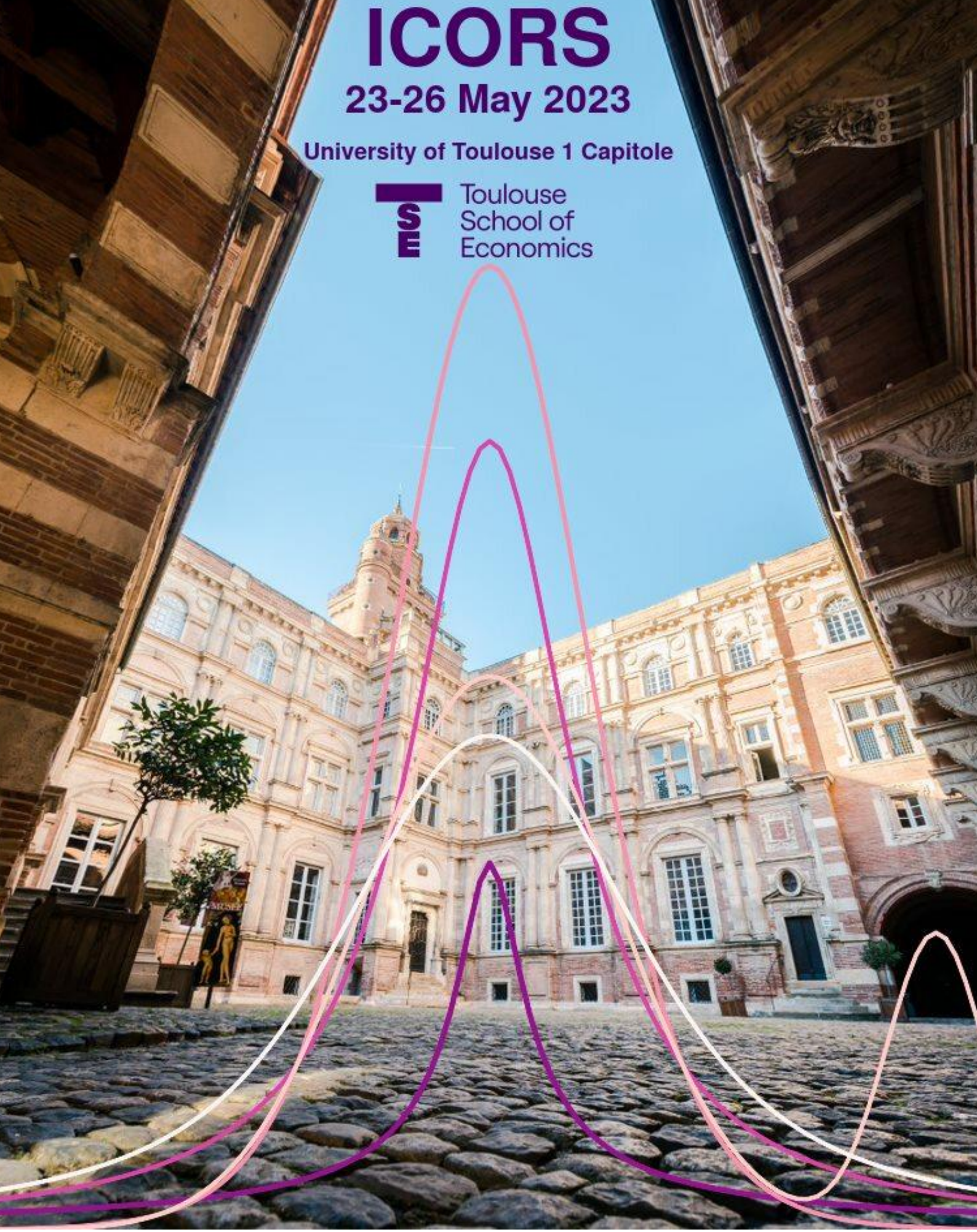


Book of Abstracts

Table of contents

Preface

Welcome to the International Conference on Robust Statistics 2023 in Toulouse, hosted by the vibrant city of Toulouse and organized at Toulouse School of Economics!

This conference will gather researchers, scholars, and practitioners from various corners of the globe, presenting over 80 contributions on the exciting field of robust statistics.

Within the pages of this booklet, you will find a compilation of all the submitted and accepted abstracts for ICORS2023. This collection represents the diverse range of ideas and discoveries that will be shared during the conference, offering a glimpse into the cutting-edge research being conducted in the field of robust statistics.

The scientific and organizing committees hope that this conference will prove to be a fruitful and enriching experience for all participants, fostering new collaborations, triggering innovative ideas, and providing a platform for meaningful discussions.

On behalf of the organizing committee, we would like to express our sincere gratitude to all those who have contributed to making this conference a reality: Airbus, Erasmus School of Economics, Insee, Meetings Toulouse, Société Française de Statistique, Toulouse School of Economics and Université Toulouse Capitole.

We wish you an inspiring and memorable conference experience in Toulouse.

Bienvenue à Toulouse!

For the scientific and organizing committees,
Anne Ruiz-Gazen

Steering committee

- Claudio Agostinelli
- Olcay Arslan
- Ayan Basu
- Claudia Becker
- Ana Bianco
- Graciela Boente
- Andrea Cerioli
- Holger Cevallos
- Ray Chambers
- Shoja'eddin Chenouri
- Christophe Croux
- Juan Cuesta
- Rudolf Dutter
- Luisa Fernholz
- Chris Field
- Peter Filzmoser
- Luis-Angel Garcia-Escudero
- Ursula Gather
- Alfonso Gordaliza
- Marc Hallin
- Xuming He
- Mia Hubert
- Jana Jurečková
- Ricardo Maronna
- Carlos Matran
- Agustin Mayo-Iscar
- Stephan Morgenthaler
- Klaus Nordhausen
- Hannu Oja
- Daniel Peña
- Rosario Oliveira
- Marco Riani
- Isabel Rodrigues
- Elvezio Ronchetti
- Peter Rousseeuw
- Garth Tarr
- David E. Tyler
- Stefan Van Aelst
- Tim Verdonck
- Maria-Pia Victoria-Feser
- Roy Welsch
- Alan Welsh
- Victor Yohai
- Ruben H. Zamar

Scientific committee

- Andreas Alfons (Erasmus School of Economics (ESE), The Netherlands)
- Aurore Archimbaud (Erasmus School of Economics (ESE), The Netherlands)
- Graciela Boente (University of Buenos Aires, Argentina)
- Eva Cantoni (Université de Genève, Switzerland)
- Christophe Croux (EDHEC Business School, France)
- David Haziza (University of Ottawa, Canada)
- Mia Hubert (University of KU Leuven, Belgium)
- Rik Lopuhaä (Delft University of Technology, Netherlands)
- Klaus Nordhausen (University of Jyväskylä, Finland)
- Davy Paindaveine (Université Libre de Bruxelles, Belgium)
- Anne Ruiz-Gazen (Toulouse School of Economics)
- Marco Riani (University of Parma, Italy)
- Matías Salibián-Barrera (University of British Columbia, Canada)

Organizing committee

- Anne Ruiz-Gazen - Co-Chair (Toulouse School of Economics (TSE))
- Aurore Archimbaud - Co-Chair (Erasmus School of Economics (ESE))
- Aline Soulié (TSE)
- Corinne Vella (TSE)
- Thibault Laurent (TSE)
- Sandrine Casanova (TSE)
- Abdelaati Douia (TSE)
- Eve Leconte (TSE)
- Zaïneb Smida (TSE)
- Cristine Thomas-Agnan (TSE)

Tuesday 23 May 2023

Keynote - Lan Wang (A3)

Statistical Learning for Individualized Decision Rules: A Quantile Approach

Lan Wang

Centennial Endowed Professor, Department of Management Science, Miami Herbert Business School, University of Miami, USA

The problem of finding the optimal individualized decision rule (IDR) or a series of sequential individualized decision rules based on individual characteristics is important for applications in precision medicine, government policies, targeted marketing, and other areas. Existing work has been mainly focused on the mean-optimal IDR, which if followed by the whole population would yield the largest average outcome (assuming a larger outcome is preferable). For a variety of applications, the mean may not be the most sensible metric, for example, when the outcome has a skewed distribution. It has also been observed that due to the heterogeneity in treatment response, the estimated mean-optimal IDR may be suboptimal or even detrimental to certain disadvantaged subpopulations. This talk will discuss how the quantile criterion can be used independently or in conjunction with the mean criterion to address these challenges, the related new methodology, and statistical theory.

I1: Robust multivariate statistics I (A3)

The cellwise Minimum Covariance Determinant estimator

Jakob Raymaekers[1] and Peter J. Rousseeuw[2*]

[1] *Department of Quantitative Economics, Maastricht University, The Netherlands; j.raymaekers@maastrichtuniversity.nl*
[2] *Section of Statistics and Data Science, University of Leuven, Belgium; peter@rousseeuw.net*
[*] *Presenting author.*

Keywords. *Cellwise outliers; Covariance matrix; Likelihood; Missing values; Sparsity.*

The usual Minimum Covariance Determinant (MCD) estimator of a covariance matrix is robust against casewise outliers. These are cases (that is, rows of the data matrix) that behave differently from the majority of cases, raising suspicion that they might belong to a different population. On the other hand, cellwise outliers are individual cells in the data matrix. When a row contains one or more outlying cells, the other cells in the same row still contain useful information that we wish to preserve. We propose a cellwise robust version of the MCD method, called cellMCD. Its main building blocks are observed likelihood and a sparsity penalty on the number of flagged cellwise outliers. It possesses good breakdown properties. We construct a fast algorithm for cellMCD based on concentration steps (C-steps) that always lower the objective. The method performs well in simulations with cellwise outliers, and has high finite-sample efficiency on clean data. It is illustrated on real data with visualizations of the results.

Robust Fitting for Generalized Additive Models for Location, Scale and Shape

E. Cantoni[1*]

[1] *Research Center for Statistics and Geneva School of Economics and Management, University of Geneva; eva.cantoni@unige.ch*
[*] *Presenting author*

Keywords. *Bounded influence function; Nonparametric distributional regression; Penalized smoothing splines; Robust smoothing parameter selection; Robust information criterion.*

1 Abstract

The validity of estimation and smoothing parameter selection for the wide class of generalized additive models for location, scale and shape (GAMLSS) relies on the correct specification of a likelihood function. Deviations from such assumption are known to mislead any likelihood-based inference and can hinder penalization schemes meant to ensure some degree of smoothness for non-linear effects. We propose a general approach to achieve robustness in fitting GAMLSSs by limiting the contribution of observations with low log-likelihood values. Robust selection of the smoothing parameters can be carried out either by minimizing information criteria that naturally arise from the robustified likelihood or via an extended Fellner-Schall method. The latter allows for automatic smoothing parameter selection and is particularly advantageous in applications with multiple smoothing parameters. We also address the challenge of tuning robust estimators for models with non-linear effects by proposing a novel median downweighting proportion criterion.

Examples of applications from for [Aeberhard et al. , 2021] and [Ranjbar et al. , 2021] for continuous and discrete responses variables will be presented.

This is joint work with W. Aeberhard, V. Chavez-Demoulin, K. Jaton, G. Marra, R. Radice and S. Ranjbar.

References

W. Aeberhard, E. Cantoni, G. Marra & R. Radice (2021). Robust Fitting for Generalized Additive Models for Location, Scale and Shape. *Statistics and Computing*, **31 (11)**.

S. Ranjbar, E. Cantoni, V. Chavez-Demoulin, G. Marra, R. Radice & K. Jaton (2022). Modelling the extremes of seasonal viruses and hospital congestion: The example of flu in Swiss hospital. *Journal of the Royal Statistical Society - Series C (Applied Statistics)*, **71 (4)**, 884–905.

S-estimation in linear models with structured covariance matrices

H.P. Lopuhaä[1*], V. Gares[2] and A. Ruiz-Gazen[3]

[1] *DIAM, Delft University of Technology;* `h.p.lopuhaa@tudelft.nl`.
[2] *Institut National des Sciences Appliqués de Rennes*
[3] *TSE, University of Toulouse Capitole*
[*] *Presenter*

Keywords. *S-estimators and S-functionals; Breakdown point; Influence function; Asymptotic distribution, Structured covariance.*

Linear models are widely used and provide a versatile approach for analyzing correlated responses, such as longitudinal data, growth data or repeated measurements. In such models, each subject i, $i = 1, \ldots, n$, is observed at k_i occasions, and the vector of responses $\mathbf{y}_i$ is assumed to arise from the model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i,$$

where $\mathbf{X}_i$ is the design matrix for the ith subject and $\mathbf{u}_i$ is a vector whose covariance matrix can be used to model the correlation between the responses. One possibility is the linear mixed effects model, in which the random effects together with the measurement error yields a specific covariance structure depending on a vector $\boldsymbol{\theta}$ consisting of some unknown covariance parameters. Other covariance structures may arise, for example if the $\mathbf{u}_i$ are the outcome of a time series.

To be resistant against outliers, robust methods have been investigated for linear mixed effects models. This mostly concerns S-estimators, originally introduced in the multiple regression context by Rousseeuw & Yohai [1984] and extended to multivariate location and scatter in Davies [1987], to multivariate regression in Van Aelst & Willems [2005], and to linear mixed effects models in Copt & Victoria-Feser [2006].

In this talk I will provide a unified approach to S-estimation in balanced linear models with structured covariance matrices. The balanced setup is already quite flexible and includes several specific multivariate statistical models. Of main interest are S-estimators for linear mixed effects models, but our approach also includes S-estimators in several other standard multivariate models, such as multiple regression, multivariate regression, and multivariate location and scatter. I will provide sufficient conditions for the existence of S-functionals and S-estimators, establish

their asymptotic properties, such as consistency and asymptotic normality, and derive their robustness properties in terms of breakdown point and influence function. All results are obtained for a large class of identifiable covariance structures, and are established under very mild conditions on the distribution of the observations, which goes far beyond models with elliptically contoured densities. In this way, some of the results are new and others are more general than existing ones in the literature.

Existence of S-estimators and S-functionals is established under mild conditions. Although existence of the estimators seems a basic requirement, such results are missing for instance for multivariate regression and for linear mixed effects models. Robustness properties for S-estimators, such as breakdown point and influence function, are obtained under mild conditions on collections of observations and under mild conditions on the distribution of the observations. High breakdown and a bounded influence function seem basic requirements for a robust method, but both properties are not available for linear mixed effects models. For multivariate regression, the influence function is only determined at distributions with an elliptical contoured density. Finally, consistency and asymptotic normality for S-estimators are established under mild conditions on the distribution of the observations. A rigorous derivation is missing for multivariate regression, or is only available for observations from a normal distribution.

The asymptotic results, such as influence function and asymptotic normality, are applied to the special case for which the distribution of the observations corresponds to an elliptically contoured density. Somewhat surprisingly, the asymptotic variances of our S-estimators for linear mixed effects models in which the response has an elliptically contoured density, differ from the ones found in Copt & Victoria-Feser [2006]. This difference is investigated by means of a simulation study.

References

Copt, S. & Victoria-Feser, M.-P. (2006). High-breakdown inference for mixed linear models. *Journal of the American Statistical Association*, **101**(473):292–300, 2006.

Davies,P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *Ann. Statist.*, **15**(3):1269–1292, 1987.

Rousseeuw, P.J. & Yohai, P. (1984). Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis (Heidelberg, 1983)*, volume **26** of *Lect. Notes Stat.*, pages 256–272. Springer, New York, 1984.

Van Aelst, S. & Willems, G. (2005) Multivariate regression S-estimators for robust estimation and inference. *Statist. Sinica*, **15**(4):981–1001, 2005.

C1: Non and semi parametrics (A5)

Robust density estimation in total variation distance under a shape constraint

Y. Baraud[1], H. Halconruy[1] and **G. Maillard**[1]*

[1] *University of Luxembourg; yannick.baraud@uni.lu, helene.halconruy@devinci.fr, guillaume.maillard@uni.lu.*
**Presenting author*

Keywords. *Robust Estimation; Density Estimation; Nonparametric Estimation; Shape Constraints.*

1 Introduction

We solve the problem of estimating the distribution of presumed i.i.d observations for the total variation loss. Our approach is based on density models and is versatile enough to cope with many different ones, including some density models for which the Maximum Likelihood Estimator (MLE for short) does not exist. We mainly illustrate the properties of our estimator on models of densities on the line that satisfy a shape constraint. We show that it possesses some similar optimality properties, with regard to some global rates of convergence, as the MLE does when it exists. It also enjoys some adaptation properties with respect to some target densities in the model for which our estimator is proven to converge at a parametric rate. More important is the fact that our estimator is robust, not only with respect to model mis-specification, but also to contamination, the presence of outliers among the dataset and the equidistribution assumption. This means that the estimator performs almost as well as if the data were i.i.d with density p in a situation where these data are only independent and the average of the marginal densities is close in total variation to a distribution with density p. We also show that our estimator converges to the average density of the data, when this density belongs to the model, even when none of the marginal densities belong to it. Our main result on the risk of the estimator takes the form of an exponential deviation inequality which is non-asymptotic and involves explicit numerical constants. We deduce from it several global rates of convergence, including some bounds for the minimax L^1 risks over the sets of concave and log-concave densities. These bounds derive from some specific results on the approximation of densities which are monotone, convex, concave and log-concave. Such results may be of independent interest.

2 Main results

Let $X_1, \ldots, X_n$ be independent real-valued random variables with marginal densities $(p_i^*)_{1\leq i\leq n}$. Let

$$p^* = \frac{1}{n}\sum_{i=1}^{n} p_i^*.$$

Consider the models $\mathcal{M}_k, \mathcal{M}_k^1$ of densities which are respectively piecewise monotone and piecewise convex or concave on a subdivision of $\mathbb{R}$ into k intervals, as well as the model $\mathcal{M}^{LC}$ of log-concave densities on the line. Let $\mathcal{O}_D(\mathcal{M})$ denote the subsets of $\mathcal{M}$ consisting of piecewise constant, piecewise affine and piecewise log-affine densities with D pieces, when $\mathcal{M} = \mathcal{M}_k, \mathcal{M}_k^1$ or $\mathcal{M}^{LC}$, respectively. Our density estimators $\hat{p}_{\mathcal{M}}$ on $\mathcal{M} \in \left\{\mathcal{M}_k, \mathcal{M}_k^1, \mathcal{M}^{LC}\right\}$ are such that

$$\mathbb{E}\left[\|\hat{p}_{\mathcal{M}} - p^*\|_1\right] \leq \inf_{D\geq 1}\left\{3\inf_{p\in\mathcal{O}_D(\mathcal{M})}\|p-p^*\|_1 + c\sqrt{\frac{D+k}{n}}\right\},$$

where c is a numerical constant (independent of k) and $\|\cdot\|_1$ denotes the L^1 norm [Baraud et al., 2022, Theorems 2,5,8].

This oracle inequality implies that our method is robust and adaptive. Bounds on the approximation error yield results which match the (minimax-optimal) performance of the MLE when it is defined, both in the log-concave and monotone settings. We are also able to handle unbounded densities with infinite support, including heavy-tailed ones [Baraud et al., 2022, Corollary 2], which is new to the best of our knowledge. In the case of bounded convex or concave densities supported on a bounded interval, we also obtain the following new risk upper bound [Baraud et al., 2022, Theorem 7]

$$\mathbb{E}\left[\|\hat{p}_{\mathcal{M}_3^1} - p^*\|_1\right] \leq \frac{C}{n^{4/5}}\log^{2/5}\left(1+\sqrt{2\Gamma LV}\right) + \frac{C}{\sqrt{n}},$$

where C is a numerical constant, L is the length of the support, V the total variation of p^* on its support and $\Gamma \in [0,1]$ is zero when p^* is affine on its support.

References

Baraud, Y., Halconruy, H. & Maillard, G. (2022). Robust density estimation with the $\mathbb{L}^1$ loss. Applications to the estimation of a density on the line satisfying a shape constraint. *arXiv e-prints:* 2205.10524.

Robust estimators for semiparametric moment condition models

A. Keziou[1], **A. Toma**[2,3*]

[1] *Laboratoire de Mathematiques de Reims, Universite de Reims, Champagne Ardenne, UFR Sciences, Moulin de la Housse, B.P. 1039, 51687 Reims, France; amor.keziou@univ-reims.fr*
[2] *Department of Applied Mathematics, Bucharest University of Economic Studies, Piaţa Romană, no. 6, Bucharest, Romania; aida.toma@csie.ase.ro)*
[3] *"Gh. Mihoc - C. Iacob" Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy, Bucharest, Romania*
[*] *Presenting author*

Keywords. *Moment condition models; Divergences; Robustness.*

1 Abstract

We present robust minimum empirical divergence estimators for moment condition models, based on truncated orthogonality functions and dual forms of divergences. Asymptotic properties of these estimators are presented and discussed. For moment condition models invariant with respect to additive or multiplicative transformations groups, these estimators are also equivariant. For models invariant with respect to additive groups, Pitman type estimators are proposed. Approximations of the Pitman estimator are given and it is shown that these approximations represent robust estimators for the model parameter. Some examples based on Monte Carlo simulations illustrate the performance of the estimation methods.

Robust estimation under a semiparametric propensity model for nonignorable missing data

Samidha Shetty[1], Yanyuan Ma[1] and **Jiwei Zhao**[2*]

[1] *Penn State University*
[2] *University of Wisconsin-Madison; jiwei.zhao@wisc.edu*
[*] *Presenting author*

Keywords. *Efficient influence function; Model misspecification; Nonignorable missing data; Robust estimation; Semiparametric statistics.*

1 General Information

Handling missing data is an inevitable issue in many empirical studies, especially when the data are directly collected from human beings or strongly linked to the subjects' behaviors. It is called missing at random or ignorable if the propensity of missing data depends only on the observed values. In applications, however, the propensity is usually nonignorable in the sense that its distribution not only depends on the observed data but also the unobserved ones. Nonignorable missing data are prevalent in surveys in social sciences as well as in patient reported outcomes in biomedical studies.

In this paper, we denote Y as the outcome variable and $\mathbf{X}$ a p-dimensional covariate. We consider the situation that $\mathbf{X}$ is fully observed but Y is subject to nonignorable missingness. We encode R as the indicator of observing Y in that $R = 1$ if Y is observed and $R = 0$ otherwise. Throughout, the propensity model is defined as

$$\pi(y, \mathbf{x}) \equiv \mathrm{pr}(R = 1 \mid y, \mathbf{x}),$$

which does not necessarily depend on all variables in $\mathbf{X}$ but does depend on Y because of the nonignorable missingness. The key question is: what type of assumption is appropriate on the propensity model $\pi(y, \mathbf{x})$ for nonignorable missing outcome data?

By constructing sufficiently many estimating equations, Shao & Wang [2016] showed that the parameter $\boldsymbol{\beta}$ can be consistently estimated if one can find some part of $\mathbf{X}$, say, $\mathbf{Z}$, that is not involved in $\pi(y, \mathbf{x})$; that is, if $\mathbf{X} = (\mathbf{U}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}})^{\mathrm{T}}$ and

$$\text{logit } \pi(y, \mathbf{x}) = \text{logit } \pi(y, \mathbf{u}, \boldsymbol{\beta}, g) = h(y, \boldsymbol{\beta}) + g(\mathbf{u}), \tag{1}$$

where the dimensions of $\mathbf{U}$ and $\mathbf{Z}$ are q and $p-q$, respectively. The variable $\mathbf{Z}$ is termed nonresponse instrument or shadow variable in the nonignorable missing data literature. In the estimation method presented in Shao & Wang [2016], however, an estimator of $g(\mathbf{u})$ is required. Although $g(\mathbf{u})$ per se does not have missing data, its estimation is not stand-alone and involves modeling of missing data or estimation of other unknown components. Shao & Wang [2016] adopted the profiling approach, and estimated $\widehat{g}(\mathbf{u}, \boldsymbol{\beta})$ via the standard kernel estimation for every fixed value of $\boldsymbol{\beta}$. Since $\widehat{g}(\mathbf{u}, \boldsymbol{\beta})$ has to be repeatedly estimated in the algorithm, the whole procedure is computationally expensive. Unfortunately, this approach brings tremendous complexity for end-users and it limits the applicability of the semiparametric propensity model to wider practice.

2 Our Contribution

In this paper, we adopt the model in (1) and propose a completely novel estimation framework which consistently estimates the unknown quantities of interest, without either estimating or modeling the nonparametric component $g(\mathbf{u})$. In other words, our method has the robustness property against the misspecification of $g(\mathbf{u})$.

We first consider the estimation of $\boldsymbol{\beta}$. Through extensive and careful derivations, we discover that our framework allows the consistent estimation of $\boldsymbol{\beta}$ without correctly estimating $g(\mathbf{u})$. We then extend our framework to the estimation of $\boldsymbol{\theta}$ that satisfies $E\{\boldsymbol{\zeta}(\mathbf{X}, Y, \boldsymbol{\theta})\} = \mathbf{0}$ for a given function $\boldsymbol{\zeta}(\cdot)$. We show that, once $\boldsymbol{\beta}$ can be consistently estimated without estimating $g(\mathbf{u})$, $\boldsymbol{\theta}$ can also be consistently estimated without estimating $g(\mathbf{u})$. Note that, $\boldsymbol{\theta}$, such as $E(Y)$, is the quantity of interest in most applications, but here we single out $\boldsymbol{\beta}$ as a quantity of interest as well because its estimation is crucial for estimating $\boldsymbol{\theta}$. In other words, $\boldsymbol{\beta}$ is of interest due to its supporting role.

The nonparametric component $g(\mathbf{u})$ is regarded as a nuisance in a semiparametric model, and its effect to estimating $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is projected to an orthogonal direction via the semiparametric treatment. Our estimation procedure only needs a working model of $g(\mathbf{u})$ for the implementation, and this working model does not have to contain the true $g(\mathbf{u})$. Thus, our method has the robustness property against the misspecification of $g(\mathbf{u})$. Importantly, we find out that such a robustness does not inherit from the standard robustness in the traditional semiparametric literature, where often the orthogonal projection itself is sufficient to guarantee the robustness.

References

Shao, J. & Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, **103**, 175–187.

I2: Functional outlier detection with industrial applications (A3)

Anomaly detection using data depth: functional setting

P. Mozharovskyi[1]

[1] *LTCI, Telecom Paris, Institut Polytechnique de Paris;*
pavlo.mozharovskyi@telecom-paris.fr

Keywords. *Anomaly detection; Functional data depth; Robustness; Visualization; Computational statistics.*

Anomaly detection [Chandola et al., 2009] is a branch of machine learning which aims at identifying observations that exhibit abnormal behavior. Be it measurement errors, disease development, severe weather, production quality default(s) (items) or failed equipment, financial frauds or crisis events, their on-time identification, isolation and explanation constitute a necessary task in almost any branch of industry and science. Since contemporary technological level allows for recording large amounts of data at potentially high frequency, the question of anomaly detection in the functional setting becomes increasingly important. This is amplified by the richness of the infinitely dimensional space and non-negligible occurrence probability of unexpected types of anomalies, *i.e.* those not present in the training sample. This multitude of abnormalities demands a non-parametric methodology for their identification, while robustness requirement suggests data treatment directly in the functional space (different to first projecting it onto a finite-dimensional basis).

Current presentation advocates that—in a number of practical situations—functional data depth [Nieto-Reyes & Battey, 2016, Gijbels & Nagy, 2017] appears to be an efficient tool for anomaly detection. Data-depth-based methodology treats observations directly as functions, thus transferring depth's robustness properties, being crucial for the anomaly detection task, to functional data. Though still retaining computational challenges, today, data depth methodology includes a number of functional depth notions that possess (together with robustness) such attractive properties as non-parametricity and desired invariances, with functional halfspace [Claeskens et al., 2014], area-of-the-convex-hull [Staerman et al., 2020], or curve [Lafaye De Micheaux et al., 2021] depths being only a few examples.

A natural question arises: which depth notions are better suited for the functional anomaly detection problem at hand? Hubert et al. [2015] suggest a taxonomy of abnormal observations, but complexity of the functional space (i) embroils attribution of real-data anomalies to pre-defined types and (ii) generates a multitude of

case-specific sorts of anomalies; see Staerman et al. [2022] for a detailed benchmark study that involves real data. This work is thus an attempt to provide practically important insights into choice and application of depth notions by show-casing their usefulness for anomaly detection in different settings and by benchmarking with the state-of-the-art methods. Simulated and real data explored in the experiments here are expected to attract attention and gain applicant's trust to the depth methodology.

References

Chandola, V., Banerjee, A. & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, **41**(3):15, 1–58.

Claeskens, G., Hubert, M., Slaets, L. & Vakili, K. (2014). Multivariate functional halfspace depth. *Journal of the American Statistical Association*, **109**(505), 411–423.

Gijbels, I. & Nagy, S. (2017). On a general definition of depth for functional data. *Statistical Science*, **32**(4), 630–639.

Hubert, M., Rousseeuw, P.J. & Segaert, P. (2015). Multivariate functional outlier detection. *Statistical Methods & Applications*, **24**(2), 177–202.

Lafaye De Micheaux, P., Mozharovskyi, P. & Vimond, M. (2021). Depth for curve data and applications. *Journal of the American Statistical Association*, **116**(536), 1881–1897.

Nieto-Reyes, A. & Battey, H. (2016). A topologically valid definition of depth for functional data. *Statistical Science*, **31**(1), 61–79.

Staerman, G., Adjakossa, E., Mozharovskyi, P., Hofer, V., Sen Gupta, J. & Clémençon, S. (2022). Functional anomaly detection: a benchmark study. *International Journal of Data Science and Analytics*, `https://doi.org/10.1007/s41060-022-00366-5`.

Staerman, G., Mozharovskyi, P. & Clémençon, S. (2020). The area of the convex hull of sampled curves: a robust functional statistical depth measure. In: Chiappa, S & Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (AISTATS 2020), **108**, 570–579.

Functional Shape based Features in Multivariate Functional Data Applied to Atmospheric Turbulence Online Prediction

Lejeune Clément[1*], Li Tianyi[2*]

[1] *Thales. clement-c.lejeune@thalesgroup.com*
[2] *Airbus and Institut de Recherche en Informatique de Toulouse. tianyi.li@airbus.com*
[*] *Presenting authors*

Keywords. *functional data analysis; atmospheric turbulence prediction; time series; anomaly; outlier*

1 Abstract

Multivariate functional data refer to a sample of multivariate functions generated by a system involving dynamic parameters depending on continuous variables (e.g., multivariate time series). Outlier detection in such a context is a challenging problem because both the individual behavior of the parameters and the dynamic correlation between them are important. We propose identifying outliers in multivariate functional data whereby different outlying features are captured based on mapping functions from differential geometry Lejeune et al. [2020a,b]. In this regard, we extract shape features reflecting the outlyingness of a curve with a high degree of interpretability. As an industrial application, we consider the task of flight online turbulence detection since the relationships between multiple parameters (i.e. aircraft sensors) contain useful information indicating upcoming turbulence Li et al. [2021].

Although it is possible to apply the offline method in real time by adding a sliding window, such a way of approximating raw time series needs to store all the data in the window. Indeed, the usual functional data approximation uses least squares to calculate the coefficients of the basis functions which demands the availability of the entire time series. To tackle this problem, we propose a new method called FUTURA for FUnctional shape feature for real time TURbulence Alerting. Compared to the original shaped based anomaly detection original method, FUTURA couples steady state Kalman Filter, instead of least squares, for functional approximation and extract the functional shape feature with mapping functions in a recursive way. FUTURA not only makes real-time turbulence prediction become possible due to its incremental nature and low computation complexity, but also captures the dynamic

relation between multiple variables which helps identify and predict turbulence. The experimental results show that FUTURA can predict 36.5% of the severe turbulence cases (true positive rate) thirty seconds in advance while keeping a zero false positive rate, which meets the zero false alarm requirement for optimizing the passenger experience and the aircraft operational reliability.

References

Clément Lejeune, Josiane Mothe, Adil Soubki, and Olivier Teste. Shape-based outlier detection in multivariate functional data. *Knowledge-Based Systems*, 198: 105960, 2020a.

Clément Lejeune, Josiane Mothe, and Olivier Teste. Outlier detection in multivariate functional data based on a geometric aggregation. 2020b.

Tianyi Li, Philippe Goupil, Josiane Mothe, and Olivier Teste. Early detection of atmospheric turbulence for civil aircraft: A data driven approach. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 1087–1093. IEEE, 2021.

Robust Profile Monitoring for Multivariate Functional Data

C. Capezza[1], **F. Centofanti**[1*], A. Lepore[1] and B. Palumbo[1]

[1]*Department of Industrial Engineering, University of Naples Federico II, Naples, Italy; christian.capezza@unina.it, fabio.centofanti@unina.it, antonio.lepore@unina.it, biagio.palumbo@unina.it*
[*]*Presenting author*

Keywords. *Statistical Process Monitoring; Functional Data Analysis; Cellwise Outliers*

1 Abstract

Profile monitoring is a statistical quality control technique used to assess the stability across time of a process, i.e., to identify the presence of special sources of variation, when one (univariate) or more (multivariate) quality characteristics are in the form of functional data. Large volumes of profile data are collected using modern production techniques in Industry 4.0 applications. However, these data are usually contaminated by anomalous observations in the form of casewise and cellwise outliers. Outliers must thus be taken into account by profile monitoring methods since they substantially impact the monitoring performance. In order to do this, we provide a novel framework called robust multivariate functional control charts (RoMFCC) that can monitor a multivariate functional quality characteristic while being robust to both functional casewise and cellwise outliers. The RoMFCC framework is made of four elements, i.e., a univariate filter to find functional cellwise outliers that are replaced by missing components, a robust functional data imputation technique, a casewise robust dimensionality reduction, and a monitoring strategy for the multivariate functional quality characteristic. A Monte Carlo simulation study is performed to evaluate the monitoring performance of the RoMFCC compared to competing approaches already proposed in the literature. The RoMFCC is then used in a real-case study to monitor a resistance spot welding process in the automotive industry.

Acknowledgements This work has been done in the framework of the RD project of the multiregional investment programme "REINForce: REsearch to INspire the Future" (CDS000609) with Hitachi Rail STS, supported by the Italian Ministry for Economic Development (MISE) through the Invitalia agency.

MacroPCA for analysing high-dimensional and functional data

M. Hubert[1], P.J. Rousseeuw[1], W. Van den Bossche

[1] *Section of Statistics and Data Science, Department of Mathematics, KU Leuven, Belgium; Mia.Hubert@kuleuven.be, peter@rousseeuw.net .*

Keywords. *Cellwise outliers; Process control*

When a multivariate dataset contains many variables one often reduces its dimension by principal component analysis (PCA). In its basic form PCA is not robust to outliers.

If contamination only occurs in a minority of the objects, 'standard' robust PCA methods that can handle rowwise outliers can be applied. However in high-dimensional or functional data it is very likely that most or even all objects contain some corrupted values (cells), hence these rowwise methods will fail.

MacroPCA is a recent PCA method that is robust against both cellwise and rowwise outliers (Hubert et al. 2019). At the same time, the algorithm can cope with missing values. Several simulations and real datasets illustrate its robustness. New residual maps are introduced, which help to determine which variables are responsible for the outlying behavior. The method is also well-suited for online process control.

We illustrate its performance on high-dimensional and functional data, such as videos where robust PCA can be applied for foreground-background separation.

References

Hubert, M., Rousseeuw, P.J. & Van den Bossche, W. (2019). MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers. *Technometrics*, **61**, 459–473.

C2: Linear models and regression (A5)

Robust and Adaptive Functional Logistic Regression

I. Kalogridis[1]

[1] *Department of Mathematics, KU Leuven; ioannis.kalogridis@kuleuven.be*

Keywords. *Functional logistic regression; Robustness; Regularization; Asymptotics.*

We introduce and study a family of robust estimators for the functional logistic regression model whose robustness automatically adapts to the data thereby leading to estimators with high efficiency in clean data and a high degree of resistance towards atypical data. The estimators are based on the concept of power divergence between densities and may be formed by any combination of lower rank approximations and penalties, as the need arises. For these estimators we prove uniform convergence and high rates of convergence with respect to the commonly used prediction error under only mild assumptions. The highly competitive practical performance of our proposal is illustrated on a simulation study and a real data example involving atypical observations.

Robust Estimation in Exponential Families

Y. Baraud[1*] and J. Chen[1]

[1] *Department of Mathematics (DMATH), University of Luxembourg, Maison du nombre, 6 avenue de la Fonte, L-4364 Esch-sur-Alzette, Grand Duchy of Luxembourg;*
yannick.baraud@uni.lu,
juntong.chen@uni.lu.
[*] *Presenting author*

Keywords. *Generalized linear model; Logit (logistic) regression; Poisson regression; Robust estimation.*

1 Abstract

We observe n of pairs of random variables $X_1 = (W_1, Y_1), \ldots, X_n = (W_n, Y_n)$ that are presumed to be i.i.d. and we consider the problem of estimating the conditional distribution $\mathcal{L}(Y|W)$ of the second coordinate given the first. We model this conditional distribution $\mathcal{L}(Y|W)$ as an element of a given single-parameter exponential family $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ for which the value of the parameter $\theta = \theta(W)$ is an unknown function of the first coordinate of the pair. We provide an estimator of the conditional distribution $\mathcal{L}(Y|W)$ based on our observations and analyse its performance not only when the statistical model is exact, as commonly done in statistics, but also when it is possibly misspecified (the pairs $X_1, \ldots, X_n$ are independent but not exactly i.i.d., the data set contain outliers, the true conditional distribution $\mathcal{L}(Y|W)$ does not belong to the chosen exponential family $\mathcal{P}$, etc). The estimator is based on new estimation strategy, called ρ-estimation, the theory of which has been developed in the series of paper Baraud and Birgé [2018] and Baraud *et al.* [2017]. We establish non-asymptotic risk bounds and show that our estimator is robust to a possible departure from the hypotheses we started from. Finally we provide an algorithm to compute the estimator in low or medium dimensions and compare its performance to that of the celebrated maximum likelihood estimator.

References

Baraud, Y. and Birgé, L. (2018). Rho-estimators revisited: General theory and applications. *Ann. Statist.*, 46(6B):3767–3804.

Baraud, Y., Birgé, L., and Sart, M. (2017). A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517.

Heteroscedastic partially liner model with the skew Laplace normal distribution

F.Z. Doğru[1*] and O. Arslan[2]

[1] *Giresun University, Department of Statistics, Giresun/Turkey;*
fatma.dogru@giresun.edu.tr
[2] *Ankara University, Department of Statistics, Ankara/Turkey;*
olcay.arslan@ankara.edu.tr
[*] *Presenting author*

Keywords. *ECM algorithm; HPLM; ML estimation; SLN.*

1 Introduction

Partially linear models (PLMs) have still been studied with interest by researchers known as semi-parametric models. These models put in an extra non-parametric component to the linear relation between response and explanatory variables, which are considered fundamental tools for modeling economic and biometric data sets (for more detail see Vanegas et al. [2015], Relvas et al. [2016], Ferreira et al. [2017], Ferreira et al. [2022], Doğru & Arslan [2022] and etc.). Generally, it is assumed that the error term in the PLM has a normal distribution. However, the data sets can have skewness and/or heavy-tailedness and heteroscedasticity problems; so modeling the PLM under normality will be ruined by these problems. Therefore, in this study, we consider proposing the heteroscedastic PLM (HPLM) under skew Laplace normal (SLN) distribution (Gómez et al. [2007]) to model skewness and heavy-tailedness simultaneously under the existence of heteroscedasticity. The SLN distribution is a very flexible distribution thanks to its huge range of skewness.

2 HPLM based on the SLN distribution

The PLM based on the SLN distribution can be defined as

$$y_i = \boldsymbol{x}_i{}^T\boldsymbol{\beta} + f(t_i) + \epsilon_i, i = 1, 2, \cdots, n \tag{1}$$

where y_i is the response, $\boldsymbol{x}_i$ is a $p \times 1$ vector of the explanatory variable, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameter, t_i is a scalar which shows a value of continuous

covariate, $f(\cdot)$ is a smoothing function, and ϵ_i shows the error term which has an SLN distribution ($\epsilon_i \sim SLN(0, \sigma^2, \lambda)$) given with the following probability density function (pdf):

$$f(y) = 2f_L(y; \mu, \sigma)\Phi\left(\lambda\frac{y-\mu}{\sigma}\right), \quad -\infty < y < \infty,$$

where $f_L(y; \mu, \sigma)$ represents the pdf of the Laplace distribution which is defined as:

$$f_L(y; \mu, \sigma) = \frac{1}{2\sigma}\exp -\frac{|y-\mu|}{\sigma},$$

and Φ is the cumulative distribution function of the standard normal distribution. We extend the PLM under the SLN distribution in the existence of heteroscedasticity and the error term will form as:

$$\epsilon_i \sim SLN(0, {\sigma_i}^2, \lambda),\ {\sigma_i}^2 = \sigma^2 m_i(\boldsymbol{\rho}, \boldsymbol{z}_i), i = 1, \cdots, n, \tag{2}$$

where $m_i(\boldsymbol{\rho}, \boldsymbol{z}_i)$ is known continuously differentiable positive function, $\boldsymbol{z}_i$ consists of the values of the explanatory variables, and $\boldsymbol{\rho}$ is the unknown parameter vector. Now, we can call the proposed model HPLM-SLN.

In this study, for the model given in (1) with the error term in (2), we will estimate the parameters of the HPLM-SLN using the ML estimation method. Furthermore, the ML estimators will be obtained via the expectation/conditional maximization (ECM) algorithm. We will give some applications for illustrating the applicability of the proposed HPLM-SLN.

References

Doğru, F.Z. & Arslan, O. (2022). Parameter estimation of the partially linear models with skew heavy-tailed error distributions. CFE-CMStatistics 2022, London, UK.

Ferreira, C.S., & Paula, G.A. (2017). Estimation and diagnostic for skew-normal partially linear models. *Journal of Applied Statistics*, **44**(16), 3033–3053.

Ferreira, C.S., Borelli Zeller, C., & de Oliveira Garcia, R.R. (2022). Heteroscedastic partially linear model under skew-normal distribution with application in ragweed pollen concentration. *Journal of Applied Statistics*, 1–28.

Gómez, H.W., Venegas, O. & Bolfarine, H. (2007). Skew-symmetric distributions generated by the distribution function of the normal distribution. *Environmetrics*, **18**, 395–407.

Relvas, C.E.M., & Paula, G.A. (2016). Partially linear models with first-order autoregressive symmetric errors. *Statistical Papers*, **57**, 795–825.

Vanegas, L.H., & Paula, G.A. (2015). A semiparametric approach for joint modeling of median and skewness. *Test*, **240**, 110–135.

Estimation of expected shortfall in linear model

J. Jurečková

The Czech Academy of Sciences, Institute of Information Theory and Automation and Charles University, Faculty of Mathematics and Physics; jureckova@utia.cas.cz

Keywords. *Expected shortfall; Linear model; Regression quantile.*

1 Introduction

Consider the linear regression model

$$\mathbf{Y}_n = \mathbf{X}_n\boldsymbol{\beta} + \mathbf{Z}_n$$

with with observations $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^\top$ and variables $\mathbf{Z}_n = (Z_1, \ldots, Z_n)^\top$, i.i.d. and unobservable, distributed according unknown distribution function F. The designed matrix $\mathbf{X}$ of order $n \times (p+1)$ is known and $x_{i0} = 1$ for $i = 1, \ldots, n$ (i.e., β_0 is an intercept). All inference on $\mathbf{Z}$ and on F is possible only by means of $\mathbf{Y}$, either after estimating the unknown parameter $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$, or by using a procedure invariant to $\boldsymbol{\beta}$. Our problem is to estimate the possible loss of an asset or a portfolio Z in a given period and with a particular confidence level α by means of the expected shortfall equal to $\mathsf{CVaR}_\alpha(\mathsf{Z}) = (1-\alpha)^{-1}\int_\alpha^1 \mathsf{F}^{-1}(\mathsf{t})\mathsf{dt}$.
It has been shown by Bassett et al. (2004) that

$$\mathsf{CVaR}_\alpha(Z) = \frac{1}{1-\alpha}\min_{\xi\in\mathbb{R}} \rho_\alpha(Z-\xi) + \mathbf{E}Z$$

where $\rho_\alpha(z) = z\,(\alpha - I[z<0])\,, \quad z \in \mathbb{R}$ is the quantile criterion function such that the solution of the minimization $\min_{\xi\in\mathbb{R}} \rho_\alpha(X-\xi)$ is the α-quantile of Z. Hence, if independent observations $Z_1, Z_2, \ldots, Z_n$ of Z were available, the estimate of $\mathsf{CVaR}_\alpha(Z)$ could be obtained from the empirical quantile function based on the order statistics $Z_{n:1} \le X_{n:2} \le \ldots \le Z_{n:n}$. It would have the form:

$$\widehat{\mathsf{CVaR}}_\alpha(X) = \frac{1}{\lfloor n(1-\alpha)\rfloor}\sum_{i=\lfloor n(1-\alpha)\rfloor}^{n} \mathsf{Z}_{n:i}.$$

However, because only the observations of Y are available, we should look for an alternative solution of explicit estimating of $\mathsf{CVaR}_\alpha(\mathsf{Z})$. A possible estimate can be

based on the regression quantile of the model or on its functional, e.g. on its intercept component, on the average regression quantile or on the two-step regression quantile with an R-estimate of its slope componens.

References

[1] Bassett, G.W., Jr., Koenker, R., Kordas, W. (2004). Pessimistic portfolio allocation and Choquet expected utility. *Journal Financial Economics* **2/4**, 477–492.

[2] Jurečková, J. and Picek, J. (2005). Two-step regression quantiles. *Sankhya* **67/2**, 227–252.

[3] Jurečková, J., Kalina, J., Večeř, J. (2022). Estimation of expected shortfall under various experimental conditions. arXiv: 22.12419v1 [stat.ME].

Outlier Robust Inference in (Weak) Linear Instrumental Variable Models

J. Klooster[1*] and M. Zhelonkin[1]

[1] *Erasmus University Rotterdam; klooster@ese.eur.nl, zhelonkin@ese.eur.nl*
[*] *Presenting author*

Keywords. *Infuence function; Robust inference; Outlier; Robust test; Weak instrument.*

Abstract

We propose a general robust framework to construct weak instrument robust testing procedures that are also robust to outliers in the linear instrumental variable model. The framework is constructed upon M-estimators and we show that the classical weak instrument robust tests, such as the Anderson & Rubin [1949] test and the Moreira [2003] conditional likelihood ratio (CLR) test can be obtained by specifying the M-estimators to be the Least Squares estimators. As it turns out that the classical testing procedures are not robust to outliers, we show how to construct robust alternatives. In particular, we show how to construct a robust CLR statistic based on Mallows type M-estimators and show that its asymptotic distribution is the same as the (classical) CLR statistic. The theoretical results are corroborated by a simulation study. Finally, we revisit three empirical studies affected by outliers and apply the robust CLR test to re-evaluate their results.

References

Anderson, T.W. & Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Statistics*, **20**, 46–63

Moreira, M.J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, **71**, 1027-1048.

Extremal index robust estimators based on Negative Binomial regression

M. Souto de Miranda[1*], M.C. Miranda[1] and M.I. Gomes[2]

[1] *CIDMA, University of Aveiro, Portugal; manuela.souto@ua.pt, cristina.miranda@ua.pt*
[2] *CEAUL, University of Lisbon; migomes@ul.pt*
[*] *Presenting author*

Keywords. *Estremal index; Robust estimation; Negative Binomial regression.*

1 Abstract

Statistical knowledge about the occurrence of extreme phenomena is increasingly necessary in different fields, namely, in economics, financial investments, epidemiology, weather conditions, floods or droughts. Often those phenomenons happen in successive observations, constituting clusters of dependent observations that exceed some fixed thresholds. It is known that under stationary and dependence conditions the limit distribution of the maxima can depend on a parameter named the extremal index. When it exists, the extremal index is related with the limit mean clusters dimension, *i.e.*, the mean number of successive exceedances (Ferro & Segers [2003]). We propose to estimate the extremal index with a robust estimator which was developed for counting processes with too many zeros (see Aeberhardt et al. [2014]), as it happens when dealing with exceedances from a fixed high threshold. The estimator is defined in the framework of a Negative Binomial regression. The inverse of the link function corresponds to the estimate of the mean clusters size dimension and it is taken as the reciprocal of the extremal index estimate. We carried out a simulation study considering different types of dependence structures. The performance of the proposed estimator is compared with traditional estimation procedures both in the assumed models and in the contaminated models generated by mixture models.

References

Aeberhard, W., Cantoni, E. & Heritier, S. (2014). Robust Inference in the Negative Binomial Regression Model with an Application to Falls Data. *Biometrics*, **70**,

920–931.

Ferro, C.A.T. & Segers, J. (2003). Inference for Clusters of Extreme Values. *J. R. Statist. Soc. B*, **65**, 545–556.

Posters (Cafeteria)

A consistent robust regression algorithm with the usage of prior information

Zheyi Fan[1,2*], Qingpei Hu[1,2] and Ng Szu Hui[3]

[1]*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China*
[2]*School of Mathematical Sciences, University of Chinese Academy of Sciences, China*
[3]*Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore .*
[1,2]*{fanzheyi,qingpeihu}@amss.ac.cn,* [3]*isensh@nus.edu.sg*
[*]*Presenting author*

Keywords. *Robust regression; Hard thresholding; Consistent analysis*

1 Abstract

By incorporating prior information into robust regression, we propose a very robust regression algorithm, which accurately recover the true parameters under adaptive adversarial attack (AAA). The common existence of noise and data corruption has led to the rapid development of robust regression. However, the AAAs aiming at data characters are more destructive and difficult to be detected, especially when a certain proportion of the response variables have been corrupted. Most robust regression algorithms fail to achieve good results under this attack.

This work is an extension of the previous work Fan et al. [2022], which can resist AAAs effectively but will have an unavoidable estimation error. Current work further improves the breakdown point of the algorithm when facing AAAs by incorporating prior information. Meanwhile, we innovatively propose an iterative algorithm, which eliminate the estimation error through iteration steps. This modification greatly improves the accuracy of the algorithm, resulting in the excellent performance in both oblivious adversarial attack and adaptive adversarial attack. Extensive experiments show that, under different dataset attacks, our algorithms achieve state-of-the-art results compared with other benchmark algorithms, demonstrating the robustness of the proposed approach.

References

Zheyi Fan, Zhaohui Li & Qingpei Hu (2022). Robust Bayesian Regression via Hard Thresholding. *Advances in Neural Information Processing Systems*.

Anomaly Detection in Financial Time Series by Principal Component Analysis and Neural Networks

S. Crépey[1], N. Lehdili[2], **N. Madhar**[1,2*] and M.Thomas[3]

[1] *Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Université Paris Cité, 75013 Paris, France; stephane.crepey@lpsm.paris, madhar@lpsm.paris.*
[2] *Natixis Entreprise Risk Management department, 75013 Paris, France; noureddine.lehdili@natixis.com.*
[3] *LPSM, Sorbonne Université, 75005 Paris, France; maud.thomas@sorbonne-universite.fr.*
[*] *Presenting author*

Abstract. A major concern when dealing with financial time series involving a wide variety of market risk factors is the presence of anomalies. These induce a miscalibration of the models used to quantify and manage risk, resulting in potential erroneous risk measures. We propose an approach that aims at improving anomaly detection in financial time series, overcoming most of the inherent difficulties. Valuable features are extracted from the time series by compressing and reconstructing the data through principal component analysis. We then define an anomaly score using a feedforward neural network. A time series is considered to be contaminated when its anomaly score exceeds a given cutoff value. This cutoff value is not a hand-set parameter but rather is calibrated as a neural network parameter throughout the minimization of a customized loss function. The efficiency of the proposed approach compared to several well-known anomaly detection algorithms is numerically demonstrated on both synthetic and real data sets, with high and stable performance being achieved with the PCA NN approach. We show that value-at-risk estimation errors are reduced when the proposed anomaly detection model is used with a basic imputation approach to correct the anomaly.

Keywords. *Anomaly detection; Financial time series; Principal component analysis; Neural network; Value at risk.*

Benchmark Of Clustering Methods Applied On A Rotating Machine Vibration Modes Identification In A Nuclear Industrial Environment

S. Arfaoui[1] and L. Makong[1*]

[1] *Orano Recyclage La Hague; senda.arfoui@hotmail.com, ludovic.makong-hell-nkatack@orano.group.*
[*] *Presenting author*

Keywords. *Functional data analysis (FDA); Clustering; Nuclear industry.*

1 Introduction

In this work, various clustering approaches were applied on irregular vibration time series sensor measurements of a rotating machine in a highly radioactive industrial environment. Our aim is to identify uncontrolled anomalous high vibration modes usually leading to production delays in order to prevent machine failure. This study represents the first step towards the development of predictive maintenance models.

2 Methodology and Results

The study raw dataset consisted of 740 chronologically ordered vibration sequences sampled at 30 seconds where a sequence represents a machine operating cycle with a 24 hours average total duration.

Given the right dataset representation, clustering methods were applied on time series, scalar and functional data respectively. The main goal was to detect all the machine vibration modes especially those anomalous high vibration modes. Those methods were evaluated and compared as shown in Table 1.

In the end, functional data clustering led to better insights relevant to the machine vibration modes given their high silhouette scores. The resulted clusters were also validated by plant operators and process engineers as nominal, alert and alarm real operating ranges of the machine (see Fig.1).

Time series clustering		
Algorithm	**Optimal number of clusters**	**Silhouette score**
Kmeans + DTW (Tavenard et al. [2006])	4	0.4
Clustering with scalar data		
Kmeans + Euclidian	4	0.32
Dbscan	6	0.4
Functional data clustering		
FunHDDC (Jacques & Preda [2014])	3	0.45
FunFEM (Bouveyron & Jacques [2015])	3	0.47
Kmeans + Lp-metric (Ramsay et al. [2014])	3	0.52

Table 1: Evaluation of studied clustering approaches.

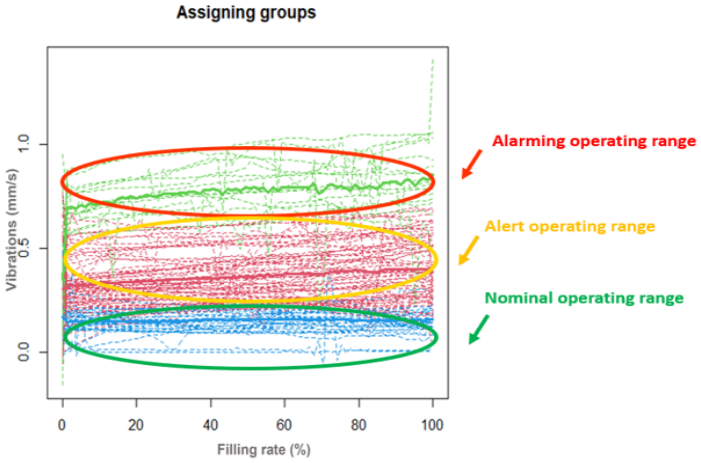


Figure 1: Identification of the machine vibration modes with Kmeans + Lp-metric.

References

Tavenard, R., et al (2020). Tslearn, a machine learning toolkit for time series data. *The Journal of Machine Learning Research*, **21.1**, 4686-4691.

Jacques, J. & Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, **8**, 231–255.

Bouveyron, C. & Jacques, J. (2015). funFEM: an R package for functional data clustering.

Ramsay, J., et al. (2014). fda: Functional data analysis. URL http://CRAN.R-project.org/package=fda.

Co-clustering contaminated data: a robust model-based approach

E. Fibbi[1,2,*] *et al.*

[1] *KU Leuven, Department of Mathematics; edoardo.fibbi@kuleuven.be*
[2] *European Commission, Joint Research Centre*
[*] *Presenting author*

Keywords. *Co-clustering; Robustness; Trimming; Latent Block Model; CEM.*

Abstract

The exploration and analysis of large high-dimensional data sets calls for well-thought techniques to extract the salient information from the data, such as co-clustering [Govaert & Nadif, 2014]. Latent block models cast co-clustering in a probabilistic framework that extends finite mixture models to the two-way setting. In addition to being high-dimensional, real-world data sets often contain anomalies which could be of interest *per se* and may make the results provided by standard, non-robust procedures unreliable. Also the estimation of latent block models can be heavily affected by contaminated data. Therefore, we propose a method to compute robust estimates for latent block models. The proposed algorithm combines impartial trimming [Cuesta-Albertos *et al.*, 1997] with a block Classification Expectation-Maximisation (CEM) algorithm [Celeux & Govaert, 1991], which aims to maximise the complete-data likelihood of the model. Experiments on both simulated and real data show that our method is able to resist high levels of contamination and can provide additional insight into the data by highlighting possible anomalies.

References

Govaert, G. & Nadif, M. (2014). Co-Clustering: models, algorithms and applications. ISTE Ltd, London, England.

Celeux, G. & Govaert, G. (1991). A classification EM algorithm for clustering and two stochastic versions. Research Report RR-1364, INRIA.

Cuesta-Albertos, J. A. *et al.* (1997). Trimmed k-means: an attempt to robustify quantizers. *The Annals of Statistics*, 25(2):553 – 576.

Comparison of functional outlier detection methods

Caroline Le Gall[1*], Aurore Archimbaud[3], Théo Condette[1], Clément Jonca[2], Anne Ruiz-Gazen[2], Mengqi Wang[2] and Chérif Yaker[2]

[1] *Airbus, France; caroline.le-gall@airbus.com*
[2] *Toulouse School of Economics, France;*
[3] *Erasmus School of Economics, The Netherlands;*
[*] *Presenting author*

Keywords. *Anomaly detection; ICS; Functional isolation forest; MUOD*

Anomaly detection is of particular interest in many areas like industrial applications. This is a challenging unsupervised task as it involves identifying observations that are only outlying on a subspace of the original variables. Nowadays, with the rise of sensor data, the task is even more complex as the measurement values become some curves. This functional framework leads to new challenges for identifying outliers, i.e curves that behave differently from the main bulk of the data either in magnitude, in amplitude or in shape, in the entire curve or only in some parts. We compare three main approaches suitable to this context. First, we investigate the Massive Unsupervised Outlier Detection (MUOD) method introduced by Azcorra et al. [2018] which computes some simple interesting metrics for functional data. Then, we focus on the functional version of the Invariant Coordinate Selection (ICS) method suggested by Archimbaud et al. [2022], which has been proven useful for outlier detection. Finally, we consider the functional isolation forest (FIF) method introduced by Staerman et al. [2019], which is an extension of the isolation forest algorithm. We illustrate which methods are the most suitable to identify some kinds of outliers on some simulated examples.

References

Archimbaud, A., Boulfani, F., Gendre, X., Nordhausen, K., Ruiz-Gazen, A., and Virta, J. (2022). ICS for multivariate functional anomaly detection with applications to predictive maintenance and quality control. *Econometrics and Statistics*. (In press.)

Azcorra, A., Chiroque, L. F., Cuevas, R., Fernández Anta, A., Laniado, H., Lillo, R. E., Romo J, Sguera, C. (2018). Unsupervised scalable statistical method for identifying influential users in online social networks. *Scientific reports*, **8**(1):6955.

Staerman, G., Mozharovskyi, P., Clémençon, S., and d'Alché-Buc, F. (2019, October). Functional isolation forest. In *Asian Conference on Machine Learning* (pp. 332-347). PMLR.

Minimum Distance Estimators in Poisson hurdle model

C. Amado[1*] and M. Souto de Miranda[2]

[1] *CEMAT, IST-Universidade de Lisboa, Portugal; conceicao.amado@tecnico.ulisboa.pt*
[2] *CIDMA, Universidade de Aveiro, Portugal; manuela.souto@ua.pt*
[*] *Presenting author*

Keywords. *Hurdle model; Minimum distance estimators; Robustness, Zero truncated Poisson.*

1 Hurdle Poisson Model

Several situations are described by counting processes that include a great number of zeroes. For instance, the number of times each person visits the doctor per month; the number of days that temperature in a specific location exceeds a determined degree; or, in general, the number of exceedances of a threshold along a fixed period. Typically, that type of occurrences is modelled with finite mixture models, namely, with a zero-inflated model or with a hurdle model (also called two-steps or conditional model). The knowledge of the real data generating process and the goals should point out the choice.

2 Estimation

Under precise and strict stochastic assumptions, maximum likelihood is employed to fit the Poisson hurdle model; however, when these assumptions are not validated, the performance of these estimators decays.

This work discusses minimum distance estimation method in the Poisson hurdle model. A numerical study compares the minimum distance estimator's performance with other well-known estimators.

Robust Regression with Discrete Covariates

Otso Hao[1*] and Bent Nielsen[2]

[1,2] *University of Oxford; otso.hao@economics.ox.ac.uk, bent.nielsen@nuffield.ox.ac.uk*

Keywords. *Robust regression; Discrete regressors; Least trimmed squares; Leverage.*

A common empirical methodology is to model an outcome variable as a linear function of discrete and continuous covariates as

$$y_i = \beta_0 + z_i'\beta_z + w_i'\beta_w + \varepsilon_i,$$

where y_i is a scalar outcome, z_i a vector of discrete covariates, and w_i a vector of continuously distributed covariates. In social sciences, a prominent example of a discrete covariate z_i is a set of controls for an individual's background characteristics, such as gender, age, and years of education.

Researchers need outlier robust methods for estimation and inference in such models. If non-robust methods, such as least squares, are used, a small share of outlying observations can have a large influence on the resulting estimates. Least trimmed squares (LTS) [Rousseeuw, 1984] is a popular robust estimator that is reported to have desirable properties, such as a high breakdown point and a resistance to bad leverage points.

Unfortunately, previous literature has not fully explored the use of LTS in models that include discrete covariates. In particular, we are not aware of previous studies that would explicitly explain how to compute the LTS estimator and use it to conduct statistical inference in such models.

We show that LTS can be used for statistical inference in linear regression models that have both discrete and continuous covariates. We propose an updated definition of LTS that is particularly suited for models that include discrete covariates. The computation of this LTS estimator can be done with a simple modification of the popular fast-LTS algorithm [Rousseeuw and van Driessen, 2006]. Leveraging a recently proposed LTS framework [Berenguer-Rico et al., forthcoming], we establish a set of conditions needed for asymptotic LTS inference in the presence of discrete covariates.

Our updated LTS definition is a simple modification of standard LTS that explic-

itly restricts the estimator to search over subsets with non-singular design matrices. This sidesteps a known issue with non-invertibility that often arises when the model includes discretely supported covariates. This updated estimator preserves the core features of standard LTS, which means we can use the LTS framework of Berenguer-Rico et al. [forthcoming] to derive its asymptotic distribution, and a simple adaptation of the fast-LTS algorithm to compute it in practice.

Our main theoretical contribution is to offer applied researchers a clear set of assumptions they can use to validate LTS asymptotic inference in models that include discrete covariates. The key condition we require is a restriction on the share of observations that have covariates lying on a lower dimensional hyperplane. This condition mirrors existing literature, where a related hyperplane condition has been shown to determine the finite sample breakdown point of LTS and other robust estimators [Mili and Coakley, 1996].

References

Berenguer-Rico, Vanessa & Johansen, Søren & Nielsen, Bent (forthcoming). A Model Where the Least Trimmed Squares Estimator is Maximum Likelihood. *Journal of the Royal Statistical Society, Series B*.

Mili, Lamine & Coakley, Clint W. (1996). Robust Estimation in Structured Linear Regression. *The Annals of Statistics*, **24(6)**, 2593–2607.

Rousseeuw, Peter J. & van Driessen, Katrien (2006). Computing LTS Regression for Large Data Sets. *Data Mining and Knowledge Discovery*, **12**, 29–45.

Rousseeuw, Peter J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association*, **79(388)**, 871–880.

Robustness of scatter depth

G. Louvet[1*] and G. Van Bever[1]

[1] *Department of Mathematics and Namur Institute for Complex Systems (Naxys), Université de Namur, Belgium; gaetan.louvet@unamur.be, germain.vanbever@unamur.be.*
[*] *Presenting author*

Keywords. *Robustness; Influence function; Scatter depth.*

1 Abstract

Statistical depth provides robust nonparametric tools to analyze distributions. Depth functions indeed measure the adequacy of distributional parameters to underlying probability measures. In the location case, the celebrated (Tukey) halfspace depth has been widely studied and its robustness properties amply discussed. Recently, depth notions for scatter parameters have been defined and studied. The robustness properties of this latter depth function remain, however, largely unknown.
In this paper, we discuss the robustness of several scatter depth functions. For example, we consider the scatter halfspace depth, whose expression is given by

$$HD_{\text{sc}}(\Sigma, \text{P}) = \inf_{||u||=1} \min \left(\text{P}\left[|u'(X - T_{\text{P}})| \leq \sqrt{u'\Sigma u} \right], \text{P}\left[|u'(X - T_{\text{P}}) \geq \sqrt{u'\Sigma u} \right] \right),$$

with T_{P} a location estimator. In the known location case, the influence function is bounded and takes the form

$$\text{IF}(z, HD_{\text{sc}}(\Sigma, \text{P})) = I[z \in A] - HD_{\text{sc}}(\Sigma, \text{P}),$$

for $A = A(\text{P})$, an appropriate set that depends on the distribution. In the general case of absolutely continuous distribution with unknwon location T_{P}, we provide mild conditions on P under which we bound the influence function as

$$-HD_{\text{sc}}(\Sigma, \text{P}) + g(\text{IF}(z, T_{\text{P}})) \leq \text{IF}(z, HD_{\text{sc}}(\Sigma, \text{P})) \leq 1 - HD_{\text{sc}}(\Sigma, \text{P}) + g(\text{IF}(z, T_{\text{P}})).$$

Subsequently, we also discuss the behaviour of the scatter depth when the distribution is discrete. The influence function of the scatter halfspace depth median, i.e. the matrix maximizing the scatter halfspace depth is also provided for elliptical distributions. We conclude by conducting some simulations to compare the efficiency of the different scatter depth functions.

Wednesday 24 May 2023

I3: Dimension reduction (A3)

Tandem clustering with invariant coordinate selection

A. Alfons[1*], A. Archimbaud[1], **K. Nordhausen**[2] and A. Ruiz-Gazen[3]

[1] *Erasmus School of Economics, Erasmus University Rotterdam, Netherlands, alfons@ese.eur.nl, archimbaud@ese.eur.nl*
[2] *Department of Mathematics and Statistics, University of Jyväskylä, Finland, klaus.k.nordhausen@jyu.fi*
[3] *Toulouse School of Economics, Université de Toulouse Capitole, France, anne.ruiz-gazen@tse-fr.eu*

Keywords. *Principal Component Analysis; Linear Discriminant Analysis; Scatter Matrices; Minimum Covariance Determinant.*

Tandem clustering is a well-known technique for dealing with high-dimensional or noisy data to better identify clusters. This is a sequential approach based on first reducing the dimension of the data and then performing the clustering. The most common method, based on principal component analysis (PCA), has been criticized for only focusing on maximizing inertia and not necessarily preserving the structure of interest for clustering. Therefore, we suggest a new tandem clustering approach based on invariant coordinate selection (ICS). This multivariate method is designed to identify the structure of the data by jointly diagonalizing two scatter matrices, while maintaining the affine invariance of the new coordinates. More specifically, some theoretical results proved that under some elliptical mixture models, the first and/or last components are carrying the information regarding the clustering structure. However, despite the attractive properties of ICS, the method has not been studied much in the context of clustering but mostly for outlier detection purposes. The issues of choosing the pair of scatter matrices and the components to keep are the two challenges that must be addressed. For clustering purposes, we suggest that the best scatter pairs consist of one matrix which captures the within-cluster structure and another which captures the global structure. To this end, the local shape or pairwise scatters prove to be good choices for estimating the within-structure. In addition, we also investigate the use of the well-known minimum covariance determinant (MCD) estimator based on a smaller-than-usual subset size. The performance of ICS as a dimension reduction method is evaluated to determine its ability to preserve the cluster structure of the data. We conducted a large simulation study and applied it to benchmark data sets. We tested various combinations of scatter matrices, component selection criteria, and the effects of the presence of outliers. Results indicate that the ICS-based tandem clustering method has superior performance over PCA, and thus is a promising approach.

Robust and sparse CCA: An algorithm for dimension reduction via sparsity inducing penalties

P. Pfeiffer[1*], A. Alfons[2] and P. Filzmoser[1]

[1] *Institute of Statistics and Mathematical Methods in Economics, TU Wien, Wiedner Hauptstraße 8–10, 1040 Vienna, Austria; pia.pfeiffer@tuwien.ac.at, peter.filzmoser@tuwien.ac.at.*
[2] *Econometric Institute, Erasmus School of Economics, Erasmus Universiteit Rotterdam, PO Box 1738, 3000 DR Rotterdam, The Netherlands; alfons@ese.eur.nl*
[*] *Presenting author*

Keywords. *Robust CCA; Sparse CCA; Constrained Optimization.*

CCA (Canonical Correlation Analysis) is widely applied to measure the association between multivariate data sets, but the classical method is neither robust in the presence of atypical observations, nor does it lead to sparse canonical vectors, and thus it is not suitable for high-dimensional data with more variables than observations. While there are several approaches to achieve robustness or sparsity, only the alternating regression method proposed by Wilms & Croux [2015] combines both objectives. Higher-order canonical correlations, however, cannot be derived directly using this algorithm.

We propose to reformulate the CCA objectives as an optimization problem with constraints, which allows for a direct statement of regularization conditions and a flexible choice of a covariance estimator. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ denote a p- and q-dimensional random variable, respectively, and $\boldsymbol{\Sigma}_{xx}, \boldsymbol{\Sigma}_{yy}$ and $\boldsymbol{\Sigma}_{xy}$ the corresponding covariance matrices. The first canonical correlation coefficient ρ_1 and the first pair of canonical vectors $(\boldsymbol{a}_1, \boldsymbol{b}_1)$ are given as a solution of the optimization problem

$$\max_{\boldsymbol{a}\in\mathbb{R}^p, \boldsymbol{b}\in\mathbb{R}^q} \boldsymbol{a}'\boldsymbol{\Sigma}_{xy}\boldsymbol{b} \tag{1}$$

under the constraints

$$\boldsymbol{a}'\boldsymbol{\Sigma}_{xx}\boldsymbol{a} = 1 \quad \text{and} \quad \boldsymbol{b}'\boldsymbol{\Sigma}_{yy}\boldsymbol{b} = 1. \tag{2}$$

The k-th canonical correlation coefficient ρ_k and the respective pair of canonical vectors $(\boldsymbol{a}_k, \boldsymbol{b}_k)$ maximize (1) under the condition that they are uncorrelated with the previous $1, \ldots, k-1$ directions, denoted as the constraints

$$\boldsymbol{a}'\boldsymbol{\Sigma}_{xx}\boldsymbol{a}_i = 0 \quad \text{and} \quad \boldsymbol{b}'\boldsymbol{\Sigma}_{yy}\boldsymbol{b}_i = 0, \text{ for } i = 1, \ldots, k-1. \tag{3}$$

For a sparse setting, we add penalty terms as further constraints,

$$P_{\alpha_1}(\boldsymbol{a}) \leq c_1 \quad \text{and} \quad P_{\alpha_2}(\boldsymbol{b}) \leq c_2 \tag{4}$$

where c_1 and c_2 denote positive constants, and the penalty terms (4) are given as elastic net [Zou & Hastie, 2005] penalties with mixing parameters α_1, α_2. The augmented Lagrangian with $\boldsymbol{\lambda}$ denoting the Lagrange multiplier, and H summarizing the constraints is then given as

$$\mathcal{L}_\rho(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\lambda}) = -|\boldsymbol{a}'\boldsymbol{\Sigma}_{xy}\boldsymbol{b}| + \boldsymbol{\lambda}' \cdot H(\boldsymbol{a}, \boldsymbol{b}) + \frac{\rho}{2}\|H(\boldsymbol{a}, \boldsymbol{b})\|_2^2. \tag{5}$$

Then, a solution to (1)-(4) can be found by minimizing (5). For the optimization algorithm, the method of multipliers (see e.g. Boyd et al. [2011]) is combined with an adaptive gradient descent algorithm as described by Reddi et al. [2018] for the minimization in step 1. Given starting values $\boldsymbol{a}_k^0, \boldsymbol{b}_k^0$ and $\boldsymbol{\lambda}_k^0$, the update is conducted in an alternating fashion:

1. $(\boldsymbol{a}_k^{t+1}, \boldsymbol{b}_k^{t+1}) \leftarrow \text{argmin}_{\boldsymbol{a},\boldsymbol{b}} \mathcal{L}_\rho(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\lambda}_k^t)$
2. $\boldsymbol{\lambda}_k^{t+1} \leftarrow \boldsymbol{\lambda}_k^t + \rho H(\boldsymbol{a}_k^{t+1}, \boldsymbol{b}_k^{t+1})$

The advantage of this approach is its flexibility (in the choice of the covariance estimator, level of sparsity) and its direct way of computation for higher-order canonical correlations. By choosing appropriate covariance estimators for $\boldsymbol{\Sigma}_{xx}, \boldsymbol{\Sigma}_{yy}$ and $\boldsymbol{\Sigma}_{xy}$, the robustness of the resulting canonical correlations can be controlled. The penalty terms (4) induce sparsity in the resulting canonical directions. Conditions (3) ensure the uncorrelatedness of higher-order directions to lower-order canonical vectors. For the higher-order directions, again, a suitable level of sparsity can be chosen.

In a simulation study, we show the robustness and suitability of our approach for high-dimensional data in different simulation scenarios. Empirical applications from tribology underline the usefulness of this approach. Furthermore, the algorithm can be adapted to other robust multivariate methods in connection with high-dimensional data.

References

Wilms, I., Croux, C., 2015. Robust sparse canonical correlation analysis. *BMC Systems Biology* **10**, 72.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67(2)**, 301-320.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning* **3**, 1-122.

Reddi, S.J., Kale, S., Kumar, S., 2018. On the Convergence of Adam and Beyond. 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018.

The Influence Function of Graphical Lasso Estimators

G. Louvet[1], J. Raymaekers[2*], G. Van Bever[1] and I. Wilms[2]

[1] *University de Namur, Belgium; gaetan.louvet@unamur.be, germain.vanbever@unamur.be.*
[2] *Maastricht University, The Netherlands; j.raymaekers@maastrichtuniversity.nl, i.wilms@maastrichtuniversity.nl*
[*] *Presenting author*

Keywords. *graphical models; regularization; robust methods*

Graphical models are nowadays often estimated using regularization that is aimed at reducing the number of edges in a network. By relying on edge-sparsity as a simplifying structure, the conditional dependency network among (potentially a large number of) variables can then be presented in a compact manner. The Graphical Lasso (Glasso) is a common choice to obtain such sparse graphical models. Glasso lacks, however, robustness to outliers. To overcome this problem, one typically applies a robust plug-in procedure where the Glasso is computed from, for instance, an initial pairwise robust covariance/correlation estimate instead of the classical sample covariance estimate, thereby providing protection against outliers. We derive and compare the influence function of the classical Glasso to various robustified versions, as well as their corresponding asymptotic variances. Simulation results provide further insights into their finite sample performance.

References

Louvet, G., Raymaekers, J., Van Bever, G. & Wilms, I. (2022). The Influence Function of Graphical Lasso Estimators. *arXiv:2209.07374*.

C3: Outlier detection (A5)

Simultaneous feature selection and outlier detection with optimality guarantees

L. Insolia[1*], A. Kenney[2], F. Chiaromonte[3] and G. Felici[4]

[1] *University of Geneva; luca.insolia@unige.ch*
[2] *University of California Berkeley; ajk5910@psu.edu*
[3] *Pennsylvania State University & Sant'Anna School of Advanced Studies; fxc11@psu.edu*
[4] *National Research Council of Italy; giovanni.felici@iasi.cnr.it*
[*] *Presenting author*

Keywords. *Breakdown point; Mixed-integer programming; Robust regression; Sparse estimation; Strong oracle property.*

High-dimensional linear regression models are nowadays pervasive in most research domains. However, as studies become larger, the likelihood of having redundant features or contaminated data (outlying values) increases, which can create serious challenges. To address these issues, researchers have focused on developing efficient methods for sparse estimation in the presence of outliers. We contribute to this area considering high-dimensional models contaminated by multiple mean-shift outliers affecting both the response and the design matrix – leading to the exclusion of outlying cases from the fit. Our novel framework leverages mixed-integer programming techniques to simultaneously perform feature selection and outlier detection with provably optimal guarantees (i.e., the global optimum of the underlying "double" combinatorial problem is indeed achievable) [1]. We also prove theoretical properties for our approach, such as a necessary and sufficient condition for the robustly strong oracle property, where the number of features can increase exponentially with the sample size, and the breakdown point of the resulting estimates. Moreover, we provide computationally efficient procedures to tune integer constraints and warm-start the algorithm. Our extensive simulations and real-world applications demonstrate the superiority of our proposal over existing heuristic methods. Additionally, we discuss its extensions to a broader class of models, as well as the use of down-weighting schemes and adaptive procedures for outlier detection.

References

[1] Insolia, L., Kenney, A., Chiaromonte, F. & Felici, G. 2022. Simultaneous feature selection and outlier detection with optimality guarantees. Biometrics, 78, 1592–1603.

Outlier detection and explanation for matrix-valued data

M. Mayrhofer[1], U. Radojicic[1], H. Lewitschnig[2], and P. Filzmoser[1]

[1] *TU Wien; marcus.mayrhofer@tuwien.ac.at, una.radojicic@tuwien.ac.at, peter.filzmoser@tuwien.ac.at*
[2] *Infineon Technologies Austria AG; Horst.Lewitschnig@infineon.com*

Keywords. *Functional data; Matrix-valued data; Outlier explanation; Robust covariance estimation*

Matrix-valued data are a common data type with images, multivariate times series, and multivariate functional data being some common examples. Outlier detection techniques are not as commonly available for matrix-valued data as for multivariate data. In practice, such data are often treated in a vectorized form, i.e., by arranging the pixel information of images column-wise on top of each other, resulting in a possible loss of important information.

Let $\boldsymbol{X} \in \mathbb{R}^{p\times t}$ follow a matrix normal distribution, $\boldsymbol{X} \sim \mathcal{MN}(\boldsymbol{M}, \boldsymbol{\Sigma}_r, \boldsymbol{\Sigma}_c)$, with mean $\boldsymbol{M} \in \mathbb{R}^{p\times t}$ and covariance matrices $\boldsymbol{\Sigma}_r \in \mathbb{R}^{p\times p}$ for the rows and $\boldsymbol{\Sigma}_c \in \mathbb{R}^{t\times t}$ for the columns. In the vectorized version $\mathrm{vec}(\boldsymbol{X})$ we would have $\mathrm{vec}(\boldsymbol{X}) \sim \mathcal{N}(\mathrm{vec}(\boldsymbol{M}), \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r)$, where $\otimes$ is the Kronecker product [Gupta and Nagar, 1999]. The parameters $\boldsymbol{M}$, $\boldsymbol{\Sigma}_r$, and $\boldsymbol{\Sigma}_c$ can be estimated by an iterative scheme using the ML method introduced by Dutilleul [1999], where the number of observations needs to be at least $\max(p/t, t/p) + 1$. We propose a robust procedure for estimating the parameters, which follows the idea of the minimum covariance determinant (MCD) estimator [Rousseeuw, 1985], and refer to it as *MMCD* estimator. The MMCD estimator can be computed if the h-subset consists of at least $\max(p/t, t/p) + 1$ samples.

Based on the connection between multivariate and matrix normal distribution, the *matrix Mahalanobis distance (MMD)* can be defined as

$$\begin{aligned}\mathrm{MMD}^2(\boldsymbol{X}) &:= \mathrm{tr}(\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{X}-\boldsymbol{M})'\boldsymbol{\Sigma}_r^{-1}(\boldsymbol{X}-\boldsymbol{M})) \\ &= \mathrm{vec}(\boldsymbol{X}-\boldsymbol{M})'(\boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r)^{-1}\mathrm{vec}(\boldsymbol{X}-\boldsymbol{M}) = \mathrm{MD}^2(\mathrm{vec}(\boldsymbol{X})),\end{aligned}$$

see Glanz and Carvalho [2018] for example. Plugging in the robustly estimated parameters results in a robust tool for identifying outliers for matrix-valued observations.

Taking image data as an example, the number of available samples is often a limiting factor in data analysis. Hence, the ability to deal with data sets with few samples can be of particular importance in such cases. Another possible application is multivariate functional data analysis, where the lack of distribution functions poses a problem for outlier detection. In this setting, we connect the MMCD procedure to the coefficient matrix of smoothed multivariate functional data to extend the matrix-variate case to outlier detection in a functional setting. Moreover, we show that outlier explanation based on Shapley values as introduced in Mayrhofer and Filzmoser [2022] can also be applied for matrix-valued data. We illustrate the performance of the MMCD estimator in those settings on simulated data as well as real-world examples.

Acknowledgements: This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 101007326. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Germany, Austria, Belgium, Czech Republic, Italy, Netherlands, Lithuania, Latvia, Norway.

References

Gupta, A. and Nagar, D. (1999). Matrix Variate Distributions. Monographs and Surveys in Pure and Applied Mathematics. Taylor & Francis.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, Vol. B, pages 283–297.

Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64(2):105–123.

Mayrhofer, M. and Filzmoser, P. (2022). Multivariate outlier explanations using Shapley values and Mahalanobis distances. arXiv preprint arXiv:2210.10063.

Glanz, H. and Carvalho, L. (2018). An expectation–maximization algorithm for the matrix normal distribution with an application in remote sensing. *Journal of Multivariate Analysis*, 167:31–48.

Least Trimmed Squares Regression: Consistent estimation of the number of outliers

V. Berenguer-Rico[1] and **B. Nielsen**[1*]

[1] *University of Oxford;*
vanessa.berenguer-rico@economics.ox.ac.uk, bent.nielsen@nuffield.ox.ac.uk.
[*] *Presenting author*

Keywords. *Asymptotic theory; Least Trimmed Squares; Maximum likelihood; Model selection.*

Classicial regression procedures based on least squares or least absolute deviation procedures are highly sensitive to atypical observations. At worst, regression slopes are severely distorted in the presence of leverage points. Robust regression estimators such as the least trimmed squares estimator (LTS) [Rousseeuw, 1984] are resistant to outliers and leverage points. Recently, a model has been proposed in which the LTS is maximum likehood [Berenguer-Rico et al., 2022] and in which standard least squares inference apply [Berenguer-Rico et al., 2023]. Those results require that the number of outliers is known. Determining the number of outliers is therefore a question of model selection among a large number of models. We show that the proportion of outliers can be estimated consistently by minimizing an information criteria.

The LTS estimator is defined as follows. Suppose there are h 'good' observations in a sample of n observation. The set of 'good' observations is found by minimizing over the least squares residual sum of squares over all possible h-subsets of observations. The LTS estimator is the least squares estimator on the estimated set of 'good' observations [Rousseeuw, 1984].

The traditional probability framework for analyzing the LTS estimator and other robust regression estimators is an i.i.d. model, where the regression errors satisfy an ϵ-contaminated model mixing a normal distribution with a contamination distribution. The asymptotic properties of the LTS estimator has been analyzed by for instance Čížek [2005], Víšek [2006]. The asymptotic distribution of the LTS is found to be normal albeit with a variance depending on the mixing unknown mixing distribution and where the limit of the scale estimator also depends on the contamination. Inference therefore depends on nuisance parameters. Moreover, the least squares estimator remains consistent and efficient as leverage effects are ruled out in the ϵ-contamination model.

In the LTS model the h 'good' observations have i.i.d. normal errors that are independent of the regressors. The $n - h$ outliers have errors with support outside the realized range of the 'good' errors but are otherwise unrestricted. The 'outlier' errors are also allowed to depend on the regressors, which can generate the bad leverage effects. The LTS estimator is maximum ϵ-likelihood in the LTS model [Berenguer-Rico et al., 2022]. Moreover, the LTS estimator has the same asymptotic distribution as the infeasible least squares estimator on the unknown set of 'good' observations [Berenguer-Rico et al., 2023] Here, the concern will be to estimate the number of 'good' observations.

Estimating the number of 'good' observations is a question of model selection among a family of LTS models with different number of 'good' observations. These LTS models are not nested. As the number of observations increases the possible number of models increases towards a continuum of models. We show that it is possible to estimate the proportion h/n of 'good' observations consistently.

References

Rousseeuw, P. J., (1984). Least median of squares regressions. *Journal of the American Statistical Association*, **79**, 871–880.

Berenguer-Rico, V. and Johansen, S. and Nielsen, B., (2022). A model where the Least Trimmed Squares estimator is maximum likelihood. *Journal of the Royal Statistical Society, Series B*, to appear.

Berenguer-Rico, V. and Nielsen, B., (2023). Least Trimmed Squares asymptotics: Regression with leverage. Mimeo.

Čížek, P., (2005). Least trimmed squares in nonlinear regression under dependence. *Journal of Statistical Planning and Inference*, **136**, 3967–3988.

Víšek, J. A., (2006). The least trimmed squares; Part III: Asymptotic normality. *Kybernetika*, **42**, 203–224.

A spatially smoothed MRCD estimator for local outlier detection

P. Puchhammer[1*] and P. Filzmoser[1]

[1] *TU Wien; patricia.puchhammer@tuwien.ac.at, peter.filzmoser@tuwien.ac.at.*
[*] *Presenting author*

Keywords. *Local outlier detection; Multivariate data; Spatial data; MRCD estimation.*

Many methods are available for multivariate outlier detection but until now only a hand full are developed for spatial data where there might be observations differing from their neighbors, so-called local outliers. Although there are methods based on a pairwise Mahalanobis distance approach, the type of the covariance matrices used is not yet agreed upon. For example, Filzmoser et al. [2013] propose a global covariance while Ernst and Haesbroeck [2016] suggest a very local structure by estimating one covariance matrix per observation.

To bridge the gap between the global and local approach by providing a refined covariance structure we develop spatially smoothed covariance matrices based on the MRCD estimator [Boudt et al., 2020] for pre-defined neighborhoods $a_1, \ldots, a_N$. As well known from the MCD literature, a subset of observations, the so-called H-set, is obtained by optimizing an objective function. In our case we obtain a set of optimal H-sets $\mathcal{H} = (H_1, \ldots, H_N)$ from minimizing the objective function

$$f(\mathcal{H}) = \sum_{i=1}^{N} \det\left((1-\lambda)\boldsymbol{K}_i(\mathcal{H}) + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij}\boldsymbol{K}_j(\mathcal{H})\right).$$

While $\boldsymbol{W} = (w_{ij})_{i,j=1,\ldots,N}$ represents the closeness of the neighborhoods, the parameter λ is essential for the degree of locality of the covariance matrices. The local covariance matrices $\boldsymbol{K}_i(\mathcal{H})$ are based on the MRCD convex combination of the sample covariance matrix of an H-set of the neighborhood a_i and a global target matrix. For the optimal set of H-sets $\mathcal{H}^* = (H_i^*)_{i=1,\ldots,N}$ of the objective function, the final covariance estimate for neighborhood a_i is defined as $\hat{\boldsymbol{\Sigma}}_{SSM,i} = (1-\lambda)\boldsymbol{K}_i(\mathcal{H}^*) + \lambda \sum_{j=1,j\neq i}^{N} \omega_{ij}\boldsymbol{K}_j(\mathcal{H}^*)$.

A heuristic algorithm based on the notion of a C-step is developed to find the optimal set of H-sets which also shows stable convergence properties in general. We

demonstrate the applicability of the new covariance estimators and the importance of a compromise between locality and globality for local outlier detection with simulated and real world data, and compare the performance with other state-of-the-art methods from statistics and machine learning.

Acknowledgements: This project has received funding by the European Commission within the Horizon 2021 programme under grant agreement ID 101057741.

References

Filzmoser, P., Ruiz-Gazen, A., and Thomas-Agnan, C. (2013). Identification of local multivariate outliers. *Statistical Papers*, **55**, 29–47.

Boudt, K., Rousseeuw, P. J., Vanduffel, S., and Verdonck, T. (2020). The minimum regularized covariance determinant estimator. *Statistics and Computing*, **30**, 113–128.

Ernst, M. and Haesbroeck, G. (2016). Comparison of local outlier detection techniques in spatial multivariate data. *Data Mining and Knowledge Discovery*, **31**, 371–399.

I4: Robustness for categorical data (A3)

Robust Inference for Categorical Response Model

Anna Clara Monti[1]

[1] *Department of Law, Economics, Management and Quantitative Methods, University of Sannio, Italy; acmonti@unisannio.it.*
[*] *Anna Clara Monti*

Keywords. *Anomalous data; Categorical response models; Link functions; M estimation; Robustness*

1 Abstract

Categorical responses arise in many fields, such as marketing, finance, medicine, social and biomedical sciences. The interest focuses on models that describe the dependence of the response on the subjects' covariates. Since the support of the response is discrete and finite, it is often assumed that classical estimation and testing techniques are not affected by deviations from the stochastic assumptions. Nevertheless, outlying covariates as well as anomalous responses can strongly affect the reliability of likelihood based inferential procedures. Two approaches are considered in order to handle anomalous data in the cumulative model for ordered responses: the choice of a robust link function and the application of M estimators. Robust estimation for nominal response models and alternative models for ordinal responses is also discussed.

References

Iannario, M., Monti A. C., Piccolo, D. and Ronchetti, E. (2017). Robust inference for ordinal response models. *Electronic Journal of Statistics*, **11**, 3407–3445.

Iannario, M., Monti A. C. (2023). Robust logistic regression for ordered and unordered responses. *Manuscript.*

Scalera, V., Iannario, M., Monti, A.C. (2021). Robust link functions. *Statistics*, **55**, 963–977.

Robust Correspondence Analysis and its applications

Marco Riani[1*], Anthony C. Atkinson[2], **Francesca Torti** [3] and Aldo Corbellini[1]

[1] *Dipartimento di Scienze Economiche e Aziendale and Interdepartmental Centre for Robust Statistics, Università di Parma, Parma 43100, Italy; marco.riani@unipr.it, aldo.corbellini@unipr.it.*
[2] *The London Schoolof Economics, London WC2A 2AE, UK; a.c.atkinson@lse.ac.uk*
[3]*European Commission, Joint Research Centre (JRC), Ispra 21027, Italy; francesca.torti@ec.europa.eu*

Keywords. *Robust correspondence analysis; Contingency table; Minimum covariance determinant estimation; Outlier detection; Informative plotting*

1 Abstract

Correspondence analysis is a method for the visual display of information from two-way contingency tables (Greenacre [2017]).

We (Riani et al. [2022]) introduce a robust form of correspondence analysis based on minimum covariance determinant estimation. This leads to the systematic deletion of outlying rows of the table and to plots of greatly increased informativeness. Our examples are trade flows and consumer evaluations of the perceived properties of cars.

The robust method requires that a specified proportion of the data be used in fitting. To accommodate this requirement we provide an algorithm that uses a subset of complete rows and one row partially, both sets of rows being chosen robustly. We prove the convergence of this algorithm.

Figure 1 is an example of the plots we produce, with outliers highlighted as filled blue circles and 99.9% confidence intervals as red ellipses.

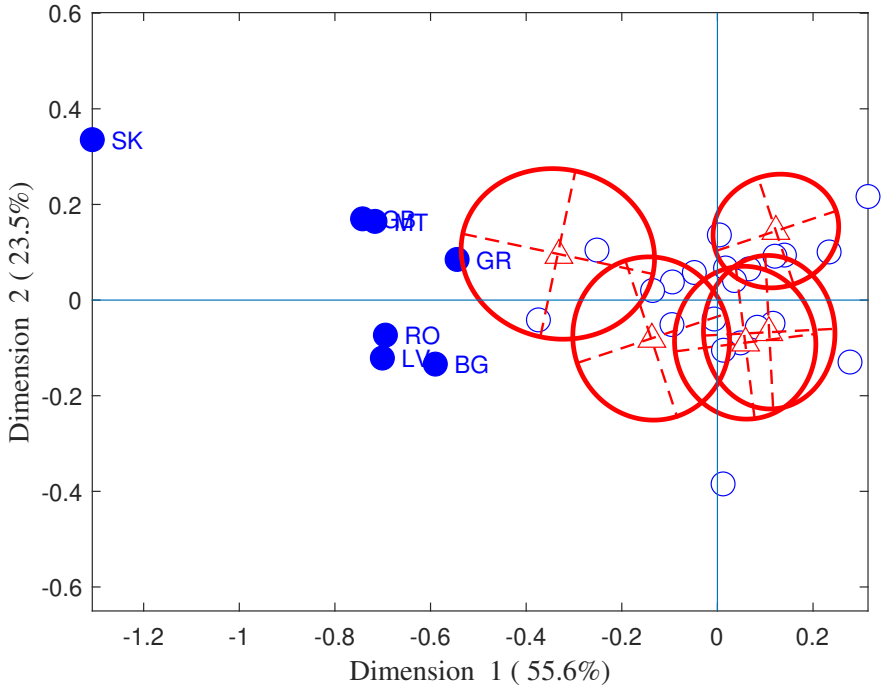


Figure 1: Trade data. Correspondence analysis plot from Minimum Covariance Determinant analysis. Filled circles are the seven countries declared as outliers. The ellipses give a 99.9% confidence interval for the positions of five levels of prices.

References

Greenacre, M. (2017) *Correspondence analysis in practice*, 3rd edition. Boca Raton, FL: Chapman and Hall CRCPress.

Riani, M., Atkinson A.C., Torti, F. & Corbellini, A. (2022). Robust correspondence analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **71**, 1063–2041.

C4: Tests (A5)

A robust multivariate combined test for comparison studies

M. Marozzi

Ca' Foscari University of Venice; marco.marozzi@unive.it.

Keywords. *Hypothesis testing; Nonparametric tests; Ranks; Combined tests.*

1 The new test

High–dimensional low sample size data with complex dependence structure are often encountered in fields like genomics, proteomics and metabolomics. The aim of this research is to propose a robust test for comparison studies that can be applied even when the number of variables p is much larger than the number of units N, and the underlying population distributions are heavy–tailed or skewed. To address such comparison studies, Marozzi [2016] proposed an exact, unbiased, consistent and powerful test based on interpoint distances. In theory, if a particular underlying distribution is assumed, it is possible –at least for some specific alternative hypotheses– to find the most powerful test for that particular distribution; unfortunately there is no uniformly most powerful test for all distributions. Different tests perform differently for the same distribution. Generally in practice, in particular when the sample sizes are not large, it is difficult to see how parent populations are distributed and then the problem of which test should be selected arises. This research aims at contributing to this issue by extending the interpoint distance based test proposed by Marozzi [2016].

Let $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_m)$ and $\mathbf{Y} = (\mathbf{Y}_1, ..., \mathbf{Y}_n)$ be two–independent random samples from p–variate populations with continuous cumulative distribution functions $F(\mathbf{Z})$ and $F(\mathbf{Z}-\boldsymbol{\mu})$ respectively. $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)'$ is the location difference parameter with $-\infty < \mu_h < \infty$, $h = 1, ..., p$. $N = m + n$. We test $H_0 : \boldsymbol{\mu} = \mathbf{0}$ against $H_1 : \boldsymbol{\mu} \neq \mathbf{0}$, ie whether the two samples come from the same unknown population. $\mathbf{0}$ denotes a vector of length p of all zeros. Let $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_N)$ be the pooled sample with $\mathbf{Z}_i = \mathbf{X}_i = (X_{i1}, ..., X_{ip})', i = 1, ..., m$ and $\mathbf{Z}_{m+j} = \mathbf{Y}_j = (Y_{j1}, ..., Y_{jp})', j = 1, ..., n$. Jureckova & Kalina [2012] proposed an unbiased distribution free test that can be computed when $p \gg N$. The steps to compute this test are: (i) compute the $N-1$ Euclidean interpoint distances ${}^2l_{ik}$, $k \neq i$ between $\mathbf{X}_i$ and the other elements of the pooled sample $\mathbf{Z}$, where i is randomly selected from $\{1, ..., m\}$; (ii) compute the ranks

${}^2R_{ik}$ of ${}^2l_{ik}$, $k = 1, ..., N$, $k \neq i$; (iii) compute the test statistic as ${}^2J_i = \sum_{k=m+1}^{N} {}^2R_{ik}$. Large values of ${}^2J_i$ speak against H_0. The ${}^2J_i$ test is the one–sided Wilcoxon rank sum test on interpoint distances. H_0 is tested by testing $\tilde{H}_0$ that the distributions of the interpoint distances are the same. Marozzi [2016] improves the Jureckova & Kalina [2012] test by (i) performing its permutation version ${}^2PJ$ (exchangeability holds) and (ii) combining m permutation ${}^2PJ$ tests, one for each element of the first sample. The observed value of the Marozzi [2016] ${}^2M$ test statistic is defined as ${}^2_0M = \min_{i=1,...,m}({}^2_0QPJ_i)$ where ${}^2_0QPJ_i$ denotes the observed p–value of the ${}^2PJ_i$ test, $i = 1, ..., m$. The test rejects for small values of its statistic.

This research improves the ${}^2M$ test by applying an additional round of combination. Consider the Minkovski distances of order $1 \leq d \leq \infty$. Note that for $0 < d < 1$ L_d is a quasi–metric because it violates the triangular property. However, we consider also the $L_{-\infty}$ distance ${}^{-\infty}l_{ik} = \min_{1 \leq h \leq p} |X_{ih} - Z_{kh}|$. Note that $L_{-\infty}$ is not a metric nor a quasi–metric (the term distance is used loosely). We can combine $2 < D < \infty$ different combined M tests considering their p–values, denoted by ${}^dQM$. The proposed test is defined using again the minimum p-value rule as

$$C_{MIN} = \min\left({}^{d_a}QM,\ a = 1, ..., D,\ 2 < D < \infty\right),$$

where ${}^{d_a}QM$, is the p–value of the ${}^{d_a}M$ test. The C_{MIN} test p-value can be computed using permutations.

A pilot power comparative study based on simulations of several ${}^{d_a}M$ tests showed that ${}^1M$ and ${}^2M$ tests always perform very similarly and are more powerful than the ${}^{-\infty}M$ test under normal distributions whereas the contrary happens under Cauchy distributions. The ${}^{\infty}M$ test is never the most powerful test. Therefore we propose this simple combined test $C = \min_{d \in \{-\infty, 2\}} \left({}^dM\right)$. The C test is unbiased and consistent. Simulations show that the test is appealing in practice being very powerful under normal, heavy–tailed and skewed distributions, while in contrast ${}^2M$ and ${}^{-\infty}M$ tests are powerful only for particular types of distributions. Moreover the power of the C test is not affected by the dependence structure of the data and tends to increase as p increases.

Possible directions for further research are the adaptation of the selection of ${}^{d_a}QM$ tests and D on the data at hand and a quality control chart based on the proposed test framework.

References

Jureckova, J. & Kalina, J. (2012). Nonparametric multivariate rank tests and their unbiasedness. *Bernoulli*, **18**, 229–251.

Marozzi, M. (2016). Multivariate tests based on interpoint distances with application to magnetic resonance imaging. *Statistical Methods in Medical Research*, **25**, 2593–2610.

FKWC tests for differences in the covariance structure of functional data

K. Ramsay[1*] and S. Chenouri[2]

[1] *York university; kramsay2@yorku.ca*
[2] *University of Waterloo; schenouri@uwaterloo.ca*
[*] *Presenting author*

Keywords. *Robust; Nonparametric; Covariance operator; Functional data; Hypothesis testing*

We present a new class of robust, nonparametric multi-sample hypothesis tests for differences in the covariance operator of functional data, called functional Kruskal-Wallis for covariance (FKWC) tests. FKWC tests use a pooled center-outward ordering of the sampled functions to measure differences in covariance structure between samples. The ordering is based on functional data depth, which we show is connected to the covariance operator. We will present some theoretical aspects and simulation study results of the FKWC tests. We also discuss how FKWC tests handle various challenges associated with functional data, such as computation, high-dimensionality, and outliers.

Power enhancement for dimension detection of Gaussian signals

Gaspard Bernard[1*] and **Thomas Verdebout**[1]

Université libre de Bruxelles; bernard.gaspard@ulb.be, thomas.verdebout@ulb.be

Keywords. *Signal dimension; Hypothesis testing; Latent roots.*

1 Abstract

We consider in the present paper the classical problem of testing

$$\mathcal{H}_{0q}^{(n)} : \lambda_q^{(n)} > \lambda_{q+1}^{(n)} = \ldots = \lambda_p^{(n)},$$

where $\lambda_1^{(n)}, \ldots, \lambda_p^{(n)}$ are the ordered latent roots of covariance matrices $\mathbf{\Sigma}^{(n)}$. We show that the usual Gaussian procedure $\phi^{(n)}$ for this problem essentially shows no power against alternatives of weaker signals of the form

$$\mathcal{H}_{1q}^{(n)} : \lambda_q^{(n)} = \lambda_{q+1}^{(n)} = \ldots = \lambda_p^{(n)}.$$

This is very problematic if the latter procedure is used to perform inference on the true dimension of the signal. We show that the same test $\phi^{(n)}$ enjoys some local and asymptotic optimality properties to detect alternatives to the equality of the $p-q$ smallest roots of $\mathbf{\Sigma}^{(n)}$ provided that $\lambda_q^{(n)}$ and $\lambda_{q+1}^{(n)}$ are sufficiently separeted. We obtain tests $\phi_{\text{new}}^{(n)}$ for the problem that keep the local and asymptotic optimality properties of $\phi^{(n)}$ when $\lambda_q^{(n)}$ and $\lambda_{q+1}^{(n)}$ are sufficiently separeted and properly detect alternatives of the form $\mathcal{H}_{1q}^{(n)}$. We also show how our tests can be turned into tests that are robust to Gaussian assumptions. Our results are illustrated via simulations and on a gene expression dataset from which we also discuss the problem of estimating the dimension of the signal.

Thursday 25 May 2023

Keynote - John H.J. Einmahl (A3)

Extreme value inference for heterogeneous power law data

John H.J. Einmahl

Tilburg University, The Netherlands

We extend extreme value statistics to independent data with possibly very different distributions. In particular, we present novel asymptotic normality results for the Hill estimator, which now estimates the extreme value index of the average distribution. Due to the heterogeneity, the asymptotic variance can be substantially smaller than that in the i.i.d. case. As a special case, we consider a heterogeneous scales model where the asymptotic variance can be calculated explicitly. The primary tool for the proofs is the functional central limit theorem for a weighted tail empirical process. We also present asymptotic normality results for the extreme quantile estimator. A simulation study shows the good finite-sample behavior of our limit theorems. We also present an application to assess the tail heaviness of earthquake energies. Joint work with YI He (University of Amsterdam).

S2: In honor of David Tyler (A3)

High breakdown regularized covariance matrices

David E. Tyler[1], Mengxi Yi[2] and Klaus Nordhausen[3]

[1] *Department of Statistics, Rutgers University, USA; dtyler@stat.rutgers.edu*
[2] *School of Statistics, Beijing Normal University, China; mxyi@bnu.edu.cn*
[3] *Dept of Math and Stat, Univ. of Jyväskylä, Finland; klaus.k.nordhausen@jyu.fi*

Keywords. *Affine Equivariant; Cross Validation; High Breakdown; Regularization; Scatter Matrix.*

Abstract

A median likelihood based cross validation criterion is proposed for selecting the tuning parameter within a a class or regularized scatter matrix estimates. This cross validation criterion helps assure the resulting tuned scatter matrix estimate is a good fit to the data as well as possessing a high breakdown point.

A motivation for this new median likelihood based criterion is that when it is optimized over all positive definite matrices, rather than only over the regularized candidates, the resulting scatter matrix estimate represents a newly introduced high breakdown point affine equivariant multivariate scatter statistic.

Directional distributions and the half-angle principle

J. T. Kent[1*]

[1] *Department of Statistics, University of Leeds LS2 9JT, UK; j.t.kent@leeds.ac.uk*
[*] *Presenting author*

Keywords. *Angular central Gaussian distribution; Gnomonic projection; Möbius transformation; Stereographic projection; Wrapped Cauchy distribution.*

1 Overview

The wrapped Cauchy (WC) distribution on the circle is a remarkable distribution that appears in a wide variety of seemingly unrelated settings in probability and statistics. The angular central Gaussian (ACG) distribution is another important distribution in directional statistics. It was used by Tyler [1987a] and Tyler [1987b] to construct and study a robust estimator of a scatter matrix for q-dimensional multivariate data.

Angle halving, or alternatively the reverse operation of angle doubling, is a useful tool when studying directional distributions. It is especially useful on the circle ($q = 2$) where, in particular, it yields an identification between the wrapped Cauchy and the angular central Gaussian distributions. That is, doubling a random angle following an ACG distribution (a two-to-one mapping) yields a random angle following a WC distribution. This identification is obvious, but several other relationships between the two distributions are more subtle. Some of these are listed below with more details in Kent [2023]

First, both the WC and ACG distributions are closed under suitable transformation groups (the Möbius and rescaled linear transformations, respectively). The identification between these two groups seems novel and surprising.

Next, the EM algorithm can be used to estimate the parameters of the ACG distribution, by treating the angular observations as incomplete observations from a bivariate normal distribution with mean 0. Similarly, an EM algorithm can be constructed for the WC distribution by treating the angular data as incomplete observations from a certain "squared complex normal distribution". See, e.g., Kent

& Tyler [1988] for the algorithm, though the EM interpretation was not properly recognized there.

Finally, both distributions are related to the Cauchy distribution on the line. Under stereographic projection from the circle to the line, the WC distribution turns into a Cauchy distribution. Similarly, under gnomonic projection from the circle to the line, the ACG distribution also turns into a Cauchy distribution.

References

Kent, J. T. (2023). Directional distributions and the half-angle principle. in *Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler*, eds. M. Yi & K. Nordhausen. Springer, to appear.

Kent, J. T. & Tyler, D. E. (1988). Maximum likelihood estimation for the wrapped Cauchy distribution. *Journal of Applied Statistics*, **15**, 247–254.

Tyler, D. E. (1987a). A distribution-free M-estimator of multivariate scatter. *Annals of Statistics*, **15**, 234–251.

Tyler, D E (1987b). Statistical analysis for the angular central Gaussian distribution on the sphere. *Biometrika*, **74**, 579–589.

C5: Processes and likelihood methods (A5)

Asymptotic behavior of the Laplacian quasi-maximum likelihood estimator of affine causal processes

Jean-Marc Bardet[1*] and Yakoub Boularouk[2]

[1] *SAMM EA 4543 Université Paris 1 Panthéon Sorbonne (France), FR 2036 CNRS; bardet@univ-paris1.fr *Presenting author*
[2] *Université de Mila (Algérie), y.boularouk@centre-univ-mila.dz*

Keywords. *Laplacian Quasi-Maximum Likelihood Estimator; Robust estimation; Strong consistency; Asymptotic normality; ARMA-ARCH processes*

1 General Information

We prove the consistency and asymptotic normality of the Laplacian Quasi-Maximum Likelihood Estimator (QMLE) for a general class of causal time series including ARMA, AR(∞), GARCH, ARCH(∞), ARMA-GARCH, APARCH, ARMA-APARCH,..., processes. We notably exhibit the advantages (moment order and robustness) of this estimator compared to the classical Gaussian QMLE. Numerical simulations confirms the accuracy of this estimator.

References

Bardet, J-M., Boularouk,Y. and Djaballah, K. (2017) Asymptotic behaviour of the Laplacian quasi-maximum likelihood estimator of affine causal processes. *Electronic Journal of Statistics*, **11**, 452–479.

Bardet, J.-M. and Wintenberger, O. (2009) Asymptotic normality of the Quasi-Maximum likelihood estimator for multidimensional causal process, *Ann. Statist.*, **37**, 2730–2759.

Davis, R. and Dunsmuir, W. (1997) Least Absolute Deviation Estimation for Regression with ARMA Errors. *Journal of Theoretical Probability*, **10**, 481–497.

Francq, C. and Zakoian, J.-M. (2004) Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes, *Bernoulli*, **10**, 605–637.

Francq, C., Lepage, G. and Zakoian, J-M. (2011) Two-stage non Gaussian QML estimation of GARCH models and testing the efficiency of the Gaussian QMLE. *Journal of Econometrics*, **165**, 246–257.

Weighted likelihood methods for torus data

L. Greco[1*], C. Agostinelli[2] and G. Saraceno[3]

[1] *University Giustino Fortunato, Benevento, Italy; l.greco@unifortunato.eu*
[*] *Presenting author.*
[2] *Department of Mathematics, University of Trento, Italy; claudio.agostinelli@unitn.it.*
[3] *Department of Biostatistics, University of Buffalo, USA.*

Keywords. *Circular; EM; Outliers; Pearson residual; Robust distance.*

1 Introduction

Torus data are multivariate circular observations that arise as measurements on a periodic scale and recorded as angles measured in degrees or radians, either clockwise or counterclockwise from some origin. They can be obtained with instruments such as the compass, protractor, weather vane, sextant, theodolite. Furthermore, circular data also stem from the time of day measured on a 24-hour clock. Multivariate circular data arise commonly in many different fields, from envirometrics, to protein bio-informatics, from meteorology to robotics, as recorded in Mardia and Jupp [2000], Jammalamadaka and SenGupta [2001], Pewsey et al. [2013]. The data can be thought as points on the surface of a p-torus $\mathbb{T}^p$, embedded in a $p+1$-dimensional space, whose surface is obtained by revolving the unit circle in a $p-$dimensional manifold. When $p = 2$, the torus is obtained by topologically gluing both pairs of opposite edges of a square together with no twists. First, we get a hollow cylinder by joining the top and bottom sides of the square. Then, one end of the cylinder is stacked on top of the other to form the torus. The key to understanding torus data is periodicity, that reflects in the boundedness of the sample space.

The problem of modeling circular data has been tackled through suitable distributions, such as the von Mises and the wrapped normal. Wrapping is a popular method to define distributions for torus data. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ be a *linear* random vector with probability density function $m(\boldsymbol{x}; \theta)$, with $\boldsymbol{x} \in \mathbb{R}^p$ and $\theta \in \Theta \subseteq \mathbb{R}^p$. Consider that each component is wrapped around the unit circle according to $Y_j = X_j \bmod 2\pi,\ j = 1, 2, \ldots, p$, where mod denotes the modulus operator. Then, the distribution of $\boldsymbol{Y} = \boldsymbol{X} \bmod 2\pi$ is a $p-$variate wrapped distri-

bution with probability density function

$$m^{\circ}(\boldsymbol{y};\theta) = \sum_{\boldsymbol{j}\in\mathbb{Z}^p} m(\boldsymbol{y} + 2\pi\boldsymbol{j};\theta) \ ,$$

$\boldsymbol{y} = (y_1, y_2, \ldots, y_p) \in [0, 2\pi)^p$, $\boldsymbol{j} = (j_1, j_2, \ldots, j_p) \in \mathbb{Z}^p$.

Torus data are not immune to the occurrence of outliers, that is unexpected values, such as angles or directions, that do not share the main pattern of the bulk of the data. In particular, circular outliers differ from linear ones in that angular distributions have bounded support. One single outliers can lead the mean to minus or plus infinity. In contrasts, breakdown occurs in directional data when contamination causes the mean direction to change by π [Davies and Gather, 2005, 2006].

Here, we outline a general strategy to robust estimation of the parameters of a multivariate wrapped distribution that is based on a Classification Expectation Maximization algorithm, whose M-step is enhanced by the computation of a set of data dependent weights aimed to down-weight outliers. In particular, the attention is focused on the weighted likelihood methodology [Markatou et al., 1998]. We compare three alternative down-weighting schemes: one from Saraceno et al. [2021], another from Greco et al. [2021] and a third new proposal. The objective is to provide a reliable and non computationally demanding approach to achieve accurate robust model fitting but also to perform torus outliers detection based on formal rules and a robust fit.

References

P Laurie Davies and Ursula Gather. Breakdown and groups. *The Annals of Statistics*, 33(3):977–1035, 2005.

P Laurie Davies and Ursula Gather. Addendum to the discussion of" breakdown and groups". *The Annals of Statistics*, pages 1577–1579, 2006.

Luca Greco, Giovanni Saraceno, and Claudio Agostinelli. Robust fitting of a wrapped normal model to multivariate circular data and outlier detection. *Stats*, 4(2):454–471, 2021.

S.R. Jammalamadaka and A. SenGupta. *Topics in Circular Statistics*, volume 5 of *Multivariate Analysis*. World Scientific, Singapore, 2001.

K.V. Mardia. *Statistics of directional data*. Academic press, 1972.

K.V. Mardia and P.E. Jupp. *Directional Statistics*. Wiley, New York, 2000.

M. Markatou, A. Basu, and B. G. Lindsay. Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association*, 93(442): 740–750, 1998.

A. Pewsey, M. Neuhäuser, and G.D. Ruxton. *Circular statistics in R*. Oxford University Press, 2013.

Giovanni Saraceno, Claudio Agostinelli, and Luca Greco. Robust estimation for multivariate wrapped models. *Metron*, 79(2):225–240, 2021.

New persectives in sample complexity of Robust Markov Decicion Processes

P. Clavier[1,2], E. Le Pennec [1] and M. Geist[3]

[1] *Ecole polytechnique, CMAP, France ;* pierre.clavier@polytechnique.edu.
[2] *INRIA Paris ;*
[3] *Google Brain, Paris.*

Keywords. *Robust Markov Decision Processes; Sample Complexity, Regularised Markov Decision Processes*

1 Abstract

In this work, we study the sample complexity of obtaining an ϵ-optimal policy in Robust discounted Markov Decision Processes (RMDPs), assuming we have only access to a generative model of the nominal kernel. This problem is well studied in the non-robust case, and it is known that any planning approach applied to an empirical MDP estimated with $\tilde{\mathcal{O}}(\frac{H^3|S||A|}{\epsilon^2})$ samples provides an ϵ-optimal policy, which is minimax optimal. Results in the robust case are much more scarce, [Yang, W. , 2022]. For *sa-* (resp *s-*)rectangular uncertainty sets, the best known sample complexity is $\tilde{\mathcal{O}}(\frac{H^4|S|^2|A|}{\epsilon^2})$ (resp. $\tilde{\mathcal{O}}(\frac{H^4|S|^2|A|^2}{\epsilon^2})$), for specific algorithms and when the uncertainty set is based on the total variation (TV), the KL or the Chi-square divergences. In this paper, we consider uncertainty sets defined with an L_p-ball (recovering the TV case), and study the sample complexity of *any* planning algorithm (with high accuracy guarantee on the solution) applied to an empirical RMDP estimated using the generative model. In the general case, we prove a sample complexity of $\tilde{\mathcal{O}}(\frac{H^4|S||A|}{\epsilon^2})$ for both the *sa-* and *s*-rectangular cases (improvements of $|S|$ and $|S||A|$ respectively). When the size of the uncertainty is small enough, we improve the sample complexity to $\tilde{\mathcal{O}}(\frac{H^3|S||A|}{\epsilon^2})$, recovering the lower-bound for the non-robust case for the first time and a robust lower-bound when the size of the uncertainty is small enough. Our analysis is based on a relation between robust MDPs and regularised MDPs which is the key of our analysis.

References

Zhang, L., Zhang, Z. Toward theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. The Annals of Statistics, 50(6), 3223-3248

Robust estimation for Markovian mixing processes

A. Lecestre[1]

[1] *University of Luxembourg; alexandre.lecestre@uni.lu*

Keywords. *robust estimation; dependence; mixing; Markov processes.*

The ρ-estimators developed by Baraud et al. [2017] and Baraud & Birgé [2018] are quite general and are proven to be robust in the independent setting. Our aim is to extend their framework to handle dependent observations without losing the robustness properties.

We use the variational formula of the Kullback-Leibler divergence to show that an exponential deviation bound in the independent context translates into the same bound in the dependent context, but in expectation and with an additional term that accounts for the dependency within the data. This additional term is the Kullback-Leibler divergence of the distribution of the data from the product distribution of the marginals.

This allows to get rid of the independence assumption in many results. Specifically, we apply it to ρ-estimators and get a non-asymptotic assumption-free bound in expectation. It echoes to the idea of robustness to dependence introduced in Chérief-Abdellatif & Alquier [2022], but in a more general framework.

Typically, one can use this robustness to dependence in order to estimate the stationary distribution of a mixing processes. Our idea to exploit the mixing is to select a sub-sample in a way that reduces the dependency within this subsample. Then the problem is to realise a compromise betweeen the sample size and the distance from independence. We apply this to different examples of Markovian processes for which we have an idea of the mixing regime, such as finite state space hidden Markov models and some discretely observed diffusion processes. We show for our examples that ρ-estimators achieves the optimal rate, up to potential logarithmic factors, conserving their robustness properties with respect to misspecification, outliers or contamination.

References

Baraud, Y. & Birgé, L. (2018). Rho-estimators revisited: General theory and applications. *The Annals of Statistics*, **46**, 3767–3804.

Baraud, Y., Birgé, L. & Sart, M. (2017). A new method for estimation and model selection:ρ-estimation. *Inventiones mathematicae*, **207**, 425–517.

Chérief-Abdellatif, B. & Alquier, P. (2022). Finite sample properties of parametric MMD estimation: Robustness to misspecification and dependence. *Bernoulli* **28**, 181 – 213.

I5: Robustness in survey sampling (A3)

Robust imputation procedures in the presence of influential units in surveys

Jia Ning Zhang[1], Sixia Chen[2] and **David Haziza**1∗

[1] *University of Ottawa; jzhan434@uottawa.ca; dhaziza@uottawa.ca)*
[2] *University of Oklahoma; Sixia-Chen@ouhsc.edu ;*
∗*Presenting author*

Keywords. *Adaptive tuning constant; Finite population; Imputed estimator; Influential unit; Item nonresponse.*

1 Abstract

Every time data are collected, it is virtually certain that we will face the problem of missing data. Missing data are undesirable because they make estimates vulnerable to nonresponse bias. In surveys, it is customary to distinguish unit nonresponse from item nonresponse. The former occurs when no usable information is collected on a sample unit, whereas the latter is characterized by the absence of information limited to some survey variables only. Unit nonresponse is usually handled through weight adjustment procedures methods, whereas item nonresponse is typically treated by some form of single imputation. Single imputation consists of constructing one replacement value to fill in for the missing value. The imputation process starts with postulating an imputation model, which is a set of assumptions about the conditional distribution of the survey variable requiring imputation. The estimator of a population mean, called an imputed estimator, is consistent provided that the first moment of the imputation model is correctly specified. However, the imputed estimator may be highly unstable in t the presence of influential units in the sample. A unit is said to be influential if its inclusion or exclusion from the computation has a large impact on the resulting estimate. We distinguish influential units from gross measurement errors. The latter are identified and corrected at the data-editing stage. In contrast, an influential unit corresponds to a respondent who exhibits a correctly recorded value. An influential unit may thus represent other similar units in the set of nonrespondents or in the non-sampled part of the population. This type of units has been called representative outliers by Chambers (1986) and are the focus of the talk. The issue of influential units is common in business surveys: on the one hand, the distribution of economic variables is typically highly skewed, which generates a conducive ground for the presence of influential units. On the

other hand, an influential unit can arise when the measure of size recorded on the sampling frame and used to stratify the population is considerably smaller than the size recorded on the field. This unit is then placed in a stratum with smaller units. As a result, it will generally exhibit a large y-value combined with a large weight, which makes it potentially harmful. These units are often referred to as stratum jumpers.

The rationale behind the treatment of influential values is to produce more stable but biased robust estimators. Therefore, we face a trade-off between bias and variance. The hope is that the mean square error of robust estimators would be smaller than that of the corresponding non-robust version. In this talk, we consider deterministic linear regression imputation. In the presence of influential unit, it may be tempting to replace the customary weighted least-squares estimator by a robust version; e.g. an M-estimator based on the Huber function and tuning constant equal to 1.345. However, classical robust estimators are generally not satisfactory as they may lead to imputed estimators with significant negative bias. This is due to the fact that influential units are generally not unique as mentioned above. A more appropriate would be to use an adaptative tuning constant; i.e., a constant whose value increases as the sample size increases. We propose an adaptive tuning constant based on the concept of conditional bias of a unit, which is an appropriate measure of influence in a finite population setting. We will present the results of a simulation study that assesses the performance of the proposed method in terms of bias and efficiency, for a wide class of distributions.

References

Chambers, R. L. (1986). Outliers Robust Finite Population Estimation. *Journal of the American Statistical Association*, **81**, 1063–1069.

Bias control for M-quantile-based small area estimators

F. Schirripa Spagnolo[1*], Gaia Bertarelli[2], Raymond Chambers[3], David Haziza[4], and Nicola Salvati[1]

[1] *University of Pisa; francesco.schirripa@unipi.it, nicola.salvati@unipi.it.*
[2] *Sant'Anna School of Advanced Studies; gaia.bertarelli@santannapisa.it*
[3] *University of Wollongong; ray@uow.edu.au*
[4] *University of Ottawa; dhaziza@uottawa.ca*
[*] *Presenting author*

Keywords. *Small area estimation; Robust estimators; Bias correction.*

Abstract

Representative outlier units occur frequently in surveys. As a result, several methods have been proposed to mitigate the effects of them in survey estimates. If outliers are a concern for estimation of population quantities, it is even more necessary to pay attention to them in a small area estimation (SAE) context, where sample size is usually very small and the estimation in often model-based. Chambers and Tzavidis [2006] explicitly addressed this issue of outlier robustness in SAE, using an approach based on fitting outlier robust M-quantile models to the survey data. More recently, Sinha and Rao [2009] also addressed this issue from the perspective of linear mixed models. Both these approaches, however, use plug-in robust prediction, i.e. they replace parameter estimates in optimal, but outlier-sensitive, predictors by outlier robust versions (a robust-projective approach). Unfortunately, these predictors are efficient under correct model specification and assumptions, but they may be sensitive to the presence of outliers because they use plug-in robust prediction which usually leads to a low prediction variance and a considerable prediction bias. Dongmo Jiongo et al. [2013] and Chambers et al. [2014] proposed bias corrected method to reduce the prediction bias when the response variable is continuous. In this talk, we focus on M-quantile approach and we propose two general methods (i.e., for continuous and discrete data) to reduce the prediction bias of the robust M-quantile predictors in SAE context. The first estimator is based on the concept of conditional bias and extends the results of Beaumont et al. [2013] and Favre-Martinoz [2015]. Then, we propose an unified approach to M-quantile predictors for continuous and discrete data which is based on a full bias correction and it could be viewed as a generalization of Chambers [1986] approach. A Monte-Carlo simulation

study is conducted and its results suggest that our approaches mainly improve the efficiency and they control the bias prediction error of M-quantile predictors when the population contains units that may be influential if selected in the sample. The methodology proposed is applied to Italian annual Labour Force Survey data for estimating the proportion of the unemployed in local labour market areas.

References

Beaumont, J.-F., Haziza, D. & Ruiz-Gazen, A. (2013). A unified approach to robust estimation infinite population sampling. *Biometrika,* **100**, 555-569.

Chambers, R. L. (1986). Outlier robust finite population estimation. *J. Am. Statist. Ass.,* **81**, 1063–1069.

Chambers, R. & Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, **93**, 255–268.

Chambers, R., Chandra, H., Salvati, N., & Tzavidis, N. (2014) Outlier robust small area estimation. *J. Roy. Stat. Soc. B,*, **76** 47–69

Dongmo Jiongo, V., Haziza, D. & Duchesne, P. (2013). Controlling the bias of robust small area estimators. *Biometrika,* **100**, 843–858.

Favre-Martinoz C. (2015) Estimation robuste en population finie et infinite. PhD Thesis.

Sinha, S. K. & Rao, J. N. K. (2009). Robust small area estimation. *Canadian Journal of Statistics,* **37**, 381-399.

Revivals: An R Package for Robust Estimation in Survey Sampling

C. Favre-Martinoz[1*], A. Ruiz-Gazen[2] and Fortó Cornella[2]

[1] *Insee, Département de la Démographie, France, cyril.favre-martinoz@insee.fr*
[2] *Toulouse School of economics, Université toulouse 1; second.author@email*
[*] *Presenting author*

Keywords. *Winzorisation ; Conditional bias; Robust estimation, R package*

1 Introduction

The presence of influential values in the estimation of population totals is an important problem that numerous polling organizations have to commonly deal with. While winsorisation methods seem to have become the most common technique to treat them, Beaumont et al. [2013] proposed to use the conditional bias to construct an equivalent robust estimator that lessens the effect of these outliers. This presentation illustrates how to use the R package revivals to implement this robust estimation. The latter also incorporates separate functions to compute conditional biases, robust weights, and the associated tuning constants. After a brief introduction on the state of the theory of robustness estimation, we provide a quick implementation example of the main function of the package. For that, we use real data about the French property values open database.

2 Robustness in survey sampling

The notion of robustness in survey sampling differs from that of classical statistics for several reasons. The first one is conceptual. In classical statistics, it is assumed that the data generating process for the majority of the data differs from the process generating the remaining part of observations, that are then considered as outliers. In this framework, the primary interest and hence the estimation is only focused on the distribution of the main bulk of data.
In survey sampling, outliers cannot simply be excluded from the group of observations of a sample, because we consider that an editing phase has already been done beforehand. We have already made sure that there are no missing or inconsistent

values (e.g. an observation given in euros instead of thousands of euros), nor any invalid input errors left, and so we cannot remove them because they are part of our population. This means that, contrarily to classical statistics, they must also intervene in the interest parameter computation that we are seeking to estimate, just as any other non-outlier observation.
Thus, the use of estimators that may drastically downweight some valid data is not generally recommended, as they can lead to biased estimators. We will favour robust estimators in the sense that (i) they are more stable than usual estimators against influential observations and turn out to be nearly as efficient as classical estimators, plus (ii) they converge towards classical estimators as the sample and the population sizes increase.

It turns out that the conditional bias, introduced by Moreno-Rebollo et al. [1999], is a good measure of the contribution of a unit i to the variance of the total estimator. It was indeed used later on by Beaumont et al. [2013] in order to construct a consistent robust estimator in a design-based framework. In particular, they made the Narain-Horvitz-Thompson (HT) estimator $\hat{\theta} = \sum_{i \in s} d_i y_i$ robust to influential values. This estimator is essentially equivalent to the forms of winsorisation presented by Dalén [1987] and Tambay [1988]. What varies, though, is the associated tuning constant, as well as the system of weights used to obtain each of these estimators.
The revivals package introduces several separate functions to compute each of these metrics.

3 Application on a data set about property values

The package comes with another data set from the property values open database. The PVD (Property Values Data set) does not contain personal data, such as the vendor or buyer names. It only contains data on transactions: property type, surface, selling price, and so on.
The data set provided in the package corresponds to a narrower version of the base available in open data. Only the following variables have been kept and the rows with missing data on any of these variables have been deleted:

- property type (appartment, house,...) ;
- property value ;
- building surface.

The main purpose of the presentation is to provide a quick implementation example of the main function of the package on this dataset.

References

J.-F. Beaumont, D. Haziza, and A. Ruiz-Gazen. A unified approach to robust estimation in finite population sampling. *Biometrika*, 100(3):555–569, 2013.

J. Dalén. *Practical estimators of a population total which reduce the impact of large observations*. Statistiska centralbyrån, 1987.

J.-L. Moreno-Rebollo, A. Muñoz-Reyes, and J. Muñoz-Pichardo. Miscellanea. influence diagnostic in survey sampling: conditional bias. *Biometrika*, 86(4):923–928, 1999.

J.-L. Tambay. An integrated approach for the treatment of outliers in sub-annual economic surveys. In *Proceedings of the Section on Survey Research Methods: American Statistical Association*, pages 229–234, 1988.

C6: Robust inference (A4)

Robust and Efficient Post-selection Inference

Anand N. Vidyashankar

Department of Statistics, George Mason University authors; avidyash@gmu.edu.

Keywords. *Data splitting; Dependent data; Distributional robustness; Divergence measures;*

In recent years, there has been an increasing interest in accounting for model-selection uncertainty for various inferential tasks. If the proposed model includes an i.i.d. component, this can be achieved using data splitting (Khalili, A. and Vidyashankar, A.N. [2018]) and an additional regularization (if necessary). However, if the data are dependent and the true distribution is only known up to an *ambiguity set*, such methods tend to yield biased results. This presentation describes a new methodology that accounts for data dependence and model ambiguity in inference. Specifically, utilizing predictive information criteria and ambiguity sets defined via $\phi-$divergences, the methodology enables robust and efficient inference for independent and dependent data. Applications of the proposed methodology include post-selection inference for stationary and non-stationary time series models, Hawkes processes, and supervised clustering via an implicit network.

References

Khalili, Abbas and Vidyashankar, Anand N. (2018). Hypothesis testing in finite mixture of regressions: Sparsity and model selection uncertainty. *Canadian Journal of Statistics*, **3**, 429-457.

Universal closed-form confidence intervals for the ratio of two general population means in the paired design

Tsung-Shan Tsou[1*]

[1] *Institute of Statistics, National Central University, Taiwan*
Email: chopinmozart0422@gmail.com
ORCID: https://orcid.org/0000-0001-5537-2854
[*] *Presenting author*

Keywords. *Paired non-negative data; Bivariate lognormal; Model misspecification Method of variance estimates recovery; Generalized confidence interval.*

We propose a closed-form formula for the confidence interval of the ratio of means of two general nonnegative populations in the paired design. The confidence interval is likelihood-based, robust, and very easy to calculate. Most of all, no knowledge about the joint distribution of the data is required. We compare our interval with the methods based on the generalized confidence intervals and the variance estimates recovery technique via simulations. Numerical studies show that our interval is the shortest for the data considered and outperforms the two competitors for paired Poisson data. We present data analysis to illustrate the usefulness of our new robust technique.

Robustness under missing data: a comparison with special attention to inference

C. Baum[1*], H. Cevallos-Valdiviezo[2] and A. Van Messem[1]

[1] *Department of Mathematics, University of Liège, Allée de la Découverte 12, 4000 Liège, Belgium; carole.baum@uliege.be, arnout.vanmessem@uliege.be.*
[2] *ESPOL Polytechnic University, Escuela Superior Politécnica del Litoral, ESPOL, Facultad de Ciencias Naturales y Matemáticas, Campus Gustavo Galindo Km. 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador; holgceva@espol.edu.ec.*
[*] *Presenting author*

Keywords. *Robustness; Missing data; Imputation; Inference.*

The size of datasets is increasing at a rapid pace, both in terms of the number of observations as in the amount of included observed characteristics. Along with this, the probability that these datasets contain missing values rises as well. However, certain statistical processes and machine learning techniques are incapable of dealing with incomplete data. As such, it is of the utmost importance that these missing values are dealt with in an adequate way.

Missing value imputation is a highly studied topic. A plethora of techniques have been proposed over the years to find suitable values to replace missing data, ranging from very simple techniques, such as mean or median imputation, to more complicated methods [Lin & Tsai , 2020], such as the popular Multiple Imputation by Chained Equations method (MICE) [van Buuren & Groothuis-Oudshoorn , 2011].

With larger datasets, it is also more likely to observe a number of atypical or extreme data due to measurement and/or encoding errors. These outliers can, to a varying degree, influence statistical analyses. To alleviate this problem, robust techniques have been introduced.

Nowadays, imputation techniques are widely in use, but a large-scale comparison of these methods – and especially in terms of their robustness against outliers – seems to be missing. During a first attempt to fill this gap, we evaluate a large selection of imputation techniques, involving classic and robust procedures, by means of a simulation study with continuous data and different configurations of missing data and outliers. To evaluate the imputation capability and robustness of the imputation techniques we computed the mean prediction error between the actual data values and the predictions obtained by the imputation method. In this study, we also evaluated computational speed of the imputation methods. Our simulations indicate

that, among the single imputation methods, robust linear regression using the MM-estimator and random forest imputation are among the most efficient and robust imputation methods, but these advantages naturally come at a cost, namely a higher computation time.

However, often, the main concern is on the analysis that is performed after imputation. Therefore, in the second phase of our research, we evaluated the inferences and predictions made by different robust regression methods combined with an imputation technique in a simulation study with different configurations of outliers and missing data. For the simulations, we used a similar setting as in Öellerer et al. [2016]. Both rowwise and cellwise outliers were generated, so we considered in the evaluation rowwise robust regression techniques as well as cellwise robust regression techniques. To evaluate the combined regression and imputation strategies in terms of inference capability, we measured the bias and variance of the estimated regression coefficients. To evaluate the prediction capability, we computed the mean prediction error.

References

Stefan van Buuren and Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45**(3), 1–67.

Wei-Chao Lin and Chih-Fong Tsai (2020). Missing value imputation: a review and analysis of the literature (2006-2017). *The Artificial Intelligence Review*, **53**(2), 1487–1509.

Viktoria Öellerer, Andreas Alfons and Christophe Croux (2016). The shooting S-estimator for robust regression. *Computational Statistics*, **31**, 829 - 844.

A Computationally Efficient Framework for Robust Estimation

Y. Zhang[1*], S. Orso[1], M. Victoria-Feser[1] and S. Guerrier[1]

[1] *University of Geneva, Switzerland; Yuming.Zhang@unige.ch, Samuel.Orso@unige.ch, Maria-Pia.VictoriaFeser@unige.ch, Stephane.Guerrier@unige.ch.*
[*] *Presenting author*

Keywords. *Simulation based estimation; Bias correction; Generalized linear models*

Constructing estimators that are robust to data contamination is non-trivial. Indeed, to be consistent, these estimators typically rely on a non-negligible correction term with no closed-form expression. Numerical approximation to this term can introduce finite sample bias, especially when the number of parameters p is relatively large compared to the sample size n. To address these challenges, we propose a simulation-based bias correction framework, which allows us to easily construct robust estimators with reduced finite sample bias. The key advantage of the proposed framework is that it bypasses the computation on the correction term in the standard procedure. The resulting estimators also enjoy consistency and asymptotic normality, and can be obtained computationally efficiently even when p is relatively large compared to n. The advantages of the method are highlighted with different simulation studies, such as logistic regression and negative binomial regression models. We also observe empirically that our estimators are actually comparable, in terms of finite sample mean squared error, to classical maximum likelihood estimators under no data contamination.

C7: High dimension and regularization (A5)

Robust estimation for high dimensional generalized linear models

C. Agostinelli[1*] and M. Valdora[2]

[1] *Department of Mathematics, University of Trento, Trento, Italy; claudio.agostinelli@unitn.it.*
[2] *Instituto de Cálculo and Department of Mathematics, University of Buenos Aires, Buenos Aires, Argentina; mvaldora@dm.uba.ar.*
[*] *Presenting author*

Keywords. *Elastic Net; High-dimension; GLM; Lasso; MT-estimators; Ridge; Robustness.*

1 Introduction

Generalized linear models (GLM) are an important tool in data analysis. In high dimensional problems, traditional methods fail, because they are based on the assumption that the number of observations is larger than the number of covariates. The problem of high dimensional data has been widely studied and penalized procedures have been proposed, see e.g. Friedman et al. [2001]. All the above proposals have a very good performance if all the observations follow the assumed model. However, if a small proportion of the observed data are atypical, they become unreliable. Robust estimators for high dimensional linear models have been proposed by Maronna [2011] and Smucler and Yohai [2017], among others. Avella-Medina and Ronchetti [2018] introduced penalized robust M-estimators for GLM, Bianco et al. [2019, 2022] proposed penalized robust estimators for logistic regression. MT-estimators Valdora and Yohai [2014] are particularly suited for GLM however they need good initial estimates Agostinelli et al. [2019]. In this work we present penalized MT-estimators for GLM we illustrate their theoretical properties and computational methods with particular attention to the initial estimates. Simulations and real examples confirm the robust properties of the propsed procedure.

References

C. Agostinelli, M. Valdora, and V.J. Yohai. Initial robust estimation in generalized linear models. *Computational Statistics & Data Analysis*, 134:144–156, 2019.

M. Avella-Medina and E. Ronchetti. Robust and consistent variable selection in high-dimensional generalized linear models. *Biometrika*, 105(1):31–44, 2018.

A.M. Bianco, G. Boente, and G. Chebi. Penalized robust estimators in logistic regression with applications to sparse models, 2019.

A.M. Bianco, G. Boente, and G. Chebi. Asymptotic behaviour of penalized robust estimators in logistic regression when dimension increases, 2022.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer, 2001.

R.A. Maronna. Robust ridge regression for high-dimensional data. *Technometrics*, 53(1):44–53, 2011.

E. Smucler and V.J. Yohai. Robust and sparse estimators for linear regression models. *Computational Statistics & Data Analysis*, 111:116–130, 2017.

M. Valdora and V.J. Yohai. Robust estimators for generalized linear models. *Journal of Statistical Planning and Inference*, 146:31–48, 2014.

The SgenoLasso, a new Lasso method dedicated to extreme observations in genomics

Charles-Elie Rabier[1*], Céline Delmas[2]

[1] *IMAG, Université de Montpellier, CNRS, France; charles-elie.rabier@umontpellier.fr.*
[2] *Université de Toulouse, INRAE, UR MIAT, F-31320, Castanet-Tolosan, France; celine.delmas.toulouse@inrae.fr*
[*] *Presenting author*

Keywords. *High-dimensional linear model; Variable selection; RALasso; Extreme data; Genomics*

In genomics, selective genotyping consists in genotyping (collecting DNA information at specific positions) only the individuals with extreme traits (i.e. with the largest or smallest trait values). This famous concept was introduced by [Lebowitz et al. , 1987]: the authors noticed that the highest or the lowest observations contain most of the signal on Quantitative Trait Loci (QTL), i.e. genes with quantitative effect on a trait. Later, [Lander & Botstein , 1989] elaborated this concept.

Nowadays, although the genotyping costs have drastically dropped, selective genotyping is still heavily used since we can optimize the statistical experiment by focusing on extreme individuals instead of random individuals. There is still a lack of tools to analyze properly this kind of data since classical penalized regressions (e.g. Lasso [Tibshirani , 1996]) are not dedicated to extreme observations.

From a statistical point of view, the linear model we are dealing with, presents the particularity of incorporating some correlation between the errors ε and the regressors, due to selective genotyping. As a consequence, we introduce the SgenoLasso [Rabier & Delmas , 2021], a new L1 penalized regression that models explicitly this correlation. The SgenoLasso relies on the "Interval Mapping" [Lander & Botstein , 1989], a famous concept in genetics that consists in scanning the genome by testing the presence of a QTL at each location. SgenoLasso is based on new limiting results on stochastic processes along the genome. SgenoLasso enjoys all known statistical properties of Lasso since the problem has been replaced in a classical L1 penalized regression framework. Typically, it is not the case for Lasso in presence of extreme data.

We compared the SgenoLasso with the "Robust Approximate quadratic Lasso" (RALasso) of Fan et al. [2017], which incorporates the Huber loss and a L1 penalty.

The RALasso can be viewed as a more flexible method than the Lasso: the loss function can be either quadratic or linear, depending on the error values. The tuning parameter helps to handle errors with different shapes and tails. Recall the [Huber , 1964] loss considered in the R package hqreg :
$\text{loss}(t) = \frac{t^2}{2M} 1_{|t|\leq M} + (|t| - M/2)\, 1_{|t|\geq M}$, where M is a tuning parameter. As soon as we multiply by $2M$ and that we replace M by α^{-1}, we obtain the RALasso loss described in formula (2.2) of Fan et al. [2017].

On the basis of a simulation study where only the largest individuals were selected, the RALasso, that models heavy tails and asymmetry, gave better results than classical methods such as Lasso, GroupLasso [Yuan & Lin , 2006] and Bayesian Lasso [Park et al. , 2008]. However, in that case, our SgenoLasso performed better than the RALasso. Last, we will show the superiority of the Adaptive version of the SgenoLasso, called the AdaptSgenoLasso, that allows to put more weights on some regressors of interests (e.g. well known genes).

References

Lander, E.S. & Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **138**, 235–240 .

Lebowitz, R.J., Soller, M., & Beckmann, J.S. (1987). Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Genetics*, **73**, 556–562.

Fan, J., Li, Q. & Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society B*, **79(1)**, 247–265.

Huber, P.J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, **35**, 73–101.

Park, T. & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, **103(482)**, 681–686.

Rabier, C.E. & Delmas, C. (2021). The SgenoLasso and its cousins for selective genotyping and extreme sampling: application to association studies and genomic selection. *Statistics*, **55(1)**, 18–44.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58(1)**, 267–288.

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, **68(1)**, 49–67.

Robust flexible GLM for high-dimensional data containing mixed variable types penalized with a combination of various penalty terms

L. Tubex[1*], S. Van Aelst[2], T. Verdonck[1,2]

[1] *Department of Mathematics, University of Antwerp - imec, Middelheimlaan 1, 2020 Antwerp, Belgium; lise.tubex@uantwerpen.be, tim.verdonck@uantwerpen.be*
[2] *Department of Mathematics, KU Leuven, Celestijnenlaan 200B, 3001 Leuven, Belgium; stefan.vanaelst@kuleuven.be*
[*] *Presenting author*

Keywords. *Generalized linear model; Various penalty terms; Mixed variable types; High-dimensional data; Robustness*

1 Abstract

The Generalized Linear Model (GLM) is a popular class of regression models that generalizes ordinary linear regression by allowing a large variety of distributions for the response variable. Robust estimators enable to reliably estimate the parameters even when a minority of the data may deviate arbitrarily far from the postulated model. Modern data may have a huge number of observations as well as a very large number of dimensions with different types of variables. New advanced robust estimation methods are required for this type of data. In this paper, we incorporate sparsity using a combination of various penalty terms in the robust GLM framework to extract the most relevant information from high-dimensional data containing mixed variable types. A computationally efficient algorithm is presented. The good performance is illustrated on simulated and real data, focusing on Poisson and logistic regression.

Elasso in estimating the signal dimension of ICA

M. Yi[1*] and K. Nordhausen[2]

[1] *School of Statistics, Beijing Normal University, China; mxyi@bnu.edu.cn.*
[2] *Department of Mathematics and Statistics, University of Jyväskylä, Finland; klaus.k.nordhausena@jyu.fi*
[*] *Presenting author*

Keywords. *Noisy ICA; Elasso; Order determination.*

Independent component analysis is often considered in a framework where the p observed variables are a mixtures of only $d < p$ latent independent components which are contaminated by white noise. The goal is then to estimate the number of latent components as well as the components itself. Meanwhile several approaches exist to estimate d which are all are based on the eigenvalues of the covariance matrix. However all these approaches were developed and tested in scenarios where p is moderately small. If p is however large the estimation of the eigenvalues suffers. To improve the estimation of d by better estimation of the eigenvalues we employ the recently suggested Elasso which penalizes the eigenvalue structure and groups them together when possible. We show how the Elasso can be used for estimation of d and show in simulations and in an example that it is better than the competing methods when p is large.

I6: Robust clustering I (A3)

Robust estimation and clustering under heavy tails

L. Barabesi[1], **A. Cerioli**[2*], L.A. García-Escudero[3] and A. Mayo-Iscar[3]

[1] *Department of Economics and Statistics, University of Siena, Italy; lucio.barabesi@unisi.it.*
[2] *Department of Economics and Management and University Centre "Robust Statistics Academy" (Ro.S.A.), University of Parma, Italy; andrea.cerioli@unipr.it*
[3] *Department of Statistics and Operations Research, University of Valladolid, Spain; lagarcia@uva.es; agustin.mayo.iscar@uva.es.*
[*] *Presenting author*

Keywords. *Consistency factor; Multivariate Student-t distribution; Robust distance; TCLUST; Trimming.*

1 Framework and Goals

It is well known that trimmed estimators of multivariate scatter, such as the Minimum Covariance Determinant estimator, are inconsistent unless an appropriate factor is applied to them in order to take the effect of trimming into account [Croux and Haesbroeck, 1999, Rousseeuw and Van Driessen, 1999, Cator and Lopuhaä, 2010]. The consistency factor is widely recommended and applied when uncontaminated data are assumed to come from a multivariate Normal model, but few applications exist outside this scenario.

In the first part of this contribution, drawing from Barabesi *et al.* [2023], we address the problem of computing a consistency factor for trimmed estimators of multivariate scatter in a heavy-tail scenario, when uncontaminated data come from a multivariate Student-t distribution. For such a purpose, we first derive a remarkably simple computational formula for the appropriate factor and show that it reduces to an even simpler analytic expression in the bivariate case. Exploiting our formula, we then develop a robust Monte Carlo procedure for estimating the usually unknown number of degrees of freedom of the assumed and possibly contaminated multivariate Student-t model, which is a necessary ingredient for obtaining the required consistency factor. We also provide substantial simulation evidence about the proposed procedure and apply it to data from image processing and financial markets.

In the second part of our work, we study the applicability of the suggested consistency factor in a multi-population framework, when the TCLUST algorithm is adopted to robustly cluster data generated from a contaminated mixture of multivariate Student-t distributions. Our approach takes advantage of the reweighting scheme for robust clustering developed by Dotto *et al.* [2018], which is then extended to a heavy-tail scenario. The connections with adaptive trimming and the forward search [Cerioli et al., 2019] are also considered.

References

Barabesi, L., Cerioli, A., García-Escudero, L.A. and Mayo-Iscar, A. 2023. Trimming heavy-tailed multivariate data. *Submitted*.

Cator, E.A. and H.P. Lopuhaä. 2010. Asymptotic expansion of the minimum covariance determinant estimator. *Journal of Multivariate Analysis* 101: 2372–2388.

Cerioli, A., A. Farcomeni, and M. Riani. 2019. Wild adaptive trimming for robust estimation and cluster analysis. *Scandinavian Journal of Statistics* 46: 235–256.

Croux, C. and G. Haesbroeck. 1999. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis* 71: 161–190.

Dotto, F., Farcomeni, A., García-Escudero, L.A. and Mayo-Iscar, A. 2018. A reweighting approach to robust clustering. *Statistics and Computing* 28: 477–493.

Rousseeuw, P.J. and K. Van Driessen. 1999. A fast algorithm for the Minimum Covariance Determinant estimator. *Technometrics* 41: 212–223.

On simulating skewed and cluster-weighted data for studying performance of clustering algorithms

V. Melnykov[1], Y Wang[1], Y. Melnykov[1], F. Torti[2], **D. Perrotta**[2*] and M. Riani[3]

[2] *European Commission, Joint Research Centre (JRC); francesca.torti@ec.europa.eu, domenico.perrotta@ec.europa.eu.*
[1] *The University of Alabama, Culverhouse College of Business; vmelnykov@culverhouse.ua.edu, wangy4@cofc.edu, ymelnykov@culverhouse.ua.edu*
[3] *University of Parma, Department of Economics and Management; mriani@unipr.it*
[*] *Presenting author*

Keywords. *Finite Mixture Model; Cluster Analysis; Skewed Clusters; Cluster Weighted Model; MixSim.*

1 Background

The objective of cluster analysis is to partition observations so that similar data points are grouped together while formed clusters are relatively distinct. There is rich computer science and statistics literature devoted to developing various clustering procedures. The performance comparison of various techniques is usually conducted based on either so-called classification data sets or simulated synthetic data. The former approach has an important advantage related to the fact that the comparison is carried out on real data. On the other hand, the assessment of the systematic clustering performance in various settings is hardly realistic without simulated data.

The discussion of several approaches to generating data for studying clustering techniques can be found in Maitra, R. & Melnykov, V. [2010] who also proposed simulating Gaussian mixture models according to the prespecified degree of pairwise overlap employed as a measure of components proximity. Algorithms related to the pairwise overlap have been implemented in the R package MixSim [Melnykov, V. & Chen, W.-C. & Maitra, R. , 2012] as well as MATLAB modules MixSim [Riani, M. & Cerioli, A. & Perrotta, D. & and Torti, F. , 2015] and MixSimReg [Torti, F. & Perrotta, D. & Riani, M. & and Cerioli, A. , 2019].

2 Contribution

Mixture modeling simulation procedures allowing to achieve the desired level of maximum or average overlap have been proposed in Maitra, R. & Melnykov, V. [2010]. While indisputably useful for studying clustering algorithms, elliptical clusters simulated from Gaussian mixtures represent just a fraction of a variety of settings modern state-of-the-art algorithms need to be tested on. In this contribution, we discuss important extensions of the proposed methodology beyond simulating elliptical data groups. We first discuss the application of transformation mixture models to simulating mixtures with pre-specified overlap. These mixtures can be effectively used for generating skewed data with desired clustering complexity. Then we explain how to measure overlap in the cluster-weighted modeling framework. We finally consider an extension of the proposed techniques capable of simulating skewed cluster-weighted data.

References

Maitra, R. & Melnykov, V. (2010). Simulating Data to Study Performance of Finite Mixture Modeling and Clustering Algorithms. *Journal of Computational and Graphical Statistics*. **2:19**, 354–376.

Melnykov, V. & Chen, W.-C. & Maitra, R. (2012). MixSim: An R Package for Simulating Data to Study Performance of Clustering Algorithms. *Journal of Statistical Software*. **51**, 1–25.

Riani, M. & Cerioli, A. & Perrotta, D. & and Torti, F. (2015). Assessing trimming methodologies for clustering linear regression data. Advances in Data Analysis and Classification. *Advances in Data Analysis and Classification*. **9**, 461-481.

Torti, F. & Perrotta, D. & Riani, M. & and Cerioli, A. (2019). Assessing trimming methodologies for clustering linear regression data. Advances in Data Analysis and Classification. *Advances in Data Analysis and Classification*. **13**, 227-257.

Mendelian randomization: A new robust causal effect estimator using summary data

A. García-Pérez

Departamento de Estadística, I.O. y C.N., Universidad Nacional de Educación a Distancia (UNED); agar-per@ccia.uned.es

Keywords. *Mendelian randomization; Von Mises expansion; Saddlepoint approximation.*

1 Abstract

Mendelian randomization (MR) is an increasingly used method of Causal Inference to study the effect of an Exposure X on an Outcome Y, by using genetic variants (usually single nucleotide polymorphisms, SNPs) as instrumental variables Z.

MR is used to avoid possible biases in the regression of Y on X due to lack of complete randomization in data, or the presence of reverse causation, or confounders. The use of MR leads us to a two stage linear regression process: first, for every genetic variant $Z_j, j = 1, ..., L$, a linear regression of X on Z_j is done, where for individuals $i = 1, ..., n$ is

$$X_i|Z_{ij} = \beta_{X_0} + \beta_{X_j} Z_{ij} + e_{X_{ij}}$$

from which we obtain the fitted values $\hat{X}$, used in a second regression of Y on these $\hat{X}$, obtaining finally (Pires Hartwig et al. [2017])

$$Y_i|Z_{ij} = \beta_{Y_0} + (\beta \cdot \beta_{X_j} + \alpha_j) Z_{ij} + e_{Y_{ij}} = \beta_{Y_0} + \beta_{Y_j} Z_{ij} + e_{Y_{ij}}$$

where β_{X_j} and β_{Y_j} represent the association of Z_j with the Exposure and the Outcome (only through X), respectively. The parameter $\beta \cdot \beta_{X_j}$ represents the effect of Z_j on Y through X, where β is the causal effect of X on Y we wish to estimate.

The two-stage least squares estimator of β (using variant j alone) is the quotient of the two linear regression slope estimators: of Y on Z_j and of X on Z_j,

$$\hat{\beta}_{R_j} = \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}}$$

that is equal to the sample covariances (or correlations) quotient. The inputs we only have, in observational studies using summary data, are the linear regression slope estimators and their standard errors, the latter used to weight the $\hat{\beta}_{R_j}$ estimators for all the L genetic variants, defining the inverse-variance weighted (IVW) estimator of β, as $IVW = \sum_{j=1}^{L} \omega_j \hat{\beta}_{R_j} / (\sum_{j=1}^{L} \omega_j)$, assuming that the L genetic variants are mutually independent, and the usual normality. This classic, and widely used estimator, has a 0% breakdown point.

There are several alternatives to this estimator, such as $M = median_{j=1,,,L}\{\hat{\beta}_{R_j}\}$, or a weighted median of the $\hat{\beta}_{R_j}$.

In this contributed paper we propose a new robust estimator of β in this context, first, obtaining an approximation for the distribution of the $\hat{\beta}_{R_j}$ under a scale contaminated normal model, for the independent but not identically distributed observations, using a VOM (von Mises) + SAD (saddlepoint approximation). This VOM+SAD approximation is

$$P_{F_1,\ldots,F_n}\left\{\hat{\beta}_{R_j} > t\right\} = 1 - \Phi(-\mu_s/\sigma_s) + \epsilon \sum_{i=1}^{n} [\Phi(-\mu_s/\sigma_s) - \Phi(-\mu_s/\sigma_s^{g_i})]$$

where Φ is the cumulative standard normal distribution function; $\mu_s = \mu_1 + ... + \mu_n$; $\sigma_s = \sqrt{\sigma_1^2 + ... + \sigma_n^2}$; $\sigma_s^{g_i} = \sqrt{\sigma_1^2 + ... + g_i^2\sigma_i^2 + ... + \sigma_n^2}$; ϵ is the percentage of contamination; g_i^2 is the contamination in scale and where μ_i and σ_i depend on t. Then, we replace in IVW the estimators $\hat{\beta}_{R_j}$ by the medians Me_j of distributions $P_{F_1,\ldots,F_n}$. Furthermore, the weights are replaced by the inverse of the $\hat{\beta}_{R_j}$ MADs, $v_j = 1/median\{|\hat{\beta}_{R_j} - M|\}$, defining the new estimator, based on the $\hat{\beta}_{R_j}$ distribution, as

$$\hat{\beta}_D = \frac{\sum_{j=1}^{L} v_j \, Me_j}{\sum_{j=1}^{L} v_j}$$

In the paper we study the properties of this new estimator and we also include simulations and real data examples.

References

Pires Hartwig, F., Smith, G.D. & Bowden, J. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, 1985–1998.

Hunting bias through trimming

I. Barrio[1,3], P. Gordaliza[1,2], H. Inouzhe[1*], J.M. Quintana[4] and M.X. Rodríguez-Álvarez[5]

[1] *Basque Center for Applied Mathematics (BCAM); pgordaliza@bcamath.org, hinouzhe@bcamath.org*
[2] *Universidad Pública de Navarra (UPNA)*
[3] *Universidad Del País Vasco (UPV/EHU); irantzu.barrio@ehu.eus*
[4] *Osakidetza - Servicio Vasco de Salud ; josemaria.quintanalopez@osakidetza.eus*
[5] *Universidad de Vigo; mxrodriguez@uvigo.es*
[*] *Presenting author*

Keywords. *Algorithmic Fairness; Optimal Transport; Contamination Neighbourhood; Generalized Trimming.*

1 Introduction and Notations

The generalization of Artificial Intelligence (AI) based-systems in the decision-making process in a wide variety of fields, particularly in the everyday and professional life, have raised serious ethical concerns about the implications of the adoption of such technologies. In a supervised fair learning setting, the aim of an algorithm is to learn the relationships between characteristic variables X and a target variable Y with the additional challenge of dealing with the presence of a protected attribute S, that conveys sensitive information about the observations X that should not be used for the prediction of $\hat{Y}$. In this sense, this variable S models the bias and we assume that it is observed. A similar situation occurs in the unsupervised learning problem (fair clustering), where the goal is to hide sensitive attributes during data partition by balancing the distribution of protected subgroups in each cluster.

Following the independence-based approach, an algorithm is called fair or unbiased when its outcome $\hat{Y}$ does not depend on S. The well-known criterion Demographic Parity (DP) requires the statistical independence between the outcome and the protected attribute $\hat{Y} \perp S$. In certain scenarios, the ground truth is available and the above definition could be weaken into a conditional independence that is given by the Equalized Odds (EO) criterion $\hat{Y} \perp S \mid Y$. A less restrictive setting, allows for X to mediate some legitimate dependence of $\hat{Y}$ with respect to S, but does not allow for direct influence from S to $\hat{Y}$. Hence, the fairness criteria of interest is

conditional independence of the form $\hat{Y} \perp S \mid X$.

In most situations, algorithms are inaccessible and the most practical approach is trying to obtain fairness by constraining the training sample. In this work we propose to remove the influence of the sensitive variable by trimming (partially or fully removing) a proportion of the input data as a pre-processing step to any further learning mechanism or bias analysis. Let us consider a binary protected attribute $S \in \{0, 1\}$, meaning a population divided into two categories, $S = 1$ for the minority (the unfavoured class), and $S = 0$ for the majority (the favoured class). The idea is to check if there is group bias in Y with respect to S. However, not all individuals in S are comparable, therefore some differences in the response variable Y may arise from genuine differences in the data, i.e, in $X|S = 0$ vs. $X|S = 1$. In order to eliminate these cases we want to trim an α proportion of the data to obtain the two closest possible conditional distributions. On this 'similar enough' population we can check for discrimination, in the sense of DP. A similar strategy can be used to obtain clusters that preserve spatial information and produce more fair partitions.

The model behind the procedure we propose is the popular contamination model:

$$\begin{aligned} \mu_0 &= \mathcal{L}(X|S = 0) = (1 - \alpha_0)P + \alpha_0 Q_0 \\ \mu_1 &= \mathcal{L}(X|S = 1) = (1 - \alpha_1)P + \alpha_1 Q_1 \end{aligned} \tag{1}$$

where P represents the common structure between the marginals, Q_0 and Q_1 the structure that produces the differences between them, and $\alpha_0, \alpha_1 \geq 0$ are the levels of mixing for each distribution. Under (1), there are, not necessarily unique, μ_0^* and μ_1^* which are α-trimmed versions of μ_0 and μ_1 and which have the same distribution, i.e., $d_{TV}(\mu_0^*, \mu_1^*) = 0$. Indeed, this characterizes the complete absence of bias in the training sample in the DP sense as proven in Gordaliza et al. [2019]. Thus, any unsupervised learning algorithm that learns on these α-trimmed versions of the conditionals should produce results that are independent from S. In the supervised case this is not guaranteed, due to possible distribution shifts, but at least one has a good starting point. Alternatively, in the case of equal conditional distributions one can check for direct bias due to S.

Recently, computation of multivariate trimmings has been made much more efficient. Using this, we provide methods to improve fairness in algorithmic considerations and also procedures to check bias. We provide a meaningful real world example, where we analyse the difference in healthcare outcomes during the COVID-19 pandemic in people with different nursing home status in the Basque Country in Spain.

References

Gordaliza, P., Del Barrio, E., Gamboa, F. & Loubes, JM. (2019). Obtaining fairness using optimal transport theory. International Conference on Machine Learning, PMLR, 2357–2365.

C8: Robust multivariate data analysis (A5)

Robust Estimation of Conditional ROC Curves

A. M. Bianco[1*] and G. Boente[1]

[1] *Universidad de Buenos Aires and CONICET; abianco@dm.uba.ar, gboente@dm.uba.ar.*
[*] *Presenting author*

Keywords. *ROC Curves; Robustness; Indirect Method; Functional Covariates.*

Diagnostic tests based on a continuous marker are a key tool in medical decisions. For that reason, it is of extreme importance to evaluate the ability of the test to distinguish between different states, such as healthy individuals from diseased ones. Receiver Operating Characteristic (ROC) curves are a very helpful instrument to assess the performance of a test based on a continuous marker.

Several factors, such as gender or blood pressure, may improve the discriminatory ability of the marker. In this cases, conditional ROC curves are useful to include covariate effects in the ROC analysis and to avoid oversimplification. The indirect method provides a way of adjusting ROC curves to covariates by means of regression models. When an infinite-dimensional covariate is measured, the indirect methodology is still suitable, but it would involve a functional covariate.

Aware of the impact that outliers may have on the diagnostic test accuracy, we focus on the robust aspects of the estimation procedures of the conditional ROC curve. In fact, since regression models are involved in the indirect approach, atypical data among the marker or the covariates may severely affect the estimation methods. With this motivation, we generalize the proposal given in Bianco *et al.* (2022) to a very general scenario in which the markers are modelled in terms of a functional partially linear model. The considered situations include the functional linear regression model and also the nonparametric or additive regression ones with real valued covariates.

The given approach enables us to cover a wide range of cases using a robust perspective. We obtain consistency results under standard regularity conditions. Through a Monte Carlo study, we compare the performance of the proposed estimators with that of the classical ones in clean and different scenarios of contamination.

References

Bianco, A. M., Boente, G. & González–Manteiga, W. (2022). Robust consistent estimators for ROC curves with covariates. *Electronic Journal of Statistics*, **16**, 4133–4161.

Robust classification tool for three-way data based on the SIMCA methodology

V. Todorov[1*], V. Simonacci[2] and M. Gallo[3]

[1] *United Nations Industrial Development Organization; valentin@todorov.at*
[2] *University of Naples Federico II;violetta.simonacci@unina.it*
[3] *University of Naples-L'Orientale;mgallo@unior.it*
[*] *Presenting author*

Keywords. *Three-way data; Outliers; PARAFAC; SIMCA*

In many research fields one is confronted with data sets that have repeated observations collected for the same variables on several occasions which are arranged in a data cube rather than in a matrix. Modeling such data, usually defined as three-way data, has gained importance in chemometrics and other fields and different models have been developed for this purpose [Smilde et al., 2004, Tomasi and Bro, 2006], however, not much attention was given to classification methods for multi-way arrays. In the few available papers the usual way is to unfold the multi-way data array into an ordinary matrix and then to apply traditional multivariate tools for classification. Another possibility is to use a decomposition method like Candecomp/Parafac which decomposes the trilinear structure into one score and two loading matrices with a given number of factors and then to perform Fisher's LDA or SIMCA on the score matrix [Kroonenberg, 2008].

Durante et al. [2011] were the first to propose a true multi-way approach to classification of such data by defining a procedure which they call N-SIMCA. They extend the traditional SIMCA method to three-way arrays and provide code in MATLAB for performing the computation, considering different alternatives for class allocation. Similarly as in two-way SIMCA, a separate decomposition model, say CANDECOMP/PARAFAC, will be build for each class. The number of factors can be chosen different for the different classes, using one of the known methods for this purpose or evaluating the misclassification error rate by cross validation. The classification rule is defined based on the distance $d_{i,j}$ of a sample i to a class j, which is constructed as a linear combination of two distances defined for the three-way model: orthogonal distance (OD) and score distance (SD). The orthogonal distance measures the distance of the sample to the model space in terms of squared residuals and the score distance measures how much the estimated scores of a sample deviate from the center of the scores.

The PARAFAC model and the principal component analysis (PCA) which is used for decomposition of the data matrix in two-way SIMCA, both rely on least squares techniques for conducting the computations and as such will be influenced by outliers present in the data. To cope with this problem in the two-way case Vanden Branden and Hubert [2005] proposed a robust version of SIMCA which essentially replaces the classical PCA by robust PCA (ROBPCA) and redefines the classification rule in such way that possibly outlying samples will not influence much the results. In the three-way case we replace the classical PARAFAC by a robust version which was introduced by Engelen and Hubert [2011] and adapt the classification rule to use the standardized robust distances generated by this model. The standardization is done by dividing the distances by suitable cut-off values which, apart from being used in the classification rule, also define a critical region which can be visualized in a diagnostic plot as introduced and discussed for the robust PARAFAC in Engelen and Hubert [2011].

The proposed algorithm is evaluated on simulated data from PARAFAC model in which the mode one scores are randomly generated by considering a structure with several classes. Different types of outliers are included to demonstrate the stability of the model estimation and misclassification errors are estimated and reported. The performance of the new method is also illustrated on a real-life example. Robust as well as classical N-SIMCA functions are implemented in the R package **rrcov3way** available on CRAN. All computations were carried out with this package.

References

Smilde, A., Bro, R. & Geladi, P. (2006). Multi-way analysis with applications in the chemical sciences. John Wiley & Sons, Chichester.

Kroonenberg, P. M. (2008). Applied multiway data analysis. John Wiley & Sons, Hoboken, NJ.

Tomasi, G. & Bro, R. (2006). A comparison of algorithms for fitting the PARAFAC model. *Computational Statistics & Data Analysis*, **50**, 1700–1734.

Durante, C., bro, R & Cochi, M. (2011). A classification tool for N-way array based on SIMCA methodology. *Chemometrics and Intelligent Laboratory Systems*, **106**, 73–85.

Vanden Branden, K. and Hubert, M. (2005). Robust classification in high dimensions based on the SIMCA method. *Chemometrics and Intellegent Laboratory Systems*, **79**, 10–21.

Engelen, S. & Hubert, M. (2011). Detecting outlying samples in a parallel factor analysis model. *Analytica Chemica Acta*. **705**, 155–165.

Multigroup classification by a robust trace ratio method

M. Rosário Oliveira[1*], Giulia Ferrandi[2], Igor Kravchenko[1], and Michiel E. Hochstenbach[2]

[1] *CEMAT and Department of Mathematics, Instituto Superior Técnico, University of Lisbon, Portugal; rosario.oliveira@tecnico.ulisboa.pt, igor.kravchenko@tecnico.ulisboa.pt*
[2] *Department of Mathematics and Computer Science, TU Eindhoven, The Netherlands; g.ferrandi@tue.nl, M.E.Hochstenbach@tue.nl.*
[*] *Presenting author*

Keywords. *Trace ratio method; Fisher's discriminant analysis; multigroup classification; linear dimensionality reduction; regularized MCD*

Classification is an important statistical task where an observation is assigned to one of the non-overlapping known groups, based on the statistical properties of the data characterizing these groups. Statistics, machine learning, data science, and pattern recognition are some of the areas that use this family of methods to solve practical problems. Recently, trace ratio (TR) optimization (see Ngo et al. [2012]) has gained in popularity due to its computational efficiency for high-dimensional data, as well as occasionally better classification results. Like Fisher's discriminant analysis (FDA), TR uses linear dimensionality reduction strategies for the multigroup classification problem. However, a statistical understanding is still incomplete.

In this work, we propose a robust TR method, obtained by exploiting MCD robust estimates family of the within and between covariance matrices. The method can deal with high-dimensional data since it uses regularized MCD estimates (MRCD, Boudt et al. [2020]). However, when the number of observations per class is lower than the number of variables, a high number of irrelevant variables for the classification problem has a negative impact on the performance of the estimation methods, even on the robust ones.

We compare TR and FDA on synthetic and real datasets. Synthetic scenarios consider cases where one method performs better than the others. In this case, FDA and TR are used as classifiers and are compared with two different criteria. While the first one is based on the performance of the associated classification rules, the second criterion is related to the proximity between the true solution of one method and the estimated one. Real datasets have been chosen from the UCI and KEEL platforms, and they illustrate the performance of FDA and TR as dimensionality reduction methods, used before the construction of a classifier. In many of these datasets, FDA is as good as or better than TR. Moreover, robust TR shows clear

improvements compared to classical TR in several datasets.

Acknowledgment: This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 812912. It has also received support from Fundação para a Ciência e Tecnologia, Portugal, through the project UIDB/04621/2020.

References

Boudt, K., Rousseeuw, P. J., Vanduffel, S., & Verdonck, T. (2020). The minimum regularized covariance determinant estimator. *Stat. Comput.*, **30**, 113–128.

Ngo, T. T., Bellalij, M., & Saad, Y. (2012). The trace ratio optimization problem. *SIAM Rev.*, **54**, 545–569.

Robustness Properties of Correlation Measures in Ordinal Discrete Data

M. Welz[1,2*], A. Archimbaud[1] and A. Alfons[1]

[1] *Erasmus University Rotterdam; welz@ese.eur.nl, archimbaud@ese.eur.nl, alfons@ese.eur.nl.*
[2] *Erasmus University Medical Center (Erasmus MC); m.welz@erasmusmc.nl*
[*] *Presenting author*

Keywords. *Rating-scale data; Careless responding; Association estimation; Maximum bias; Breakdown value*

Ordinal discrete data obtained from questionnaires are common in many scientific fields, such as psychology, business, and social sciences. Such data are frequently called *rating-scale* data because they are measured by having respondents choose one answer category (the rating) out of a fixed number of answer categories (the scale) in a given questionnaire item. Answer categories are typically numeric such that a larger value corresponds to stronger agreement to the item, thereby rendering rating-scale data ordinal. Rating-scale data are often a discrete measurement of a latent continuous variable (such as a personality trait), and multiple items measuring the same latent variable are called a *construct*.

Not much is known about the robustness properties of standard estimators when applied to rating-scale data. Due to the bounded and discrete nature of such data, existing results from robustness theory are often not applicable because of their implicit assumptions that outlying data points may be characterized by an arbitrarily large magnitude. In rating-scale data, outliers are item responses of individuals with *"low or little motivation to comply with questionnaire instructions, correctly interpret item content, and provide accurate responses"* [Huang et al. , 2012]. Such individuals are called careless respondents, and there is substantial empirical evidence that even a low prevalence of careless respondents of about 5–10% can jeopardize the validity of research findings in questionnaire-based studies [e.g., Arias et al. , 2020, Credé , 2010].

In this paper, we study from a theoretical perspective the robustness properties of popular location, scale, and association estimators in rating-scale data (where contamination occurs through careless responding). We focus on association estimation because correlation plays a crucial role in confirmatory factor analyses, which are commonly employed when the data measure multiple constructs. In particular,

we derive maximum bias curves and breakdown values of correlation estimators such as Pearson's. Furthermore, we study how robustness properties are affected by questionnaire design, for instance through the number of answer categories, construct size, and construct reliability.

References

Arias, V. B., Garrido, L., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods*, **52**, 2489–2505.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, **70**, 596–612.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, **27**, 99–114.

Robust Maximum Association Estimators of a General Regression Model

C. Croux[1*] and R. Crevits[2]

[1] *Edhec Business School, Lille, France*
[2] *OMP and KU Leuven, Belgium*
[*] *Presenting author*

Keywords. *Binary regression; Censored regression; Influence function; Rank correlation; Robustness.*

1 Summary

We study asymptotic properties of a class of maximum association estimators for a general regression model. We maximize an association measure R between a linear combination of covariates and a univariate respons. This association measure R can be the Pearson, Spearman, Kendall, Quadrant or another type of association. The generalized regression model includes the linear transformation model, the binary choice model, and the censored regression model. We show that Fisher consistency holds for a fairly large class of association measures and under mild distributional assumptions. We derive expressions for the influence function and compute asymptotic variances for several models and associations measures. Finite sample efficiencies are investigated by means of simulation.

2 Model

The general regression model we consider is the same as in Han [1987]. The response is a one dimensional random variable Y and the covariates are collected in a multivariate random variable $\mathbf{X}$. The general regression model states that there exist functions D and G such that

$$Y = D \circ G(\boldsymbol{\beta}_0^T \mathbf{X}, \varepsilon). \tag{1}$$

with $\boldsymbol{\beta}_0$ the true parameter. The random variable ε is an error term independent of $\mathbf{X}$. The transformation $D \circ G$ is a composite function where $D : \mathbb{R} \to \mathbb{R}$ is

nonconstant and nondecreasing and $G : \mathbb{R}^2 \to \mathbb{R}$ is strictly monotonic increasing in both arguments. Many well known models are special cases of this general regression model. We list a number of models which we will discuss in more detail.

1. Linear transformation regression model: $F(Y) = \boldsymbol{\beta}_0^T \mathbf{X} + \varepsilon$, with F monotone increasing.
2. Binary regression model: $Y = I(\boldsymbol{\beta}_0^T \mathbf{X} + \varepsilon > 0)$, with $I(\cdot)$ the indicator function. If ε follows a normal distribution this is a Probit model. If ε has a logistic distribution, it is a Logit model.
3. Censored regression model: $Y = \max(\boldsymbol{\beta}_0^T \mathbf{X} + \varepsilon, 0)$. If ε has a normal distribution, this is the Tobit model.

3 Estimator

The purpose is to estimate the parameter without knowing or specifying D and G. It is only possible to estimate $\boldsymbol{\beta}_0$ up to a constant factor, and therefore we assume $||\boldsymbol{\beta}_0|| = 1$. The estimator we study is defined as the maximizer of an association measure between the linear predictor $\boldsymbol{\beta}^T \mathbf{X}$ and the response Y. The association measure R can for example be the Pearson, the Spearman, the Kendall or the Quadrant correlation. Robustness properties of these association measures are presented in Croux & Dehon [2010].

Han [1987] showed that if R is the Kendall correlation, and under very mild distributional assumptions on $\mathbf{X}$, the estimator is consistent for $\boldsymbol{\beta}_0$ at the model distribution. Sherman [1993] showed $\sqrt{n}$-consistency if R is the Kendall or the Spearman correlation, respectively. In Alfons et al. [2017] Fisher consistency is shown for a general association measure R, but only for the linear transformation regression model. They compute the influence function and asymptotic variances assuming that $(\mathbf{X}, Y)$ has a jointly elliptical distribution, so excluding for instance the binary regression model.

References

Alfons, A., Croux, C. & Filzmoser, P. (2017). Robust maximum association estimators. *Journal of the American Statistical Association*, **112**, 436–445.

Croux, C. & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Stat Methods and Applications*, **19**, 497–515.

Han, A.K. (1987). Non-parametric analysis of a generalized regression model. *ournal of Econometrics*, **35**, 303–316.

Sherman, R.P. (1993). The Limiting Distribution of the Maximum Rank Correlation Estimator. *Econometrica*, **61**, 123–137.

Generalized Spherical Principal Component Analysis

S. Leyder[1*], T. Verdonck[1] and J. Raymaekers[2]

[1] *University of Antwerp (Middelheimlaan 1 2020 Antwerp Belgium); Sarah.Leyder@uantwerpen.be; Tim.Verdonck@uantwerpen.be*
[2] *Maastricht University; j.raymaekers@maastrichtuniversity.nl*
[*] *Presenting author*

Keywords. *Principal component analysis; Robustness; Influence functions; Efficiency; Breakdown value.*

1 Abstract

Outliers contaminating data sets are a challenge to statistical estimators. Even a small fraction of outlying observations can heavily influence most classical statistical methods. In this talk we propose *generalized spherical principal component analysis (GSPCA)*, a new robust version of principal component analysis that is based on the generalized spatial sign covariance matrix [Raymaekers & Rousseeuw, 2019]. We discuss supporting theoretical properties of the proposed method including influence functions, breakdown values and asymptotic efficiencies, and show the results of a simulation study to compare our new method to existing methods. We also propose an adjustment of the generalized spatial sign covariance matrix to achieve better Fisher consistency properties. We illustrate that generalized spherical principal component analysis, depending on a chosen radial function, has both great robustness and efficiency properties in addition to a low computational cost.

References

Raymaekers, J. & Rousseeuw, P. (2019) A generalized spatial sign covariance matrix. *Journal of Multivariate Analysis*, **171**, 94–111.

Friday 26 May 2023

Keynote - Emmanuel J. Candès (A3)

Conformal Prediction in 2023

Emmanuel J. Candès

Stanford University, USA

Conformal inference methods are becoming all the rage in academia and industry alike. In a nutshell, these methods deliver exact prediction intervals for future observations without making any distributional assumption whatsoever other than having iid, and more generally, exchangeable data. This talk will review the basic principles underlying conformal inference and survey some major contributions that have occurred in the last 2-3 years or. We will discuss enhanced conformity scores applicable to quantitative as well as categorical labels. We will also survey novel methods which deal with situations, where the distribution of observations can shift drastically — think of finance or economics where market behavior can change over time in response to new legislation or major world events, or public health where changes occur because of geography and/or policies. All along, we shall illustrate the methods with examples including the prediction of election results or COVID19-case trajectories.

I7: Robust regression (A3)

Robust estimation for functional logistic regression models based on $B-$splines

G. Boente[1*] and M. Valdora[1]

[1] *Universidad de Buenos Aires and CONICET; gboente@dm.uba.ar, mvaldora@gmail.com*
[*] *Presenting author*

Keywords. *B-splines; Functional Data Analysis; Logistic Regression Models; Robustness*

1 Abstract

Functional data analysis aims to provide tools for analysing data collected in the form of functions or curves which appear in fields such as chemometrics, image recognition and spectroscopy, among others. Functional data are intrinsically infinite–dimensional and, as mentioned for instance in Wang et al. [2016], this infinite–dimensional structure is indeed a source of information. For that reason, even when recorded at a finite grid of points, functional observations should be considered as random elements of some functional space rather than multivariate observations. In this manner, some of the theoretical and numerical challenges posed by the high dimensionality may be solved. This framework led to the extension of some classical multivariate analysis concepts, such as linear regression or logistic regression, to the context of functional data, usually through some regularization tool. An overview of different tools for analysing this type of data may be see in Ramsay & Silverman [2005], see also Horváth & Kokoszka [2012].

The functional logistic regression model assumes that (y_i, X_i), $1 \leq i \leq n$ are independent observations such that $y_i \in \{0, 1\}$ and $X_i \in L_2(\mathcal{I})$ with $\mathcal{I}$ a compact interval, that, without loss of generality, we assume to be $\mathcal{I} = [0, 1]$ and that the model relating the responses to the covariates is given by

$$\mathbb{P}(y_i = 1|X_i) = \frac{\exp\{\alpha_0 + \langle X_i, \beta_0\rangle\}}{1 + \exp\{\alpha_0 + \langle X_i, \beta_0\rangle\}},$$

where $\alpha_0 \in \mathbb{R}$, $\beta_0 \in L^2(\mathcal{I})$ and $\langle u, v\rangle = \int_{\mathcal{I}} u(t)v(t)dt$ stands for the usual inner product in $L_2(\mathcal{I})$.

As mentioned in Wang et al. [2016], one of the challenges in functional regression is the inverse nature of the problem, which causes estimation problems mainly

generated by the compactness of the covariance operator of X. The usual practices to solve this problem is regularization which can be achieved in several ways, either reducing the set of candidates for estimating β_0 to those belonging to a finite–dimensional space spanned by some bases, such as spline, Fourier, or wavelet bases or by adding a penalty term as when considering $P-$splines or smoothing splines. A special case of the former procedure corresponds to selecting the principal direction basis, that is, the one generated by the eigenfunctions of the covariance operator. However, a key difference with the method considering fixed basis such a $B-$splines is that basis functions need to be estimated rather than pre–specified. This data–dependence of the basis introduces additional technical difficulties when deriving consistency results. In this talk, we will focus on the situation where the basis is not data–dependent but known. Clearly the regularization process involves the selection of the basis dimension which should increase with the sample size at a given rate.

Taking into account the sensitivity of these estimators to atypical observations and based on the ideas given for euclidean covariates by Bianco & Yohai [1996] and Croux & Haesbroeck [2003], we define robust estimators of the intercept α_0 and the slope β_0 following a sieve approach combined with weighted $M-$estimators. Theoretical assurances regarding the consistency and convergence rates of our proposal will be presented. Besides, through the results of a numerical study, we will illustrate the sensitivity of the classical estimator and the stability of the proposed method.

References

Bianco, A. and Yohai, V. (1996). Robust estimation in the logistic regression model. *Lecture Notes in Statistics*, **109**, 17-34.

Croux, C. and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics and Data Analysis*, **44**, 273-295.

Horváth, L. & Kokoszka, P. (2012). *Inference for Functional Data with Applications*, Springer, New York.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*, Springer, Berlin.

Wang, J. L., Chiou, J. & Müller, H. G. (2016). Functional Data Analysis. *Annual Review of Statistics and Its Application*, **3**, 257-295.

Robust parameter estimation and variable selection in joint regression modelling for location, scale and skewness of the skew normal distribution

Y. Güney[1] and **O. Arslan**[1*]

[1] *Ankara University, Turkey; Yesim.Done@ankara.edu.tr, oarslan@ankara.edu.tr*

Keywords. *Lq-likelihood; Heteroscedastic regression; Penalized estimation; Joint modelling*

Abstract

In many real data examples, not only location but also scale and even skewness of the response variable may depend on some explanatory variables. In these cases, joint modeling of location, scale, and skewness may be needed to fully determine and take into account all features of the data set under consideration. The joint location, scale, and skewness model of the skew-normal distribution provide a useful tool for such responses where the normality assumption can be relaxed to allow for skewness in the data. However, in the literature, the parameter estimation methods used for these models are typically limited to classical approaches which are sensitive to outliers. However, since the parameter estimation methods used for these models in the literature are typically limited to classical approaches, the obtained estimators can be drastically affected by possible outliers in the data. Therefore, using robust estimation methods may be necessary to obtain robustness against outliers. Another challenging problem for these models is the selection of important explanatory variables for each model. In this study, we will consider the joint regression modeling of the location, scale, and skewness of the skew-normal distributed responses with two main objectives. One of these objectives is to use the maximum Lq-likelihood (MLq) estimation method to achieve robustness in estimating all the parameters of the sub-models in joint modeling of the location, scale, and skewness parameters of the skew-normal distribution. The other is to combine MLq estimation with the penalized estimation methods to carry on variable selection in all these models in order to determine the important explanatory variables for the models and therefore the model parameters corresponding to these models. An expectation-maximization (EM) algorithm is given to compute the MLq and penalized MLq estimates, and a simulation study and an application to real data are provided to demonstrate the performance of the proposed methods over the classical methods in the presence of outliers.

C9: Robust clustering II (A5)

Choice of input parameters in robust clustering based on trimming

L. A. García-Escudero[1], C. Hennig[2], A. Mayo-Iscar[1], **G. Morelli**[3*] and M. Riani[3]

[1] *Department of Statistics and Operational Research and IMUVA University of Valladolid; lagarcia@uva.es, agustinm@eio.uva.es.*
[2] *Department of Statistical Sciences "Paolo Fortunati" University of Bologna; christian.hennig@unibo.it*
[3] *Department of Economics and Management University of Parma; gianluca.morelli@unipr.it, marco.riani@unipr.it*

[*] *Presenting author*

Keywords. *Robust clustering; Trimming; Constraints.*

Abstract

The use of the so-called "classification trimmed likelihood" curves has been proposed as an useful heuristic tool to determine the number of clusters and trimming proportion in trimmed-based robust clustering methods L. García-Escudero et al. [2011]. However, these curves needs a careful visual inspection and the corresponding choice of parameter is far from being fully automated since it requires subjective decisions [L. García-Escudero et al. , 2016]. This work is intended to provide a theoretical support of their use and better understand the elements involved in their derivation. A parametric bootstrap approach will be presented so that the choice of parameter is automated. This idea provide a small list of sensible choices for the parameters where the user can hopefully find the one that better fits his/her aims.

References

García-Escudero, L., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2011). Exploring the number of groups in robust model-based clustering. *Statistics and Computing*, **21**, 585–599.

García-Escudero, L., Gordaliza, A., Matrán, C., Mayo-Iscar, A. & Hennig, C. (2016). Handbook of Cluster Analysis: Robustness and outliers. Serie Chapman & Hall/CRC Handbooks of Modern Statistical Methods

A robust model-based clustering based on the geometric median and the Median Covariation Matrix

Antoine Godichon-Baggioni[1*] and Stéphane Robin[1]

[1] *Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, antoine.godichon_baggioni@upmc.fr, stephane.robin@sorbonne-universite.fr*

Keywords. *EM algorithm; Geometric median; Median Covariation Matrix; Mixture models; Robust statistics*

Grouping observations into homogeneous groups (or "clusters") is one of the most typical tasks in statistical data analysis. Among the many methods that have been proposed over the years, model-based clustering is one of the most popular [McLahan and Peel, 2000]. Model-based clustering relies on the assumption that the observed data come from a mixture model, which means that the observations can be divided into a finite number of clusters characterized by a specific distribution.

One reason for the popularity of model-based clustering is that the distributions of the clusters are usually chosen with a parametric class (e.g., a multivariate Gaussian), which makes the interpretation of the results particularly easy. Another reason for this popularity is that the maximum likelihood estimates of the parameters can be obtained via the well-known EM algorithm (Dempster et al. [1977]).

Nevertheless, one of the weaknesses of model-based clustering methods is their sensitivity to misspecification of emission distributions or to the presence of (possibly numerous) outliers. In both cases, this results in a high proportion of misclassified observations or a poor estimate of the number of clusters García-Escudero et al. [2010].

Here we focus on the robustness of model-based clustering methods to the presence of outliers, meaning that we make no assumptions about how outliers deviate from prescribed emission distributions. To this end, we adopt a fully parametric model-based clustering framework, but modify the EM algorithm (specifically, the M-step) to ensure robustness. Our method is valid for any symmetric emission distribution and uses the estimation of the median vector and median covariation matrix instead of the mean vector and covariance matrix [Vardi and Zhang, 2000, Cardot et al., 2013, Cardot and Godichon-Baggioni, 2015]. In particular, it was proven (see Kraus and Panaretos [2012]) that for symmetric distributions, the MCM and the usual covariance have the same eigenvectors. Nevertheless, although recursive estimation

of the MCM has been studied in Cardot and Godichon-Baggioni [2015], no method to construct the covariance from the MCM has been proposed. We first propose methods to obtain robust estimates of covariance before applying it to model-based robust clustering. The detailed procedure is described in Godichon-Baggioni and Robin [2022] and all the methods are available in the R package RGMM on CRAN.

References

Cardot, H., Cénac, P., and Zitt, P.-A. (2013). Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19(1):18–43.

Cardot, H. and Godichon-Baggioni, A. (2015). Fast estimation of the median covariation matrix with application to online robust principal components analysis. *TEST*, pages 1–20.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38.

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification*, 4(2):89–109.

Antoine Godichon-Baggioni and Stéphane Robin (2022). A robust model-based clustering based on the geometric median and the median covariation matrix. *arXiv preprint arXiv:2211.08131*.

Kraus, D. and Panaretos, V. M. (2012). Dispersion operators and resistant second-order functional data analysis. *Biometrika*, 99:813–832.

McLahan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley.

Vardi, Y. and Zhang, C.-H. (2000). The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA*, 97(4):1423–1426.

Semi-continuous time series for sparse data with volatility clustering

Š. Hudecová[1] and **M. Pešta**[1*]

[1] *Charles University, Prague, Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics; sarka.hudecova@mff.cuni.cz, michal.pesta@mff.cuni.cz.*
[*] *Presenting author*

Keywords. *Non-negative time series; Sparse data; Hurdle GARCH model; Multiplicative error models; Robust quasi-likelihood.*

Abstract

Time series containing non-negligible portion of possibly dependent zeros, whereas the remaining observations are positive, are considered. They are regarded as GARCH processes consisting of non-negative values. Our first aim lies in estimation of the omnibus model parameters taking into account the semi-continuous distribution. The hurdle distribution together with dependent zeros cause that the classical GARCH estimation techniques fail. Two different robust quasi-likelihood approaches are employed. Both estimators are proved to be strongly consistent and asymptotically normal. The second goal consists in the proposed predictions with bootstrap add-ons. The considered class of models can be reformulated as multiplicative error models. The empirical properties are illustrated in a simulation study, which demonstrates computational efficiency of the employed methods. The developed techniques are presented through an actuarial problem concerning sparse insurance claims.

I8: Robust multivariate statistics II (A3)

Robust second-order stationary spatial blind source separation

Mika Sipilä[1], Christoph Muehlmann[2], Klaus Nordhausen[1] and **Sara Taskinen**[1*]

[1] *Department of Mathematics and Statistics, University of Jyväskylä, Finland; mika.e.sipila@jyu.fi, klaus.k.nordhausen@jyu.fi, sara.l.taskinen@jyu.fi.*
[2] *Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Austria; christoph.muehlmann@tuwien.ac.at*
[*] *Presenting author*

Keywords. *Affine equivariance; Bias; Multivariate spatial data; Scatter matrix; Spatial signs.*

Abstract

Assume a spatial blind source separation model in which the observed multivariate spatial data is assumed to be a linear mixture of latent stationary spatially uncorrelated random fields. The goal is then to recover an unknown mixing procedure as well as latent uncorrelated random fields. Recently, spatial blind source separation methods that are based on simultaneous diagonalization of two or more scatter matrices were proposed. In case of uncontaminated data such methods are capable of solving the blind source separation problem, but in presence of outlying observations the methods perform poorly. We propose a robust blind source separation method which uses robust global and local scatter matrices based on generalized spatial signs in simultaneous diagonalization. Simulation studies are used to illustrate robustness and efficiency properties of proposed methods in various scenarios.

L_p inference for multivariate location based on data-based simplices

A. Dürre[1] and **D. Paindaveine**[2*]

[1] *Leiden University, Mathematical Institute; a.m.durre@math.leidenuniv.nl*
[2] *Université libre de Bruxelles, ECARES and Department of Mathematics; Davy.Paindaveine@ulb.be*
[*] *Presenting author*

Keywords. *Affine equivariance/affine invariance; L_p loss functions; Oja median; Random simplices; Spatial median.*

The fundamental problem of estimating the location of a d-variate probability measure under an L_p loss function is considered. The naive estimator, that minimizes the usual empirical L_p risk, has a known asymptotic behaviour but suffers from several deficiencies for $p \neq 2$, the most important one being the lack of equivariance under general affine transformations. We introduce a collection of L_p location estimators that minimize the size of suitable ℓ-dimensional data-based simplices. For $\ell = 1$, these estimators reduce to the naive ones, whereas, for $\ell = d$, they are equivariant under affine transformations and generalize the famous [Oja, 1983] median. The proposed class also contains the spatial median. Under very mild assumptions, we derive an explicit Bahadur representation result for each estimator in the class and establish asymptotic normality. Under a centro-symmetry assumption, we also introduce companion tests for the problem of testing the null hypothesis that the location μ of the underlying probability measure coincides with a given location μ_0. We compute asymptotic powers of these tests under contiguous local alternatives, which reveals that asymptotic relative efficiencies with respect to traditional parametric Gaussian procedures for hypothesis testing coincide with those obtained for point estimation. Monte Carlo exercises confirm our asymptotic results.

References

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics and Probability Letters*, **1**, 327–332.

Dürre, A. & Paindaveine, D. (2022). Affine-equivariant inference for multivariate location under L_p loss functions. *Annals of Statistics*, **50**, 2616–2640.

Dürre, A. & Paindaveine, D. (2023). Multivariate L_p location testing: Wald tests and Lagrange multipler tests based on simplices. Submitted.

Robust Elastic Net estimators

Gabriela Cohen-Freue[1], David Kepplinger[2], **Matias Salibian-Barrera**[1*], and Ezequiel Smucler[3]

[1] *Department of Statistics, The University of British Columbia; {gcohen, matias}@stat.ubc.ca*
[2] *Department of Statistics, George Mason University, dkepplin@gmu.edu*
[3] *e.smucler@aristas.com.ar*
[*] *Presenting author*

Keywords. *Robust regression; High-dimensional regression; Regularized estimation*

1 Robust Elastic Net Estimators

Regularized linear regression estimators are routinely used in situations where some variable selection or shrinkage may be advantageous due to multicollinearity in the training data, or when there are more explanatory variables than observed items. Although both of these problems can be addressed by using penalized regresssion estimators, the underlying goals of the analyses may be different. For example, in the first case we might be content with finding a subset of explanatory variables that is identifiable and produces stable regression estimates and good future predictions. Alternatively, we may be interested in identifying **all** explanatory variables that are correlated with the response, and not just "a" subset of them that produces good predictions, even when the number of observations is smaller than the number of covariates. In the last 25 years, these problems have driven the development of a very rich and extensive literature on different variants of regularized estimators. For example, while the LASSO estimator [Tibshirani , 1996] typically works well for the first type of problems, it is well known to not be suitable for the second situation. The Elastic Net [Zou & Hastie , 2005] and other modifications of the LASSO provide alternatives that are designed to work well in a variety of scenarios, and which can be tuned to achieve the different goals mentioned above. It is well known that in general these estimators can be seriously affected by a small proportion of atypical observations in the training set. Many robust regularized linear regression estimators have been proposed in the literature in recent years, e.g. Alfons et al [2013], Loh and Wainwright [2015], Loh [2017], Smucler and Yohai [2017], Avella-Medina and Ronchetti [2018], Cohen Freue et al [2019] and Kepplinger & Cohen Freue [2023].

In this talk I will discuss some of these proposals with a focus on effective variable selection.

References

Alfons, A., Croux, C. and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, **7**, 226-248.

Avella-Medina, M. and Ronchetti, E. (2018). Robust and consistent variable selection in high-dimensional generalized linear models. *Biometrika*, **105**(1), 31-44.

Cohen Freue, G.V., Kepplinger, D., Salibian Barrera, M. and Smucler, E. (2019). Robust elastic net estimators for variable selection and identification of proteomic biomarkers. *The Annals of Applied Statistics*, **13**(4), 2065-2090.

Kepplinger, D. and Cohen Freue, G.V. (2023). Robust Prediction and Protein Selection with Adaptive PENSE. In: Burger, T. (eds) *Statistical Analysis of Proteomic Data. Methods in Molecular Biology*, vol 2426. Humana, New York, NY. DOI: 10.1007/978-1-0716-1967-4_14.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics*, **45**, 866-896.

Loh, P.-L. and Wainwright, M.J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, **16**, 559-616.

Smucler, E. and Yohai, V. J. (2017). Robust and sparse estimators for linear regression models. *Computational Statistisics and Data Analysis*, **111**, 116-130.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, **67**, 301-320.

C10: Cellwise and rowwise outliers (A5)

Robust PARAFAC for cellwise and rowwise outliers

M. Hirari[1], M. Hubert[1]

[1] *Section of Statistics and Data Science, Department of Mathematics, KU Leuven, Belgium; Mehdi.Hirari@kuleuven.be, Mia.Hubert@kuleuven.be.*

Keywords. *Multi-way; PARAFAC; Cellwise Outliers*

Multi-way data extend two-way matrices to a higher dimensional tensor. In many fields, it is relevant to pursue the analysis of such data by keeping it in its initial form without unfolding it into a matrix. Often multi-way data are explored by means of dimensional reduction techniques. Here, we study the Parallel factor analysis (PARAFAC) model, which expresses the multi-way data in a more compact way through a small set of loading matrices and scores.

The most common algorithm to fit this model is by means of an Alternating Least Squares (ALS) algorithm. However, it is well known that ALS is not robust to outliers. Robust alternatives which are resistant towards rowwise outliers have been proposed in the past [Engelen and Hubert, 2011, Hubert et al., 2012], the latter approach being able to cope with missing values as well.

These methods are however not resistant towards cellwise outliers that might contaminate all observations. A new algorithm is proposed that generalizes the MacroPCA method for two-way data [Hubert et al., 2019] towards multi-way data. We show with simulations and the analysis of real data that this MacroPARAFAC method is robust to rowwise and cellwise outliers while also being able to handle missing elements.

References

Engelen, S. & Hubert, M. (2011). Detecting outlying samples in a parallel factor analysis model. *Analytica Chimica Acta*, **75**, 155–165.

Hubert, M., Rousseeuw, P.J. & Van den Bossche, W. (2019). MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers. *Technometrics*, **61**, 459–473.

Hubert, M., Van Kerckhoven, J. & Verdonck, T. (2012). Robust PARAFAC for incomplete data. *Journal of Chemometrics*, **26**, 290–298.

Robust variable selection under cellwise contamination

Peng Su[1*], Samuel Muller[2] and Garth Tarr[1]

[1] *School of Mathematics and Statistics, The University of Sydney, NSW 2006, Australia; peng.su@sydney.edu.au, garth.tarr@sydney.edu.au.*
[2] *School of Mathematical and Physical Sciences, Macquarie University; samuel.muller@mq.edu.au*
[*] *Presenting author*

Keywords. *robust variable selection; robust correlation; Gaussrank correlation; Adaptive Lasso.*

1 Introduction

Consider a linear regression model where n observations are modelled through,

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where y_i is the response, $\mathbf{x}_i$ is a p-vector of predictors, $\boldsymbol{\beta}$ is a p-vector of regression parameters and ε_i is an independent random error with variance σ^2. The propagation of cellwise outliers under the independent contamination model is dramatic [Alqallaf et al., 2009]. For a cell contamination rate e, the expected row contamination proportion is $1-(1-e)^p$, which on average rapidly exceeds 50% with increasing dimension p. Traditional robust methods may fail when applied to datasets under cellwise contamination. We propose to use the Gaussian rank correlation [Boudt et al., 2012] to obtain an initial empirical correlation/covariance matrix among the response and potential active predictors. We re-parameterise the design matrix and the response vector of the original linear regression model so that we are able to take advantage of these robustly estimated components before applying the adaptive Lasso to obtain variable selection results.

2 Methodology

Given a cellwise robust positive definite covariance matrix estimate $\hat{\boldsymbol{\Sigma}}$ (from the Gaussian rank correlation [see e.g., Boudt et al., 2012, Amengual., 2022] and robust

scale estimates), we denote $\hat{\Sigma}_{yy}, \hat{\boldsymbol{\Sigma}}_{xy}, \hat{\boldsymbol{\Sigma}}_{xx}$ to be the components of the empirical covariance matrix among the predictors and the response,

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\Sigma}_{yy} & \hat{\boldsymbol{\Sigma}}_{xy} \\ \hat{\boldsymbol{\Sigma}}_{xy}^{\top} & \hat{\boldsymbol{\Sigma}}_{xx} \end{bmatrix}. \tag{2}$$

Then we define $[\hat{\mathbf{z}}, \hat{\mathbf{W}}] = \hat{\boldsymbol{\Sigma}}^{1/2}$, where $\hat{\mathbf{z}}$ denotes the first column of $\hat{\boldsymbol{\Sigma}}^{1/2}$ and $\hat{\mathbf{W}}$ denotes the remaining columns. For a linear regression model, the quadratic loss function can be expressed as

$$\begin{aligned} &\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^{\top}\boldsymbol{\beta})^2 \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \hat{\Sigma}_{yy} + \boldsymbol{\beta}^{\top}\hat{\boldsymbol{\Sigma}}_{xx}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\top}\hat{\boldsymbol{\Sigma}}_{xy} \right\} \qquad (3) \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\hat{\mathbf{z}} - \hat{\mathbf{W}}\boldsymbol{\beta}\|_2^2. \qquad (4) \end{aligned}$$

To obtain variable selection results, we combine the objective loss (4) with the adaptive Lasso [Zou., 2006]

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \|\hat{\mathbf{z}} - \hat{\mathbf{W}}\boldsymbol{\beta}\|_2^2 + \lambda \|\hat{\boldsymbol{\omega}}\boldsymbol{\beta}\|_1 \right\}, \tag{5}$$

where λ is the tuning parameter, $\hat{\boldsymbol{\omega}} = \mathrm{Diag}(1/\tilde{\boldsymbol{\beta}})$ and $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_{xx}^{-1}\hat{\boldsymbol{\Sigma}}_{xy}$ is an initial non-regularized robust consistent estimate of $\boldsymbol{\beta}$.

This procedure is robust to cellwise outliers in low and high dimensional settings. Empirical results show good selection results and competent prediction results compared to other robust techniques such as [Bottmer et al., 2022], particularly in a challenging environment when contamination rates are high but the magnitude of outliers is moderate.

References

Alqallaf, F., Van Aelst, S., Yohai, V. J., & Zamar, R. H. (2009). Propagation of outliers in multivariate data. *Annals of Statistics*, 311-331.

Amengual, D., Sentana, E., & Tian, Z. (2022). Gaussian rank correlation and regression. *Essays in Honor of M. Hashem Pesaran: Panel Modeling, Micro Applications, and Econometric Methodology*, 43, 269-306.

Bottmer, L., Croux, C., & Wilms, I. (2022). Sparse regression for large data sets with outliers. *European Journal of Operational Research*, 297(2), 782-794.

Boudt, K., Cornelissen, J., & Croux, C. (2012). The Gaussian rank correlation estimator: robustness properties. *Statistics and Computing*, 22, 471-483.

Loh, P. L., & Wainwright, M. J. (2011). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Advances in neural information processing systems*, 24.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476), 1418-1429.

Minimum Regularized Covariance Trace Estimator and Outlier Detection for Functional Data

J. Oguamalam[1], P. Filzmoser[1] and **U. Radojicic**[1*]

[1] *TU Wien, Wiedner Hauptstraße 8-10, Vienna, Austria; jeremy.oguamalam@tuwien.ac.at*

[2] *TU Wien, Wiedner Hauptstraße 8-10, Vienna, Austria peter.filzmoser@tuwien.ac.at*

[3] *TU Wien, Wiedner Hauptstraße 8-10, Vienna, Austria una.radojicic@tuwien.ac.at.*

[*] *Presenting author*

Keywords. *Functional outlier detection; Robust covariance operator; Regularization; RKHS*

Abstract

Outlier detection is a big part of functional data analysis as it is crucial to find atypical curves to prevent bias in subsequent analysis. This paper proposes a new method for finding irregular functional data, the minimum regularized covariance trace estimator (MRCT). It searches for a subset of the data for which the standardization results in the covariance with minimal trace. This framework includes inverting the singular covariance operator by the Tikhonov regularization. The proposed iterative algorithm consists of concentration steps in which the dissimilarity is based on a functional Mahalanobis distance defined on the reproducing kernel Hilbert space. Furthermore, the selection of the Tikhonov regularization parameter is automated. The method converges fast in practice and performs favorably compared to other functional outlier detection methods.

References

Berrendero, J. R., Bueno-Larraz, B., and Cuevas, A.: On Mahalanobis distance in functional settings. *Journal of Machine Learning Research* **21.**, 9 (2020), 1–33.

Mas, A.: Weak convergence in the functional autoregressive model. *Journal of Multivariate Analysis* **98.**, 6 (2007), 1231–1261.

Rousseeuw, P. J. and Driessen, K.: A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41.**, 3 (1999), 212–223.

Changepoint detection in structured time-dependent functional profiles

M. Maciak[1*] and S. Vitali[2]

[1] *Department of Probability and Mathematical Statistics, Charles University, Sokolovská 83, Prague, 187 75, Czech Republic; maciak@karlin.mff.cuni.cz.*
[2] *Department of Economics, University of Bergamo, Via dei Caniana 2, 24127, Bergamo, Italy; ; sebastiano.vitali@unibg.it*
[*] *Presenting author*

Keywords. *Robust estimation; Changepoint; Panel data; Implied volatility*

1 Motivation

A proper analysis of financial markets always relies on the ability to detect and estimate all kinds of sudden shocks—*changepoints*—which randomly and repeatedly occur over time. On the other hand, the changepoint detection is, in general, well known as a complex, sophisticated, and rather difficult problem in statistics. Moreover, the overall complexity of the problem even increases if the underlying shocks are assumed to be of different nature and some basic distinction between the corresponding types of the occurring changepoints is of the main interest.

2 General Framework

In Maciak & Vitali [2023], we introduce a unique methodological approach to recognize and detect a specific type of stochastically relevant (significant) changepoints within a sequence of time-dependent functional profiles—the options' implied volatility (IV) smiles in particular. The main focus is on changes caused by various exogenous effects (induced not by the market itself but rather by of some human-made interactions). A standard implied volatility tool (commonly used for the option market analysis by practitioners) is shown to be insufficient for a proper detection and analysis of this type of the market risk. This is mainly because the exogenous changes are typically dominated by endogenous effects coming from a specific trading mechanism or a natural market dynamics (such as the so-called "getting-close-to-maturity" effect for example).

A unique methodological approach based on "artificial options" that always have a constant (over time) maturity is proposed. The key principle is to use interpolated volatilities that are shown to be able to effectively eliminate instabilities due to the natural market dynamics while the changes caused by the exogenous causes stay preserved. In addition, a robust semi-parametric estimation of the time-dependent IV smile profiles can be employed to postulate an underlying model that fully complies with the financial theory on arbitrage-free markets (see, for instance, Maciak & Vitali [2023] or Maciak et al. [2020]).

Formal statistical tests for detecting significant changepoints are proposed under different theoretical assumptions and various practical scenarios. The overall performance of the proposed tests is investigated from both—the theoretical as well as the empirical perspective. Finally, important applicational issues are addressed and real data examples are given for some illustration.

References

Maciak, M. (2019). Quantile lasso with changepoints in panel data models applied to option pricing. *Econometrics and Statistics*, 20(11/2021):166–175.

Maciak, M., Pešta, M., & Vitali, S. (2020). Implied volatility surface estimation via quantile regularization. *In Maciak et al. (Eds), Analytical Methods in Statistics, Springer Proceedings in Mathematics & Statistics 329*, Springer Nature Switzerland AG, 73–87.

Maciak, M. & Vitali, S. (2023). Exogenous market changes analysis using artificial options volatility. *Computational Economics*, (submitted after revision).

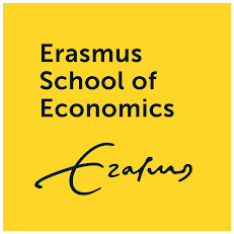
List of sponsors

Toulouse School of Economics
Toulouse, France

AIRBUS

AIRBUS - GOLD sponsor
Advanced Analytics & Artificial Intelligence
Toulouse, France

Insee - Bronze sponsor
Institut national de la Statistique
et des Études Économiques
France



Erasmus School of Economics
Rotterdam, Netherlands

Université Toulouse Capitole
Toulouse, France

SFdS

Société Française de Statistique
French Statistical Society

Toulouse

Meetings Toulouse
Toulouse, France

Author Index