Comput Stat (2014) 29:403–405
DOI 10.1007/s00180-013-0468-8

EDITORIAL

Sparse matrices in data analysis

Nickolay Trendafilov · Martin Kleinsteuber · Hui Zou

Published online: 24 December 2013
© Springer-Verlag Berlin Heidelberg 2013

In the last decade, the demand for statistical and computation methods for data analysis that involve sparse matrices has grown dramatically. The main reason for this is that the classical approaches produce solutions in a form of linear combinations of all variables involved in the problem. However, the nowadays applications deal with huge data sets and the interpretation of linear combinations of tens of thousands of variables is virtually an impossible task. The natural escape is to modify the standard techniques to produce sparse solutions which involve only few of the original variables, but still providing competitive goodness-of-fit to the data. Another reason is the increasing number of problems for analysis of sparse data where a portion of the data entries are missing or grossly corrupted. Such problems require modification of the standard approaches to produce robust solutions which, in turn, may also need to be sparse.

This special issue comprises 11 invited contributions of scientists working actively in the area of statistical computing, data analysis, and machine learning. Our main objective is to collect and present recent developments and different aspects of sparse data analysis, both in terms of modeling and numerical realization.

A nice review and empirical study of the high-dimensional regression problem is provided by Bühlmann and Mandozzi (2014). The authors examine an important

N. Trendafilov (✉)
Department of Mathematics and Statistics, The Open University, Walton Hall,
Milton Keynes MK7 6AA, UK
e-mail: nickolay.trendafilov@open.ac.uk

M. Kleinsteuber
TU München, Munich, Germany
e-mail: kleinsteuber@tum.de

H. Zou
University of Minnesota, Minneapolis, MN, USA
e-mail: zouhui2007umn@gmail.com

problem concerning inference for problems with high-dimensional variable screening. Most attention in this topic has focused on prediction, but this paper addresses the parallel problem of identifying the key variables and performing inference. A thorough and realistic simulation experiment is performed. Disappointingly, most of the existing methods perform similarly, and none are highly successful. Nevertheless, this is an important question and the paper stimulates further research.

Principal component analysis (PCA) of high-dimensional data is another fundamental topic in modern multivariate analysis. This special issue has three interesting papers on this topic. The article by Trendafilov (2014) is a good expository article on PCA and sparse PCA. The emphasis is on the interpretation of PCA. He reviews and compares different modern sparse PCA proposals in the last ten years. Fang et al. (2014) consider the problem of estimating the principal components of heredity (PCH) with modern genomic data. They propose a sparse regularized PCH estimator and study its numeric and theoretic performance. Hage and Kleinsteuber (2014) propose a new robust PCA approach in which the core algorithm is an intrinsic conjugate gradient method on the Grassmann manifold. It allows to employ non-convex sparsity penalties that show improved robustness compared to state-of-the-art methods in identifying the underlying principal components in the presence of sparse noise.

The next two contributions are related to high-dimensional classification problems. The article by Bouveyron and Brunet (2014) is an extension of earlier works of the authors on Fisher-EM algorithms such that the new algorithm can produce sparse loading matrices. The Fisher-EM algorithm includes a F-step (F for Fisher) between the E and the M steps, which finds the Fisher's linear discriminants for the current clusters. The article incorporates existing methods for sparse singular value decomposition (SVD) into the F-step of the algorithm to perform a selection of the discriminative variables.

Choy and Meinshausen (2014) use a least absolute shrinkage and selection operator (LASSO) type regularization method to construct a sparse distance metric. Such a construct will be very welcome in different classification techniques applied to high-dimensional problems to discard a great number of unimportant (spam) variables. The authors show that the best sparse metric can be recovered with an exponential deviation bound.

In some real applications the covariance matrix is only partially identifiable. G'Sell et al. (2014) present a method based on semidefinite programming for computing the range of possible equal-likelihood inferred values for the missing entries and affine functions of such a covariance matrix.

Frame theory is a rather new approach in signal modeling in the context of sparse matrices, worthwhile introducing to the statistical community. The basic construct is that linear measurements of a signal are described as a product of a matrix and the signal, where the rows of the matrix consist of the dual of the measurement vectors. A frame is a collection of such measurement vectors and frame theory is about the design of frames for a given class of signals and the reconstruction of the signal, given the measurements. It aims, for example, at reducing dimensionality of the signal to achieve data compression or feature extraction. In this context, the design of sparse frames allows particularly effective numerical implementations and also opens the door for handling high dimensional data. The article by Krahmer et al. (2014) surveys the sparse frame theory and carefully introduces this interesting new topic, while presenting recent research results.

The development of efficient and reliable general algorithms in data analysis is a challenging problem, in particular when large and high-dimensional data are considered. One promising approach is to employ the underlying geometry of the particular problem at hand for designing optimization methods. Such Riemannian manifold approaches often show desirable numerical properties, e.g., numerical stability and minimal dimension of the search space, and, at the same time, having nice convergence behavior in practice. These issues are discussed in the next two contributions by Absil et al. (2014) and by Mishra et al. (2014) which are quite similar in spirit. They provide a geometric framework for optimization on the set of fixed rank matrices. Potential applications include low-rank matrix completion and other low-rank matrix approximation problems.

Another type of optimization methods are considered by Chen and Ye (2014). They study a regularized estimator of multiple predictive functions from a dictionary of basis functions where the coefficient matrix is penalized by the trace norm and the ℓ_1-norm simultaneously. Two efficient algorithms are proposed which are based on the accelerated gradient method and the alternating direction method, respectively, to find the proposed estimator.

We are very grateful to all reviewers for their help in the refereeing process of the contributions and for sharing their time and knowledge. Unfortunately, due to the Journal policy, we cannot list their names. We also want to thank again all authors who have enthusiastically contributed with interesting original works to the Special Issue.

References

Absil P-A, Amodei L, Meyer G (2014) Two Newton methods on the manifold of fixed-rank matrices endowed with Riemannian quotient geometries. Comput Stat 29. doi:10.1007/s00180-013-0441-6

Bouveyron C, Brunet C (2014) Discriminative variable selection for clustering with the sparse Fisher-EM algorithm. Comput Stat 29. doi:10.1007/s00180-013-0433-6

Bühlmann P, Mandozzi J (2014) High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. Comput Stat 29. doi:10.1007/s00180-013-0436-3

Chen J, Ye J (2014) Sparse trace norm regularization. Comput Stat 29. doi:10.1007/s00180-013-0440-7

Choy T, Meinshausen N (2014) Sparse distance metric learning. Comput Stat 29. doi:10.1007/s00180-013-0437-2

Fang Y, Feng Y, Yuan M (2014) Regularized principal components of heritability. Computat Stat 29. doi10.1007/s00180-013-0444-3

G'Sell MG, Shen-Orr S, Tibshirani R (2014) Sensitivity analysis for inference with partially identifiable covariance matrices. Comput Stat 29. doi:10.1007/s00180-013-0451-4

Hage C, Kleinsteuber M (2014) Robust PCA and subspace tracking from incomplete observations using ℓ_0-surrogates. Comput Stat 29. doi:10.1007/s00180-013-0435-4

Krahmer F, Kutyniok G, Lemvig J (2014) Sparse matrices in frame theory. Comput Stat 29. doi:10.1007/s00180-013-0446-1

Mishra B, Meyer G, Bonnabel S, Sepulchre R (2014) Fixed-rank matrix factorizations and Riemannian low-rank optimization. Comput Stat 29. doi:10.1007/s00180-013-0464-z

Trendafilov N (2014) From simple structure to sparse components: a review. Comput Stat 29. doi:10.1007/s00180-013-0434-5