

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



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



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



© 2019 CUZZOLIN s.r.l.

Traversa Michele Pietravalle, 8 - 80131 Napoli

Tel. 081 5451143 - Fax 081 7707340

cuzzolineditor@cuzzolin.it

www.cuzzolineditore.com

ISBN: 978-88-86638-65-4

Tutti i diritti riservati.

Questa opera è protetta dalla Legge sul diritto d'autore.

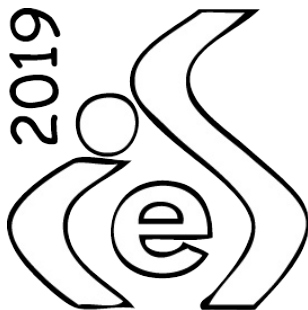
Tutti i diritti, in particolare quelli relativi alla traduzione, alla citazione, alla riproduzione in qualsiasi forma, all'uso delle illustrazioni, delle tabelle e del materiale software a corredo, alla trasmissione radiofonica o televisiva, alla registrazione analogica o digitale, alla pubblicazione e diffusione attraverso la rete internet sono riservati, anche nel caso di utilizzo parziale.

La riproduzione di quest'opera, anche se parziale o in copia digitale, è ammessa solo ed esclusivamente nei limiti della legge ed è soggetta all'autorizzazione dell'editore.

La violazione delle norme comporta le sanzioni previste dalla Legge.

CUZZOLIN EDITORE

2019



IES 2019


Innovation & Society

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

 Società
Italiana di
Statistica

Table of Contents

1 - Preface	1
2 - Plenary Sessions	6
● A general approach to regression modelling using I-priors. <i>Wicher Bergsma, Haziq Jamil</i>	8
● Latent variable models for evaluation systems. <i>Francesco Bartolucci</i>	14
3 - Solleccitated Sessions	19
New Perspectives in Evaluation	19
● Parceling in Multilevel Structural Equation Models for the measure of a latent construct. <i>Sciandra, M., Boscaino, G., Genova, V.</i>	21
● Classification of Italian schools via semi-parametric mixed-effects models. <i>Masci, C., Ieva, F., Paganoni, A.</i>	27
● The ANVUR's system assessing the perceived quality of professors' teaching effectiveness: defining a suitable performance indicator. <i>D'Ambrosio, A., Conversano, C., Ingrassia, S.</i>	32
Student mobility	38
● The analysis of the International Erasmus student mobility network. Insights from empirical data. <i>Primerano, I., Restaino, M., Vitale, M.</i>	40
● Exploring the Italian student mobility flows in higher education. <i>Columbu, S., Porcu, M., Primerano, I., Sulis, I., Vitale, M.</i>	46
● Profiles and educational pathways of students with migratory background: evidences from an Italian Higher Education Institution. <i>Norton, L., Santini, I., Giudici, C., Trappolini, E., Miaci, E.</i>	51
Functional Analysis for Smart Data Applications	56
● Evaluating tourism demand using search queries and functional models. <i>Fortuna, F., Di Battista, T., Gattone, S., Caruso, G.</i>	58
● Functional data analysis of seismic reports from smartphone apps for the real time estimation of earthquake parameters. <i>Finazzi, F., Fassò, A., Wang, Y.</i>	63
● The HOG-FDA Approach with Mobile Phone Data to Modeling the Dynamic of People's Presences in the City. <i>Metulini, R., Carpita, M.</i>	68

itENBIS session: Consistency and Robustness issues in the analysis of customer's preferences 73

- Towards the definition of a Human-Robot Collaboration scale.
Gervasi, R., Mastrogiacomo, L., Franceschini, F. 75
- Robust multivariate analysis of mixed type data.
Granè, A., Salini, S. 82
- A new approach to assess consensus and consistency in customers' preferences.
Vanacore, A., Pellegrino, M., Marmor, Y., Bashkansky, E. 87

4 - Contributed Sessions 92

Regression analysis and applications 1 92

- Analysis of the financial performance in Italian football championship clubs by Marginal Models (GEE) and diagnostic measures.
Venezuela, M., Crisci, A., D'Ambra, L. 94
- Dealing with multicollinearity and outliers in ordinal logit model.
Camminatiello, I., Lucadamo, A. 100
- The concept of food quality in children: a survey on middle school students in Brescia. 105
Golia, S., Simonetto, A., Maio, B., Gilioli, G.
- A two-part finite mixture quantile regression model for semi-continuous longitudinal data.
Maruotti, A., Merlo, L., Petrella, L. 110

Statistical methods for Labour Market 115

- Robust analysis of the labour market.
Corbellini, A., Morelli, G., Magnani, M. 117
- An analysis of the requirements market for job profiles in the Italian manufacturing sector.
Mariani, P., Marletta, A., Masserini, L., Mussini, M., Zenga, M. 122
- From university to job market: an entrepreneurs' view.
Crippa, F., Mariani, P., Marletta, A., Zenga, M. 127
- Integrated Indicators for the Analysis of the Italian Social Security System.
Frenda, A., Scippacercola, S. 132

Environment 136

- Mines and quarries: An analysis of withdrawals determinants in Italy.
Vignani, D., Auci, S. 138
- Climate Change and Italian Main Cities: some Evidence from Meteo-climatic Statistics and Indices on Extreme Events.
Vignani, D., Budano, F., Buseti, C. 143
- A spatial stochastic frontier model for evaluating water use efficiency: the case of fruit and vegetable farming in Apulia (Italy)
Benedetti, I., Laureti, T. 149
- Hierarchical Disjoint Non-Negative Factor Analysis: environment and waste management in the EU.
Cavicchia, C., Sarnacchiaro, P., Vichi, M. 154

Tourism: destinations, household, service evaluation	159
● An index for crowdsourced data on multipoint scales in tourism services evaluation. <i>Tomaselli, V., Cantone, G.</i>	161
● Local authorities and tourist use of the territory. <i>Sergio, S.</i>	166
● Analysing Tourist Destination Image through Topic Modeling. <i>Polizzi, G., Oliveri, A.</i>	171
● Attendance motivations and benefits arising from participation to cultural events. A case study based on the Artecinema documentary festival in Naples (Italy). <i>Ercolano, S., Gaeta, G., Guarino, M., Parenti, B.</i>	176
Statistical methods for Business	181
● Performance evaluation of Italian Firms in the last decade: a Latent Growth Models approach. <i>Bini, M., Masserini, L., Zeli, A.</i>	183
● Network modeling the bike-sharing intention: an empirical analysis of non user needs. <i>Guglielmetti Mugion, R., Musella, F., Vicard, P., Vitale, V.</i>	188
● A Quantile Regression perspective on consumer heterogeneity. <i>Davino, C., Romano, R., Vistocco, D.</i>	193
● How network strategies affect innovation performances in R&D intense industries: a patent-based perspective. <i>Cammarano, A., Caputo, M., La Rocca, M., Michelino, F., Vitale, M.</i>	198
Educational World	203
● Economies of scale and scope in European universities: a large-scale analysis. <i>Bonaccorsi, A., Secondi, L.</i>	205
● New developments in the evaluation of goodness of fit for multidimensional IRT models based on posterior predictive assessment: Results from the INVALSI data. <i>Matteucci, M., Mignani, S.</i>	210
● Women, Education and Empowerment: the case of India. <i>Rinaldi, A., Sciarelli, F.</i>	215
● Classifying Italian Students by Mobility. <i>Casacci, S.</i>	221
New methods for ordinal data	226
● Alternative cumulated chi-squared-type statistics for ordered correspondence analysis. <i>D'Ambra, A., Amenta, P., D'Ambra, L.</i>	228
● The association in three-way contingency tables: Log-ratio analysis vs RC(M) models. <i>Lombardo, R., Camminatiello, I., D'Ambra, A., Beh, E.</i>	233
● A mixture model for ordinal variables measured on semantic differential scales. <i>Manisera, M., Zuccolotto, P.</i>	238
● Partial cumulative correspondence analysis. <i>Amenta, P., D'Ambra, A., Lucadamo, A., D'Ambra, L.</i>	243

Education University	248
● Does joining to groups or Facebook pages help students' retention within the first university year? <i>Bini, M., Masserini, L.</i>	250
● A graduates' multilevel satisfaction index for the evaluation of the university external efficacy. <i>Bacci, S., Bertaccini, B., Dorgali, V., Petrucci, A.</i>	255
● The effect of ICT on the academic performance of Italian students. <i>De Luca, G., Longobardi, S., Pagliuca, M., Regoli, A.</i>	260
Socio demographic analysis	265
● The New Era of Italian Continuous Censuses: Strategies and Perspectives. <i>Bernardini, A., Bonardo, D., Dentini, A., Giacummo, M., Mazziotta, M., Preti, A., Quondamstefano, V.</i>	267
● Nonparametric spatio-temporal interpolation for an Italian real estate index. <i>Cappello, C., De Iaco, S., Palma, M., Pellegrino, D., Posa, D.</i>	272
● A bridge from past to future: the new perspective of the population Census. <i>Falorsi, S., Bernardini, A., Cibella, N., Solari, F.</i>	277
● Life after the storm: the effect of L'Aquila earthquake on marriage rates. <i>Cicatiello, L., Ercolano, S., Gaeta, G., Gallo, M., Parenti, B.</i>	282
Statistical methods 1	287
● A note on the Difference in Differences method with multiplicative effects and unequal size units. <i>Montanari, G., Doretti, M.</i>	289
● Identification of point of attractions and network analysis for GPS tracking data. <i>Abbruzzo, A., Ferrante, M., De Cantis, S.</i>	294
● Interactions in three-way contingency tables: likelihood-ratio vs chi-squared statistic. <i>Rossi, L., Lombardo, R., D'Ambra, A.</i>	299
● Performance permutation analysis of Equality, Convergence and their Combination under the Delphi method. <i>Auciello, M., Bolzan, M., Pesarin, F.</i>	304
Indicators for Micro and Macro Economics	308
● Teaching evaluations: a structural equation modelling application. <i>Basile, A., Cataldo, R., Fano, S., Venitelli, T.</i>	310
● Early estimates of the macroeconomic indicators by electronic payments data in Italy: services value added and turnover index. <i>Righi, A., Ardizzi, G., Gambini, A., Moauro, F., Renzi, N.</i>	315
● Sustainability's dimensions: a bibliometric analysis to create an indicators system. <i>Cataldo, R., Grassia, M., Marino, M., Voytsekhovska, V.</i>	324
● Models for measuring well-being. Which and why?. <i>Mazziotta, M., Pareto, A.</i>	329

Environmental processes, human activities and their interaction	334
● Evaluating the attitudes of europeans towards the environment <i>Punzo, G., Castellano, R., Pagliuca, M., Panarello, D.</i>	336
● Evaluation of geophysical data from campi flegrei caldera with a vector autoregressive model. <i>Scippacercola, S., Petrillo, Z., Mangiacapra, A., Tripaldi, S.</i>	341
● A new hierarchical model-based composite indicator on climate change <i>Cavicchia, C., Vichi, M., Zaccaria, G.</i>	346
Statistical methods 2	351
● Steel in tweets: an analysis to disambiguate short texts. <i>Zola, P., Cortez, P., Brentari, E.</i>	353
● Dealing with outliers in high dimensional data: a COMALS procedure. <i>Di Palma, A., Gallo, M.</i>	358
● Robustness of k-type coefficients as measures of rater consistency. <i>Vanacore, A., Pellegrino, M.</i>	363
● CP model estimation with incorrect rank of factorization on large data sets. <i>Simonacci, V., Gallo, M., Ciavolino, E.</i>	368
Education School	373
● Student satisfaction for Teaching and its impact on drop-out rate. <i>Cafarelli, B., D'Uggento, A., Petrucci, A.</i>	375
● Determinants of student performance in higher education. <i>Basile, A., Cataldo, R., Fano, S.</i>	380
● Teachers job satisfaction in school management system: a multidimensional data analysis. <i>Belfiore, P., Malafronte, P., Sarnacchiaro, P.</i>	385
● Biplot and unfolding models for evaluation of teacher behaviour in the classroom. <i>Bove, G., Catalano, G., Perucchini, P., Serafini, A., Vecchio, G.</i>	390
Statistical methods for Public Services	395
● An application of deep learning to chest disease detection using images and clinical data. <i>Crobu, F., Di Ciaccio, A.</i>	397
● Dynamic efficiency evaluation of Italian judicial system using DEA based Malmquist productivity indexes. <i>Nissi, E., Giacalone, M., Cusatelli, C.</i>	402
● Evaluation of health conditions using POSET approach. <i>Furfaro, E., Pagani, L., Zandarotti, M.</i>	407
● Log-ratio analysis for monitoring tax process. <i>Camminatiello, I., Lombardo, R., Bucicco, C., Valenzano, M.</i>	412

Statistics for financial risks	417
● A statistical model for the investigation of the retail investors' risk assessment ability.	
<i>Castellano, R., Mancinelli, M., Sarnacchiaro, P.</i>	419
● Port throughput allocation within the portfolio framework.	
<i>Resta, M., Persico, L., Parola, F., Satta, G.</i>	426
● Cluster Analysis for mixed data: an application to credit risk evaluation.	
<i>Caruso, G., Gattone, S., Di Battista, T., Fortuna, F.</i>	431
● Evaluate people's behaviour towards risk: a multidimensional problem.	
<i>Bollani, L., Rossi, G., Sciascia, I.</i>	436
Regression analysis and applications 2	441
● The environmental impact of financial insecurity and conservatism in Germany.	
<i>Panarello, D.</i>	443
● Model-based analysis of a crossover study on aircraft passenger comfort.	
<i>Vanacore, A., Percuoco, C.</i>	448
● Reduced K-means Principal Component Multinomial Regression with external information.	
<i>Lucadamo, A., Amenta, P.</i>	453
● Diagnostic tools for Ordered Logit Models in an analysis of relationships between Public Service Motivation and Individual Performance.	
<i>Crisci, A., D'Ambra, L., Palma, R.</i>	458



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 algorithms, 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

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 $COM(X, Y) = \text{med}((X - \text{med}(X))(Y - \text{med}(Y)))$; while in case of $X = Y$ the measure turns out to be $COM(X, Y) = \text{MAD}^2(X)$, where MAD is the median absolute deviation: $(\text{MAD}(X) = \text{med}(|X - \text{med}(X)|))$. 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 build 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