International Conference on Robust Statistics, 2017

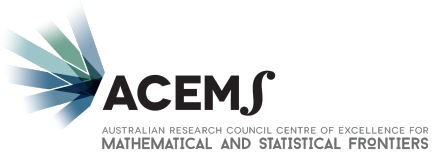
Wollongong, Australia

3–7 July 2017

Book of Abstracts

Premier sponsors

Sponsors



Hosts

Scientific committee

Luke Prendergast, La Trobe University (Chair)
Peter Filzmoser, Vienna University of Technology
Yanyuan Ma, University of South Carolina
Christophe Croux, KU Lueven
Elvezio Ronchetti, University of Geneva
Matías Salibián-Barrera, University of British Columbia
Brenton Clarke, Murdoch University

Local organising committee

Ray Chambers, University of Wollongong (Chair)
Garth Tarr, University of Newcastle
Samuel Mueller, University of Sydney
Alan Welsh, Australian National University
Walter Davis, University of Wollongong
Anica Damcevski, University of Wollongong
Karin Karr, University of Wollongong
Clint Shumack, University of Wollongong

Contents

IV Robust Survey Sampling Workshop

Part I

Keynote Presentations

Robust inference in functional data analysis

Graciela Boente

Universidad de Buenos Aires and CONICET, Argentina

Keywords: functional canonical correlation; functional principal components; functional semi–linear models; robustness

Functional data analysis provides modern analytical tools for data that are recoded as images or as a continuous phenomenon over a period of time. Because of the intrinsic nature of these data, they can be viewed as realizations of random functions often assumed to be in a Hilbert space such as $L^2(\mathcal{I})$, with $\mathcal{I}$ a real interval or a finite dimensional Euclidean set.

In particular, functional principal components or functional canonical correlation are statistical procedures developed to reduce the dimensionality retaining as much information as possible with respect to the measure of interest. To be more precise, the first q functional principal components provide the best q-dimensional approximation to random elements in Hilbert spaces, while functional canonical correlation is a tool to quantify correlations between pairs of observed random curves for which a sample is available.

On the other hand, partial linear modelling ideas have recently been adapted to situations in which functional data are observed. More precisely, two generalizations have been considered to deal with the problem of predicting a real-valued response variable using explanatory variables that include a functional element, usually a random function, and a random variable. The *semi-functional partial linear regression model* allows the functional explanatory variables to act in a free nonparametric manner, while the scalar covariate corresponds to the linear component. On the other hand, the so–called *functional partial linear model* assumes that the scalar response is explained by a linear operator of a random function and a nonparametric function of a real–valued random variable.

We will briefly discuss some approaches leading to obtain estimators of the principal directions in these situations less sensitive to atypical observations. In particular, the robust procedures developed to estimate the principal directions are used to develop robust methods under a *functional partial linear model*. If possible, we will also discuss methods to provide robust inferences for the canonical functions.

Robust models for small area estimation – random group effects vs. random group indexing

James Dawber, **Ray Chambers**

University of Wollongong, Australia

The standard approach to small area estimation based on unit level data is to assume that between area heterogeneity is a consequence of the values of random area effects. There is a very well-developed body of theory that addresses estimation of a regression function in this case and its use in prediction of small area characteristics of interest. However, there is another way of characterising between area heterogeneity that does not depend on recourse to a latent variable to distinguish differences between areas. Instead, a suitable ensemble regression function that covers the full spectrum of variability for the characteristic of interest is first used to index the population. Area heterogeneity is present if these index values cluster within areas, and small area estimation is based on that particular regression function within the ensemble that corresponds to an area-specific 'average' index. There is no random effect, with its consequent distributional assumptions, to complicate matters. In this context robust M-quantile ensemble models have seen considerable development in recent years, with a population unit's index defined by the index of that component M-quantile regression function with value equal to the unit's value for the characteristic of interest. In this presentation we will briefly discuss these two paradigms and then describe extensions of M-quantile ensemble models to binary and categorical responses. The extension of random indexing of population values based on the categorical M-quantile model will also be described and applied to small area estimation.

Robust statistical methods in the geosciences

Noel Cressie

University of Wollongong, Australia

Our Earth can be considered, at least for the moment, to be a sample of size one, although other planets in the habitable zone of distant solar systems are being discovered. Inference on geophysical parameters and processes rely on finding replication in spatial and temporal data. Global data from the Earth system are notoriously heterogeneous, so probability-distributional assumptions require care and circumspection. Further, the resulting statistical methods should not rely too heavily on those assumptions. In this talk, I give geophysical examples where robust (often median-based) methods are essential to the success of the scientific investigation.

Quantile regression in varying coefficient models

Irène Gijbels

Department of Mathematics and Leuven Statistics Research Centre, KU Leuven, Belgium

Quantile regression is an important tool for describing the characteristics of conditional distributions. Similarly to median versus mean estimates, median regression estimates are more robust against outliers in the response variable than mean regression estimates.

In both, mean and quantile regression, flexible models are often needed to capture the complexity of the underlying stochastic phenomenon, of which one wants to find the influence of covariates on a response variable. Among flexible models encountered in a multivariate covariate regression setting are additive models, single-index models, and varying coefficient models. The latter introduce additional flexibility as compared to multiple linear regression models by allowing the coefficients to vary with, for example, another covariate. Coefficients are no longer real parameters, but entire unknown functions.

In this talk we focus on quantile regression in such flexible varying coefficient models. As a major working tool we use B-spline approximations for the unknown coefficient functions. We discuss various aspects of quantile regression estimation: methods to prevent that the estimated quantile curves cross; homoscedasticity versus heteroscedasticity, including estimation under heteroscedasticity; and testing procedures for testing for specific shapes (constancy, monotonicity, convexity, ...) of coefficient functions. Estimation and testing procedures will be illustrated with finite-sample simulations and real data applications.

Part II

Invited Presentations

Marginal estimation under a general regression model with missing responses and covariates

Ana Bianco[1], Graciela Boente[1], Wenceslao González-Manteiga[2], Ana Pérez González[3]

[1] Universidad de Buenos Aires and CONICET, Argentina
[2] Universidad de Santiago de Compostela, Spain
[3] Universidad de Vigo, Spain

Keywords: missing at random; marginal estimation; regression; robustness

We consider a general regression model where missing data occur, in the responses and in the covariates. Our aim is to discuss estimation procedures for any marginal functional of the responses such as the median or any $\alpha-$quantile of the response variable. A missing at random condition is assumed in order to prevent from bias in the estimation of the marginal measures under a non-ignorable missing mechanism. Different approaches for the estimation of the marginal functional of interest are presented, including inverse probability weighting, the convolution of the estimators of the distributions of the errors and of the regression function and also double robust estimators, which protect against misspecification of the regression or the missing probability models. The proposals may be combined with robust procedures for the estimation of the regression function. The small sample behaviour of the proposals is illustrated through a Monte Carlo study on a partially linear model.

Relative error accurate statistic based on nonparametric likelihood

Lorenzo Camponovo, Taisuke Otsu

This paper develops a new test statistic for parameters defined by moment conditions that exhibits desirable relative error properties for the approximation of tail area probabilities. Our statistic, called the tilted exponential tilting (TET) statistic, is constructed by estimating certain cumulant generating function under exponential tilting weights. We show that the asymptoticp-value of the TET statistic can provide an accurate approximation to thep-value of an infeasible saddlepoint statistic, which is asymptotically chi-squared distributed with a relative error of order n^{-1} both in normal and large deviation regions. Numerical results illustrate the accuracy of the proposed TET statistic. Our results cover both just- and over-identified moment condition models.

Robust semiparametric estimators

Eva Cantoni[1], Xavier de Luna[2]

[1] Research Center for Statistics and Geneva School of Economics and Management, University of Geneva, Switzerland
[2] Umeå School of Business and Economics, Umeå University, Sweden

Keywords: outcome regression imputation; inverse probability weighting; double robust estimator; influence function; location-scale parameters; BMI

We aim at estimating location and scale parameters from a distribution law of interest in the context of missing data. We focus on situations where, while observations may be missing for variables of primary interest (responses/outcomes), there is a set of auxiliary variables (covariates) which are observed for all units. Under the assumption of missingness at random, commonly used semiparametric estimators are inverse probability weighted estimators (IPW), where observations are weighted with respect to the inverse of their probability to be observed given the covariates. A modified version of IPW is the augmented inverse probability weighted estimator (AIPW), also called doubly robust estimator. AIPW uses two auxiliary models (specified up to nuisance parameters), one as IPW for the missingness mechanism, and also a model for the outcome given the covariates.

We introduce robust (bounded influence) versions of these semiparametric estimators, which, in addition, can deal with contamination of the distribution law of interest. We describe the asymptotic properties of our new estimators and support the theoretical findings with a large simulation in a realistic setting. We finally present the conclusions of our analysis of a longitudinal study on BMI combining data from an intervention study and population wide record linked data, which has motivated this research.

Robust clustering tools based on optimal transportation

Tasio del Barrio

IMUVA, Universidad de Valladolid, Spain

Keywords: cluster analysis; optimal transportation; trimming methods; structured data

A robust clustering method for probabilities in Wasserstein space is introduced. This new 'trimmed k-barycenters' approach relies on recent results on barycenters in Wasserstein space that allow intensive computation, as required by clustering algorithms. The possibility of trimming the most discrepant distributions results in a gain in stability and robustness, highly convenient in this setting. As a remarkable application we consider a parallelized estimation setup in which each of m units processes a portion of the data, producing an estimate of k-features, encoded as k probabilities. We prove that the trimmed k-barycenter of the $m \times k$ estimates produces a consistent aggregation. We illustrate the methodology with simulated and real data examples.

Comparison of robust saddlepoint tests: the negative binomial regression case

Stephane Heritier[1], William Aeberhard[2], Eva Cantoni[3]

[1] Monash University, Australia
[2] Dalhousie University, Canada
[3] Research Center for Statistics and Geneva School of Economics and Management, University of Geneva, Switzerland

Keywords: exponential tilting; negative binomial regression; robust bounded influence tests; small samples

Inference on regression coefficients when the response variable consists of overdispersed counts is traditionally based on Wald, score and likelihood ratio tests. As the accuracy of the p-values of such tests becomes questionable in small samples, three recently developed saddlepoint tests based on general M-estimators are adapted to the negative binomial regression model. Under regularity conditions, these tests feature a relative error property of $O(1/n)$ under the null hypothesis with respect to the asymptotic chi-squared distribution. Extensive simulations show how these new tests outperform the traditional ones in small samples in terms of actual level with comparable power. Moreover, inference based on robust (bounded influence) versions of these tests remains reliable when the sample does not entirely conform to the model assumptions. An R package implementing all tests is readily available.

References

Aeberhard W., Cantoni E. and Heritier S. (2017). Saddlepoint tests for accurate and robust inference on overdispersed count data. *Computational Statistics & Data Analysis*, **107**, 162–175.

A robust Bayesian approach on mixture joint models for skew-longitudinal and survival data

Yangxin Huang[1], Jiaqing Chen[2], Chunning Yan[3]

[1] Department of Epidemiology and Biostatistics, College of Public Health, University of South Florida, USA
[2] Department of Statistics, College of Science, Wuhan University of Technology, P. R. China
[3] School of Management, Shanghai University, P. R. China

Keywords: Bayesian inference; longitudinal data analysis; mixture of hierarchical joint models; skew-t distributions; time-to-event

In longitudinal studies it is of interest to investigate how repeatedly measured markers in time are associated with a time to an event of interest and, in the mean time, the repeated measurements are often observed with the features of a heterogeneous population, non-normality and covariate measured with error due to longitudinal nature. Statistical analysis may complicate dramatically when one analyzes longitudinal-survival data with these features together. Recently, a mixture of skewed distributions has received increasing attention in the treatment of heterogeneous data involving asymmetric behaviors across subclasses, but there are relatively few studies accommodating heterogeneity, non-normality and measurement error in covariate simultaneously arose in longitudinal-survival data setting. Under the umbrella of a robust Bayesian inference, this article explores a finite mixture of semiparametric mixed-effects joint models with skewed distributions for longitudinal measures with an attempt to mediate homogeneous characteristics, adjust departures from normality and tailor accuracy from measurement error in covariate as well as overcome shortages of confidence in specifying a time-to-event model. The Bayesian mixture of joint modeling offers an appropriate avenue to estimate not only all parameters of mixture joint models, but also probabilities of class membership. Simulation studies are conducted to assess the performance of the proposed method, and a real example is analyzed to demonstrate the methodology. The results are reported by comparing potential models with various scenarios.

Saddlepoint approximations in the frequency domain

Davide La Vecchia, Elvezio Ronchetti

Research Center for Statistics and Geneva School of Economics and Management, University of Geneva, Switzerland

Keywords: Edgeworth expansion; generalized linear model in the frequency domain; p-value; short and long memory; testing in the presence of nuisance; tilting

Saddlepoint techniques provide accurate, higher order, small sample approximations to the distribution of estimators and test statistics. Except for a few simple models, these approximations are not available in the framework of stationary time series. We contribute to fill this gap by developing new saddlepoint approximations for frequency domain statistics. Our method is based on tilting devices of the Edgeworth expansion and it can be applied to frequency domain statistics admitting a valid Edgeworth expansion. Under short or long range serial dependence, for Gaussian and non Gaussian processes, we show how to derive and implement our saddlepoint techniques (density approximation and test in the presence of nuisance) for two relevant classes of statistics: ratio statistics and Whittle's estimator. A Monte Carlo study for different models illustrates the theory and compares (for Whittle's estimator) the new approximations with those obtained by first order asymptotic theory and the frequency domain bootstrap. An example based on data about the European Central Bank assets concludes the paper.

References

Andrews, D. W. and Lieberman, O. (2005). Valid Edgeworth expansions for the Whittle maximum likelihood estimator for stationary long-memory Gaussian time series. *Econometric Theory*, **21**(04), 710–734.

Beran, J. (1993). Fitting long-memory models by generalized linear regression. *Biometrika*, **80**(4), 817–822.

Dahlhaus, R. and Janas, D. (1996). A frequency domain bootstrap for ratio statistics in time series analysis. *Annals of Statistics*, **24**(5), 1934–1963.

Daniels, H. E. (1956). The approximate distribution of serial correlation coefficients. *Biometrika*, **43**(1/2), 169–185.

Easton, G. S. and Ronchetti, E. (1986). General saddlepoint approximations with applications to L-statistics. *Journal of the American Statistical Association*, **81**(394), 420–430.

Field, C. and Robinson, J. (2013). Relative errors for bootstrap approximations of the serial correlation coefficient. *Annals of Statistics*, **41**(2), 1035–1053.

Kim, Y. M. and Nordman, D. J. (2013). A frequency domain bootstrap for Whittle estimation under long-range dependence. *Journal of Multivariate Analysis*, **115**, 405–420.

Phillips, P. (1978). Edgeworth and saddlepoint approximations in the first-order noncircular autoregression. *Biometrika*, **65**, 91–98.

Robust functional principal components for irregularly spaced longitudinal data

Ricardo Maronna

University of La Plata, Argentina

Keywords: MM-estimators; B-splines

Consider a data set x_{ij}, with $i = 1, ..., n$ and $j = 1, ..., p_j$, where x_{ij} is the $j-$th observation of the random function $X_i(.)$ observed at time t_{ij}. We propose a parsimonious representation of the data by a linear combination of a set of q smooth functions $H_k(.)$ $(k = 1, .., q)$ in the sense that $x_{ij} \approx \sum_{k=1}^{q} \beta_{ki} H_k(t_{ij})$, such that it (a) is resistant to atypical X_i's ("case contamination"), (b) is resistant to isolated gross errors at some t_{ij} ("cell contamination"), and (c) can be applied when the set $\{t_{ij}\}$ depends on i ("irregularly spaced data").

Among the abundant literature on this subject, Boente et al. (2015) Lee et al. (2013) and Cevallos Valdiviezo (2016) deal with item (a), and Yao et al (2005) deal with (c).

Our approach to deal with all three problems, which is similar to MM-estimation, is defined as follows. Let $B_l(.)$ be a basis of B-splines; for $\alpha = \{\alpha_{kl}\}$ and $\beta = \{\beta_{ki}\}$. Put

$$\widehat{x}_{ij}(\alpha.\beta) = \sum_{k=1}^{q} \beta_{ki} H_k(t_{ij})$$

with $H_k(t) = \sum_{l=1}^{m} \alpha_{kl} B_l(t)$. Then the estimator is given by

$$\left(\widehat{\mathbf{a}}, \widehat{\beta}\right) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \sum_{k=1}^{q} \widehat{\sigma}_j^2 \rho\left(\frac{x_{ij} - \widehat{x}_{ij}(\alpha, \beta)}{\widehat{\sigma}_j}\right).$$

where $\widehat{\sigma}_j$ are previously computed local scales.

The parameters are computed by an iterative algorithm starting from deterministic initial values, which are the most complex part of the procedure.

References

Boente, G. and Salibián-Barrera, M. (2015). S-Estimators for Functional Principal Component Analysis. *Journal of the American Statistical Association*, **110**, 1100–1111.

Cevallos Valdiviezo, H. (2016). On Methods for Prediction Based on Complex Data with Missing Values and Robust Principal Component Analysis, PhD thesis, Ghent University (supervisors Van Aelst S. and Van den Poel, D.).

Lee, S., Shin, H. and Billor, N. (2013). M-type smoothing spline estimators for principal functions. *Computational Statistics and Data Analysis*, **66**, 89–100.

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

Sliced average variance estimation for multivariate time series

Markus Matilainen, Klaus Nordhausen, Hannu Oja

University of Turku, Finland

Keywords: joint diagonalization; non-linear regression; prediction; sliced inverse regression; supervised dimension reduction

Linear supervised dimension reduction has a long tradition for iid data with a rich literature. The idea in this setup is to find all linear combinations of a predictor vector $\mathbf{x}$ which are needed to model a response y, where the functional relationship between the response and the explaining variables is assumed to be unknown. Multivariate time series occur also more and more often in a regression context, where the goal is to model one time series as a function of several other time series. Supervised dimension reduction in time series context is much more difficult as the dependency between the response series and the explaining series might also be lagging in time. Using supervised iid dimension reduction methods by adding lagged time series as new variables to the data increases the dimension of the problem dramatically and at the same time reduces the sample size.

Matilainen et al. (2017) recently proposed a supervised dimension reduction approach aimed especially for time series. Under quite week assumptions authors use for their TSIR approach the approximate joint diagonalization of several supervised matrices, which consider the temporal nature of the data. These matrices were inspired by the iid supervised dimension reduction method sliced inverse regression (SIR) (Li, 1991) and have similar drawbacks as the standard SIR. Similar as SIR for iid data was improved by SAVE (sliced average variance estimation) (Cook, 2000), we would like to suggest a SAVE for time series version TSAVE here. Also a hybrid of TSIR and TSAVE is introduced.

Our methods are first illustrated by examples. Then their predictive performance compared to other methods, including TSIR, is measured with a simulation study. Finally robustness properties of the algorithms are briefly discussed.

References

Cook, R. D. (2000). Save: a method for dimension reduction and graphics in regression, *Communications in Statistics - Theory and Methods*, **29**, 2109–2121.

Li K.-C. (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86**(414), 316–327.

Matilainen M., Croux C., Nordhausen K. and Oja, H. (2017). Supervised dimension reduction for multivariate time series, *Submitted*.

Robust Bayesian hypothesis testing

John Ormerod, Sarah Romanes

School of Mathematics and Statistics, University of Sydney, Australia

Keywords: Bartlett's paradox; cake priors; likelihood ratio tests; mixture models

Natural approaches to hypothesis testing in the Bayesian paradigm leads to Lindley's and Bartlett's like paradoxes. We develop "cake priors", a new class of priors to circumvent these problems and show how Bayesian tests can be constructed to mimic likelihood ratio tests controlling for type I error under the null hypothesis. Such tests are brittle in the presence of outliers. We handle outliers by including an additional mixture component to detect outliers leading to a small modification of our Bayesian tests leading to a finite mixture of likelihood ratio test statistics. Preliminary empirical evidence suggests that these tests are robust to outliers whilst controlling for type I error under the null hypothesis.

Cleaning a large set of time series for common, cluster and specific outliers

Daniel Peña, Pedro Galeano

Universidad Carlos III de Madrid, Spain

Keywords: dynamic factor models; time series clustering; generalized dynamic principal components; additive outliers

We presents a procedure to clean large set of time series from outliers that is based on analyzing linear transformations of the data. Three types of outliers are considered. First common outliers that affect to all the time series. Second, cluster outliers that affects to groups of time series and third specific outliers that affect individual time series. We assume that the vector of series has been generated from a dynamic factor model with cluster structure and that the outliers can affect either the latent common factors, the factors affecting the clusters or the idiosyncratic noise. The effects of the outliers on the observed time series are analyzed and it is shown that outliers can be detected using linear transformations of the observed series constructed from eigenvectors of robust estimates of the autocovariance matrices of the observed time series. We propose a simple and fast procedure based on these linear transformations for outlier detection that it is illustrated with simulations and the analysis of a real data example. We also explore the Generalized Dynamic Principal Components for finding common factors and clean the common outliers.

References

Ando, T. and Bai, J. (2017). Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association*, DOI:10.1080/01621459.2016.1195743

Peña, D. and Yohai, V. J. (2016). Generalized dynamic principal components. *Journal of the American Statistical Association*, **111**(515), 1121–1131.

Hierarchical mixture models for longitudinal immunologic data with multiple features

Tingting Qin, Ping Yin

Department of Epidemiology and Biostatistics, School of Public Health, Huazhong University of Science and Technology, P. R. China

Keywords: Bayesian inference; longitudinal immunologic data; finite mixture model; heterogeneity; skew distributions; missing data

It is a common practice to analyze longitudinal data frequently arisen in medical studies using various mixed-effects models in the literature. However, the following issues may standout in longitudinal data analysis: (i) In clinical practice, the profile of each subject's response from a longitudinal study may follow a "broken stick" like trajectory, indicating multiple phases of increase, decline and/or stable in response. Such multiple phases (with change-points) may be an important indicator to help quantify treatment effect and improve management of patient care. To estimate change-points, the various mixed-effects models become a challenge due to complicated structures of model formulations; (ii) an assumption of homogeneous population for models may be unrealistically obscuring important features of between-subject and within-subject variations; (iii) normality assumption for model errors may not always give robust and reliable results, in particular, if the data exhibit non-normality; and (iv) the response may be missing and the missingness may be non-ignorable. In the literature, there has been considerable interest in accommodating heterogeneity, non-normality or missingness in such models. However, there has been relatively little work concerning all of these features simultaneously. There is a need to fill up this gap as longitudinal data do often have these characteristics. In this article, our objectives are to study simultaneous impact of these data features by developing a Bayesian mixture modeling approach-based Finite Mixture of Change-point Mixed-Effects (FMCME) models with skew distributions, allowing estimates of both model parameters and class membership probabilities at population and individual levels. Simulation studies are conducted to assess the performance of the proposed method, and a real clinical example is analyzed to demonstrate the proposed methodologies and to compare modeling results of potential mixture models under different scenarios.

Nonparametric tests for multi-dimensional M-estimators

John Robinson

University of Sydney, Australia

Keywords: weighted bootstrap; saddlepoint approximation; empirical exponential family

Given a set of M-estimation equations from some parametric model, perhaps with some robust adjustments, a test statistic for a test of a subset of the parameters, based on the M-estimates, is derived from the cumulant generating function of the score functions. When the distribution, and so the cumulant generating function, of the data is not certain, a weighted empirical distribution and an empirical cumulant generating function can be used to obtain a nonparametric test statistic and weighted bootstrap sampling can be used to approximate its distribution and so to obtain p- values for the test. A saddlepoint approximation, with good relative error properties, is given for this bootstrap distribution. Numerical examples illustrate the properties of the tests and approximations for linear regression and generalized linear models.

Detecting deviating data cells

Peter Rousseeuw, Wannes Van den Bossche

KU Leuven, Belgium

Keywords: algorithms; outliers; visualization

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 situation, outliers may be (a) undesirable errors which can adversely affect the data analysis, or (b) valuable nuggets of unexpected information. In statistics and data science the word outlier usually refers to a row of the data matrix, and the methods to detect such outliers only work when at least half the rows are clean, see e.g. Rousseeuw and Leroy (1987). But often many rows have a few contaminated cell values, especially in high-dimensional data, which implies few rows are entirely clean (Alqallaf et al., 2009).

Such contaminated cells may not be visible by looking at each variable (column) separately. We propose the first method to detect deviating data cells in a multivariate sample which takes the correlations between the variables into account. It has no restriction on the number of clean rows, and can deal with high dimensions. Other advantages are that it provides predicted values of the outlying cells, while imputing missing values at the same time.

The results are visualized by *cell maps* in which the colors indicate which cells are suspect and whether their values are higher or lower than predicted.The software allows to block cells, to zoom in on a part of the data, and to adjust the contrast.

We illustrate the method on several real data sets, where it uncovers more structure than found by purely columnwise methods or purely rowwise methods. The proposed method can help to diagnose why a certain row is outlying, e.g. in process control. It may also serve as an initial step for estimating multivariate location and scatter matrices, as in Agostinelli et al. (2015).

References

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

Alqallaf F., Van Aelst S., Yohai V. J., and Zamar R. H. (2009). Propagation of outliers in multivariate data. *Annals of Statistics*, **37**, 311–331.

Rousseeuw P. J. and Leroy A. M. (1987), *Robust Regression and Outlier Detection*. Wiley, New York.

On the interplay between semiparametric optimality and robustness

Michael Stewart[1], Alan Welsh[2]

[1] University of Sydney, Australia
[2] Australian National University, Australia

A powerful theme in robust statistics is the "trade-off" between efficiency and robustness. In a sense these two terms are antagonistic in that, at least classically, one is usually obtained at the expense of the other. However, semiparametric theory provides an opportunity to qualitatively specify the desired type and degree of robustness and then within that envelope, derive procedures which are "as efficient as possible" in a certain pleasing sense. We firstly provide a selective survey of the literature and point out some interesting recent work in this direction. Secondly we examine a simple problem, that of inference regarding location, in fine detail in a way which clearly illustrates the main ideas we wish to communicate. Finally we report on some recent progress along these lines for mixed models.

Envelope quantile regression

Zhihua Su[1], Shanshan Ding[2], Guangyu Zhu[3], Lan Wang[4]

[1] University of Florida, USA
[2] University of Delaware, USA
[3] University of British Columbia, Canada
[4] University of Minnesota, USA

Keywords: envelope model; quantile regression; dimension reduction; reducing subspace; generalized method of moments

Quantile regression offers a valuable complement of classical mean regression for robust and comprehensive data analysis in a variety of applications. We propose a novel *envelope quantile* regression method (EQR) that adapts a nascent technique called *enveloping* (Cook et al., 2010) to improve the efficiency of standard quantile regression. The new method aims to identify material and immaterial information in a quantile regression model and use only the material information for estimation. By excluding the immaterial part, the EQR method has the potential to substantially reduce the estimation variability with standard quantile regression. Unlike existing envelop model approaches which mainly rely on the likelihood framework, our proposed estimator is defined through a set of nonsmooth estimating equations. We facilitate the estimation via the generalized method of moments and derive the asymptotic normality of the proposed estimator by applying empirical process techniques. Furthermore, we establish that EQR is asymptotically more efficient than (or at least as asymptotically efficient as) the standard quantile regression estimators without imposing stringent conditions. Hence, our work advances the envelope model theory to general distribution-free settings. We demonstrate the effectiveness of the proposed method via Monte-Carlo simulations and a real data example.

References

Cook, R.D., Li, B. and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression, *Statistica Sinica*, **20**, 927–1010.

Blind source separation based on robust autocovariance matrices

Sara Taskinen[1], Jari Miettinen[1,2], Klaus Nordhausen[3], David E. Tyler[4]

[1] Department of Mathematics and Statistics, University of Jyväskylä, Finland
[2] Department of Mathematics and Statistics, University of Turku, Finland
[3] Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Austria
[4] Department of Statistics, Rutgers University, USA

Keywords: affine equivariance; M-autocovariances; stationarity; time series; SOBI

Assume that the observed p time series are linear combinations of p latent uncorrelated weakly stationary time series. The aim in blind source separation (BSS) is then to find an estimate for the unmixing matrix which transforms the observed time series back to uncorrelated latent time series. In the classical SOBI (Second Order Blind Identification) method, approximate joint diagonalization of the sample covariance matrix and sample autocovariance matrices with several lags is used to estimate the unmixing matrix (Belouchrani et al. 1997). For a thorough discussion on SOBI estimators and their statistical properties, see Miettinen et al. (2016).

It is well known that in the presence of outliers, the sample covariance matrix and sample autocovariance matrices perform poorly and yield to unreliable unmixing matrix estimates. Ilmonen et al. (2015) tackle this problem by using a robust scatter matrix and spatial sign autocovariance matrices to solve the BSS problem. In this talk we generalize their method by using so-called M-autocovariance matrices in the estimation. The M-autocovariance matrices are similar to the classical M-estimators (Maronna, 1976) in that they downweight the outliers using some preselected, bounded weight function. We use finite-simple simulation studies and a real data example to illustrate the performance of the robust SOBI method.

References

Belouchrani, A., Abed-Meriam, K., Cardoso, J.F. and Moulines, R. (1997). A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, **45**, 434–444.

Ilmonen, P., Nordhausen, K., Oja, H. and Theis, F.J. (2015). An affine equivariant robust second order blind source separation method. In Vincent, E., Yeredor, A., Koldovsky, Z. and Tichavsky, P. (eds). *Latent Variable Analysis and Signal Separation: 12th International Conference, LVA/ICA 2015*, 328–335, Springer.

Maronna, R.A. (1976). Robust M-estimators of multivariate location and scatter. *Annals of Statistics*, **4**, 51–67.

Miettinen, J., Illner, K., Nordhausen, K., Oja, H., Taskinen, S. and Theis, F. J. (2016), Separation of uncorrelated stationary time series using autocovariance matrices. *Journal of Time Series Analysis*, **37**, 337–354.

Robust high-dimensional principal component analysis and variable screening

Stefan Van Aelst

KU Leuven, Belgium

Many robust principal component analysis methods are not suitable for high-dimensional data when most or all of the observations are contaminated in some of their cells. However, this situation is not uncommon in practice and can be formalized by the independent contamination model. To robustly estimate principal components in this setting, we consider the componentwise S-estimator of Boente and Salibián-Barrera (2015) and a least trimmed squares (LTS) variant. We propose an efficient algorithm to calculate these estimators by using estimating equations and deterministic starting values. We then illustrate how this method can be used in a robust variable selection procedure for ultra-high dimensional regression analysis. We propose a robust version of Factor Profiled Sure Independence Screening (Wang, 2012). By assuming that the predictors can be represented by a few latent factors, this method can handle correlation among the candidate predictors. We use robust componentwise principal components to estimate the factors. Then, a robust regression method is applied on the profiled variables to screen for the most important predictors.

Parts of this work is in collaboration with Holger Cevallos-Valdiviezo, Matías Salibián-Barrera and Yixin Wang, respectively.

References

Boente G. and Salibián-Barrera M. (2015). S-estimators for functional principal component analysis. *Journal of the American Statistical Association*, **110**(511), 1100–1111.

Wang, H. (2012). Factor profiled sure independence screening. *Biometrika*, **99**(1), 15–28.

Joint diagonalisation of scatter operators: functional fourth order blind identification

Germain Van Bever[1], Bing Li[2], Hannu Oja[3], Radka Sabolová[4], Frank Critchley[4]

[1] Ecares & Département de Mathématique, Université libre de Bruxelles, Belgium
[2] Penn State University, USA
[3] University of Turku, Finland
[4] The Open University, United Kingdom

Keywords: invariant coordinate selection; functional data; symmetric component analysis; independent component analysis

With the increase in measurement precision, functional data is becoming common practice. Relatively few techniques for analysing such data have been developed, however, and a first step often consists in reducing the dimension via Functional PCA, which amounts to diagonalising the covariance operator. Joint diagonalisation of a *couple* of scatter functionals has proved useful in many different setups, such has Independent Component Analysis (ICA), Invariant Coordinate Selection (ICS), etc.

The main part of this talk consists in extending the Fourth Order Blind Identification procedure to the case of data on a separable Hilbert space (with classical FDA setting being the go-to example). In the finite-dimensional setup, this procedure provides a matrix Γ such that ΓX has independent components, if one assumes that the random vector X satisfies $X = \Psi Z$, where Z has independent marginals and Ψ is an invertible mixing matrix. When dealing with distributions on Hilbert spaces, two major problems arise: (i) the notion of marginals is not naturally defined and (ii) the covariance operator is, in general, non invertible. These limitations are tackled by reformulating the problem in a coordinate-free manner and by imposing natural restrictions on the mixing model.

The proposed procedure is shown to be Fisher consistent and affine invariant. A sample estimator is provided and its convergence rates are derived. The procedure is amply illustrated on simulated and real datasets.

References

Cardoso, J.-F. (1989). Source separation using higher moments. *Proceedings of IEEE international conference on acoustics, speech and signal processing*, 2109–2112.

Tyler, D., Critchley, F., Dumbgen, L. and Oja, H. (2009), Invariant co-ordinate selection. *Journal of the Royal Statistical Society, Series B*, **71**(3), 549–592.

Ramsay, J. and Silverman, B.W. (2006), *Functional Data Analysis*, 2nd Edition, Springer, New-York.

Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F. and Moulines, E. (1997), A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, **45**(2), 434–444.

Bootstrap inference in high dimensional ANOVA with skewed data and small sample size

Haiyan Wang, Richard Opoku-Nsiah

Department of Statistics, Kansas State University, USA

Keywords: hypothesis testing; local alternative; Edgeworth expansion; bootstrap inference

The bootstrap inference in classical ANOVA has been well studied by Fisher and Hall (1990) and others. Fisher and Hall (1990) showed that the bootstrap distribution approximates the distribution of their pivotal statistics in large sample sizes while the number of factor levels is small. In the Big Data world now, however, researchers often need to compare a large number of treatments (for example, cultivars or genotypes) with very small number of replications per treatment. The bootstrap distribution of the statistics studied in Fisher and Hall no longer approximates the distribution of the test statistic in this setting. In fact, the bootstrap statistics and the original test statistic could have drastically different support and very different stochastic behavior. This study presents the bootstrap inference in the large p, small n setting with heteroscedastic data. We discuss appropriate test statistics that can be bootstrapped in this setting. We give Edgeworth expansion and Cornish-Fisher expansion of the test statistics and provide the convergence rate of the bootstrap approximation to the distribution of the test statistics in the large p, small n setting. Numerical studies will be presented along with the theoretical results.

References

Fisher, N. I. and Hall, P. (1990). On bootstrap hypothesis testing. *Australian Journal of Statistics*, **32**, 177–190.

Robustness and efficiency of quadratic inference function estimators

Suojin Wang[1], Samuel Müller[2], A.H. Welsh[3]

[1] Texas A&M University, USA
[2] University of Sydney, Australia
[3] Australian National University, Australia

Keywords: Generalized estimating equations; longitudinal data; ridged estimators

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 and 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 and other properties of quadratic inference function estimators that show that the issues are more subtle than we had anticipated. A ridged version of the quadratic inference function estimators is also considered.

References

Qu, A., Lindsay, B.G. and Li, B. (2000). Improving generalized estimating equations using quadratic inference functions. *Biometrika*, **87**(4), 823–836.

Qu, A. and Song, P.X.-K. (2004). Assessing robustness of generalized estimating equations and quadratic inference functions. *Biometrika*, **91**(2), 447–459.

Regression phalanxes

Ruben Zamar[1], Fred (Hongyang) Zhang[2], William Welch[1]

[1] University of British Columbia, Canada
[2] Splunk Vancouver, Canada

Keywords: model ensembling, hierarchical clustering, Lasso, random forest

Tomal, et al. (2015) introduced the notion of "phalanxes" in the context of rare-class detection in two-class classification problems. A phalanx is a subset of features that work well for classification tasks. In this paper, we propose a different class of phalanxes for application in regression settings. We define a "regression phalanx" – a subset of features that work well together for prediction. We propose an algorithm which automatically chooses Regression Phalanxes from high-dimensional data sets using hierarchical clustering and builds a prediction model for each phalanx for further ensembling. Through extensive simulation studies and several real-life applications in various areas (including drug discovery, chemical analysis of spectra data, microarray analysis and climate projections) we show that an ensemble of regression phalanxes generally improves prediction accuracy when combined with effective prediction methods like Lasso or random forests.

References

Tomal, J. H., Welch, W. J., and Zamar, R. H. (2015). Ensembling classification models based on phalanxes of variables with applications in drug discovery. *The Annals of Applied Statistics*, **9**(1), 69–93.

Part III

Contributed Presentations

Detecting influential observations using forward search method in optimal dynamic treatment regimes

Nur Raihan Abdul Jalil, Nur Anisah Mohamed

Institute of Mathematical Sciences, Faculty of Science, University of Malaya, Malaysia

Keywords: quadratic inference functions; forward search; optimal dynamic treatment regimes; influential observations

A dynamic treatment regime is a set of decision rules that assign treatment only when or if the individual need the treatment. An optimal dynamic treatment regime is a semiparametric method which determines a treatment decision over a series of time to maximize the mean response. The regret-regression is an alternative method to estimate the decision rules from observational data. Myopic regret-regression (MRR) is a short-term strategy derived from the regret-regression method. Quadratic inference functions (QIF) is a method that can be used to analyze longitudinal data. The combination of MRR with the QIF is possible and yet more efficient to analyze the correlation data in optimal dynamic treatment regimes even with misspecified working correlation structure. We proposed a forward search method for the QIF-MRR to investigate the influential observations in the data studies. A simulation study and an application to anticoagulation data are applied to illustrate this work.

References

Murphy, S. (2003). Optimal dynamic treatment regimes, *Journal of the Royal Statistical Society, Series B*, **65**(2), 331–355.

Henderson, R., Ansell, P. and Alshibani, D. (2010). Regret-regression for optimal dynamic treatment regimes, *Biometrics*, **66**(4), 1192-1201.

Qu, A., Bruce, G. and Bing, L. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, **87**(4), 823-836.

Robustness of all possible comparisons criterion for analyzing screening experiments

Abu Zar Md. Shafiullah, Arden Miller, Chris Triggs

The University of Auckland, New Zealand

Keywords: APC method; false positive rates; orthogonal designs; screening experiments

The All Possible Comparisons (APC) method (Miller, 2005) considers the problem of selecting the active effects when orthogonal designs are used in the analysis of screening experiments.Based on the APC method of Miller (2005) an AIC-style criterion called the APC-criterion is developed which controls the false positive rates, viz. individual error rate, experimentwise error rate (Hamada and Balakrishnan, 1998) or false discovery rate (Kimel, Benjamini and Steinberg, 2008) at the specified level set by the experimenter.Our simulation studies reveal the robustness of the APC-criterion to (i) non-normality of error distribution and (ii) small fraction of outliers in the response variable.In order to capture the trade off between level of error control and power of screening, we estimate the accuracy rate as a performance measure.We observe that the effect size, proportion of active effects, type and level of error control, etc. affect the performance but the robustness of APC-criterion is maintained nicely over a wide range of experimental setups, specially for bigger designs.

References

Hamada, M. and Balakrishnan, N. (1998). Analyzing unreplicated factorial experiments: a review with some new proposals. *Statistica Sinica*, **8**(1), 1–41.

Kimel, M.T., Benjamini, Y. and Steinberg D.M. (2008). The false discovery rate for multiple testing in factorial experiments. *Technometrics*, **50**(1), 32–39.

Miller, A. (2005). The analysis of unreplicated factorial experiments using all possible comparisons. *Technometrics*, **47**(1), 51–63.

Robust comparisons of variation using ratios of interquantile ranges

Chandima Arachchige, Luke Prendergast, Robert Staudte

Department of Mathematics and Statistics, La Trobe University, Australia

Keywords: interquantile range ratio; interval estimators; partial influence functions

The ratio of variance test (F-test) plays an important role among the tests available to compare the variance between two populations. However, it is assumed that the data is sampled from normal distributions. In addition, this test is not robust since the variance estimator is highly affected by outliers. Therefore, the motivation of our study was to introduce a robust estimator that can be used to compare variation between populations. In our study, we propose the IQR (interquantile range) ratio estimator as a robust alternative to the F-test. We start by discussing existing hypothesis tests based on the IQR (Shoemaker, 1999; Marozzi, 2012), before introducing interval estimators for the ratio of two IQRs. Using the Quantile Optimality Ratio (QOR) approach (Prendergast and Staudte, 2015) we show that excellent coverage can be achieved for a large choice of distributions (e.g. normal, log normal, exponential, chi-square and Pareto to name just a few). We also derive the influence function and show how it can be used to gain insight into various properties of the estimator.

References

Marozzi, M. (2012). A combined test for differences in scale based on the interquantile range. *Statistical Papers*, **53**(1), 61–72.

Prendergast, L. A. and Staudte, R. G. (2015). Exploiting the quantile optimality ratio to obtain better confidence intervals for quantiles. *arXiv preprint*, arXiv:1505.04234.

Shoemaker, L. H. (1999). Interquartile tests for dispersion in skewed distributions. *Communications in Statistics – Simulation and Computation*, **28**(1), 189-205.

Maximum Lq-likelihood estimation for the parameters of multivariate t-distribution

Olcay Arslan[1], Y. Murat Bulut[2], Fatma Zehra Dogru[3]

[1] Ankara University, Turkey
[2] Eskişehir Osmangazi University, Turkey
[3] Giresun University, Turkey

Keywords: EM; ML; MLq; multivariate t-distribution

The t-distribution (univariate or multivariate) has many useful applications in robust statistical analysis. The parameter estimation of the t-distribution is carried out using the maximum likelihood (ML) estimation method, and the ML estimates are obtained via the Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977). However, it is well-known that when estimating the degrees of freedom parameter along with the other parameters the estimators become no longer locally robust due to the unboundedness of the score function for the degrees of freedom parameter. Therefore, to obtain robust estimators the degrees of freedom parameter is usually assumed to be known and considered as a robustness tuning constant (e.g., see, Lange et al. 1989). In this study, we give alternative estimators for all the parameters of the multivariate t distribution using the maximum Lq (MLq) likelihood estimation method introduced by Ferrari and Yang (2010). We show that unlike the ML case, the score function for the degrees of freedom parameter obtained from the MLq estimation method is also bounded so that the resulting estimators for all the parameters gain local robustness property measured by the influence function. We adapt the EM algorithm to obtain the MLq estimates for all the parameters of the multivariate t-distribution. We provide a simulation study and a real data example to illustrate the performance of the MLq likelihood estimators over the ML estimators and observe that the MLq likelihood estimators have considerable superiority over the ML estimators in terms of MSE and bias values.

References

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

Ferrari, D. and Yang, Y. (2010). Maximum Lq-likelihood estimation. *Annals of Statistics*, **38**(2), 753–783.

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989). Robust statistical modeling using the t-distribution. *Journal of the American Statistical Association*, **84**(408), 881–896.

Testing for anomalies in intertemporal choice: a nonparametric approach

Stefano Bonnini[1], Isabel Maria Parra Oller[2]

[1] Department of Economics and Management, University of Ferrara, Italy
[2] Department of Economics and Business, University of Almeria, Spain

Keywords: intertemporal choice; magnitude effect; delay effect; sign effect; permutation test

Intertemporal choice (IC) concerns the study of how people make their decisions through time. A typical example concerns the question whether it is better to save money now, in order to consume more in the future, or to consume today by giving up a greater future consumption. The discounting utility model (DU model) (Samuelson, 1937) assumes that people preferences maintain constant across time and amount. This model is similar to the capitalization function used in the investment and banking practice. In the DU model, people "discount" the value of future outcomes at a constant discount rate. Some empirical studies have shown anomalous people's behaviors that go against the DU model premises, called anomalies in IC (Thaler, 1981; Benzion et al., 1989; Loewenstein and Prelec, 1992). For example, the discount rate of future gains/losses seems to be a decreasing function of the amount (*magnitude effect*); the discount rate varies inversely to the time delay (*delay effect*); the discount rates of gains and losses are different (*sign effect*).

Usually, empirical studies on IC consist of lab or field experiments where the interviewed subjects are asked to answer a question such as "what is the minimum amount you would like to receive in return of postponing at time t the receiving of a reward equal to m?" or "what is the maximum amount you are willing to pay in return of anticipating at time t the payment of a fine equal to m?", for different values of t and m. The methods used to test for anomalies in IC, are parametric techniques such as t-test, ANOVA, regression analysis or simple rank tests, which don't take into account the multivariate nature of responses, the articulation of some tested hypothesis (e.g. stochastic ordering), the presence of confounding factors and the possible violation of the assumption about the underlying distribution of data, especially for small sample sizes.

We propose a nonparametric method to test for anomalies in IC, based on combined permutation tests (see Bonnini et al., 2014), which overcomes these limitations. It is robust because based on less stringent assumptions than the cited methods. A simulation study proves the good performance of the proposed method. The results of the analysis of real experimental data are also presented.

References

Benzion, U., Rapoport, A. and Yagil, J. (1989). Discount rates inferred from decisions: An experimental study. *Management Science,* **35**(2), 270–284.

Bonnini, S., Corain, L., Marozzi, M. and Salmaso, L. (2014). *Nonparametric Hypothesis Testing. Rank and Permutation Methods with Applications in R.* Wiley:Chichester.

Loewenstein, G. and Prelec, D. (1992). Anomalies in intertemporal choice: evidence and an intepretation. *Quarterly Journal of Economics,* **107**(2), 573–597.

Samuelson, P.A. (1937). A note on measurement of utility. *The Review of Economic Studies,* **4**(2), 155–161.

Thaler, R. H. (1981). Some empirical evidence on dynamic inconsistency. *Economic Letters,* **8**, 201–207.

Confidence intervals for quantiles from grouped data

Dilanka Dedduwakumara, Luke Prendergast

Department of Mathematics and Statistics, La Trobe University, Australia

Keywords: histograms; generalized lambda distribution; quantile density

The estimation of confidence intervals for quantiles have been discussed mainly for continuous data in the literature. Until now, and to the best of our knowledge, viable methods for obtaining such intervals when the data is in grouped format were yet to be discovered. Motivated by that fact, we introduce several methods to obtain confidence intervals for quantiles when availed with grouped data. Grouped data may include traditional histograms and frequency data within unequal width bins. Our preferred method for interval estimation is to approximate the underlying density using the Generalized Lambda distributions (GLD). GLD distributions are often encountered in fields such as finance where it may be used to approximate a very large number of well-known distributions. Use of the bda package in R statistical software enables the GLD density estimation based on grouped (or pre-binned) data. We compare the GLD estimation method with two other methods that we also introduce which are based on a frequency approximation approach and a linear interpolation approximation of the density considered by Lyon et al (2016). These other methods can also be easily implemented in statistics packages like R, SAS etc. as well as packages such as Excel. Our methods are strongly supported by simulations that show that excellent coverage can be achieved for a wide number of distributions which includes even highly-skewed distributions like the log-normal, dagum, Singh-Maddala and Pareto distributions. We also apply our methods to real data sets.

References

Wang, B. (2015). `bda`: density estimation for grouped data. R package version 5.1.6.

Lyon, M, Cheung, L.C. and Gastwirth, J.L. (2016). The advantages of using group means in estimating the Lorenz curve and Gini index from grouped data, *The American Statistician*, **70**(1), 25–32.

Some alternatives for inhomogeneous Poisson point processes for presence only data

Hassan Doosti[1], Lewi Stone[2], Yan Wang[2], Aihua Xia[3]

[1] Macquarie University, Australia
[2] RMIT University, Australia
[3] The University of Melbourne, Australia

Keywords: species distribution models; inhomogeneous Poisson point processes; negative binomial point processes; imperfect detection; presence only data

Inhomogeneous Poisson point (IPP) process plays a vital role in species distribution modelling. An IPP model can be interpreted as a point process whose counts of points in disjoint sets are independent Poisson random variables Brown and Xia (2002). In the first glance, the independence of counts of points in disjoint sets makes IPP models unsuitable in species distribution modelling as animals don't behave independently, e.g., offsprings can't independently move away from their habitat. However, in this talk, we investigate a more general alternative class of point processes introduced in Xia and Zhang (2012) and show that these processes can respectively capture negative and positive dependence in species distribution modelling, and also are more flexible than IPP models. Furthermore weprove that when we have access to only one realisation the estimation of parameters based on IPP is robust. A numerical study support our findings.

References

Brown, T. C. and Xia, A. (2002). How many processes have Poisson counts? *Stochastic Processes and their Applications*, **98**, 331–339.

Xia, A. and Zhang, F. (2012). On the asymptotics of locally dependent point processes. *Stochastic Processes and their Applications*, **122**, 3033–3065.

Pyramid quantile regression

Yanan Fan[1], Thais Rodrigues[1], Jean-Luc Dortet-Bernadet[2]

[1] University of New South Wales, Sydney, Australia
[2] Université de Strasbourg, France

Keywords: Bayesian inference; Bayesian nonparametrics; quantile regression; pyramid quantiles

Quantile regression models provide a wide picture of the conditional distributions of the response variable by capturing the effect of the covariates at different quantile levels. In most applications, the parametric form of those conditional distributions is unknown and varies across the covariate space, so fitting the given quantile levels simultaneously without relying on parametric assumptions is crucial.

We consider the use of pyramid quantiles (Hjort and Walker, 2009) in the Bayesian treatment of the quantile regression problem. We show how to flexibly construct a the base distribution in the context of quantile regression, and obtain inference at multiple quantile levels simultaneously. We discuss the implementation in both linear and nonlinear quantile regression case. Finally, we demonstrate the advantages of this approach through some simulations.

References

Hjort, N. L. and Walker, S. G. (2009). Quantile pyramids for Bayesian nonparametrics. *Annals of Statistics*, **37**(1), 105–131.

LassoPCA – a robust powerful tool to accurately predict disease risk and estimate cancer mortality

Wenjiang Fu[1], Beverly Fu[2]

[1] University of Houston, USA
[2] Okemos High School, USA

Keywords: dimension reduction; Lasso; PCA; prediction; robust

Accurate prediction of disease risk and estimation of disease mortality are of crucial importance in medicine, public health research and healthcare management. Although numerous computational tools have been developed, many are limited to special study designs and may not perform well in different designs. It is desirable to develop a powerful computational tool for various study designs, such as predicting disease risk of individual patients and estimating disease mortality of a country.

We develop a powerful tool – the LassoPCA by combining two statistical procedures, the Lasso for variable selection (Tibshirani, 1996) and the principal component analysis (PCA) (Jolliffe, 1986). We evaluate the performance of the LassoPCA using data from two studies, a clinical study of kyphosis in children after spinal surgery aiming to predict the kyphosis risk, and a public health study of lung cancer mortality rate in the US to estimate the mortality trend.

We demonstrate that the LassoPCA outperforms the current best model in predicting kyphosis risk and provides accurate cancer trend estimation.

We conclude that the LassoPCA is a powerful tool for accurate estimation and prediction. It applies to various study designs, including regressions and supervised learning, and has broad applications to clinical studies, economic studies, public health research and social studies.

References

Jolliffe, I. T. (1986). *Principal Component Analysis.* Springer, New York.

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

Meta-analysis of medians

Charles Gray, Luke Prendergast

La Trobe University, Australia

Keywords: meta-analysis; median; metafor

In practice, it is common to report medians and interquartile ranges when faced with outliers or skewed data. However, conventional methods for performing meta-analyses of location summary measures of continuous data require means and standard deviations. Thus, studies reporting medians with interquartile ranges are typically excluded from meta-analyses. Given that it is often impractical for practitioners to obtain more detailed information other than the reported summary statistics, and that an analyses of means may not always be appropriate in the first place, it is important that further consideration be given to the analysis of medians. We present an approximate estimator for the variance of the sample median based on limited information. Via this estimator we can carry out a meta-analysis of medians, difference of medians, or ratios of medians. Through simulations we show that our interval estimators of the median has excellent coverage for a very wide range of underlying distributions, and is more suitable than existing methods that involve converting the median and interquartile range to means and standard deviations. Finally, we demonstrate how this also allows practitioners to combine effects presented as means and effects presented as medians in meta-analyses. This can be achieved via a meta-regression with the addition of a dichotomous covariate allowing for a difference in means and medians based effects. Our approach is simple to implement in existing packages such as the `R` package `metafor`.

References

Viechtbauer, W. (2010). Conducting meta-analyses in R with the `metafor` package. *Journal of Statistical Software*, **36**(3), 1–48.

Robust principal components of allocation fund weights

Eka Irianti[1], Nuha Gunawan[2], Melisa Ayuningtyas[3], Kusni Rohani[4]

[1] STIS 52 Computer Division
[2] STIS 53 Social Division
[3] STIS 44 Computer Division
[4] STIS 44 Social Division

Keywords: dependency ratio; allocation fund; principal components; projection pursuit

According to Presidential Regulation number 96 year 2012 allocation fund is weighted among others by Index of Physical Construction Cost (IKK), Gross Domestic Product (GDP), and Human Development Index (HDI).

Gross national income is used to measure standard of living dimension for HDI. However, Indonesia uses adjusted per capita expenditure for HDI together with mean years of schooling and life expectancy at birth. The incorporation of adjusted per capita expenditure as contributors of HDI and GDP is increasing weights of allocation fund. If allocation fund is supposed to be received by regions with lower expenditure, then it may be more accurate to use one of two expenditure variates. Instead of HDI and GDP, to find robust principal components, contributors of HDI are used.

Out of 497 regions, nation capital is separated. Then, 24 richest regions having more than fifty trillion Rupiahs GDP are separated. 23 most developed regions in terms of human development are separated next after GDP separation. Next 34 highest IKK regions are separated that is regions where it is expensive to build physical infrastructure such as bridge, hospital, traditional low rise market, school, office, pedestrian road, play ground park.

The remaining 410 regions are subject to outlier detection and exclusion by median absolute deviation (MAD) for all four variates. There are 31 outliers. Projection pursuit is applied to find largest principal components of 379 regions close to median. Principal Component 1 of 379 regions explains 95.73589 per cent of variance is dominated by IKK. Principal component 2 of 379 regions explains 4.264112 per cent of variance is dominated by life expectancy.

To have more ideas about robustness we find median of IKK and sort all four variates around median of IKK for 350 regions and 150 regions. Outliers detection and exclusion, followed by principal components analysis, are performed using projection pursuit.

Conclusion: Robust Principal Components is of two approaches: robust data set and robust method. Robust data set is approached by outlier detection and exclusion. Robust method is approached by iterative projection pursuit.

Empirical regression quantile process in the risk model

Jana Jurečková[1], Martin Schindler[2], Jan Picek[2]

[1] Charles University, Prague, Czech Republic
[2] Technical University Liberec, Czech Republic

Keywords: averaged regression quantile process; one-step regression quantile process; R-estimator; risk measurement

We describe the empirical averaged regression quantile process as a useful inference tool. Its trajectories are monotone step-functions, and its inversion approximates the distribution function of model errors. The two-step modification of this process involves an R-estimator of the slope components of the regression model. Both processes are asymptotically equivalent and asymptotically independent of the covariates, and hence asymptotically equivalent to the quantile process of the model errors. As such they provide a useful tool in the analysis and measuring of the risk in the situation when the return depends on some exogenous variables. The applications are in the financial analysis, in insurance and social statistics, and also in environment analysis dealing with exposures to toxic chemicals (coming from power plants, road vehicles, from the agriculture), and elsewhere. Possible applications are also in testing hypotheses under nuisance regression, including goodness-of-fit testing, and in estimating various functionals of the risk.

References

Jurečková, J. (2016). Averaged extreme regression quantile. *Extremes*, **19**, 41–49.

Jurečková, J. (2017). Regression quantiles and averaged regression quantile processes. In: *Analytical Methods in Statistics* (J. Antoch et al., Eds.). *Springer Proceedings in Mathematics and Statistics*, **193**, 53–62.

Jurečková, J. and Picek, J. (2014). Averaged regression quantiles. In: Contemporary Developments in Statistical Theory (S. Lahiri et al., Eds.). *Springer Proceedings in Mathematics and Statistics*, **68**, 203–216.

Molak, V. (Ed.) (1997). *Fundamentals of Risk Analysis and Risk Management.* CRC Press.

Rockafellar, R. T. and Uryasev, S. (2013). The fundamental risk quadrangle in risk management, optimization and statistical estimation, *Surveys in Operations Research and Management Science*, **18**, 33–53.

Rockafellar, R. T., Royset, J. O., Miranda, S. I. (2014). Superquantile regression with applications to buffered reliability, uncertainty quantification, and conditional value-at-risk. *European Journal of Operational Research*, **234**, 140–154.

Robust estimate with Partial L_2E for Multivariate data

Fikriye Kabakci[1], Umashanger Thayasivam[2]

[1] Recep Tayyip Erdogan University, Turkey
[2] Rowan University, USA

Keywords: robustness; fast MCD; L_2E; MCD; multivariate data; multivariate M estimators; MVE; partial L_2E; R

The classical multivariate analysis influenced by outlying points in the data. Many methods that are resistant to the outliers have been studied in the literature. When the number of outlying objects become, larger and have similar or different type distribution from the major part of the data, they can be considered as another component from the mixture distribution and in this case parameter estimates can be compared with the proposed robust methods. On the other hand, when the number of outlying objects are smaller, we should be able to detect the major part of the data. Finding estimates which are highly efficient when there is no data contamination and at the same time, high resistance to outliers, i.e. provide lower bias is not always an easy task. In this study, we plan to perform a comparison study of robustness based on minimum integrated square error estimation (L_2E) including the partial L_2E ,with the well-known multivariate covariance and location estimates namely: minimum covariance determinant; fast algorithm for minimum covariance determinant; minimum volume ellipsoid; and multivariate M-estimators. We hope to establish estimators using L_2E, which is efficient for normal data and resistant to outliers when data is contaminated. We plan to perform multiple simulation study with different types of data contamination.

References

Chen, Y., and Gupta, M. R. (2010). *EM demystified: An expectation-maximization tutorial.* Department of Electrical Engineering, University of Washington, Seattle, WA.

Hubert, M., and Debruyne, M. (2010). Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**(1), 36–43.

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

Scott, D. W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics*, **43**(3), 274–285.

Scott, D. W. (2004). Partial mixture estimation and outlier detection in data and regression. *In: Theory and Applications of Recent Robust Methods*, 297-306, Birkhäuser, Basel.

Umashanger, T., Sriram, T. N. and Lee, J. (2012). Simultaneous robust estimation in finite mixture: the continuous case. *Journal of Indian Statistical Association*, **50**(1–2), 277–295.

Performance analysis and robustness for sequential testing of parametric hypotheses under deviations in data models

Alexey Kharin[1], Ton That Tu[1,2]

[1] Deptartment of Probability Theory and Mathematical Statistics, Belarusian State University, Belarus
[2] Faculty of Mathematics, Da Nang University of Education, Vietnam

Keywords: hypotheses; sequential test; deviation; performance; robustness; asymptotic expansion

Sequential analysis (Wald, 1947) is an efficient approach to solve real-life problems of decision making in different areas (Mukhopadhyay and de Silva, 2009). Under deviations in data models ("outliers", functional distortions of probability distributions, incomplete data) statistical procedures often lose the performance optimality (Hampel et al., 1986; Huber and Ronchetti, 2009). Here we develop our recent results to analyze performance of sequential tests (operative characteristic, error probabilities, expected sample sizes) and robustness (Kharin, 2008; Kharin, 2016; Kharin and Ton, 2017) for different models of data, and to construct robust tests by the minimax risk criterion (Kharin and Galinskij, 1999) under deviations from the hypothetical model.

The data models considered are: Markov chains, time series with trends, autoregressive time series. The following deviations from the hypothetical models are discussed: "outliers" in data, ϵ-deviations of probability density functions in the weighted C-metrics, L_1- and L_2-metrics, missing values, and simultaneous distortions for simple and composite hypotheses settings.

Asymptotic expansions for operative characteristic, error probabilities, and conditional expected sample sizes are obtained under the deviated models for general families of modified sequential tests. The robust sequential tests are constructed.

Theoretical results are illustrated numerically by computer experiments on simulated and real data sets.

The research is supported by the Project "Computer Data Analysis and Modeling for Complex Stochastic Systems" via the IMPULSE Program of the Austrian Agency for International Cooperation in Education and Research.

References

Wald, A. (1947). *Sequential Analysis.* Wiley.

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

Huber, P. and Ronchetti, E. (2009). *Robust Statistics.* Wiley.

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

Kharin, A. (2008). Robustness analysis for bayesian sequential testing of composite hypotheses under simultaneous distortions of priors and likelihoods, *Austrian Journal of Statistics*, **40**(1-2), 65–73.

Kharin, A. (2016). Performance and robustness evaluation in sequential hypotheses testing, *Communications in Statistics – Theory and Methods*, **45**(6), 1693–1709.

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

Kharin, A. and Galinskij V. (1999). On minimax robustness of Bayesian statistical prediction, *Probability Theory and Mathematical Statistics*, TEV, 259–266.

Estimation of thresholds in optimal stopping problems via a cross-entropy method

Thilini Dulanjali Kularatne[1], Georgy Sofronov[2]

Macquarie University, Australia

Keywords: optimal stopping, cross-entropy method

A sequential data set is a collection of records which are ordered with respect to time. There are frequent situations where data are sequentially collected over time, and it is necessary to make decisions based on already obtained information while future observations are not known yet. The important question is, therefore, how a decision can be made when the data are still coming and changing the picture with every moment. Analyses of these data are conducted sequentially as new data become available without fixing the sample size in advance. Further sampling of these data are terminated according to a pre-defined stopping rule in order to maximize an expected gain. Problems like this are faced everywhere and everyday in many areas including industrial quality control, econometrics, biomedical signal processing, and analyses of financial systems. We carry out a simulation study to identify an optimal sequential procedure and the value of a game by developing a Cross-Entropy method. We illustrate the effectiveness of the method using simulated data.

References

Chow, Y. S., Robbins, H. and Siegmund, D. (1971). *Great Expectations: The Theory of Optimal Stopping*, Boston: Houghton Mifflin.

Rubinstein, R. Y. and Kroese, D. P. (2004). *The Cross-Entropy Method: a Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*, Springer Science and Business Media.

DeGroot, M. H. (2005). *Optimal Statistical Decisions*, John Wiley and Sons.

Milan, P. and Carmona, T. R. (2007). *Monte Carlo Methods for Financial Instruments with American Exercises.*

Robust estimation and outlier detection on panel data: an application to environmental science

Ahmed Al Sayed[1], Anas Aljarah[1], **Sek Sok Kun**[1], Zaidi Isa[2]

[1] School of Mathematical Sciences, National University of Malaysia, Malaysia
[2] University Science Malaysia, Malaysia

Keywords: outlier detection; robust estimators; panel data regression

The existence of outliers in the dataset might have significant effects towards the estimated coefficients and it may lead to misestimating on the relationship between variables. Nevertheless, some of the researchers have applied the common estimator ordinary less square (OLS) regardless the assumption of the error terms must be normally distributed which could be violated when outliers exist. On the other hand, outliers might be actual values and it is not recommended to exclude them from the analysis as they may contain significant information towards the detection on the relationship between the variables. To tackle that problem, the existing robust estimators considering the outliers in analysis such as; (M, Median, S and MM-estimator). The main objective of this study is to detect the preferable estimator to model the relationship between CO2 emissions, energy consumption and gross domestic product by considering the influence of different types of outliers in panel data. The dataset takes the annual period over 1960-2008 for two groups of countries; developed and developing countries. The main significant finding of this study is that the M-estimator is the preferable robust estimator which could represent the dataset by considering different types of outliers in the analysis. Applying the Mahalanobis distances to detect the outliers, the results indicates that the dataset contains three types of outliers: leverage point, block concentrated vertical outliers and block concentrated leverage points. In conclusion, the robust M-estimator is robust estimator handling data with high efficiency and high breakdown point with the existence of different types of outliers.

References

Baltagi, B. H. (2005) *Econometric Analysis of Panel Data*, John Wiley & Sons.

Rousseeuw, P. J., and A. M. Leroy. (2003). *Robust Regression and Outlier Detection*. New York: Wiley.

Modelling and estimating change-points in time series processes

Lijing Ma, Georgy Sofronov

Macquarie University, Australia

Keywords: change-point estimate; cross-entropy method; time series process

Many macroeconomic variables are often modelled as highly persistent non-stationary integrated processes. For example, the variable such as inflation may be subject to changes in government policy which may cause structural breaks in the data. In this paper we develop a methodology to estimate change-points in time series. We compare the statistical performance of a number of computational methods for estimating unknown parameters of first order autoregressive data with structural breaks. Specifically, we consider the Cross Entropy method for modelling break points using Monte Carlo simulation to estimate change-points as well as parameters of the process on each segment. Numerical experiments illustrate the robustness of this approach. We obtain estimates for the locations of change-points in artificially generated sequences and compare the accuracy of these estimates to those obtained with other methods. We also provide examples of analysis of real data to illustrate the usefulness of the proposed method.

References

Cho, H. and Fryzlewicz, P. (2012). Multiscale and multilevel technique for consistent segmentation of nonstationary time series, *Statistica Sinica*, **22**, 207–229.

Davis, R. A., Lee, T. C. and Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, **101**, 223–239.

Rubinstein, R. Y. and Kroese, D. P. (2004). *The Cross-Entropy Method: a Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*, Springer Science and Business Media.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, **18**, 1–22.

A comparison between predictive and estimative approach for exponential families using various type of estimators

Avijit Maji[1], Ayanendranath Basu[2]

[1] Reserve Bank of India and Indian Statistical Institute, India
[2] Indian Statistical Institute, India

Keywords: bootstrap; minimum density power divergence estimator; predictive distribution

The basic problem considered is the one where the observed data are $x_1, \cdots, x_n$, all realizations of a random variable X and some probability statement about a future random variable Z from the same distribution is desired. A predictive distribution for Z is used for this purpose. Harris (1989) suggested a predictive distribution based on bootstrapping using the maximum likelihood estimator of an unknown parameter. Basu and Harris (1994) introduced robust estimative and bootstrap predictive distributions by using the minimum Hellinger distance estimator of the unknown parameter. The present paper considers robust predictive distributions based on the minimum density power divergence (DPD) estimator developed by Basu et al.(1998) as compared to Hellinger distance estimator using the Kullback-Leibler divergence as a measure of the predictive fit and brings out several advantages of the DPD based method. Monte Carlo simulations suggest that DPD bootstrap predictive distributions are attractive robust alternatives to the usual predictive distributions, and have distinct advantage in certain scenarios.

References

Basu, A. and Harris, I. R. (1994). Robust predictive distributions for exponential families. *Biometrika*, **81**, 790–794.

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**, 549–559.

Harris, I. R. (1989). Predictive fit for natural exponential families. *Biometrika*, **76**, 675–684.

Robust Outlier Approach to Detect Faulty Digital Circuit

Kenan Matawie[1], Nethal Jajo[2]

[1] School of Computing Engineering and Mathematics, Western Sydney University, Australia
[2] University of Sydney, Australia

Keywords: outliers; eigenvalue; influence matrix; IDDQ test; digital circuit

Identifying faulty part(s) of a digital circuit is always an interesting, active, complicated and challenging research. One of the recent and most popular method used to detect and improve the quality of the fabricated circuits is the IDDQ testing. This test offers high fault coverage with a small set of test vectors (McEuen, 1991). Usually the approach is to separate the defective chips from the faulty-free ones using a threshold IDDQ value, however, there is a possibility that this test can pass components while they are actually faulty, or fail the test while they are faulty-free. This adds more complication to the problem of detecting the real faulty components. Therefore, a robust, more advanced, reliable and effective IDDQ test criteria is needed.

In this paper we will use and extend an approach based on robust outlier technique proposed in Jajo and Matawie (2009) to identify and detect the defective (outliers) and faulty-free components/chips. This method is demonstrated and validated using real IDQQ data from IBM. The results are also compared with other approaches.

References

Jajo, N.K. and Matawie, K.M. (2009). Eigenvalues application in robust outlier detection. *International Workshop on Statistical Modelling*, pp. 182–186.

McEuen, S.D. (1991). IDDQ benefits. *IEEE VLSI test Symposium*, pp. 34–39.

Influential observations in optimal dynamic treatment regimes

Nur Anisah Mohamed

Institute of Mathematical Sciences, Faculty of Science, University of Malaya, Malaysia

Keywords: dynamic treatment regimes; generalised estimating equations; influential observations

Dynamic treatment regimes is a personalized treatment strategy that assign treatment based on the individual needs. Optimal dynamic treatment regime is an approach for estimating optimal decision rules which carried out over a series of time which produce the highest mean response at the end of the study. Generalised Estimating Equations for Myopic regret-regression (GEE-MRr) is a semiparametric approach that combined the generalized estimating equations (GEE) method with the regret functions from Murphy (2003) and Henderson et al. (2009) to estimate the optimal dynamic treatment regime for correlated data. In this paper, we introduce a one-case deletion diagnostic for GEE-MRr approach to identify the influence of observations on the estimated parameters. We will use a Cook's distance measure to assess the influence of deleting an observation and introduce a feasible one-step approximation for GEE-MRr approach. We illustrate these methods using simulation and application to anticoagulation data.

References

Murphy, S. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B*, **65**(2), 331-355.

Henderson, R., Ansell, P. and Alshibani, D. (2010). Regret-regression for optimal dynamic treatment regimes. *Biometrics*, **66**(4), 1192-1201.

Robustness and robust optimization

Stephan Morgenthaler

Ecole Polytechnique Fédérale de Lausanne; Switzerland

Keywords: robust optimization; finance

The Markowitz asset allocation model for multivariate normal distributions is equivalent to minimizing the value at risk at level alpha. The minimization is over the investment weights, that is, the fraction of the invested sum given to the different assets. This optimization problem, can be generalized by considering uncertainty in the distribution and the parameters of the asset returns. We will discuss this step in detail and compare it to other instances of robustification in statistical estimation.

Robust multi-view graph embedding

Akifumi Okuno[1,2], Hidetoshi Shimodaira[1,2]

[1] Graduate School of Informatics, Kyoto University, Japan
[2] RIKEN Center for Advanced Intelligence Project, Japan

Keywords: correlation analysis with many-to-many association; multi-view graph embedding

Various types of data, such as texts, images, and sounds, become easily obtainable these days. Different types of data are referred to as "views", "modalities" or "domains", and approaches to integrate the multi-view data $\{\boldsymbol{x}_i^d\}_{i=1}^{n_d} \subset \mathbb{R}^{p_d}$ $(d = 1, 2, \ldots, D)$ into an unified representation have attracted much attention.

In practice, multi-view data have a many-to-many relationship but not one-to-one. An example is a image set with their multiple tags. CCA, which is one of the best-known approaches for integrating 2 different types of vectors, cannot utilize the complex data structure. Therefore, Huang et al. (2012) extended CCA as Cross-view Graph Embedding (CvGE), that utilizes a strength of association w_{ij} between $\boldsymbol{x}_i^1$ and $\boldsymbol{x}_j^2$. CvGE finds low-dimensional representation $\boldsymbol{y}_i^d := (\boldsymbol{A}^d)^\top \boldsymbol{x}_i^d \in \mathbb{R}^K$ by minimizing

$$\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} w_{ij} \|(\boldsymbol{A}^1)^\top \boldsymbol{x}_i^1 - (\boldsymbol{A}^2)^\top \boldsymbol{x}_j^2\|_2^2, \tag{1}$$

with respect to $\boldsymbol{A}^d \in \mathbb{R}^{p_d \times K}$. A quadratic constraint is imposed on $\boldsymbol{A}^1, \boldsymbol{A}^2$ to prevent Eq.(1) from being degenerated.

The objective function (1) is similar to graph embedding (Yan et al., 2007). In graph embedding methods, the strength of association w_{ij} between data vectors is often determined by their similarity. In multi-view setting, however, the associations should be specified by users because their similarity cannot be measured across different domains; then incorrect associations, or outliers, may be included. These incorrect associations may influence the embedding result. Our aim is to alleviate their negative effect. Therefore, we propose a novel iterative method for down-weighting such an incorrect association. Our method multiplies the strength of association by a weight function depending on the current embedding result,

$$w_{ij}^{(t)} \leftarrow w_{ij} \cdot \exp(-\gamma \|(\hat{\boldsymbol{A}}_{(t)}^1)^\top \boldsymbol{x}_i^1 - (\hat{\boldsymbol{A}}_{(t)}^2)^\top \boldsymbol{x}_j^2\|_2^2), \tag{2}$$

where $\gamma > 0$ is a tuning parameter. By applying CvGE to the weight (2), we obtain updated transformation matrices $(\hat{\boldsymbol{A}}_{(t+1)}^1, \hat{\boldsymbol{A}}_{(t+1)}^2)$. Repeating this update derives a robust embedding result against the incorrect associations. We show that our method is an MM algorithm for minimizing the loss function analogous to the γ-divergence (Fujisawa and Eguchi, 2008). It ensures the convergence of the iterative update. Moreover, we generalize these results to D-view setting.

References

Huang, Z., Shan, S., Zhang, H., Lao, S., and Chen, X. (2012). Cross-view graph embedding. In *Asian Conference on Computer Vision* (pp. 770-781). Springer Berlin Heidelberg.

Yan, S., Xu, D., Zhang, B., Zhang, H. J., Yang, Q., and Lin, S. (2007). Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, **29**(1).

Fujisawa, H., and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, **99**(9), 2053-2081.

Robust portfolio optimization under multiperiod mean-standard deviation criterion

Spiridon Penev[1], Pavel Shevchenko[2], Wei Wu[1]

[1] University of New South Wales, Sydney, Australia
[2] Macquarie University, Australia

Keywords: model risk; robust portfolio optimization; Kullback-Leibler divergence

We quantify model risk for an optimized portfolio of stocks in which as a selection criterion, a multi-period mean-standard deviation criterion is used. This is a specific dynamic portfolio allocation problem in which a robustification is needed due to uncertainty in the underlying distribution of the asset returns in the portfolio. The uncertainty is measured by Kullback-Leibler divergence to certain nominal model for the joint distribution of the assets. The optimal robust strategy is obtained in a semi-analytical form. This semi-analytic solution contains some integrals that can be approximated with a Monte Carlo approach thus the solution can be obtained easily.

We present several numerical results which compare the performance of the optimal robust portfolio with the nominal model. The radius of the ball of uncertainty plays important role in the performance of the robust solution hence we also deal with estimating this radius.

References

Bannister, H., Goldys, B., Penev, S. and Wu, W. (2016). Multiperiod mean-standard-deviation time consistent portfolio selection. *Automatica,* **73**, 15–26.

Glasserman, P. and Xu, X. (2014). Robust portfolio control with stochastic factor dynamics. *Operations Research*, **61**(4), 874–893.

Identifying boundaries of domains in spatial binary data

Nishanthi Raveendran, Georgy Sofronov

Department of Statistics, Macquarie University, Australia

Keywords: spatial clustering; change-point detection; binary segmentation; binary data

Spatial clustering is an important component of spatial data analysis. The aim is to identify the boundaries of domains and their number. It is commonly used in disease surveillance, spatial epidemiology, population genetics, landscape ecology, crime analysis and many other fields. We focus on identifying homogeneous sub-regions in an ecology dataset. We use binary data indicating the presence or absence of a certain plant species which are observed over a two-dimensional lattice. The problem of finding regional homogeneous domains is known as segmentation, partitioning or clustering. To solve this problem we propose to use change-point methodology. We develop new methods based on a binary segmentation algorithm which is a well-known multiple change-point detection method. We use both simulated and real datasets to illustrate the usefulness of these methods.

References

Chen, J. and Gupta, A.K. (2011). *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer Science and Business Media.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley and Sons.

Eckley, I.A., Fearnhead, P. and Killick, R. (2011). Analysis of change-point models. *In: Bayesian Time Series Models*, Chapter 10, pp. 205-224, Cambridge University Press.

López, I., Gámez, M., Garay, J., Standovár, T. and Varga, Z. (2010). Application of change-point problem to the detection of plant patches. *Acta Biotheoretica*, **58**(1), 51-63.

Yang, T.Y and Swartz, T.B. (2005). Applications of binary segmentation to the estimation of quantal response curves and spatial intensity. *Biometrical Journal*, **47**(4), 489-501.

Multiscale Bayesian state space model for Granger causality analysis, with application to intracranial electroencephalogram data

Olivier Renaud[1], Sezen Cekic[1], Didier Grandjean[2]

[1] Group of Methodology and Data Analysis, Department of Psychology, University of Geneva, Switzerland
[2] Neuropsychology of Emotion and Affective Dynamics Laboratory, Department of Psychology, University of Geneva, Switzerland

Keywords: à trous Haar wavelets; multiple trials; neuroscience data; nonstationarity; time-frequency; variational methods

This talk concerns the modelling of time-varying and frequency-specific relationships between two signals, with a focus on intracerebral signals measuring neural activities. Many researchers in neuroscience would like to assess what is called a frequency Granger causality that may vary in time to evaluate the functional connections between two brain regions during a task. We propose the use of an adaptive Kalman filter type of estimator of a linear Gaussian vector autoregressive model with coefficients evolving over time. The estimation procedure is achieved through variational Bayesian approximation and can be extended for multiple trials. This Bayesian State Space (BSS) model provides a dynamical Granger-causality statistic that is quite natural. The Bayesian nature of the model provides a criterion for model order selection and allows us to include prior knowledge in the model. We propose to extend the BSS model to include the *à trous* Haar decomposition. This wavelet-based forecasting method, based on a multiple resolution decomposition of the signal using the redundant *à trous* wavelet transform, captures short- and long-range dependencies between signals and is further used to derive the desired dynamical and frequency-specific Granger-causality statistic.

Often, brain signal data are recorded during an experimental situation where stimuli are presented at fix time and are expected to induce a subject reaction. The causal links between recorded signals (e.g. from different part of the brain) may therefore vary in time and be frequency specific. I will present an application of this methodology to real intracranial electroencephalogram data recorded in the regions of amygdala and medial orbitofrontal cortex during an experimental task of emotional auditory stimuli recognition. The analysis will show the complex frequency based cross-talk between these two regions.

References

Cassidy, M.J. and Penny, W. (2002). Bayesian nonstationary autoregressive models for biomedical signal analysis. *IEEE Transactions on Biomedical Engineering*, **49**(10), 1142–1152.

Cekic, S. , Grandjean, D., and Renaud, O. (submitted). Multiscale Bayesian state space model for Granger causality analysis of brain signal.

Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, **77**(378), 304–313.

Granger, C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **37**(3), 424–438.

Renaud, O., Starck, J.-L., and Murtagh, F. (2003). Prediction based on a multiscale decomposition. *International Journal of Wavelets, Multiresolution and Information Processing*, **1**(2), 217–232.

Statistical modeling for unemployment rates using nonparametric geographically weighted regression with truncated spline approach

Sifriyani[1,2], Haryatmi[1], Budiantara[3], Gunardi[1]

[1] Gadjah Mada University, Indonesia
[2] Mulawarman University, Indonesia
[3] Institut Teknologi Sepuluh Nopember, Indonesia

Keywords: East Java; nonparametric geographically weighted regression; spatial; truncated spline approach; unemployed rate

In this work we developed nonparametric geographically weighted regression models using truncated spline approach. This method is applied to the rate of unemployment data in East Java, as there are problems of high population growth impacting the high unemployment rate. This method was chosen due to the fact that the data used had an unknown regression curve and is influenced by the geographical element. Truncated spline approach is used because this approach has the function of a flexible mathematics and models that tend to look for the shape of the curve regression estimate objectively, without being influenced by the subjective factor of researchers. We first analysis was to find model estimators using truncated spline approach. Furthermore the selection of optimum knot points was done by selecting the minimum value of the Generalized Cross Validation (GCV). This study has been successfully mapped and modeled the unemployment rate with variables that affect it. The results using the nonparametric geographically weighted regression methods using truncated spline approach has the best statistical model and is satisfactory with 98.95% coefficient of determination and the MSE is equal to 0.0047.

References

Demsar, U., Fotheringham, A. S.,and Charlton, M. (2008). Exploring the spatio-temporal dynamics of geographical processes with geographically weighted regression and geovisual analytics. *Inference*, **7**(3-4), 181–197.

Golan, A. and Perloff, J., M. (2004). Superior forecasts of the U.S. unemployment rate using a nonparametric method. *The Review of Economics and Statistics*, **86**(1), 433–438.

Huang, B., Wu, B. and Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science*. **24**(3), 383–401.

Mei, C. L., Wang, N. and Zhang, W. X. (2006). Testing the importance of the explanatory variables in a mixed geographically weighted regression model. *Journal Environment and Planning A*, **38**(3), 587–598.

Miskolczi, M., Langhamrova, J., and Fiala, T. (2011). Dependency between gross domestic product and unemployment in the Czech Republic. *Research Journal of Economics, Business and ICT*, **4**, 47–51.

Nakaya, T., Fotheringham, A. S., Brunsdon, C. and Charltron, M. (2005). Geographically weighted Poisson regression for disease association mapping. *Statistics In Medicine*, **24**(17), 2695–2717.

Wrenn, D.H. and Sam, A.G. 2014. Geographically and temporally weighted likelihood regression: exploring the spatiotemporal determinants of land use change. *Regional Science and Urban Economics*, **44**, 60–74.

Yang, L. (2008). Confidence band for additive regression model. *Journal of Data Science*, **6**, 207–217.

Robust modified invisible fence

Connor Smith, Samuel Mueller

University of Sydney, Australia

Keywords: robust model selection; variable selection in regression; subtractive lack-of-fit measure

The Invisible Fence (Jiang et al 2011) is a promising new model selection method that is suitable for complex problems. Having already been used for inference for Mixed Models in particular, this method has yet to be implemented in a robust setting for regression type models including linear and generalized linear regression. Using a subtractive lack-of-fit measure, in this talk, we consider different versions of the Invisible Fence which utilize a combination of robust regression coefficient estimates, robust estimates of scale and robust quasi-deviances (Cantoni and Ronchetti 2001). The suggested methods will be applied to both simulated and real examples with comparisons about both time and effectiveness made to selections through alternative selection procedures.

References

Jiang, J., Nguyen, T., and Rao, J. S. (2011). Invisible fence methods and the identification of differentially expressed gene sets. *Statistics and its Interface*, **4**(3), 403-415.

Cantoni, E., and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, **96**(455), 1022-1030.

Assessing wool fibre diameter distributions

Robert Staudte, Chandima Arachchige, Luke Prendergast

La Trobe University, Australia

Keywords: influence function; quantile density; relative asymptotic variance

Thousands of raw wool samples are routinely classified using characteristics of the distribution of fibre diameters, measured in microns and obtained by various technologies. Of main interest to wool assessors are the mean and coefficient of variation of fibre diameter data, as well as the percentages of relatively large fibres (rough edge) that can cause skin irritation in finished woolen goods. The accuracy of these estimators is usually not questioned because the sample size is large (in the thousands), and reproducibility of results is assumed. The statistical methodology has not much changed for the last 60 years.

The statistics based on sample moments can be affected by the presence of a few outliers, so here we investigate the extent to which statistics based on quantile estimators can provide adequate information for wool assessment while remaining resistant to anomolous fibres. In particular, we investigate the median (in lieu of the mean) and the interquartile range, divided by the median, (in lieu of the coefficient of variation).

In addition to comparing the influence functions of the competing estimators and their asymptotic biases and variances, we compare interval estimators using simulation studies. The results are illustrated on examples of Alpaca and Merino data. The important question of how to convince wool testing authorities to change their wool assessment methods is also discussed.

References

Groeneveld, R. A. (2011). Influence functions for the coefficient of variation, its inverse, and CV comparisons. *Communications in Statistics – Theory and Methods*, **40**(23), 4139–4150.

Prendergast, L. A. and Staudte, R. G. (2016a). Exploiting the quantile optimality ratio in finding confidence intervals for quantiles. *Stat*, **5**, 70–81.

Prendergast, L. A. and Staudte, R. G. (2016b). Quantile versions of the Lorenz curve. *Electronic Journal of Statistics*, **10**(2), 1896–1926.

Staudte, R. G. (2017). The shapes of things to come: probability density quantiles. *Statistics: a Journal of Theoretical and Applied Statistics*, 1–19. DOI:10.1080/02331888.2016.1277225.

Response modelling approach to robust parameter design methodology using supersaturated designs

Stelios Georgiou

Royal Melbourne Institute of Technology, Australia

Keywords: experimental designs; robust parameter designs; LASO; simulations

In recent years, both robust parameter designs (RPDs) and supersaturated designs (SSDs) have attracted a great deal of attention. In this talk, a common sense of both fields is considered. More precisely, we propose a construction of an effective SSD along with an analysis method, in order to deal with the significant problem of the robust parameter design methodology (RPDM). Combining iterative SIS variable selection and a penalized method, namely SCAD, we perform the analysis of the SSDs developed in the present work. The proposed methodology applied in different models so as to show its effectiveness in many different scenarios, assuming both first and second-order models in a sense of a response surface design. Two illustrative examples as well as numerous numerical experiments are conducted for plenty cases. The results imply that the proposed method is highly effective for identifying the active effects of main factors, two-factor interactions, three-factor interactions as well as the pure quadratic ones, under the assumption of effect sparsity.

References

Box, G.E.P., Meyer, R.D. (1986). An analysis for unreplicated fractional factorials. *Technometrics*, **28**, 11–18.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

Fan, J., Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B*, **70**(5), 849–911.

Georgiou, S.D. (2014). Supersaturated designs: a review of their construction and analysis. *Journal of Statistical Planning and Inference*, **144**, 92–109.

Kunert J., Auer C., Erdbrugge M., Ewers R. (2007). An experiment to compare Taguchi's product array and the combined array. *Journal of Quality Technology*, **39**(1), 17–34.

Matsuura, S., Suzuki, H., Iida, T., Kure H., Mori, H., (2011). Robust parameter design using a supersaturated design for a response surface model. *Quality and Reliability Engineering International*, **27**, 541–554.

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

Robust screening by edge designs

Stella Stylianou

Royal Melbourne Institute of Technology, Australia

Keywords: experimental designs; robust designs; screening

When a large number of variables (factors) is examined in experimental situation it is often anticipated that only few of these will be important. Usually it is not known which of the variables will be the important ones, so it is not known which columns of the experimental design will be of further interest. Many designs have been proposed to be used for screening experiments and to identify the relevant variables. Recently, a new class of experimental designs, called edge designs, was introduced. These designs allow a model-independent estimate of the set of relevant variables, thus providing more robustness than traditional designs. A construction for edge designs having $n-1$ edges, $n = 0 \pmod 4$ from skew conference matrices of order n.

We describe an algorithm to find four $(1, -1)$ vectors (A, B, C, D), which have zero periodic autocorrelation function. From these vectors we constructed the appropriate circulant matrices and use them to build new inequivalent skew Conference matrices. We embed the new conference matrices in the suggested structure and we generate new edge designs. An illustrative simulated example using the analysis of edge design is provided.

References

Elster C. and Neumaier A. (1995). Screening by conference designs. *Biometrika*, **82**(3), 589–602.

Georgiou S., Koukouvinos C. and Stylianou S. (2004). On good matrices, skew Hadamard matrices and optimal designs. *Computational Statistics and Data Analysis*, **41**(1), 171–184.

Outlier detection in a complex linear mixed model

Emi Tanaka

School of Mathematics and Statistics, University of Sydney, Australia

Keywords: outlier; linear mixed model; field trial; residual analysis

Outlier detection is an important preliminary step in data analysis. An outlier in a complex data, such as those that are analysed by linear mixed models, offers additional challenges since observations can be flagged as outliers in different contexts. For example, in a repeated measures data, there may be an outlying subject and consequently all observations from the outlying subject are outliers.

Crop field trials are routinely conducted each year across multiple locations. These so-called multi-environmental trials (MET) is commonly analysed using linear mixed models which involve many random effects as well as terms to account for spatial trends within a field trial. An outlier detection for such a complex linear mixed model is often conducted by some form of residual analysis (Haslett and Haslett, 2007; Guzmede et al., 2010). Alternative methods, which is not as commonly used in the MET analysis, are local influence by measuring the effect of perturbing subset of data or effects (Lesaffre and Verbeke, 1998), and some form of deletion diagnostic such as the (generalised) Cook's distance (Enea and Plaia, 2017). In this talk, I will compare various outlier detection for the analysis of a MET data.

Haslett, J. and Haslett, S. (2007). The three basic types of residuals for a linear model. *International Statistical Review*, **75**(1), 1–24

Gumedze, F., Welham, S., Gogel, B., and Thompson, R. (2010) A variance shift model for detection of outliers in the linear mixed model. *Computational Statistics and Data Analysis*, **54**, 2128–2144

Enea, M. and Plaia, A. (2017) A gradient-based deletion diagnostic measure for generalized linear mixed models. *Communications in Statistics - Theory and Methods*, **45**(4), 1972–1982

Lesaffre, E. and Verbeke, G. (1998) Local influence in linear mixed models. *Biometrics*, **54**(2), 570–582

Fast and approximate exhaustive variable selection for GLMs with APES

Kevin Wang[1], Samuel Mueller[1], Garth Tarr[2], Jean Yang[1]

[1] University of Sydney, Australia
[2] The University of Newcastle, Australia

Keywords: variable selection; all subset selection; fast exhaustive algorithms; generalised linear models

Traditionally, obtaining maximum likelihood estimates for generalised linear models (GLMs) is time consuming and remains as the major obstacle for performing all subsets variable selection. The non-linearity of GLMs meant exhaustive exploration of model spaces, even for a moderately large number of covariates, remains a formidable challenge for modern computing capabilities. For linear models, on the other hand, fast algorithms exist, most notably the leaps and bound algorithm.

In this talk, we will present APES (APproximated Exhaustive Search) a new method that approximates all subset selection for a given GLM by reformulating the problem as a linear model instead. The method works by transforming the problem into a linear model and learning from observational weights in a correct/saturated generalised linear regression model. Robust regression is used in two ways: (i) to limit the influence of outliers that are introduced in the linear regression approximation space that have extreme observation weights inherited from their large standardized Pearson residual in the GLM space, and (ii) to obtain more robust fitted models in either the original GLM or the approximated linear regression space. The simulation study results are promising, APES is competitive with exhaustive searches for selection of the true data generating model.

Cellwise robust regularized discriminant analysis

Ines Wilms[1], Stéphanie Aerts[2]

[1] Leuven Statistics Research Centre (LStat), KU Leuven, Belgium
[2] HEC-Liège, University of Liège, Belgium

Keywords: cellwise robust precision matrix; classification; discriminant analysis; penalized estimation

Quadratic and Linear Discriminant Analysis (QDA/LDA) are the most often applied classification rules under normality. In QDA, a separate covariance matrix is estimated for each group. If there are more variables than observations in the groups, the usual estimates are singular and cannot be used anymore. Assuming homoscedasticity, as in LDA, reduces the number of parameters to estimate. This rather strong assumption is however rarely verified in practice. Regularized discriminant techniques that are computable in high-dimension and cover the path between the two extremes QDA and LDA have been proposed in the literature. However, these procedures rely on sample covariance matrices. As such, they become inappropriate in presence of cellwise outliers, a type of outliers that is very likely to occur in high-dimensional datasets. In this talk, we propose cellwise robust counterparts of these regularized discriminant techniques by inserting cellwise robust covariance matrices. Our methodology results in a family of discriminant methods that (i) are robust against outlying cells, (ii) provide, as a by-product, a way to detect outliers, (iii) cover the path between LDA and QDA, and (iv) are computable in high-dimension. The good performance of the new methods is illustrated through simulated and real data examples.

References

Danaher, P., Wang, P. and Witten, D. (2014). The joint graphical Lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B*, **76**(2), 373–397.

Öllerer, V. and Croux, C. (2015). Robust high-dimensional precision matrix estimation. *In: Nordhausen, K., Taskinen, S. (Eds.), Modern Nonparametric, Robust and Multivariate Methods: Festschrift in Honour of Hannu Oja,* Springer, 325–350.

Fast quantile process regression

Yonggang Yao

SAS Institute Inc., USA

Keywords: conditional distribution estimation; linear programming; binary tree

Quantile process regression (QPR) is a powerful methodology for estimating the entire distribution of a response variable conditional on a set of explanatory variables. QPR often surpasses other regression methodologies in its estimation richness and veracity, especially for analysis of heteroscedastic data. However, a major difficulty for QPR in practice is in its computation cost. To approximate a QPR on a grid of q- quantile levels, current QPR algorithms separately fit q- single-level quantile regression models. For large q-, the QPR computation is often so substantial that it overwhelms the advantages of QPR's estimation richness and veracity.

This paper proposes a fast quantile process regression (FQPR) algorithm to tame the QPR computational difficulty. FQPR uses a divide-and-conquer strategy to fit q- quantile regression models in approximately the amount of time that is required to separately fit $\log_2(q)$ single-level quantile regression models. For example, $q = 1024$ has $\log_2(q) = 10$, so that FQPR can be 100 times faster than current QPR algorithms. Such a vast improvement in computational efficiency may encourage for more applications of QPR in practice.

References

Chen, C. (2007). A finite smoothing algorithm for quantile regression. *Journal of Computational and Graphical Statistics*, **16**, 136–164.

Gass, S. and Saaty, T. (1955). The computational algorithm for the parametric objective function. *Naval Research Logistics Quarterly*, **2**, 39–45.

Gutenbrunner, C. and Jureckova, J. (1992). Regression rank scores and regression quantiles. *Annals of Statistics*, **20**, 305–330.

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

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

Koenker, R. and D'Orey, V. (1987). Algorithm AS 229: computing regression quantiles. *Journal of the Royal Statistical Society, Series C*, **36**, 383–393.

Koenker, R. and D'Orey, V. (1994). Remark AS R92: a remark on algorithm AS 229: computing dual regression quantiles and regression rank scores. *Journal of the Royal Statistical Society, Series C*, **43**, 410–414.

Koenker, R. and Ng, P. (2005). A Frisch-Newton algorithm for sparse quantile regression. *Acta Mathematicae Applicatae Sinica*, **21**, 225–236.

Portnoy, S. and Koenker, R., (1997). The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, **12**, 279–300.

Yang, J., Meng, X., and Mahoney, M. W. (2014). Quantile regression for large-scale applications. *Society for Industrial and Applied Mathematics (SIAM) Journal on Scientific Computing*, **36**, s78–s110.

On estimating the variances of the satellite remote sensing data

Bohai Zhang

NIASRA, University of Wollongong, Australia

Keywords: atmospheric carbon dioxide; errors in variables; OCO-2; semivariogram

The column-averaged carbon dioxide (CO_2) concentrations collected by the Orbiting Carbon Observatory-2 (OCO-2) satellite mission provide an unprecedented opportunity to characterize and understand the spatio-temporal variability of CO_2 in Earth's atmosphere. Because of the additional challenges of making precise and accurate measurements from space, it is essential to validate the OCO-2 data with highly precise and accurate ground-based measurements. The errors-in-variables (EIV) regression model provides a straightforward way for modeling the linear relationship between the less-precise OCO-2 datasets and the more-precise ground-based datasets. To obtain consistent estimates of the regression coefficients in the EIV model, it is critical to estimate the error variance of each datum in the regression consistently. We will discuss how to estimate the variances of the median-based OCO-2 datum. A spatial-process model is fitted, where the variance-covariance parameters are estimated from robust semivariograms. Numerical results for analyzing the OCO-2 target-mode observations corresponding to the Lamont ground station with the orbit number 3590 illustrate our procedure. This research is based on joint work with Noel Cressie (University of Wollongong) and Debra Wunch (University of Toronto).

Robust estimation of treatment effects in a latent-variable framework

Mikhail Zhelonkin

Erasmus University Rotterdam, Netherlands

Keywords: observational studies; sample selection model; treatment effect

The policy evaluation is one of the central problems in modern economics. Unfortunately, it is usually impossible to perform a randomized experiments in order to evaluate the treatment effects. Hence, the data from observational studies has to be used. In this case the sample is typically non-random and one has either to correct for selectivity or to impose (conditional) independence assumption. Since this assumption is often unrealistic, the structural latent variable model is used. The parametric estimators (although, they are straightforward to compute and to interpret) have been criticized for sensitivity to the departures from the distributional assumptions. The alternative semi- and nonparametric estimators have complex identification and are limited to estimation of certain parameter(s) of interest but do not allow for the general evaluation and interpretation of the model. In this work we employ the latent-variable framework (Heckman, Tobias, and Vytlacil 2003). We study the robustness properties of the estimators of four principal parameters: average treatment effect, average treatment effect on the treated, local average treatment effect and marginal treatment effect (Heckman, Tobias, and Vytlacil 2001), and propose the robust alternatives.

References

Heckman, J. J., Tobias, J. L., Vytlacil, E. (2001). Four parameters of interest in the evaluation of social programs. *Southern Economic Journal*, **68**(2), 210–223.

Heckman, J. J., Tobias, J. L., Vytlacil, E. (2003). Simple estimators for treatment parameters in a latent-variable framework. *The Review of Economics and Statistics*, **85**(3), 748–755.

Part IV

Robust Survey Sampling Workshop

Exploring the robustness of log-gamma vs. normal for random effect distributions: the example of small area estimation

Jarod Lee, **James Brown**

University of Technology Sydney, Australia

When making small area estimates, it is common to utilise a GLM model with Gaussian random effects. In this work, we contrast this approach with a Poisson model combined with log-gamma random effects. The latter has advantages with respect to the form of the likelihood function and we also explore whether it has advantages regarding robustness to outliers and skewness with respect to the random effect distribution.

Estimating population totals using imperfect administrative data and a survey subject to non-ignorable non-response

James Chipperfield

Australian Bureau of Statistics, Australia

Methods for estimating population totals often assume that survey non-response can be ignored and that imperfections in register data (e.g. contain incorrect information or miss/double-count population units), used as benchmarks for survey weights, can be ignored. This paper explores a way to relax both of these assumptions at the same time. One application is estimating the Australian Resident Population using the Post Enumeration Survey and the Australian Census of Population and Housing.

Imputation using robust regression

John Preston, **James Chipperfield**

Australian Bureau of Statistics, Australia

Many national statistical offices embody the ideal of an "industrialisation" of the production of official statistics, and hence need to implement standardised and automated imputation methods into their generalised systems. Continuous variables in business surveys are usually imputed using deterministic regression imputation methods, such as mean imputation, historical ratio imputation and auxiliary ratio imputation. This deterministic regression imputation method is predicated by the well-known ordinary least squares linear regression model, which can be severely affected by outliers, leading to estimates of regression parameters that do not accurately reflect the true underlying relationship between the variable of interest and the auxiliary variables. While these generalised systems will generally allow the functionality to remove outliers from the calculation of the imputed values, these influential units are often ignored in practice because the manual identification of outliers can be an inconsistent and inefficient process. Robust deterministic regression imputation methods which can be implemented as a standardised and automated process have been proposed. A simulation study using continuous variables from a typical business survey found that these robust imputation methods performs well compared with other alternatives.

An empirical likelihood based estimator for respondent driven sampled data

Sanjay Chaudhuri[1], Mark Handcock[2]

[1] National University of Singapore, Singapore
[2] The University of California – Los Angeles, USA

We discuss an empirical likelihood based estimator of population means applicable to data obtained from a respondent driven sampling procedure. Our estimator directly uses the second order weights of selection and constructs a composite empirical likelihood to estimate the parameter of interest. This estimate is asymptotically unbiased and normally distributed. Analytic expression of the asymptotic standard errors can be obtained which can also be estimated from the data using a sandwich estimator. Using real life social network data, we show that, our estimator produces confidence intervals with far better coverages than the existing estimators.

One-sided Winsorization in sample surveys

Robert Clark[1], Phil Kokic[2]

[1] University of Wollongong, Australia
[2] Australian National University, Australia

Sample surveys have the distinctive feature of "representative outliers". These are extreme sample values which should not be downweighted too far when estimatingpopulation means or totals, because they are also influential in the population. Some theory on one-sided Winsorization cutoffs is presented. A simulation study based on business survey data finds that one-sided Winsorization with estimated optimal cutoffs performs well compared to other alternatives. We argue that theconcept of representative outliers deserves greater attention outside sample surveys, because the distribution including outliers and non-outliers is sometimes thecorrect target of inference, albeit a challenging one.

Bias-correction under a semiparametric model for small area estimation

Laura Dumitrescu

Victoria University Wellington, New Zealand

In recent years, several robust estimation techniques for estimating a unit-level model have been developed in the context of small area estimation. We consider a semiparametric framework and use a bias correction technique to obtain efficient robust predictors of the area means. The proposed predictor can be used when outliers occur in the random effects and/or possible outliers in the individual residuals, but can also in the case when data is derived from a mixture for which the mean of the outliers and the mean of the non-outliers are different.

Exploring the use of time series modelling in state space form for detection ofoutliers and structural changes

Oksana Honchar

Australian Bureau of Statistics, Australia

In this paper we first explain a strategy for removing the sample error component from the structural time series model. We then extend the model to incorporate multiple series to improve estimation and prediction of specific months in the series. We then demonstrate how this approach can detect the presence of an unusual movements in estimate in the series, and that detection is more effective with this approach.

Setting tuning parameters in one- and two-sided Winsorization in sample surveys

Phil Kokic[1], Robert Clark[2]

[1] Australian National University, Australia
[2] University of Wollongong, Australia

The choice of tuning parameters is particularly important in outlier treatment in sample surveys, because the aim is to predict the non-sample total of both outliers and non-outliers. A simplifying theoretical result is available for setting tuning parameters in one-sided Winsorization, but we show that no such result is possible in two-sided Winsorization. A number of alternatives for the two-sided case are explored. A simulation study evaluates different methods of setting tuning parameters for both one- and two-sided Winsorization for both levels and movements.

Robust population health

Alice Richardson

National Centre for Epidemiology and Population Health, Australian National University, Australia

Robust methods in population health research are very popular, but the types of robustness catered for form only a small subset of possible departures from a model. In this talk I will discuss what robustness means in a population health research context, describe the models considered, and compare the methods implemented. Finally I will offer some thoughts on how to embed a wider range of robust modelling into biostatistics as it is applied in population health research.

Small area estimation of expenditure proportions

Janice Scealy

Australian National University, Australia

Compositional data are vectors of proportions defined on the unit simplex and this type of constrained data occur frequently in Government surveys. It is also possible for the compositional data to be correlated due to the clustering or grouping of the observations within small domains or areas. We propose a new class of mixed model for compositional data based on the Kent distribution for directional data, where the random effects also have Kent distributions. One useful property of the new directional mixed model is that the marginal mean direction has a closed form and is interpretable. The random effects enter the model in a multiplicative way via the product of a set of rotation matrices and the conditional mean direction is a random rotation of the marginal mean direction. In small area estimation settings the mean proportions are usually of primary interest and these are shown to be simple functions of the marginal mean direction. For estimation we apply a quasi-likelihood method which results in solving a new set of generalised estimating equations and these are shown to have low bias in typical situations. For inference we use a nonparametric bootstrap method for clustered data which does not rely on estimates of the shape parameters (shape parameters are difficult to estimate in Kent models). We analyse data from the 2009–10 Australian Household Expenditure Survey CURF (confidentialised unit record file). We predict the proportions of total weekly expenditure on food and housing costs for households in a chosen set of domains. The new approach is shown to be more tractable than the traditional approach based on the logratio transformation.

Robustifying inference for probabilistically linked data with population auxiliary information

Suojin Wang[1], Nicola Salvati[2], Enrico Fabrizi[3], Ray Chambers[4]

[1] Texas A&M University, USA
[2] Università di Pisa, Italy
[3] Catholic University of the Sacred Heart, Italy
[4] University of Wollongong, Australia

Linkage errors occur when probability-based methods are used to link or match records from two or more distinct data sets corresponding to the same target population. These errors can lead to biased analytical decisions when they are ignored. We investigate an estimating equations approach to develop a bias correction method for secondary analysis of probabilistically linked data, using the missing information principle to accommodate the more realistic scenario of dependent linkage errors in both linear and logistic regression settings. We also develop the maximum likelihood solution when population auxiliary information in the form of population summary statistics is available. We examine how incorporation of population auxiliary information can robustify inference for linked sample data by removing measurement or linkage error bias. Our simulation results show that an incorrect assumption of independent linkage errors can lead to insufficient linkage error bias correction, while an approach that allows for correlated linkage errors appears to fully correct this bias.

Improving robustness of estimates from non-probability online samples

Dina Neiger, **Andrew Ward**, Darren Pennay

Social Research Centre, Australian National University, Australia

Weighting is a common method to reduce the total survey error for probability samples by adjusting for different chances of selection and by enforcing the population distribution across key demographic characteristics. There is no agreement on the efficacy of similar weighting adjustments for correcting bias of non-probability samples, however, given non-probability selection methods, the enforcement of quotas and the proprietary mechanisms used by sample providers to ensure that their sample resembles the population. Alternative methods, such as blending and calibration (e.g. DiSogra et al., 2011) and propensity-based weighting (e.g. Schonlau et al., 2003) have shown benefit but there is limited research available comparing the impact of different methods on the total survey error. The recent establishment of Australia's first probability online panel, Life in Australia, presents the opportunity to assess a range of weighting adjustments to reduce the bias of survey estimates from non-probability samples. With non-probability panels still representing the most common online collection methodology in Australia, the Life in Australia panel, in conjunction with data from the Australian Online Panels Benchmarking Study (Pennay et al., 2016), enables the evaluation of a number of different approaches to incorporate and improve the results of non-probability panels. By comparing estimates of key outcome variables with independent benchmarks, we are able to develop some general guidance for reducing bias and improving the robustness of survey estimates from non-probability online surveys.

Robust model-based sampling designs

Alan Welsh

Australian National University, Australia

A general issue in statistics, including survey sampling, is that the optimal design under a model often represents over-commitment to the model in the sense that using it can produce very good estimates when the model holds and very poor estimates when it does not. Moreover, optimal designs may not allow the possibility of either checking the model or fitting more general models. One way to approach at least the first problem is to consider robust designs which produce estimates that perform well when the designer's assumed model holds and also remain reasonably accurate in a neighbourhood of this central model. We will discuss this approach to sample design in the context of predicting the finite population total of asurvey variable that isrelated to anauxiliary variable that is available for all units in the population. The design problem is to specify a selection rule to select theunits for the sample, using only the values of the auxiliary variable,so that the predictor has optimal robustness properties. We will discuss the general issues in approaching this problem and describe theoptimally robust ('minimax') design approach of Welsh and Wiens (2012, Stat. Comput.).