Book of Abstracts

International Conference on Robust Statistics

ICORS 2021

20-24 September 2021

TU Wien, Austria

http://www.icors.eu/

Sponsors:

International Association for Statistical Computing (IASC)
TU Wien
Bernoulli Society

Conference Venue

The conference will be held in a hybrid form. It takes place at TU Wien, at the Campus Gusshaus, Gusshausstr. 27-29, 1040 Vienna. The lecture hall is called “Kontaktraum”, and it is in the 6th floor of this building.

Registration, all coffee breaks, etc. will be in front of this lecture hall.

Important information for participants **entering a TU Wien building** (and restaurants/hotels): You must present a proof of a low epidemiological risk to the security service.

- Vaccinated: Proof of vaccination (complete immunisation) with an EU-approved vaccine against COVID-19 is required.
- Tested: Vienna has different testing offers (entry tests: PCR self-tests, gargle boxes, test lanes, pharmacies). Actually, tests are valid for 48 hours.
- Recovered: Proof is the medical confirmation of a SARS-CoV-2 infection recovered from in the last 180 days (confirmed by molecular biology), proof of neutralising antibodies (not older than 90 days) or your certificate of segregation.

The **registration desk** will be open on September 21, from 8:00 to 09:20. It will be in front of the “Kontaktraum”. Late registration will be possible also on the other conference days.

On-site participants for the R Workshop can register on September 20 from 12:00-13:00 in front of the “Kontaktraum”.

The on-site telephone number of the conference administrations: **+43 (1)58801 10560**.

For the **virtual participation** we will use *zoom*. Virtual participants will receive a zoom link via email a few days before the conference start. There will be separate zoom links for the workshop on September 20, and the ICORS conference from Sep. 21-24.

Guidelines for Speakers:

Talks in presence: Please, bring your presentation as pdf-file on a USB memory-stick and contact the person responsible for the conference computer *before the session starts.*

Virtual presenters: Please, test the zoom environment beforehand. In particular, test screen sharing and your audio and video devices.

Scientific Program Committee

Claudio Agostinelli, University of Trento, Italy
Ana Bianco, Universidad de Buenos Aires and CONICET, Argentina
Shoja Chenouri, University of Waterloo, Canada
Peter Filzmoser, TU Wien, Austria
Xuming He, University of Michigan, USA
Klaus Nordhausen, University of Jyväskylä, Finland
Peter Rousseeuw, KU Leuven, Belgium
Anne Ruiz-Gazen, Toulouse School of Economics, France
Stefan Van Aelst, KU Leuven, Belgium
Maria-Pia Victoria-Feser, University of Geneva, Switzerland

Local Organizing Committee

Peter Filzmoser, TU Wien, Austria
Daniela Vater, TU Wien, Austria
Helmut Schwarz, TU Wien, Austria
Dominika Miksova, TU Wien, Austria
Christopher Rieser, TU Wien, Austria
Alexandra Posekany, TU Wien, Austria

Workshop on "Robustness & R"

Monday, 20 September 2021

The workshop will take place at TU Wien, Gusshausstr. 27-29, 6^{th} floor ("Kontaktraum")
Virtual participants will receive a zoom link via email.

13:00 – 14:15 Valentin Todorov

Robust Principal Component and Discriminant Analysis

PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) are two key techniques for dimension reduction. The former is an unsupervised algorithm which tries to find linear combinations (components) that capture the maximal variance within the data while the latter is a supervised method which preserves as much of the class discrimination information as possible. Both methods are based on a kind of sample covariance matrix of the data which makes them very sensitive to the presence of even a few outliers in the data. To cope with this problem robust methods were proposed, the most straightforward one being to replace the sample covariance matrix by a robust version of it. While appealing by its simplicity, this approach has disadvantages in case of higher dimensions and methods based on projection pursuit, sparsity and regularization were proposed. We will start by a simple example, will demonstrate the issues of outliers and contamination in the data and will walk through the most popular robust methods for PCA and LDA as implemented in the R package 'rrcov'.

14:30 – 15:45 Alexander Dürre

Robust Time Series Analysis

The lecture consists of three parts, namely exploratory analysis, modelling, and predicting. In the first section we start by presenting robust tests for stationarity. Subsequently we look at methods describing the dependence structure of the time series, mainly robust estimators of the autocorrelation function. We conclude the first part by talking about robust spectral analysis. The second section mainly covers linear time series models. We present methods to robustly fit ARMA models, evaluate the goodness of fit and detect outliers. The last section covers parametric and non-parametric forecasting.

Throughout the talk we will visualize the effects of outliers on non-robust and robust methods and apply the presented methods to real data sets.

Coffee Break: 15:45 – 16:15

16:15 – 17:30 Andreas Alfons

Sparse Robust Regression and Model Selection

For regression analysis in high-dimensional settings, variable selection is a crucial task to improve prediction performance by variance reduction, to increase interpretability of the resulting models due to the smaller number of variables, and to avoid computational issues with standard methods due to the rank deficiency of the design matrix. Common strategies are to either obtain a sequence of important variables and fit a series of regression models, or to apply regularized regression estimators that simultaneously perform variable selection and coefficient estimation. However, when outliers are present in the data, robust methods are necessary to prevent unreliable results. This tutorial provides an overview of robust methods for regression and variable selection for high-dimensional data, such as robust least angle regression and robust lasso and elastic net. Moreover, the practical application of these methods is illustrated using R packages such as robustHD and pense.

Conference Schedule ICORS, September 21-24, 2021

Time		Sep 21 Tuesday
8:00	9:20	Registration
9:20	9:30	Opening
9:30	10:50	CP1
10:50	11:20	Coffee Break
11:20	12:20	CP2
12:20	14:00	Lunch
14:00	15:20	CP3
15:20	15:50	Coffee Break
15:50	17:20	IS1
17:25	18:25	Keynote Fan

Time		Sep 22 Wednesday
14:00	15:20	CP4
15:20	15:50	Coffee Break
15:50	16:30	Short Papers
16:35	18:05	IS2

Time		Sep 23 Thursday
9:30	10:50	CP5
10:50	11:20	Coffee Break
11:20	12:20	CP6
12:20	14:00	Lunch
14:00	15:20	CP7
15:20	15:50	Coffee Break
15:50	17:20	IS3
17:25	18:25	Keynote Rousseeuw

Time		Sep 24 Friday
9:30	10:20	Invited Biggio
10:20	10:50	Coffee Break
10:50	11:40	Invited Suykens

Program Overview ICORS 2021

All presentations take place at TU Wien, Gusshausstr. 27-29, 6th floor ("Kontaktraum")
Virtual participants will receive a zoom link via email.

Tuesday, 21 September 2021

Opening 09:20 – 09:30

General robustness 09:30 – 10:50

Contributed Paper Session: CP1 Tuesday 21
Chair: Jana Jurečková

Christian Hennig
Is there a role for model assumption testing in applied statistics? (see p. 36)
09:30 – 09:50

Tino Werner
Global quantitative robustness of instance ranking problems (see p. 63)
09:50 – 10:10

Pavlina Kalcheva Jordanova and Milan Stehlik
Tails and probabilities for p-outside values (see p. 39)
10:10 – 10:30

Vanessa Berenguer-Rico, Søren Johansen, and Bent Nielsen
A model where the Least Trimmed Squares estimator is maximum likelihood (see p. 19)
10:30 – 10:50

Coffee Break: 10:50 – 11:20

Tests and depth 11:20 – 12:20

Contributed Paper Session: CP2 Tuesday 21
Chair: Luis Angel Garcia-Escudero

Dennis Malcherczyk, Kevin Leckey, and Christine H. Müller
The K-sign depth and generalizations (see p. 47)
11:20 – 11:40

Christine H. Müller
K-sign depth tests: Some properties and some open problems (see p. 50)
11:40 – 12:00

Jana Jurečková, Yesim Güney, Martin Schindler, Jan Picek, Olcay Arslan, and Yetkin Tuaç

Rank tests in linear model with autoregressive errors (see p. 41)
12:00 – 12:20

Lunch Break: 12:20 – 14:00

Linear models and extensions 14:00 – 15:20

Contributed Paper Session: CP3 Tuesday 21
Chair: Christine Müller

Giovanni Saraceno, Abhik Ghosh, Ayanendranath Basu, and Claudio Agostinelli
Robust estimation under Linear Mixed Models: a Minimum Density Power Divergence approach (see p. 59)
14:00 – 14:20

Kamila Fačevicová, Christoph Muehlmann, Klaus Nordhausen, Martin Žídek, and Ondřej Bábek
Use of a robust blind source separation approach for XRF core scanning of soft sediments (see p. 30)
14:20 – 14:40

Graciela Boente and Alejandra Mercedes Martinez
A B-spline robust approach for partially linear additive models (see p. 23)
14:40 – 15:00

William H. Aeberhard, Eva Cantoni, Giampiero Marra, and Rosalba Radice
Robust fitting for Generalized Additive Models for location, scale and shape (see p. 16)
15:00 – 15:20

Coffee Break: 15:20 – 15:50

Robustness for functional data 15:50 – 17:20

Invited Paper Session: IS1 Tuesday 21
Chair: Stefan Van Aelst

Xinyi Li, Lily Wang, and Huixia Judy Wang
Sparse learning and structure identification for ultra-high-dimensional image-on-scalar regression (see p. 46)
15:50 – 16:20

Graciela Boente and Nadia Kudraszow
A robust smoothed approach to functional canonical correlation analysis (see p. 21)
16:20 – 16:50

Xiaomeng Ju and Matias Salibian Barrera
Robust Boosting for functional regression (see p. 40)
16:50 – 17:20

Keynote presentation 17:25 – 18.25

Chair: Xuming He Tuesday 21

Jianqing Fan
High-dimensional robust inference: Farming significant and important variables (see p. 31)

Wednesday, 22 September 2021

Functional and high-dimensional data — 14:00 – 15:20

Contributed Paper Session: CP4 — Wednesday 22
Chair: Karel Hron

Ioannis Kalogridis and Stefan Van Aelst
Robust optimal estimation of location from discretely sampled functional data (see p. 42)
14:00 – 14:20

Luis A. Garcia-Escudero, Diego Rivera-Garcia, Agustin Mayo-Iscar, and Joaquin Ortega
Cluster Analysis with cellwise outliers with applications to robust functional clustering (see p. 32)
14:20 – 14:40

Gianna Serafina Monti and Peter Filzmoser
A robust approach to classification and regression tasks for microbiome data (see p. 48)
14:40 – 15:00

Abhik Ghosh, Maria Jaenada, and Leandro Pardo
Robust adaptive variable selection in ultra-high dimensional linear regression models (see p. 33)
15:00 – 15:20

Coffee Break: 15:20 – 15:50

Short Papers — 15:50 – 16:30

Short oral presentations: — Wednesday 22
Chair: Peter Filzmoser

Barbara Brune, Irene Ortner, and Peter Filzmoser
A comparison study of robust Mixed Effects Models for analyzing degradation of photovoltaic modules (see p. 25)
15:50 – 16:00

Lukas Neubauer and Peter Filzmoser
Robust functional principal component regression: a comparison (see p. 52)
16:00 – 16:10

Ebru Ergun and Onder Aydemir
A robust firefly algorithm based feature selection method for EEG signal classification (see p. 28)
16:10 – 16:20

Virgilio Pérez, Jose M. Pavia, and Cristina Aybar
Over time robust estimation of subjective latent variables from cross-section repeated surveys under measurement error (see p. 55)
16:20 – 16:30

Time series analysis 16:35 – 18:05

Invited Paper Session: IS2 Wednesday 22
Chair: Alexander Dürre

Sara Taskinen, Klaus Nordhausen, and David E. Tyler
Blind source separation based on M autocovariance matrices (see p. 61)
16:35 – 17:05

Victor Jaime Yohai and Daniel Peña
Forecasting multiple time series with robust one-sided dynamic principal components (see p. 53)
17:05 – 17:35

Marc Hallin, Davide La Vecchia, and Hang Liu
Inference for multivariate time series models: a measure transportation approach (see p. 34)
17:35 – 18:05

Thursday, 23 September 2021

Outliers and anomaly detection 09:30 – 10:50

Contributed Paper Session: CP5 Thursday 23
Chair: Ioannis Kalogridis

Dana Rahbani, Andreas Morel-Forster, and Thomas Vetter
A robust acquisition function for sequential Gaussian Process inference (see p. 56)
09:30 – 09:50

Max Welz and Andreas Alfons
Outlier detection in rating-scale data via autoencoders (see p. 62)
09:50 – 10:10

Luca Insolia, Francesca Chiaromonte, Runze Li, and Marco Riani
Doubly robust feature selection with mean and variance outlier detection and oracle properties (see p. 38)
10:10 – 10:30

Jakob Raymaekers and Peter J. Rousseeuw
Transforming variables to central normality (see p. 57)
10:30 – 10:50

Coffee Break: 10:50 – 11:20

Time series 11:20 – 12:20

Contributed Paper Session: CP6 Thursday 23
Chair: Bernhard Spangl

Alexey Kharin, Ton That Tu, Hongqiang Zhao, and Yu Li
Robustness in sequential decision making on parameters of stochastic data flows (see p. 43)
11:20 – 11:40

Yuriy Kharin
Discrete-valued time series: parsimonious models and statistical analysis (see p. 44)
11:40 – 12:00

Ieva Axt, Alexander Dürre, and Roland Fried
Robust scale estimation under shifts in the mean (see p. 18)
12:00 – 12:20

Lunch Break: 12:20 – 14:00

Linear regression models 14:00 – 15:20

Contributed Paper Session: CP7 Thursday 23
Chair: Agostin Mayo-Iscar

Viktorie Nesrstová, Karel Hron, Josep Antonio Martin-Fernández, Peter Filzmoser, Javier Palarea-Albaladejo, and Ines Wilms
Variable selection in compositional data using balance coordinates based on robust PLS (see p. 51)
14:00 – 14:20

Olcay Arslan and Senay Ozdemir
Robust penalized empirical likelihood estimation method for linear regression (see p. 17)
14:20 – 14:40

Hira L. Koul and Pei Geng
Weighted empirical minimum distance estimators in errors in variables linear regression models (see p. 45)
14:40 – 15:00

Sukru Acitas, Peter Filzmoser, Birdal Senoglu, and Gamze Guven
A new robust Liu-type estimator for regression based on RAMML estimators (see p. 15)
15:00 – 15:20

Coffee Break: 15:20 – 15:50

Cellwise robustness and sparsity 15:50 – 17:20

Invited Paper Session: IS3 Thursday 23
Chair: Christophe Croux

Karel Hron, Nikola Stefelova, Andreas Alfons, Javier Palarea-Albaladejo, and Peter Filzmoser
Cellwise robust regression with compositional and real-valued covariates (see p. 37)
15:50 – 16:20

Lea Bottmer, Christophe Croux, and Ines Wilms
A cellwise robust lasso estimator (see p. 24)
16:20 – 16:50

Anthony-Alexander Christidis, Stefan Van Aelst, and Ruben H. Zamar
Data-driven diverse logistic regression (see p. 26)
16:50 – 17:20

Keynote presentation 17:25 – 18:25

Chair: Andreas Alfons Thursday 23

Peter J. Rousseeuw and Jakob Raymaekers
Flagging cellwise outliers using a robust covariance matrix (see p. 58)

Friday, 24 September 2021

Invited Session 09:30 – 10:20

Chair: Max Welz Friday 24

Battista Biggio
Machine learning (for) security: Lessons learned and future challenges (see p. 20)

Coffee Break: 10:20 – 10:50

Invited Session 10:50 – 11:40

Chair: Christian Hennig Friday 24

Johan A.K. Suykens
Deep learning, kernel machines and robustness (see p. 60)

A new robust Liu-type estimator for regression based on RAMML estimators

Sukru Acitas^a, Peter Filzmoser^b, Birdal Senoglu^c and Gamze Guven^d

^{a} *Department of Statistics, Eskisehir Technical University, Turkey,* ^{b} *Computational Statistics Research Unit, Vienna University of Technology, Austria,* ^{c} *Department of Statistics, Ankara University, Turkey,* ^{d} *Department of Statistics, Eskisehir Osmangazi University, Turkey,*

Multicollinearity among the predictors and nonnormality of the error terms are mostly encountered problems in the linear regression model, see e.g. Acitas and Senoglu [1]. In this study, we propose a new robust Liu-type estimator using robust adaptive modified maximum likelihood (RAMML) estimators [2] by following the similar lines as in Filzmoser and Kurnaz [3]. Liu-type RAMML estimators inherits the properties of the RAMML estimators, i.e. they are robust to $x-$ and/or $y-$outliers and also easy to compute. Furthermore, it is able to cope with the multicollinearity problem. It is shown that Liu-type RAMML estimators work quite well under different simulation scenarios.

Keywords: Multicollinearity, outlier, RAMML estimators.

References

[1] S. Acitas, B. Senoglu (2019). Ridge-Type MML Estimator in the Linear Regression Model, *Iranian Journal of Science and Technology, Transactions A: Science*,. **43**(2), 589–599.

[2] S. Acitas, P. Filzmoser, B. Senoglu (2021). A robust adaptive modified maximum likelihood estimator for the linear regression model. *Journal of Statistical Computation and Simulation*, **91**(7), 1394–1414.

[3] P. Filzmoser, F.S. Kurnaz (2018). A robust Liu regression estimator. *Communications in Statistics-Simulation and Computation*, **47**(2), 432–443.

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

W. H. Aeberhard[a], E. Cantoni[b], G. Marra[c], R. Radice[d]

[a] *Department of Mathematical Sciences, Stevens Institute of Technology,* [b] *Research Center for Statistics and GSEM, University of Geneva,* [c] *Department of Statistical Science, University College London,* [d] *Cass Business School, City, University of London*

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. This enables a fair comparison with existing robust estimators for the special case of generalized additive models, where our estimator competes favorably. The overall good performance of our proposal is illustrated by further simulations in the GAMLSS setting and by an application to functional magnetic resonance brain imaging using bivariate smoothing splines.

Keywords: Peanlized non-parametric regression; Robust smoothing parameter selection; Robust Information Criterion.

Robust penalized empirical likelihood estimation method for linear regression

Olcay Arslan^a and Senay Ozdemir^b

^{a}*Ankara University,* ^{b}*Afyon Kocatepe University*

Likelihood estimation method is a crucial part of the parameter estimation in regression models. However, since in some data sets it may not be possible to make any distributional assumptions on the error term, likelihood method cannot be used to estimate the parameters of interest. In those data sets, one can use the empirical likelihood estimation method to estimate the parameters of a linear regression model ([1], [2], [3]). The aim of this study is to propose a robust penalized empirical likelihood estimation method to estimate the regression parameters and select significant variables, simultaneously, for the data sets that a well-defined likelihood function may not be available. This will be achieved by combining robust empirical estimation method and the bridge penalty function. We investigate the asymptotic properties of the proposed estimator and explore the finite sample behavior with a simulation study and a real data example.

Keywords: Empirical likelihood; Robust estimation; Variable selection.

References

[1] A.B. Owen (1991). Empirical likelihood for linear models. *Ann. Statist*, 19, 1725-1747.

[2] S. Ozdemir and O. Arslan (2018). Combining Empirical Likelihood and Robust Estimation Methods for Linear Regression Models. *Comm. Stat.-Sim-Computing,* https://doi.org/10.1080/03610918.2019.1659968.

[3] S. Ozdemir and O. Arslan (2018). Empirical likelihood-MM (EL-MM) estimation for the parameters of a linear regression model. *STATISTICS*, 2021, 1, 45–67.

Robust scale estimation under shifts in the mean

I. Axt[a], A. Duerre[b] and R. Fried[a]

[a] *TU Dortmund University, Faculty of Statistics, Dortmund, Germany,* [b] *ECARES, Universite libre de Bruxelles, Brussels, Belgium*

In many applications the standard deviation of the observations needs to be estimated, e.g. for standardization. In the presence of outliers and jumps in the mean suitable estimation procedures are required, since ordinary scale estimators are biased in such situations. Therefore, we investigate a modified version of the well known median absolute deviation (MAD) to account for both sources of contamination. The formula of the MAD involves the sample median, which is not a good estimator of location in the presence of level shifts. Our proposal is to calculate the sample median in non-overlapping blocks and to consider absolute differences involving blockwise medians instead of a single median calculated on the whole sample. In this way only some blocks are affected by level shifts and the resulting modified MAD is robust against outliers and level shifts simultaneously. We proved strong consistency and asymptotic normality for independent random variables under some conditions on the number of change-points and the number of blocks. The Bahadur representation of the proposed estimator is shown to be the same as in the case of the ordinary MAD, resulting in the same asymptotic variance. Some suggestions on the choice of the block size are given. In a simulation study the modified MAD provides very good results. The proposed estimator performs well as compared to other robust methods, which are discussed for comparison, in many simulation scenarios.

Keywords: change-point, outliers, median absolute deviation

References

[1] I. Axt, A. Duerre, and R. Fried (2021+). Robust scale estimation under shifts in the mean. To appear in *Statistics.*

[2] S. Mazumder, and R. Serfling (2009). Bahadur representations for the median absolute deviation and its modifications. *Statistics & Probability Letters*, **79**(16), 1774–1783.

[3] R. Serfling (1980). Approximation theorems of mathematical statistics. *John Wiley & Sons*, New York.

[4] R. Serfling, and S. Mazumder (2009). Exponential probability inequality and convergence results for the median absolute deviation and its modifications. *Statistics & Probability Letters*, **79**(16), 1767–1773.

A model where the Least Trimmed Squares estimator is maximum likelihood

V. Berenguer-Rico[a], S. Johansen[b] and B. Nielsen[a]

[a] *University of Oxford,* [b] *University of Copenhagen*

The Least Trimmed Squares (LTS) estimator finds a sub-sample of h 'good' observations among n observations and applies ordinary least squares (OLS) on that sub-sample. We formulate a model in which this estimator is maximum likelihood. The model has 'outliers' of a new type, drawn from a distribution with values outside the realized range of h 'good', normal observations. The LTS estimator is found to be $h^{1/2}$ consistent and asymptotically standard normal in the location-scale case. Consistent estimation of h is discussed. The model differs from the commonly used ϵ-contamination models and opens the door for discussion on contamination schemes, methodological developments on tests for contamination and inferences based on the estimated 'good' data.

The suggested model is distinctive in that the errors are not i.i.d. Rather, the h 'good' errors are i.i.d. normal, whereas the $n - h$ 'outlier' errors are i.i.d., conditionally on the 'good' errors, with distributions assigning zero probability to the realized range of the 'good' errors. When $h = n$, we have a standard i.i.d. normal model, just as the LTS estimator reduces to the OLS estimator. The model is semi-parametric, so we use an extension of traditional likelihoods, in which we compare pairs of probability measures and consider probabilities of small hyper-cubes including the data.

In practice, it is of considerable interest to develop a theory for inference for LTS. Within a framework of i.i.d. ϵ-contaminated errors, the asymptotic theory depends on the contamination distribution and the scale estimator requires a consistency correction. Since the contamination distribution is unknown in practice, inference is typically done using the asymptotic distribution of the LTS estimator derived as if there is no contamination. This seems fine for an infinitesimal deviation from the central normal model. However, these approaches would lead to invalid inference in case of stronger contamination. Within the present framework, we derive the asymptotic properties of the LTS estimator for a location-scale version of the presented model and find that the LTS estimator has the same asymptotic theory as the infeasible OLS estimator computed from the 'good' data, when it is known which data are 'good'. As the asymptotic distribution does not depend on the contamination distribution, inference is much simpler.

In all cases, the practitioner must choose h, the number of 'good' observations. In our reading, this remains a major issue in robust statistics. We propose a consistent estimator for the proportion of 'good' observations, h/n, in a location-scale model.

The Least Median of Squares (LMS) estimator is closely related to LTS. Replacing the normal distribution in the LTS model with a uniform distribution gives a model in which LMS is maximum likelihood. We show that the LMS estimator is h-consistent and asymptotically Laplace in the location-scale case. This is at odds with the slow $n^{1/3}$ consistency rate found in the context of i.i.d. models.

Keywords: Contamination, Least Trimmed Squares, Leverage, Maximum Likelihood.

Machine Learning (for) Security: Lessons Learned and Future Challenges

B. Biggio[a]

[a] *University of Cagliari, Italy and Pluribus One*

In this talk, I will briefly review some recent advancements in the area of machine learning security [2] with a critical focus on the main factors which are hindering progress in this field. These include the lack of an underlying, systematic and scalable framework to properly evaluate machine-learning models under adversarial and out-of-distribution scenarios, along with suitable tools for easing their debugging. The latter may be helpful to unveil flaws in the evaluation process [7], as well as the presence of potential dataset biases and spurious features learned during training. I will finally report concrete examples of what our laboratory has been recently working on to enable a first step towards overcoming these limitations [3, 1], in the context of Android [6] and Windows malware detection [5, 4].

Keywords: Machine Learning, Computer Security, Adversarial Machine Learning, Malware Detection.

References

[1] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., Roli, F.: Evasion attacks against machine learning at test time. In: Blockeel, H., et al. (eds.), ECML PKDD, Part III. LNCS, vol. 8190, pp. 387–402. Springer (2013)

[2] Biggio, B., Roli, F.: Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition **84**, 317–331 (2018)

[3] Biggio, B., Nelson, B., Laskov, P.: Poisoning attacks against support vector machines. In: 29th ICML. pp. 1807–1814. Omnipress (2012)

[4] Demetrio, L., Biggio, B., Lagorio, G., Roli, F., Armando, A.: Functionality-preserving black-box optimization of adversarial windows malware. IEEE Transactions on Information Forensics and Security **16**, 3469–3478 (2021).

[5] Demetrio, L., Coull, S.E., Biggio, B., Lagorio, G., Armando, A., Roli, F.: Adversarial EXEmples: A survey and experimental evaluation of practical attacks on machine learning for Windows malware detection. ACM Trans. Priv. Secur. (2021)

[6] Demontis, A., Melis, M., Biggio, B., Maiorca, D., Arp, D., Rieck, K., Corona, I., Giacinto, G., Roli, F.: Yes, machine learning can be more secure! a case study on android malware detection. IEEE Trans. on Dep. Sec. Comp. **16**(4), 711–724 (2019).

[7] Pintor, M., Demetrio, L., Sotgiu, A., Manca, G., Demontis, A., Carlini, N., Biggio, B., Roli, F.: Indicators of attack failure: Debugging and improving optimization of adversarial examples. CoRR **abs/2106.09947** (2021)

A robust smoothed approach to functional canonical correlation analysis

G. Boente[a], and N. Kudraszow[b]

[a] *Universidad de Buenos Aires and CONICET,* [b] *Universidad Nacional de La Plata and CONICET*

In recent years, data collected in the form of functions or curves received considerable attention in fields such as chemometrics, image recognition and spectroscopy, among others. These data are known in the literature as functional data, see [3] for a complete overview. Functional data are intrinsically infinite–dimensional and, as mentioned for instance in [4], 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 more 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 dimension reduction techniques, to the context of functional data, usually through some regularization tool.

In this talk, we will focus on functional canonical correlation analysis, where data consist of pairs of random curves and the analysis tries to identify and quantify the relation between the observed functions. Under a Gaussian model, [2] showed that the natural extension of multivariate estimators to the functional scenario fails, motivating the introduction of regularization techniques which may combine smoothing through a penalty term and/or projection of the observed curves on a finite–dimensional linear space generated by a given basis, see [1] and [3]. The classical estimators use the Pearson correlation as measure of the association between the observed functions and for that reason they are sensitive to outliers.

To provide robust estimators for the first functional canonical correlation and directions, we will introduce two families of robust consistent estimators that combine robust association and scale measures with basis expansion and/or penalizations as a regularization tool. Both families turn out to be consistent under mild assumptions. We will present the results of a numerical study that shows that, as expected, the robust method outperforms the existing classical procedure when the data are contaminated A real data example will also be presented.

Keywords: Functional Canonical Correlation Analysis, Robust estimation, Smoothing techniques.

References

[1] He, G., Müller, H. G. and Wang, J. L. (2004). Methods of canonical analysis for functional data. *Journal of Statistical Planning and Inference*, **122**, 141-159.

[2] Leurgans, S. E., Moyeed, R. A. and Silverman, B. W. (1993). Canonical correlation analysis when the data are curves, *Journal of the Royal Society, Series B*, **55**, 725-740.

[3] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, Springer,

Berlin.

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

A $B-$spline robust approach for partially linear additive models

G. Boente[a], and A. M. Martinez[b]

[a] *CONICET and Universidad de Buenos Aires,* [b] *CONICET and Universidad Nacional de Lujan*

To deal with the curse of dimensionality, partially linear additive models provide a flexible and interpretable approach to build predictive models. They combine features in an additive manner, allowing each of them to have either a linear or non–linear effect on the response. More precisely, under a partially linear additive model both parametric and nonparametric components coexist. Among others, [2] describes some of the advantages of partially linear additive models, including the facts that they are easily interpretable, and that the estimators for the parametric components are more efficient.

Under a partially linear additive model, we deal with independent and identically distributed observations $(Y_i, \mathbf{Z}_i^{\mathrm{T}}, \mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R} \times \mathbb{R}^q \times \mathbb{R}^p$, $1 \leq i \leq n$, such that

$$Y_i = \mu + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{Z}_i + \sum_{j=1}^{p} \eta_j(X_{ij}) + \sigma \varepsilon_i \,,$$

where $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^{\mathrm{T}}$ and the errors ε_i are centered and independent from the covariates. Then, the univariate unknown functions $\eta_j : \mathbb{R} \to \mathbb{R}$ $(1 \leq j \leq p)$, the coefficients $\mu \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^q$ and the scale parameter $\sigma > 0$ are the quantities to be estimated. Classical estimation procedures based on least squares assume that $\mathbb{E}(\varepsilon_i) = 0$ and $\mathrm{V\small{AR}}(\varepsilon_i) = 1$ and may be found in [1], for instance.

In this presentation, we introduce a family of robust estimators that combines $B-$splines with robust $MM-$regression estimators and avoids moment conditions for the errors. Consistency results, rates of convergence of the robust estimators as well as the asymptotic distribution of the estimators of $\boldsymbol{\beta}$ are obtained under mild assumptions. To select the dimension of the $B-$spline basis, a robust version of the BIC criteria is considered. Through the results of a Monte Carlo experiment we will show the advantage of the proposed methodology over the classical one for finite samples. Finally, we will also illustrate the robust proposal on a real data set.

Keywords: B-splines, Partially Linear Additive Models, Robust estimation.

References

[1] Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. Springer.

[2] Opsomer, J. D. and Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *Journal of Computational and Graphical Statistics*, **8**, 715–732.

A cellwise robust lasso estimator

L. Bottmer^{a,b}, C. Croux^{c} and I. Wilms^{a}

^{a}*Maastricht University,* ^{b}*Stanford University,* ^{c}*EDHEC Business School*

The high-dimensional multiple regression model is an important workhorse for data scientists. The lasso is a popular estimator to reduce the dimensionality by imposing sparsity on the estimated regression parameters. The lasso is, however, not a robust estimator. Nevertheless, outliers frequently occur in high-dimensional datasets. In this talk, we propose the sparse shooting S, a cellwise robust lasso estimator [1]. The resulting regression coefficients are sparse, meaning that many of them are set to zero, hereby selecting the most relevant predictors. As such, the sparse shooting S is computable in high-dimensional settings with more predictors than observations. Moreover, a distinct feature of this estimator is its ability to deal with cellwise contamination, where many cells of the design matrix of the predictor variables may be outlying. We compare its performance to several other sparse and/or robust regression estimators.

Keywords: Lasso, Cellwise Outliers, Robust Regression

References

[1] L. Bottmer, C. Croux and I. Wilms (2021). Sparse regression for large data sets with outliers. *European Journal of Operational Research*, doi: https://doi.org/10.1016/j.ejor.2021.05.049.

A Comparison Study of Robust Mixed Effects Models for Analyzing Degradation of Photovoltaic Modules

B. Brune[a,b], I. Ortner[a] and P. Filzmoser[b]

[a] *Applied Statistics GmbH, Vienna, Austria,* [b] *TU Wien, Austria*

There are different approaches to the robustification of mixed effects models in the literature: Rank-based approaches [1, 2], methods that are based on down-weighing observations (M- and S-estimation, see e.g. [3, 4, 5]), as well as ones that replace the normal error distribution with more heavy-tailed distributions such as the t-distribution [6]. We compare those approaches in terms of their efficiency and robustness in a simulation study. Here, we place special emphasis on small and unbalanced data sets. We identify differences, similarities and shortcomings of the methods and inspire further research in this area.

The comparison is motivated by a data set obtained from accelerated aging experiments that were performed on photovoltaic systems under different climate conditions. The data set is challenging in the sense that it consists of very short time series (corresponding to the different modules of interest) with varying measurement time points and distances. The tested number of modules differs for the various climate conditions, producing an unbalanced data set. Some of the PV modules under consideration are faulty, and thus produce outlying observations (or outlying observation series).

Keywords: Mixed effects models, Unbalanced data sets.

References

[1] J. D. Kloke, J. W. McKean, and M. M. Rashid (2009). Rank-Based Estimation and Associated Inferences for Linear Models With Cluster Correlated Errors. *Journal of the American Statistical Association* **104**(485), 384–390.

[2] Y. K. Bilgic (2012). Rank-Based Estimation and Prediction for Mixed-Effects Models in Nested Designs. *Dissertation, Western Michigan University.*

[3] S. Copt, and M.-P. Victoria-Feser (2006). High-Breakdown Inference for Mixed Linear Models. *Journal of the American Statistical Association* **101**(473), 292–300.

[4] M. Koller (2013). Robust Estimation of Linear Mixed Models. *Dissertation, ETH Zurich.*

[5] C. Agostinelli and V. J. Yohai (2016). Composite Robust Estimators for Linear Mixed Models. *Journal of the American Statistical Association* **111**(516), 1764—1774.

[6] J. C. Pinheiro, C. Liu, and Y. N. Wu (2001). Efficient Algorithms for Robust Estimation in Linear Mixed-Effects Models Using the Multivariate t Distribution. *Journal of Computational and Graphical Statistics* **10**(2), 249—276.

Data-Driven Diverse Logistic Regression Ensembles

A. A. Christidis[a], S. Van Aelst[b] and R. H. Zamar[a]

[a] *Department of Statistics, University of British Columbia,* [b] *Department of Mathematics, K. U. Leuven*

A novel framework for statistical learning is introduced which combines ideas from regularization and ensembling. This framework is applied to learn an ensemble of logistic regression models for high-dimensional binary classification. In the new framework the models in the ensemble are learned simultaneously by optimizing a multi-convex objective function. To enforce diversity between the models the objective function penalizes overlap between the models in the ensemble. Measures of diversity in classifiers ensembles are used to show how our method learns the ensemble by exploiting the accuracy-diversity trade-off for ensemble models. In contrast to other ensembling approaches, the resulting ensemble model is fully interpretable as a logistic regression model, asymptotically consistent, and at the same time yields ex- cellent prediction accuracy as demonstrated in an extensive simulation study and gene expression data applications. The models found by the proposed ensemble methodology can also reveal alternative mechanisms that can explain the relationship between the predictors and the response variable.

References

[1] Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794.

[2] Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li (2020). xgboost: Extreme Gradient Boosting. R package version 1.1.1.1.

[3] Christidis, A.-A., L. Lakshmanan, E. Smucler, and R. Zamar (2020). Split regularized regression. Technometrics 62 (3), 330-338.

[4] Yang, Y., M. Pesavento, Z.-Q. Luo, and B. Ottersten (2019). Inexact block coordinate descent algorithms for nonsmooth nonconvex optimization. IEEE Transactions on Signal Processing.

[5] Yang, Y. and H. Zou (2017). gcdnet: LASSO and Elastic Net (Adaptive) Penalized Least Squares, Logistic Regression, HHSVM, Squared Hinge SVM and Expectile Regression using a Fast GCD Algorithm. R package version 1.0.5.

[6] Yu, B., W. Qiu, C. Chen, A. Ma, J. Jiang, H. Zhou, and Q. Ma (2020). Submito-xgboost: predicting protein submitochondrial localization by fusing multiple feature information and extreme gradient boosting. Bioinformatics 36 (4), 1074-1081.

[7] Zahoor, J. and K. Zafar (2020). Classification of microarray gene expression data using an infiltration tactics optimization (ito) algorithm. Genes 11 (7), 819.

A Robust Firefly Algorithm Based Feature Selection Method for EEG Signal Classification

E. Ergün[a], Ö. Aydemir[b]

[a] *Recep Tayyip Erdoğan University,* [b] *Karadeniz Technical University*

The main control center of the human body is the brain [1]. This control occurs when millions of nerve cells (neurons) that form the structural unit of the central nervous system communicate with each other. The electrical activity of the neurons forms the electroencephalography (EEG) signals, also known as the electrical picture of the brain [2]. EEG, which is recorded depending on electrical potential changes with electrodes placed in the skull, makes it possible to obtain and interpret the underlying information in the brain [3], [4]. Also, EEG signals form the basis of brain computer interface (BCI) systems [5]. In addition to use in different areas, BCI systems facilitate the lives of patients who cannot move any muscles but have no cognitive disorder. The basic BCI system commonly consists of pre-processing, feature extraction and classification steps. In recent years, studies have been focused on feature selection algorithms because reducing the dimensionality of features contributes to improve the classification performance in BCI approaches [6]. It can also save storage and computation time and increase comprehensibility. Particularly, nature inspired heuristic optimization algorithms became popular [7]. In this study, we applied firefly algorithm (FA) to 2-class motor imaginary (MI) based EEG signals in order to enhance the performance of a classifier and obtain a rapid and high performance BCI system without redundant features.

In recent years, especially nature inspired heuristic optimization algorithms became popular in order to eliminate unnecessary features. This paper addresses a crucial factor for effective classification of motor imaginary-based EEG signals that are an optimal selection of relevant EEG features using firefly algorithm. Firefly algorithm (FA) works on the principle of directing the less shiny than the light intensity emitted by fireflies in nature towards the bright. The algorithm can adaptively select the best subset of features and improve classification accuracy. In this study, following extracted Katz Fractal Dimension based features, effective feature(s) were selected by FA. The proposed method successfully applied on open access dataset which was collected from 29 subjects. We obtained an average 76.14% classification accuracy (CA) using k-nearest neighbor classifier. This is 4.4% higher than the CA calculated by using all features. These results proved that used method is robust for this dataset. We used a FA for feature selection to choose minimal number of features and to obtain even better classification accuracy by utilizing all features. The achieved results verified that FA is an effective search algorithm for feature selection problems. Further, effective feature(s) selected by FA can enhance the performance of right/left hand opening-closing motor imagery BCI application.

Keywords: electroencephalography, brain computer interface, katz fractal dimension, firefly algorithm, k-nearest neighbor

References

[1] Ö. Aydemir and E. Ergün (2019). A robust and subject-specific sequential forward

search method for effective channel selection in brain computer interfaces. *Journal of Neuroscience Methods*, **313**, 60-67.

[2] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy and F. Yger (2018). A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of Neural Engineering*, **15**(3), 031005.

[3] E. Yavuz and Ö. Aydemir (2016, August). Olfaction recognition by EEG analysis using wavelet transform features. *In 2016 International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)* (pp. 1-4). IEEE.

[4] Z. Jin, G. Zhou, D. Gao and Y. Zhang (2018). EEG classification using sparse Bayesian extreme learning machine for brain–computer interface. *Neural Computing and Applications*, 1-9.

[5] H. Zhiping, C. Guangming, C. Cheng, X. He and Z. Jiacai (2010, November). A new EEG feature selection method for self-paced brain-computer interface. *In 2010 10th International Conference on Intelligent Systems Design and Applications* (pp. 845-849). IEEE.

[6] M. Anbu and G. A. Mala (2019). Feature selection using firefly algorithm in software defect prediction. *Cluster Computing*, **22**(5), 10925-10934.

[7] G. Chandrashekar and F. Sahin (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, **40**(1), 16-28.

Use of a robust blind source separation approach for XRF core scanning of soft sediments

K. Fačevicová[a], C. Muehlmann[b], K. Nordhausen[c], M. Žídek[a] and O. Bábek[a]

[a] *Department of Mathematical Analysis and Applications of Mathematics, Palacký University Olomouc, Czech Republic,* [b] *Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Austria,* [c] *Department of Mathematics and Statistics, University of Jyväskylä, Finland*

In this contribution, methods of blind source separation (BSS) of compositional time series are proposed as a useful tool for analysis of element intensities measured by portable energy-dispersive X-ray fluorescence (EDXRF) device on split sediment cores. Core scanning by portable EDXRF device, which allows non-destructive and easily obtainable extraction of element intensities in soft sediments, is a low-cost alternative to the traditional destructive EDXRF analysis of powdered sample aliquots. Element concentrations measured by the portable and traditional apparatus differ due to the sediment porosity, variable water content and grain size distribution, what prevents from their direct statistical analysis. A calibration method based on orthogonal regression of pairwise log-ratios was suggested in [3]. This approach is from our perspective not optimal, since it does not respect the multivariate and relative nature of measured concentrations as well as possible presence of outliers and correlations between measurements from adjoining depths.

These features are respected if the robust BSS algorithm [1] is used and combined with principles of compositional data analysis [2]. This allows us to find latent variables, signals, mapping the stratigraphic trends imprinted to the geochemical structure of the sediment. In addition, also a noise component is modelled, and a source of the inaccuracy between element intensities given by the portable and traditional apparatus can be revealed.

The pros and cons of the proposed method will be demonstrated on a set of EDXRF scans of underwater sediment cores taken from the Nosice dam reservoir, Váh River, Slovakia.

Keywords: Blind source separation, Compositional data, Time series.

References

[1] P. Ilmonen, K. Nordhausen, H. Oja and F. Theis (2015). An Affine Equivariant Robust Second-Order BSS Method. In E. Vincent , A. Yeredor, Z. Koldovský and P. Tichavský (eds.), *Latent Variable Analysis and Signal Separation.* Springer, 328 – 335.

[2] K. Nordhausen, G., Fischer and P. Filzmoser (2020). Blind Source Separation for Compositional Time Series. *Mathematical Geosciences*, DOI:10.1007/s11004-020-09869-y.

[3] G.J. Weltje and R. Tjallingii (2008). Calibration of XRF Core Scanners for Quantitative Geochemical Logging of Sediment Cores: Theory and Application. *Earth and Planetary Science Letters*, 274(3-4), 423 – 438.

High-dimensional robust inference: Farming significant and important variables

J. Fan[a]

[a] *Princeton University*

Heavy-tailed distributions arise easily from high-dimensional data and they are at odd with commonly used sub-Gaussian assumptions. This talk introduces robust principle from a new perspective. The key observation is that the robustification parameter should adapt to sample size, dimension and moments for an optimal bias-robustness tradeoff. As an important application, we propose a factor-adjusted robust procedure for large-scale simultaneous inference with control of the false discovery proportion. We demonstrate that robust factor adjustments are extremely important in both improving the power of the tests and controlling FDP. We identify general conditions under which the proposed method produces consistent estimate of the FDP. Extensive numerical experiments demonstrate the advantage of the proposed method over several state-of-the-art methods especially when the data are generated from heavy-tailed distributions. We also apply the robust principle to high-dimensional variable selection and prediction.

Keywords: High-dimensional data, Inference, Heavy-tailed distributions.

Cluster Analysis with cellwise outliers with applications to robust functional clustering

L. A. García-Escudero[a], D. Rivera-García[b], A. Mayo-Iscar[a] and J. Ortega[c]

[a] Universidad de Valladolid, [b] Centro en Investigación en Matemáticas Guanajuato, [c] King Abdullah University of Science and Technology

A robust clustering procedure based on cellwise trimming is proposed. The approach follows by extending a robust PCA procedure with cellwise trimming introduced in [2]. In moderate and high dimensional problems, cellwise trimming is more appealing than the traditional casewise trimming because it avoids the use of very large and unrealistic trimming levels. It also prevents a significant loss of information associated with completely discarding the information from the non-outlying cells that occurs when trimming entire observations. The procedure can be seen as an extension for the casewise trimming method introduced in [1] for robust clustering.

A feasible algorithm is proposed that is based on alternated weighted regressions, a modified "concentration step" and the use of a short version of the regression LTS for the reassignment of observations.

The approach is particularized in the case of functional clustering in such a way that "pieces" of curves can be trimmed. A simulation study and two real data sets are considered to illustrate the methodology.

Keywords: Cluster Analysis, Trimming, Cellwise

References

[1] L. A. García-Escudero, A. Gordaliza, A., R. San Martín, S. Van Aelst and R. Zamar (2009). Robust linear clustering. *Journal of the Royal Statistical Society. Series B*, **71**(1), 301–318.

[2] H. Cevallos-Valdiviezo (2016). *On methods for prediction based on complex data with missing values and robust principal component analysis. PhD thesis, Ghent University.*

Robust adaptive variable selection in ultra-high dimensional linear regression models

A. Ghosh[a], M. Jaenada[b] and L. Pardo[b]

[a] *Indian Statistical Institute, Kolkata, India,* [b] *Complutense University of Madrid, Spain.*

We consider the problem of simultaneous variable selection and estimation of the corresponding regression coefficients in an ultra-high dimensional linear regression models, an extremely important problem in the recent era. The adaptive penalty functions are used in this regard to achieve the oracle variable selection property along with easier computational burden. However, the usual adaptive procedures (e.g., adaptive LASSO) based on the squared error loss function is extremely non-robust in the presence of data contamination which are quite common with large-scale data (e.g., noisy gene expression data, spectra and spectral data). In this paper, we present a regularization procedure for the ultra-high dimensional data using a robust loss function based on the popular density power divergence (DPD) measure along with the adaptive LASSO penalty. We theoretically study the robustness and the large-sample properties of the proposed adaptive robust estimators for a general class of error distributions; in particular, we show that the proposed adaptive DPD-LASSO estimator is highly robust, satisfies the oracle variable selection property, and the corresponding estimators of the regression coefficients are consistent and asymptotically normal under easily verifiable set of assumptions. Numerical illustrations are provided for the mostly used normal error density. Finally, the proposal is applied to analyze an interesting spectral dataset, in the field of chemometrics, regarding the electron-probe X-ray microanalysis (EPXMA) of archaeological glass vessels from the 16th and 17th centuries.

Keywords: Robustness, High-dimensional Statistics, Density Power Divergence, Adaptive LASSO estimator, Oracle property.

Inference for multivariate time series models: a measure transportation approach

M.Hallin [a], D.La Vecchia[b] and H.Liu[c]

[a] *ECARES Université libre de Bruxelles,* [b] *University of Geneva,* [c] *University of Science and Technology of China*

Measure transportation theory is a growing field of mathematics and it is popular in many other disciplines. For instance, optimal transportation mappings, as related to the results of Monge (1781) and Kantorovich (1942), are widely-applied in bioengineering (e.g. for image analysis), in economics (e.g. for resources allocation), in physics (e.g. for fluids dynamics analysis), in machine learning (e.g. for classification and network analysis), just to name a few; see e.g. Santambrogio (2015). In this talk, I focus on statistics and I explain how one can hinge on the solution to the Kantorovich primal problem to devise novel inference procedures (estimation and testing) for multivariate time series models. The talk contains two parts. (i) I present the results available in Hallin et al. (2020b), where we provide, for the first time in the literature, an application of measure transportation ideas to semiparametric VARMA models. The proposed *R-estimators* build on novel concepts of centered-outward ranks and signs (see Chernozhukov et al. (2017) and Hallin et al. (2020a)), which allow to overcome the classical problem of lack of canonical order in R^d, $d \geq 2$. (ii) I present the results available in Hallin et al. (2020b). I explain how centered-outward ranks and signs can be applied to define a novel class of *tests* for semiparametric VAR models. Numerical results (Monte Carlo simulations and real data analysis of macroeconomic time series) illustrate the theoretical findings of the two parts of the talk.

Keywords: Multivariate ranks, Distribution-freeness, Local asymptotic normality, Macroeconomic time series, Measure transportation.

References

[1] Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). Monge- Kantorovich depth, quantiles, ranks, and signs, *Annals of Statistics*, 45, 223–256.

[2] Kantorovich, L. V. (1942). On the translocation of masses. (Dokl.) *Acad. Sci. URSS*, 37(3):199– 201.

[3] Hallin, M. (2017). On distribution and quantile functions, ranks and signs in R^d. ECARES Working Paper.

[4] Hallin, M. and Paindaveine, D. (2004). Rank-based optimal tests of the adequacy of an elliptic VARMA model. *Annals of Statistics*, 6, 2642–2678.

[5] Hallin, M. and La Vecchia, D. (2017). R-estimation in semiparametric dynamic location-scale models. *Journal of Econometrics,* 2, 233–247.

[6] Hallin, M. and Werker, B. (2003). Semiparametric efficiency, distribution-freeness, and invariance. *Bernoulli*, 9, 137–165.

[7] Hallin, La Vecchia & Liu, *Center-outward R-estimation for semiparametric VARMA*

models, Journal of the American Statistical Association, (2020a), Published online: 07 Dec 2020.

[8] Hallin, La Vecchia & Liu, *Rank-based testing for semiparametric VAR models: a measure transportation approach*, (2020b), working paper.

[9] Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris.*

[10] Santambrogio, F. (2015). *Optimal transport for applied mathematicians*, Birkaüser, NY, 55(58-63):94.

[11] Villani, C. (2009). *Optimal Transport: Old and New*, Springer-Verlag, Berlin.

Is there a role for model assumption testing in applied statistics?

C. Hennig[a]

[a] *Dipartimento di Scienze Statistiche "Paolo Fortunati", University of Bologna*

The aim of robust statistics is to introduce procedures that perform well if standard model assumptions are fulfilled, but that do not deteriorate much if assumptions are violated and the truth is in a suitably defined neighbourhood of the standard model. There are also nonparametric procedures that aim at performing well over a wide class of models.

An approach that is often recommended and applied in practice is to use a standard procedure after having tested (and not rejected) the model assumptions by a model misspecification test. This comes with problems. The performance of the standard procedure is often affected, because conditionally on not rejecting the model assumptions it may differ from the unconditional performance under the assumed model [1, 2]. Furthermore, a non-rejection of the model assumptions does by no means imply that they are indeed fulfilled.

One may think that robust and nonparametric procedures make misspecification testing obsolete, but, as [4] argued, such procedures still come with model assumptions that may be even more difficult to check. Another line of thought is that a decision about what procedures to use and implicitly what model assumptions to tolerate should be made by more flexible exploratory tools, particularly data visualisation, but this arguably makes the conditioning problem even worse, because it cannot be investigated systematically to what extent decisions from preliminary exploratory data analysis affect the performance characteristics of later applied procedures.

Rather than conveying a simple message about whether or not (and how) model assumptions should be tested, I will go through a number of aspects of the issue and present some results that indicate that the matter is very dependent on the specifics of the situation in question [3]. Misspecification testing can be helpful as well as worse than useless, and at least some rough guidelines on when it helps will be given.

Keywords: Misspecification test, misspecification paradox, data visualisation

References

[1] T. A. Bancroft (1944). On biases in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics* **15**, 190–204

[2] C. Hennig (2007) Falsification of propensity models by statistical tests and the goodness-of-fit paradox. *Philosophia Mathematica* **15**, 166–192

[3] M. I. Shamsudheen, and C. Hennig (2021) Does Preliminary Model Checking Help With Subsequent Inference? A Review And A New Result. `arXiv:1908.02218`

[4] A. Spanos (2018). Mis-specification testing in retrospect. *Journal of Economic Surveys* **32**, 541–577

Cellwise robust regression with compositional and real-valued covariates

K. Hron[a], N. Štefelová[a], A. Alfons[b], J. Palarea-Albaladejo[c] and P. Filzmoser[d]

[a] Palacký University, [b] Erasmus University Rotterdam, [c] Biomathematics and Statistics Scotland, [d] Vienna University of Technology

Compositional data are multivariate observations carrying relative information [2]. Consequently, the potential source of outlyingness are not absolute values of components, but rather (log-)ratios between them. We propose a robust procedure to estimate a linear regression model with compositional and real-valued explanatory variables. The proposed procedure yields reliable results even when the data matrix contains outliers in individual cells (cellwise outliers), as well as entire outlying observations (rowwise outliers) [1]. Cellwise outliers are first filtered using the information from pairwise logratios and then imputed by robust estimates. Afterwards, rowwise robust compositional regression is performed to obtain model coefficient estimates. Simulations show that the proposed procedure generally outperforms traditional rowwise-only robust regression estimators such as the MM-estimator, as well as some other cellwise robust regression methods (shooting S-estimator, 3-step regression). An application to bio-environmental data reveals that the proposed procedure - compared to other regression methods - leads to conclusions that are more aligned with established scientific knowledge.

Keywords: Cellwise contamination, Compositional data, Pivot coordinates, Model-based imputation.

References

[1] N. Štefelová, A. Alfons, J. Palarea-Albaladejo, P. Filzmoser and K. Hron (2021). Robust regression with compositional covariates including cellwise outliers. *Advances in Data Analysis and Classification.* DOI: 10.1007/s11634-021-00436-9.

[2] P. Filzmoser, K. Hron, and M. Templ (2018). *Applied Compositional Data Analysis.* Springer Series in Statistics. Springer, Cham.

Doubly Robust Feature Selection with Mean and Variance Outlier Detection and Oracle Properties

L. Insolia[a,b], F. Chiaromonte[a,c], R. Li[c] and M. Riani[d]

[a] *Sant'Anna School of Advanced Studies,* [b] *Scuola Normale Superiore,* [c] *Penn State University,* [d] *University of Parma*

High-dimensional linear regression models are nowadays pervasive in most research domains. We propose a general approach to handle data contaminations that might disrupt the performance of feature selection and estimation procedures. Specifically, we consider the co-occurrence of mean-shift and variance-inflation outliers, which can be modeled as additional fixed and random components, respectively, and evaluated independently. Our proposal performs feature selection while detecting and down-weighting variance-inflation outliers, detecting and excluding mean-shift outliers, and retaining non-outlying cases with full weights. Feature selection and mean-shift outlier detection are performed through a robust class of nonconcave penalization methods. Variance-inflation outlier detection is based on the penalization of the restricted posterior mode. The resulting approach satisfies a robust oracle property for feature selection in the presence of data contamination – which allows the number of features to exponentially increase with the sample size – and detects truly outlying cases of each type with asymptotic probability one. This provides an optimal trade-off between a high breakdown point and efficiency. Effective and computationally efficient heuristic procedures are also presented. We illustrate the finite-sample performance of our proposal through an extensive simulation study and a real-world application.

Keywords: Nonconvex penalties, Robustness, Variable selection.

Tails and Probabilities for p-Outside values

P. K. Jordanova[a] and M. Stehlik[b]

[a] *Faculty of Mathematics and Computer Science, Konstantin Preslavsky University of Shumen, Bulgaria,* [b] *Department of Applied Statistics and Linz Institute of Technology, Johannes Kepler University, Linz, Austria.*

The task for a general and useful classification of the heaviness of the tails of probability distributions still has no satisfactory solution. Due to lack of information outside the range of the data the tails of the distribution should be described via many characteristics. Index of regular variation is a good characteristic, but it puts too many distributions with very different tail behavior in one and the same class. One can consider for example Pareto(α), Fr*é*chet(α) and Hill-horror(α) with one and the same fixed parameter $\alpha > 0$. The main disadvantage of VaR, expectiles, and hazard functions, when we speak about the tails of the distribution, is that they depend on the center of the distribution and on the scaling factor. Therefore, they are very appropriate for predicting ”big losses”, but after a right characterization of the distributional type of ”the payoff”. When analyzing the heaviness of the tail of the observed distribution we need some characteristic which does not depend on the moments because in the most important cases of the heavy-tailed distributions theoretical moments do not exist and the corresponding empirical moments fluctuate too much. In this paper, we show that probabilities for different types of outliers can be very appropriate characteristics of the heaviness of the tails of the observed distribution. They do not depend on increasing affine transformations and do not need the existence of the moments. The idea origins from Tukey’s box plots, and allows us to obtain one and the same characteristic of the heaviness of the tail of the observed distribution within the whole distributional type with respect to all increasing affine transformations. These characteristics answer the question:

At what extent we can observe ”unexpected” values?

Keywords: Classification, Probability distributions, Heavy-tails.

References

[1] P. Jordanova, and M. Stehlik, (2020). IPO estimation of heaviness of the distribution beyond regularly varying tails. *Stochastic Analysis and Applications*, **38**(1), 76–96. TAYLOR & FRANCIS INC

[2] P. K. Jordanova. *Probabilities for p-outside values and heavy tails*. Editors: Prof. Kosto Mitov, PhD, Prof. Dr Mladen Savov, Konstantin Preslavsky Publishing House, Shumen, Bulgaria, 2020, ISBN 978-619-201-381-3(print), ISBN 978-619-201-401-8(online)178 pages

Acknowledgements: The work is supported by the Project RD-08-75/27.01.2021 from the Scientific Research Fund in Konstantin Preslavsky University of Shumen, Bulgaria.

Robust Boosting for functional regression

X. Ju[a], M. Salibian Barrera[a]

[a] *The University of British Columbia*

We present a robust non-parametric regression estimator for models with a functional explanatory variable and a scalar response. Our proposal is based on the Robust boosting algorithm for regression proposed in [1]. To fix ideas, let (Y_i, X_i), $1 \leq i \leq n$, be a sample following a regression model of the form $Y = F(X) + \varepsilon$. The interest is in estimating F, often with the goal of obtaining predictions for future observations of the explanatory variable X. Given a loss function $L : \mathbb{R}^2 \to \mathbb{R}_+$, gradient boosting [2] iteratively approximates the solution of

$$\min_G \frac{1}{n} \sum_{i=1}^{n} L(Y_i, G(X_i)) \,,$$

using "base" functions h^k that fit the negative gradient $(g_1^k, g_2^k, \ldots, g_n^k)^\top$, evaluated at the current estimate F^k: $g_i^k = \partial L(Y_i, b)/\partial b|_{b=F^k(X_i)}$. Typically the loss function is chosen to be $L(a,b) = (a-b)^2$, and robust gradient boosting algorithms can be constructed naturally by using a different L. Robust boosting [1] is a two-stage estimator that first minimizes an M-scale of the residuals and then uses a bounded ρ-function for the loss functon L above (as in [3]).

When $X \in \mathbb{R}^d$ regression trees are constructed by selecting at each node the coordinate of X and split that produces the best fit. In the functional case we build regression trees using optimal projections $\langle X, \beta \rangle$ and splits instead. Here $\langle \cdot, \cdot \rangle$ is the usual inner product in $\mathcal{L}^2$ and, in principle, $\beta \in \mathcal{L}^2$. There are at least two natural ways of choosing these optimal projections and we discuss both here.

Our initial numerical experiments are very encouraging in terms of being able to approximate general regression functions F, and also resist the damaging effect of potential outliers in the training data.

Keywords: Robustness, Functional Regression, Non-parametric regression.

References

[1] X. Ju and M. Salibian-Barrera. (2021). Robust Boosting for Regression Problems. *Computational Statistics and Data Science*, **153**. DOI: 10.1016/j.csda.2020.107065

[2] J. H. Friedman (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, **29**(5), 1189–1232.

[3] V. J. Yohai. (1987). High Breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, **15**(2), 642–656.

Rank Tests in Linear Model with Autoregressive Errors

J. Jurečková[a], Y. Güney[b], M. Schindler[b], J. Picek[b], and Y. Tuaç[c]

[a] *Charles University and Academy of Sciences, Prague,* [b] *Ankara University,* [c] *Technical University of Liberec*

In the linear regression model with possibly autoregressive errors, we propose a family of nonparametric tests for regression under a nuisance autoregression. The tests avoid the estimation of nuisance parameters, in contrast to the tests proposed in the literature. We consider the linear regression model of order s, whose model errors follow a p-th -order stationary autoregressive process:

$$y_t = \beta_0 + \mathbf{x}_t^\top \boldsymbol{\beta}^* + \varepsilon_t = \beta_0 + x_{t1}\beta_1 + ... + x_{tp}\beta_p + \varepsilon_t,$$
$$\varepsilon_t = \varphi_0 + \varphi_1\varepsilon_{t-1} + u_t + \ldots + \varphi_p\varepsilon_{t-p},\ t = 1, 2, ..., n,$$

where $\varphi_0, \varphi_1, \ldots, \varphi_p$ are unknown autoregression parameters. The innovations u_t are assumed being independently and identically distributed $(i.i.d.)$ with a continuous distribution function F and density f exponentially tailed, satisfying $E\,(u_t) = 0, \quad Var\,(u_t) = \sigma^2 < \infty$, otherwise generally unknown.

We construct the tests of the hypothesis:

$\mathbf{H}_0 :\ \boldsymbol{\beta}^* = \mathbf{0},\ \text{ with }\ \beta_0,\ (\varphi_0, \varphi_1, \ldots, \varphi_p)^\top \neq \mathbf{0}\ \text{ unspecified}$

The test of $\mathbf{H}_0$ is based on the autoregression rank scores and on the linear autoregression rank statistics for the hypothetical model. It is asymptotically equivalent to the rank test of $\mathbf{H}_0$ in the situation without nuisance autoregression.

Joint work with O. Arslan, Y. Güney, M. Schindler, J. Picek, Y. Tuaç

Robust optimal estimation of location from discretely sampled functional data

I. Kalogridis[a] and S. Van Aelst[a]

[a] *KU Leuven*

Estimating location is a central problem in functional data analysis, yet most current estimation procedures either unrealistically assume completely observed trajectories or lack robustness with respect to the many kinds of anomalies one can encounter in the functional setting. To remedy these deficiencies we introduce the first class of optimal robust location estimators based on discretely sampled functional data. The proposed method is based on M-type smoothing spline estimation with repeated measurements and is suitable for both commonly and independently observed trajectories that are subject to measurement error. We show that under suitable assumptions the proposed family of estimators is minimax rate optimal both for commonly and independently observed trajectories and we illustrate its highly competitive performance and practical usefulness in a Monte-Carlo study and a real-data example involving recent Covid-19 data.

Keywords: Functional data, M-estimators, smoothing splines. **References**

[1] I. Kalogridis and S. Van Aelst (2021). Robust optimal estimation of location from discretely sampled functional data. *https://arxiv.org/abs/2008.00782*.

[2] I. Kalogridis (2020). Asymptotics for M-type smoothing splines with non-smooth objective functions. *Test*, to appear.

Robustness in sequential decision making on parameters of stochastic data flows

A. Y. Kharin[a], Ton That Tu[b], H. Zhao[a], Y. Li[a]

[a] *Belarusian State University,* [b] *University of Danang*

Data flows monitoring and analysis are real parts of modern life. During these activities, problems of estimation, forecasting, and decision making appear quite often. To solve the mentioned problems, as models of data are behind the growing real data complexity, it is important [1] to analyze robustness [2] and to develop robust versions of relevant statistical procedures [3], [4].

In the problem of decision making about the parameters of stochastic data flows observed, sequential analysis [5] is the optimal methodology. Unfortunately, under deviations of the data in the flow from the used hypothetical model, the performance characteristics of sequential statistical decision rules suffer [6].

Three models of data flows are considered here: (i) independent homogeneous observations (including the special case of discrete distributions); (ii) inhomogeneous independent data (including time series following a trend); (iii) data with Markovian type of dependencies.

For the considered models the performance characteristics of the classical sequential statistical decision rules are analyzed under distortions of different types (“contamination”, ε-neighbourhoods in functional spaces, misspecification of the dependency model). Families of generalized sequential decision rules are proposed, and algorithms for robust versions construction within those families are developed.

The results are applied for COVID-19 data flows monitoring in the Republic of Belarus.

Keywords: Data flow, Sequential decision rule, Robustness.

References

[1] P. J. Huber, and E. M. Ronchetti (2009). *Robust Statistics.* Wiley, Hoboken.

[2] A. Y. Kharin (2013). *Robustness of Bayesian and Sequential Statistical Decision Rules.* BSU, Minsk. (In Russian)

[3] A. Y. Kharin (2005). Robust Bayesian prediction under distortions of prior and conditional distributions. *Journal of Mathematical Sciences*, **126**(4), 992–997.

[4] A. Y. Kharin, and D. V. Kishylau (2015). Robust sequential test for hypotheses about discrete distributions in the presence of “outliers”. *Journal of Mathematical Sciences*, **205**(1), 68–73.

[5] A. Wald (1947). *Sequential Analysis.* Chapman & Hall, London.

[6] A. Kharin, and Ton That Tu (2017). Performance and robustness analysis of sequential hypotheses testing for time series with trend. *Austrian Journal of Statistics*, **46**(3-4), 23–36.

Discrete-valued time series: parsimonious models and statistical analysis

Yu. S. Kharin[a]

[a] *Belarusian State University*

Time series analysis is deep developed for “continuous” data, but in practice observed time series x_t, $t = 1, 2, \ldots$, are usually discrete-valued [1]: $x_t \in A$, where A is some discrete observation space with cardinality $2 \leq N = |A| \leq +\infty$.

An universal base model for discrete-valued time series $x_t \in A$ is the homogeneous Markov chain MC(s) of some order s determined by some $(s+1)$-dimensional matrix of one-step transition probabilities: $P = \left(p_{i_1, i_2, \ldots, i_s, i_{s+1}}\right)$, $i_1, \ldots, i_{s+1} \in A$. Number of independent parameters for the MC(s) model increases exponentially w.r.t. the memory depth s: $D_{\mathrm{MC}(s)} = N^s(N-1)$. To identify this model we need to have huge data set and the computation work of size $O\left(N^{s+1}\right)$. To avoid this “curse of dimensionality” we propose [2] to use the parsimonious (“small-parametric”) models of high-order Markov chains that are determined by small number of parameters $d \ll D_{\mathrm{MC}(s)}$. We propose two approaches to construction of parsimonious matrix P: 1) squeezing of the set of different values of elements in matrix P; 2) using of some generation equation for the conditional probability distribution of the future state x_t under its prehistory $\{x_{t-1}, \ldots, x_{t-s}\}$.

We present two constructed parsimonious models: Markov chain of order s with r partial connections and Binomial conditionally nonlinear autoregressive time series. For these models we give statistical estimators of parameters, statistical tests and forecasting statistics. Theoretical results are illustrated on simulated and real data.

Keywords: Discrete data, Time series, Estimation.

References

[1] C. Weiss (2018). *An Introduction to Discrete-valued Time Series.* Springer, N.Y.

[2] Yu. Kharin (2013). *Robustness in Statistical Forecasting.* Springer, N.Y.

Minimum Distance Estimation in Linear Errors-in-Variables Regression Model

H. L. Koul[a] and P. Geng[b]

[a] *Michigan State University,* [b] *Illinois State University*

We develop analogs of a class of weighted empirical minimum distance estimators of the underlying parameters in errors-in-variables linear regression models, when the regression error distribution and the conditional distribution of conditionally centered measurement error, given the surrogate, are symmetric around the origin. This class of estimators is defined as the minimizers of integrals of the square of a certain symmetrized weighted empirical process of the residuals. It includes the least absolute deviation and an analog of the Hodges-Lehmann-Sen estimators. We first develop this class of estimators when the distributions of the true covariates and measurement errors are known, and then extend them to the case when these distributions are unknown but validation data is available. An example of the distributions of the errors and covariate is given where the Pitman's asymptotic relative efficiency of some m.d. estimators, relative to the bias corrected LSE, increases to infinity as the ME variance increases to infinity.

Findings of a simulation study show significant superiority of some members of the proposed class of estimators over the bias corrected least squares estimator, in finite samples. In particular, the analog of the Hodges-Lehmann-Sen estimator is seen to be much more robust against the increasing measurement error variance compared to the bias corrected least squares estimator, when the regression error distribution is t_2.

Keywords: errors-in-variables models, minimum distance estimators

Sparse Learning and Structure Identification for Ultra-High-Dimensional Image-on-Scalar Regression

Xinyi Li[a], Lily Wang[b] and Huixia Judy Wang[c]

[a] *The Statistical and Applied Mathematical Sciences Institute,* [b] *Iowa State University,* [c] *The George Washington University*

We consider high-dimensional image-on-scalar regression, where the spatial heterogeneity of covariate effects on imaging responses is investigated via a flexible partially linear spatially varying coefficient model. To tackle the challenges of spatial smoothing over the imaging response's complex domain consisting of regions of interest, we approximate the spatially varying coefficient functions via bivariate spline functions over triangulation. We first study estimation when the active constant coefficients and varying coefficient functions are known in advance. We then further develop a unified approach for simultaneous sparse learning and model structure identification in the presence of ultra-high-dimensional covariates. Our method can identify zero, nonzero constant and spatially varying components correctly and efficiently. The estimators of constant coefficients and varying coefficient functions are consistent and asymptotically normal for constant coefficient estimators. The method is evaluated by Monte Carlo simulation studies and applied to a dataset provided by the Alzheimer's Disease Neuroimaging Initiative.

Keywords: Bivariate splines; Imaging data; Triangulation; Varying coefficient models.

The K-sign depth and generalizations

D. Malcherczyk, K. Leckey and C. H. Müller

TU Dortmund

The K-sign depths describe the fit of given parameters $\theta \in \mathbb{R}^p$ in models based on residuals (e.g. regression models) by counting all ordered K-tupels of the residuals $R_1, \ldots, R_N$ with $K-1$ sign changes. Since the test statistic is distribution free under the true model parameter, the K-sign depth leads to a test for

$$H_0 : \theta \in \Theta_0$$

with an arbitrary subclass $\Theta_0 \subseteq \mathbb{R}^p$. We only assume that the errors of the model are independent with a continuous distribution, i.e., neither assumptions on the moments, homoscedasticity nor identical distributions are required.

We discuss an asymptotic equivalent representation of the K-sign depth based on a functional of the random walk associated to the sum of signs of the residuals. This representation will be used to derive the asymptotic distribution under the null hypothesis by applying a functional central limit theorem for random walks.

Furthermore, the sign function can be replaced by various score functions as a generalization of the test statistic where the sign function is understood as a special case. This generalization allows us to have more information from the residuals than the signs while having still robust properties. For a large class of score functions, the asymptotic asymptotic does not change except of a scale factor which has to be estimated. E.g., the scores can be chosen as weight functions which are used for M-estimators or as ranks.

Keywords: sign-test, distribution free, M-estimator, rank, regression, depth.

References

[1] Kustosz, Ch.P., Leucht, A. and Müller, Ch.H. (2016). Tests based on simplicial depth for AR(1) models with explosion. Journal of Time Series Analysis **37**, 763-784.

[2] Leckey, K., Malcherczyk, D., and Müller, C.H. (2020). Powerful generalized sign tests based on sign depth. *Discussion papers SFB 823.* URL: `https://www.statistik.tu-dortmund.de/2630.html`

[3] Malcherczyk, D., Leckey, K., and Müller, C.H. (2021). K-sign depth: From asymptotics to efficient implementation. *Journal of Statistical Planning and Inference* **215**, 344-355.

A Robust Approach to Classification and Regression Tasks for Microbiome Data

G. S. Monti[a] and P. Filzmoser[b]

[a] *Dep. of Economics, Management and Statistics, University of Milano–Bicocca, Milan, Italy,* [b] *Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria*

The frontier in genomic research states that the human microbiome is strictly linked to many essential functions. Variation in its composition may have important effects on human health and could be associated to status disease. Microbiome data, obtained from high-throughput DNA sequencing techniques, usually are reported as tables of counts, or proportions. The total number of reads per sample depends on the capacity of the instrument, in other words is constrained by the maximum number of sequences reads, resulting in constrained compositional data. Consequently, the abundance values are not informative per se, and the relevant information is contained in the ratios of abundances between different components of the microbiome. Such data are typically high-dimensional, and usually only a small subset of taxa is related to an external property. A naive application of the classical technique to microbiome data may lead to unreliable results, such as spurious correlations and subcompositional incoherences. Thus, there is a great demand of rigorous predictive modelling techniques for human microbiome data taking into account the special statistical scale of compositions.

With this in mind, we have proposed two novel robust regression models for microbiome-based compositional data as explanatory variables and a non-compositional response that preserve the principles of compositional data analysis (CoDa).

By combining the least trimmed squares [1] objective function with the elastic-net penalty, we have introduced the sparse least trimmed squares estimators with zero-sum constraint tailored for high dimensional data with continuous and binary response [3, 2, 5, 4] , namely the Robust Zero-Sum Regression (RobZS) and the Robust Logistic Zero-Sum Regression (RobLZS) estimators.

A comparison of the performance of the RobZS and RobLZS estimators with a non-robust counterpart and with other competitors by Monte Carlo simulation studies demonstrates their usefulness. Several microbiome data applications are considered to investigate the stability of the estimators to the presence of outliers. Robust Zero-Sum Regression and Robust Logistic Zero-Sum Regression are available as an R package that can be downloaded at https://github.com/giannamonti/RobZS.

Keywords: Compositional data, Penalized Regression, Metagenomics, Sparsity.

References

[1] Alfons, A., Croux, C., Gelper, S. (2013) *Sparse least trimmed squares regression for analyzing high dimensional large data sets.* Ann Appl Stat, **7**(1), 226–248

[2] Altenbuchinger, M., Rehberg, T., Zacharias, H.U., Stammler, F., Dettmer, K., Weber, D., Hiergeist, A., Gessner, A., Holler, E., Oefner, P.J., Spang, R. (2017) *Reference point insensitive molecular data analysis.* Bioinformatics, **33**(2), 219–226

[3] Lin, W., Shi, P., Feng, R., Li, H., 2014. *Variable selection in regression with*

compositional covariates. Biometrika, **101**, 785–797.

[4] Lu, J., Shi, P., Li, H. (2019) *Generalized linear models with linear constraints for microbiome compositional data*. Biometrics, **75**(1), 235–244

[5] Zacharias, H.U., Rehberg, T., Mehrl, S., Richtmann, D., Wettig, T., Oefner, P.J., Spang, R., Gronwald W., Altenbuchinger, M. (2017) *Scale-invariant biomarker discovery in urine and plasma metabolite fingerprints*. J Proteome Res, **16**(10), 3596–3605

K-sign depth tests: Some properties and some open problems

C. H. Müller

Department of Statistics, TU Dortmund

Up to now, powerful outlier robust tests for linear models are based on M-estimators and R-estimators and are quite complicated. On the other hand, the simple robust classical sign test based on signs of residuals usually provides a very bad power for certain alternatives. K-sign depth leads to robust tests, the K-sign depth tests or shortly K-depth tests, which are similarly easy to comprehend as the classical sign test. In particular for $K = 2$, they are equivalent with the classical sign test, but for $K > 2$, they are much more powerful. Originally, K-sign depth appeared in special situations of the simplicial regression depth introduced by Rousseeuw and Hubert [3] who proposed regression depth and simplicial regression depth as a measure of fit of a regression model. The simplicial regression depth becomes more manageable when it is equivalent to K-sign depth where sufficient conditions for this equivalence are given in [1].

In [2], some properties of the K-sign depth and the K-sign depth tests are given. In particular, a block implementation is given which leads to a linear implementation of K-sign depth while a naive implementation has a compexity of N^K for sample size N. Additionally, some conjectures for the K-sign depth for special block structures of the signs of the residuals are given which are used to explain the good power of the K-sign depth tests. However up to now, these conjectures could be proved only for $K = 3$ although they have a simple formulation for any $K \geq 3$. Especially, these conjectures are given in the talk.

Keywords: Regression depth, K-sign depth, robust tests.

References

[1] C. P. Kustosz, C. H. Müller, and M. Wendler (2016). Simplified simplicial depth for regression and autoregressive growth processes. *Journal of Statistical Planning and Inference*, **173**, 125–146.

[2] K. Leckey, D. Malcherczyk, M. Horn, and C. H. Müller (2021). Simple powerful robust tests based on sign depth. *Submitted.*.

[3] P. J. Rousseeuw and M. Hubert (1999). Regression depth. *Journal of the American Statistical Association*, **94**, 388–402.

Variable selection in compositional data using balance coordinates based on robust PLS

V. Nesrstová[a], K. Hron[a], J. A. Martín-Fernández[b], P. Filzmoser[c], J. Palarea-Albaladejo[d], I. Wilms[e]

[a] *Palacký University Olomouc,* [b] *University of Girona,* [c] *Vienna University of Technology,* [d] *Biomathematics and Statistics Scotland,* [e] *Maastricht University*

Compositional data consist of multivariate observations that carry relative information in the ratios between parts.For statistical analysis, compositional data are commonly expressed as log-ratio coordinates with respect to an orthonormal basis of their sample space. The choice of such a basis is relevant for the interpretability of the results, particularly in a high-dimensional context. A class of log-ratio coordinates called principal balances (PB) has been proposed as a suitable, interpretable choice in this respect. PB are constructed so that the first few log-ratio coordinates explain most of the data variability. Moreover, the fact that PB represent contrasts between subsets of compositional parts facilitates their interpretation.

In practice, compositional data sets frequently contain outliers, observations which are characterized by one or more aberrant pairwise log-ratios. In this contribution, we extend the originally proposed PB procedure [2] to the regression analysis context and make it robust, with a high-dimensional composition acting in an explanatory role. To this end, we embed robust partial least squares (PLS) estimation [3] into the construction of the balances. We start by performing robust PLS in in full-ranked orthonormal log-ratio coordinates following [1]. Next, we transform the resulting loadings to centered log-ratio coefficients which are then used to construct the first balance. To construct the other balances, we maximize the (absolute) robust MCD covariance between a candidate balance and a response, which can be either real-valued or dichotomous. The resulting balances can be used for both dimension reduction and variable selection in regression and classification problems with compositional covariates.

Keywords: Compositional Data, PLS regression, Principal Balances.

References

[1] P. Filzmoser, K. Hron, C. Reimann (2009). Principal component analysis for compositional data with outliers. *Environmetrics*, **20**(6), 621-632.

[2] J.A. Martín-Fernández, V. Pawlowsky-Glahn, J.J. Egozcue, R. Tolosana-Delgado (2018). Advances in principal balances for compositional data. *Mathematical Geosciences*, **50**(3), 273-298.

[3] S. Serneels, C. Croux, P. Filzmoser, P.J. Van Espen (2005) Partial robust M-regression. *Chemometrics and Intelligent Laboratory System*, **79**(1-2), 55–64.

Robust functional principal component regression: a comparison

L. Neubauer[a] and P. Filzmoser[a]

[a] *TU Wien, Austria*

Functional principal component regression is based on functional data where we observe an underlying stochastic process. In a regression setting we want to regress a scalar response y onto such stochastic process X resulting in a model $y = \langle X, \beta \rangle + \epsilon$, where the coefficient function β is unknown. As in a multivariate setting this regression is sensitive to outliers in both response and explanatory variables, and thus we want to robustify this procedure. A common technique for such model is to use functional principal components of the corresponding process and use the resulting scores to explain the response.

Two different types of estimators are compared. One is made for regular, densely observed data [2] whereas a new approach for irregular, longitudinal data is proposed which is based on [1]. In a simulation study all estimators are applied in various settings, covering regular and irregular as well as dense and sparse data. The results of this simulation study reveal acceptable performance of the robust estimators in clean and contaminated scenarios, especially in regular settings. However, in very sparse, irregular settings clearer performance differences are visible.

Keywords: Functional Regression, Functional Principal Component Analysis, Robustness.

References

[1] Graciela Boente and Matías Salibián-Barrera. Robust functional principal components for sparse longitudinal data. *Metron*, Feb 2021.

[2] Ioannis Kalogridis and Stefan Van Aelst. Robust functional regression based on principal components. *J. Multivariate Anal.*, 173:393–415, 2019.

Forecasting Multiple Time Series with Robust One-Sided Dynamic Principal Components

D. Peña[a] and V. J. Yohai[b]

[a] Department of Statistics and Institute of Financial Big Data, Universidad Carlos III de Madrid, Spain, [b] University of Buenos Aires and CONICET

Given a vector time series $\mathbf{z}_t = (z_{t,1}, ..., z_{t,m})', 1 \leq t \leq T$, we consider the class of robust one-sided dynamic principal components (ODPC). Since these principal components depend only of past and present values, they are apt for forecasting purposes. Then, given $k_1 \geq 0$, the first one-sided principal component is of the form.

$$f_t = \sum_{h=0}^{k_1} \mathbf{a}'_h \mathbf{z}'_{t-h}, \quad t = k_1^1 + 1, \dots, T, \tag{0.1}$$

This component can be used to reconstruct the series using k_2 lags, that is $z^R_{t,j}(\mathbf{a}, \mathbf{B}) = \sum\limits_{h=0}^{k_2} b_{h,j} f_{t-h}$, where $\mathbf{a} = (\mathbf{a}'_1, ., \mathbf{a}'_{k_1})'$ and $\mathbf{B}$ the $(k_2+1) \times m$ matrix with $\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_m]$ and $\mathbf{b}_j = (b_{0,j}, ..., b_{k_1})'$. In [1], the first one-sided principal component $\mathbf{f}$ =$(f_{k_1}, ..., f_T)'$ is defined with $\mathbf{a}$ and $\mathbf{B}$ obtained by minimizing the mean square reconstruction error, i.e., by

$$(\widehat{\mathbf{a}}, \widehat{\mathbf{B}}) = \arg \min_{\|\mathbf{a}\|=1, \mathbf{B}} \frac{1}{T - k_1 - k_2} \sum_{t=(k_1+k_2)+1}^{T} \left\| \mathbf{z}_t - \mathbf{z}^R_t(\mathbf{a}, \mathbf{B}) \right\|^2,$$

where $\mathbf{z}^R_t = (z^R_{t,1}, ..., z^R_{t,m})'$. Observe that if $k_1 = 0$ and $k_2 = 0$, this definition of first principal component coincides with the classical Hotelling's definition. Since this procedure minimizes an l_2 reconstruction error, outlier observations can have a large influence on $\mathbf{f}$. For this reason we propose a class of robust ODPC replacing the l_2- reconstruction errors by the sum of the squares of robust M scales of each variable . The reconstruction error of the variable j at the period t is given by $r_{t,j}(\mathbf{a}, B) = z_{t,j} - z^R_{t,j}(\mathbf{a}, \mathbf{B})$, $k_1^1 + k_2^1 + 1 \leq t \leq T$. Let $\mathbf{r}^j(\mathbf{a}, \mathbf{B}) = (r_{t,j}(\mathbf{a}, \mathbf{B}))_{k_1^1 + k_2^1 + 1 \leq t \leq T}$, then we define the first component as $(\widehat{\mathbf{a}}_S, \widehat{\mathbf{B}}_S) = \arg\min_{\|\mathbf{a}\|=1, \mathbf{B}} \sum_{j=1}^{m} s^2_M(\mathbf{r}^j(\mathbf{a}, \mathbf{B}))$. We call to this procedure S-ODPC. A second principal component can be defined applying a similar procedure, but now with the goal of reconstructing the residuals obtained with first component. Similarly, higher order principal components can be obtained. We propose an alternating weighted least squares algorithm to compute the S-ODPC procedure. A Monte Carlo study shows that the S-ODPC can be successfully used for forecasting high-dimensional multiple time series, even in the presence of outlier observations.

Keywords: Robustness, Time series, Dynamic Principal components.

References

[1] D. Peña, E. Smucler and V. J Yohai (2019). Forecasting Multiple Time Series with One-Sided Dynamic Principal Components. *J. Amer. Stat. Asoc*, **114**(1683), 1683–1694.

Over time robust estimation of subjective latent variables from cross-section repeated surveys under measurement error

V. Perez[a], J. M. Pavia[a] and C. Aybar[a]

[a] *University of Valencia*

The individual measurement of subjective latent variables is exposed to the impact of multiple factors, including the responder internal mood and the effect of the external environment. When the objective is to measure the aggregate evolution of summaries statistics (e.g., the means of population subgroups) of these variables in the whole population using repeated cross-section surveys over time, the problem escalates due to the accumulating effects of contextual effects across subjects (all biased in the same direction) and the unfeasibility of correcting these systematic deviations that longitudinal surveys offer. To this, we need to add the small sample sizes that finer granularities impose. This is the case when we want to measure ideology by age and gender through auto-self positioning responses.

In electoral and political surveys, ideology is usually observed as a relevant variable that can be used to improve election outcome forecasts and to answer many substantive sociological and political research questions. At the individual level, ideology is a variable that smoothly evolves over time but that is also short-term impacted by contextual effects, being intensely exposed to measurement error. At the aggregate level, the average ideology of subgroups of the population is similarly strongly influenced by general opinion states and, therefore, shows unexplained jagging evolutions.

To solve this problem, we propose to use robust statistics that through the use rolling windows exploit adjacent information in two dimensions: over time and at the individual level. That is, we admit that subjects with similar characteristics to subgroup of interest surveyed close in time can be treated as member of the target population. We borrow strength from the neighbours to build robust estimators more balanced from a bias-variance perspective by increasing sample sizes, attaining smoother (over-age and over-time) estimates. For this, we have assessed different approaches using a large database with more than 150 variables and more than 700,000 observations, collected in Spain over more than 30 years.

Keywords: Rolling windows, Sample size, Estimation error

A Robust Acquisition Function for Sequential Gaussian Process Inference

D. Rahbani, A. Morel-Forster and T. Vetter

University of Basel

We formulate robust inference in terms of sequential optimization and propose an acquisition function for robust Gaussian Process inference. In sequential optimization, observations are gradually introduced. The underlying assumptions are that the surrogate model is correct and the observations have Gaussian noise [1]. Outliers do not follow this noise assumption, causing masking effects and reducing accuracy. Instead of updating the noise model, we address this problem by providing a robust acquisition function. The acquisition function uses the variance in the predictive posterior distribution to choose the next observations to be used in inference. The function therefore filters outlier observations out, keeping the Gaussian noise model valid and enabling closed-form inference. We use the Neal dataset [2] to compare to robust GP inference strategies that use a Student-t likelihood with and without a trimmed estimator or that use a Gaussian likelihood with a Huber-loss M-estimator [2, 3, 4]. The proposed method performs as well as the robust approaches do but does not require a manual threshold and does not assume a known outlier distribution. The advantages directly arise from the robust acquisition function which uses the residual errors to estimate an outlier threshold alongside model inference. The proposed method can be extended to robust GP regression and outlier detection. We show this using a medical example, in which the healthy shape is reconstructed from a target organ with outliers caused by pathological surface deviations. The implementations using GPy in Python and Scalismo in Scala are provided with publication.

Keywords: Outlier Detection, Gaussian Process Inference, Bayesian Optimization.

References

[1] Seo, Sambu and Wallat, Marko and Graepel, Thore and Obermayer, Klaus (2000). Gaussian process regression: Active data selection and test point rejection. In *Sommer G., Krüger N., Perwass C. Mustererkennung 2000.*, Berlin, 27–34.

[2] Li, Z-Z and Li, Lu and Shao, Zhengyi (2021). Robust Gaussian process regression based on iterative trimming. *Astronomy and Computing*, **36**, 100483.

[3] Martinez-Cantin, Ruben and Tee, Kevin and McCourt, Michael (2018). Practical Bayesian optimization in the presence of outliers. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, Playa Blanca, PMLR **84**, 1722–1731.

[4] Gerig, Thomas and Morel-Forster, Andreas and Blumer, Clemens and Egger, Bernhard and Luthi, Marcel and Schönborn, Sandro and Vetter, Thomas (2018). Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, Xi'an, 75–82.

Transforming variables to central normality

J. Raymaekers[a] and P. J. Rousseeuw[a]

[a] *University of Leuven, Belgium*

Many real data sets contain numerical features (variables) whose distribution is far from normal (Gaussian). Instead, their distribution is often skewed. In order to handle such data it is customary to preprocess the variables to make them more normal. The Box–Cox and Yeo–Johnson transformations are well-known tools for this. However, the standard maximum likelihood estimator of their transformation parameter is highly sensitive to outliers, and will often try to move outliers inward at the expense of the normality of the central part of the data.

We propose a modification of these transformations as well as an estimator of the transformation parameter that is robust to outliers [1], so the transformed data can be approximately normal in the center and a few outliers may deviate from it. It compares favorably to existing techniques in an extensive simulation study and on real data. An implementation of the proposed method is available in the `R` package `cellWise` [2].

Keywords: Data transformation, Anomaly detection, Robustness

References

[1] Raymaekers, J. and Rousseeuw, P.J. (2021). Transforming variables to central normality. Machine Learning. *https://doi.org/10.1007/s10994-021-05960-5*

[2] Raymaekers, J. and Rousseeuw, P.J. (2021). cellWise: Analyzing Data with Cellwise Outliers. R package version 2.2.5.

Flagging cellwise outliers using a robust covariance matrix

P.J. Rousseeuw[a] and J. Raymaekers[a]

[a] *KU Leuven*

In recent years there has been much research on the topic of cellwise outliers, as witnessed by several talks at this conference. The techniques used are often quite different from earlier work on casewise outliers.

We propose a data-analytic method for detecting cellwise outliers. Given a robust covariance matrix, outlying cells (entries) in a row are found by the new **cellHandler** technique which combines lasso regression with a stepwise application of constructed cutoff values. The penalty term of the lasso has a physical interpretation as the total distance that suspicious cells need to move in order to bring their row into the fold. Moreover, the cellHandler technique provides estimates of the flagged cells, yielding cell residuals.

For actually *estimating* a cellwise robust covariance matrix, we also construct the **Detection-Imputation method** which alternates between flagging outlying cells and updating the covariance matrix as in the EM algorithm.

The proposed methods are illustrated by simulations and on real data about volatile organic compounds in children.

Keywords: Cellwise outliers, Robust covariance.

To appear in *Journal of Data Science, Statistics, and Visualization* (JDSSV).

Robust estimation under Linear Mixed Models: a Minimum Density Power Divergence approach

G. Saraceno[a], A. Ghosh[b], A. Basu[b] and C. Agostinelli[a]

[a] *Department of Mathematics, Università degli studi di Trento, Trento, Italy,*
[b] *Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata, India*

Many real-life data sets can be analyzed using Linear Mixed Models. They include multilevel models, longitudinal data and ANOVA models with repeated measures. Since linear mixed models are based on the normality assumption, the classical methods estimates can be greatly affected even by small deviations. Furthermore, data sets modeled using mixed effects are often large, so the identification of this contamination could be difficult. Recently, robust estimators to fit linear mixed models have been proposed. [1] introduced a high breakdown point S-estimator, namely CVFS-estimator, while the estimator given by [2], namely SMDM-estimator, suggests to achieve robustness by a robustification of the scoring equations. On the other hand, the density power divergence (DPD) family, which measures the discrepancy between two probability density functions, has been used to build a robust estimator in case of independent and identically distributed observations. [3] extended the construction of an estimator based on DPDs to the case of independent but not identically distributed data. They also showed that the Minimum Density Power Divergence Estimator (MDPDE) can easily applied to the linear regression problem. In the presented work, a robust estimator for linear mixed models, based on the MDPDE for independent but non identically distributed observations, has been developed. We proved the theoretical properties of the estimator, such as consistency and asymptotic normality. As well, the influence function and sensitivity measures were computed in order to verify the robustness properties. We also propose two candidates as "optimal" choices of the tuning parameter α of the MDPDE, where "optimal" is intended as the best compromise between robustness and efficiency. We conducted a simulation study comparing the proposed MDPDE, for different values of α, with the CVFS-, SMDM-estimators and the classical MLE. Different levels of case-wise contamination have been considered. Finally, we explored the performance of our estimator when used in a real-data example.

Keywords: Case-wise contamination, Linear Mixed Models, MDPDE.

References

[1] S. Copt and M. P. Victoria-Feser (2006). High breakdown inference in the mixed linear model. *Journal of American Statistical Association*, **101**:292–300.

[2] M. Koller (2013). *Robust Estimation of Linear Mixed Models.* PhD thesis, ETH Zürich.

[3] A. Ghosh and A. Basu (2013). Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. *Electronic Journal of Statistics*, **7**:2420–2456.

Deep learning, kernel machines and robustness

J. Suykens[a]

[a] *KU Leuven, ESAT-Stadius and Leuven.AI Institute*

A wide range of problems in supervised and unsupervised learning with kernel-based methods can be understood in terms of least squares support vector machines (LS-SVM). They may serve as core models and possess primal and dual model representations, which are represented by feature maps and kernel functions, respectively. Robust versions have been obtained by (re)-weighting schemes. Recently, new connections have been established between restricted Boltzmann machines (RBM), kernel principal component analysis (KPCA) and least squares support vector machines (LS-SVM), through duality principles. Based on the restricted kernel machines (RKM) representations, a new framework for deep learning and kernel machines is conceived. One can work either with explicit (e.g. multi-layered, convolutional) feature maps or implicit feature maps in connection to kernel functions. Within this framework we show how robust generative kernel machines are obtained with robust generative RKMs.

Keywords: support vector machines, restricted Boltzmann machines, robust generative kernel machines.

Blind source separation based on M autocovariance matrices

S. Taskinen[a], K. Nordhausen[a] and D.E. Tyler[b]

[a] *University of Jyväskylä,* [b] *Rutgers University*

Assume that the observed p time series are linear combinations of p latent uncorrelated weakly stationary time series. The aim of blind source separation (BSS) is to find an estimate for the unmixing matrix which transforms the observed time series back to uncorrelated latent time series. Classical AMUSE (Algorithm for Multiple Unknown Signals Extraction) method solves the BSS problem by jointly diagonalizing the sample covariance matrix and the sample autocovariance matrix with chosen lag. A natural extension of AMUSE is SOBI (Second Order Blind Identification) method, which approximately jointly diagonalizes the sample covariance matrix and several sample autocovariance matrices with chosen lags to solve the unmixing matrix. It is well known that in the presence of outliers, the sample covariance matrix and sample autocovariance matrices perform poorly and yield to unreliable unmixing matrix estimates. In this paper we propose a robust blind source separation method which utilizes so-called M autocovariance matrices. The M autocovariance matrices are similar to the classical M estimators in that they downweight the outliers using some preselected, bounded weight function. Simulation studies and a real data example are used to illustrate robustness and efficiency properties of proposed methods.

Keywords: ϵ-contaminated ARMA processes, Efficiency, Maximum bias.

Outlier Detection in Rating-Scale Data via Autoencoders

M. Welz and A. Alfons

Erasmus University Rotterdam

Rating-scale datasets collected from surveys are ubiquitous in empirical research, in particular in the social sciences. However, the presence of outliers may pose a threat to the reliability of the data and the validity of subsequent analyses. Interestingly, rating-scale outliers have not been awarded with much attention by the statistical literature. Common outlier detection methods from robust statistics are based on (multivariate) normality assumptions and are not suitable for rating-scale data due to their discrete and bounded nature. While some methods for the detection of rating-scale outliers have been proposed in the psychological literature, our simulations suggest that those methods can only detect certain types of outliers, but struggle to detect others. Moreover, we could not find any study on the effect of outliers in rating-scale data on statistical inference.

Our aims are twofold: (*i*) to quantify the impact of rating-scale outliers on statistical inference, and (*ii*) to propose a method that can reliably detect all types of rating-scale outliers. For assigning an outlyingness score to each individual, we propose to use the mean squared reconstruction errors returned by an autoassociative neural network (*autoencoder*) [1]. Surveys typically collect information on various *constructs* (e.g., certain values or attitudes), with each construct being captured via several rating-scale questions. This structure is conceptually reflected by autoencoders through the use of a compression layer for a lower-dimensional representation. For the decision rule to flag observations as outliers, we propose to transform the outlyingness scores to central normality by using a modified version of the reweighted maximum likelihood estimator for the Box-Cox transformation [2], and use the square root of a χ^2-quantile as cutoff.

By means of extensive simulation studies, we find that (*i*) even a few outliers in rating-scale data can bias parameter estimates in subsequent analyses, to the extent of making type I and type II errors in hypotheses tests on those parameters; and (*ii*) our proposed method outperforms all benchmark techniques for outlier detection, both in terms of detection of true outliers as well as avoidance of false positives.

Keywords: Autoassociative Neural Networks, Outlier Detection, Rating-Scale Data.

References

[1] M. Kramer (1992). Autoassociative neural networks. *Computers & Chemical Engineering*, **16**(4), 313–328.

[2] J. Raymaekers and P. Rousseeuw (2021). Transforming variables to central normality. *Machine Learning*, DOI 10.1007/s10994-021-05960-5.

Global quantitative robustness of instance ranking problems

Tino Werner[a]

[a] *Carl von Ossietzky University Oldenburg*

Consider the problem to order instances in a data set. If responses Y_i are available, instance ranking problems intend to learn a real-valued, here parametric, scoring function which makes a prediction $\hat{Y}_i$ for each instance. In contrast to regression, the goal is to minimize a pair-wise loss function, i.e., one considers all pairs of responses and their predictions and checks whether their ordering coindices, i.e., whether $Y_i - Y_j$ and $\hat{Y}_i - \hat{Y}_j$ have the same sign, leading to a coarser problem than regression. Such problems have a variety of applications in for example scientific, social, document retrieval and financial contexts.

Robust statistics studies the behaviour of estimators in the presence of perturbations of the data resp. the underlying distribution and provides different concepts to characterize local and global robustness. Here, we concentrate on the global robustness of parametric instance ranking problems in terms of a breakdown point (BDP) which represents the fraction of instances that need to be perturbed in order to let the estimator take unreasonable values. The regression BDP is understood in the sense that such a fraction of outliers has full control of the parameter's norm, more precisely, any finite bound on the norm can be exceeded. Recently, a BDP for classification and for multiclass-classification have been proposed. However, all existing BDP notions do not cover ranking problems so far.

It is out of question that for a ranking prediction, the worst case is to predict an inverted ordering which is even worse than to perform random guessing. This motivates our definition of a breakdown of the estimator as a sign-reversal of all components which causes the predicted orderings to be inverted, therefore we call our concept the order-inversal breakdown point (OIBDP) which is the minimum fraction of perturbed data points so that the predicted ordering is guaranteed to be inverted due to all components of the parameter being sign-reverted. We will study the OIBDP, based on a linear model, for several different instance ranking problems that we carefully distinguish, and provide least favorable outlier configurations, characterizations of the OIBDP as well as asymptotic upper bounds. We also discuss the case of sparse model selection and outline the case of SVM-type instance ranking estimators.

Our contribution is threefold: **i)** We propose the definition of the order-inversal BDP for instance ranking problems which embeds the BDP concept of robust statistics into this important subfield of machine learning; **ii)** we provide explicit worst-case outlier configurations for all types of instance ranking problems; **iii)** we compute (asymptotic) upper bounds for the corresponding OIBDPs for all types of instance ranking problems.

Keywords: Breakdown point, instance ranking problems.