15.

Library of Congress Cataloging-in-Publication Data

International Workshop on Perspectives on High-Dimensional Data Analysis (2nd : 2012 : Montréal, Québec)

Perspectives on big data analysis : methodologies and applications : International Workshop on Perspectives on High-Dimensional Data Analysis II, May 30–June 1, 2012, Centre de recherches mathématiques, Université de Montréal, Montréal, Québec, Canada / S. Ejaz Ahmed, Editor.

pages cm.– (Contemporary mathematics / Centre de recherches mathématiques proceedings; volume 622)

Includes bibliographical references.

ISBN 978-1-4704-1042-1 (alk. paper)

1. Multivariate analysis–Congresses. 2. Artificial intelligence–Congresses. 3. Big data–Congresses. I. Ahmed, S. E. (Syed Ejaz), 1957-editor of compilation. II. Title.

QA278.I585 2012

519.5′35–dc23 2014000814

Contemporary Mathematics ISSN: 0271-4132 (print); ISSN: 1098-3627 (online)

DOI: http://dx.doi.org/10.1090/conm/622

10 9 8 7 6 5 4 3 2 1 19 18 17 16 15 14

Contents

Preface

This book comprises a collection of research contributions that were presented, with one exception, at the International Workshop on the Perspectives on High-Dimensional Data Analysis II, 2012 at the Centre de recherches mathématiques (CRM), Université de Montréal, Canada. One goal of the workshop was to improve the understanding of high-dimensional modeling from an integrative perspective and to bridge the gap among statisticians, computer scientists and applied mathematicians in understanding each other's tools. It provided a venue for participants to meet leading researchers of this field in a small group in order to maximize the chance of interaction and discussion. The objectives included: (1) to highlight and expand the breadth of existing methods in high-dimensional data analysis and their potential to advance both mathematical and statistical sciences, (2) to identify important directions for future research in the theory of regularization methods, in algorithmic development, and in methodology for different application areas, (3) to facilitate collaboration between theoretical and subject-area researchers, and (4) to provide opportunities for highly qualified personnel to meet and interact with leading researchers from countries around the world.

The discipline of statistical science is ever changing and evolving from investigation of classical finite-dimensional data to high-dimensional data analysis. Indeed, we are commonly experiencing data sets containing huge numbers of variables where in some cases the number of variables exceeds the number of sample observations. Many modern scientific investigations require the analysis of enormous, complex high-dimensional data that is beyond the classical statistical methodologies developed decades ago. For example, genomic and proteomic data, spatial-temporal data, social network data, and many others fall into this category. Modeling and making statistical decisions of high-dimensional data is a challenging problem. A range of different models with increasing complexity can be considered, and a model that is optimal in some sense needs to be selected from a set of candidate models. Simultaneous variable selection and model parameters estimation plays a central role in such investigations. There is a massive literature on variable selection, and penalized regression methods are becoming increasingly popular. Many interesting and useful developments have been published in recent years in scientific and statistical journals.

The application of regression models for high-dimensional data analysis is a challenging task. Regularization/penalization techniques have attracted much attention in this arena. Penalized regression is a technique for mitigating difficulties arising from collinearity and high dimensionality. This approach necessarily incurs an estimation bias, while reducing the variance of the estimator. A tuning parameter is needed to adjust the effect of the penalization so that a balance between model parsimony and goodness-of-fit can be achieved. Different forms of penalty functions have been studied intensively over the last two decades. More recently, some of the regularization techniques have been extended to deal with the estimation of large covariance matrices and the analysis of complex dependence structures such as networks and graphs.

However, the development in this area is still in its infancy. For example, methods may require the assumption of sparsity in the model where most coefficients are exactly 0 and nonzero coefficients are big enough to be separated from the zero ones. There are situations where noise cannot easily be separated from the signal. Furthermore, penalty estimators are not efficient when the number of variables is extremely large compared with the sample size. There are still many theoretical and computationally challenging problems when the number of variables increases at a nonpolynomial rate with the sample size.

This book collates applications and methodological developments in high-dimensional statistics dealing with interesting and challenging problems concerning the analysis of complex, high-dimensional data with a focus on model selection and data reduction. The chapters contained in this book deal with submodel selection and parameter estimation for an array of interesting models. 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 provide interesting applications. All the papers were thoroughly refereed. A brief description of the contents of each of the thirteen papers in this book follows.

Chapter 1 presents frameworks where principal component analysis (PCA) is an effective tool for the analysis of very high-dimensional data. In contrast to the many papers giving negative results on the lack of consistency of the covariance estimates that are the basis of PCA, this chapter shows that certain methods based on PCA are successful in detecting signals while controlling the probability of false discoveries. This chapter presents three such methods and justifies them by asymptotic arguments and by simulation results.

In particle physics, a system of equations known as the Lattice QCD equations plays a Nobel Prize-winning role. Such equations help physicists determine the weight of sub-atomic particles, like the famous Higgs Boson. These equations are archetypal because they also arise in biology, chemistry and engineering, to name a few. One characteristic of these equations is the feature that they form a large p (parameters) small n (number of equations) system. Chapter 2 introduces an unknown hidden parameter which makes the system of equations telescopic, and in doing so mitigates the effects of the large p, small n problem. A prior distribution is then imposed on the hidden parameter, and by using a Bayesian approach coupled with a hierarchical Markov Chain Monte Carlo algorithm, a solution to the system of equations is achieved.

A main goal of a Bayesian analysis of a spatial point pattern is to estimate the underlying intensity function. The estimation can proceed by either specifying a parametric form for the function or by a Bayesian nonparametric model. A common nonparametric model is the Poisson/gamma random field model introduced in 1998 by Wolpert and Ickstadt; they also provided a Levy construction algorithm to estimate the intensity function. In this chapter the authors propose a slice sampling algorithm for a hierarchical Poisson/gamma random field model for multi-type point patterns under the condition that multiple, independent, point patterns of each type are observed. In Chapter 3 it is demonstrated that the slice sampling algorithm is computationally more efficient than a Levy construction algorithm. The authors then demonstrate their model and algorithm on a functional neuroimaging meta-analysis of emotions.

Hidden Markov models (HMMs) have been used in a wide range of applications, including speech recognition and DNA sequence analysis. In some applications, the background of the problem readily indicates the number of states (or order) of the HMM to be fitted. Often, however, the knowledge of the researcher or practitioner is limited, and the order of the model needs to be estimated from the data. In Chapter 4, a new penalized quasi-likelihood method for order estimation in HMMs is proposed. Starting with an HMM with a large number of states, the method clusters and combines similar states through two penalty functions, yielding a model of a suitable lower order. The performance of the new method is assessed theoretically and via extensive simulation studies. Two well-known data sets are also examined to illustrate the use of the new method.

Chapter 5 considers estimation of the linear model regression coefficients as in a partially linear model (PLM) with a diverging number of predictors in a high-dimensional data analysis. In this chapter, a high-dimensional shrinkage estimation strategy is proposed to improve the prediction performance of a PLM based only on a predefined subset. The asymptotic property of the high-dimensional shrinkage estimator is developed, and its relative performances with respect to the full model and submodel estimators using the quadratic loss function are critically assessed. Furthermore, it is shown both analytically and numerically that the proposed high-dimensional shrinkage estimator performs better than the full model estimator, and in many instances, it performs better than penalty estimators.

Different kinds of shrinkage estimators have been proposed in situations where the number of variables dominates the sample size. Interestingly, ridge regression, which is one of the earliest shrinkage methods developed, has been studied minimally under such a paradigm. It would seem natural to do so since correlation between predictors abounds in such situations. Chapter 6 develops a novel geometric interpretation for a generalized ridge regression estimator which lends insights into its properties. Interestingly, it is shown that useful properties seem to exist in truly sparse settings but are not guaranteed in nonsparse settings. This chapter also develops a computationally efficient numerical algorithm for estimation and studies the performance of the procedure in a real data setting, looking at predictors of disease progression in diabetes.

Chapter 7 is concerned with hypothesis testing for a very general class of regression models where the number of covariates p is allowed to exceed the sample size n substantially, leading to ultra-high-dimensional problems. Specifically, motivated by large sample considerations, a new test statistic (which, in applications, can also be based on efficient scores) controlling for the family-wise error rate is developed. A novel multiplier bootstrap methodology with Rademacher weights is introduced for computing the critical values of the test.

Chapter 8 develops a new single-step multiple comparison procedure and derives the related simultaneous confidence intervals for arbitrary contrasts of means in high-dimensional repeated measures designs that are, by construction, consonant and coherent. The key approach is to combine parametric bootstrap with the modern matrix regularization techniques such as banding and thresholding for estimation of covariance matrices. The numerical studies indicate that the new regularized multiple contrast procedures deliver an accurate estimate of Type I error without sacrificing the power of the test, even for a small number of subjects and a substantial number of variables. The proposed testing procedure is also illustrated by an application to a sleep disorder study.

The asymptotic properties of cubic smoothing splines are well known. For example, when utilized as a nonparametric function estimation tool, the rate of convergence of the cubic smoothing splines can be proven to be nearly optimal. The derivation of the cubic smoothing splines depends on a special property of the penalty that is the integral of the squared second-order derivative. Such a definition requires that the second derivative of the underlying function is square integrable, which could sound arbitrary and restrictive to some researchers. Chapter 9 shows that an alternative approach, whose derivation is completely based on data, hence circumventing the aforementioned restriction, can preserve nearly all the nice asymptotic properties that the cubic smoothing splines enjoy, such as conditioning on the predictor variable following an equally spaced design. The key idea in this chapter is to derive the decaying rate of eigenvalues in a newly derived smoothing matrix. The derivation uses some results from functional analysis and operator theory. Numerical experiments demonstrate better performance of the proposed method, as compared with existing counterparts.

Chapter 10 concerns the question of whether the SCAD procedure is applicable to logistic regression models where the dimension of covariates diverges in an exponential rate of the sample size, and theoretically justifies this applicability. In addition, an optimization algorithm is developed by combining the concave convex procedure and the coordinate descent algorithm in solving regularization problems. The numerical study shows the promise of the proposed procedure in various high-dimensional logistic regression settings.

Chapter 11 considers the variable selection and estimation problems for a logistic regression model using the shrinkage and penalty methods. The shrinkage method relies on prior information of inactive predictors when estimating the coefficients of active predictors. On the other hand, the penalty methods identifies inactive variables by producing zero solutions for their associated regression coefficients. Both methods are shown to have higher efficiency than the classical methods for a wide class of models. A large sample theory for the shrinkage estimators, including asymptotic distributional biases and risks, is developed. A Monte Carlo simulation study is conducted for different combinations of inactive predictors, and the performance of each method is evaluated in terms of a simulated mean squared error. This study indicates that if the number of inactive predictors is correctly specified, the shrinkage method would be expected to do better than the penalty method, but if the number of inactive predictors is incorrectly specified, the penalty methods would be expected to do better than the shrinkage methods. A real data example is presented to illustrate the proposed methodologies.

In Chapter 12 a dimensionality reduction method is proposed which has a low computational cost. This method is inspired by the observation that reasonably large chunks of a high-dimensional dataset can be approximated by low-dimensional patches over its underlying manifold.

In summary, several directions for statistical inference in high-dimensional statistics were highlighted by the talks, papers, and the discussion. This volume conveys some of the surprises, puzzles and success stories in big data analysis.

As an ending thought, I would like to thank all the authors who submitted their papers for possible publication in this special issue as well as all the reviewers for their valuable input and constructive comments on all submitted manuscripts. I will take this opportunity to thank all the participants for their amazing contributions and support for the workshop. I would like to express my special thanks to the superb staff at CRM for the encouragement and support in the organization of this workshop. The hospitality was unparalleled and equally appreciated by participants and organizers. Louis Pelletier provided superb local arrangements for the workshop, and André Montpetit outstanding technical support for the production of this volume. Last but not least, I am thankful to Galia Dafni, Christine Thivierge and the staff of the AMS for their support in the completion of this volume.

S. Ejaz Ahmed
St. Catharines, Canada
November 2013

Published Titles in This Subseries

622 **S. Ejaz Ahmed, Editor,** Perspectives on Big Data Analysis, 2014

606 **Chantal David, Matilde Lalín, and Michelle Manes, Editors,** Women in Numbers 2, 2013

605 **Omid Amini, Matthew Baker, and Xander Faber, Editors,** Tropical and Non-Archimedean Geometry, 2013

CRM

The Centre de Recherches Mathématiques (CRM) was created in 1968 to promote research in pure and applied mathematics and related disciplines. Among its activities are special theme years, summer schools, workshops, postdoctoral programs, and publishing. The CRM receives funding from the Natural Sciences and Engineering Research Council (Canada), the FRQNT (Québec), the NSF (USA), and its partner universities (Université de Montréal, McGill, UQAM, Concordia, Université Laval, Université de Sherbrooke and University of Ottawa).

CONM 622

This volume contains the proceedings of the International Workshop on Perspectives on High-dimensional Data Analysis II, held May 30–June 1, 2012, at the Centre de Recherches Mathématiques, Université de Montréal, Montréal, Quebec, Canada.

This book collates applications and methodological developments in high-dimensional statistics dealing with interesting and challenging problems concerning the analysis of complex, high-dimensional data with a focus on model selection and data reduction. The chapters contained in this book deal with submodel selection and parameter estimation for an array of interesting models. The book also presents some surprising results on high-dimensional data analysis, especially when signals cannot be effectively separated from the noise, it provides a critical assessment of penalty estimation when the model may not be sparse, and it suggests alternative estimation strategies. Readers can apply the suggested methodologies to a host of applications and also can extend these methodologies in a variety of directions. This volume conveys some of the surprises, puzzles and success stories in big data analysis and related fields.

Big Data Analysis • Ahmed, Editor

American Mathematical Society
www.ams.org

Centre de Recherches Mathématiques
www.crm.math.ca

ISBN 978-1-4704-1042-1

9 781470 410421

CONM/622

AMS/CRM