Matilde Bini, Pietro Amenta, Antonello D'Ambra, Ida Camminatiello
Editors

Società Italiana di Statistica

Statistical Methods for Service Quality Evaluation

Book of short papers

9th International Conference **IES 2019** - Innovation & Society -

Statistical evaluation systems at 360°: techniques, technologies and new frontiers

organized by Statistics for the Evaluation and Quality in Services Group of the Italian Statistical Society and European University of Rome

Società Italiana di Statistica

Cuzzolin

Matilde Bini - European University of Rome, Italy
Pietro Amenta - University of Sannio, Italy
Antonello D'Ambra - University of Campania "L. Vanvitelli", Italy
Ida Camminatiello - University of Campania "L. Vanvitelli", Italy
Editors

Prima Edizione: Luglio 2019

Traversa Michele Pietravalle, 8 - 80131 Napoli
Tel. 081 5451143 - Fax 081 7707340
cuzzolineditor@cuzzolin.it
www.cuzzolineditore.com

ISBN: 978-88-86638-65-4

CUZZOLIN EDITORE

2019
IES 2019
Innovation & Society
Statistical evaluation systems at 360°:
techniques, technologies and new frontiers
European University of Rome
Rome, Italy
July 4-5, 2019
SIS Società Italiana di Statistica

Table of Contents

Dealing with outliers in high dimensional data: a COMALS procedure

Maria Anna Di Palma and Michele Gallo

Abstract The growing interest in high dimensional data contributes to the development of new statistical techniques aimed at reducing dimensionality when data are influenced by deviating points. An extreme observation or outlier deviates from the model assumption and severely affects the estimates; as the data quality plays an important role in terms of feasible results, it is thus preferable underweights extremeness. In this context, the Candecomp/Parafac model, a decomposition techniques for high dimensional arrays, is not exempted to be sensible to the presence of extreme observations. The algorithm at the base of the model (Alternating Least Square - ALS) is extremely sensitive to the influence of extremeness reproducing flaw results in the analysis. In this context a robust COMedian algorithm (COMALS) is proposed. The algorithm is based on an incredible fast and accurate procedure able to manage the high dimensionality of the data reporting efficient results at any contamination level.

Key words: Outliers, robust algoritms, robust ALS, CP model

1 Introduction

In recent years, the increasing number of domains in which the multidimensional arrays are matter of interest raised discussion on the proper statistical techniques able to effectively summarize the information when diverging points occurred in data.

M.A. Di Palma
University of Naples "L'Orientale", Department of Human and Social Sciences, P.zza S.Giovanni 30, 80134 Napoli e-mail: mariaannadipalma@gmail.com

M. Gallo
University of Naples "L'Orientale", Department of Human and Social Sciences, P.zza S.Giovanni 30, 80134 Napoli e-mail: mgallo@unior.it

Bini M., Amenta P., D'Ambra A., Camminatiello I. (Editors). Statistical Methods for Service quality evaluation. Cuzzolin Publisher.

A multidimensional array is a block of repeated measurements collected for the same variables on the same occasions (i.e. conditions, times, locations); as each aspect pertains to one dimension, the manifold structure is defined as an N-dimensional array, or if only three dimensions are considered, in a three-way array. Different proposals exist to reduce the three-way array in a few informative factors Kroonenberg (2008) while keeping separate the source of variances of each dimension; here the Candecomp/Parafac model, abbreviated in CP, is considered. The model, independently proposed by Carroll and Chang (1970) and Harshman (1970) solves the complexity of the interrelations between entities describing the variation in several matrices simultaneously with different proportions according to the occasions. The interpretability of results, the uniqueness of the solution under mild conditions and the possibility to represent a complex trilinear structure keeping separate the source of variances of each mode, makes the CP model particularly attractive.

From a mathematical perspective, the model is quite simple, it defines a best low rank approximation of the original array through the Alternating Least Squares algorithm (ALS) which correctly processes the multilinear structure decomposing the array into a number of two-way loading matrices, one for each mode by using the same number of factors. Despite the fast computational power, the full informative procedure, the guaranteed convergence and the improving fit at each iteration step which makes the ALS the preferred algorithm to solve the CP model. In terms of data quality preservation, the deviating points, defined as outliers (Hawkins, 1980), required to be efficiently detected by an algorithm able to mitigate their corruption effect on the estimates including them into the analysis as a source of information (i.e revealing atypical patterns) instead of discarding them.

Different attempts to underweight outliers by the means of a robustification of the ALS algorithm in the CP model were made in literature, however, the only suitable procedure largely studied and applied to the different field of studies (Engelen et al, 2009; Hubert et al, 2012) is the ROBust Principal Component Analysis (ROBPCA) developed by Hubert et al (2005). The procedure even if combines the Fast Minimum Covariance Determinant (FastMCD) (Rousseeuw and Driessen, 1999) with the projection pursuit technique, in order to overcome some deficiencies dealing with high dimensions, inherits all the features of the FastMCD estimator including the affine equivariance property. Its fulfillment, even if desirable, is an interesting turning point in case of three-way data applications. In fact, affine equivariant estimators are particularly sensitive to the increase of the data dimensionality and level of contamination (fraction of outliers exhibits in data) (Hubert et al, 2014), that may imply an over identification of outliers (False Positives or Type I error) and an higher computational time.

Based on these considerations, a new robust algorithm is put forward as a solution to the shortage of efficiency and speed inadequacy; the procedure here proposed, defined COMedian-ALS (COMALS), is conceived to properly estimate the CP parameters and correctly identify outliers in three-way arrays. To improve the computational speed of the robust algorithm and obtain a greater resistance of the estimator to the extreme observations, the affine equivariance property is relaxed;

in fact as (Lopuhaa and Rousseeuw, 1991) found, the affine equivariance property is required only for orthogonal matrices (rigid motion equivariance) and thus no problem arises in case of decomposition techniques, as in case of the CP model.

2 The COMALS algorithm overview

The algorithm relies on two robust measures of location and dispersion, the median (med) and the comedian (COM) (Falk, 1997), respectively. Given two random variables X and Y, the robust scatter measure is defined as $\mathrm{COM}(X,Y) = \mathrm{med}((X - \mathrm{med}(X))(Y - \mathrm{med}(Y)))$; while in case of $X = Y$ the measure turns out to be $\mathrm{COM}(X,Y) = \mathrm{MAD}^2(X)$, where MAD is the median absolute deviation: $(\mathrm{MAD}(X) = \mathrm{med}(|X - \mathrm{med}(X)|))$. Different features makes the comedian and MAD highly interesting (Falk, 1997, 1998), holding some of the desirable properties of a robust estimator). Those measures are efficiently combined in a three-stage procedure.

An overview on the COMALS method of operating is introduced in a three stage process.In (Stage 1) the median and comedian are used to built a robust location vector and a robust scatter matrix; the transformation considered by Maronna and Zamar (2002) turns the covariance matrix into a positive semidefinite matrix so that the robust Mahalanobis distance is computed. According to the robust distance, deviating points are identified and removed from the original data. The clean subset is then processed by the ALS-CP algorithm and parameters are estimated - (Stage 2). Finally, all points are classified into different types of outliers (Hubert et al, 2012) according to their distance with respect to the score space and to the initial data (Score Distance - SD and Residual Distance - RD) - (Stage 3).

The COMALS, as not iterative in its initial identification step (Stage 1), results in a more simple algorithm, coherent against the robust distance methods in literature for high dimensional arrays, still retaining efficiency and robustness.

3 Discussion

The procedure was tested in a simulation study. Simulations findings demonstrated less affected and incredibly accurate estimates at different level of contamination when the fraction of outliers rises in data, a greater computational simplicity when dimensionality increases, a reduced computational time and the absence of arbitrariness in the definition of the initial set of parameters at the base of the algorithm estimates. The procedure is preferred to the benchmark algorithm (ROBALS). Additional analysis were performed even on real datasets (Aspirin data and OECD data) revealing COMALS as a robust method particularly suitable in high dimension. Its computational complexity and the execution time are reduced (compared to ROBALS algorithm); in case of heavy contamination its efficiency is not eroded.

Further studies and development of the algorithm should consider the possibility to mitigate the approximately affine equivariance of the COMALS and extend its use to a broad range of methods as the regression analysis.

References

Carroll J, Chang J (1970) Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition. Psychometrika 35:283–319

Engelen S, Hubert M (2011) Detecting outlying samples in a parallel factor analysis model. Analytica chimica acta 705(1):155–165

Engelen S, Frosch S, Jørgensen BM (2009) A fully robust parafac method for analyzing fluorescence data. Journal of chemometrics 23(3):124–131

Falk M (1997) On MAD and comedians. Annals of the Institute of Statistical Mathematics 49(4):615–644

Falk M (1998) A note on the comedian for elliptical distributions. Journal of Multivariate Analysis 67(2):306–317

Hall P, Welsh A, et al (1985) Adaptive estimates of parameters of regular variation. The Annals of Statistics 13(1):331–341

Harshman R (1970) Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. Tech. Rep. 16: 84. No. 10,085, UCLA Working Papers in Phonetics, University of California, Los Angeles

Hawkins DM (1980) Identification of outliers, vol 11. Springer

Hubert M, Rousseeuw P, Vanden Branden K (2005) ROBPCA: a new approach to robust principal components analysis. Technometrics 47:64–79

Hubert M, Van Kerckhoven J, Verdonck T (2012) Robust parafac for incomplete data. Journal of Chemometrics 26(6):290–298

Hubert M, Rousseeuw P, Vakili K (2014) Shape bias of robust covariance estimators: an empirical study. Statistical Papers 55(1):15–28

Kroonenberg P (2008) Applied Multi-way Data Analysis. Wiley-Interscience, Hoboken, NJ, USA

Lopuhaa HP, Rousseeuw PJ (1991) Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. The Annals of Statistics pp 229–248

Maronna R, Zamar R (2002) Robust estimates of location and dispersion for high-dimensional datasets. Technometrics 44(4)

Rousseeuw PJ, Driessen KV (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41(3):212–223