Adv Data Anal Classif (2010) 4:85–87
DOI 10.1007/s11634-010-0071-6

PREFACE

Special Issue on Robust Methods for Classification and Data Analysis

Preface by the Guest Editors

Marco Riani · Andrea Cerioli · Peter J. Rousseeuw

Published online: 6 August 2010

It is now widely recognized that most statistical techniques are not resistant to outliers or other deviations from the classical assumptions. Therefore, the development of highly robust and efficient statistical methods has become a goal of paramount importance in both theoretical and applied statistics. The demand for such methods has been driven by the increasing availability of data in almost any area of scientific research. These data sets are not only becoming larger in size, but also in complexity. The extraction of essential features and the discovery of structures and relations in complex data sets must not break down when atypical observations are present. Moreover, there is a need for the development of effective diagnostics that can help to pinpoint these outliers. With many variables at hand, outlying observations can be hard to detect. Outliers need not necessarily be gross errors, but may instead contain valuable information. For instance, by means of robust methods one may discover the existence of several populations instead of one. While robust statistical methods and diagnostic tools are well established for studying data sets in simple univariate models, this is not yet the case for more complicated multivariate situations which may contain multiple groups.

M. Riani (✉) · A. Cerioli
Department of Economics, Faculty of Economics, University of Parma,
Via Kennedy 6, 43125 Parma, Italy
e-mail: mriani@unipr.it
URL: http://www.riani.it

A. Cerioli
e-mail: andrea.cerioli@unipr.it
URL: http://economia.unipr.it/docenti/cerioli

P. J. Rousseeuw
Department of Mathematics, Katholieke Universiteit Leuven, Leuven, Belgium
e-mail: peter@rousseeuw.net
URL: http://www.agoras.ua.ac.be

To promote research in these areas, every year an International Conference on Robust Statistics (ICORS) is held. In 2009 this conference took place in Parma (Italy) and welcomed almost 200 participants from all over the world.

Shortly before the ICORS conference in Parma we reached an agreement with the Editors of Advances in Data Analysis and Classification (ADAC) to publish a Special Issue on topics related to Robust Multivariate Data Analysis and Classification. After the conference a call for papers was circulated widely. Topics of particular interest included but were not limited to:

- Methodological innovations in the fields of robust multivariate data analysis, robust classification and clustering, robust regression, and multivariate outlier detection.
- The development of computational and graphical tools that efficiently implement robust statistical methods in the abovementioned fields.
- Applications of robust multivariate methods in specific domains such as bioinformatics, business, finance, signal processing, data mining, etc.

As a result, 25 papers were submitted. All the papers were refereed following the usual peer-review process, which resulted in at least two reports for each manuscript. The acceptance rate was below 50%. Seven of the accepted papers are published in this Special Issue, including a review paper sketching the state-of-the-art of robust classification. The other accepted papers will appear in subsequent issues of ADAC.

The first paper in this Special Issue, by L.A. Garcia-Escudero, A. Gordaliza, C. Matran and A. Mayo-Iscar, reviews various robust clustering approaches in the literature. It illustrates that deviations from theoretical assumptions and/or the presence of a number of outlying observations are common in cluster analysis, and may lead to unsatisfactory classifications. The paper pays special attention to trimming-based methods which discard outlying data during the clustering process.

The paper by P. Coretto and C. Hennig contains a simulation study which compares several methods for robust clustering based on mixture models. The authors conclude that, although different methods "win" under different setups in the simulation study, the Robust Improper Maximum Likelihood Method can be recommended as optimal in some situations, and acceptable in others.

C. Croux, C. Dehon, and A. Yadine propose a k-step version of the robust Sign Covariance Matrix, which improves its efficiency while keeping the maximal breakdown property. For increasing k, this orthogonally equivariant estimator approaches affine equivariance.

In the next paper, M. Debruyne and T. Verdonck present kernel versions of three robust PCA algorithms: Spherical PCA, Projection Pursuit, and ROBPCA. These robust Kernel PCA (KPCA) algorithms are analyzed in a classification context, by applying discriminant analysis to the KPCA scores. The performance of the various robust KPCA algorithms is studied in a simulation study comparing misclassification percentages, both on clean and contaminated data.

A. Marazzi and V. Yohai consider optimal robust M-estimates of a multidimensional parameter using Hellinger distance in Hampel's infinitesimal approach. The optimal estimates are derived by maximizing an efficiency measure at the model, subject to a bound on the sensitivity to contamination.

S. Van Aelst and G. Willems consider canonical variate analysis based on robust estimates of the group centers and joint scatter matrix. The paper shows how the fast and robust bootstrap method can be used to obtain inference for the robustly estimated canonical variates. It constructs robust confidence intervals to investigate which variables contribute significantly to the canonical variate.

The last paper, by A. Van Messem and A. Christmann, illustrates how support vector machines can be successful methods of classification and regression for analyzing even large data sets with unknown, complex, and high-dimensional dependency structures. More specifically, the authors show how, by shifting the loss function, it is possible to extend the applicability of support vector machines to situations where the output space is unbounded.

As a concluding remark, we are grateful to the Editors of ADAC, and in particular to Professor Hans-Hermann Bock, for giving us the opportunity to produce this Special Issue which we hope will disseminate the concepts related to robust classification and robust multivariate data analysis. Professor Bock's encouragement and wisdom have been an essential support for our work. We also would like to warmly thank the many distinguished scholars who have acted as referees for this Special Issue.

Last but not least, we express our gratitude to both the Minerva Foundation (through Prof. Luisa Fernholz) and the Joint Research Centre (JRC) of the European Commission (through Dr. Domenico Perrotta and Dr. Spyros Arsenis of the Institute for the Protection and Security of the Citizen), because without their support the ICORS Conference in Parma, and consequently this Special Issue, could not have happened.

We look forward to seeing everyone who has contributed to this Special Issue at one of the next ICORS meetings!