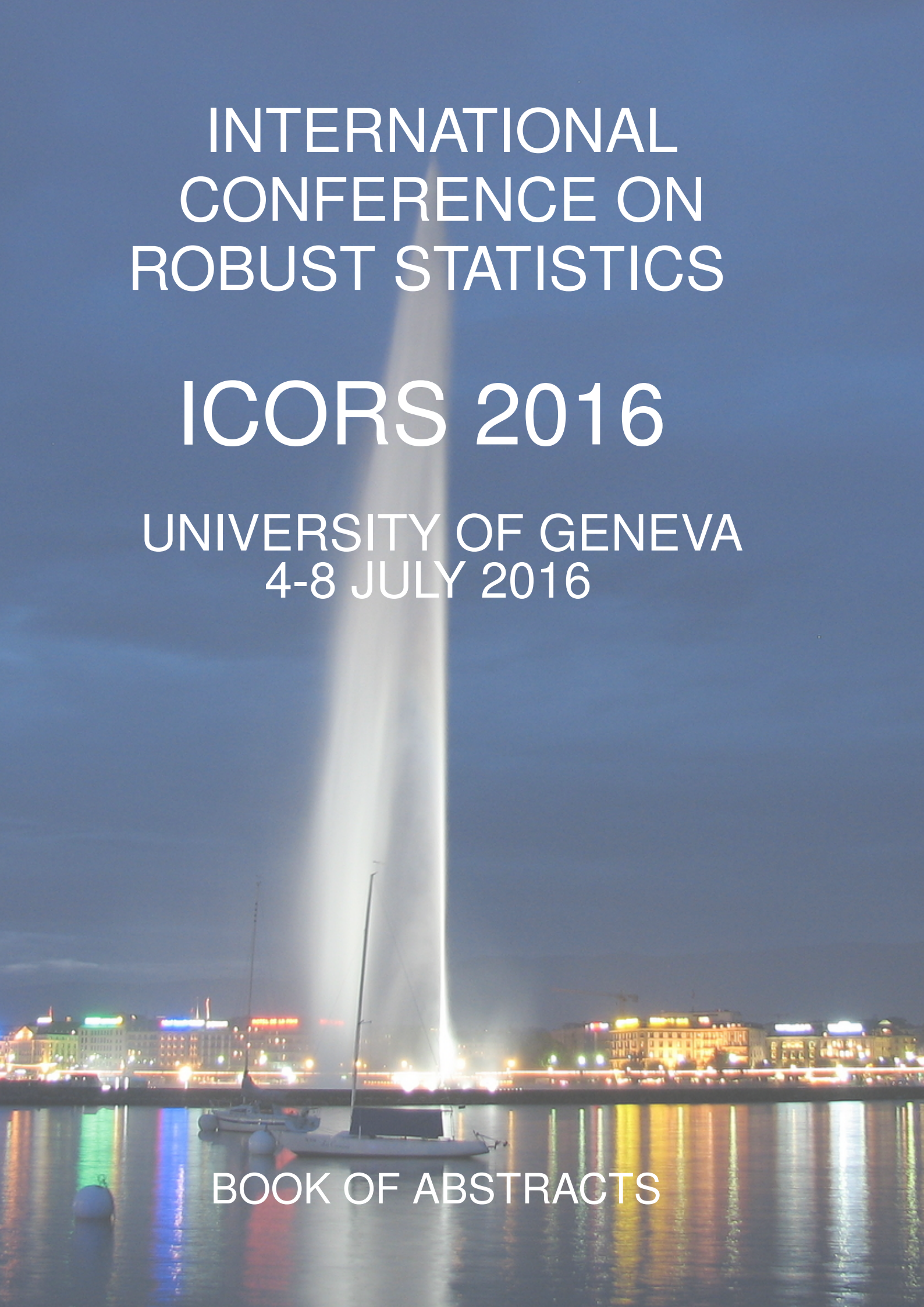
INTERNATIONAL CONFERENCE ON ROBUST STATISTICS

ICORS 2016

UNIVERSITY OF GENEVA
4-8 JULY 2016



BOOK OF ABSTRACTS

We gratefully acknowledge the support of the following organizations

THE MINERVA RESEARCH FOUNDATION
PRINCETON, NEW JERSEY

The Minerva Research Foundation
Princeton, New Jersey
www.minerva-foundation.org

University of Geneva
www.unige.ch

Geneva School of Economics and Management (GSEM)
University of Geneva
www.unige.ch/gsem/

Research Center for Statistics, GSEM
University of Geneva
www.unige.ch/gsem/rcs/

Swiss National Science Foundation
www.snf.ch

Preface

Welcome to ICORS2016, to Geneva, to the Geneva School of Economics and Management of the University of Geneva!

ICORS2016 will host more than 120 contributions on (robust) statistics coming from many places around the world. The organising and scientific committees are convinced that the conference will bring many new insights in our favourite discipline and even beyond.

We would like to express our gratitude to all attendees, participants in the organisation, and partner institutions, in particular the Minerva Research Foundation, that contributed to build up ICORS2016.

This booklet contains all submitted and accepted abstracts for the conference. The abstracts are ordered according to the name of the presenting author.

We wish you a fruitful conference and a pleasant stay in Geneva.

Geneva, 4 July 2016

Maria-Pia Victoria-Feser
Chair of the Organizing
Committee

Steering Committee

Peter Rousseeuw – Chairman	(Belgium)
Claudio Agostinelli	(Italy)
Olcay Arslan	(Turkey)
Claudia Becker	(Germany)
Ana Bianco	(Argentina)
Ayanendranath Basy	(India)
Graciela Boente	(Argentina)
Andrea Cerioli	(Italy)
Frank Critchley	(UK)
Christophe Croux	(Belgium)
Juan Cuesta Albertos	(Spain)
Laurie Davies	(Germany)
Rudolf Dutter	(Austria)
Luisa Fernholz	(USA)
Chris Field	(Canada)
Peter Filzmoser	(Austria)
Luis-Angel Garcia-Escudero	(Spain)
Ursula Gather	(Germany)
Alfonso Gordaliza	(Spain)
Marc Hallin	(Belgium)
Xuming He	(USA)
Mia Hubert	(Belgium)
Jana Jureckova	(Czech Republic)
Agustin Mayo	(Spain)
Ricardo Maronna	(Argentina)
Carlos Matran	(Spain)
Stephan Morgenthaler	(Switzerland)
Diganta Mukherjee	(India)
Hannu Oja	(Finland)
Daniel Peña	(Spain)
Ana Pires	(Portugal)
Marco Riani	(Italy)
Isabel Rodrigues	(Portugal)
Elvezio Ronchetti	(Switzerland)
Georgy Shevlyakov	(Russia)
David Tyler	(USA)
Stefan Van Aelst	(Belgium)
Roy Welsch	(USA)
Victor Yohai	(Argentina)
Ruben Zamar	(Canada)

Scientific Committee

Elvezio Ronchetti (Chair),	University of Geneva, Switzerland
Claudio Agostinelli,	University of Trento, Italy
Eva Cantoni,	University of Geneva, Switzerland
Luisa Fernholz,	Temple University and Minerva Research Foundation, USA
Yanyuan Ma,	University of South Carolina, USA
Maria-Pia Victoria-Feser,	University of Geneva, Switzerland

Organizing Committee

Maria-Pia Victoria-Feser (Chair)
Claudio Agostinelli
Marc-Olivier Boldi
Eva Cantoni
Davide La Vecchia
Dilara Yilmaz

Robust and Consistent Estimation of Fixed Parameters in General State-Space Models

W. H. Aeberhard[1*], *J. Mills Flemming*[1], *E. Cantoni*[2], *C. Field*[1] *and X. Xu*[3]

[1] *Department of Mathematics and Statistics, Dalhousie University, Canada; william.aeberhard@dal.ca, joanna.flemming@dal.ca, chris.field@dal.ca.*
[2] *GSEM and Research Center for Statistics, University of Geneva, Switzerland; eva.cantoni@unige.ch.*
[3] *Institute of Statistics, Nankai University, People's Republic of China; ximing@nankai.edu.cn.*
[*] *Presenting author*

Keywords. *Dynamic Models; Laplace Approximation; M-Estimation; Random Effects*

State-Space Framework

State-space models (SSMs) encompass a wide range of popular models encountered in various fields such as mathematical finance, control engineering and ecology. SSMs are essentially characterized by a hierarchical structure, with latent (unobserved) variables governed by Markovian dynamics. Fixed parameters in these models are traditionally estimated by maximum likelihood and typically include regression, auto-regression and scale parameters. The sensitivity of these estimates to deviations from the assumed model is problematic, all the more so since distributional assumptions about latent variables cannot be verified by the data analyst.

Robust Estimates of Fixed Parameters

Standard robust estimation techniques from generalized linear and time series models cannot be directly adapted to SSMs, and this mainly because of high-dimensional integrals that generally need to be approximated. We propose a robust estimating method inspired by the unpublished work of Eguchi and Kano [2001], where we downweight observations on the joint log-likelihood scale and then approximate the marginal robustified log-likelihood by Laplace's method. Computing a Fisher consistency correction term involves further approximations at the joint likelihood level to recover a typical M-functional form. Encouraging simulation results are presented with an application to a fish stock assessment.

References

Eguchi, S. & Kano, Y. (2001). Robustifing Maximum Likelihood Estimation. Unpublished Technical Report.

Robust Weighted Partial Least Squares Regression

A. Alin[1*] *and C. Agostinelli*[2]

[1] *Department of Statistics, Dokuz Eylul University, Turkey ; aylin.alin@deu.edu.tr*
[2] *Dipartimento di Matematica, Università di Trento, Italy; claudio.agostinelli@unitn.it*
[*] *Presenting author*

Keywords. *Outliers; SIMPLS; Weighted Likelihood*

As technology develops, it gets much easier to reach and collect data. In (multivariate) linear regression model setting, large data sets may be in the form of $n \gg k$ or $n \ll k$ where n is the sample size and k is the number of independent variables. One common problem in these two scenarios is the multicollinearity causing large standard errors for the least squares parameter estimates. The result is the unreliable model as well as harmed hypothesis testing and estimation. Partial least squares regression is a well known multivariate technique for modelling data sets with multicollinearity problem. The idea is to solve multicollinearity and build regression model simultaneously. It is based on calculating components for both explanatory variables and response variables which will have maximum covariance. These components are calculated using well known algorithms. In this study we focus on the SIMPLS algorithm [DeJong, 1993] and, despite its popularity, it is vulnerable against the outlying observations. This vulnerability is the result of the least squares estimation method or using nonrobust covariance matrix to obtain the components. In the literature, there are proposals that use robust estimators for both covariance matrix and parameter estimates and hence they are more resistant to outlying observations. Here, we propose a robust weighted SIMPLS algorithm which is based on iteratively reweighting approach with robust weights calculated using weighted likelihood methodology Markatou et al. [1998]. Another improvement with respect to the original SIMPLS algorithm is in the calculation of loadings. The performance of the proposed method, the ordinary SIMPLS and Partial robust M-regression [Serneels et al., 2005] methods are compared with an extensive simulation study.

References

De Jong, S. (1993). SIMPLS:an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, **18**, 251–263.

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

Serneels, S., Croux, C., Filzmoser, P. & Van Espen P.J. (2005). Partial robust M-regression. *Chemometrics and Intelligent Laboratory Systems*, **79**, 55–64.

Interpolation of Order Statistics in Stable Distribution

I. Almasi[1*], *A. Mohammadpour*[1] *and M. Mohammadi*[2]

[1] *Department of Statistics, Faculty of Mathematics Computer Science, Amirkabir University of Technology (Tehran Polytechnic); i.almasi@aut.ac.ir, adel@aut.ac.ir.*
[2] *Department of Statistics, Faculty of Basic Science, Behbahan Khatam Alanbia University of Technology, Behbahan, Iran; mohammadi@bkatu.ac.ir*
[*] *Presenting author*

Keywords. *Best linear unbiased interpolation or prediction; Projection; Hilbert space; Order statistics.*

1 Introduction

In this paper we derive the Best Linear Unbiased Interpolation for the missing order statistics from a stable distribution using the well-known *projection theorem*. The proposed interpolation method only needs the first two moments of both sides of a missing order statistic. A simulation study is performed to compare the proposed method with a few interpolation methods for some stable distributions.

2 Best linear interpolation

All results in this section will be stated for a Hilbert space $L^2 = \{X, \ \ E(X^2) < \infty\}$ with the inner product $< X_1, X_2 >= E(X_1 X_2)$.

Let $X_1, \ldots, X_n$ be a sequence of random variables with common pdf $f(x)$ and cdf $F(x)$ and let $Y_1 \leq \cdots \leq Y_n$ be the corresponding order statistics. Using the projection theorem, the conditional expectation $E_{\mathcal{M}(X_1,\ldots,X_n)}(X)$ is the best mean square predictor of X in $\mathcal{M}(X_1, \ldots, X_n)$, see Brockwell and Davis [1991].
Lemma 2.1 (Best order statistics interpolator). Let L^2 be a Hilbert space and $P_{\mathcal{M}}$ denote the projection mapping onto a closed subspace $\mathcal{M}$.
If $Y_1 \leq \cdots \leq Y_r \leq Y_s \leq \cdots \leq Y_n$ is order statistics of a distribution, then

$$P_{\mathcal{M}(Y_1,\ldots,Y_r,Y_s,\ldots,Y_n)}(Y_l) = P_{\mathcal{M}(Y_r,Y_s)}(Y_l), \tag{1}$$

where $r < l < s$.

Lemma 2.2 Best linear interpolator of l'th order statistic ($r < l < s$) can be obtained as follows

$$P_{\bar{sp}\{1,Y_r,Y_s\}}(Y_l) = a_0 + a_1 Y_r + a_2 Y_s, \tag{2}$$

where

$$a_1 = \frac{\rho_{r,l} - \rho_{r,s}\rho_{l,s}}{(1-\rho_{r,s}^2)}\frac{\sigma_{Y_l}}{\sigma_{Y_r}}, \quad a_2 = \frac{\rho_{l,s} - \rho_{r,s}\rho_{r,l}}{(1-\rho_{r,s}^2)}\frac{\sigma_{Y_l}}{\sigma_{Y_s}}, \quad a_0 = \mu_{Y_l} - a_1\mu_{Y_r} - a_2\mu_{Y_s}, \tag{3}$$

and $\mu_{Y_i} = E(Y_i)$, $\sigma_{Y_i}^2 = \mathrm{var}(Y_i)$ and $\rho_{i,j} = \mathrm{corr}(Y_i, Y_j), i, j = r, l, s$.

Corollary 2.1 Let $\hat{Y}_{l,\mathrm{BLUI}}$ stands for the best linear unbiased interpolation of Y_l. Under a few conditions, the BLUI is reduce to

$$\hat{Y}_{l,\mathrm{BLUI}} = \begin{cases} \mu_{Y_l} + \rho_{r,l}\frac{\sigma_{Y_l}}{\sigma_{Y_r}}(Y_r - \mu_{Y_r}), & \text{if } |l-s| \text{ is large,} \\ \mu_{Y_l} + \rho_{l,s}\frac{\sigma_{Y_l}}{\sigma_{Y_s}}(Y_s - \mu_{Y_s}), & \text{if } |r-l| \text{ is large,} \\ \mu_{Y_l} + \rho_{r,l}\frac{\sigma_{Y_l}}{\sigma_{Y_r}}(Y_r - \mu_{Y_r}) + \rho_{l,s}\frac{\sigma_{Y_l}}{\sigma_{Y_s}}(Y_s - \mu_{Y_s}), & \text{if } |r-s| \text{ is large,} \\ \mu_{Y_l}, & \text{if } |l-r| \text{ and } |l-s| \text{ are large.} \end{cases}$$

References

Brockwell, P. J., Davis, R. A. (1991). Time series: theory and methods, second ed., Springer Series in Statistics, Springer-Verlag, New York.

Robust estimation of high dimensional covariance and precision matrices

M. Avella-Medina[1*], *H. Battey*[2,3], *J. Fan*[2] *and Q. Li*[4]

[1] *Research Center for Statistics and GSEM, University of Geneva; marco.avella@unige.ch*
[2] *Department of Operations Research and Financial Engineering, Princeton University; hbattey@princeton.edu, jqfan@princeton.edu*
[3] *Department of Mathematics, Imperial College London; h.battey@imperial.ac.uk*
[4] *Department of Biostatistics, University of North Carolina; quefeng@email.unc.edu*
[*] *Presenting author*

Keywords. *Constrained ℓ_1-minimization; M-estimators; Optimal rates; Thresholding; Sparsity*

High dimensional data are likely to be drawn from distributions with heavy tails, at least in some coordinates. When this is the case, the performance of popular covariance and precision matrix estimators like adaptive thresholding [Cai & Liu , 2011] and adaptive CLIME [Cai et al. , 2012] is not guaranteed. We propose robust counterparts to these procedures that achieve the same minimax convergence rates attained in the above references but under only a finite fourth moment assumption. The key technical step in this work is in establishing the elementwise max norm convergence rates of functionals of the robust pilot estimators. The numerical performance of our estimators is investigated using both simulated and real data. In particualar, our simulations demonstrate the finite sample improvements achieved through the use of our robustification procedure over a range of data generating processes.

References

Cai, T. & Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. Journal of the American Statistical Association, **106**, 672–684.

Cai, T., Liu, W. & Zhou, H. (2012). Estimating Sparse Precision Matrix: Optimal Rates of Convergence and Adaptive Estimation. *arXiv preprint arXiv:1212.2882*.

Estimation in Accelerated Failure Time (AFT) Model

I.G. Balay[1*] *and B. Senoglu*[2]

[1] *Yıldırım Beyazıt University, Department of Banking and Finance, Ankara, Turkey; iklimgdk@gmail.com*
[2] *Ankara University, Department of Statistics, Ankara, Turkey; senoglu@science.ankara.edu.tr*
[*] *Presenting author*

Keywords. *AFT model, Skew-t; Censoring; Modifed likelihood; Iteratively Reweighting Algorithm*

Balay and Senoglu (2016) obtained the estimators of the parameters in the accelerated failure time (AFT) model under type II censoring when the distribution of the error terms is skew normal (SN). In this study, Balay and Senoglu (2016) is extended to Skew-t (ST) distribution which is more flexible for modeling the symmetric and the skew data sets then SN. Similar to Balay and Senoglu (2016), we use maximum likelihood (ML) methodology based on iteratively reweighting algorithm (IRA) to obtain the estimators of the model parameters in AFT. We also use the modified ML methodology known as MML as an alternative to ML methodology since it gives the closed form estimators of the parameters. At the end of the study, efficiencies of the ML and the MML estimators of the parameters are compared via an extensive Monte Carlo simulation study. As an application, we use a simulated data and obtain the parameter estimates of the AFT model based on this data.

References

Maronna, R., Martin, D. & Yohai, V. (2006). Robust Statistics: Theory and Methods. John Wiley & Sons, Chichester.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. Journal of Statistics*, **12**, 171–178.

Azzalini, A. and Genton, M.G. (2008). Robust likelihood methods on the skew t and related distributions. *International Statistical Review*, **76**, 106–129.

Balay, I.G. (2014). Robust Parameter Estimation for the Life-Stress Models and the Accelerated Failure Time Model Under Skew Distributions. *Ph.D. Thesis*. Ankara University, Turkey.

Balay, I.G. and Senoglu, B. (2016). Inference for the Accelerated Failure Time (AFT) Model Under Type II Censoring. *Submitted to ALT2016:6th International Conference on Accelerated Life Testing and Degradation Models*.

Wei, L.J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis *Statistics in Medicine*, **11**, 1871–1879.

Tiku, M.L. (1967). Estimating the mean and standard deviation from censored normal sample. *Biometrika*, **54**, 155–165.

Tiku, M.L. (1968). Estimating the parameters of normal and logistic distributions from censored samples. *Austral. J. Statist*, **10**, 64–74.

Chan, P.S., Ng, H.T., Balakrishnan, N. and Zhou, Q. (2008). Point and interval estimation for extreme-value regression model Type-II censoring. *Computational Statistics and Data Analysis*, **52**, 4040–4058.

A Minimum Distance Weighted Likelihood Method of Estimation

A.K. Kuchibhotla[1] *and* ***A. Basu***[2*]

[1] *Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia 19104-6340, USA; arunku@wharton.upenn.edu.*

[2] *Interdisciplinary Statistical Research Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700108, India; ayanbasu@isical.ac.in*

[*] *Presenting author*

Keywords. *Minimum Disparity Estimation; Weighted Likelihood Estimation; Residual Adjustment Function; Higher Order Influence Analysis.*

Over the last several decades, minimum distance (or minimum divergence, minimum disparity, minimum discrepancy) estimation methods have been studied in different statistical settings as an alternative to the method of maximum likelihood. The initial motivation was probably to exhibit that there exists other estimators apart from the maximum likelihood estimator (MLE) which has full asymptotic efficiency at the model. As the scope of and interest in the area of robust inference grew, many of these estimators were found to be particularly useful in that respect and performed better than the MLE under contamination. See Lindsay [1994]. Later, a weighted likelihood variant of the method was developed in the same spirit, which was substantially simpler to implement. See Markatou et al. [1998]. In the statistics literature the method of minimum disparity estimation and the corresponding weighted likelihood estimation methods have distinct identities. Despite their similarities, they have some basic differences. In this paper we propose a method of estimation which is simultaneously a minimum disparity method and a weighted likelihood method, and may be viewed as a method that combines the positive aspects of both. We refer to the estimator as the minimum distance weighted likelihood (MDWL) estimator, investigate its properties, and illustrate the same through real data examples and simulations. We briefly explore the applicability of the method in robust tests of hypothesis.

References

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Annals of Statistics*, **22**, 1081–1114.

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

Statistical Functionals of Residuals

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

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

Keywords. *Empirical Processes; Residuals; Robust procedures; Specification testing; Statistical Functionals.*

In this paper, we provide a theoretical framework to analyze a class of statistical functionals of residuals from the model

$$y_i = \beta' x_i + \varepsilon_i,$$

where the regressors can be stationary or non-stationary time series. The framework is built using empirical process theory. Specifically, we introduce a new class of weighted and marked empirical distribution functions that makes the analysis of statistical functionals of residuals general and systematic. Of particular interest are functionals used in specification testing since the stochastic properties of these tests can be analyzed in a simple manner within our framework.

As an example of statistical functional, consider the third standardized moment used in normality testing, that is,

$$T(\hat{F}_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{\varepsilon}_i}{\hat{\sigma}}\right)^3 = \int_{-\infty}^{+\infty} c^3 d\hat{F}_n(c),$$

where $\hat{\sigma}$ is the residual variance and

$$\hat{F}_n(c) = \frac{1}{n}\sum_{i=1}^{n} 1_{(\hat{\varepsilon}_i \leq \hat{\sigma} c)},$$

is the empirical distribution function of the standarized residuals.

The asymptotic properties of $T(\hat{F}_n)$ in its summation form can be studied using binomial expansions. However, this approach is somewhat tedious and not easily generalizable to other statistics involving different functionals. In this paper, we make the analysis of these functionals less involved using empirical processes theory. In particular, from the asymptotic behaviour of $\hat{F}_n$, asymptotic results for $T(\hat{F}_n)$ in its integral form can be more generally obtained. A difficulty that our approach overcomes is that the integrand, c^3 in the example above, may be unbounded. We overcome this difficulty by introducing a new class of weighted and marked empirical distributions functions of the form

$$\hat{F}_n^{g,h}(c) = \frac{1}{n}\sum_{i=1}^{n} g(x_{in}) h(\hat{\varepsilon}_i/\hat{\sigma}) 1_{(|\hat{\varepsilon}_i| \leq \hat{\sigma} c)},$$

where $g(x_{in})$ is a weight function and $h(\hat{\varepsilon}_i/\hat{\sigma})$, for some smooth function h, is the mark. In the case of the third cumulant, $g(x_{in}) = 1$ and $h(\hat{\varepsilon}_i/\hat{\sigma}) = (\hat{\varepsilon}_i/\hat{\sigma})^3$. Empirical process techniques allow us to control the interplay between the tail behaviour of the empirical distribution functions

and the integrand, which in turn implies the control of the asymptotic behaviour of $T(\hat{F}_n)$. It is worth emphasizing that the theory is general enough as to allow for a variety of estimators for β and σ, including standard least squares as well as robust estimators.

We apply our theoretical framework to study the asymptotic properties of several specification tests in the context of both stationary and non-stationary regressors –some of these results are new in the literature. Firstly, the cumulant based test and the Kolmogorov-Smirnov test for normality are considered. Secondly, White's test for heteroskedasticity, which involves cross moments between regressors and functions of residuals such as $\sum_{i=1}^{n} |x_i|^q \hat{\varepsilon}_i^p$, is analyzed. Thirdly, specification tests are examined in the context of robust methods. To be more precise, we propose a two stage approach where we first detect outliers using a robust procedure and then, after eliminating them, standard specification tests are applied to the retained observations.

A new and fast block bootstrap based prediction intervals for GARCH processes with application to exchange rates

B. H. Beyaztas[1,3*], *U. Beyaztas*[1,3], *S. Bandyopadhyay*[2] *and W. M. Huang*[2]

[1] *Dokuz Eylul University Department of Statistics Izmir - Turkey*
[2] *Lehigh University Department of Mathematics Bethlehem 18015 PA-USA*
[3] *Istanbul Medeniyet University Department of Statistics Istanbul - Turkey*
beste.sertdemir@deu.edu.tr, ufuk.beyaztas@deu.edu.tr, sob210@lehigh.edu, wh02@lehigh.edu
[*] *Presenting author*

Keywords. *Financial time series; Prediction; Resampling methods.*

Measuring volatility and construction of valid predictions for future returns and volatilities have an important role in assessing risk and uncertainty in the financial market. To this end, the generalized autoregressive conditionally heteroscedastic (GARCH) model is one of the most commonly used technique for modeling volatility and obtaining dynamic prediction intervals for returns as well as volatilities. Technically, construction of prediction intervals requires some distributional assumptions which are generally unknown in practice. Moreover, the constructed prediction intervals along with the estimated parameter values can be affected due to any departure from the assumptions and may lead us to unreliable results. One of the remedy to construct prediction intervals without considering distributional assumptions is to use the well known resampling methods, e.g., the bootstrap.

In this study, we propose a new bootstrap algorithm to obtain prediction intervals for GARCH processes which can be applied to construct prediction intervals for future returns and volatilities of conditionally heteroskedastic time series models. The advantages of the proposed method are two-fold: (a) it often exhibits improved performance and, (b) is computationally more efficient compared to other available resampling methods. Also, it is more robust to model disturbances than the existing resampling methods. The superiority of this method over the other resampling method based prediction intervals is explained with Spearman's rank correlation coefficient. The finite sample properties of the proposed method are also illustrated by an extensive simulation study.

A weighted likelihood ratio test for change point analysis in time series

U. Beyaztas[1,2*], *B. H. Beyaztas*[1,2] *and A. Alin*[1]

[1] *Dokuz Eylul University Department of Statistics Izmir - Turkey*
[2] *Istanbul Medeniyet University Department of Statistics Istanbul - Turkey*
ufuk.beyaztas@deu.edu.tr, beste.sertdemir@deu.edu.tr, aylin.alin@deu.edu.tr
[*] *Presenting author*

Keywords. *Change point; Weighted likelihood; Time series.*

1 Abstract

Homogeneity of parameters as well as structural stability are important aspects for time series data analysis. The common way to evaluate the structural stability is to test the model parameters for a possible change at an unknown time point, generally named as change point. It is an important problem in many scientific fields; such as, financial market analysis, quality control, medical researches, etc. It is not surprising that it has received great attention in the literature. Many test statistics including likelihood ratio (LR) based tests have been developed to detect unknown change points in the time series. The LR based test statistics may give unsatisfactory results on the change point detection and parameter estimations in the presence of unusual data points since such points have a great effect on the estimates obtained by the likelihood function. Using robust methods is a possible solution to deal with this problem.

In this study, we focus on the detection of change points in autoregressive of order p (AR(p)) models, and we restrict our attention on one change-point at an unknown time point. We propose a weighted likelihood based ratio test statistic to estimate the change point. The proposed test statistic is asymptotically equivalent to the conventional LR test under the null hypothesis of no change. The finite sample properties of the test statistic are illustrated by an extensive simulation study. Under considered scenarios, our proposed test statistic has better performance compared the conventional LR when there is a change in the model with possible outliers.

Robust estimation in single index models with asymmetric errors

C. Agostinelli[1], ***A. Bianco***[2*] *and G. Boente*[3]

[1] *Dipartimento di Matematica, Università degli Studi di Trento, Trento, Italy; claudio.agostinelli@unitn.it.*
[2] *Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Ciudad Universitaria, Pabellón 21, 1428, Buenos Aires, Argentina; abianco@dm.uba.ar.*
[3] *Departamento de Matemáticas, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and IMAS, CONICET, Ciudad Universitaria, Pabellón 1, 1428, Buenos Aires, Argentina; gboente@dm.uba.ar.*
[*] *Presenting author*

Keywords. *Kernel Weights; Local Polinomials; Single Index Models; Robust Estimation.*

Parametric and nonparametric models are two important branches of of regression analysis. An alternative between them is given by semiparametric models, that combine parametric components with nonparametric ones, retaining the advantages of both types of models and avoiding their drawbacks.

A relevant topic in this broad class of models is given by Simple Index Models (SIM) in which the response variable y is related to the covariate vector x through the equation

$$y = \eta(\beta' x) + \epsilon\,, \tag{4}$$

where $\beta, x \in R^q$ and $\eta : R \to R$ is a univariate real valued function. For the sake of identifiability, it is assumed with no loss of generality that $\|\beta\| = 1$ and the first component of β is positive, where $\|\cdot\|$ denotes the Euclidean norm.

Theses models reduce the dimensionality of the covariates by means of the single index $\beta' x$, capturing at the same time a possible nonlinear trend by means of the function η. In this way these models cope with the *curse of dimensionality*. They can also be seen as a technique of dimension reduction since, if β can be estimated in an efficient way, variable $\beta' x$ can be use as a carrier to estimate nonparametrically the function η.

Most of the literature assumes that the errors distribution has finite second moment and mean zero. However, in the robust framework this assumption is generally replaced by the symmetry of the errors term distribution, in order to achieve Fisher–consistent estimators. In some situations the practitioner faces the problem of asymmetric errors, as it is the case when the error term distribution belongs to a class of exponential families, for instance the log–gamma distribution. We focus on the problem of robust estimating the parametric and nonparametric components of model (6) when the density of the error ϵ is of the form

$$g(\epsilon, \alpha) = Q(\alpha) \exp^{\alpha t(\epsilon)}\,,$$

with $\alpha > 0$ a nuisance parameter and $t : R \to R$ a continuous function with unique mode at ϵ_0, which includes the Gamma distribution with a log link.

A family of robust estimators for η and β based on a three–step procedure related to a profile approach is proposed. When nuisance parameters are present, they may be estimated using a

preliminary S-estimator which will allow to define also the tuning constant. In particular, we have introduced a robust consistent estimator of the nuisance parameter for the usual regression model with symmetric errors and for the log–gamma regression model.

Consistency results for the robust profile proposal is obtained. A preliminary simulation study is presented so as to validate the good behaviour of the estimators under the central model and under different contaminated scenarios.

A Simulation Study on Resampling Based Methods for Comparing k Independent Groups

U. Binzat[1*], *E. Yildiztepe*[1]

[1] *Department of Statistics, Dokuz Eylül University, Turkey; ugur.binzat@deu.edu.tr, engin.yildiztepe@deu.edu.tr*

[*] *Presenting author*

Keywords. *Permutation test; k independent groups; Null resampling; Bootstrap-t.*

1 Abstract

The F-test is the most well-known method for comparing more than two independent groups. The F-test assumes that all groups are normally distributed and their variances are homogenous. But the data usually does not hold these assumptions. Another well-known method, Kruskal-Wallis test, is the ranked based alternative of the F-test for comparing k independent groups. Also, resampling based methods, which are highly computer intensive, can be used to compare these groups.

In this study, a Monte Carlo simulation is conducted to compare five methods in terms of their power and ability to control Type I error using various simulation settings. We consider five methods: the F-test, Kruskal-Wallis test [Kruskal & Wallis, 1952], permutation test [Ernst, 2004], null resampling based test [Martin, 2007], bootstrap-t trimmed mean Welch test [Wilcox , 2012]. The F-test and Kruskal-Wallis test are included as a reference in this work.

References

Kruskal, W. H. & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, **47(260)**, 583–621.

Ernst, M. D.(2004). Permutation methods: a basis for exact inference. *Statistical Science*, **19(4)**, 675–685.

Martin, M.(2007). Bootstrap hypothesis testing for some common statistical problems: A critical evaluation of size and power properties. *Computational Statistics & Data Analysis*, **51(12)**, 6321–6342.

Wilcox, R. R.(2012). Robust Statistics: Theory and Methods. Academic Press & Sons.

Robust Inference for Dynamic Semiparametric Factor Models

J. Bodelet[1*] *and D. La Vecchia*[1]

[1] *Research center in statistics and Geneva School of Economics and Management, University of Geneva, Blv. Pont d'Arve 40CH-1211, Geneva, Switzerland; Julien.Bodelet@unige.ch, Davide.LaVecchia@unige.ch.*
[*] *Presenting author*

Keywords. *Robust statistics; Dynamic factor model; Functional Magnetic Resonance Imaging; M-estimators on sieves; Outliers; Semiparametric inference.*

The aim of this paper is to define a robust statistical methodology to conduct inference on high dimensional data, collected in a dynamic context and observed at changing locations. A typical example of this kind of data arises in neuroscience, where the functional Magnetic Resonance Imaging (fMRI) datasets convey information about the brain's response (the so called, blood-oxygen-level-dependent, BOLD, signals) to certain stimuli, over time and over different activation areas (the so called, voxels, namely volumetric pixels).

To control for the high-dimensionality of the data, Dynamic Semiparametric Factor Models (DSFM) have been introduced by Park et al. [2009], where the aim is to isolate a limited number L of (latent) factors, explaining the behavior of the response variable (say, Y, representing e.g. the BOLD signals). Specifically, we observe $(X_{t,i}, Y_{t,i})$ for $i = 1, ..., n$ (cross-sectional dimension) and $t = 1, ..., T$ (time series dimension) and we set the model:

$$Y_{t,i} \quad = \quad m_0(X_{t,i}) + \sum_{l=1}^{L} Z_{t,l} m_l(X_{t,i}) + \epsilon_{t,i}, \tag{5}$$

where $(Z_{t,1}, ..., Z_{t,L})$ is an unobservable L-dimensional process (representing the $L \ll n$ latent factors). The $(L+1)$-functions $(m_0, ..., m_L)$ are unknown real-valued functions, defined on a (compact) subset of $\mathbb{R}^d$. The variables $X_{1,1}; ...; X_{T,n}; \epsilon_{1,1}; ...; \epsilon_{T,n}$ are independent, while the errors have zero mean and finite variance. Estimation methods for the model in (5) have been developed (and implemented) in Fengler, M. R., Härdle, W. K. and Mammen, E. [2007], Park et al. [2009] and Härdle, W. K. & Majer, P. [2014]. The extant inferential procedures rely on the least squares method and make use of nonparametric techniques (e.g., kernels or splines), to estimate both the latent factors and the functions m_l, for $l = 0, 1, ..., L$. A challenging open problem related to the inference about DSFM is the treatment of outliers. Indeed, our simulation exercises emphasize that even a small number of anomalous records can largely distort the inference about the underlying latent factors. This problem is particularly relevant in the neurological applications, where diagnoses are obtained from the statistical analysis of fMRI datasets typically contaminated by motion artifacts and/or recording device failures; see Muschelli et al. [2014] and Power et al. [2014].

To cope with this statistical issue, we consider the class of $M-$estimators on sieves (see Chen, X. & Shen, X. [1997]) for DSFM and we define a class of robust inferential procedures which, by design, mitigates the impact of extreme observations. Since in the considered context a formal characterization of a robustness principle is missing, we first need to introduce a criterion which identifies the class of robust $M-$estimators for the DSFM. Then, we: (i) investigate the

asymptotic properties of our new estimators, following [van der Geer, S. A. , 2000]; (ii) define an algorithm to implement our procedure, following Park et al. [2009]; (iii) develop a methodology for selecting the constant which controls the degree of robustness of our $M-$estimates, following La Vecchia, D., Camponovo, L. & Ferrari, D. [2015]. Monte-Carlo simulations provide numerical evidence of the good performance of our methodology, under different contamination settings. An application to fMRI data analysis concludes the paper.

References

Chen, X. & Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, 289–314.

Fengler, M. R., Härdle, W. K. & Mammen, E. (2007). A semiparametric factor model for implied volatility surface dynamics. *Journal of Financial Econometrics*, **5**, 189–218.

Härdle, W. K. & Majer, P. (2014). Yield curve modeling and forecasting using semiparametric factor dynamics. *The European Journal of Finance*, 1–21.

La Vecchia, D., Camponovo, L. & Ferrari, D. (2015). Robust heart rate variability analysis by generalized entropy minimization. *Computational Statistics & Data Analysis*, **82**, 137–151.

Muschelli, J., Nebel, M. B., Caffo, B. S, Barber, A. D., Pekar, J. J. & Mostofsky, S. H. (2014). Reduction of motion-related artifacts in resting state fMRI using aCompCor. *Neuroimage*, **96**, 22–35.

Park, B. U., Mammen, E., Härdle, W. & Borak,S. (2009). Time series modelling with semiparametric factor dynamics. *Journal of the American Statistical Association*, **104**, 284–298.

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage*, **84**, 320–341.

van der Geer, S. A. (2000). Empirical Processes in $M-$estimation. Cambridge university press.

Robust testing for superiority between two regression curves

G. Boente[1*] *and J.C. Pardo–Fernández*[2]

[1] *Departamento de Matemáticas, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and IMAS, CONICET, Ciudad Universitaria, Pabellón 1, 1428, Buenos Aires, Argentina; gboente@dm.uba.ar.*
[2] *Departamento de Estatística e Investigación Operativa, Universidade de Vigo, Campus Universitario As Lagoas-Marcosende, Vigo, 36310, Spain; juancp@uvigo.es*
[*] *Presenting author*

Keywords. *Hypothesis testing; Nonparametric regression models; Robust inference; Smoothing techniques.*

Let us assume that the random vectors $(X_j, Y_j)^{\mathrm{T}} \in \mathbb{R}^2$, $j = 1, 2$, follow the homoscedastic nonparametric regression models given by

$$Y_j = m_j(X_j) + \varepsilon_j = m_j(X_j) + \sigma_j U_j, \tag{6}$$

where $m_j : \mathbb{R} \to \mathbb{R}$ is a nonparametric smooth function and the error ε_j is independent of the covariate X_j. As is usual in a robust framework, we will assume that the errors ε_j are such that $\varepsilon_j = \sigma_j\ U_j$, where U_j has a symmetric distribution $G_j(\cdot)$ with scale 1, so that we are able to identify the error's scale, σ_j. When second moments exist, as it is the case of the classical approach, these conditions imply that $\mathbb{E}(\varepsilon_j) = 0$ and $\textsc{Var}(\varepsilon_j) = \sigma_j^2$, which means that m_j represents the conditional mean, while σ_j^2 equals the residuals variance, i.e., $\sigma_j^2 = \textsc{Var}(Y_j - m_j(X_j))$. The nonparametric nature of model (6) offers more flexibility than the standard linear model when modelling a complicated relationship between the response variable and the covariate. In many situations, it is of interest to compare the regression functions m_1 and m_2 to decide if the same functional form appears in both populations. In particular, we will focus on testing the null hypothesis of equality of the regression curves versus a one-sided alternative. Let $\mathcal{R}$ be the common support of the covariates X_1 and X_2 where the comparison will be performed. The null hypothesis to be considered is

$$H_0 : m_1(x) = m_2(x) \text{ for all } x \in \mathcal{R},$$

while the alternative hypothesis is of the following one-sided type

$$\begin{aligned} H_1 :\ & m_1(x) \le m_2(x) \text{ for all } x \in \mathcal{R} \text{ and } m_1(x) < m_2(x) \text{ for } x \in \mathcal{A}, \\ & \text{where } \mathcal{A} \subset \mathcal{R} \text{ is such that } \mathbb{P}(X_j \in \mathcal{A}) > 0, \text{ for } j = 1, 2. \end{aligned} \tag{7}$$

To protect against atypical observations, the test statistic to be considered is based on the residuals obtained by using a robust estimate for the regression function under the null hypothesis. More precisely, our proposal combines the ideas of robust smoothing with those given in Neumeyer & Pardo-Fernández [2009] to obtain a procedure detecting root$-n$ alternatives. The asymptotic distribution of the test statistic is studied under the null hypothesis and under root$-n$ contiguous alternatives.The results of a Monte Carlo study performed to compare the finite sample behaviour of the proposed tests with the classical one obtained using local averages will be described.

References

Neumeyer, N & Pardo-Fernández, J. C. (2009). A simple test for comparing regression curves versus one-sided alternatives. *Journal of Statistical Planning and Inference*, **139**, 4006-4016.

Quantile regression for dependent data using a working odds ratios matrix

D. Bossoli[1*], *M. Bottai*[2]

[1] *Department of Statistical Sciences, University of Padua, Padova ; bossoli@stat.unipd.it.*
[2] *Unit of Biostatistics, Institute of Environmental Medicine, Karolinska Institutet, Stockholm; matteo.bottai@ki.se*
[*] *Presenting author*

Keywords. *Quantile regression; Dependent data; Generalized estimating equations; Working odds ratios matrix*

1 Abstract

Dependent data arises frequently in applied research. A common solution to adjusting for correlation among observations within clusters is to use generalized estimating equations. When the interest lies on quantiles of the conditional distribution of an outcome variable of interest, the working correlation matrix is no longer computed from the regression residuals but instead from the residuals' sign. Because correlation between binary variables is bounded, we propose an unconstrained alternative, the odds ratio. In addition to computational advantages, odds ratios allow flexible modeling. Different working structures can be estimated easily by appropriate logistic regression models. Simulations show similar results to generalized estimating equations applied to regression for the mean. We illustrate the proposed method with data from a randomized trial on cognitive behavior therapy for treatment of obsessive compulsive disorder.

2 Introduction and Method

Longitudinal and clustered data represent two frequent analytical situations in which observations within a cluster are not independent. In these cases, the statistical assumption about independent observations of traditional regression models is violated. Ignoring the dependency between observations within the same cluster generally leads to unbiased estimators but wrong standard errors. Generalized estimating equation (GEE) is a population-averaged (or marginal) method used to account for the dependency induced by the clustering. The dependency between observations within the same cluster is modeled through a correlation matrix, which is usually considered the same for all clusters. In the literature this matrix is called working correlation matrix, because estimated parameters and their standard errors are correct even if the correlation matrix is misspecified.

Let $\{y_{ij}, x_{ij}\}$, $i = 1, \ldots, n$, $j = 1, \ldots, T$ be the longitudinal data set, where $y \in \mathbb{R}$ is the response variable and $x \in \mathbb{R}^P$ is the covariate vector.
Consider the set of repeated measurement of the i-th individual, denoted by $y_i = (y_{11}, \ldots, y_{1T})$,

and its design matrix $x_i = (x_{11}, \ldots, x_{1T})$. Marginal quantiles can be obtained solving the following estimating equation: (Potdar & Shirke [2013])

$$\tilde{U}_Q(\beta) = \sum_{i=1}^{n} x_i^T W_i^{-1}(\eta) \tilde{\psi}_\tau(\epsilon_i) = 0,$$

where $\tilde{\psi}_\tau(\epsilon_i) = \left(1 - \Phi\left(\frac{y_{i1} - x_{i1}^T\beta}{r_{i1}}\right), \ldots, 1 - \Phi\left(\frac{y_{iT} - x_{iT}^T\beta}{r_{iT}}\right)\right)^T$, $\Phi(\cdot)$ is the standard normal cumulative distribution, $r_{ik} = (x_{ik}^T \Omega x_{ik})^{1/2}$, Ω is a smoothing parameter and $W_i(\eta)$ is the working correlation matrix. The main difference between mean and quantile GEE is related to the estimation of the working correlation matrix $W_i(\eta)$. In the quantile approach, it is estimated from the regression residuals' signs. However, correlation is not a good measure of dependency between binary variables because it is bounded by their marginal frequencies. An unconstrained alternative is given by the odds ratio. Let $S_{it} = I(y_{it} \leq x_{it}^T \beta_\tau)$ be the residual sign of the i-th individual at time t and $S_t = (S_{1t}, \ldots, S_{nt})$ be the set of residual signs at time t. The odds ratio between S_z and S_u is obtained by

$$\eta_{zu} = \frac{P(S_z = 1, S_u = 1)/P(S_z = 0, S_u = 0)}{P(S_z = 0, S_u = 1)/P(S_z = 1, S_u = 0)}.$$

Let $\mathcal{A} = \{(S_z, S_u),\ z = 1, \ldots, T,\ u = 1, \ldots, T,\ z > u\}$ be the set of pairwise comparisons required to estimate all the odds ratios of the working correlation matrix $W_i(\eta)$. Consider the augmented dataset (V_z, V_u, z, u, c), where $(V_z, V_u) = \{(S_z, S_u) \in \mathcal{A}\}$ and $c = 1, \ldots, \binom{T}{2}$ indicates the pairwise comparison between V_z and V_z . For any working structure of $W_i(\eta)$, the respective set of odds ratios can be estimated simultaneously through an appropriate choice of the linear predictor in a logistic regression of V_z on V_u, $V_z|V_u \sim Be(\mu)$:

- Exchangeable: $logit(\mu) = \alpha + \eta V_u$;
- Toeplitz: $logit(\mu) = \alpha + \sum_{i=1}^{T-1} \eta_i I_{z-u=i} V_u + \sum_{i=1}^{T-1} I_{z-u=i}$;
- Unstructured: $logit(\mu) = \alpha + \sum_{i=1}^{\binom{T}{2}} \eta_i I_{c=i} V_u + \sum_{i=1}^{\binom{T}{2}} I_{c=i}$.

We illustrate the proposed method with data from a randomized trial on cognitive behavior therapy for treatment of obsessive compulsive disorder.

References

Fu, L. & Wang, Y.G. (2012). Quantile regression for longitudinal data with a working correlation model. *Computational Statistics and Data Analysis*, **8**, 2526–2538.

A weighted bootstrap procedure for divergence minimization problems

M. Broniatowski[1*]

LSTA, Université Pierre et Marie Curie, Paris, France; michel.broniatowski@upmc.fr.
[*] *Presenting author*

Keywords. *Minimum divergence estimation, Large deviation, Weighted bootstrap*

Sanov type results hold for some weighted versions of empirical measures, and the rates for those Large Deviation principles can be identified as divergences between measures, which in turn characterize the form of the weights. This correspondance is considered within the range of the Cressie-Read family of statistical divergences, which covers most of the usual statistical criterions. We propose a weighted bootstrap procedure in order to estimate these rates. To any such rate we produce an explicit procedure which defines the weights, therefore replacing a variational problem in the space of measures by a simple Monte Carlo procedure.

Inhomogeneous large-scale data: a call for (non-conventional) robust statistics

P. Bühlmann[1*]

[1] *Seminar for Statistics, ETH Zürich; buhlmann@stat.math.ethz.ch*
[*] *Presenting author*

Keywords. *Aggregation; Causal inference; High-dimensional regression; Magging; Maximin effects.*

Large-scale or "big" data usually refers to scenarios with potentially very many variables (dimension p) and very large sample size n. Such data is most often of "inhomogeneous" nature, i.e., neither being i.i.d. realizations from a distribution nor being generated from a stationary distribution. We highlight some new methodology and a corresponding aggregation algorithm to deal with such homogeneity issues [Meinshausen and Bühlmann, 2015, Bühlmann and Meinshausen, 2016]. For the special case with ordered data (e.g. time ordering in streaming data), we can incorporate recent approaches in high-dimensional change point detection [Leonardi and Bühlmann, 2016]. We provide statistical accuracy guarantees for computationally efficient methods, in scenarios where n and/or p are large, and we illustrate the methodology on real data examples. If time permits, we will also demonstrate the benefits of inhomogeneous data for causal inference [Peters et al., 2016].

References

Bühlmann P, Meinshausen N (2016) Magging: maximin aggregation for inhomogeneous large-scale data. Proceedings of the IEEE 104:126–135

Leonardi F, Bühlmann P (2016) Computationally efficient change point detection for high-dimensional regression. Preprint arXiv:1601.03704

Meinshausen N, Bühlmann P (2015) Maximin effects in inhomogeneous large-scale data. Annals of Statistics 43:1801–1830

Peters J, Bühlmann P, Meinshausen N (2016) Causal inference using invariant prediction: identification and confidence intervals. To appear in the Journal of the Royal Statistical Society: Series B (with discussion). Preprint arXiv:1501.01332

Efficient Use of EMR for Discovery Research

A. Chakrabortty[1], *J. Gronsbell*[1] *and* ***T. Cai***[1*]

[1] *Dept of Biostatistics, Harvard University*
abhi060988@gmail.com; jlg735@mail.harvard.edu; tcai@hsph.harvard.edu
[*] *Presenting author*

Keywords. *Classification; Electronic Medical Records Data; Model mis-specification; Robustness; Semi-supervised learning*

In clinical practice, patients with the same disease diagnosis often differ in outcomes and response to treatment. The ability to both classify and predict disease phenotypes would be a valuable asset in clinical decision-making. Large datasets containing both a wealth of clinical and experimental data now exist as a result of the increasing adoption of electronic medical records (EMR) linked with specimen bio-repositories. These datasets allow for data driven classification and prediction of sub-phenotypes and investigation of shared risk factors across a group of phenotypes. In this talk, I?ll discuss various statistical and informatics methods that illustrate both the challenges and potential opportunities that arise from analyzing EMR data. For example, obtaining validated phenotype information is a major bottleneck in EMR research, as it requires laborious medical record review. Thus gold standard labels are typically available only in a small training set nested in a large cohort. In contrast, data on the clinical predictors of the phenotype are often available on all subjects. To improve phenotype definition, we developed robust semi-supervised learning methods that can leverage such rich source of auxiliary information. These methods are illustrated with an EMR cohort of RA patients.

Robust Variable Selection for Functional Linear Regression

Guanqun Cao[1*] *and Yichen Qin*[2]

[1] *Auburn University, Auburn, Alabama 36849 USA; gzc0009@auburn.edu.*
[2] *University of Cincinnati, Cincinnati, Ohio 45221 USA; qinyn@ucmail.uc.edu.*
[*] *Presenting author*

Keywords. *Multiple functional predictors; Penalized estimation; Robustness; Variable selection.*

Functional linear regression has been widely used to explore the relationship between a scalar response and functional predictors. In this article, we consider the situation where multiple functional predictors are observed, but only a few of these predictors are actually useful in predicting the response. Several recent literatures investigate the variable selection in such model. For example, Gertheiss et al. [2013] used a penalized likelihood method that controls the sparsity of the model and the smoothness of the corresponding coefficient functions without considering any outliers. Our objective is to develop an outlying-resistant variable selection procedure to identify the important functional predictors and estimate the corresponding coefficient functions simultaneously even in presence of a significant proportion of contanimated observations.

We assume a functional linear regression, $Y_i = \alpha + \sum_{j=1}^{p} \int \beta_j(t) X_{i,j}(t)dt + \epsilon_i$. To estimate the parameter $\boldsymbol{\beta}(t) = (\beta_1(t), ..., \beta_p(t))^T$, we propose to minimize the following objective function,

$$\sum_{i=1}^{n} \phi\Big(Y_i - \alpha - \sum_{j=1}^{p} \int \beta_j(t) X_{i,j}(t)dt\Big) + \sum_{j=1}^{p} P_{\lambda,\tau}(\beta_j(t)),$$

where $\phi(r) = 1 - \exp(-r^2/c)$ is the exponential squared loss function [Wang et al., 2013] with tuning parameter c. Note that when $c \to \infty$, $\phi \approx r^2/c$, which corresponds to the traditional least square estimate. Due to its boundedness, ϕ can effectively limit the influence of outliers, therefore, provides robust estimate of the coefficient functions. Furthermore, we can show the $\sqrt{n}$-consistence of the proposed estimators. We further incorporate a penalty function $P_{\lambda,\tau}(\beta_j(t)) = \lambda(||\beta_j(t)||_2^2 + \tau||\beta_j''(t)||_2^2)^{1/2}$ and $||\beta_j(t)||_2^2 = \int \beta_j(t)^2 dt$, $\beta_j''(t) = \partial^2\beta_j(t)/\partial t^2$, which controls the sparsity of the model and smoothness of coefficients. To solve the above optimization problem, we first approximate the coefficient functions $\beta_j(t)$ using the B-spline basis functions $\mathbf{b}_j(t) = (b_{j1}(t), ..., b_{jq}(t))^T$, i.e., $\beta_j(t) \approx \sum_{r=1}^{q} \gamma_{jr} b_{jr}(t)$, where γ_{jr} are the corresponding basis coefficients. When the functional predictors, $X_{i,j}(\cdot)$, are observed without measurement errors and at an equally spaced dense grid of points, $\{t_{j,1}, \ldots, t_{j,N_j}\}$, then $\int \beta_j(t) X_{i,j}(t)dt$ can be approximated by the Riemann sum, i.e. $\int \beta_j(t) X_{i,j}(t)dt = \mathbf{Z}_{ij}^T \boldsymbol{\gamma}_j$, where $\mathbf{Z}_{ij} = (Z_{ij,1}, \ldots, Z_{ij,q})^T$, $Z_{ij,r} = (t_{j,l} - t_{j,l-1}) \sum_l X_{i,j}(t_{j,l}) b_{j,r}(t_{j,l})$ and $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jq})^T$. Hence, the functional linear model can be approximated by a typical linear regression model $Y_i \approx \alpha + \sum_{j=1}^{p} \mathbf{Z}_{ij}^T \boldsymbol{\gamma}_j + \epsilon_i$.

We conduct simulation to demonstrate the superior performance of the proposed method. We simulate data sets of the form $\{X_{i,1}(t), ..., X_{i,10}(t), Y_i\}$, $i = 1, ..., 1000$, where each covariate $X_{i,j}$ is observed on the set of 300 equidistant points in $(0, 300)$. In particular, the generating model

Contamination Ratio		0.00	0.10	0.20	0.30
SE	Proposed Method	1.80	1.99	2.35	2.78
	Least Square	1.58	29.48	80.78	159.76
TPR	Proposed Method	0.99	0.98	0.97	0.97
	Least Square	0.80	1.00	1.00	1.00
TNR	Proposed Method	0.78	0.80	0.82	0.85
	Least Square	1.00	0.01	0.00	0.00

Table 1: Comparison of performance of the proposed method and the least squared method at difference levels of contamination

is $Y_i = \alpha + \sum_{j=1}^{10} \int_0^{300} \beta_j(t) X_{i,j}(t) dt + \epsilon_i$, where $\epsilon_i \overset{i.i.d.}{\sim} (1-\delta)N(0, 0.01) + \delta N(0.01, 0.1)$ and δ is the contamination ratio. $\beta_1(t)$, $\beta_2(t)$, and $\beta_3(t)$ have Gamma-density like shape with effect sizes decreasing with increasing j, and $\beta_4(t)$ and $\beta_5(t)$ have exponential like shape with $\beta_5(t)$ being more linear. Only signals $j = 1, ..., 5$ are assumed to be relevant. We run 100 replications.

Table 1 presents the comparison of the proposed method and traditional least square method in terms of (1) square errors (SE), SE $= \sum_{j=1}^{10} \int_0^{300} (\beta_j(t) - \hat{\beta}_j(t))^2 dt$; (2) true positive rate (TPR); (3) true negative rate (TNR). As the figure shows, when there is no contamination, the proposed method outperforms the least square method slightly. As the contamination becomes more serious, the proposed method totally dominates least square method in all three categories. Therefore, we can see the clear advantage of the proposed method.

References

Gertheiss, J., Maity, A. and Staicu, A.-M. (2013). Variable Selection in Generalized Functional Linear Models, *Stat*, **2**, 86-101.

Wang, X., Jiang, Y., Huang, M. and Zhang, H. (2013). Robust Variable Selection with Exponential Squared Loss, *Journal of American Statistical Association*, **108**, 632-643.

Outlier detection and the distribution of residuals in robust regression

A. Cerioli[1*], *S. Salini*[2], *M. Riani*[1], *F. Laurini*[1] *and A. Ghiretti*[3]

[1] *Department of Economics, University of Parma, Italy; andrea.cerioli@unipr.it, mriani@unipr.it, fabrizio.laurini@unipr.it.*
[2] *Department of Economics Management and Quantitative Methods, University of Milan, Italy; silvia.salini@unimi.it.*
[3] *Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Italy; a.ghiretti@disia.unifi.it.*
[*] *Presenting author*

Keywords. *Forward Search; LTS; Masking and Swamping; Residuals; S-estimation*

1 Introduction

Recent work in robust statistics has focused on the attempt to reconcile the two enemy brothers of high-breakdown estimation: robustness against a large fraction of masked outliers and good statistical properties, comparable to those of classical estimators, when the normal model holds for all the data. From the point of view of estimation, the goal of this body of work has been the construction of estimators that can achieve both a high breakdown point and high efficiency at the normal distribution [see, e.g., Van Aelst et al, 2013]. From a diagnostic perspective, reaching satisfactory statistical properties under the normal model also implies good control of the number of false discoveries in situations of practical interest. There are many application fields, such as high-dimensional genomics, quality control and anti-fraud analysis, where such a property is highly desirable [see, e.g., Cerioli and Farcomeni, 2011, Cerioli and Perrotta, 2014]. However, high-breakdown techniques may produce a potentially large number of spurious outliers. The main target of the present work is to address the diagnostic behaviour of high-breakdown techniques at the normal model from a regression perspective, by considering a wide variety of alternative estimators and different approximations to the null distribution of the resulting robust residuals.

2 Framework and main results

Our basic diagnostic quantities in robust regression are the squared scaled residuals

$$\hat{s}_i^2 = \frac{\hat{\epsilon}_i^2}{\hat{\sigma}^2}, \qquad i = 1, \ldots, n, \tag{8}$$

where $\hat{\epsilon}_i$ is the estimated regression residual for observation i, $\hat{\sigma}^2$ is the model-based estimate of the error variance, and n is the sample size. Precise outlier identification is based on the asymptotic approximation

$$\hat{s}_i^2 \simeq \chi_1^2. \tag{9}$$

Also informal diagnostic methods, such as Q-Q plots of squared scaled residuals, rely on (9). However, the reference χ^2_1 distribution holds only in the limit and may provide poor approximations in small or moderate samples when parameters are estimated by high-breakdown techniques. Additional problems may occur due to the effect of alternative tuning choices in the algorithm used to compute the parameter estimates.

One goal of our work is to investigate to what extent the most popular high-breakdown regression methods provide accurate rules for outlier detection using the squared scaled residuals $\hat{s}^2_i$. We compute appropriate corrections when the null performance of the resulting procedure is poor. In particular, we place outlier detection in a testing scenario and we develop robust regression diagnostics that are able to control empirical test sizes at a prescribed level for all the procedures that we analyze. We also evaluate the loss of power that can be expected from our corrections under different contamination schemes and we show that this loss is often not dramatic. See Salini et al [2016] for details. A second goal of our work is to find simple and accurate approximations to the finite sample distribution of (8), thus extending the results of Cerioli [2010] to the case of regression. The availability of these approximations will provide more flexible outlier detection rules and more accurate diagnostics tools to be used, e.g., in Q-Q plots of squared scaled residuals.

References

Cerioli A (2010) Multivariate outlier detection with high-breakdown estimators. Journal of the American Statistical Association 105:147–156

Cerioli A, Farcomeni A (2011) Error rates for multivariate outlier detection. Computational Statistics and Data Analysis 55:544–553

Cerioli A, Perrotta D (2014) Robust clustering around regression lines with high density regions. Advances in Data Analysis and Classification 8:5–26

Salini S, Cerioli A, Laurini F, Riani M (2016) Reliable Robust Regression Diagnostics. International Statistical Review, in press, doi:10.1111/insr.12103

Van Aelst S, Willems G, Zamar RH (2013) Robust and efficient estimation of the residual scale in linear regression. Journal of Multivariate Analysis 116:278–296

A Robust Method (LTED) for Imputation Missing Values and Location Estimate

C. Chatzinakos[1*] *and G.Zioutas*[1]

[1] *Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece ; chatzin@auth.gr, zioutas@eng.auth.gr*

[*] *Presenting author*

Keywords. *Robust Statistic; Missing Values; Imputation; Location Estimate.*

Estimating the mean (location) of an incomplete dataset and filling in missing values with imputed values is a common problem in data analysis. There will always be times when something goes wrong, resulting in gaps in data. Some statistical procedure will not work as well, or at all, with some data missing. Deleting an entire row implies that we loose available information for the analysis, for that reason is preferable to impute these missing values. Different approaches can be used to handle the missing values, such as Last Observation Carried Forward (LOCF), Multiple Imputations, Simple Mean Imputation, Simple Median Imputation, Expectation Maximization Algorithm Approach. But these methods can be greatly affected by the presence of outliers in the data. This paper introduces a new robust imputation method, for imputing missing values in data, Least Trimmed Euclidean Distance Imputation (LTED-IM), and finally estimating the location of the data. The method based on an iterative algorithm which use the Euclidean norm and propose a solution technique for the resulting combinatorial optimization problem, based on a necessary condition, that results in a high convergent local search algorithm. Simulation studies on a real and artificial data indicate that our proposed method outperforms existing methods in accuracy and robustness.

References

Hron, K. and Templ, M. and Filzmoser, P. (2010). Imputation of Missing Values for Compositional Data Using. *Comput. Stat. Data Anal.*, **54**, 3095–3107.

Park, M., Lai, D., Du, X.L., Delclos, G.P., and Moye, L.A. (2015). General Linear Models in a Missing Outcome Environment of Clinical Trials Incorporating with Splines for Time-Invariant Continuous Adjustment. *American Journal of Biostatistics*, **5**, 7–51.

Siddiqui, O.I. (2015). Methods for Computing Missing Item Response in Psychometric Scale Construction. *American Journal of Biostatistics*, **5**, 1–6.

The Beta- Fisher Snedecor Distribution with applications to Cancer Remission Data

A.U. Chukwu[1*] *and K.A. Adepoju*[1]

[1] *Department of Statistics, University of Ibadan, Ibadan, Nigeria. unnachuks2002@yahoo.co.uk*
[*] *Presenting author*

Keywords. *Fisher-Snedecor distribution; Beta-F distribution; Outlier; Maximum likelihood method.*

n this paper a new four-parameter generalized version of the Fisher Snedecor distribution called Beta- F distribution is introduced. The comprehensive account of the statistical properties of the new distributions was considered. Formal expressions for the cumulative density function, moments, moment generating function and maximum likelihood estimation as well as its Fisher information were obtained. The flexibility of this distribution as well as its robustness using cancer remission time data was demonstrated. The new distribution can be used in most applications where the assumption underlying the use of other life time distributions is violated.

Robust semiparametric estimation of single-index binary choice models under missclassification

P. Čížek[1*]

[1] *Department of Econometrics & OR, Tilburg University, The Netherlands; P.Cizek@uvt.nl.*
[*] *Presenting author*

Keywords. *Binary-choice model; Breakdown point; Indirect inference; Misclassification; Single-index model.*

In this paper, binary-response models are considered in the situations when the binary dependent variable contains wrong, that is, misclassified responses for some observations. Misclassified binary responses have been found in many surveys (e.g., Meyer and Mittag, 2014) and the effects of misclassification in the binary-choice models have therefore been studied both in the parametric (Hausman et al., 1998) and nonparametric context (Lewbel, 2000). There is also related literature concentrating on the sensitivity to erroneous observations in binary-choice regression (e.g., Cizek, 2008). Both the traditional maximum likelihood estimation (MLE) as well as the various semiparametric estimators, being typically based locally on averages, least squares, or maximum likelihood criteria, are sensitive to misclassification of the responses (Neuhaus, 1999; Cizek, 2008), especially if the misclassified observations have unlikely values of covariates.

Accounting for misclassification as in Hausman et al. (1998), for instance, however requires assumptions concerning the misclassification probability and Meyer and Mittag (2014) document that these methods do not perform well in real-data applications relative to the standard MLE estimator that ignores the presence of misclassification. We therefore attempt to develop estimation methods that also allow for modelling of misclassification probability, but are rather insensitive to deviations from the specified misclassification process and more generally to any kind of data contamination.

For this purpose, a new class of semiparametric estimators of single-index binary choice models is introduced and based on the bias correction by means of indirect inference. The estimators in this class can be based on a variety of auxiliary estimators such as nonlinear least squares or nonlinear least absolute deviation estimators. For some suitably chosen auxiliary estimators, the proposed estimator can offer a unique combination of properties: root-n convergence, robustness to index-based heteroscedasticity, robustness to misclassified leverage points, and very good finite-sample performance compared to existing semiparametric estimators. Both consistency and asymptotic normality of the proposed semiparametric estimators are derived. The finite-sample performance is compared with existing semiparametric and robust estimators of binary-choice models by means of Monte Carlo simulations.

References

Čížek, P. (2008) Robust and efficient adaptive estimation of binary-choice regression models, *Journal of the American Statistical Association*, **103**, 687–696.

Hausman, J. A., Abrevaya, J. & Scott-Morton, F. M. (1998) Misclassification of the dependent variable in a discrete-response setting, *Journal of Econometrics*, **87**, 239–269.

Lewbel, A. (2000) Identification of the binary choice model with misclassification, *Econometric Theory*, **16**, 603–609.

Meyer, B. & Mittag, N. (2014) Misclassification in binary choice models. NBER Working Paper No. 20509.

Neuhaus, J. M. (1999) Bias and efficiency loss due to misclassified responses in binary regression, *Biometrika*, **86**, 843–855.

A Comparison of the L_2 Minimum Distance Estimator and the EM-Algorithm when Fitting k-Component Univariate Normal Mixtures

B.R. Clarke[1*], *T. Davidson*[2] *and R. Hammarstrand*[1]

[1] *Mathematics and Statistics, School of Eng. and I.T., Murdoch University, Murdoch, Western Australia, 6150; B.Clarke@murdoch.edu.au, R.Hammarstrand@murdoch.edu.au.*
[2] *Australian Bureau of Statistics, Perth, Western Australia, 6000; tom.davidson@abs.gov.au*
[*] *Presenting author*

Keywords. *EM algorithm; Minimum distance estimation; Robust estimation; Monte Carlo simulation.*

The method of maximum likelihood using the EM-algorithm for fitting finite mixtures of normal distributions is the accepted method of estimation ever since it has been shown to be superior to the method of moments. Recent books testify to this. There has however been criticism of the method of maximum likelihood for this problem, the main criticism being when the variances of component distributions are unequal the likelihood is in fact unbounded and there can be multiple local maxima. Another major criticism is that the maximum likelihood estimator is not robust. Several alternative minimum distance estimators have since been proposed as a way of dealing with the first problem. This paper deals with one of these estimators which is not only superior due to its robustness, but in fact can have an advantage in numerical studies even at the model distribution. Importantly, robust alternatives of the EM-algorithm, ostensibly fitting t distributions when in fact the data are mixtures of normals, are also not competitive at the normal mixture model when compared to the chosen minimum distance estimator. It is argued for instance that natural processes should lead to mixtures whose component distributions are normal as a result of the Central Limit Theorem. On the other hand data can be contaminated because of extraneous sources as are typically assumed in robustness studies. This calls for a robust estimator.

Relevant references include Clarke and Heathcote [1978], Clarke [1989], Clarke [1983], Clarke [2000], Clarke and Heathcote [1994], and Clarke et al. [2016].

References

Clarke, B. R. (1983). Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations. *Annals of Statistics*, **11**, 1196-1205.

Clarke, B. R. (1989). An unbiased minimum distance estimator of the proportion parameter in a mixture of two normal distributions. *Statist. Prob. Lett.*, **7**, 275-281.

Clarke, B. R. (2000). A review of differentiability in relation to robustness with an application to seismic data analysis. *PINSA*, **66A**, 467-482.

Clarke, B. R., Davidson, T. & Hammarstrand, R. (2016). A Comparison of the L_2 Minimum Distance Estimator and the EM-Algorithm when Fitting $\boldsymbol{k}$-Component Univariate Normal Mixtures. *Statistical Papers*, http://link.springer.com/article/10.1007/s00362-016-0747-x

Clarke, B. R. & Heathcote, C. R. (1978). Comment on "Estimating mixtures of normal distributions and switching regressions" by Quandt, R.E. and Ramsey, J.B.. *J. Am. Statist. Ass.*, **73**, 749-750.

Clarke, B. R. & Heathcote, C. R. (1994). Robust estimation of $\boldsymbol{k}$-component univariate normal mixtures. *Ann. Inst. Statist. Math.*, **46**, 83-93.

Outlier Robust Filtering

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

[1] *KU Leuven, ORSTAT; ruben.crevits@kuleuven.be, christophe.croux@kuleuven.be*
[*] *Presenting author*

Keywords. *Time Series; State-space Model; Particle Filter.*

In time series analysis state space models are very popular. Often it is interesting to sequentially estimate the distribution of the hidden states given all observations available. That problem is called filtering. For linear Gaussian state space models the filtering distribution can be exactly computed with the Kalman filter. For nonlinear or nongaussian models one can rely on particle methods to estimate an approximation of the filtering distribution. Standard optimal Bayesian filtering with the Kalman or bootstrap particle filter is not an outlier robust method. Several proposals to robustify the filters for general nonlinear nongaussian state space models have been made (Maiz et al. [2012]; Calvet et al. [2015]). We introduce an Outlier Robust Filter which has the lowest possible efficiency cost for a certain impact bound of a new observation on the estimated filtering density. The performance of the new filter is compared with the existing methods. It turns out that our filter performs very well with contaminated time series, while still performing adequately with clean time series.

References

Maiz, C. S., Molanes-Lopez, E., Miguez, J. & Djuric, P. (2012). A particle filtering scheme for processing time series corrupted by outliers *IEEE transactions on Signal Processing*, **60**(9), 4611–4627.

Calvet, L., Czellar, V., & Ronchetti, E. (2015). Robust Filtering *Journal of the American Statistical Association*, **110**(512), 1591–1606.

Robust forecasting of short time series

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

[1] *KU Leuven, Faculty of Economics and Business; christophe.croux@kuleuven.be, ruben.crevits@kuleuven.be*
[*] *Presenting author*

Keywords. *Robustness; Smoothing; Time Series; Trends*

Simple forecasting methods, such as exponential smoothing, are very popular in business analytics. This is not only due to their simplicity, but also because they perform very well, in particular for shorter time series. Incorporating trend and seasonality into an exponential smoothing method is standard. Many real time series, including the short ones, show seasonal patterns that should be exploited for forecasting purposes. Including a trend or not may be less clear. For instance, weekly sales (in units) may show an increasing trend, but the sales will not grow to infinity. Here, the damped trend model gives an outcome. Damped trend exponential smoothing gives excellent results in forecasting competitions.

In a highly cited paper, Hyndman et al. [2008] develop an automatic forecasting method, available as an R-package and very easy to use. Within a class of 15 different types of exponential smoothing methods, the best one (according to an information criterion) is selected for a given time series, and a prediction is made. The damped trend model is one of these 15 types. In our paper we provide a robust version of this automatic forecasting procedure. The approach we take generalizes Gelper et al. [2010]. We show that the robust automatic forecasting method performs well on simulated data, and outperforms the non-robust approach in presence of additive outliers . But for out-of-sample forecasting of real time series, results are less convincing.

References

Gelper, S., Fried, R., & Croux C. (2010). Robust forecasting with exponential and Holt-Winters smoothing. *Journal of Forecasting*, **29**,285-300.

Hyndman, R. & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, **27**, 1–22.

Method of moments approach to nuisance scale estimation for Huber M-quantiles

J. Dawber[1*], *N. Tzavidis*[2] *and N. Salvati*[3]

[1] *University of Wollongong, New South Wales, Australia; jdawber@uow.edu.au.*
[2] *Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK; n.tzavidis@soton.ac.uk*
[3] *Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa, Italy; salvati@ec.unipi.it*
[*] *Presenting author*

Keywords. *M-estimation; Influence functions; Robust inference; Scale estimation.*

M-quantiles are a quantile-type generalization of M-estimation using influence functions. M-quantile regression using a Huber influence function provides an adjustable middle ground between quantile and expectile regression through the use of a tuning constant. This adjustability determines the level of robustness of the regression model. Such M-quantile regression models have been used as robust alternatives to random effects models especially in small area estimation. To ensure that these Huber M-quantiles are scale invariant a nuisance scale parameter is required. This scale parameter must also be robust to avoid compromising the robustness of the location parameter. The most commonly used robust scale estimator is the median absolute deviation (MAD), which was originally proposed when Huber M-quantile regression was introduced. Another scale estimator using maximum likelihood was also suggested recently which uses an Asymmetric Least Informative (ALI) distribution which is derived from the M-quantile loss function. This talk introduces a third scale estimator approach to Huber M-quantile regression using the method of moments based on the ALI distribution, and shows why it is better than the other two approaches. The appropriateness of each approach is assessed in a range of different contexts using simulations and real data.

Robust parallelized inference based on wide consensus

E. del Barrio[1*]

IMUVA, Universidad de Valladolid; tasio@eio.uva.es
[*] *Presenting author*

Keywords. *Trimmed barycenters; Wide consensus; Wasserstein distance; Trimmed distributions; Parallelized inference.*

We develop a general theory to address a consensus-based combination of estimations in a parallelized or distributed estimation setting. Taking into account the possibility of very discrepant estimations, instead of a full consensus we consider a "wide consensus" procedure. The approach is based on the consideration of trimmed barycenters in the Wasserstein space of probability distributions on $\mathbb{R}^d$ with finite second order moments. We include general existence and consistency results as well as characterizations of barycenters of probabilities that belong to (non necessarily elliptical) location and scatter familes. On these families, the effective computation of barycenters and distances can be addressed through a consistent iterative algorithm. Since, once a shape has been chosen, these computations just depend on the locations and scatters, the theory can be applied to cover with great generality a wide consensus approach for location and scatter estimation or for obtaining confidence regions.

References

Álvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A. and Matrán, C. (2015). Wide consensus for parallelized inference. ArXiv:1511.05350v1.

Robust semi-parametric estimators: missing data and causal inference

E. Cantoni[1] *and* ***X. de Luna***[2*]

[1] *Research Center for Statistics and Geneva School of Economics and Management, University of Geneva, Switzerland; Eva.Cantoni@unige.ch*
[2] *Department of Statistics, USBE, Umeå University, Umeå, Sweden; xavier.deluna@umu.se.*
[*] *Presenting author*

Keywords. *bounded influence function; inverse probability weighting; double robust.*

We consider situations where we aim at estimating location and scale parameters from a distribution law of interest, from which a random sample has been drawn. We introduce semi-parametric estimators, which are able to deal simultaneously with two common challenges within this general context: (i) not all observations from the random sample intended are available (incomplete data due to dropout, selection, potential outcomes framework), and (ii) some of the available observations in the sample may be contaminated (generated by a nuisance distribution, outliers). Under an assumption of ignorable missingness, popular semi-parametric estimators of the parameters of interest are augmented inverse probability weighted (AIPW, doubly robust) estimators (e.g., Tsiatis [2007]). They use two auxiliary models, one for the missingness mechanism, and another for an outcome of interest, both given observed covariates. AIPW estimators are then robust to misspecification of one of these two models (but not both simultaneously - a so called double robustness property). We introduce versions of AIPW, which provide, moreover, robustness to contamination of the distribution of interest. Asymptotic properties are described and finite sample results are presented. We motivate the need for robust AIPW estimators with a follow up study on BMI combining data from an intervention study and population wide record linked data.

References

Tsiatis, A. (2007). Semiparametric theory and missing data. Springer Science & Business Media, Berlin.

Smooth Plus Rough Variation of Random Functions: the Interplay Between Rank, Resolution, and Scale

M-H. Descary and V.M. Panaretos*

Section de Mathématiques, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; marie-helene.descary@epfl.ch, victor.panaretos@epfl.ch

Keywords. *Covariance operator; Functional Data Analysis; Functional PCA; Matrix completion.*

Functional data analyses typically proceed by smoothing, followed by functional PCA. This paradigm implicitly assumes that any roughness is due to nuisance noise. Nevertheless, relevant functional features such as time-localised or short scale variations may indeed be rough. These will be confounded with the smooth components of variation by the smoothing/PCA steps, potentially distorting the parsimony and interpretability of the analysis. We will explore how both smooth and rough variations can be recovered on the basis of discretely observed functional data. Assuming that a functional datum arises as the sum of two uncorrelated components, one smooth and one rough, we develop identifiability conditions for the estimation of the two corresponding covariance operators. We construct nonlinear estimators of the smooth and rough covariance operators and their spectra via matrix completion, and establish their consistency and rates of convergence. We then use them to recover the smooth and rough components of each functional datum, effectively producing separate functional PCAs for smooth and rough variation.

References

Descary, M-H. & Panaretos, V.M. (2015). Smooth plus rough variation of random functions: the interplay between rank, resolution, and scale. *Submitted.*

GLM-lasso explores the cosmos

J. Diaz Rodriguez[1*], *S. Sardy*[1] *D. Eckert*[2] *and S. Paltani*[2]

[1] *Mathematics Section, University of Geneva; jairo.diaz@unige.ch, sylvain.sardy@unige.ch*
[2] *Astronomy Department, University of Geneva; dominique.eckert@unige.ch, stephane.paltani@unige.ch*
[*] *Presenting author*

Keywords. *Poisson Inverse Problem; High-Dimensions; GLM-lasso; Galaxy Clusters; Abel Transform.*

An important problem in Cosmology is to reconstruct the mass distribution of a galaxy cluster based on telescope images for instance using X-ray Multi-Mirror techniques. Such images include point sources spread behind the cluster which although there are of importance, also affect the estimation of the mass distribution. Therefore it is important to robustify the reconstruction of the galaxy cluster by identifying at the same time the point sources, outliers in the response. We cast this challenging problem into a linear inverse problem involving blurring, Abel and wavelet transforms for Poisson counts. Owing to the fact that the high-dimensional (more than two millions) estimands are sparse, we employ GLM-lasso. We address the issues of selecting two regularization parameters and of solving the high-dimensional optimization problem. We show the strength of our method on simulated and real images.

Multivariate Moment Based Extreme Value Index Estimators

M. Heikkilä[1], ***Y. Dominicy***[2*] *and P. Ilmonen*[1]

[1] *Department of Mathematics and Systems Analysis, Aalto University School of Science, Espoo, Finland; matias.heikkila@aalto.fi, pauliina.ilmonen@aalto.fi.*
[2] *Université libre de Bruxelles, Solvay Brussels School of Economics and Management, ECARES, Brussels, Belgium; yves.dominicy@ulb.ac.be.*
[*] *Presenting author*

Keywords. *Extreme value index; Elliptical distribution; Hill estimator; Moment estimator; Mixed moment estimator.*

Modelling extreme events is of paramount importance in various areas of science – biostatistics, climatology, finance, geology, and telecommunications, to name a few. Most of these application areas involve multivariate data. Estimation of the extreme value index plays a crucial role in modelling rare events. An affine invariant multivariate generalization of the well known Hill estimator was recently proposed by Dominicy et al. [2015]. Their idea is based on the distance between a tail probability contour and the observations outside this contour. To select the extreme points, the observation outside the tail probability contour, they use the Minimum Covariance Determinant (MCD) approach. However, the Hill estimator is only suitable for heavy tailed distributions. As in Dominicy et al. [2015], we consider estimation of the extreme value index under the assumption of multivariate ellipticity. We provide affine invariant multivariate generalizations of the moment estimator and the mixed moment estimator. These estimators are suitable for both: light and heavy tailed distributions. Asymptotic properties of the new extreme value index estimators are derived under known location and scatter, and the effect of replacing true location and scatter by estimates is examined in a thorough simulation study.

References

Dominicy, Y., Ilmonen, P. & Veredas, D. (2015). Multivariate Hill Estimators. Forthcoming in *International Statistical Review.*

Measuring Correlation in the Presence of Spikes

D.J. Dupuis[1*]

[1] *Department of Decision Sciences, HEC Montréal; 3000, chemin de la Côte-Sainte-Catherine, Montréal (Québec) Canada H3T 2A7; debbie.dupuis@hec.ca*
[*] *Presenting author*

Keywords. *Exponential weighting; Robust correlation.*

The cost of electricity varies across the zones of the New York State electric system. While fair and open access to the electrical grid is sought, we show that that residents currently do not equally benefit, or suffer, from price changes. Upcoming major investments in the grid offer an opportunity to rectify these inequalities, but only if we understand the price-change propagation dynamics for the current underlying infrastructure. We study these dynamics, estimating the partial correlations between changes in electricity prices in connected zones. We develop and investigate a robust exponentially weighted correlation estimator that performs well in the presence of electricity price spikes and can track a rapidly-changing time-varying correlation. We show that price-change partial correlations are mostly positive, but can also be negative, and provide new insight into price-change dynamics within the grid that cannot be extracted from the price-setting algorithm or obtained from available transmission capability data.

Robustness in practice

P. Filzmoser[1]∗

[1] *Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria; P.Filzmoser@tuwien.ac.at*
∗*Presenting author*

Keywords. *Robust statistics; Applications; R; robustbase.*

1 Abstract

During the last decades, a lot of effort has been made in the development of robust statistical methods. Nowadays, many robust counterparts to traditional non-robust methods are available, and there is a lot of knowledge about the theoretical properties of these methods [Maronna et al., 2006]. Moreover, robust methods have been widely implemented in software packages, for example in the statistical software environment R [R Core Team, 2015]. Thus, the availability of those methods could even lead to the hypothesis that robust tools will at some point replace the traditional tools. However, so far we cannot see this tendency.

Thinking about regression analysis, several robust counterparts to least-squares regression have been developed, and they have been implemented for example in the R package `robustbase` [Rousseeuw et al., 2015]. We usually teach our students that if no outliers are present in the data (and if the usual data requirements are met), robust regression leads to about the same answer as least-squares regression; however, in presence of outliers, least-squares regression results could be heavily spoiled. This is often demonstrated with simulations and with "real" data, which are frequently "toy" data sets, or data sets which served as test data also in other scientific papers. One could thus ask if real "real" data show the same features. Is there still a striking difference in the analysis when using non-robust and robust methods? Is it even possible to use the code implementation of the robust methods in real applications without any difficulties? Do the parameters of the routines require special adaptations or tuning? Is the runtime of these algorithms comparable to the runtime of algorithms for traditional methods?

We will present some use-cases and focus on different robust methods for outlier detection, regression and discrimination. Available routines in R are tested, and results (if any) are compared to non-robust counterparts. Recommendations for practical use are given, and possible directions for future developments are provided.

References

Maronna, R., Martin, D. & Yohai, V. (2006). Robust Statistics: Theory and Methods. John Wiley & Sons, Chichester.

R Core Team (2015). R Foundation for Statistical Computing. R: A Language and Environment for Statistical Computing. Vienna, Austria. `https://www.R-project.org/`.

Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M. & Maechler, M. (2015). robustbase: Basic Robust Statistics. R package version 0.92-5, `http://CRAN.R-project.org/package=robustbase`.

Detection of Genomic Imprinting Effects for Qualitative Traits on the X Chromosome

W.K. Fung[1*]

[1] *Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong; wingfung@hku.hk*
[*] *Presenting author*

Keywords. *Affected daughter; Imprinting effects; Parental-asymmetry test; Qualitative traits; X chromosome.*

Genomic imprinting is an important epigenetic factor which can lead to complete or partial deactivation of the maternally or paternally inherited copy of a gene. Some complex human diseases such as diabetes, obesity, Beckwith-Wiedemann, Prader-Willi and Angelman syndromes are demonstrated to be connected with imprinting. Methods for detecting imprinting effects have been developed primarily for autosomal markers. However, no method is available to test for imprinting effects on the X chromosome. Therefore, it is necessary to suggest methods for detecting such imprinting effects. In this talk, the parental-asymmetry test on the X chromosome (XPAT) is first developed to test for imprinting for qualitative traits in the presence of association, based on family trios each with both parents and their affected daughter. Then, we propose 1-XPAT to handle parent-daughter pairs. By simultaneously considering family trios and parent-daughter pairs, C-XPAT is constructed to test for imprinting. Moreover, we extend the proposed methods to accommodate nuclear families having multiple daughters of which at least one is affected. The performance of the proposed methods are investigated by simulations. Simulation results demonstrate that the proposed methods control the size well, whether or not Hardy-Weinberg equilibrium holds. The proposed methods are applied to analyze the rheumatoid arthritis data.

On Robustness for Spatial Data

A. García-Pérez[1*] *and Y. Cabrero-Ortega*[2]

[1] *Departamento de Estadística, I.O. y C.N., Universidad Nacional de Educación a Distancia (UNED), Paseo Senda del Rey 9, 28040-Madrid, Spain; agar-per@ccia.uned.es*
[2] *C.A. UNED-Madrid, Spain; ycabrero@madrid.uned.es*
[*] *Presenting author*

Keywords. *Robustness;Spatial outliers;GAMs;GIS;Spatial Influence Function.*

1 Identification of local outliers using Robust GAMs and Geographical Information Systems

In the first part of the paper we propose two different methodologies for detecting possible local outliers, that we call *hotspots*. The first one is based in using Geographical Information Systems (GIS) considering a map for the observations where the *heights* of the ground have been replaced by the data $z(s) = z(x, y)$ of the observed variable. We do a TIN interpolation and then, with GRASS through QGIS, we compute the slopes of the triangles thus obtained, concluding with the detection of outlying slopes with GRASS again. Areas with big slopes will indicate zones of possible outliers. This method works with a Big amount of Data. These ideas have been used (with some variants) by Felicísmo [1994].

The second technique consists in fitting a robust Generalized Additive Model (GAM) to the observations. Then we do the previous process (interpolation plus detection of outlying slopes) to the residuals of this robust fit where the Longitude, x and the Latitude, y, are used as covariates in the model. We use QGIS again because the statistical package R can be run inside, obtaining so, layers that can overlapped on other pre-existing ones. This second detecting method has been used in Liu et al. [2001]. Here we extend their ideas considering a more general model, a GAM one, because this is the common model considered in a spatial data fit.

After we have obtained a reduced set of *hotspots*, we compute the probability of obtaining such outlying slopes according to a classical GAM. Those hotspots for which we obtain a small probability, will be labeled as local outliers. We apply these techniques to Guerry data, as Filzmoser et al. [2014] did, obtaining the same conclusions than they.

2 A Spatial Influence Function

We know that the Hampel's Influence Function IF of a functional T at a model F, is a very useful tool to measure the effect of an outlier z on an estimator (on a functional, really). For instance, the Hampel's influence function of the Mean is the function of z, $IF(z; \text{Mean}, F) = z - \mu$, where $\mu = \int u\, dF(u)$. With this function we see that the outlier z influences the Mean

linearly and in an unbounded way. Nevertheless, in this standard definition of the IF, the coordinates of the outlying observation z are omitted and, as we have seen, these could be very important.

Using the three-dimensional notation of this paper, we can rewrite the IF for the Mean as $IF(z(s_0); T, F) = IF(h(x_0, y_0); T, F) = z(s_0) - \mu$, where h is a smooth function used in the GAM fit that is expressed in terms of a basis.

If $z(s)$ is a local outlier and not a global one, it will affect the estimator because, al least, it is a none locally expected value but, because it is not outside of the bulk of the data in the OZ axis, it will pass unnoticed for the Hampel's Influence Function IF, i.e, the IF does not measure the influence of local outliers that are not global, on estimators (or functionals) because these local outliers act on the x-y plane and not on the OZ axis. Nevertheless, they are influential observations and they do affect the value of the estimator.

We need, then, to modify (extend) the IF to take into account both local and global outliers because, all of them affect the estimator. This new influence function will be called *Spatial Influence Function*, SIF. It is defined using the ideas of the previous section and its main properties studied.

References

Felicísmo, A.M. (1994). Parametric statistical method for error detection in digital elevation models. *ISPRS Journal of Photogrammetry and Remote Sensing*, **49**, 29–33.

Filzmoser P., Ruiz-Gazen, A. & Thomas-Agnan, C. (2014). Identification of local multivariate outliers. *Statistical Papers*, **55**, 29–47.

Liu, H., Jezek, K.C. & O'Kelly, M.E. (2001). Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and GIS. *International Journal of Geographical Information Science*, **15**, 721–741.

Tukey g-and-h Random Fields

G. Xu[1] *and* ***M.G. Genton***[2*]

[1] *Binghamton University, USA; gang@math.binghamton.edu*
[2] *King Abdullah University of Science and Technology, KSA; marc.genton@kaust.edu.sa*
[*] *Presenting author*

Keywords. *Heavy tails; Non-Gaussian random field; Skewness; Spatial outliers; Spatial statistics.*

We propose a new class of trans-Gaussian random fields named Tukey g-and-h (TGH) random fields to model non-Gaussian spatial data. The proposed TGH random fields have extremely flexible marginal distributions, possibly skewed and/or heavy-tailed, and, therefore, have a wide range of applications. The special formulation of the TGH random field enables an automatic search for the most suitable transformation for the dataset of interest while estimating model parameters. An efficient estimation procedure, based on maximum approximated likelihood, is proposed and an extreme spatial outlier detection algorithm is formulated. The probabilistic properties of the TGH random fields, such as second-order moments, are investigated. Kriging and probabilistic prediction with TGH random fields are developed along with prediction confidence intervals. The predictive performance of TGH random fields is demonstrated through extensive simulation studies and an application to a dataset of total precipitation in the south east of the United States.

References

Xu, G. & Genton, M. G. (2015). Efficient maximum approximated likelihood inference for Tukey's g-and-h distribution. *Computational Statistics & Data Analysis*, **91**, 78–91.
Xu, G. & Genton, M. G. (2016). Tukey g-and-h random fields. Manuscript.

Robust Wald-Type Tests under Random Censoring

Abhik Ghosh[1*], *Ayanendranath Basu*[2] *and Leandro Pardo*[3]

[1] *Department of Bio-Statistics, Institute of Basic Medical Sciences, University of Oslo, Norway; abhik.ghosh@medisin.uio.no.*
[2] *Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata, India; ayanbasu@isical.ac.in*
[3] *Department of Statistics and Operations Research I, Complutense University, Madrid, Spain; lpardo@mat.ucm.es.*
[*] *Presenting author*

Keywords. *Robsut Test of Hypothesis; Random Censoring; Wald-Type Test; M-Estimator; Density Power Divergence; Influence Function.*

Randomly censored survival data are frequently encountered in several applied sciences including biomedical sciences and reliability applications; analyzing these data are very important to obtain inferences for all these applications. The underlying censoring mechanism may or may not be known and several semi-parametric approaches are available in the literature for the analysis of such data. However, these semi-parametric procedures are not as efficient as the classical maximum likelihood procedures which can be implemented under some parametric assumptions of the distribution and random censoring scheme. However, as in the case of several other types of data, the maximum likelihood methods for censored data are also highly non-robust with respect to the outlying observations. Since it is common to have some potential outliers in several real-life applications with survival data, either due to erroneous input or some unknown underlying mechanism (like some data points might come from a different group), suitable robust procedures having good efficiency are very useful in practice.

Under the context of survival data with random censoring, there are very few approaches of robust estimation that consider the fully parametric set-up to gain more efficiency; Wang (1999) and Basu et al. (2006) proposed two such estimation approaches. The first one develops the general theory of M-estimation under the set-up of randomly censored data while the second one considers a particular M-estimator based on the density power divergence (DPD; Basu et al., 1998) exhibiting highly efficient and robust performances. Recently, Ghosh and Basu (2014) have extended these approaches of M-estimation and the minimum DPD estimation for the cases where we have some stochastic covariates along with the randomly censored response. However, there is no literature on the robust test of hypothesis under the fully parametric set-up with censored data, although this is very important for any real-life inference problems.

In this paper, we propose the Wald-type test for testing common statistical hypothesis for randomly censored data. We consider both the simple and composite hypotheses and use the minimum DPD estimators under fully parametric set-up; the high efficiency of the estimator used yield high power for our testing procedure. We have also derived the general theory and properties of our proposed Wald-type test statistics for the general M-estimators and demonstrated the usefulness of its particular member, the Minimum DPD estimator based test statistics, through appropriate theoretical and numerical illustrations. In this way, we have proposed a consistent estimator of the asymptotic variance of the M-estimator based on the sample data without any assumption on the form of censoring scheme. We have also presented the robustness of our proposal theoretically through suitable influence function analysis and

numerically through appropriate simulations and real data examples. Finally, we have briefly indicated how to extend our proposed Wald-type test statistics for the two sample problem with random censoring and stochastic covariates.

References

Basu, S., Basu, A., and Jones, M. C. (2006). Robust and efficient parametric estimation for censored survival data. *Annals of the Institute of Statistical Mathematics*, **58(2)**, 341–355.

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85(3)**, 549–559.

Ghosh, A. & Basu, A. (2014). Robust and Efficient Parameter Estimation based on Censored Data with Stochastic Covariates. *ArXiv Pre-print*, arXiv:1410.5170 [math.ST].

Wang, J. L. (1999). Asymptotic Properties of M-Estimators Based on Estimating Equations and Censored Data. *Scandinavian journal of statistics*, **26(2)**, 297–318.

Adapting Robust Estimators to Iterative Parameter Estimation and Model Selection in Linear Mixed Models

F. Gokalp Yavuz[1*] *and O. Arslan*[2]

[1] *Yildiz Technical University; Department of Statistics A1-19, 34220, Istanbul, Turkey; fulyagokalp@gmail.com.*
[2] *Ankara University; oarslan@ankara.edu.tr.*
[*] *Presenting author*

Keywords. *Scad; Robust mixed models, Model selection.*

Linear mixed models (LMM) are extended regression types models and several methods are used to estimate parameters for longitudinal data, clustered data and cross-sectional data. The existence of outliers and the complication of the detection of these outliers for multi-dimensional data are required to use robust methods for parameter estimations in LMM. Mixed effect models are robustified with two different methods: One is to use heavy-tailed distributions as alternatives to the normal distribution, and the other is to use robust estimation methods. Peng and Lu (2012) introduce a simple iterative procedure for estimating parameters as an alternative to the optimization required methods such as Expectation Maximization and Newton-Raphson for LMM. However, their method is based on the classical LS estimator so that the resulting estimators for the parameters of LMM will be sensitive to the outliers. In this study, we extend Peng and Lu (2012)'s iterative methods with robust initial parameters as a contribution to the field of model selection in LMM. In addition to parameter estimations, LMM is required to implement model selection procedures. The smoothly clipped absolute deviation penalty function (SCAD), which is a non-concave penalized likelihood used to shrinkage the coefficients of unimportant variables to 0, is preferred to implement parameter estimation and model selection simultaneously for this study. Procedure adaptation results are tested with simulation studies.

References

Peng, H. & Lu, Y. (2012). Model selection in linear mixed effect models. *Journal of Multivariate Analysis*, **109**, 109–129.

Analytics in Insurance: Opportunities and Profit Drivers

Y.-L. Grize[1*]

[1] *Pricing Nonlife, Baloise Insurance; yves-laurent.grize@baloise.ch*
[*] *Presenting author*

Keywords. *General insurance; Pricing; Analytics; Business intelligence; Statistics in industry.*

Insurance companies are known to be "slow movers". Indeed the potential impact of analytics on the insurance business has been known for years, but it is only now that most insurers are starting to "feel the heat". I will first describe in general terms the main profit drivers of analytics for insurers and then illustrate specifically their impact using real business applications at Baloise. Finally I will briefly mention the resulting technological changes currently taking place at Baloise emphasizing thereby the importance of company culture.

Robust Ranking Using Heavy-tailed Prior Distributions

H. He[1], *T. Kenney*[1] *and* ***H. Gu***[1*]

[1] *Department of Mathematics and Statistics, Dalhousie University:*
hao.he@dal.ca, tkenney@mathstat.dal.ca, hgu@dal.ca
[*] *Presenting author*

Keywords. *Ranking; Prior Distributions; Posterior Mean.*

Ranking is the problem of ordering a collection of random variables, based on observation. The aim is to determine which variable has the higher mean. This can be complicated when the standard errors associated with different data points are different.

Ranking of these variables usually depends fundamentally on assumptions of how the true values are distributed — that is, the prior distribution. Posterior mean is a popular criterion for this sort of ranking, and has been applied to a number of real-life problems of this type. For selection purposes, using posterior mean maximises the expected value of the means of selected variables. It is possible to adapt this approach to use a different loss function that could more accurately reflect the relevant criteria for ranking and selection. Indeed there are examples of application of more general loss functions to this problem.

In most research on this topic, little work has gone into the question of robustness to choice of prior. In many cases, a parametric form for the prior is simply chosen (seemingly arbitrarily) and applied. Little work has been done on questioning whether the prior fits the data well, and if it does not, what the consequences are. This is different from a lot of model misspecification cases, because ranking is mostly focussed on the tail. Even if a certain prior fits most of the data very well, if it fits the tail badly there can be serious implications for ranking based on the posterior distribution.

In a simple example of this, we simulated a sample with a heavy-tailed prior, and attempted to rank the samples by posterior mean, using a normal prior. In this simulation, the posterior mean ranking discounted the points with large standard error, even if the observed values were also very large. The reason for this was that the observed values were too extreme under the normal distribution, so the posterior distribution too heavily discounted the observed values.

We study a range of prior distributions — ranging from light-tailed to heavy-tailed, both for simulation of MLE and estimation of posterior mean. We compare the results across a range of measures — Firstly, we look at the overall MSE of the posterior means, to confirm that using the correct distribution gives the best results, and that the difference between using too heavy a tail and too light a tail is relatively small. We then look at the MSE for the top 5% and the top 1%. Here we see that using too heavy a tail causes less harm than using too light a tail. The key point is that if the tail is too heavy, the results will end up close to the MLE, so there is a limit to how bad the results can be. Using a tail that is too light, there is no limit to how bad the results might be. We also examine the average of the means of selected variables, selecting the top 1% and top 5% by posterior mean under each prior. Again, this shows that using too heavy a prior has less potential harm than using too light a prior.

We also examine parameter estimation for the prior distribution. In terms of overall estimation of posterior means, the parameters should be as close to the best fitting values as possible. However, for estimating the posterior means of the largest points, in cases where the true prior distribution is more heavy-tailed than the prior distribution used, we show that it can be advantageous to use parameter estimates that are not the best-fitting ones. For example, using a larger variance than the true prior distribution can reduce the MSE for the points with largest true mean.

Tools for outlier detection inspired by topology

M. Heikkilä[1*]

[1] *Aalto University School of Science; matias.heikkila@aalto.fi*
[*] *Presenting author*

Keywords. *Outlier Detection, Topological Data Analysis, Nonlinear Structure*

Conventional outlier detection methods often fail in the presence of significant nonlinear structure in the data: Consider a set of points densely distributed on the unit circle and add a single point at the origin. For example, by simply considering the distances to the sample mean, the point at the origin is misclassified.

In a seminal paper [Carlsson, 2009], Carlsson demonstrated the feasibility of computational topology in the qualitative analysis of multivariate data: An approach that is naturally suited for data with arbitrarily complicated structure. We discuss prospects of similar ideas in outlier detection.

References

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society* **46**, **2**, 255–308

Estimating the number of clusters in OTRIMLE robust Gaussian mixture clustering

C. Hennig[1*], *P. Coretto*[2]

[1] *Department of Statistical Science, University College London; c.hennig@ucl.ac.uk.*
[2] *Department of Economics and Statistics, University of Salerno; pcoretto@unisa.it.*
[*] *Presenting author*

Keywords. *Mixture model; parametric bootstrap; Optimally Tuned Robust Improper Maximum Likelihood.*

Coretto and Hennig (2015a, 2015b) developed the method of Robust Improper Maximum Likelihood (RIMLE; "OTRIMLE" stands for "Optimally Tuned RIMLE") for clustering based on a Gaussian mixture model but allowing for some observations that could not reasonably be assigned to any cluster.

RIMLE is based on modelling the data (from p-dimensional Euclidean space) as i.i.d. generated by a Gaussian mixture model with an additional improper mixture component, namely a uniform distribution with density level c over the whole Euclidean space ("noise component"), and to fit such an improper distribution by (pseudo-)maximum likelihood. This allows for a smooth classification of points as outliers/noise or being generated by one of the clusters, i.e., the Gaussian mixture components.

The OTRIMLE is defined by minimising a "Mahalanobis criterion". This amounts to choosing c in such a way that the distribution of Mahalanobis distances to the corresponding cluster mean of the portion of the data classified as non-outlying is optimally approximated by a χ^2-distribution, as should be the case if the non-outliers came indeed from a Gaussian mixture.

Coretto and Hennig (2015a, 2015b) assume the number of mixture components k as fixed.

In this presentation we explore model diagnostic and estimation of the number of clusters by the following parametric bootstrap principle: we generate many datasets from the Gaussian mixture (non-outlying) part of the estimated mixture, using the estimated parameters, and we compare the distribution of the values of the Mahalanobis criterion mentioned above to the value achieved for the dataset under study. This allows us to see which numbers of clusters yield models that are consistent with the data. This is inspired by Davies's (1995) Data Features and can be applied to more general model selection and diagnostic problems in robust model-based clustering.

References

Coretto, P. & Hennig, C. (2015a). Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust Gaussian clustering. arXiv:1406.0808. To appear in *Journal of the American Statistical Association.*

Coretto, P. & Hennig, C. (2015b). A consistent and breakdown robust model-based clustering method. arXiv:1309.6895.

Davies, P. L. (1995). Data Features. *Statistica Neerlandica*, **49**, 185–245.

Robust and sparse multiclass classification by the optimal scoring approach

I. Hoffmann[1*], *P. Filzmoser*[1], *C. Croux*[2]

[1] *Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Wiedner Hauptstrasse 8-10, 1040 Vienna, Austria; irene.hoffmann@tuwien.ac.at, p.filzmoser@tuwien.ac.at*
[2] *Faculty of Business and Economics, KU Leuven, Leuven, Belgium; christophe.croux@econ.kuleuven.be*
[*] *Presenting author*

Keywords. *Multiclass discriminant analysis; Outlier detection; Variable selection;*

1 Introduction

In regression analysis a great variety of so-called sparse methods have been developed, which perform simultaneous model estimation and variable selection due to restrictions on the coefficient estimate. Thereby estimation precision can be increased and the selected models are easier to interpret. This is especially useful for data sets with a large number of predictor variables.

1.1 Sparse LTS regression

Sparse least trimmed squares (LTS) introduced by Alfons et al. [2013] is a robust regression estimator with a lasso penalty (L_1 norm penalty on the coefficient estimate), which has good robustness properties and a fast algorithm.

1.2 Optimal Scoring

Several formulations for linear discriminant analysis (LDA) exist leading to the same discriminant rules. Optimal scoring being one of them recasts the classification problem into a regression framework and iteratively models the class-membership as continuous variable. Let Q be smaller than the number of groups G, commonly $Q = G - 1$. Then solve for $q = 1, ..., Q$

$$\min_{\boldsymbol{\beta}_q, \boldsymbol{\theta}_q} \{\|\boldsymbol{Y}\boldsymbol{\theta}_q - \boldsymbol{X}\boldsymbol{\beta}_q\|^2\} \quad \text{s.t.} \quad \frac{1}{n}\boldsymbol{\theta}_q^T \boldsymbol{Y}^T \boldsymbol{Y} \boldsymbol{\theta}_q = 1, \quad \boldsymbol{\theta}_q^T \boldsymbol{Y}^T \boldsymbol{Y} \boldsymbol{\theta}_l = 0 \quad \forall l < q. \tag{10}$$

where $\boldsymbol{X}$ is the data matrix with n observations and p variables and $\boldsymbol{Y}$ an $n \times G$ matrix of dummy variables coding the class membership of the observations. Adding an L_1 penalty for $\boldsymbol{\beta}_k$ to the minimization problem leads to sparse discriminant analysis as proposed by Clemmensen et al. [2011].

2 Methodology

We propose to recast the classification problem into a robust regression framework by optimal scoring in order to obtain a robust discriminant model. The minimization problem for $\boldsymbol{\beta}_q$ and $\boldsymbol{\theta}_q$ is solved iteratively. For fixed $\boldsymbol{\theta}_q$ (a vector of random values in the first iteration) $\boldsymbol{\beta}_q$ is estimated robustly by a sparse fast LTS algorithm (using starting observations representing all classes). Then $\boldsymbol{\theta}_q$ is calculated using $\boldsymbol{X}\boldsymbol{\beta}_q$ as response but excluding the observations detected as outliers by sparse LTS in the former step. Robust LDA is then applied to $(\boldsymbol{X}\boldsymbol{\beta}_1, ..., \boldsymbol{X}\boldsymbol{\beta}_Q)$. For the selection of the sparsity parameter cross validation is performed and a robust misclassification rate (excluding potential outliers) is used to decide for the best model.

3 Evaluation

The proposed algorithm is evaluated on simulated data with few relevant variables and a large number of noise variables. Outliers are included to demonstrate the stability of the model estimation. False negative and false positive rates for the selected variables and the detected outliers are reported. The prediction performance is compared to classical sparse LDA and to robust LDA evaluated only on the relevant variables (oracle estimator).

References

Alfons, A., Croux, C. & Gelper, S. (2013). Sparse least trimmed squares regression for analysing high-dimensional large data sets. *The Annals of Applied Statistics*, **7(1)**, 226–248.

Clemmensen, L., Hastie, T., Witten, D. & Ersboll, B. (2011). Sparse discriminant analysis. *Technometrics*, **53**, 406–413.

Acknowledgement: This work is supported by the Austrian Science Fund (FWF), project P 26871-N20.

Sparse PCA for high-dimensional data with outliers

M. Hubert[1*], *T. Reynkens*[1], *E. Schmitt*[1,2] *and T. Verdonck*[1]

[1] *Department of Mathematics, KU Leuven, Leuven, Belgium; mia.hubert@kuleuven.be, tom.reynkens@kuleuven.be, tim.verdonck@kuleuven.be*
[2] *Protix, 5107 NC Dongen, The Netherlands; eric.schmitt@protix.eu*
[*] *Presenting author*

Keywords. *Dimension reduction; Outlier detection; Robustness*

A new sparse PCA algorithm is presented which is robust against outliers. The approach is based on the ROBPCA algorithm which generates robust but non-sparse loadings.

The construction of the new ROSPCA method is detailed, as well as a selection criterion for the sparsity parameter. An extensive simulation study and a real data example are performed, showing that it is capable of accurately finding the sparse structure of datasets, even when challenging outliers are present.

In comparison with a projection pursuit based algorithm, ROSPCA demonstrates superior robustness properties and comparable sparsity estimation capability, as well as significantly faster computation time.

See Hubert et al. [2016] for full details.

References

Hubert, M. & Reynkens, T. & Schmitt, E. & Verdonck, T. (2016). Sparse PCA for high-dimensional data with outliers. *Technometrics*, in press.

Modeling Homophily in ERGMs for Bipartite Networks

Rashmi P. Bomiriya[1], *Shweta Bansal*[2] *and* ***David R. Hunter***[3*]

[1] *Remote Sensing Metrics Asia (Pvt) Ltd.; rashmi.bomiriya@gmail.com.*
[2] *Georgetown University; shweta@sbansal.com.*
[3] *Pennsylvania State University; dhunter@stat.psu.edu*
[*] *Presenting author*

Keywords. *Exponential-family random graph model; Bipartite network; Homophily*

Bipartite networks, in which edges only exist between two disjoint sets of nodes, represent an important tool for modeling processes such as affiliations, collaborations, and co-location. Frequently, we would like to model the propensity of similar nodes to form links among themselves, a property referred to as homophily. Modeling homophily in a bipartite network is complicated by the prohibition of direct ties between nodes in the same subset. This paper introduces a method for modeling homophily in the commonly used exponential-family random graph model (ERGM) framework.

If we allow the $n \times n$ matrix Y to encode the status of all the binary edges of the network, where Y_{ij} equals 0 or 1 according to whether the (i, j)th edge is absent or present, then the basic ERGM may be written

$$P_\theta(Y = y) = \frac{\exp\{\sum_{i=1}^{p} \theta_i s_i(y)\}}{\kappa(\theta)}, \quad y \in \mathcal{Y}, \tag{11}$$

where $s_1(y), \ldots, s_p(y)$ are user-defined statistics measured on the network y and we denote the vector of all network statistics by $s(y)$. When covariates X should be included in the model, we may add X to the notation and write $s(y, X)$, where we allow these statistics to depend on any available known covariates. The parameters $\theta_1, \ldots, \theta_p$ are the corresponding unknown coefficients to be estimated, $\mathcal{Y}$ is the set of all allowable networks, and $\kappa(\theta)$ is merely a normalizer necessary to ensure that Equation (11) defines a legitimate probability distribution.

We argue that the "natural" approach to modeling homophily in a bipartite network, in which we incorporate the total count of matching two-stars into an ERGM as one of the $s(y, X)$ statistics, might be problematic due to potential degeneracy issues. We introduce a new set of model terms for the ergm package in R designed to model homophily while mitigating such issues. We demonstrate that these model terms can be expressed in a curved exponential family form, then discuss real-world applications.

Outliers in the power exponential model

M. Ibazizen[1*] *and M. Mehiri*[2]

[1] *Université de Poitiers, Laboratoire de Mathématiques et Applications, France; mohamed.ibazizen@univ-poitiers.fr*
[2] *Université Mouloud Mammeri, Tizi-Ouzou, Algérie; m2mehiri@yahoo.fr*
[*] *Presenting author*

Keywords. *Kurtosis; Outliers; Power Exponential Model.*

Outliers are observations that do not follow the pattern of the majority of the data and show extremeness relative to some basic model.

For the $p-$ variate normal distribution $\mathcal{N}_p(\mu, \Sigma)$ as a model, where $\mu \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{pxp}$, Σ positive definite, Becker and Gather (1997) gave the general concept of an α outlier : an α outlier with respect to $\mathcal{N}_p(\mu, \Sigma)$ is defined as an element of the set

$$out(\alpha, \mu, \Sigma) = \{x \in \mathbb{R}^p : (x-\mu)'\Sigma^{-1}(x-\mu) > \chi^2_{p,1-\alpha}\} \tag{1}$$

for $\alpha \in]0, 1[$, with $\chi^2_{p,1-\alpha}$ denoting the $(1-\alpha)$ quantile of the χ^2_p distribution. We can write

$$P[X \in out(\alpha, \mu, \Sigma)] = \alpha \quad \text{for} \quad X \sim \mathcal{N}_p(\mu, \Sigma).$$

For usual choices of α ($\alpha = 0.05$, $\alpha = 0.10$), this reflects the idea of an outlier being an observation that is rather unlikely under the assumed model and also situated "outside the main mass of the distribution".
Our objective is to extend this approach for a more general distribution having higher or lower tails than those of the normal distribution, namely "the Multivariate Power Exponential distribution". The definition and properties of this distribution are given in Gömez & al. (1998).
A random vector $X = (X_1, ..., X_p)'$, with $p \geq 1$, has a p-dimensional power exponential distribution, with μ, Σ and β parameters ($\beta > 0$) if its density function is

$$f(x; \mu, \Sigma, \beta) = k|\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[(x-\mu)'\Sigma^{-1}(x-\mu)\right]^{\beta}\right\}, \tag{2}$$

with $k = \dfrac{p\Gamma(\frac{p}{2})}{\pi^{\frac{p}{2}}\Gamma(1+\frac{p}{2\beta})2^{1+\frac{p}{2\beta}}}$.

This distribution, denoted by $PE_p(\mu, \Sigma, \beta)$, is a member of the family of elliptically symmetric distributions. Its mean, covariance matrix and kurtosis are :

$$E(X) = \mu, \quad Var(X) = \frac{2^{\frac{1}{\beta}}\Gamma(\frac{p+2}{2\beta})}{p\Gamma(\frac{p}{2\beta})}\Sigma, \quad \gamma_2(X) = \frac{p^2\Gamma(\frac{p}{2\beta})\Gamma(\frac{p+4}{2\beta})}{\Gamma^2(\frac{p+2}{2\beta})} - p(p+2) \tag{3}$$

The parameter β indicates, in terms of kurtosis, the disparity from (2) to the normal case, which corresponds to $\beta = 1$.

Consider the quadratic form : $R(x, \mu, \Sigma) = (x-\mu)'\Sigma^{-1}(x-\mu)$.

The distribution of the r.v. R^β is $\Gamma(\frac{1}{2}, \frac{p}{2\beta})$ (see Gómez & al. (1998)).
By analogy with the multivariate normal case, let us define the set

$$out(\alpha, \mu, \Sigma, \beta) = \left\{ x \in \mathbb{R}^p : R(x, \mu, \Sigma) > \left[\gamma_{(\frac{1}{2}, \frac{p}{2\beta})}(1-\alpha) \right]^{1/\beta} \right\}$$

Likewise, $\gamma_{(\frac{1}{2}, \frac{p}{2\beta})}(1-\alpha)$ is here the $(1-\alpha)$ quantile of the Gamma distribution with parameters $\frac{1}{2}$ and $\frac{p}{2\beta}$.
An α outlier with respect to the power exponential distribution $PE_p(\mu, \Sigma, \beta)$ is then an element of the set $out(\alpha, \mu, \Sigma, \beta)$.

Recall that the parameter β in the $PE_p(\mu, \Sigma, \beta)$ model is related to the kurtosis which reflects the tails of the distribution.

In this note, we perform a numerical study to confirm that an observation may be an outlier for a given model but not be for yet another neighbor of the first model, such that the same identification rule led to different conclusions in the two models considered to be relatively close. The most interesting fact is that the number of outliers depends heavily on the model assumption.

References

Barnett, T., Lewis, T. (1994). Outliers in Statistical Data. J. Wiley & Sons, New York.

Becker, C., Gather, U. (1999). The Masking Breakdown Point of Multivariate Outlier Identification Rules. *JASA*, **94(447)**, 947–55.

Gómez, E., Gómez-Villegas, M.A., Marïn, J.M. (1998). A Multivariate Generalization of the Power Exponential Family of Distributions. *Communications in Statistics-Theory and Methods* **27(3)**, 589–600.

Lopuhaä, H. & Rousseeuw, P. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Annals of Statistics*, **19**, 229–248.

Maronna, R., Martin, D. & Yohai, V. (2006). Robust Statistics: Theory and Methods. John Wiley & Sons, Chichester.

Testing the heteroscedastic error structure in quantile varying coefficient models

I. Gijbels[1], ***M. A. Ibrahim***[2*] *and A. Verhasselt*[2]

[1] *Department of Mathematics, KU Leuven, Belgium; Irene.Gijbels@wis.kuleuven.be.*
[2] *Censtat, Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Belgium; mohammed.ibrahim@uhasselt.be, anneleen.verhasselt@uhasselt.be.*
[*] *Presenting author*

Keywords. *Consistency; Heteroscedasticity; Likelihood-ratio test; Penalized splines; Quantile regression.*

Regression quantiles generalize naturally mean regression for Gaussian linear models, while substantially out-performing the least-squares estimator over a wide class of non-Gaussian error distributions. Varying coefficient models (VCMs) are considered, they are an extension of classical linear regression models that allow the coefficients to depend on other variables. Assuming longitudinal data, the regression coefficients are allowed to vary with time.

We consider VCMs with various structures for the variance of the errors (the variability function) in order to allow for heteroscedasticity. The coefficient functions and the variability function are estimated with P-splines. Consistency of the proposed estimators is proved. Further, likelihood-ratio-type tests are considered for comparing the variability functions. The performance of the testing procedure is illustrated on simulated and real data.

Regression quantile and averaged regression quantile processes

J. Jurečková[1*]

[1] *Charles University in Prague, Faculty of Mathematics and Physics, Czech Republic; jurecko@karlin.mff.cuni.cz*

Keywords. *Regression quantile; Averaged regression quantile; Asymptotic representation; Brownian bridge.*

Consider the linear regression model $\mathbf{Y}_n = \mathbf{X}_n\beta + \mathbf{U}_n$ with observations $\mathbf{Y}_n = (Y_1, \dots, Y_n)^\top$, i.i.d. errors $\mathbf{U}_n = (U_1, \dots, U_n)^\top$ with an unknown distribution function F, and unknown parameter $\beta = (\beta_0, \beta_1, \dots, beta_p)^\top$. The $n \times (p+1)$ matrix $\mathbf{X} = \mathbf{X}_n$ is known and $x_{i0} = 1$ for $i = 1, \dots, n$ (i.e., β_0 is an intercept). The α-regression quantile $\widehat{\beta}_n(\alpha)$ is a solution of the minimization $\sum_{i=1}^n \rho_\alpha(Y_i - \mathbf{x}_i^\top \mathbf{b}) := \min$ with respect to $\mathbf{b} = (b_0, \dots, b_p)^\top \in \mathbb{R}^{p+1}$, where $\mathbf{x}_i^\top$ is the i-th row of $\mathbf{X}_n$, $i = 1, \dots, n$ and $\rho_\alpha(z) = |z|\{\alpha I[z > 0] + (1-\alpha) I[z < 0]\}$, $z \in \mathbb{R}^1$. The population counterpart of $\widehat{\beta}_n(\alpha)$ is the vector $\widetilde{\beta}(\alpha) = (\beta_0 + F^{-1}(\alpha), \beta_1, \dots, \beta_p)^\top$.
Koenker & Bassett [1978] used a linear programming algorithm for calculation of $\widehat{\beta}_n(\alpha)$. They also used the following dual algorithm as a computational device:

$$\begin{aligned} \mathbf{Y}_n^\top \widehat{\mathbf{a}} &:= \max \\ \text{under the constraint } \ \mathbf{X}_n^\top \widehat{\mathbf{a}} &= (1-\alpha)\mathbf{X}_n^\top \mathbf{1}_n, \ \ \widehat{\mathbf{a}} \in [0,1]^n, \ 0 \le \alpha \le 1. \end{aligned} \tag{12}$$

The components of the optimal solution of (12), $\widehat{\mathbf{a}}_n(\alpha) = (\hat{a}_{n1}(\alpha), \dots, \hat{a}_{nn}(\alpha))^\top$, were named the *regression rank scores* by Gutenbrunner & Jurečková [1992], who used them for construction of the rank tests in the linear model.
Consider the regression quantile process $\mathbf{Z}_n = \left\{\mathbf{Z}_n(\alpha) = n^{1/2}(\widehat{\beta}_n(\alpha) - \widetilde{\beta}_n(\alpha))\right\}$ and the ordinary quantile process $Z_n^{(0)} = \left\{Z_n^{(0)}(\alpha) = n^{1/2}(F_n^{-1}(\alpha) - F^{-1}(\alpha))\right\}$, $0 < \alpha < 1$ where F_n^{-1} is the empirical quantile function. Under conditions on $\mathbf{X}$ and F, $\mathbf{Z}_n \xrightarrow{\mathcal{D}} \left(f \circ F^{-1}\right)^{-1} \mathbf{Q}^{-1} \mathbf{W}_{(p)}^*$ as $n \to \infty$, where $\mathbf{W}_{(p)}^*$ is a vector of p independent Brownian bridges on $(0,1)$ (see Gutenbrunner & Jurečková [1992]).

The scalar statistic

$$\bar{B}_n(\alpha) = \bar{\mathbf{x}}_n^\top \widehat{\beta}_n(\alpha), \quad \bar{\mathbf{x}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_{ni}$$

is called the *averaged regression α-quantile*. It is scale equivariant and regression equivariant. It was shown by Jurečková [2016] that, for every fixed n, $\bar{B}_n(\alpha)$ equals to a linear combination of $p+1$ components of vector of observations $\mathbf{Y}$, corresponding to the optimal base of the linear programming (12). Asymptotically is $\bar{B}_n(\alpha)$ equivalent to the $[n\alpha]$-quantile of the location model, as was proved in Jurečková & Picek [2014]. Under mild conditions, the process $\bar{\mathcal{B}}_n$ admits the asymptotic representation

$$\bar{\mathcal{B}}_n = \frac{1}{\sqrt{n} f(F^{-1}(\alpha))} \sum_{i=1}^n \Big(I[U_i > F^{-1}(\alpha)] - (1-\alpha)\Big) + o_p^*(1).$$

Moreover, $\bar{\mathcal{B}}_n \xrightarrow{\mathcal{D}} (f \circ F^{-1})^{-1} W^*$ as $n \to \infty$ where W^* is the Brownian bridge on (0,1); here $o_p^*(1)$ means uniform convergence on $(\varepsilon, 1-\varepsilon)$ for any $\varepsilon \in (0, 1/2)$.
The weak convergence of process $\bar{\mathcal{B}}_n$ applies also to its various functionals, what in turn leads to various useful applications under nuisance regression. Moreover, its representation coincides with the representation of the ordinary quantile process, hence $\bar{\mathcal{B}}_n$ is asymptotically equivalent to the same. This, by Csörgő & Révész [1978], is in turn approximated by a sequence of Brownian bridges.

References

Csörgő, M. & Révész, P. (1978). Strong approximation of the quantile process. *Ann. Statist.* **6**, 882–894.

Gutenbrunner, C. & Jurečková, J. (1992). Regression rank scores and regression quantiles. *Ann.Statist.* **20**, 305–330.

Jurečková J. (2016). Finite sample behavior of averaged extreme regression quantile. *Extremes* **19**, 41–49.

Jurečková, J. & Picek, J. (2014). Averaged regression quantiles. In: *Contemporary Developments in Statistical Theory* (S. N. Lahiri et al. (eds.), Springer Proceedings in Mathematics & Statistics 68, Chapter 12, pp.203–216.

Koenker, R. & Bassett, G. (1978). Regression quantiles. *Econometrica* **46**, 33–50.

Nonparametric Quantile Regression via a New MM Algorithm

B. Kai[1*], *M. Huang*[2] , *W. Yao*[3] *and Y. Dong*[4]

[1] *Department of Mathematics, College of Charleston, Charleston, SC, USA; kaib@cofc.edu*
[2] *School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China; huang.mian@mail.shufe.edu.cn*
[3] *Department of Statistics, University of California, Riverside, Riverside, CA, USA; weixin.yao@ucr.edu*
[4] *Department of Statistics, Temple University, Philadelphia, PA, USA; ydong@temple.edu*
[*] *Presenting author*

Keywords. *Quantile regression; MM algorithms; Descent property; Nonparametric regression.*

Nonparametric quantile regression is a widely used statistical model in many research fields and applications. However, its optimization is very challenging due to non-differentiable objective functions. Though MM algorithms adapted from Hunter and Lange [2000] can be implemented for nonparametric quantile regression, it suffers from several limitations such as instability, non-smoothness, and even discontinuity. In this work, we propose a new MM algorithm for nonparametric quantile regression. The proposed algorithm simultaneously updates the quantile function and yields a more stable and smooth estimate of the quantile function. We also show that the monotone descent property of the new algorithm maintains in an asymptotic sense. The finite sample performance of the new algorithm is assessed via Monte Carlo simulation studies.

References

Hunter, D.R. & Lange, K. (2000). Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics*, **9**, 60–77.

Robust Tests for Exponential Distribution Data based on Repeated Median Estimator

D. Karagöz[1*]

[1] *Department of Statistics, Hacettepe University, Turkey email: deryacal@hacettepe.edu.tr*
[*] *Presenting author*

Keywords. *ANOVA, Exponential Distribution, Brown-Forsythe, Modified Brown-Forsythe, Repeated Median.*

In this study, we use the robust Brown-Forsythe and robust Modified Brown-Forsythe test statistics proposed by [Karagöz & Saraçbaşı(2016)] for the one-way ANOVA under heteroscedasticity for the exponentially distributed data with outliers. In order to get these robust test statistics, repeated median robust estimator of the mean and variance of exponential distribution are obtained. We consider balanced and unbalanced sample sizes with homogeneous and heterogeneous variances. The simulations results show up that the proposed robust tests have a good performance.

1 Introduction

ANOVA is one of the most commonly used models in many fields such as medicine, engineering, psychology, sociology, etc. In general, the main interest of this paper is testing the homogeneity of group means of non-normal data. One-way ANOVA is based on assuming the normality of the observations and the homogeneity of group variances. The classical ANOVA uses the F-test statistic. If the assumptions of normality and homogeneity of variances are invalid and there are also outliers are present, classical ANOVA does not give accurate results. Therefore, test statistics based on robust methods should be used instead of the classical ANOVA.

The main theme of the paper is the estimation of the parameters λ in the exponential density function $f_\lambda(x) = \lambda \exp[-(\lambda x)]$, where $x, \lambda > 0$ and the parameter λ is called a rate parameter. The reciprocal $1/\lambda$ is known as the scale parameter.

The one-way classification of analysis of variance for non-normal data with heteroscedastic variance has been studied for a long time. In the case of disruption of assumptions, instead of ANOVA F-test, many test statistics have been developed such as Welch, Brown-Forsythe and Modified Brown-Forsythe. The difference of this study from the other studies are the RBF and $RMBF$ test statistics based on the repeated median estimator of mean and variance of the exponential distribution are considered. Robust Brown-Forsythe test statistics is given by [Karagöz & Saraçbaşı(2016)] as following

$$RBF = \frac{\sum_{i=1}^{k} n_i(\hat{\mu}_{ri.} - \hat{\mu}_{r..})^2}{\sum_{i=1}^{k}(1 - n_i/N)\hat{\sigma}_{ri}^2}. \tag{13}$$

RBF test statistic has F_{k-1,v_r} distribution with $k-1$ and v_r degrees of freedom. v_r is defined

as $v_r = \frac{[\sum_{i=1}^{k}(1-n_i/N)\hat{\sigma}_{ri}^2]^2}{\sum_{i=1}^{k}(1-n_i/N)^2\hat{\sigma}_{ri}^4/(n_i-1)}$.

Robust modified Brown-Forsythe test statistics is given by [Karagöz & Saraçbaşı(2016)] as following

$$RMBF = \frac{\sum_{i=1}^{k} n_i(\hat{\mu}_{ri.} - \hat{\mu}_{r..})^2}{\sum_{i=1}^{k}(1 - n_i/N)\hat{\sigma}_{ri}^2} \tag{14}$$

$RMBF$ test statistic has $F_{v_{r_1},v_r}$ distribution with v_{r_1} and v_r degrees of freedom. The numerator degrees of freedom v_{r_1} is defined as

$$v_{r_1} = \frac{[\sum_{i=1}^{k}(1 - n_i/N)\hat{\sigma}_{ri}^2]}{\sum_{i=1}^{k}\hat{\sigma}_{ri}^4 + [\frac{\sum_{i=1}^{k} n_i\hat{\sigma}_{ri}^2}{N}]^2 - 2\frac{\sum_{i=1}^{k} n_i\hat{\sigma}_{ri}^4}{N}}$$

In the above equations, $\hat{\mu}$ and $\hat{\sigma}$ are the robust repeated median estimators of mean and variance of exponential distribution.

2 Conclusion

In this study we work on the robust test statistics RBF and RGBF for the exponentially distributed data by using the robust repeated median estimators. In the simulation study, using different experimental designs, type I error risks of the robust test statistics for the exponential distribution are obtained for k=4,8 groups. We consider balanced and unbalanced sample sizes with homogeneous and heterogeneous variances. The simulations results show up that the proposed robust tests have a good performance.

References

Boudt, K., Caliskan, D. & Croux, C. (2011). Robust explicit estimators of weibull parameters,*Metrika*, **73**, 187—209.

Brown, M. & Forsythe, A. (1974). Small sample behavior of some statistics which test the equality of several means. *Technometrics*, **16**, 129—132.

D. Karagöz,D. & Saraçbaşı, T., (2016). Robust Brown-Forsythe and Robust Modified Brown-Forsythe ANOVA Tests Under Heteroscedasticity for Contaminated Weibull Distribution. *Revista Colombiana de Estadística* , **39-1**, 17-32.

Maronna, R., Martin, D. & Yohai, V. (2006). Robust Statistics: Theory and Methods. John Wiley & Sons, Chichester.

Model Selection in Multiple Linear Regression

D.N. Kashid[1*]

[1] *Department of Statistics, Shivaji University, Kolhapur, India(MS), 416004; dnk_stats@unishivaji.ac.in*
[*] *Presenting author*

Keywords. *Outlier; Multicollinearity; M-estimator;Robust criterion.*

Model selection is the most persuasive problem in multiple linear regression. The performance of least squares parameter estimation based model selection methods are reasonably not well in the presence of outlier, multicollinearity, leverage, etc. due to non-robustness of least squares estimator. To overcome problems in the data, various alternative parameter estimation procedures like M-estimator, Rank-estimator, Ridge, Liu, Liu-M, JRM, LRM estimators are proposed by the researchers. Using some alternative parameter estimation methods, researchers have proposed robust model selection methods [Ronchetti & Staudte, 1994, Sommer & Huggins, 1996]. However, only a few model selection methods cope with simultaneous occurrence of two or more problem in the data [Jadhav et al., 2014]. In this article, we study the existing methods and propose a new model selection method resistant to simultaneously occurrence of outlier and multicollinearity. A superiority of a proposed method is illustrated through simulation study.

References

Ertas, H., Kaciranlar, S. & Guler, H. (2015). Robust Liu-type Estimator for Regression Based on M-estimator. *Communications in Statistics-Simulation and Computation*, DOI: 10.1080/03610918.2015.1045077.

Jadhav, N. H., Kashid, D. N. & Kulkarni, S. (2014). Subset selection in multiple linear regression in the presence of outlier and multicollinearity. *Statistical Methodology*, **19**, 44–59.

Ronchetti, E. & Staudte, R.G. (1994). A Robust Version of Mallows' C_p. *Journal of the American Statistical Association*, **89(426)**, 550–559.

Sommer, S. & Huggins, R. M. (1996). Variables Selection Using the Wald Test and a Robust C_p. *Applied Statistics*, **45(1)**, 15–29.

The Adequate Bootstrap — A new Method for Measuring Model Uncertainty

T. Kenney[1*] *and H. Gu*[1]

[1] *Department of Mathematics and Statistics, Dalhousie University: tkenney@mathstat.dal.ca, hgu@dal.ca*
[*] *Presenting author*

Keywords. *Model Adequacy; Robust Inference; Bootstrap.*

Model adequacy testing is less ubiquitous than it ought to be. Any parametric analysis should be acompanied by model adequacy testing. However, in practice, this is not always the case. There are several reasons for this. One particular reason is the fundamental disconnect between what is tested, and what we would like to test. The usual approach to testing model adequacy is to set up an hypothesis test. The null hypothesis is "Model M is the true model." However, when we consider the famous words of Box [1976]: *"All models are wrong. Some models are useful."* we see the problem with this approach. We already know that the null hypothesis is false, and our model is wrong. What we want to know from our test is whether the model is useful. A model might still be useful even if we have enough data to reject it.

We consider the context where we are confident that our model reflects some part of the underlying process, but some further process (such as data contamination or sampling bias) results in observed data that do not follow the model distribution. The question we ask ourselves is how much uncertainty in our parameter estimates is caused by the difference between the model distribution and the actual data distribution.

Our solution to this problem is to use bootstrap inference on samples of a smaller size, for which the model cannot be rejected. We use the model adequacy test to choose a bootstrap size with limited probability of rejecting the model (we use probability 0.5 for analytical convenience). The intuitive idea is that if we have a sample size for which the model adequacy test is not often rejected, and our inference at this sample size gives a certain confidence interval, then we should be happy with this inference, because we might have been confident in it if our original dataset had been this size.

This approach has parallels with the *credibility index* of Lindsay and Liu [2009], which uses subsampling and a model adequacy test to measure the extent to which the model matches the data. However, the credibility index is simply a measure of how much data is needed to falsify the model. It does not give such an easily interpretable assessment of the goodness of fit, in terms of its effect on parameter estimates. That is, merely knowing that about 2,000 data points is sufficient to falsify a given model does not give a clear impression of whether the model is useful — in some cases this makes the model useful, and in others it does not. A confidence interval incorporating uncertainty due to model misspecification is often much easier to relate to usefulness.

We demonstrate the theory and application of the adequate bootstrap in two common situations — contamination and sampling bias. In both of these situations, we show that the adequate bootstrap greatly improves our coverage under the misspecified model cases. Meanwhile, when

the model is not misspecified, the adequate bootstrap is able to recover the same confidence interval as inference based on the full data.

References

G. E. P. Box (1976), Science and Statistics, *Journal of the American Statistical Association* **71** 791–799

B. Lindsay and J. Liu (2009) Model Assessment Tools for a Model False World *Statistical Science,* **24**, 303–318

Performance analysis and robustification of sequential statistical decision rules

A. Kharin[1*] *and T. Ton*[1]

[1] *Department of Probability Theory and Mathematical Statistics, Belarusian State University, Independence av. 4, Minsk 220030, Belarus; KharinAY@bsu.by, tthattu@gmail.com.*
[*] *Presenting author*

Keywords. *Sequential decision rule; Distortion; Robustness; Incomplete data; Asymptotic expansion.*

1 Introduction

Problems of statistical decision rules construction appear in many fields of real life. With sequential approach [Wald , 1947] it is possible to construct optimal decision rules w.r.t. the expected number of observations (sample size) provided the requested accuracy level (given small values for error probabilities upper bounds) is satisfied [Mukhopadhyay & de Silva , 2009]. This feature of sequential statistical decision rules plays an important role for practical use, especially in quality control, medicine, risk analysis.

In practice, the hypothetical model is often distorted [Huber & Ronchetti , 2009], [Hampel et al. , 1986]: observations contain "outliers", prior probability distributions are not true, the data can be incomplete. For these situations the mentioned optimal property of sequential decision rules is not valid [Kharin , 2008], [Kharin , 2013], and in some situations the traditionally used sequential decision rules are even not defined.

As a result, the problem of sequential decision rules performance analysis under distortions of different types is important [Kharin , 2002]. This problem should be considered together with the problem of robustified sequential decision rules construction [Kharin & Kishylau , 2015].

2 Short Description of Data Models, Distortions and Results

Three types of data models were considered: sequences of independent identically distributed observations; Markov chains; time series with a trend. Simple and composite (in the Bayesian setting) hypotheses cases are studied.

The following distortion types were analyzed: "outliers" in the data; functional distortions of the likelihood presented by ε-neighborhoods; "contamination" of the prior probability density function. The incomplete data case is investigated for the model of time series with a trend.

Asymptotic expansions of performance characteristics (factual values of error probabilities, conditional expected sample sizes) are derived for the proposed families of generalized sequential decision rules, including the decision rules that are traditionally used. Within the proposed families the robustified sequential decision rules are constructed. Theoretical results are illustrated numerically.

References

Wald, A. (1947). Sequential Analysis. Wiley, New York

Mukhopadhyay, N. & de Silva, B. (2009). Sequential Methods and their Applications. Chapman & Hall / CRC, Boca Raton.

Huber, P. & Ronchetti, E. (2009). Robust Statistics. Wiley, New York.

Hampel, F., Ronchetti, E, Rousseeuw, P. & Stahel, W. (1986). Robust Statistics. The Approach Based on Influence Functions. John Wiley and Sons.

Kharin, A. (2002). An approach to performance analysis of the SPRT for simple hypotheses testing *Proc. of the Belarusian State University*, **1**, 92–96.

Kharin, A. (2008). Robustness evaluation in sequential testing of composite hypotheses. *Austrian Journal of Statistics*, **37** (1), 51–60.

Kharin, A. (2013). Robustness of sequential testing of hypotheses on parameters of M-valued random sequences. *Journal of Mathematical Sciences*, **189** (6), 924–931.

Kharin, A.Yu. & Kishylau, D.V. (2015). Robust sequential test for hypotheses about discrete distributions in the presence of "outliers". *Journal of Mathematical Sciences*, **205** (1), 68–73.

Computational Methods for Quantile Regression

R. Koenker[1*]

[1] *Department of Economics, U. of Illinois, Urbana, IL 61801, USA, rkoenker@uiuc.edu*
[*] *Presenting author*

Keywords. *Quantile Regression, Convex Analysis, Proximal Operators.*

A survey of recent developments of computational methods for quantile regression will be provided. After a brief review of simplex and interior point methods, we will focus most of our attention on recent developments of ADMM methods based on modified gradient descent and the proximal operator. Several applications will be described to illustrate these methods in high dimensional settings.

Testing the symmetry of multivariate distribution of stock returns

V. Kalyagin[1], *P. Koldanov*[1] *and* ***A. Koldanov***[1*]

[1] *National Research University Higher School of Economics; vkalyagin@hse.ru, pkoldanov@hse.ru, akoldanov@hse.ru*

[*] *Presenting author*

Keywords. *Stock returns distribution–Elliptical model–Properties of symmetry–Multiple hypotheses testing–Distribution free tests.*

Models of multivariate distributions of stock returns attract important attention in theoretical and applied finance. In particular multivariate distribution models are necessary for portfolio selection and risk management. One of the popular multivariate model in financial analysis is the class of elliptically contoured distributions Gupta et al. [2013]. Consistency of real data with elliptically contoured model was studied in Chicheportiche & Bouchaud [2012] where it was shown that the joint distribution of real market stock returns is not in accordance with hypothesis of elliptical distributions. However, as pointed by the authors, their methodology differs from usual hypothesis testing using statistical tools.

In the present paper we analyze this problem from multiple hypotheses testing theory point of view Lehmann [2005]. Let X_i is a random variable, corresponding to the return of stock i ($i = 1, \ldots, N$) and $X = (X_1, \ldots, X_N)$ be the random vector of stock returns. To study the consistency of real data with elliptically contoured model for X we use one important property of elliptically contoured distributions, namely, the symmetry of density function of joint distribution with respect to the vector of means. To formulate the individual hypotheses we use the following pairwise sign symmetry property of multivariate elliptical distribution:

$$p_{1,1}^{i,j} = p_{-1,-1}^{i,j}; \quad p_{1,-1}^{i,j} = p_{-1,1}^{i,j}; \quad \forall i, j = 1, \ldots, N$$

where

$$p_{k,l}^{i,j} = P(k(X_i - E(X_i)) > 0, l(X_j - E(X_j)) > 0); k, l \in \{-1, 1\}$$

In the paper we construct multiple test which control FWER (Family-Wise Error Rate) for simultaneous individual hypotheses testing

$$h_1^{ij} : p_{1,1}^{i,j} = p_{-1,-1}^{i,j} \text{ vs } k_1^{ij} : p_{1,1}^{i,j} \neq p_{-1,-1}^{i,j}; i, j = 1, \ldots, N$$

$$h_2^{ij} : p_{1,-1}^{i,j} = p_{-1,1}^{i,j} \text{ vs } k_2^{ij} : p_{1,-1}^{i,j} \neq p_{-1,1}^{i,j}; i, j = 1, \ldots, N$$

For individual hypotheses testing we construct a tests of a Neyman structure based on sign statistics. Simultaneous inferences are conducted by Holm procedure, which is known to control FWER.

Resulting distribution free multiple test procedure was applied for USA and UK stock markets for different periods of observations. In most cases pairwise sign symmetry hypotheses are

not rejected. For the case where sign symmetry hypotheses are rejected for some pairs of stocks, rejection graph has an surprising structure. Despite number of rejected hypotheses the rejection graph has only few hubs and their removing leads to acceptance of all remaining symmetry hypotheses.

Acknowledgements: This work is supported RHRF grant N 15-32-01052

References

Gupta F.K. Varga T. Bodnar T. Elliptically Contoured Models in Statistics and Portfolio Theory, Springer, 2013, ISBN: 978-1-4614-8153-9.

Chicheportiche R. Bouchaud J-P. The joint distribution of stock returns is not elliptical: International Journal of Theoretical and Applied Finance, 15, 3, 2012.

Lehmann E.L. Romano J.P. Testing statistical hypotheses. Third Edition, Springer, New York, 2005, chapter 9.

Robust threshold graph selection in random variables network

V. Kalyagin[1], *A. Koldanov*[1] *and* ***P. Koldanov***[1*]

[1] *National Research University Higher School of Economics, Laboratory of Algorithms and Technologies for Network Analysis, Russia; vkalyagin@hse.ru, akoldanov@hse.ru, pkoldanov@hse.ru*
[*] *Presenting author*

Keywords. *Random variables network–Threshold graph–Sign similarity–Elliptically contoured distributions–Distribution free multiple decision procedures.*

Network analysis became a popular area of investigations last decades. Correlation networks form an important class in this field. Such networks appear to be useful in biological and financial applications. Biological applications are mostly related with probabilistic graphical models, Gaussian graphical models, weighted correlation networks and others. Financial applications are related with market network analysis. For correlation network observations are modeled as a sample from multivariate distribution. In this setting methods and algorithms of network analysis need to be considered as statistical procedures Drton & Perlman [2007], Koldanov et al. [2013].

We distinguish a class of correlation networks which we call *random variables networks*. Random variables network is a pair (X, γ), where $X = (X_1, X_2, \ldots, X_N)$ is a random vector and γ is a measure of association of random variables. Nodes of the network are associated with random variables and weight of edge (i, j) is given by $\gamma_{i,j} = \gamma(X_i, X_j)$, $i, j = 1, 2, \ldots, N$. This class includes Gaussian graphical and market network models. For Gaussian graphical model vector X has a multivariate Gaussian distribution and γ is the partial correlation. For market network model X_i is an attribute of stock i (return, volume, price) and γ is the Pearson correlation (in most cases).

Main goal of network analysis is to study a network structures containing a key information about network. This is especially important for large scale networks, with a big number of nodes and edges. One popular network structure is a *threshold or market graph* Boginski et al. [2005]. An edge is included in threshold graph iff its weight is larger than a given threshold. Threshold graph selection (identification) problem is to select the threshold graph by sample of observations. One of the way to construct threshold graph selection procedures is to use a sample Pearson correlations. It is known that statistical procedures based on sample Pearson correlations are appropriate for Gaussian distributions. However sometimes distributions of real data (stock market returns) have a heavy tails and does not fit hypothesis of Gaussian distribution. Therefore it is important to investigate a distribution free (robust) threshold graph selection statistical procedures. Our main goal is to analyse from this point of view a statistical procedures for threshold graph selection in random variables network with elliptically contoured distributions. This class of distributions is largely used in financial analysis Gupta et al. [2013]. In this paper we study threshold graph selection statistical procedures in *sign similarity network* and compare it with selection statistical procedures in Pearson correlation network. The measure of association between random variables X_i, X_j in sign similarity network is given by the probability of sign coincidence $\gamma^S_{i,j} = P[(X_i - E(X_i))(X_j - E(X_j))] > 0$.

To study threshold graph selection problem we use a multiple decision statistical theory. The decision procedures considered in this paper are based on simultaneous application of sign tests. Three popular multiple statistical procedures are investigated: single step multiple decision procedure, step down Holm multiple testing procedure and step up Hochberg multiple testing procedure. The quality of the procedures is measured by FWER (Family-Wise Error Rate) and risk function. Our main result is: considered multiple decision procedures for threshold graph selection are distribution free in sign similarity network in the class of elliptically contoured distributions. Moreover it is shown that these procedures can be adapted for distribution free threshold graph selection in Pearson correlation network. Note that the class of elliptically contoured distributions includes Gaussian distribution and distributions with heavy tails (multivariate Student).

Acknowledgements: This work is supported RHRF grant N 15-32-01052

References

Koldanov A.P., Koldanov P.A., Kalyagin V.A., Pardalos P.M. Statistical procedures for the market graph construction.: Computational Statistics and Data Analysis 68 17–29 (2013).

Drton M., Pelman M. Multiple Testing and Error Control in Gaussian Graphical Model Selection, Statistical Science 22 3 430-449 (2007).

Gupta F.K. Varga T. Bodnar T. Elliptically Contoured Models in Statistics and Portfolio Theory, Springer, 2013, ISBN: 978-1-4614-8153-9.

Boginski V., Butenko S., Pardalos P.M. Statistical analysis of financial networks. J. Computational Statistics and Data Analysis. 48 (2), 431–443 (2005).

Ratio estimators in median ranked set and neoteric ranked set sampling

N. Koyuncu[1*]

[1] *Hacettepe University, Faculty of Science, Department of Statistics, Ankara, Turkey; nkoyuncu@hacettepe.edu.tr.*
[*] *Presenting author*

Keywords. *Ratio estimator; Ranked Set Sampling; Efficiency ; Median.*

1 General Information

Ranked set sampling, introduced by McIntyre [1952] is one of the most effective sampling design when the characteristic of interest is expensive or time consuming to measure but a few units in a set are easily ranked without full measurement. This sampling design is especially applied in agricultural, biological, ecological, engineering, medical, physical, and social sciences. To get improvement many researchers have developed and modified ranked set sampling design. Recently Zamanzade &Al-Omari [2016] have proposed neoteric ranked set sampling. Koyuncu [2015] and Koyuncu [2016] have defined ratio type estimators in various ranked set sampling design. In this paper, following Zamanzade &Al-Omari [2016] and Koyuncu [2016] we have defined ratio type estimators in median and neoteric ranked set sampling. To compare the performance of suggested estimators in each sampling design we have conducted a simulation study. In the simulation study, we have used a real data set. In this data set, we have examined a rare endemic annual plant species which is grown in Ankara-Turkey. 900 seeds of endemic plant's weight and height are measured. Using this data set as a population by setting height as study variable and weight as an auxiliary variable we selected samples under median ranked set and neoteric ranked set sampling desing. From each sample we have estimate the height of plant with suggested estimators and calculated mean square error. According to mean square error value we have decided that which sampling plan is suitable for this data set.

References

Koyuncu, N. (2015). Ratio estimation of the population mean in extreme ranked set and double robust extreme ranked set sampling. *International Journal of Agricultural and Statistical Sciences*, **11,1**, 21–28.

Koyuncu, N. (2016). New difference-cum-ratio and exponential type estimators in median ranked set sampling. *Hacettepe Journal of Mathematics and Statistics*, **45,1**, 207–225.

McIntyre, G.A. (1952). A method for unbiased selective sampling using ranked sets. *Australian Journal of Agricultural Research*, **3**, 385–390.

Zamanzade, E. & Al-Omari, A.I. (2016). New ranked set sampling for estimating the population mean and variance. *Hacettepe Journal of Mathematics and Statistics*, Doi: 10.15672/HJMS.20159213166 (online published).

Business analytics, statistical thinking and statistical engineering and their application in education at the GSEM

D. Kuonen[1*]

[1] *Statoo Consulting, Berne, Switzerland & 'Geneva School of Economics and Management' (GSEM), University of Geneva, Switzerland; @DiegoKuonen; kuonen@statoo.com & Diego.Kuonen@unige.ch*
[*] *Presenting author*

Keywords. *Business analytics; Data-driven decision making; Statistical thinking; Statistical engineering; Data science; Big data; Education; Teaching; GSEM*

More data are available than ever before, and data analytics have become a part of every major business decision today.

Business analytics refers to the methodology employed by an organisation to enhance its business and make optimised decisions based on data and by the use of statistical thinking to improve, for example, their products, services, supply chain and operations, human resources, financial management and marketing.

Business analytics is about bringing the business questions to the data. Statistical thinking and engineering assist this knowledge acquisition process in a principal and scientific manner.

This presentation starts by giving an introduction to business analytics, along with data-driven decision making, and illustrates the key link to statistical thinking and statistical engineering.

Next the speaker will give an overview and share his experiences of teaching 'Business Analytics' to students taking a major in management within the second part of GSEM's 'Bachelor in Economics and Management' (taught the first time in the fall semester 2015).

Finally, an outlook of future GSEM initiatives on business analytics and data science will be presented, in particular GSEM's forthcoming 'Business Analytics' master programme (to be started in the fall semester 2017).

A robust estimator for the mean direction of the von Mises-Fisher distribution

S. Liebscher[1*], *T. Kirschstein*[1], *G. Pandolfo*[2], *G.C. Porzio*[2] *and G. Ragozini*[3]

[1] *Martin-Luther-University Halle-Wittenberg, Große Steinstraße 73, D-06099 Halle (Saale), Germany; steffen.liebscher@wiwi.uni-halle.de, thomas.kirschstein@wiwi.uni-halle.de.*
[2] *University of Cassino and Southern Lazio, Via San Angelo, localita Folcara, I-03043 Cassino, Italy; giuseppepandolfo1@virgilio.it, porzio@unicas.it.*
[3] *University of Naples Federico II, Via L. Rodinò 22, I-80138, Napoli, Italy; giancarlo.ragozini@unina.it.*
[*] *Presenting author*

Keywords. *Directional data; Robust estimation; Von Mises-Fisher distribution.*

The von Mises-Fisher (or Langevin) distribution is commonly used to describe the distribution of data on the (hyper)sphere. The distribution has two parameters: the mean direction $\boldsymbol{\mu}$ and the concentration parameter κ (with $0 \leq \kappa \leq \infty$), is unimodal (for $\kappa > 0$), and symmetrical about $\boldsymbol{\mu}$. Estimators for both parameters are readily available for some time (Hornik & Grün [2014], Sra [2012], Banerjee et al. [2005]). While robustness issues have been taken into consideration when estimating the concentration parameter (Laha & Mahesh [2012], Ko [1992], Fisher [1982]), robustness has only recently been discussed when estimating the mean direction [Kato & Eguchi, 2014]. This is mainly due to a common and persistent misconception that non-robustness does not constitute a serious problem when working with directional data (especially when talking about the mean direction, as the directional difference between any two points on the (hyper)sphere can only be 180° or π at the maximum).

In this paper the measurement of robustness of mean direction estimators is reconsidered. A new measure to assess the robustness based on the maximum bias is introduced. This measure provides the foundation to derive a definition of breakdown of an estimator similar in concept to the well-known finite sample breakdown point. Finally, a new estimator for the mean direction of the von Mises-Fisher distribution is proposed which is shown to deliver consistent estimates as well as being robust in terms of the newly introduced measures of robustness. Results of a simulation study indicate that the new estimator is more robust than the approach by Kato & Eguchi [2014] and much more robust than the standard ML estimator for certain contamination schemes.

References

Banerjee, A., Dhillon, I.S., Ghosh, J., and Sra, S. (2005). Clustering on the Unit Hypersphere Using Von Mises-Fisher Distributions. *J. Mach. Learn. Res.*, **6**, 1345–1382.

Fisher, N. I. (1982). Robust estimation of the concentration parameter of fisher's distribution on the sphere. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **31**(2), 152–154.

Hornik, K. and Grün, B. (2014). On maximum likelihood estimation of the concentration parameter of von Mises-Fisher distributions. *Computational Statistics*, **29**, 945–957.

Kato, S. and Eguchi, S. (2014). Robust estimation of location and concentration parameters for the von Mises-Fisher distribution. *Statistical Papers*, 1–30.

Ko, D. (1992). Robust estimation of the concentration parameter of the von mises-fisher distribution. *The Annals of Statistics*, **20**(2), 917–928.

Laha, A. and Mahesh, K. (2012). Sb-robust estimator for the concentration parameter of circular normal distribution. *Statistical Papers*, **53**(2), 457–467.

Sra, S. (2012). A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of $I_s(x)$. *Computational Statistics*, **27**, 177–190.

On complex valued ICA using M-estimators

N. Lietzén[1*] *and P. Ilmonen*[1]

[1] *Aalto University School of Science, Department of Mathematics and Systems Analysis, P.O. Box 11100, 00076 Aalto, Finland;*
niko.lietzen@aalto.fi, pauliina.ilmonen@aalto.fi.
[*] *Presenting author*

Keywords. *ICA; Complex valued variables; M-estimators*

In complex valued Independent Component model (IC-model) the elements of a complex valued p-variate random vector are assumed to be linear combinations of an unobservable p-variate vector with mutually independent components. In Independent Component Analysis (ICA) the aim is to recover the independent components by estimating an unmixing matrix that transforms the observed p-variate vector to independent components. In this work we consider, in presence of outlying points, robust ICA based on two complex valued scatter matrices.

Appraising the aptness of GEE models for longitudinal binary data via graphical and numerical methods

K.C. Lin[1*] *and Y.J. Chen*[2]

[1] *Professor, Department of Business Administration, Tainan University of Technology, Taiwan; t20053@mail.tut.edu.t*
[2] *Associate Professor, Department of Statistics, Tamkang University, Taiwan; ychen@stat.tku.edu.tw.*
[*] *Presenting author*

Keywords. *Generalized estimating equations; Goodness-of-fit; Graphical method; Hierarchical clustering; Longitudinal binary data; Nonparametric smoothing.*

Longitudinal binary data are pervasively utilized in a variety of fields, and commonly fitted by the generalized estimating equations (GEE) approach. We develop two graphical methods and one numerical goodness-of-fit test for appraising the aptness of GEE fitted models under independent and unstructured working correlation matrices. Two graphical approaches are marginal model plots (MMP) and local mean deviance methods (LMDP). A goodness-of-fit test based on nonparametric smoothing approach is provided. Moreover, the estimations of mean and standard deviation functions in the MMP procedure are employed the kernel smoothing technique, and the number of groups is determined by hierarchical clustering method in the LMDP procedure. Two real data sets are used to demonstrate the application of the numerical and graphical approaches. The results show that the models are adequate by utilizing the numerical goodness-of-fit test based on both independent and unstructured working correlations. Furthermore, the graphical LMPD procedure offers more visual plots to depict that the unstructured working correlation is more adequate for the feature of data, and the MMP procedure furnishes the detailed plots for the model with or without particular covariates.

A nonparametric graphical model for functional data with application to brain networks based on fMRI

B. Li[1*], *and E. Solea*[1]

[1] *Department of Statistics, Pennsylvania State University, 326 Thomas Building, University Park, PA 16802, USA; bxl9@psu.edu, exs392@psu.edu*
[*] *Presenting author*

Keywords. *Additive conditional independence; Additive correlation operator; additive precision operator; Gaussian graphical model; Reproducing kernel Hilbert space.*

We introduce a nonparametric graphical model whose observations on vertices are functions. Many modern applications, such as electroencephalogram and functional magnetic resonance imaging (fMRI), produce data are of this type. The model is based on Additive Conditional Independence (ACI), a statistical relation that captures the spirit of conditional independence without resorting to multi-dimensional kernels. The random functions are assumed to reside in a Hilbert space. No distributional assumption is imposed on the random functions: instead, their statistical relations are characterized nonparametrically by a second Hilbert space, which is a reproducing kernel Hilbert space whose kernel is determined by the inner product of the first Hilbert space. A precision operator is then constructed based on the second space, which characterizes ACI, and hence also the graph. The resulting estimator is relatively easy to compute, requiring no iterative optimization or inversion of large matrices. We establish the consistency the convergence rate of the estimator. Through simulation studies we demonstrate that the estimator performs better than the functional Gaussian graphical model when the relations among vertices are nonlinear or heteroscedastic. The method is applied to an fMRI data set to construct brain networks for patients with attention-deficit/hyperactivity disorder.

High-dimensional consistency of robust precision matrix estimators

Po-Ling Loh[1*] *and Xin Lu Tan*[1]

[1] *University of Pennsylvania; loh@wharton.upenn.edu, xtan@wharton.upenn.edu.*
[*] *Presenting author*

Keywords. *High-dimensional statistics; Precision matrix estimation; Robust correlation; Graphical models.*

We analyze the statistical consistency of robust estimators for precision matrices in high dimensions. Such estimators, formed by plugging robust covariance matrix estimators into the graphical Lasso/CLIME machinery, were recently proposed by other authors but only studied from the point of view of breakdown behavior. As a complementary result, we provide error bounds for the precision matrix estimators based on various contamination models, revealing the interplay between the dimensionality of the problem and the degree of contamination permitted in the observed distribution. We discuss implications of our work for problems involving graphical model estimation when the uncontaminated data follow a multivariate normal distribution.

Practical Robust Statistical Methods via R

M. Mächler[1*]

[1] *Seminar für Statistik, ETH Zurich; maechler@stat.math.ethz.ch*
[*] *Presenting author*

Keywords. *Robust statistics; R; Applications; robustbase; regression.*

The R language and environment for statistical computing and graphics has become the de facto standard for most of professional statisticians. Some robust methodology has been part of S, the precursor of R, since the 1980s. Further, in 2006, a considerable group of research experts in robust statistics had decided to join forces creating a dedicated R package providing the most basic and important robust methodology, notably covering Maronna et al. (2006), and the R package `robustbase` ([Rousseeuw et al., 2015]) was created. Not much later, when CRAN task views (CTV) were initiated as a human expert guides to the exponentially growing CRAN R packages, the "Robust" CTV ([Maechler, 2014]) was created and listed dozens of R packages already in 2007 and has been listing 50 packages since the end of 2014.

Our presentation will emphasize on use-cases of regression methods — in our experience used for roughly 90% of all data analysis — and also use the `fit.models` package to compare classical (least squares) and robust model fits. We will look at some generalizations of linear models— nonlinear, mixed effects, generalize additive models (GAM— and explore how easily (or not) using robust R functions can be used instead of or in addition to classical methods.

References

Kjell Konis. (2016). fit.models: Compare Fitted Models. R package version 0.5-11, `http://CRAN.R-project.org/package=fit.models`.

Martin Maechler (2014). CRAN Task View: Robust Statistical Methods. Version 2014-12-07, `http://CRAN.R-project.org/view=Robust`

Maronna, R., Martin, D. & Yohai, V. (2006). Robust Statistics: Theory and Methods. John Wiley & Sons, Chichester.

R Core Team (2016). R Foundation for Statistical Computing. R: A Language and Environment for Statistical Computing. Vienna, Austria. `https://www.R-project.org/`.

Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M. & Maechler, M. (2015). robustbase: Basic Robust Statistics. R package version 0.92-5, `http://CRAN.R-project.org/package=robustbase`.

Distances and their role in robustness

M. Markatou[1*]

[1] *Department of Biostatistics, School of Public Health & Health Professions and Jacobs School of Medicine and Biomedical Sciences SUNY, Buffalo; markatou@buffalo.edu*
[*] *Presenting author*

Keywords. *Model selection; Quadratic information criterion (QIC); Robustness; Statistical distances*

Statistical distances, divergences and similar quantities have a large history and play a fundamental role in statistics, machine learning and associated scientific disciplines. In this talk, we discuss the role of statistical distances in robustness. First, we discuss selected properties of the distances and illustrate their robustness in testing. Then, we focus on their role on robust model selection and propose the Quadratic Information Criterion (QIC), a method for model selection derived as an appropriate estimator of the relative quadratic risk, the definition of which is based on the concept of quadratic distance between two probability distributions. We discuss the construction of QIC and illustrate its performance in linear and non-linear regression via simulation and via application to real data sets.

Can we do without subsampling?

R. Maronna[1*] *and V. Yohai*[2]

[1] *University of La Plata (rmaronna@retina.ar)*
[2] *University of Buenos Aires and CONICET (victoryohai@gmail.com)*
[*] *Presenting author*

Keywords. *Subsampling; High-dimensional data; Fast estimation; Peña-Yohai regression procedure; Peña-Prieto multivariate estimator*

High breakdown point equivariant estimators in regresssion and multivariate analysis usually require the minimization of a non-convex function of the parameters through an iterative procedure. The possible existence of local minima corresponding to "bad" solutions makes the starting point of the iterations crucial. Subsampling has been the traditional way to compute a robust equivariant starting point. It is known that the number of subsamples required to ensure a given"probabilistic breakdown point" increases exponentially with the data dimension. This fact implies that, except for small dimensionality, there is a conflict between robustness and computing time.

We want to put forward two procedures, one for linear regression and the other for the estimation of multivariate location and scatter, which have been already proposed several years ago for outlier detection, but which have never been considered as initial values for robust iterative estimators. Since their complete description is rather complex, here we describe just the main ideas behind each one.

The first one is a deterministic method proposed by Peña and Yohai (1999) to detect outliers in linear regression. Briefly said, it computes a set of p "sensitivity directions" (where p is the number of parameters) from which it derives $3p$ candidate estimates, which are used as starting values for an S-estimate. The method is very fast. Our simulations show that its use to compute MM-estimates highly improves both robustness and speed.

The second is a semi-deterministic method for multivariate estimation proposed by Peña and Prieto (2007), based on ideas similar to the Stahel-Donoho estimator. But rather than taking purely random directions, they look for two sets directions that have a high probability of exposing outliers. The first set is deterministic, and contains the $2p$ directions that maximize or minimize the kurtosis of the projections. The other set, which is random, is obtained by an elaborate "stratified sampling", and the number of its elements is proportional to p. These directions are used to compute outlyingness measures for the data points. The complete procedure is rather complex. Simulations by Maronna and Yohai (2015) show that employing this procedure as a starting point highly improves the performance of multivariate MM- and τ-estimators with respect to subsampling, both in robustness and speed.

For these reasons we propose these two procedures to be routinely employed instead of subsampling.

References

Maronna, R. & Yohai, V. (2015). Robust and efficient estimation of high dimensional scatter and location. http://arxiv.org/abs/1504.03389

Peña, D. & Prieto, F.J. (2007). Combining random and specific directions for outlier detection and robust estimation in high-dimensional multivariate data. *Journal of Computational and Graphical Statistics*, **16**, 228-254.

Peña, D. & Yohai, V. (1999). A Fast Procedure for Outlier Diagnostics in Large Regression Problems. *Journal of the American Statistical Association*, **94**, 434-445.

Finance Applications of Robust Statistics and Influence Functions

R.D. Martin[1*]

[1] *Professor Emeritus of Applied Mathematics and former CFRM Program Director, University of Washington.*
[*] *Presenting author*

Keywords. *Fundament factor exposures; Fama-French1992; Independent outliers across assets; Risk and performance measure IF's; Standard errors with serial correlation*

The first part of this talk briefly discusses the following results for cross-section factor models and mean-variance portfolio optimization. Robust distances reveal many more multi-dimensional outliers than one-dimensional outliers in fundamental factor models exposures data (EP, MB, size, etc.), a fact that is widely ignored by commercial fundamental factor model based portfolio optimization software providers. Robust regression applied to Fama-French cross-section regression asset pricing models yield results that differ from least squares in financially significant ways. In the context of mean-variance portfolio optimization we show by example that independent outliers in assets (IOA) is an appropriate model for firm specific outliers and that a robust covariance matrix based on pairwise quadrant correlation yields better out-of-sample backtesting performance than the MCD estimator as an exemplar of traditional robust covariance estimation. The second part of the talk focuses on applications of influence functions to risk measure and performance measure estimators. We show that influence functions reveal important differences between parametric maximum-likelihood estimators and non-parametric estimators of expected shortfall (ES), and that the ES MLE has a serious shortcoming that can be corrected by use of a semi-scale estimator. Finally, we discuss the use of influence functions to compute standard errors of risk and performance measure estimators for serially correlated returns.

This talk is based on joint work with Xin Chen, Chris Green and Shengyu Zhang.

Sliced inverse regression for time series

M. Matilainen[1*], *C. Croux*[2], *K. Nordhausen*[1] *and H. Oja*[1]

[1] *Department of Mathematics and Statistics, University of Turku, Finland; markus.matilainen@utu.fi, klaus.nordhausen@utu.fi, hannu.oja@utu.fi.*
[2] *Faculty of Business and Economics, KU Leuven, Belgium; christophe.croux@kuleuven.be*
[*] *Presenting author*

Keywords. *Covariance matrix; Dimension reduction; Multivariate analysis*

1 Background

When analysing data with a response variable y and explanatory variables $\mathbf{x}$, modelling may become infeasible when number of variables gets higher. It can also cause computational problems and visualization of data becomes harder.

To avoid these kind of problems we can use Sliced Inverse Regression (SIR) [Li, 1991], which is a supervised dimension reduction method and it is used to study a relationship between the response y and the variables $\mathbf{x}$. However, in case of time series SIR algorithm does not use any information on lagged values directly. One way to deal with this is to treat explanatory variables and their past values and the past values of response variable as explanatory variables. [Becker & Fried, 2003]

2 A new method for time series data

We suggest in Matilainen et al. [2016] a new method, which is based on SIR algorithm, but instead of using a regular supervised covariance matrix, we use several lagged supervised covariance matrices. We show how our new algorithm can be used to determine which lags and directions are the most important ones when trying to predict future values. Some illustrative examples are presented in order to see how our algorithm performs in different kind of situations.

References

Li, K-C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, **86**, 316–327.

Becker, C. & Fried, R. (2003). Exploratory Data Analysis in Empirical Research. Springer, Berlin Heidelberg, pp. 3–11.

Matilainen, M., Croux, C., Nordhausen, K. & Oja, H. (2016). Supervised dimension reduction for multivariate time series. *Manuscript*.

Robust model-based clustering and mixture modeling via trimming and constraints

L.A. García-Escudero[1], *A. Gordaliza*[1], *F. Greselin*[2] *and* ***A. Mayo-Íscar***[3*]

[1] *Dpto. Estadística e I.O and IMUVA, Facultad de Ciencias, Universidad de Valladolid, Paseo Belén 7, Campus Miguel Delibes, 47011 Valladolid, Spain; lagarcia@eio.uva.es, alfonsog@eio.uva.es*
[2] *Department of Statistics and Quantitative Methods, Milano Bicocca University, 20126 Milan, Italy; francesca.greselin@unimib.it*
[3] *Dpto. Estadística e I.O and IMUVA, Facultad de Medicina, Universidad de Valladolid, Avda. Ramón y Cajal 7, Campus Miguel Delibes, 47005 Valladolid, Spain; agustin@med.uva.es*
[*] *Presenting author*

Keywords. *Mixture models; Clustering; Trimming; Constraints*

1 Abstract

Trimming procedures are commonly used in many statistical settings for getting robust estimators in the presence of contamination. For achieve a robust proposal, both in model-based clustering and mixture modeling, the only adoption of trimming is not enough. It is also necessary to control the relative size of the components' scatter by applying constraints. The talk presents robust methodology based on the joint application of trimming and constraints for different clusters and mixture models settings. Proposals based in this methodology will be applied for estimating mixtures of regressions, mixtures of factor analyzers and mixtures of skew-symmetric models. Data driven tools for helping to the user in choosing the additional input parameters, related with the joint application of these tools, will be also presented.

Identifying behaviors from marine animal tracks

J. Mills Flemming[1*], *K. Whoriskey*[1] *M. Auger-Methe*[1] *and C. Albertsen*[2]

[1] *Department of Mathematics and Statistics, Dalhousie University; Joanna.Flemming@Dal.Ca, kwhoriskey@dal.ca, auger-methe@dal.ca.*
[2] *National Institute of Aquatic Resources, Technical University of Denmark; cmoe@aqua.dtu.dk.*
[*] *Presenting author*

Keywords. *Animal movement; Satellite telemetry data; State-space models; Robustness; Template Model Builder.*

Animals move in order to maximize their probability of survival and reproduction. The movement of an animal therefore reflects its response to its current physical needs and available environment. In the marine realm, where direct observation of animal movements is often impossible, researchers typically employ satellite telemetry positioning systems to obtain series of estimates of locations of animals in space through time. Each series resembles an animal path or track. Inferring behavioural states from animal tracks is possible by reasonably assuming that different types of movement, and therefore behavioural state, can be reflected by a change in characteristics of an animal path. For example, while foraging can often be characterized by a tortuous track, a more directed path may suggest travelling between foraging patches. Here I discuss robust state-space model formulations for estimating behavioural states from animal tracks and demonstrate their utility with both a simulation study and real application. Information gained from these models can be used for proper management of both species and ecosystems.

An R Package for Robust Time Series Analysis

S. Guerrier[1], ***R. Molinari***[2*] *and J. Balamuta*[1]

[1] *Department of Statistics, University of Illinois at Urbana Champaign, USA; stephane@illinois.edu, balamut2@illinois.edu.*
[2] *Research Center for Statistics, GSEM, University of Geneva, Switzerland; roberto.molinari@unige.ch*
.
[*] *Presenting author*

Keywords. *Wavelet variance; Generalized method of wavelet moments; Robust inference; State-space models.*

This work presents a new R package for the robust estimation and inference for time series models. The package is called "gmwm" and is based on the findings of Guerrier et al. [2013] who use a wavelet decomposition of the time series to obtain parameter estimates using the idea of the Generalized Method of Wavelet Moments (GMWM). Taking from these results, Guerrier et al. [2016] developed a robust version of this methodology which allows to make available a flexible and broad tool for robust time series inference.

The gmwm package makes use of the Wavelet Variance (WV) which is the variance of the wavelet coefficients issued from a wavelet decomposition. This WV is estimated robustly by adapting Huber's Proposal 2 to the stationary time series setting and allows the user to specify the desired level of efficiency compared to the standard estimator of WV. Using this quantity, the package makes available some flexible plotting tools to compare the standard and robust WV and understand if a robust inference procedure is necessary.

Once the WV analysis is carried out and a robust estimation appears to be necessary, the package provides a function which allows to robustly estimate the parameters of a wide range of Gaussian time series, going from ARMA to many (additive) State-Space models. Moreover, the package provides further plotting tools to verify whether the estimated model fits the time series well in addition to functions to carry out robust inference and model selection.

The gmwm package therefore represents the only available platform to date which implements a general framework for robust inference for a broad class of time series models.

References

Guerrier, Stéphane and Skaloud, Jan and Stebler, Yannick & Victoria-Feser, Maria-Pia (2013). Wavelet-variance-based estimation for composite stochastic processes. *Journal of the American Statistical Association*, **108**, 1021–1030.

Guerrier S. & Molinari R. (2016). Robust Inference for Time Series Models: a Wavelet-based Framework. Submitted manuscript.

Quo vadis Robustness?

S. Morgenthaler[1*]

[1] *École polytechnique fédérale de Lausanne; stephan.morgenthaler@epfl.ch.*
[*] *Presenting author*

Keywords. *Model uncertainty; Bias; Regression, Big data.*

1 Exploratory Data Analysis v. Robustness

Robustness provides an answer to model uncertainly. It has dealt successfully with distributional uncertainly in linear models and with multivariate observations, in particular through theoretical advances and resulting heuristics. But the emphasis on asymptotic theory and a narrow view of model uncertainty has held back the general use of robust procedures.

The more archaic concepts favored by JW Tukey such as sensitivity and resistance should again take precedence over asymptotic robustness. I will give examples from John Tukey's teaching materials to illustrate his thinking. He put the focus on generally applicable ideas – or technologies – that could contribute to the construction of procedures for data analysis. Examples include the use of weights, transformations, the jackknife, and borrowing strength. In any given problem, a combination of technologies can be applied, with each combination giving different results. The emphasis is not on achieving optimality, but rather on creating a variety of data analytic procedures, which could enter into competition and be compared by simulation and through examples.

This does not argue for or against robust procedures. It argues for variety and creativity. The problem it poses is rather one of dissemination. We need a collection of the basic concepts that we find useful and instructions on how to use them. We also need a depository for all the robust/resistant procedures that have been found, for the results that are known about them, and for comparisons among themselves and with other approaches.

2 New Avenues for Robustness

The basic principles of robustness can lead to new procedures in many data analytic problems where it is unclear how to think about alternatives to established stochastic models. What is a robust/resistant analysis of risk or of extreme observations? What is a robust/resistant allocation of resources in the face of uncertainty. How would a robust/resistant test of genetic risk look like? What is a robust classification.

I will use a simple situation involving binary observations to illustrate problems of bias that arise naturally in many of these applied areas and comment on the use of the basic principles

in big data applications, where it is natural to repeatedly fit subsamples and to combine the results.

Comparing Two Independent Groups Through Quantiles

G. Navruz[1*] , *A. F. Özdemir* [1]

[1] *Department of Statistics, Dokuz Eylül University, Turkey ; firat.ozdemir@deu.edu.tr, gnavruz@gmail.com*
[*] *Presenting author*

Keywords. *Two independent groups; Percentile bootstrap; Quantile estimators*

The most common approach for comparing two independent groups is on the basis of some measure of location. If the population distributions are symmetric the mean is appropriate, but if the population distributions are skewed a robust measure of location might be preferred. In addition to this, it is often of interest to determine whether the differences occur in the tails of distributions or not. Therefore, the quantiles should be considered, especially the lower and upper ones. There are great number of methods for estimating population quantiles and additional comparisons of various quantile estimators in the literature. Some well-known examples are Harrell & Davis [1982], Wilcox & Muska [2001],Sheather & Marron [1990], Wilcox [1994], Bradley [1978] as well as Hyndman & Fan [1996]. The quantile estimators may be based on one or two order statistics, rather based on all of the order statistics by taking a weighted average as is Harrell Davis estimator. Moreover, Sfakianakis and Verginis derived three quantile estimators that again use a weighted average of all the order statistics and have advantages in some situations[Wilcox, 2004]. In this study, Harrell Davis estimator and three estimators by Sfakianakis and Verginis are used in conjunction with a percentile bootstrap method with the aim of comparing two independent groups via the quantiles. When comparing the quantiles that are close to the median, Sfakianakis and Verginis estimators coped well with the Harrell Davis estimator, and when comparing the quantiles that are close to zero or one Sfakianakis and Verginis estimators performed better in terms of actual type I error rates.

References

Dielman, T., Lowry, C.& Pfaffenberger, R. (1994). A comparison of quantile estimators. *Communications in Statistics-Simulation and Computation*, **23**(1), 355–371.

Harrell, F. E. & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika*,**69**(3), 635–640.

Hyndman, R. J & Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, **50**(4), 361–365.

Kalgh, W. D. & Lachenbruch, P. A. (1982). A generalized quantile estimator. *Communications in Statistics-Theory and Methods*, **11**(19), 2217–2238.

Parrish, R. S. (1990). Comparison of quantile estimators in normal sampling. *Biometrics*, **46**(1), 247–257.

Sfakianakis, M. E.& Verginis, D. G. (2008). A new family of nonparametric quantile estimators. *Communications in Statistics - Simulation and Computation*, **37**, 337–345.

Sheather, S. J. & Marron, J. S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association*, **85**(410), 410–416.

Wilcox, R. R., Erceg-Hurn, D. M., Clark, F. & Carlson, M. (2014). Comparing two independent

groups via the lower and upper quantiles. *Journal of Statistical Computation and Simulation*, **84**(7), 1543–1551.

A Unified Theory of Confidence Regions and Testing for High Dimensional Estimating Equations

M. Neykov[1*], *Y. Ning*[1], *Jun S. Liu*[3] *and H. Liu*[1]

[1] *Department of Operations Research, Princeton University; mneykov@princeton.edu, yning@princeton.edu, hanliu@princeton.edu.*
[2] *Department of Statistics, Harvard University; jliu@stat.harvard.edu*
[*] *Presenting author*

Keywords. *Post-regularization inference; Estimating equations; Confidence regions; Hypothesis tests;*

We propose a new inferential framework for constructing confidence regions and testing hypotheses in statistical models specified by a system of high dimensional estimating equations. We construct an influence function by projecting the fitted estimating equations to a sparse direction obtained by solving a large-scale linear program. Our main theoretical contribution is to establish a unified Z-estimation theory of confidence regions for high dimensional problems. Different from existing methods, all of which require the specification of the likelihood or pseudo-likelihood, our framework is likelihood-free. As a result, our approach provides valid inference for a broad class of high dimensional constrained estimating equation problems, which are not covered by existing methods. Such examples include, noisy compressed sensing, instrumental variable regression, undirected graphical models, discriminant analysis and vector autoregressive models. We present detailed theoretical results for all these examples. Finally, we conduct thorough numerical simulations, and a real dataset analysis to back up the developed theoretical results.

Tightness of M-estimators for multiple linear regression in time series

S. Johansen[1] *and* ***B. Nielsen***[2*]

[1] *University of Copenhagen & CREATES; soren.johansen@econ.ku.dk.*
[2] *University of Oxford; bent.nielsen@nuffield.ox.ac.uk*
[*] *Presenting author*

Keywords. *M-estimator; Martingales; Huber-skip; Quantile regression.*

We show tightness for a class of regression M-estimators, where the objective function can be non-monotonic and non-continuous. A prominent example of an estimator is the Huber skip estimator, where each observation contributes to the objective function through a criterion function, which is quadratic in the central part and horizontal otherwise. The tightness result addresses a difficulty which is often met in asymptotic analysis of problems where the objective function is non-convex. A very common solution is to assume that the parameter space is compact. While such an assumption circumvents the problem, it is done through a condition on the unknown parameter and it is therefore rarely satisfactory from an applied viewpoint. Instead, our result only requires an assumption that can be justified by inspecting the observed regressors and the objective function.

We consider the multiple linear regression

$$y_i = \mu + \alpha' x_i + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{15}$$

where the innovations ε_i are independent of $\mathcal{F}_{i-1} = \sigma(x_1, \ldots, x_i, \varepsilon_1, \ldots, \varepsilon_{i-1})$. The regressors x_i have dimension m. They can be deterministic or stochastic, and stationary or stochastically trending. Often we will subsume the intercept in the regressors and use the notation

$$\beta' z_i = \mu + \alpha' x_i.$$

The M-estimator for the parameter β is the minimizer, $\hat{\beta}$, of the objective function

$$R_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - z_i'\beta), \tag{16}$$

for some criterion function ρ. M-estimators were originally introduced for location problems by Huber in 1964, but later extended to regression models. The class of M-estimators considered includes the Huber-skip estimator, which has a non-convex criterion function, as well as quantile regression estimators, in particular the least absolute deviation, and least squares estimator, which all have a convex criterion.

The asymptotic theory of the regression M-estimator is well understood for nice criterion functions ρ. Maronna et al. [2006] provide an asymptotic theory for regression M-estimators and show existence and uniqueness for the case of a convex, differentiable criterion function. Chen & Wu [1988] give two results on tightness (and consistency) for more general criterion functions. In both cases the criterion function $\rho(u)$ is continuous, non-decreasing in $u > 0$ and non-increasing for $u < 0$. Their Theorem 1 shows tightness when (y_i, z_i') are i.i.d. and $\mathsf{E}\rho(y_i - z_i'\beta)$

has a unique minimum. Their Theorem 4 shows tightness when z_i is deterministic and satisfies a condition on the frequency of small regressors.

In this paper we generalize the result of Chen & Wu [1988]. We assume ρ is semi-continuous and nonnegative with a minimum at zero and greater then $\rho^* > 0$ for large values of the argument. We also need an extra condition on the expected criterion function $h(v)$, which is assumed to take a value below ρ^* somewhere in the central part of the distribution of the error term. The only condition to the regressors is a condition on the frequency of small regressors, which is weaker than the condition of Chen & Wu [1988], albeit stronger than the conditions for the tightness of least square estimators. The latter illustrates the price we pay by leaving the least squares criterion. The condition is related to a condition for deterministic regressor used by Davies [1990] for S-estimators. Our condition is, however, formulated in a slightly different way, which seems to be easier to check for particular regressors. Indeed, we check the condition for a few situations. We give a number of examples with deterministic regressors to illustrate the condition. We also show that the condition is satisfied for stationary regressors and for random walk regressors.

It is worth noting that the innovations are neither required to have a zero expectation nor a continuous density. Thus, the results apply both when the innovations follow a non-contaminated reference distribution with density f_0, say, and when the they are contaminated so that they follow a mixture distribution with density $(1-\epsilon)\mathsf{f}_0 + \epsilon\mathsf{f}_1$, say. The proofs use martingale techniques, chaining arguments and the iterated martingale inequality from Johansen & Nielsen [2016].

References

Chen, X.R. & Wu, Y.H. (1988). Strong consistency of M-estimates in linear models. *Journal of Multivariate Analysis* **27**, 116–130.

Davies, L. (1990) The asymptotics of S-estimators in the linear regression model. *Annals of Statistics* **18**, 1651–1675.

Johansen, S. & Nielsen, B. (2016) Analysis of the Forward Search using some new results for martingales and empirical processes. *Bernoulli* **22**, 1131-1183.

Maronna, R., Martin, D. & Yohai, V. (2006). Robust Statistics: Theory and Methods. John Wiley & Sons, Chichester.

Two-step robust estimation of copulae

S. Guerrier[1], ***S. Orso***[2*] *and M.-P. Victoria-Feser*[2]

[1] *Department of statistics, University of Illinois at Urbana-Champaign; stephane@illinois.edu*
[2] *Research center for statistics, Geneva school of economics and management, University of Geneva; Samuel.Orso@unige.ch and Maria-Pia.VictoriaFeser@unige.ch*
[*] *Presenting author*

Keywords. *Two-step estimation; Indirect inference; Multivariate contamination; Fast bootstrap.*

1 Abstract

Copula is a flexible tool for modeling multivariate random variables. Inference is generally based on multi-step estimators to preserve this flexibility. We address the problem of robustness in this context. It is challenging for many reasons:

- How to built a gross error model that encompasses issues arising in multiple dimensions?
- How to built multi-step robust estimators with good asymptotic properties?
- How to obtain computationally feasible estimators and their inference?

We concentrate our efforts in the two-dimensional case. First, we propose a new gross error model from which the influence function is derived for two-step M-estimators. Second, we prove the strong consistency and asymptotic normality under weak conditions. Third, we propose a fast bootstrap procedure to obtain the covariance matrix of the two-step estimators. We illustrate the estimating procedure with a new R package under development.

Inferences About Robust Correlations With a Percentile Bootstrap Method

A.F. Özdemir[1*]

[1] *Department of Statistics, Dokuz Eylül University, Turkey ; firat.ozdemir@deu.edu.tr*
[*] *Presenting author*

Keywords. *Robust correlation; Heteroscedastic inference; Percentile bootstrap*

A measure of the linear association between two random variables X and Y is a fundamental component of statistical methods. It is clear that the most frequently applied choice in applied work is Pearson's correlation which is very weak in terms of robustness. A very small shift in one of the marginal distributions can have a large effect on it. Wilcox [2012] classified robust analogs of Pearson's correlation into two types: those that protect against outliers among the marginal distributions without taking into account the overall structure of the data (type M), and those that take into account the overall structure of the data when looking for outliers (type O). Percentage bend, biweight, Winsorized, Spearman and Kendall's tau correlations are some members of the first class.On the other hand, the outlier projection (OP) correlation can be given as an example of the second class. All robust correlations given above have conventional hypothesis testing methods of independence but all those methods are sensitive to heteroscedasticity which refers to a situation where the conditional variance of Y varies with X. In this study, the performance of a hypothesis testing procedure based on a percentile bootstrap method which is insensitive to heteroscedasticity was investigated for those six robust correlations in terms of actual significance level and power.

References

Bradley, J. V. (1994). Robustness?. *British Journal of Mathematical and Statistical Psychology*, **31**, 144–152.

Hoaglin, D. C. (1985).Summarizing shape numerically: The g-and-h distribution. In: Haoglin D., Mostseller F., Tukey J. editors. Exploring data tables trends and shapes. Wiley, New York, 461-515.

Hogg, R. V. & Craig, A. T. (1970). Introduction to mathematical statistics. Mcmillian, New York.

Wilcox, R. R. & Muska, J. (2001). Inferences about correlations when there is heteroscedasticity. *British Journal of Mathematical and Statistical Psychology*, **54**, 39–47.

Wilcox, R. R. (1994). The Percentage Bend Correlation Coefficient. *Psychometrika*, **59**(4), 601-616.

Wilcox, R. R. (2004). Inferences based on a Skipped Correlation Coefficient. *Journal of Applied Statistics*, **31**(2),131–143.

Wilcox, R. R. (2012). Introduction to robust estimation and hypothesis testing(3rd edition),Elsevier Academics Press, San Diago, California.

On the bivariate generalized linear exponential distribution

A.K. Pathak[1*]

[1] *Department of Mathematics, Indian Institute of Technology Bombay, Powai, Mumbai 400076, INDIA. ashokiitb09@gmail.com;ashok@math.iitb.ac.in*
[*] *Presenting author*

Keywords. *Bivariate generalized exponential distribution; Linear exponential distribution; Measures of association; Copulas*

In this talk, we deal with a new family of bivariate generalized linear exponential (BGLE) distribution, whose marginals are generalized linear exponential (GLE) distributions. We derive the expressions for some quantities like regression function, product moment and measure of reliability for the BGLE distribution. We also discuss some statistical properties and some measures of dependence. Further, we discuss the associated copula and its various important properties. Finally, the maximum likelihood estimators for the parameters are obtained and real data applications are also discussed.

Inference for Zero Inflated Truncated Power Series Family of Distributions

M.K. Patil[1*]

[1] *Department of Statistics, Padmabhushan Vasantraodada Patil Mahavidyalaya, Kavathe Mahankal, Dist. Sangli, Maharashtra, India, 416405;*
mkpatil_stats@rediffmail.com
[*] *Presenting author*

Keywords. *Zero inflation, Zero inflated Power Series Distribution, Zero Inflated Truncated Power Series Distribution, Zero Inflated Truncated Poisson distribution.*

Zero-inflated data indicates that the data set contains an excessive number of zeros. The word zero-inflation is used to emphasize that the probability mass at the point zero exceeds than the one allowed under a standard parametric family of discrete distributions. Sugiura et al. [2006], Jurečková & Kalina [2012], Jurečková & Milhaud [2003] and Amrhein [1995] have contributed to estimation and testing of the parameters involved in Zero Inflated Power Series Distributions. If the data set under study does not contain observations after some known point in the support, we have to modify Zero Inflated Power Series Distribution (ZIPSD) accordingly in order to get better inferential properties. Zero Inflated Truncated Power Series Distribution (ZITPSD) is one of the better options. In the present work we address problem of estimation for ZITPSD with more emphasis on statistical tests. We provide three asymptotic tests for testing the parameter of ZITPSD, using an unconditional (standard) likelihood approach, a conditional likelihood approach and the sample mean, respectively. The performance of these three tests has been studied for Zero Inflated Truncated Poisson Distribution (ZITPD). Asymptotic Confidence Intervals for the parameter are also provided. The model has been applied to a real life data.

References

Gupta P. L., Gupta R. L. & Tripathi R. C. (1995). Inflated Modified Power Series Distributions with Applications. *Comm. Statist.- Theory Meth.*, **24(9)**, 2355–2374.

Gupta P. L., Gupta R. L. & Tripathi R. C. (1995). Non-Zero-Inflated Modified Power Series Distributions. *Comm. Statist.- Theory Meth.*, **27(12)**, 3047–3064.

Patil, M.K. & Shirke, D. T. (2007). Testing parameter of the power series distribution of a zero-inflated power series model. *Statistical Methodology*, **4**, 393–406.

Patil, M.K. & Shirke, D. T. (2011). Tests for equality of inflation parameters of two zero-inflated power series distributions. *Comm. Statist.- Theory Meth.*, **40(14)**, 2539–2553.

Nonparametric Control Chart Based on Quantiles

V.Y. Pawar[1*]

[1] *Department of Statistics,*
Padmabhushan Dr. Vasantraodada Patil College,
Tasgaon (MS), 416 312 INDIA;
vypawar.stats@gmail.com
[*] *Presenting author*

Keywords. *Nonparametric control chart; Quantiles; Average run length.*

Most of the control charts are based on assumption of normality. Control charts for non-normal process distributions have also been reported in literature. In absence of any knowledge about the process distribution, nonparametric chart is a good alternative. In the recent past number of nonparametric control charts have been studied. Jurečková & Kalina [2012] have proposed a nonparametric synthetic control chart based on signed-rank statistic to monitor process location. In the present work we propose a control chart for monitoring process variability, which is based on in-control quantiles. The chart is motivated from a nonparametric control chart based on in-control quartiles due to Sugiura et al. [2006]. The proposed chart has been studied for its performance for various process distributions to monitor change in variability and has been compared with the existing nonparametric and parametric charts. It has attractive out-of-control Average Run Length performance and is very simple to use. We illustrate the chart through an example and recommend use of this chart to monitor process variability. Generalization of the chart will also be discussed in view to further improve its detection ability.

References

Amin, R. W. Reynolds, M. R. Jr & Bakir, S. T. (1995). Nonparametric Quality Control Charts Based On The Sign Statistic. *Communications in Statistic-Theory and Methods*, **24(6)**, 1597–1623.

Pawar V. Y. & Shirke D. T. (2010). Shewhart-Type Synthetic Control Chart. *Communications in Statistics-Simulation and Computations*, **39**, 1493–1505.

Outlier Detection in Large Sets of Multivariate Time Series

P. Galeano[1] *and* ***D. Peña***[1*]

[1] *Departamento de Estadística, Universidad Carlos III, Madrid, Spain; pedro.galeano@uc3m.es, daniel.pena@uc3m.es*
[*] *Presenting author*

Keywords. *Dynamic Factor Models; Big Data; Additive Outlier; Principal Components, Projection Pursuit; Robust Estimation.*

This article presents a procedure based on projections to find outliers in a large set of multivariate time series. It is assumed that the data have been generated by a Dynamic Factor Model and two types of outliers are considered. Common outliers, generated by the factors, which affect several or all of the time series, and specific or idiosyncratic outliers, which are generated by the specific components and affect a single time series. The outliers are identified by projecting the vector of time series into directions with some optimality properties and searching for univariate outliers in this directions. The procedure is fast to apply and does not require to specify a multivariate model for the data.

Locally robust density estimation and near-parametric asymptotics

S. Penev[1*] *and K. Naito*[2]

[1] *The University of New South Wales, Australia ; s.penev@unsw.edu.au*
[2] *Shimane University, Matsue, Japan; naito@riko.shimane-u.ac.jp*
[*] *Presenting author*

Keywords. *Risk; Robustness; Bregman divergence; Power divergence; Kernel.*

The original application of local likelihood as a truly semiparametric method in density estimation was proposed in papers by Loader [1996] and Hjort and Jones [1996]: "The estimators run the gamut from a fully parametric fit to almost fully nonparametric with only a single smoothing parameter to be chosen". They also give an interpretation of the procedure as one that minimizes, at each value of the argument, the locally weighted Kullback-Leibler divergence between the "true" and the model density. The infusion of local adaptation to the global likelihood by considering maximization of an expression of the form

$$\sum_{i=1}^{n} K\big(\frac{x_i - t}{h}\big) \log\big\{g(x_i, \theta)\big\} \tag{17}$$

with $K(\frac{x-t}{h})$ being some suitable kernel function centered at t and with bandwidth h, and $g(x, \theta)$ is a nominal parametric density. In any version of these modifications,the resulting local maximum likelihood estimator $\hat{\theta}_{t,h}$ could be substituted to obtain the density estimator $g(x, \hat{\theta}_{t,h})$. As opposed to the global parametric model where the substitution of the global maximum likelihood estimator $\hat{\theta}$ automatically results in a density $g(x, \hat{\theta})$, this is not the case with $g(x, \hat{\theta}_{t,h})$ and one needs to normalize to get a shape-preserving density by $\hat{g}_h(x) = g(x, \hat{\theta}_{x,h}) / \int g(t, \hat{\theta}_{t,h}) dt$.

1 From likelihoods to Bregman divergences

When the ideal parametric model does not confidently hold, other divergences are used to replace the Kullback-Leibler divergence. These divergences have been demonstrated to possess good robustness properties relative to maximum likelihood methods. Specific applications for robust density estimation have been considered in Windham [1995] and in Basu et. al. [1998], with Bregman divergence type measures, parameterised by one "robustness control" parameter $\lambda \geq 0$, with $\lambda = 0$ corresponding to no efficiency loss. However none of the mentioned works deals with **local** versions of the robust divergence measures which are our main object.

Since the class of Bregman divergences is very large (and not all of them have found useful applications), we focus on a particular class of them. Starting with the Box-Cox transformation $G_\lambda(x) = \begin{cases} \frac{1}{\lambda}(x^\lambda - 1), \lambda > 0 \\ \log x \end{cases}$ we define $U_\lambda(x) = x(G_\lambda(x) - 1)$. Note that $U_\lambda(x) \to_{\lambda \to 0} x \log x - x$ which is the $U_\lambda(.)$ function that is used to define the von Neumann divergence. The only paper known to us where the case of **large** h has been analysed is Eguchi and Copas [1998] but

it is completely devoted to the local likelihood method. In a nutshell, the results of Eguchi and Copas [1998] show that with respect to the relative entropy risk minimization, there is a benefit of using the local likelihood: little localization "always helps". We demonstrate both theoretically and numerically that such type of statement is also true with respect to robustness: with respect to the Bregman distance based risk minimization, little localization of the globally robust estimator"always helps".

We believe that the localisation proposed here offers a new view towards robustness. In the standard robustness approach (Huber and Ronchetti [2009]) the main focus is on modifying non-robust estimators of *parameters* of certain model density when it is believed that the data was not necessarily generated from the model density because there was contamination. The inference part is essentially finalized once the parameters have been estimated. In our approach we estimate the local features of the density that has generated the data. Our estimated $g(x, \hat{\theta}(x))$ does not in general belong to the class $g(x, \theta), \theta \in \Theta$ and is giving a better idea about the "true" density that has generated the data. On the other hand, there are similarities, too: the belief that the true density is "not too far away" from a model density $g(x, \theta_0)$ is common for both approaches.

References

Basu, A., Harris, I., Hjort, N. and Jones, M.C. (1998) Robust and Efficient Estimation by Minimizing a Density Power Divergence. *Biometrika*, 85, 549–559.

Eguchi, S. and Copas, J. (1998) A Class of Local Likelihood Methods and Near-Parametric Asymptotics. *J. R. Statist. Soc. B*, 60, 709–724.

Hjort, N. and Jones, M.C. (1996) Locally Parametric Nonparametric Density Estimation. *The Annals of Statistics*, 24, 1619–1647.

Huber, P. and Ronchetti, E. (2009) *Robust Statistics*, 2nd Edition. Wiley, New York.

Loader, C. (1996) Local Likelihood Density Estimation. *The Annals of Statistics*, 24, 1602–1618.

Windham, M. (1995) Robustifying Model Fitting. *J. R. Statist. Soc. B*, 57, 599–609.

Fast and robust bootstrap in seemingly unrelated regression models

K. Peremans[1*] *and S. Van Aelst*[1]

[1] *KU Leuven, Department of Mathematics, Celestijnenlaan 200B, 3001 Leuven, Belgium; kris.peremans@wis.kuleuven stefan.vanaelst@wis.kuleuven.be.*
[*] *Presenting author*

Keywords. *Seemingly unrelated regression model; Robust estimator; Influence function; Fast and robust bootstrap.*

Seemingly unrelated regression models generalize ordinary linear regression models by considering multiple regression equations that are linked by contemporaneously correlated disturbances. Traditional estimators allow this feature of correlated error terms, but they are extremely vulnerable to the presence of contamination in the data. Therefore, robust estimators for seemingly unrelated regression models are considered. S-estimators can attain a high breakdown point, but their normal efficiency can be quite low. For that reason, MM-estimators are introduced to obtain estimators that have both a high breakdown point and a high normal efficiency.

Furthermore, the problem of statistical inference is studied. Asymptotic inference relies on assumptions which are hard to verify. Moreover, this inference may be non-robust. Bootstrapping on the other hand, requires less strict assumptions but lacks speed and robustness. Therefore, a fast and robust bootstrap procedure is developed. Robust bootstrap confidence intervals of the unknown parameters in seemingly unrelated regression models are constructed and their performance is analyzed in simulation studies. In addition, hypothesis tests regarding the regression coefficients are carried out by bootstrapping a robust version of the likelihood-ratio statistic. The robust estimators and the fast and robust bootstrap procedure are illustrated on real data.

References

Bilodeau, M. & Duchesne, P. (2000). Robust estimation of the SUR model. *Canadian Journal of Statistics*, **28**, 277–288.

Salibian-Barrera, M. & Zamar, R. H. (2002). Bootrapping robust estimates of regression. *Annals of Statistics*, **30** (2), 556–582.

Van Aelst, S. & Willems, G. (2011). Robust and Efficient One-Way MANOVA Tests. *Journal of the American Statistical Association*, **106** (494), 706–718.

On the Jackknife for Regression Quantiles

S. Portnoy[1*]

[1] *Department of Statistics, University of Illinois, sportnoy@illinois.edu*
[*] *Presenting author*

Keywords. *Inference Delete-d Jackknife*

A recent paper: "The jackknifes edge: Inference for censored regression quantiles" (Portnoy, 2014) showed that the delete-d jackknife enjoyed a serious advantage for inference on censored regression quantiles, especially when the probability argument was not too far from its upper limit. In fact, it appeared that the jackknife also worked much better than expected even for small probability arguments (that is, for very little censoring). This suggested looking at the jackknife for the usual (uncensored) regression quantiles. After briefly summarizing earlier results, some simulation results will be presented. These suggest that indeed the delete-d jackknife works remarkably well when compared with the best methods available in the `quantreg` R-package. Here d could be set to exactly twice the square root of n, the sample size; and so no empirical selection of the tuning parameter was needed (at least for n not very large). The reasons for this relatively favorable result are less clear than for the case of censored regression quantiles, but the results suggest the need for further (more general) investigation.

References

Portnoy, S. (2014). The jackknife's edge: Inference for censored regression quantiles. Computational Statistics & Data Analysis, Vol. 72, 273–281.

Reliability Estimation for Generalized Inverted Scale Family of Distributions

K.G. Potdar[1*]

[1] *Department of Statistics, Ajara Mahavidyalaya, Ajara, Dist-Kolhapur, Maharashtra, India, 416505;*
potdarkiran.stat@gmail.com
[*] *Presenting author*

Keywords. *Generalized inverted half-logistic distribution; Generalized inverted Rayleigh distribution; Maximum likelihood estimation; Confidence interval; Reliability estimation.*

Estimation of reliability plays an important role in various fields, which include medical science, industry, research and development studies, etc. Reliability computation is generally based on specific lifetime distribution suitable to model of the data under consideration, which contains unknown parameter(s). Potdar & Shirke [2013] introduced a new family of lifetime distributions namely generalized inverted scale family of distributions. Generalized inverted exponential distribution, generalized inverted Rayleigh distribution, generalized inverted half-logistic distribution etc. are some members of this family. In the present work, estimation of parameters and reliability function have been considered for members of generalized inverted scale family of distributions. Simulation studies are conducted to evaluate performance of proposed estimation procedures. Illustrations with real data are also provided.

References

Potdar K. G. & Shirke, D. T. (2013). Inference for the parameters of generalized inverted family of distributions. *ProbStat Forum*, **6**, 18–28.

Robust Estimation and Variable Selection for High-Dimensional Linear Regression

S. Li[1], ***Y. Qin***[1*], *Y. Li*[2] *and Y. Yu*[1]

[1] *University of Cincinnati, Cincinnati, Ohio, USA 45221; qinyn@ucmail.uc.edu, lis6@mail.uc.edu, yuyu@ucmail.uc.edu.*
[2] *Renmin University of China, Beijing, China 100872; yang.li@ruc.edu.cn.*
[*] *Presenting author*

Keywords. *Penalized estimation; Adaptive lasso; High-dimensional data.*

In this article, we introduce a class of robust linear regression estimators for variable selection in presence of outliers. Consider a linear model $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i$, $i = 1, \ldots, n$ where $(y_i, \mathbf{x}_i)$ represents the ith observation and $y_i \in \mathbb{R}, \mathbf{x}_i \in \mathbb{R}^d$. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$ is an unknown regression coefficient vector. ϵ_i is an iid random error that is independent from $\mathbf{x}_i$ and follows a symmetric parametric distribution $f(\cdot)$ with mean 0 and constant variance σ^2. $f(\cdot)$ is assumed to be a Gaussian probability density function. Usually some of the elements in $\boldsymbol{\beta}$ are zeros. To select only important variables and estimate their coefficients robustly, we propose the following penalized likelihood estimator,

$$\hat{\boldsymbol{\beta}}_t = \arg\max_{\beta\in\mathbb{R}^d} \left\{ \underbrace{\sum_{i=1}^{n} \ln_t(f(y_i - \mathbf{x}_i^T\boldsymbol{\beta}))}_{\text{robustified likelihood}} - \underbrace{n\sum_{j=1}^{d} p_{\lambda_{nj}}(\beta_j)}_{\text{penalty}} \right\},$$

where $\ln_t(\cdot)$ is defined as: $\ln_t(u) = \ln(t) + \sum_{k=1}^{K} \frac{\ln^{(k)}(t)}{k!}(u-t)^k$ if $u < t$, and $\ln_t(u) = \ln(u)$ if $u \geq t$ or $t = 0$. Here, $t \geq 0$ is a tuning parameter and $\ln_t(u)$ is essentially a K-th order Taylor expansion of $\ln(u)$ for $u < t$. By introducing this tuning parameter t, we robustify the log-likelihood function so that it becomes insensitive to perturbation to the data [Wang et al., 2013]. Note that when $t \to 0$, $\ln_t(u) \to \ln(u)$, therefore, the proposed estimator includes the penalized least square estimator as a special case with $t = 0$.

When solving the optimization program, we essentially solve a weighted likelihood equation where observations that disagree with the assumed model receive low weights. For example, when $K = 1$, the first order condition on the robustified likelihood becomes $0 = \sum_{i=1}^{n}[\frac{\partial}{\partial\boldsymbol{\beta}} \ln(f(y_i - \mathbf{x}_i^T\boldsymbol{\beta}))] \min(1, f(y_i - \mathbf{x}_i^T\boldsymbol{\beta})/t)$. Therefore, observations whose likelihoods are below t (which more likely turn out to be outliers) receive only partial weights whereas other observations receive full weights.

In the linear regression setting, the proposed penalized estimator obtains remarkable robustness when data is contaminated and still performs well when the model is correctly specified. One can control the estimator's robustness by adjusting t. When $t \to 0$, the proposed estimator becomes the traditional penalized least square estimator. When t is sufficiently large, the proposed estimator becomes the penalized minimum L_2 distance estimator. With a moderate t, the proposed estimator can be considered as a mixture of penalized Kullback-Leibler distance estimation and penalized L_2 distance estimation, where the former is known for its desirable asymptotic properties and the latter is known for its remarkable robustness.

Table 2: Monte Carlo Simulation

n	Method	CSR	Under Fitted	Over Fitted	Miss Fitted	Model Error	
						Median	MAD
100	Proposed	0.995	0	0.005	0	0.053	0.027
	LAD	0.989	0	0.011	0	0.097	0.050
200	Proposed	1.000	0	0.000	0	0.024	0.012
	LAD	0.998	0	0.002	0	0.043	0.020
400	Proposed	1.000	0	0.000	0	0.012	0.006
	LAD	1.000	0	0.000	0	0.021	0.011

We further show that the proposed estimator is consistent and enjoys oracle property. We also establish the bound of L_2 norm of the estimation error. Furthermore, the proposed estimator achieves the highest asymptotic breakdown point of 1/2 and is equipped with a bounded influence function. In addition, we have proposed a method for adaptively selecting the tuning parameter t to guarantee the robustness as well as the its asymptotic properties.

We conduct simulation studies to demonstrate the advantage of the proposed method over the traditional method in Table 2. We generate $\mathbf{x}_i \overset{i.i.d.}{\sim} 0.8N(\mathbf{0}, \Omega_1) + 0.2N(\mathbf{2}, \Omega_2)$ where $\Omega_1 = \mathbf{I}_d, \Omega_2 = \Sigma_{d\times d}$ with $\{\Sigma\}_{ij} = 0.5^{|i-j|}$ and $\epsilon_i \overset{i.i.d.}{\sim} 0.8N(0,1) + 0.2N(10, 6^2)$, and obtain y_i by the linear regression. We apply both the proposed estimator and the least absolute deviations (LAD) estimator with adaptive lasso penalty on the simulated data and compare their variable selection performance based on the proportions of correctly selecting (i.e. CSR), under fitting, over fitting, and miss fitting the true model. We also compare the estimation performance based on model error, $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E[\mathbf{x}\mathbf{x}^T](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. As the table shows, the proposed method outperforms LAD in different scenarios in terms of both selection accuracy and estimation error.

References

Wang, X., Jiang, Y., Huang, M., Zhang, H. (2013). Robust Variable Selection with Exponential Squared Loss, *J. Amer. Statist. Assoc.*, 108(502), 632-643

Finding Outliers in Surface Data and Video

M. Hubert[1], ***J. Raymaekers***[1*], *P.J. Rousseeuw*[1] *and P. Segaert*[1]

[1] *KU Leuven, Belgium; mia.hubert@wis.kuleuven.be, jakob.raymaekers@wis.kuleuven.be, peter@rousseeuw.net, pieter.segaert@wis.kuleuven.be.*
[*] *Presenting author*

Keywords. *Adjusted outlyingness; Functional data; Image data; Multiway; Robustness.*

Surface, image and video data can be considered as functional data with a bivariate domain. It is well known that classical statistical techniques to detect outlying surfaces or to flag outlying parts of a surface are not trustworthy as the analysis itself may be distorted by the outliers.

We propose a new display based on the mean and the variability of the degree of outlyingness at each grid point. A rule is constructed to flag the outliers in the resulting functional outlier map. Heatmaps of their outlyingness indicate the regions which are most deviating from the regular surfaces. Using projection pursuit techniques, the method is applicable to univariate as well as multivariate surface data.

To illustrate the performance of the method, it is applied to fluorescence excitation-emission spectra after fitting a PARAFAC model, to MRI image data which are augmented with their gradients, and to video surveillance data.

References

Hubert, M., Raymaekers, J., Rousseeuw, P.J. & Segaert, P. (2016). Finding Outliers in Surface Data and Video. *arXiv*:1601.08133 .

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

Maximum likelihood framework for computing robust statistics

M. Rohan[1*] *and M. Jorgensen*[2]

[1] *Department of Biostatistics and Epidemiology, Auckland University of Technology, Auckland, New Zealand; mrohan@aut.ac.nz*
[2] *Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand; mjorgensen@aut.ac.nz*
[*] *Presenting author*

Keywords. *M-estimators; Finite Mixture; EM algorithm.*

Robust methods are often used to estimate model parameters when outliers are present. In contrast with the classical statistical modelling, the methodology of computing robust statistics is often *ad hoc* and lacks a unified approach. We developed a method using likelihood to make statistical modelling more robust for addressing above mentioned issues. To develop this method, the finite mixture form is being used as a mathematical tool and the weights for observations are computed at the E-step of EM algorithm. For promoting our algorithm and simplification, we explain the method of computing robust statistics for the linear model parameters. Later we will discuss results from some simulation and the real example of well known speed-of-light data. Finally, we compare our approach with classical robust methods, but our reason for developing the framework lies in its applicability to maximum likelihood modelling in general.

Detecting anomalous data cells

P.J. Rousseeuw[1*] *and W. Van den Bossche*[1]

[1] *KU Leuven, Belgium; peter@rousseeuw.net; W.VandenBossche@wis.kuleuven.be*
[*] *Presenting author*

Keywords. *Cellwise outliers; High dimension; Missing values; Multivariate statistics; Rowwise outliers.*

A multivariate dataset consists of n cases in d dimensions, and is often stored in an n by d data matrix. It is well-known that real data may contain outliers. Depending on the circumstances, outliers may be (a) undesirable errors which can adversely affect the data analysis, or (b) valuable nuggets of unexpected information. In statistics and data analysis the word outlier usually refers to a row of the data matrix, and the methods to detect such outliers only work when at most 50% of the rows are contaminated. But often only one or a few cells (coordinates) in a row are outlying, and they may not be found by looking at each variable (column) separately. We propose the first method to detect anomalous data cells which takes the correlations between the variables into account. It has no restriction on the number of contaminated rows, and can deal with high dimensions. Other advantages are that it provides estimates of the 'expected' values of the outlying cells, while imputing the missing values at the same time. We illustrate the method on several real data sets, where it uncovers more structure than found by purely columnwise methods or purely rowwise methods. Following the approach of Agostinelli et al. [2015], the proposed method can also serve as an initial step for estimating multivariate location and scatter matrices.

References

Agostinelli, C., Leung, A., Yohai, V.J. & Zamar, R.H. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test*, **24**, 441–461.

Rousseeuw, P.J. & Van den Bossche, W. (2016). Detecting anomalous data cells. *arXiv*:1601.07251 .

Robustness for Dynamic Models for Extreme Values

P. Ruckdeschel[1*], *B. Spangl*[2], *S. Desmettre*[3], *A. Mändle*[1], *and K. Rohmeyer*[1,4]

[1] *Institute for Mathematics, Oldenburg University, Oldenburg, Germany; peter.ruckdeschel@uni-oldenburg.de, andreas.maendle@uni-oldenburg.de, kornelius.rohmeyer@uni-oldenburg.de*
[2] *Institute of Applied Statistics and Computing, BOKU, Vienna, Austria; bernhard.spangl@boku.ac.at*
[3] *Department of Mathematics, Kaiserslautern University, Kaiserslautern, Germany; desmettre@mathematik.uni-kl.de*
[4] *Institute for Statistics, Bremen University, Bremen, Germany*
[*] *Presenting author*

Keywords. *GLM; Dynamic model; Extremes; Hydrology; Goodness of fit.*

River discharge data exhibit interesting dynamic patterns as to the occurrence of extreme events. For i.i.d. data the respective limit theorems of extreme value statistics give rise to and convincingly suggest parsimoniously parametrized models, i.e., the Generalized Pareto (GPD) and the Generalized Extreme Value distributions.

Passing over to dynamic situations, respective distributional limit theorems do exist as well, but parsimonious parametric models are less evident, in particular when a good fit of the empirical interarrival times is also of interest.

We discuss four different approaches to tackle this issue, i.e., dynamics introduced by (a) a state-space model for location and scale, (b) a shot-noise-type process with GPD marginals, (c) a copula-based autoregressive model with GPD marginals, and (d) a generalized linear model with GPD marginals (and previous extremal events as regressors) including a corresponding regression for the shape of the GPD as well.

In each of these models we discuss respective robustness issues and robust procedures, and some techniques to decide among the different approaches based on corresponding goodness of fit tests.

Robustness, Information Geometry and Sparse Goodness-of-Fit Testing

R. Sabolová[1*], *P. Marriott*[2], *G. Van Bever*[1] *and F. Critchley*[1]

[1] *The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom; radka.sabolova@open.ac.uk, germain.van-bever@open.ac.uk, f.critchley@open.ac.uk*
[2] *University of Waterloo, 200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada; pmarriott@uwaterloo.ca*
[*] *Presenting author*

Keywords. *Computational Information Geometry; Sparse Multinomial Model.*

Recent work in computational information geometry has the potential to provide a universal treatment of robustness, dealing with all possible models for a discrete data space. In particular, both local and global perturbations are naturally accommodated, while geometric considerations are to the fore. Within this general setting, the focus here is on the challenging issue of goodness-of-fit testing in the high dimensional, low sample size context where, potentially, boundary effects dominate. Three distinct issues are explored. First, we show how to implement a novel asymptotic-in-dimension expansion to the sampling distribution of the deviance. Secondly, we rigorously explore the relationship between information-geometric based divergences and standard goodness-of-fit statistics. Finally, we explore the Fisher metric geometry and statistical curvature near the boundary. In particular we show how, in this context, discretisation effects can dominate in sampling distributions. This contrasts markedly with skewness and other corrections flowing from standard higher-order asymptotic analysis. Simulation exercises are used throughout to validate and illustrate our theoretical results.

Robust Functional Principal Components with sparse observations

G. Boente[1], ***Matias Salibian-Barrera***[2*] *and Jane-Ling Wang*[3]

[1] *Universidad de Buenos Aires and IMAS, CONICET; gboente@dm.uba.ar.*
[2] *The University of British Columbia; matias@stat.ubc.ca*
[3] *University of California at Davis; janelwang@ucdavis.edu*
[*] *Presenting author*

Keywords. *Robust principal components, Functional data, Smoothing.*

Principal components analysis is a widely used technique that provides an optimal lower-dimensional approximation to multivariate observations. Similarly, functional principal components analysis allows us to obtain parsimonious predictions for each trajectory in the sample. A new characterization of elliptical distributions on separable Hilbert spaces (Boente *et al.* [2014]) helps us show that this property holds even when second moments do not exist. If these functional principal components are estimated robustly, the resulting lower-dimensional approximations can be very useful in identifying potential outliers among high-dimensional or functional observations.

In this talk we discuss the problem of robust estimation of functional principal components when only a few observations are available per curve. Specifically, we observe $X_i(t_{ij})$, i=1, ..., n, j=1, ..., n_i, where the n_i's can be small (e.g. between 1 and 5) for all curves. Many available methods to estimate functional principal components rely on a smoothing step of the observed trajectories, and thus require many observations per curve. A notable exception is the conditional expectation approach of Yao *et al.* [2005], which estimates the covariance function by smoothing the sparsely available cross-products, and thus is able to "combine information" from many curves. A first attempt at protecting this approach from potential outliers by using a robust smoother on the cross-products does not work because the distribution of the cross-products is generally asymmetric. However, when the stochastic process has an elliptical distribution, one can exploit the linear structure of the conditional distribution of $X_i(t)|X_i(s)$ as a function of $X_i(s)$ to obtain robust estimators of the scatter function $G(t, s)$ of the underlying random process. Furthermore, this approach allows us to use a robust smoother to combine observations at neighbouring points (t^*, s^*). In this talk, we report initial numerical experiments comparing the performance of the resulting estimates and existing alternatives.

References

Boente, G., Salibian-Barrera, M. and Tyler, D. (2014) A characterization of elliptical distributions and some optimality properties of principal components for functional data. *Journal of Multivariate Analysis*, **131**, 254-264.

Yao, F., Müller, H.-G., Wang, J.-L., (2005), Functional Data Analysis for Sparse Longitudinal Data, *Journal of the American Statistical Association*, **100**, 577-590.

Some tests of independence between two random vectors in arbitrary dimensions

S. Sarkar[1*], *M. Biswas*[2] *and A. K. Ghosh*[1]

[1] *Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata.*
[2] *Department of Statistics, Brahmananda Keshab Chandra College, Kolkata.*
email : sohamsarkar1991@gmail.com, munmun.biswas08@gmail.com, akghosh@isical.ac.in.
[*] *Presenting author*

Keywords. *Distribution-free test; High-dimensional data; Inter-point distance; Minimal spanning tree; Multiscale approach*

Given a random sample from the distribution of $\mathbf{Z} = [\mathbf{X}'\,\mathbf{Y}']'$, we want to test whether $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$ are independent. There are several methods for this test in the literature, both in parametric and nonparametric regime. Most of these methods are not applicable when the dimension of the data is larger than the sample size. Recently some new methods have been proposed for this problem, which are based on inter-point distances of $\mathbf{X}$- and $\mathbf{Y}$-samples. Among them Heller et al. [2012] proposed some tests using random traversal on minimal spanning trees, which are exactly distribution-free under the null hypothesis of independence. We identify some shortcomings of these tests and propose some modifications which rectify these problems, keeping the distribution-free property intact. Several simulated and real data sets are analyzed which demonstrate the superiority of our proposed methods.

However, all these above mentioned distribution-free tests sacrifice a lot of valuable information to achieve distribution-free property. Moreover, they may not yield the same result if the roles of $\mathbf{X}$ and $\mathbf{Y}$ are interchanged. We propose some new tests based on nearest neighbors, which use the unused information to come up with better performance, and are symmetric with respect to $\mathbf{X}$ and $\mathbf{Y}$. Multiscale versions of these tests are also developed. Performances of all these tests are evaluated using several simulated and real data sets.

Since our tests are based on ranks of inter-point distances, they are applicable to high-dimensional data and even for functional data taking values in infinite dimensional Banach spaces.

References

Heller, R., Gorfine, M., & Heller, Y. (2012). A class of multivariate distribution-free tests of independence based on graphs. *J. Statist. Plann. Inf.*, **142**, 3097-3106.

Robust bootstrap procedures for claims reserving using Generalized Linear Models

K. Peremans[1*], ***P. Segaert***[1*], *S. Van Aelst*[1] *and T. Verdonck*[1]

[1]*KU Leuven, Belgium;*
kris.peremans@wis.kuleuven.be, pieter.segaert@wis.kuleuven.be,
stefan.vanaelst@wis.kuleuven.be, tim.verdonck@wis.kuleuven.be
[*]*Presenting author*

Keywords. *Bootstrap – Generalized Linear Models – Solvency II*

Insurers are faced with the challenge of estimating the future reserves needed to handle historic and current claims that are not fully settled. Settlement delays may occur to long legal trials or medical complications. The future reserves may be estimated using generalized linear models using so called run-off triangles. However due to the specific nature of these run-off triangles it is typically difficult to derive analytic expressions for the standard deviation of the resulting reserve estimates. A popular alternative for obtaining standard deviations is then to use the bootstrap technique.

Traditional bootstrap procedures are however very sensitive to the possible presence of outliers. Even when bootstrapping a robust estimator, breakdown may occur as a bootstrap sample may contain a higher percentage of outliers than the original sample. Therefore we discuss and implement several robust bootstrap procedures in the claims reserving framework and we investigate and compare their performance on both simulated and real data.

References

Amado, C. and Pires, A. M. (2004). Robust bootstrap with non random weights based on the influence function. *Communications in Statistics - Simulation and Computation*, **33(2)**:377–396.

England, P. D. and Verrall, R. J. (2006). Predictive distributions of outstanding liabilities in general insurance. *Annals of Actuarial Science*, 1:221–270.

Salibian-Barrera, M. and Zamar, R. H. (2002). Bootstrapping robust estimates of regression. *The Annals of Statistics*, **30(2)**:pp. 556–582.

Verdonck, T., Wouwe, M. V., and Dhaene, J. (2009). A robustification of the chain-ladder method. *Nort American Actuarial Journal*, **13**(2):280 – 298.

Progressive iterative quantile thresholding for robust estimation beyond Gaussianity

Y. She[1*]

[1] *Florida State University, Tallahassee, FL 32306-4330; yshe@stat.fsu.edu*
[*] *Presenting author*

Keywords. *High-dimensional statistics; Non-Gaussian robust estimation; Outlier detection; Non-asymptotics*

The work studies how to obtain a high break-down robust estimator in non-Gaussian applications. We propose an approach for joint robust estimation and outlier detection after formulating the problem with sparse mean-shift parametrization into a high-dimensional shrinkage estimation. A simple and scalable quantile thresholding-based iterative procedure is developed for optimization. We show that the approach extends the least trimmed squares and has theoretical guarantees of accuracy and robustness. Moreover, subset sampling is not required in some less challenging problems. The nonasymptotic robust analysis reveals the power of progressive quantile thresholding. The technique applies to high-dimensional estimation.

Tests based on notion of data depth for testing equality of locations

D.T. Shirke[1*]

[1] *Department of Statistics,*
Shivaji University,
Kolhapur (MS)- 416004 INDIA;
dts_stats@unishivaji.ac.in
[*] *Presenting author*

Keywords. *Data depth; Nonparametric tests; Permutation test.*

A notion of data depth has been used to measure centrality or outlyingness of a given point in a given data cloud. The DD plots (depth vs. depth plot) introduced by Jurečková & Kalina [2012] is a two dimensional graph which is useful diagnostic tool for comparing two multivariate populations. Based on DD plot Sugiura et al. [2006] have provided tests for testing equality of locations of two populations. In the present work, we discuss various features of DD plot with respect to change in location of two populations. Limitations of the tests proposed by Sugiura et al. [2006] are discussed and modifications to the tests are provided. Measures to discriminate one sample from the other with regard to change in location are provided. Based on these measures, tests for equality of locations of two multivariate populations are proposed. These tests are implemented through idea of permutation test. Performance of proposed tests is studied by simulation. Illustration with real data is also provided. We further discuss extension of the DD-plots and related tests for more than two populations.

References

Li, J & Liu, R. Y. (2004). New nonparametric tests of multivariate locations and scales using data depth. *Statist. Sci.*, **19**, 686–696.

Li, R., Parelius, J. & Singh, K. (1999). Multivariate analysis by data depth: Descriptive Statistics, Graphics, Inference. *Ann. of Statist.*, **27**, 783–858.

Recent Advances in Directional Multiple-Output Quantile Regression

P. Boček[1] *and* ***M. Šiman***[1*]

[1] *The Institute of Information Theory and Automation of the Czech Academy of Sciences, Pod Vodárenskou věží 4, CZ-182 08 Prague 8, Czech Republic; bocek@utia.cas.cz, siman@utia.cas.cz*
[*] *Presenting author*

Keywords. *Quantile regression; Multiple-output regression; Multivariate quantile; Data depth; Regression depth.*

The standard single-response quantile regression [Koenker, 2005] has already become a powerful tool for econometric data analysis. Its extension for vector responses is, therefore, highly desirable because the reality tends to be hopelessly multivariate. A few such multiple-output generalizations has already been proposed by means of various approaches, usually based on projections, ellipsoids, hyperspherical coordinates or measure transportation.

The presentation highlights some recent achievements in the research of the two directional multiple-output quantile regression methods of Hallin et al. [2010] and Paindaveine & Šiman [2011], e.g., some properties of their weighted and local polynomial generalizations [Boček and Šiman, 2016c] (see also Hallin et al. [2015]), their use for defining process capability indices [Šiman, 2014a,b] and promising inferential statistics [Šiman, 2011], and their software implementation in Octave [Boček & Šiman, 2016a] and R [Boček & Šiman, 2016b], based on the algorithms described in Paindaveine & Šiman [2012a] and Paindaveine & Šiman [2012b].

References

Boček, P. & Šiman, M. (2016a). Directional quantile regression in Octave (and MATLAB). *Kybernetika*, in press.

Boček, P. & Šiman, M. (2016b). Directional quantile regression in R. Submitted.

Boček, P. & Šiman, M. (2016c). On weighted and locally polynomial directional quantile regression. Submitted.

Hallin, M., Paindaveine, D. & Šiman, M. (2010). Multivariate quantiles and multiple-output regression quantiles: from L_1 optimization to halfspace depth. *Annals of Statistics*, **38**, 635–669.

Hallin, M., Lu, Z., Paindaveine, D. & Šiman, M. (2015). Local bilinear multiple-output quantile/depth regression. *Bernoulli*, **21**, 1435–1466.

Koenker, R. (2005). Quantile Regression. Econometric Society Monograph Series. Cambridge University Press, New York.

Paindaveine, D. & Šiman, M. (2011). On directional multiple-output quantile regression. *Journal of Multivariate Analysis*, **102**, 193–212.

Paindaveine, D. & Šiman, M. (2012a). Computing multiple-output regression quantile regions. *Computational Statistics & Data Analysis*, **56**, 840–853.

Paindaveine, D. & Šiman, M. (2012b). Computing multiple-output regression quantile regions

from projection quantiles. *Computational Statistics*, **27**, 29–49.

Šiman, M. (2011). On exact computation of some statistics based on projection pursuit in a general regression context. *Communications in Statistics - Simulation and Computation*, **40**, 948–956.

Šiman, M. (2014a). Precision index in the multivariate context. *Communications in Statistics - Theory and Methods*, **43**, 377–387.

Šiman, M. (2014b). Multivariate process capability indices: a directional approach. *Communications in Statistics - Theory and Methods*, **43**, 1949–1955.

A General and Robust Framework for Secondary Traits Analysis

X. Song[1*], *I. Ionita-Laza*[2], *M. Liu*[3], *J. Reibman*[4], *and Y. Wei*[2]

[1] *Heilbrunn Department of Population and Family Health, Columbia University, New York, NY 10032; xs2148@cumc.columbia.edu*

[2] *Department of Biostatistics, Columbia University, New York, NY 10032*

[3] *Department of Population Health, New York University School of Medicine, New York, NY 10016*

[4] *Department of Medicine, New York University School of Medicine, New York, NY 10016*

[*] *Presenting author*

Keywords. *Secondary trait analysis; Estimating equations; Case-control studies*

Background/Aims: Case-control designs are commonly employed in genetic association studies. In addition to the primary trait of interest, data on secondary traits are often collected. Directly regressing secondary traits on genetic variants from a case-control sample often leads to biased estimation. Several statistical methods have been proposed to address this issue. Among them, the Inverse-Probability-Weighting (IPW) approach and the semi-parametric maximum likelihood (SPML) approach are the most commonly used. **Methods**: A new weighted estimating equation (WEE) approach is proposed to provide unbiased and robust estimation of genetic associations with secondary traits, by combining observed and counter-factual outcomes. Compared to the existing approaches, the new WEE approach is more robust against biased sampling and disease model misspecification. **Simulations**: We conducted simulations to evaluate the performance of the WEE approach and compare with the IPW and SPML under various models and sampling schemes. The WEE approach demonstrated robustness in all scenarios investigated, had appropriate Type I error, and was as powerful or more powerful than the IPW and SPML approaches. Applications: We applied our new approach in an asthma case-control study to estimate the association between thymic stromal lymphopoietin (TSLP) gene and two secondary traits, overweight status and serum IgE level. For the binary trait overweight, the WEE approach identified two associated SNPs in logistic regression; for the continuous trait log serum IgE levels, the WEE approach identified three associated SNPs in linear regression, and additional four SNPs that are missed in mean regression to be associated with the 75th quantile of IgE in quantile regression. **Conclusion**: The proposed WEE approach provides a general and robust secondary analysis framework. It complements the existing approaches, and can be valuable in identifying new associations with secondary traits.

The new confidence set method for statistical classification

N. Srimaneekarn[1*] *and W. Liu*[1]

[1] *Mathematical Sciences, University of Southampton; ns4e12@soton.ac.uk, w.liu@soton.ac.uk.*
[*] *Presenting author*

Keywords. *Classification; Classification Method; Confidence Set*

To classify a new case into its true class, based on some measurements, is an important problem of statistical classification. Five classification methods have been studied. They are logistic regression, classification tree, Bayesian method, support vector machine and the new confidence set method. The new method constructs a confidence set for the true class for a new case by inverting the acceptance sets. The advantage of this method is that the probability of correct classification is not less than $1-\alpha$. The methods are illustrated specifically with the well-known Iris data and applied to a data set for classifying patients as normal, having fibrosis or having cirrhosis based on some measurements on blood samples. The total misclassification error and sensitivity (true positive rate) are used for comparing the methods.

Exploring the robustness of robustness concepts and approaches

W.A. Stahel[1*]

[1] *ETH Zurich; stahel@stat.math.ethz.ch.*
[*] *Presenting author*

Keywords. *Robustness criteria, classes of robust estimators, nuisance parameters, robustification*

After fifty years of development of robust statistics, there is a zoo of procedures and a variety of criteria to measure their merits. Emphasis on one criterion has often lead to methods which had considerable flaws under other aspects. I will discuss some approaches and concepts in robust statistics with a historical perspective and an intention to reach an overview of their merits and limitations.

Robust Inequality Measures

R.G. Staudte[1*] *and L.A. Prendergast*[1]

[1] *La Trobe University, Melbourne 3086 Australia;*
r.staudte@latrobe.edu.au, luke.prendergast@latrobe.edu.au
[*] *Presenting author*

Keywords. *Gini index; Influence function; Density quantile*

The Lorenz curve and the associated Gini coefficient are routinely employed for comparisons of income inequality in various countries. These concepts have nice mathematical properties, and thus are the subject of numerous theoretical studies. But when it comes to statistical inference for them, thorny issues arise: in particular, Cowell and Victoria-Feser [1996] show that these and many other inequality measures in the econometrics literature have unbounded influence functions. While robust methods are available for income distributions with heavy tails, usually only grouped data are available for privacy reasons. To deal with such situations, we redefine the basic concept of the Lorenz curve in terms of quantiles instead of moments, and see what has been gained and lost in terms of conceptual clarity, inference and estimator resistance to contamination.

Let x_p be the pth quantile of a population of incomes, and let μ_p be the mean of smallest incomes $x \le x_p$. Then if $\mu = \lim \mu_p$ exists as p approaches one, the Lorenz curve is defined by $L_0(p) = p\,\mu_p/\mu$ for all $0 < p < 1$. What we propose is to replace in this definition the mean μ_p by the median $x_{p/2}$ of incomes $x \le x_p$; and, in the denominator to replace the mean μ by one of three quantities: $d_1(p) = x_{0.5}$, $d_2(p) = x_{1-p/2}$ or $d_3(p) = (x_{p/2} + x_{1-p/2})/2$. This leads to three quantile based versions of the Lorenz curve $L_i(p) = p\,x_{p/2}/d_i(p)$ for $i = 1, 2$ and 3. The associated inequality coefficients G_i are defined as one minus twice the area between the graph of L_i and the diagonal line for $i = 0, 1, 2$ and 3.

The quantile inequality curves measure inequality in that for any non-decreasing transfer of income function that does not increase the distance of any quantile from the median, the curve ordinates $L_i(p)$ can only increase for each p, which means that the corresponding coefficient of inequality can only decrease. The quantile inequality curves have most of the nice properties of the Lorenz curve, including convexity for all income models commonly used in the literature. In addition, they and the associated inequality coefficients have bounded influence functions.

By substituting sample quantile estimates into the formulae defining the L_i and G_i one obtains estimators $\hat{L}_i$ and $\hat{G}_i$ with standard errors depending on unknown quantile densities as well as the quantile function. Using quantile density estimators from Prendergast and Staudte [2016a], this leads to large-sample distribution-free confidence intervals for the quantile inequality coefficients that have good coverage probabilities, even for moderate sample sizes, Prendergast and Staudte [2016b].

The Gini coefficient has been criticized for placing too much emphasis on the middle incomes, and the quantile versions can be criticized for the same reason. To see why, let Y_1, Y_2 be independent, random incomes less than the median, and let $V = \max\{Y_1, Y_2\}$. Then by a

change of variable, one finds $G_1 = \mathrm{E}[(m - V)/m]$. That is, G_1 is the average relative distance of V from the median. Next define $W = x_{1-r}$ to be the $(1-r)$th quantile of incomes whenever $V = x_r$ is the rth quantile of incomes. Then it follows that $G_2 = \mathrm{E}[(W - V)/W]$ and $G_3 = \mathrm{E}[(W - V)/(V + W)]$.

The maximum $V = \max\{Y_1, Y_2\}$ arises because of the multiplier p in the definition of $L_i(p)$, as one can see by omitting it. For example, if L_1 were redefined to be $L_1^*(p) = x_{p/2}/x_{0.5}$ taking values in [0,1], and G_1 redefined to G_1^*, the area between $L_1^*(p)$ and the horizontal line over the unit interval, then $G_1^* = (m - \mathrm{E}[Y])/m$, where Y has the conditional distribution of X, given that X is less than its median. Thus G_1^* is the average relative distance of a *single* randomly chosen income less than the median from the median.

If one wants an inequality curve and associated coefficient of inequality that do not suffer from down-weighting small incomes, and is willing to give up the property of convexity, then L_1^*, G_1^* is a possibility. The analogous definitions L_2^*, G_2^* and their robust estimators are investigated in Prendergast and Staudte [2016c].

References

F.A. Cowell and M.P. Victoria-Feser. Robustness properties of inequality measures. *Econometrica*, 64(1):77–101, 1996.

L.A. Prendergast and R.G. Staudte. Exploiting the quantile optimality ratio in finding confidence intervals for quantiles. *Stat*, 5:70–81, 2016a. DOI: 10.1002/sta4.105.

L.A. Prendergast and R.G. Staudte. Quantile versions of the Lorenz curve. *http://arxiv.org/abs/1510.06085*, 2016b.

L.A. Prendergast and R.G. Staudte. A simple and Effective Inequality Measure. *http://arxiv.org/abs/1603.03481*, 2016c.

On estimation of the amount of sparsity in normal mixture models

N. Stepanova[1*] *and Y. Wang*[1]

[1] *School of Mathematics and Statistics, Carleton University, Ottawa, ON, K1S 5B6 Canada; nstep@math.carleton.ca, YiboWang@cmail.carleton.ca.*

[*] *Presenting author*

Keywords. *Normal mixture models; Sparse normal means; Minimax estimation*

This work is motivated by the problem of variable selection in high-dimensional linear regression models and Gaussian sequence models with some sparsity pattern. For such a problem, sharp conditions for the possibility of exact and almost full variable selection are available (see, for example, Genovese et al. [2012]). These conditions, as well as an asymptotically minimax procedure that provides almost full selection, depend on the amount of signal contained in the data, which is generally unknown. In this talk, we present a new estimator for the fraction of nonzero means in a Gaussian sequence model with relatively few nonzero means that are only moderately large. We show that, in the region where variable selection is possible, the new estimator dominates (in terms of the minimax rate of convergence) the estimator of Cai et al. [2007], that was proposed earlier in a similar context. Moreover, our estimator (nearly) attains the optimal rate of convergence in this setting. The results obtained analytically are supported by simulations.

References

Cai, T., Jin, J. & Low, M. (2007). Estimation and confidence sets for sparse normal mixtures. *Annals of Statistics*, **35**, 2421–2449.

Genovese, C.R., Jin, J., Wasserman, L., & Yao, Z. (2012). A comparison of the lasso and marginal regression. *Journal of Machine Learning Research*, **13**, 2107–2143.

Robust statistics by means of scaled Bregman distances

W. Stummer[1*]

[1] *Department of Mathematics, University of Erlangen-Nürnberg (FAU), Cauerstrasse 11, 91058 Erlangen, Germany; stummer@math.fau.de*
[*] *Presenting author*

Keywords. *Scaled Bregman distances; Kullback-Leibler information.*

1 Abstract

We present a method for the goal-oriented design of outlier- and inlier-robust statistical inference tools. In particular, this includes the tasks of parameter estimation, testing for goodness-of-fit resp. homogenity resp. independence, clustering, change-point detection, exploratory model search, and some Bayesian decision procedures.

In order to achieve this goal, we adapt the concept of *scaled Bregman distances between two distributions*, which was introduced in Stummer [2007], Stummer & Vajda [2012], and which generalizes the widely-used (partially non-robust) concepts of Kullback-Leibler information distance/relative entropy, Pearson's chisquare distance, Hellinger distance, Csiszar-Ali-Slivey divergences, etc. The classical (i.e., unscaled) Bregman distances – such as the L^2-distance and the more general density power divergences – are covered as well.

In order to visualize effectively and transparently the corresponding robustness properties, we present 3D-plots of associated *density-pair adjustment functions*. Numerous special cases will be illustrated. For the discrete case, some universally applicable results on the asymptotics of the underlying scaled-Bregman-distance test statistics are derived as well. Furthermore, we give some application to the robust estimation of the tail dependence coefficient of bivariate heavy-tailed distributions.

This talk is mainly based on several joint works with A.-L. Kißlinger (Erlangen-Nürnberg) respectively with B.H. Roensch (Erlangen-Nürnberg).

References

Kißlinger, A.L. & Stummer, W. (2013). Some decision procedures based on scaled Bregman distance surfaces. In: Nielsen F. & Barbaresco F. (eds) GSI 2013, Lecture Notes in Computer Science LNCS 8085, Springer, pp 479 – 486.

Kißlinger, A.L. & Stummer, W. (2015a). New model search for nonlinear recursive models, regressions and autoregressions. In: Nielsen F. & Barbaresco F. (eds) GSI 2015, Lecture Notes in Computer Science LNCS 9389, Springer, pp 693 – 701.

Kißlinger, A.L. & Stummer, W. (2015b). Robust statistical engineering by means of scaled Bregman distances. To appear in Springer volume on ICORS 2015.

Kißlinger, A.L. & Stummer, W. (2015c). A new information-geometric method of change detection. Preprint.

Stummer, W. (2007). Some Bregman distances between financial diffusion processes. *Proc Appl Math Mech (PAMM)*, **7**, 1050,503 – 1050,504.

Stummer, W. & Vajda, I. (2012). On Bregman distances and divergences of probability measures. *IEEE Trans Inf Theory* **58(3)**, 1277 – 1288.

Total Variation Depth for Functional Data: Properties and Applications

H. Huang and ***Y. Sun***[1*]

[1] *Division of Computer, Electrical and Mathematical Sciences and Engineering, King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia; ying.sun@kaust.edu.sa.*
[*] *Presenting author*

Keywords. *Data depth; Functional data; Total variation; Outlier detection; Shape outliers*

There has been extensive work on data depth-based methods for robust multivariate data analysis. Recent developments have moved to infinite-dimensional objects such as functional data. In this work, a new notion of depth for functional data, the total variation depth, is introduced. The proposed notion is well suited for shape outlier detection due to the fact that it considers the total variation and takes into account the necessary correlations in functional data. Effective outlier detection rules along with visualization tools are also developed, and the outlier detection performance is examined through simulation studies. The numerical results show that it outperforms many other notions under different types of outlier models. Finally, we illustrate our method using real data examples for detecting outliers in sample curves and images.

Robust orthogonal regression for compositional data in R

V. Todorov[1,*], *K. Hrůzová*[2], *K. Hron*[2] *and P. Filzmoser*[3]

[1] *United Nations Industrial Development Organization, Vienna, Austria; v.todorov@unido.org*
[2] *Palacky University, Olomouc, Czech Republic; klara.hruzova@gmail.com, hronk@seznam.cz*
[3] *Vienna University of Technology, Vienna, Austria; p.filzmoser@tuwien.ac.at*
[*] *Presenting author*

Keywords. *Compositional data; Orthogonal regression; Isometric logratio coordinates; MM-estimates; Bootstrap inference*

In the context of building a regression model on compositional data, orthogonal regression (as a special case of errors-in-variable models) is appropriate since all compositional parts - also the explanatory variables - are measured with errors. The classical approach to estimate the model is based on an eigenvector analysis of the joint covariance matrix of the observations. However, in the presence of outlying observations in compositional data the orthogonal regression (that is able to handle the regression problem statistically) should be replaced by its robust counterpart. Therefore, we consider also a robust version of orthogonal regression. In Zamar [1989], M- and S-estimators for robust orthogonal regression are presented. However, S-estimators are computed using inefficient algorithms and M-estimators have low breakdown point. Another possibility can be found in Croux et al. [2010], where the projection-pursuit approach is used, which is also suitable for more than one response variable. In order to benefit from the better statistical properties of the MM-estimates, we decided to follow the above mentioned (classical) approach and develop robust orthogonal regression in orthonormal coordinates using robust PCA, which is obtained through a robust estimation of the covariance matrix. Among other possibilities, the MM-estimators are employed for this purpose. The reason for choosing MM-estimators is that they are highly efficient when the errors have a normal distribution, their breakdown point is 50% and they have bounded influence function.

In order to perform statistical inference, like deriving confidence intervals or testing hypotheses, bootstrap techniques for classical and robust orthogonal regression are proposed. Although bootstrap is a very useful tool, in case of robust estimators there are two problems: computational complexity of robust estimators and the instability of the bootstrap in case of outliers. Therefore we used fast and robust bootstrap Salibian et al. [2006] which is based on the fact that the robust estimators (namely S- and MM-estimators) can be represented by smooth fixed point equations which allow to calculate a fast approximation of the estimates in each bootstrap sample.

As the estimation of parameters and statistical inferences is performed in real (unconstrained) coordinates, the resulting orthogonal regression model is not exclusively designed for compositional data, but it could be used also with any non-compositional data.

The R package *oreg* provides functions for classical and robust orthogonal regression. These functions can be applied on both compositional and non-compositional data. In case of compositional data, all regression models are estimated, one for each orthonormal basis. The regression parameters are estimated using (classical or robust) principal components and the

MM-estimates are computed by a call to the implementation in the *rrcov* package Todorov & Filzmoser [2009]. The results can be viewed by standard `print()` and `plot()` functions, while a `summary()` function presents the parameter estimates and also the corresponding statistical inference (confidence intervals and p-values for significance testing) obtained through bootstrap. In the robust version, fast and robust bootstrap from the package *FRB* Van Aelst & Willems [2013] is used.

The robustness, the efficiency and the computational performance of the procedure are studied through simulation and are illustrated with a data set from macroeconomics representing the structure of gross value added and the relation between its components. The data set comes from the World Bank database (http://data.worldbank.org) and includes observations for 131 countries in 2010 at constant 2005 USD.

References

Croux C, Fekri M & Ruiz-Gazen A. Fast and robust estimation of the multivariate errors in variables model. *Test* **19** 286–303.

Salibian-Barrera M, Van Aelst S & Willems G. (2006). PCA based on multivariate MM-estimators with fast and robust bootstrap. *J Am Stat Assoc* **101** 1198–1211.

Todorov V & Filzmoser Filzmoser P. An object oriented framework for robust multivariate analysis. *Journal of Statistical Software* **2009** 32/3.

Van Aelst S & Willems G. Fast and robust bootstrap for multivariate inference: The R package FRB. *Journal of Statistical Software* **2013** 53/3.

Zamar, RH. Robust estimation in the errors-in-variables model. *Biometrika* **76** (1) 149–160.

Robust inference with minimum dual divergence estimators for moment condition models

A. Toma[1*] *and A. Keziou*[2]

[1] *Department of Applied Mathematics, Bucharest Academy of Economic Studies, and Gh. Mihoc - C. Iacob Institute of Mathematical Statistics and Applied Mathematics, Bucharest, Romania; aida_toma@yahoo.com.*
[2] *Laboratoire de Mathematiques de Reims, Universite de Reims, Champagne Ardenne, UFR Sciences, Moulin de la Housse, B.P. 1039, 51687 Reims, France; amor.keziou@univ.reims.fr.*
[*] *Presenting author*

Keywords. *Moment condition models; Robust estimation; Divergence measures.*

The minimum dual divergence estimators and tests for moment condition models have been proposed recently in literature. The main advantage, of using a divergence based approach and duality, lays in the fact that it leads to asymptotic properties of the estimators and test statistics under the model, as well as under the alternative including misspecification, which cannot be achieved through the classical empirical likelihood context. Also, the estimators are asymptotically the best in the sense on Hansen yielding a "smallest" asymptotic covariance matrix. On the other hand, the estimators of the unknown parameter corresponding to the model have bounded influence functions if and only if the function inducing the orthogonality conditions of the model is bounded. Since in many applications this function is not bounded, it is useful to have procedures that modify the orthogonality conditions in order to obtain robust versions of the estimators. In this paper we propose robust versions of the minimum dual divergence estimators using truncated orthogonality functions. We prove robustness properties and asymptotic properties for the new estimation method, underlying some advantages of using it with respect to other known methods. The performance of the new method is illustrated through Monte Carlo simulations for some moment condition model.

Acknowledgements. This work was supported by a grant of the Romanian National Authority for Scientific Research, CNCS – UEFISCDI, project number PN-II-RU-TE-2012-3-0007.

A robust likelihood approach to inference about the difference between two multinomial distributions in paired designs

T.S. Tsou[1*]

[1] *Institute of Statistics, Institute of Systems Biology and Bioinformatics, Center for Biotechnology and Biomedical Engineering, National Central University, Jhongli, Taiwan; chopinmozart0422@gmail.com.*
[*] *Presenting author*

Keywords. *Multinomial distribution; Paired design; Robust likelihood.*

Abstract: We propose a new universal robust test to test the equality of two multinomial distributions in paired designs. This test accounts for the within-cluster correlation in a data-driven manner and is easy to compute without a full model specification. We provide theoretical justifications and use simulations and real data analysis to demonstrate the merit of the robust procedure.

Reference

Royall, R. M. & Tsou, T. S. (2003). Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *Journal of the Royal Statistical Society, Series B*, **65**, 391–404.

Joint penalization of multiple scatter matrices

D.E. Tyler[1*], *E. Ollila* [2], *I. Soloveychik*[3], *and A. Wiesel*[3]

[1] *Department of Statistics & Biostatistics, Rutgers, The State University of New Jersey, USA*
[2] *Department of Signal Processing and Acoustics, Aalto University, Finland*
[3] *School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel*
[*] *Presenting author*

Keywords. *Geodesic convexity; Regularized M-estimators of covariance matrices.*

Consider sampling p-dimensional observations from k distinct groups, with sample sizes $n_j, j = 1, \ldots k$ respectively. A common assumption in the k-group problem is the equality of the covariance matrix over the different groups. Such an assumption is helpful in the under-sampled scenario, that is when the sample sizes of the different groups are relatively small, and in particular when $n_j < p$ or even $n_j = 1$ for some groups. Rather than assume equal covariance matrices, we consider in this paper estimating the covariance or scatter matrices under the assumption that they may be simply close to each other in some metric, and hence deviate from some common positive definite "center".

We consider two penalized M-estimation approaches. The first approach begins with a pooled M-estimator of scatter based on all the data, followed by a penalized M-estimator of scatter for each group, with the penalty term chosen so that they groups scatter matrices are shrunk towards the pooled scatter matrix. In the second approach, we minimize the sum of the M-estimation cost functions over the groups along with an additive joint penalty enforcing some similarity, i.e. with shrinkage towards a mutual center.

In both approaches, we utilize the concept of geodesic convexity to prove the existence and uniqueness of the penalized solution under general conditions. We then consider three specific penalty functions based on the Euclidean, the Riemannian, and the information theoretic (Kullback-Leibler) distances In the second approach, the distance based penalties are shown to lead estimators of the mutual center that are related to the arithmetic, the Riemannian and the harmonic means of positive definite matrices, respectively. We also consider a penalty based on an ellipticity measure for positive definite matrices, which shrinks the individual estimators toward a common shape matrix rather than a common scatter matrix.

Robust principal components by least trimmed squares for high-dimensional and functional data

S. Van Aelst[1*], *M. Salibian-Barrera*[2] *and H. Cevallos-Valdiviezo*[3]

[1] *KU Leuven, Department of Mathematics; Stefan.VanAelst@wis.kuleuven.be.*
[2] *University of British Columbia, Department of Statistics; matias@stat.ubc.ca*
[3] *Ghent University, Department of Applied Mathematics, Computer Science and Statistics; Holger.CevallosValdiviezo@UGent.be*
[*] *Presenting author*

Keywords. *Principal components; Least trimmed squares; High-dimensional; Functional data.*

Classical (functional) principal component analysis can yield erroneous approximations in presence of outliers. To reduce the influence of atypical measurements in the data, we propose two methods based on trimming: a multivariate least trimmed squares (LTS) estimator and a componentwise variant. The multivariate LTS minimizes the least squares criterion over subsets of observations. This approach corresponds to the PCA method in Maronna [2005], but a different representation for the subspace is used which leads to an algorithm that is better suited for high-dimensional data. The componentwise version minimizes the sum of univariate LTS scale estimators in each of the components. This method corresponds to the approach of Boente & Salibian-Barrera [2015] by using LTS scales instead of S-scales.

The methods can directly be applied to high-dimensional multivariate data. For functional data that consist of irregularly spaced curves, first smoothing as in Boente & Salibian-Barrera [2015] is used to represent the curves in a high-dimensional space. The LTS solution is then computed on these multivariate data and the obtained solution is mapped back onto the functional space.

Observations that lie far from the estimated subspace are considered to be outliers. Outliers can thus be flagged according to their orthogonal distance from the subspace. A simulation study and real data applications show that our estimators yield competitive results, both in identifying outliers and approximating regular data when compared to other existing methods.

References

Maronna, R. (2005). Principal components and orthogonal regression based on robust scales. *Technometrics*, **47**, 264–273.

Boente, G. & Salibian-Barrera, M. (2015). S-Estimators for functional principal component analysis. *Journal of the American Statistical Association*, **110**, 1100–1111.

The Minimum Regularized Covariance Determinant estimator

Kris Boudt[1*], *Peter Rousseeuw*[2], *Steven Vanduffel*[1] *and* ***Tim Verdonck***[2]

[1] *Solvay Business School, Vrije Universiteit Brussel, Belgium; kris.boudt@vub.ac.be, steven.vanduffel@vub.ac.be.*
[2] *Department of Mathematics, KU Leuven, Belgium; peter@rousseeuw.net, tim.verdonck@wis.kuleuven.be*
[*] *Presenting author*

Keywords. *Robust covariance estimation; High-dimensional data; MCD estimator.*

The Minimum Covariance Determinant (MCD) estimator proposed by Rousseeuw [1985] is often used for high breakdown point covariance estimation in small to medium dimensions. Like most covariance estimators the MCD requires that the sample size n is larger than the dimension p of the data. In this paper we generalize the MCD approach by estimating the covariance matrix using a weighted average of a target matrix and the sample covariance on the subset, chosen in order to minimize the determinant of this regularized covariance estimate. The resulting Minimum Regularized Covariance Determinant (MRCD) estimator preserves the high breakdown point properties of the MCD estimator, and has the additional advantage that it is well-conditioned, even when $p > n$. A simulation study confirms the good properties of the estimator. Finally, we illustrate the advantages of the MRCD estimator for outlier detection in high dimensions on real data.

References

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8, 283-297.

Quantile regression in varying coefficient models: non-crossingness and heteroscedasticity

Y. Andriyana[1], *I. Gijbels*[2] *and* ***A. Verhasselt***[3*]

[1] *Statistics Department, Universitas Padjadjaran, Bandung, Indonesia; yudhie.andriyana@unpad.ac.id.*
[2] *Department of Mathematics and Leuven Statistics Research Center (LStat), KU Leuven, Leuven, Belgium; irene.gijbels@wis.kuleuven.be*
[3] *Censtat, Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Belgium; anneleen.verhasselt@uhasselt.be.*
[*] *Presenting author*

Keywords. *Crossing quantile curves; Heteroscedasticity; P-splines; Quantile regression; Varying coefficient models.*

Quantile regression is an important tool for describing the characteristics of conditional distributions. In real applications, the impact of explanatory variables on a variable of interest, leads to the study of conditional quantile functions or regression quantiles. In practice, the conditional quantile functions are estimated, from data, for various fixed values of the order of the quantile τ. Conditional quantile functions are by definition, for any given fixed values of the covariates, an increasing function in the argument τ. Unfortunately estimated regression quantile curves often violate this non-crossingness property, which can be very annoying for interpretation and further analysis. There is thus an interest to prevent this crossing to happen in finite-samples.

To describe accurately complex data, one often considers regression models that are on the one hand flexible enough to capture this complexity, but on the other hand still allow for estimation methods with good practical performance. We focus on varying coefficient models, which naturally extend linear regression models by allowing the regression coefficients to change with another covariate.

We consider flexible varying coefficient models, and develop methods (based on P-splines) for quantile regression that ensure that the estimated quantile curves do not cross. A second aim is to allow for some heteroscedasticity in the error modelling, and to also estimate the associated scaling/variability function. We investigate the finite-sample performance of the discussed methods via simulation studies. Some applications to real data illustrate the use of the methods in practical settings.

Trade-off between Efficiency and Robustness in Post-Model Selection Inference

A.N. Vidyashankar[1*]

[1] *Department of Statistics, Volgeneau School of Engineering, George Mason University, Fairfax, VA 22033, USA; avidyash@gmu.edu*

[*] *Presenting author*

Keywords. *Large deviations; Moderate Deviations; High-dimensional data; Bahadur Efficiency; Pitman Efficiency; Hellinger Distance*

It is common practice in high-dimensional data analysis that a model selection is first performed and then inference is carried out using the selected model presuming that the chosen model is the true model; that is, without accounting for model selection uncertainty. Recently, methods such as *clean and screen* are being used to account for model selection uncertainty. However, the robustness and the efficiency properties of the resulting statistical procedures are largely unknown. In this presentation, we provide a systematic account of efficiency and robustness properties of post-selection estimators. In the process we address some foundational questions concerning the role of moderate deviation theory in the study of statistical efficiency and robustness and their trade-offs.

S-weighted estimators

J.Á. Víšek[1*]

[1] *Institute of Economic Studies, Faculty of Social Sciences, Charles University in Prague*
[*] *Presenting author*

Keywords. *Robustness; Weighting the order statistics of the squared residuals plugged into a ρ-function; Consistency of S-weighted estimator.*

After an unsuccessful pursuit for a robust estimator with 50% breakdown point the task became at the end of seventies a nightmare of statisticians. Feasible versions of such estimators - the *least median of squares* and the *least trimmed squares* - fulfilled the desire but the discontinuity of objective function and the presence of order statistics of the squared residuals in their definitions were not favorable for studying the properties of estimators in question. All after, it caused that the consistency of the *least trimmed squares* was in full generality proved more than 20 years after the proposal of $\hat\beta^{(LTS,n,h)}$, Víšek (2006). *S-estimators*, defined as

$$\hat\beta^{(S,n,\rho)} = \arg\min_{\beta\in R^p}\left\{\sigma\in R^+ \,:\, \sum_{i=1}^n \rho\left(\frac{r_i(\beta)}{\sigma}\right) = b\right\} \tag{18}$$

see Rousseeuw *&* Yohai (1984), removed both these snags simultaneously preserving the high breakdown point. The extraordinary virtue of all these estimators was their "innate" *scale-* and *regression-equivariance* in contrast to *M-estimators* which require special studentization of residuals, see Bickel (1975). However, the requirement on high breakdown point (seemingly) yields the high sensitivity to a small shift of "in-liers", see Hettmansperger *&* Sheather(1992). Although, their results were - due to a bad algorithm - a bit biased, they opened a discussion on the sensitivity of (robust) estimators to "inliers". The *least weighted squares* (LWS)

$$\hat\beta^{(LWS,n,w)} = \arg\min_{\beta\in R^p}\sum_{i=1}^n w\left(\frac{i-1}{n}\right) r^2_{(i)}(\beta)$$

where

$$r^2_{(1)}(\beta) \le r^2_{(2)}(\beta) \le \ \ldots\ \le r^2_{(n)}(\beta),$$

see Víšek (2000), offered a chance to cope with this drawback. and represented an alternative to S-estimators (it is easy to see that $\hat\beta^{(LWS,n,w)}$ is not a special case of $\hat\beta^{(S,\rho,n)}$ and vice versa). The high speed of modern computational means allowed to select the weight function w just tailored to the level and even to the character of contamination. Utilization of a generalized version of Kolmogorov-Smirnov result on the convergence of empirical distribution functions to the underlying one - generalized for the regression framework, see Víšek (2011) - then simplified the proofs of consistency, $\sqrt{n}$-consistency, etc . Finally, recently proposed *S-weighted estimator*

$$\hat\beta^{(SW,n,w,\rho)} = \arg\min_{\beta\in R^p}\left\{\sigma(\beta)\in R^+ \,:\, \frac{1}{n}\sum_{i=1}^n w\left(\frac{i-1}{n}\right)\rho\left(\frac{r^2_{(i)}(\beta)}{\sigma^2}\right) = b\right\}, \tag{19}$$

see Víšek (2015), where $b = \mathbb{E}\rho(\frac{\varepsilon_1^2}{\sigma_0^2})$, $\rho : (0,\infty) \to (0,\infty)$ and nondecraesing on $(0,\infty)$, inherited plausible properties of *S-estimators* as well as of the *least weighted squares*, allowing

for a wide range of objective (unbounded) functions and simultaneously offering to adjust the estimator to the level and to the character of contamination. Notice that in (19) in contrast to (18) we need the order statistics of squared residuals. Fortunately, employing a trick from *rank statistics*, see Hájek & Šidák (1967), we can rid of the technical problems with order statistics and then to employ Kolmogorov-Smirnov result recalled above. The paper summarizes the ideas of proving the consistency and $\sqrt{n}$-consistency under heteroscedasticity of error terms. It offers also a few patterns of results of numerical studies of their behavior for moderate sample size. These patterns demonstrate that a widely spread idea that the leverage points cause much more serious problems than the outliers need not have the absolute validity.

References

Bickel, P. J. (1975): One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc. 70, 428–433.*

Hájek, J. & Šidák, Z. (1967): *Theory of Rank Test.* Academic Press, New York.

Hettmansperger, T. P. & Sheather, S. J. (1992): A Cautionary Note on the Method of Least Median Squares. *The American Statistician 46, 79–83.*

Rousseeuw, P. J. & Yohai, V. (1984): Robust regressiom by means of *S*-estimators. In: *Robust and Nonlinear Time Series Analysis. eds. J. Franke, W. Hărdle and R. D. Martin, Lecture Notes in Statistics 26, Springer Verlag, New York, 256-272.*

Víšek, J. Á. (2000): Regression with high breakdown point. *Robust 2000 (eds. Jaromír Antoch & Gejza Dohnal, published by Union of Czech Mathematicians and Physicists), 2001, 324 - 356.*

Víšek, J. Á. (2006): The least trimmed squares. Consistency. $\sqrt{n}$-consistency. Asymptotic normality and Bahadur representation. *Kybernetika 42, 1 - 36, 181 - 202, 203 - 224.*

Víšek, J. Á. (2011): Empirical distribution function under heteroscedasticity. *Statistics 45, 497-508.*

Víšek, J. Á. (2015): *S*-weighted estimators. *Proc. of the 16th Conference on the Applied Stochastic Models, Data Analysis and Demographics 2015, 1031 - 1042.*

From Sparse to Dense Functional Data and Beyond

X. Zhang[1] *and* ***J.-L. Wang***[2*]

[1] *Department of Applied Economics and Statistics, University of Delaware; xkzhang@udel.edu*
[2] *Department of Statistics, University of California, Davis; janelwang@ucdavis.edu*
[*] *Presenting author*

Keywords. *Local liner smoothing; asymptotic normality; L 2 convergence, uniform convergence, longitudinal data*

Nonparametric estimation of mean and covariance functions is important in functional data analysis. We investigate the performance of local linear smoothers for both mean and covariance functions with a general weighing scheme, which includes two commonly used schemes, equal weight per observation (OBS), and equal weight per subject (SUBJ), as two special cases. We provide a comprehensive analysis of their asymptotic properties on a unified platform for all types of sampling plan, be it dense, sparse, or neither. Three types of asymptotic properties are investigated in this paper: asymptotic normality, L^2 convergence and uniform convergence. The asymptotic theories are unified on two aspects: (1) the weighing scheme is very general; (2) the magnitude of the number N_i of measurements for the ith subject relative to the sample size n can vary freely. Based on the relative order of N_i to n, functional data are partitioned into three types: non-dense, dense, and ultra-dense functional data for the OBS and SUBJ schemes. These two weighing schemes are compared both theoretically and numerically. We also propose a new class of weighing schemes in terms of a mixture of the OBS and SUBJ weights, of which theoretical and numerical performances are examined and compared.

A robust method for ultra-high dimensional regression analysis

Y. Wang[1*] *and S. Van Aelst*[1]

[1] *Department of Mathematics, KU Leuven;*
yixin.wang@wis.kuleuven.be, stefan.vanaelst@wis.kuleuven.be
[*] *Presenting author*

Keywords. *Sure Screening; Factor Profiling; Robust Factor Analysis; Least Trimmed Squares; Robust Regression.*

To increase the estimation accuracy and reduce the computational cost in ultra-high dimensional regression analysis, Fan & Lv [2008] proposed Sure Independence Screening (SIS) which selects a subset of the variables before estimating the regression coefficients. Predictor variables are selected according to the magnitude of their marginal correlations with the response variable. Fan & Lv [2008] proved that SIS shares the Sure Screening Property that the selected model will cover all the important variables with an overwhelming probability under certain model assumptions. However, the performance of SIS deteriorates greatly with increasing dependence among the predictors. To solve this problem, Wang [2012] proposed Factor Profiled Sure Independence Screening (FPSIS) based on the assumption that the predictors' correlation structure can be represented by a low-dimensional factor model. After profiling out the correlation structure by projecting the data onto the orthogonal complement of the subspace spanned by the factors, the screening performance can be largely improved. Since FPSIS is sensitive to outliers, a robust procedure is proposed.

In this work, we first estimate the subspace robustly based on the least trimmed squares estimator. Other robust techniques, such as ROBPCA, can also be applied. To increase the statistical efficiency, we apply a reweighting step in which we select the observations that are only outlying according to their score distances. The profiled variables are obtained by projecting all variables onto the orthogonal complement of the subspace spanned by the factors estimated from the reduced data. Finally, a robust regression method is used on the profiled data to estimate the marginal contribution of each predictor to the response variable. The results given by both the classical and the robust procedures with different types of outliers are compared.

References

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, **70**, 849-911.

Wang, H. (2012). Factor Profiled Sure Independence Screening. *Biometrika*, **99**, 15-28.

Robustness of quadratic inference function estimators

Samuel Müller[1], *Suojin Wang*[2] *and* ***A.H. Welsh***[3*]

[1] *University of Sydney; samuel.mueller@sydney.edu.au*
[2] *Texas A&M University; sjwang@stat.tamu.edu*
[3] *The Australian National University; Alan.Welsh@anu.edu.au*
[*] *Presenting author*

Keywords. *Generalized estimating equations; Longitudinal data.*

Quadratic inference function estimators for the regression parameter in regression models for longitudinal data were introduced by Qu et al. [2000] to improve on the efficiency of generalized estimating equations estimators. Qu & Song [2004] argued that quadratic inference function estimators are also robust against outliers, making them preferable to generalized estimating equations estimators. In this talk, we discuss the robustness properties of quadratic inference function estimators, revisiting particular cases to understand more deeply the generality of the conclusions of Qu & Song [2004]. We show that robustness issues in generalised estimating equations estimation are more subtle than is generally believed and we had anticipated.

References

Qu, A., Lindsay, B.G. & Li, B. (2000). Improving generalized estimating equations using quadratic inference functions *Biometrika*, **87**, 823–836.

Qu, A. & Song, P.X.-K. (2004). Assessing robustness of generalized estimating equations and quadratic inference functions. *Biometrika*, **91**, 447–459.

Comparison of estimation methods for cellwise robust regression

I. Wilms[1*] *and C. Croux*[1]

[1] *Faculty of Economics and Business; ines.wilms@kuleuven.be, christophe.croux@kuleuven.be.*
[*] *Presenting author*

Keywords. *Cellwise outliers; componentwise contamination; multiple regression.*

In multiple regression analysis, a response variable is predicted based on a set of p predictor variables. We collect all available information in an n by $(p+1)$ matrix where the n observations are contained in the rows and the response and predictor variables are contained in the columns of the data set. Robust statistics has mostly focused on developing 'rowwise robust' estimation methods, methods that remain reliable in the presence of outlying rows in the data set. Such rowwise robust estimation methods flag either a whole row in the data set as outlying or not. Only recently, 'cellwise robust' estimation methods have been developed that flag a cell of the data set as outlying or not. Such cellwise robust estimation methods are likely to be more suited for situations where a large number of observations suffer from contamination in only a small number of variables. We perform an extensive simulation study where we compare the performance of several cellwise robust regression methods. We consider both low-dimensional ('thin') data sets, data sets with a large number of observations (rows) relative to the number of variables (columns) and high-dimensional ('fat') data sets, data sets with a large number of variables (columns) relative to the number of observations (rows).

References

Leung, A., Zhang, H. & Zamar, R. (2016). Robust regression estimation and inference in the presence of cellwise and casewise contamination, *Computational Statistics & Data Analysis*, 99, 1-11.

Öllerer, V., Alfons, A. & Croux, C. (2015). The shooting S-estimator for robust regression. *Computational Statistics*, DOI 10.1007/s00180-015-0593-7.

The journey of predictive analytics in Nestlé

D. Wu[1*]

[1] *Global Lead of Demand Planning, Nestlé, Vevey, Switzerland; Gang.Wu@nestle.com*
[*] *Presenting author*

Keywords. *Predictive analytics; Demand planning; Forecast: Big data; Promotion*

Nestlé, like many CPG companies, face the same challenges in increasing demand volatility and competitions, and thus facing a growing portion of sales made on promotions, which is one of the most difficult to predict events.

We are also in an environment where more data become more accessible, notably consumption data from syndicated scanners or retailers (namely sell-out data). The use of such data becomes increasingly critical to gauge the "True Demand" at the point of sales to consumers.

This in turn helps manage our supply to retailers (namely sell-in) in a far more accurate and controlled way to ensure that we provide the right services to them, in both deliveries and maintaining retailers' inventory holding. It is a win-win situation.

Over the past two years, Nestlé has established a much improved "Demand Planning" process and advanced analytical solution to fundamentally change the way we capture and utilize data, analyze the historical demand patterns and predict accurately the future demand at both sell-out and sell-in levels. Nestlé can also use the data and solution to track, measure and eventually optimize promotional sales. This is only possible with the use of advanced analytical solution such as what "SAS Forecasting Solutions" can offer.

Nestlé will share its experiences from its data segmentation and process, forecast modelling from sell-out to sell-in, to its attempts to optimize promotion effectiveness.

On High-Dimensional Cross-Validation

C.K. Ing[1], *W.C. Hsiao*[1] *and* ***W. Y. Wu***[2*]

[1] *Institute of Statistical Science, Academia Sinica, Taiwan.*
[2] *Department of Applied Mathematics, National Dong Hwa University, Taiwan.*
[*] *Presenting author*

Keywords. *High-dimensional Linear Model; Model Selection, Cross-validation, Consistency*

Cross-validation (CV) is one of the most popular methods for model selection. By splitting n data points with $n_v/n \to 1$ and $n_c \to \infty$ into a training sample of size n_c and a validation sample of size n_v, Shao (1993) showed that subset selection based on CV is consistent in a regression model of p candidate variables with $p << n$. However, in the case of $p >> n$, not only does CV's consistency remain undeveloped, but subset selection is also practically infeasible. In this paper, we fill this gap by using CV as a backward elimination tool for eliminating variables that are included by high-dimensional variable screening methods possessing sure screening property. By choosing an n_v such that n_v/n converges to 1 at a rate faster than the one in Shao's (1993) paper, we establish the consistency of our selection procedure. We also illustrate the finite-sample performance of the proposed procedure using Monte Carlo simulation.

Sufficient Forecasting Using Factor Models

J. Fan[1], ***L. Xue***[2*] *and J. Yao*[3]

[1] *Princeton University; jqfan@princeton.edu*
[2] *The Pennsylvania State University; lzxue@psu.edu*
[3] *Citadel LLC; jiaweiy@princeton.edu*
[*] *Presenting author*

Keywords. *Regression; Forecasting; Semiparametric Factor Model; Dimension Reduction; High Dimensional Asymptotic Theory.*

We consider forecasting a single time series when there is a large number of predictors and a possible nonlinear effect. The dimensionality was first reduced via a high-dimensional approximate factor model implemented by the principal component analysis. Using the extracted factors, we develop a novel forecasting method called the sufficient forecasting, which provides a set of sufficient predictive indices, inferred from high-dimensional predictors, to deliver additional predictive power. The projected principal component analysis will be employed to enhance the accuracy of inferred factors when a semi-parametric approximate factor model is assumed. Our method is also applicable to cross-sectional sufficient regression using extracted factors. The connection between the sufficient forecasting and the deep learning architecture is explicitly stated. The sufficient forecasting correctly estimates projection indices of the underlying factors even in the presence of a nonparametric forecasting function. The proposed method extends the sufficient dimension reduction to high-dimensional regimes by condensing the cross-sectional information through factor models. We derive asymptotic properties for the estimate of the central subspace spanned by these projection directions as well as the estimates of the sufficient predictive indices. We further show that the natural method of running multiple regression of target on estimated factors yields a linear estimate that actually falls into this central subspace. Our method and theory allow the number of predictors to be larger than the number of observations. We finally demonstrate that the sufficient forecasting improves upon the linear forecasting in both simulation studies and an empirical study of forecasting macroeconomic variables.

Robust Singular Spectrum Analysis

M. Yarmohammadi[1*], *M. Kalantari*[1]

Department of Statistics, Payame Noor University, PO BOX 19395-3697, Tehran, Iran.

Keywords. *Time Series; Singular Spectrum Analysis; Outlier; L_1norm.*

The Singular Spectrum Analysis (SSA) method is a powerful tool for analysis and forecasting time series data. This technique combines elements of classical time series analysis, multivariate statistics, multivariate geometry, dynamical systems and signal processing. The last decade has seen an exponential increase in the application of SSA to forecasting in various fields ranging from meteorology, biomedical science and finance, to economics.

The aim of SSA is to make a decomposition of the original series into the sum of a small number of interpretable components such as a slowly varying trend, oscillatory components and a structureless noise. Neither a parametric model nor stationarity conditions have to be assumed for the time series. This makes SSA a model-free or nonparametric method and hence enables SSA to have a very wide range of applicability.

A thorough description of the theoretical and practical foundations of the SSA technique (with several examples) can be found [Golyandina et al. , 2001] and [Golyandina & Zhigljavsky , 2013]. An elementary introduction to the subject can be found in [Elsner & Tsonis , 1996].

The most common version of SSA is called *Basic SSA*. This version of SSA is based on the *Frobenius* norm (or L_2 norm), which is very sensitive to the presence of outliers. Therefore the outliers have a significant impact on SSA reconstruction and forecasts Hassani et al. [2014].

The main object of this paper is to introduce a rubustification version of SSA which is based on the L_1 norm. The theoretical and empirical results confirm that L_1-SSA outperforms the basic SSA in reconstruction and forecasts when faced with time series which are contaminated by outliers. The performance of this approach will be investigated by simulation studies.

References

Golyandina, N., Nekrutkin, V. & Zhigljavsky, A. (2001). Analysis of Time Series Structure: SSA and Related Techniques, Chapman&Hall/CRC, Boca Raton.

Golyandina, N. & Zhigljavsky, A. (2013). Singular Spectrum Analysis for Time Series, Springer Briefs in Statistics. Springer.

Elsner, J. B. & Tsonis, A. A. (1996). Singular Spectrum Analysis: A New Tool in Time Series Analysis, Plenum Press.

Hassani, H., Mahmoudvand, R., Omer, H. N. & Silva, E. S. (2014). A Preliminary Investigation into the Effect of Outlier(s) on Singular Spectrum Analysis. *Fluctuation and Noise Letter* **13** (4) 1450029.

Non-unbiased two-sample nonparametric tests. Numerical example

X. Yermolenko[1*]

[1] *Charles University in Prague, Department of Statistics, Sokolovská 83, CZ-186 75 Prague 8, Czech Republic; Yermolenko@karlin.mff.cuni.cz*
[*] *Presenting author*

Keywords. *Unbiasedness; Two - sample problem; Wilcoxon test.*

Many tests on vector or scalar parameters against two-sided alternatives are generally not finite-sample unbiased. They are unbiased only for symmetric distributions or under similar conditions. This was already noticed by Amrhein [1995], Sugiura et al. [2006] and generally analyzed by Jurečková & Kalina [2012], Jurečková & Milhaud [2003] and later by many others. While in univariate models the tests are unbiased against one-sample alternatives, such alternatives are not clearly characterized in the multivariate models.
We shall numerically illustrate this important problem on the Wilcoxon test against two-sample alternative of shift in location, applied to a skew logistic distribution and unequal sample sizes. Namely, let $X_1, \ldots, X_n; Y_1, \ldots, Y_m$ be two random samples according to absolutely continuous distributions $F(x)$ and $G(x)$ respectively.
In order to test the hypotheses $G(x) = F(x)$, against the alternative that F is the skew logistic distribution and $G(x) = F(x-\Delta), \Delta \neq 0$, we will consider the following two - sample Wilcoxon test:

$$\phi(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = \begin{cases} 1, & \text{if } X_{(n)} < Y_{(1)} \text{ or } X_{(1)} > Y_{(m)}, \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

References

Amrhein, P. (1995). An example of a two-sided Wilcoxon signed rank test which is not unbiased. *Ann. Inst. Statist. Math.* **47** 167-–170.

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

Jurečková, J. & Milhaud, X. (2003). Derivative in the mean of a density and statistical applications. *IMS Lecture Notes* **42**, 217-–232.

Sugiura, N., Murakami, H., Lee, S.K. & Maeda, Y. (2006). Biased and unbiased two-sided Wilcoxon tests for equal sample sizes. *Ann. Inst. Statist. Math.* **58**, 93-–100.

Robust estimators for negative binomial regression.

M. Amiguet[1], *A. Marazzi*[1], *M.Valdora*[2] *and* ***V.J. Yohai*** [2*]

[1] *Université de Lausanne; Michael.Amiguet@chuv.ch, alfio.marazzi@chuv.ch*
[2] *Universidad de Buenos Aires; mvaldora@gmail.com, vyohai@dm.uba.ar*
[*] *Presenting author*

Keywords. *Kendall rank correlation coefficient; Full efficiency; Conditional maximum likelihood estimator*

In recent years, negative binomial (NB) regression has received increasing attention as a tool for modeling count data in presence of overdispersion. A convenient way to parametrize the negative binomial probability function is by the mean μ and the dispersion parameter α. We denote this distribution by NB(μ, α)

The negative binomial regression model assumes that we observe a response y and a vector of covariables $\mathbf{x} \in \mathbf{R}^p$, so that $y|\mathbf{x}$ has distribution NB($h(\beta_0^{\mathrm{T}}\mathbf{x} + \delta), \alpha_0)$), where the link function h is known while $\beta_0 = (\beta_{01}, ..., \beta_{0p})$ and α_0 are unknown parameters.

One way to estimate these parameters is by means of the maximum likelihood estimator. However these estimators are very sensitive to the presence of outliers in the sample. A robust estimator for this model was proposed by Aeberhard, Cantoni, and Heritier [2014].

We are going to introduce a new estimator which is simultaneously highly robust and fully efficient. Suppose tat we have a sample $(y_1, \mathbf{x}_1), ..., (y_n, \mathbf{x}_n)$, then the estimator we propose is defined by the following three steps.

First step. We first obtain a consistent initial estimate of $\beta_0^* = \beta_0/||\beta_0||_2$. This estimator estimator is defined by

$$\widetilde{\beta}^* = \arg\min_{\beta^*} \tau((y_1, ..., y_n)), (\beta^{*\mathrm{T}}\mathbf{x}_1), ..., (\beta^{*\mathrm{T}}\mathbf{x}_n)),$$

where if $\mathbf{z} = (z_1, ...z_n)$ and $\mathbf{w} = (w_1, ..., w_n)$, we call $\tau(\mathbf{z}\,, \mathbf{w})$ the Kendall rank correlation between $\mathbf{z}$ and $\mathbf{w}$ given by

$$\tau(\mathbf{z}\,, \mathbf{w}) = \#\left\{(i, j), 1 \le i < j \le n : \mathrm{sign}\left((z_i - w_i)(z_j - w_j)\right) \ge 0\right\}.$$

This estimator was proposed first by Han [1987] for another models. It may be proved that this estimator is consistent for any strictly increasing link function h.

Second step. In this step we obtain initial estimators of $\eta_0 = ||\beta_0^*||$, δ_0 and α_0. For that purpose, let $\widehat{z}_i = \widetilde{\beta}^{*\mathrm{T}}\mathbf{x}_i$, $1 \le i \le n$. Then we fit the following negative binomial regression model with just one covariable : the distribution of $y_i|\widehat{z}_i$ is NB$(h(\eta_0\widehat{z}_j + \delta_0))$. Observe that this model holds exactly if we replace $\widehat{z}_i$ by $z_i = \beta_0^{*\mathrm{T}}\mathbf{x}_i$. Then we obtain estimators $\widetilde{\eta} = ||\gamma_0||$, $\widetilde{\delta}$ and $\widetilde{\alpha}_0$ using an M-estimator similar to the ones proposed by Marazzi and Yohai [2004] to estimate(μ_0, α_0) given a sample of a NB(μ_0, α_0) distribution. Finally, we complete the initial estimating by taking $\widetilde{\beta}$=$\widetilde{\alpha}\widetilde{\beta}^*$ as estimator of β_0

Third step. We transform the variables y_i as follows. Put $\theta = (\mu, \alpha)$ and let $p(.,\theta)$ and $F(.,\theta)$ the probability and distribution functions of the NB(μ,α) distribution respectively. Let y with NB(μ,α) distribution and $v = F(y,\theta) - up(y,\theta)$, where u has uniform distribution in $[0,1]$ (U(0,1)) and is independent of y. Then v has U(0,1) distribution. Call $\widetilde{\theta}_i = (\widetilde{\beta}^{\mathrm{T}}\mathbf{x}_i + \widetilde{\delta}, \widetilde{\alpha})$ and let $u_1, ..., u_n$ be i.i.d. U(0,1) variables which are independent of the sample. Then $v_i = F(y_i, \widetilde{\theta}_i) - u_i p(y, \widetilde{\theta}_i)$, $1 \leq i \leq n$, are approximately i.i.d. U(0,1) variables. Then we can detect outliers comparing the empirical distribution of $r_i = |v_i - 0.5|, 1 \leq i \leq n$, with the distribution $F_0(u) = 2uI([0,1])$ of $|u-0.5|$, where u has distribution U(0,1). Let $r_{(1)} <, ..., < r_{(n)}$ be the ordered sample and for $t = 1, ..., n$ let H_t be the empirical distribution of $r_{(1)}, ..., r_{(t)}$. Put $s_0 = \min\{s : \min_{r \geq 0.5-\varepsilon}(H_{n-s}(r) - H_0(r)) \geq 0\}$, where ε is a small number, e.g., $\varepsilon = 0.05$. Then, the observations such that $|v_i - 0.5| > r_{(n-s)} = \phi$ are going to be considered outliers and eliminated from the sample

Then, the final estimators are defined by

$$(\widehat{\beta}, \widehat{\delta}, \widehat{\alpha}) = \arg\min_{\beta,\alpha,\delta} L\left(y_1, y_2, ..., y_n, \beta, \alpha, \delta | \, \mathbf{x}_1, .., \mathbf{x}_n, \max_{1 \leq i \leq n} v_i \leq \phi \right),$$

where $L(y_1, ..., y_n, \beta, \alpha, \delta | \mathbf{t})$ denotes the conditional likelihood of $\mathbf{y} = (y_1, y_2, ..., y_n)$ given $\mathbf{t}$, when the parameters are β, α, δ. It can be proved that under the model we have $s_0/n \to 0$. We show that this implies that the final estimators are fully efficient. Moreover, a Monte Carlo simulation study show that they are also highly robust.

References

Aeberhard, W.H, Cantoni, E. and Heritier, S. (2014). Robust inference in the negative binomial regression model with an application to falls data. *Biometrics*, **70**, 920-931.

Han, A.K. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics*, **35**, 303-316.

Marazzi, A. and Yohai, V.J. (2010). Optimal robust estimates based on the Hellinger distance. *Advances in Data Analysis and Classification,* **4**, 169-179.