19th International Conference on Computational Statistics

Paris - France, August 22-27, 2010

COMPSTAT'2010 Book of Abstracts

Conservatoire National des Arts et Métiers (CNAM)

and

the French National Institute for Research in Computer Science and Control (INRIA)

Title COMPSTAT'2010: Book of Abstracts

Cover design:

Authors: Several

Number of copies: 600

Preface

The 19th Conference of IASC-ERS, COMPSTAT'2010, is held in Paris, France, from August 22nd to August 27th 2010, locally organised by the Conservatoire National des Arts et Métiers (CNAM) and the French National Institute for Research in Computer Science and Control (INRIA). COMPSTAT is an initiative of the European Regional Section of the International Association for Statistical Computing (IASC-ERS), a section of the International Statistical Institute (ISI).

Keynote lectures are addressed by **Luc Devroye** (School of Computer Science, McGill University, Montreal), **Lutz Edler** (Division of Biostatistics, German Cancer Research Center, Heidelberg) and **David Hand** (Statistics section, Imperial College, London). Each COMPSTAT meeting is organised with a number of topics highlighted, which lead to Invited Sessions. The Conference program includes also contributed sessions and short communications (both oral communications and posters).

The Conference Scientific Program Committee chaired by Gilbert Saporta, CNAM, includes:

Ana Maria Aguilera, Universidad Granada
Avner Bar-Hen, Université René Descartes, Paris
Maria Paula Brito, University of Porto
Christophe Croux, Katholieke Universiteit Leuven
Michel Denuit, Université Catholique de Louvain
Gejza Dohnal, Technical University, Prag
Patrick J. F. Groenen, Erasmus University, Rotterdam
Georges Hébrail, TELECOM ParisTech, Paris
Henk Kiers, University of Groningen
Erricos Kontoghiorghes, University of Cyprus
Martina Mittlböck, Medical University of Vienna
Christian P. Robert, Université Paris-Dauphine, Paris
Maurizio Vichi, Universita La Sapienza, Roma
Peter Winker, Universität Giessen
Moon Yul Huh, SungKyunKwan University, Seoul, Korea
Djamel Zighed, Université Lumière, Lyon

who were responsible for the Conference Scientific Program, and whom the organisers wish to thank for their invaluable cooperation and permanent avaibility.

This book contains the abstracts corresponding to all the presentations at the conference.

The organisers would like to express their gratitude to all people from CNAM and INRIA who contributed to the success of COMPSTAT'2010, and worked actively for its organisation. We are very grateful to all our sponsors, for their generous support. Finally, we thank all authors and participants, without whom the conference would not have been possible.

The organisers of COMPSTAT'2010 wish the best success to Erricos Kontoghiorghes, Chairman of the 20th edition of COMPSTAT, which will be held in Cyprus in Summer 2012. See you there!

Paris, August 2010
The Local Committee Organisers

Gilbert Saporta
Yves Lechevallier

Stéphanie Aubin
Gérard Biau
Stéphanie Chaix
Marc Christine
Laurence de Crémiers
Séverine Demeyer
Thierry Despeyroux
Christian Derquenne
Vincenzo Esposito Vinzi
Ali Gannoun
Jean-Pierre Gauchi
Chantal Girodon
Pierre-Louis Gonzalez
Luan Jaupi
Ludovic Lebart
Ndeye Niang
Françoise Potier
Giorgio Russolillo
Julie Séguéla

Acknowledgements

The Editors are extremely grateful to the reviewers, whose work was determinant for the scientific quality of these proceeding. They were, in alphatbetical order:

Hervé Abdi
Ana Maria Aguilera
Massimo Aria
Josef Arlt
Avner Bar-Hen
Jean-Patrick Baudry
Younès Bennani
Petr Berka
Patrice Bertrand
Pierre Bertrand
Gerard Biau
Christophe Biernacki
Lynne Billard
Hans-Hermann Bock
Frank Bretz
Henri Briand
Maria Paula Brito
Edgar Brunner
Stephane Canu
Gilles Celeux
Andrea Cerioli
Roy Cerqueti
Ka Chun Cheung
Marc Christine
Guillaume Cleuziou
Claudio Conversano
Christophe Croux
Francisco de Assis De Carvalho
Michel Denuit
Christian Derquenne
Thierry Despeyroux
Gejza Dohnal
Antonio D'Ambrosio
Manuel Escabias
Vincenzo Esposito Vinzi
Christian Francq
Giuliano Galimberti
Ali Gannoun
Bernard Garel
Cristian Gatu
Jean-Pierre Gauchi
Pierre-Louis Gonzalez
Gérard Govaert
Patrick Groenen
Nistor Grozavu
Fabrice Guillet
Frederic Guilloux
Anne Gégout-Petit
Hakim Hacid
Peter Hall
André Hardy
Georges Hébrail
Harald Heinzl
Marc Hoffman
Moon Yul Huh
Alfonso Iodice d'Enza
Antonio Irpino
Junling Ji
François-Xavier Jollois
Henk A.L. Kiers
Dong Kim
Christine Kiss
Erricos Kontoghiorghes
Labiod Lazhar
Ludovic Lebart
Mustapha Lebbah
Yves Lechevallier
Seung Lee
Guodong Li
Olivier Lopez
Maria Laura Maag
Jean-Michel Marin
Claudia Marinica
Roland Marion-Gallois
Geoffrey McLachlan
Bertrand Michel

Martina Mittlboeck
Angela Montanari
Irini Moustaki
Shu Ng
Ndeye Niang
Monique Noirhomme
Francisco A. Ocaña
Matej Oresic
Chongsun Park
Franceso Palumbo
Fabien Picarougne
Jean-Michel Poggi
Christian Preda
Tommaso Proietti
Pierre Pudlo
Jan Rauch
Marco Riani
Christian Robert
Nicoleta Rogovschi
Rosaria Romano
Fabrice Rossi
Anthony Rossini
Judith Rousseau
Laurent Rouviere
Maurice Roux
Giorgio Russolillo
Lorenza Saitta
Ryan Skraba
Gilbert Saporta
Seisho Sato
Roberta Siciliano
Francoise Soulie Fogelman
Matthias Studer
Laura Trinchera
Brigitte Trousse
Mariano J. Valderrama
Stefan Van Aelst
Gilles Venturini
Rosanna Verde
Maurizio Vichi
Emmanuel Viennet
Cinzia Viroli
Michal Vrabec
Franois Wahl
William Wieczorek
Peter Winker
Jingyun Yang
In-Kwon Yeo
Kam Yuen
Daniela Zaharie
Djamel A. Zighed
Lihong Zhang
Xinyuan Zhao

Sponsors

We are extremely grateful to the following institutions whose support contributes to the success of COMPSTAT'2010:

- Conseil Régional Ile de France
- Mairie de Paris
- Société Française de Statistique
- Association EGC (Extraction et Gestion des Connaissances)
- Société Francophone de Classification
- Electricité de France
- Institut National de la Recherche Agronomique
- Institut National de Recherche sur les Transports et leur Sécurit
- Institut National de la Statistique et des Etudes Economiques
- IPSOS
- Orange Labs
- SAS-Institute

Programme

With the help of the Scientific Programme Committee and many reviewers, we have prepared a scientific programme which we hope very attractive, as well as the social programme.

Besides the 3 keynote lectures and the 39 invited communications, we received about 450 submissions. After the reviewing process we retained 127 'long papers', 92 'short papers' and 144 posters.

The authors come from 52 countries.

country	authors	country	authors	country	authors
Spain	115	Brazil	11	Luxembourg	2
France	94	Iran	10	Mexico	2
Japan	83	Hong Kong	8	Morocco	2
Italy	76	Korea Republic	8	Norway	2
Germany	52	China	5	Romania	2
United States	51	New Zealand	5	Slovenia	2
Belgium	39	Russian Federation	4	Bahrain	1
Czech Republic	38	Slovakia	4	Bangladesh	1
Turkey	30	Estonia	3	Colombia	1
United Kingdom	24	Finland	3	Georgia	1
Portugal	19	India	3	Greece	1
Austria	18	Netherlands	3	Hungary	1
Australia	16	Singapore	3	Indonesia	1
Taiwan	16	Sweden	3	Malaysia	1
Poland	15	Algeria	2	Niger	1
Switzerland	14	Argentina	2	Senegal	1
Canada	12	Bulgaria	2		
Tunisia	12	Lithuania	2		

Scientific Programme by Day

This book contains the abstracts corresponding to all the presentations at the conference. It is organized by day, according to the type of session: 3 keynote sessions, 14 invited sessions, 23 contributed sessions, 8 sessions of short papers and 2 posters sessions.

Monday, August 23

Tuesday, August 24

Wednesday , august 25

Thursday, August 26

Friday, August 27

Contents

CP2: Categorical Data Analysis

CP3: Computational Bayesian Methods

CP4: Multivariate Data Analysis 1

CP5: Time Series Analysis & Signal Processing 1

SP1: Biostatistics

CP6: Biostatistics & Bio-Computing 1

CP7: Clustering & Classification 1

CP8: Functional Data

CP11: Biostatistics & Bio-Computing 2

CP12: Clustering & Classification 2

CP13: Nonparametric Statistics & Smoothing

CP14: Optimization heuristics in Statistical Modelling

SP3: Econometrics & Finance

SP4: Multivariate Analysis 1 & Spatial Statistics

CP15: Multivariate Data Analysis 3

CP16: Data Visualization

CP17: Sampling Methods

CP18: Spatial Statistics

CP19: Time Series Analysis & Signal Processing 3

SP5: Classification

SP6: Data Analysis

IP6: Optimization heuristics in statistical modelling

KN2: Keynote Speaker 2

IP7: Data Stream Mining

IP8: ARS Session (Financial) Time Series

Part IV. Thursday August 26

IP9: Spatial Statistics / Spatial Epidemiology

IP10: KDD Session: Topological Learning

PS2: Poster Session 2

IP11: ABC Methods for Genetic Data

IP12: IFCS Session

CP20: Computational Econometrics & Finance

CP21: Machine Learning

CP22: Numerical Methods

CP23: Symbolic Data Analysis

SP7: Multivariate Analysis 2

SP8: Time Series & Numerical Methods

IP14: Monte Carlo Methods

KN3: Kenote Speaker 3

Part I

Monday August 23

The Laws of Coincidence

David J. Hand

Imperial College London
and
Winton Capital Management
d.j.hand@imperial.ac.uk

Abstract. Anomalous events often lie at the roots of discoveries in science and of actions in other domains. Familiar examples are the discovery of pulsars, the identification of the initial signs of an epidemic, and the detection of faults and fraud. In general, they are events which are seen as so unexpected or improbable that one is led to suspect there must be some underlying cause. However, to determine whether such events are genuinely improbable, one needs to evaluate their probability under normal conditions. It is all too easy to underestimate such probabilities. Using the device of a number of 'laws', this paper describes how apparent coincidences should be expected to happen by chance alone.

Keywords: anomalies, coincidences, hidden forces

Robust Model Selection with LARS Based on S-estimators

Claudio Agostinelli[1] and Matias Salibian-Barrera[2]

[1] Dipartimento di Statistica
Ca' Foscari University
Venice, Italy *claudio@unive.it*
[2] Department of Statistics
The University of British Columbia
Vancouver, BC, Canada *matias@stat.ubc.ca*

Abstract. We consider the problem of selecting a parsimonious subset of explanatory variables from a potentially large collection of covariates. We are concerned with the case when data quality may be unreliable (e.g. there might be outliers among the observations). When the number of available covariates is moderately large, fitting all possible subsets is not a feasible option. Sequential methods like forward or backward selection are generally "greedy" and may fail to include important predictors when these are correlated. To avoid this problem Efron et al. (2004) proposed the Least Angle Regression algorithm to produce an ordered list of the available covariates (sequencing) according to their relevance. We introduce outlier robust versions of the LARS algorithm based on S-estimators for regression (Rousseeuw and Yohai (1984)). This algorithm is computationally efficient and suitable even when the number of variables exceeds the sample size. Simulation studies show that it is also robust to the presence of outliers in the data and compares favourably to previous proposals in the literature.

Keywords: robustness, model selection, LARS, S-estimators, robust regression

Robust Methods for Compositional Data

Peter Filzmoser[1] and Karel Hron[2]

[1] Vienna University of Technology
Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria, *P.Filzmoser@tuwien.ac.at*
[2] Palacký University, Faculty of Science
tř. 17. listopadu 12, CZ-77146 Czech Republic, *hronk@seznam.cz*

Abstract. Many practical data sets in environmental sciences, official statistics and various other disciplines are in fact compositional data because only the ratios between the variables are informative. Compositional data are represented in the Aitchison geometry on the simplex, and for applying statistical methods designed for the Euclidean geometry they need to be transformed first. The isometric logratio (ilr) transformation has the best geometrical properties, and it avoids the singularity problem introduced by the centered logratio (clr) transformation. Robust multivariate methods which are based on a robust covariance estimation can thus only be used with ilr transformed data. However, usually the results are difficult to interpret because the ilr coordinates are formed by non-linear combinations of the original variables. We show for different multivariate methods how robustness can be managed for compositional data, and provide algorithms for the computation.

Keywords: Aitchison geometry, logratio transformations, robustness, affine equivariance, multivariate statistical methods

Detecting Multivariate Outliers Using Projection Pursuit with Particle Swarm Optimization

Anne Ruiz-Gazen[1], Souad Larabi Marie-Sainte[2], and Alain Berro[2]

[1] Toulouse School of Economics (Gremaq et IMT),
21, allée de Brienne, 31000 Toulouse, France
ruiz@cict.fr
[2] IRIT, 21, allée de Brienne, 31000 Toulouse, France
larabi@irit.fr, berro@irit.fr

Abstract. Detecting outliers in the context of multivariate data is known as an important but difficult task and there already exist several detection methods. Most of the proposed methods are based either on the Mahalanobis distance of the observations to the center of the distribution or on a projection pursuit (PP) approach. In the present paper we focus on the one-dimensional PP approach which may be of particular interest when the data are not elliptically symmetric. We give a survey of the statistical literature on PP for multivariate outliers detection and investigate the pros and cons of the different methods. We also propose the use of a recent heuristic optimization algorithm called Tribes for multivariate outliers detection in the projection pursuit context.

Keywords: heuristic algorithms, multivariate outliers detection, particle swarm optimization, projection pursuit, Tribes algorithm

Empirical Dynamics and Functional Data Analysis

Hans-Georg Müller

Department of Statistics, University of California, Davis
One Shields Avenue, Davis, CA 95616, U.S.A. *mueller@wald.ucdavis.edu*

Abstract. We review some recent developments on modeling and estimation of dynamic phenomena within the framework of Functional Data Analysis (FDA). The focus is on longitudinal data which correspond to sparsely and irregularly sampled repeated measurements that are contaminated with noise and are available for a sample of subjects. A main modeling assumption is that the data are generated by underlying but unobservable smooth trajectories that are realizations of a Gaussian process. In this setting, with only a few measurements available per subject, classical methods of Functional Data Analysis that are based on presmoothing individual trajectories will not work. We review the estimation of derivatives for sparse data, the PACE package to implement these procedures, and an empirically derived stochastic differential equation that the processes satisfy and that consists of a linear deterministic component and a drift process.

Keywords: dynamics, Gaussian process, drift term

Bootstrap Calibration in Functional Linear Regression Models with Applications

Wenceslao González-Manteiga[1] and Adela Martínez-Calvo[2]

[1] Departamento de Estadística e I.O., Universidad de Santiago de Compostela, Facultad de Matemáticas, Campus Sur, 15782, Santiago de Compostela, Spain
wenceslao.gonzalez@usc.es

[2] Departamento de Estadística e I.O., Universidad de Santiago de Compostela, Facultad de Matemáticas, Campus Sur, 15782, Santiago de Compostela, Spain
adela.martinez@usc.es

Abstract. Our work focuses on the functional linear model given by $Y = \langle \theta, X \rangle + \epsilon$, where Y and ϵ are real random variables, X is a zero-mean random variable valued in a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$, and $\theta \in \mathcal{H}$ is the fixed model parameter. Using an initial sample $\{(X_i, Y_i)\}_{i=1}^n$, a bootstrap resampling $Y_i^* = \langle \hat{\theta}, X_i \rangle + \hat{\epsilon}_i^*$, $i = 1, \ldots, n$, is proposed, where $\hat{\theta}$ is a general pilot estimator, and $\hat{\epsilon}_i^*$ is a naive or wild bootstrap error. The obtained consistency of bootstrap allows us to calibrate distributions as $P_X\{\sqrt{n}(\langle \hat{\theta}, x \rangle - \langle \theta, x \rangle) \leq y\}$ for a fixed x, where P_X is the probability conditionally on $\{X_i\}_{i=1}^n$. Different applications illustrate the usefulness of bootstrap for testing different hypotheses related with θ, and a brief simulation study is also presented.

Keywords: bootstrap, functional linear regression, functional principal components analysis, hypothesis test

Anticipated and Adaptive Prediction in Functional Discriminant Analysis

Cristian Preda[1], Gilbert Saporta[2], and Mohamed Hadj Mbarek[3]

[1] Ecole Polytehnique Universitaire de Lille& Laboratoire Painlevé, UMR 8524
Université des Sciences et Technologies de Lille, France,
cristian.preda@polytech-lille.fr
[2] Chaire de statistique appliquée & CEDRIC, CNAM
292 rue Saint Martin, Paris, France, *gilbert.saporta@cnam.fr*
[3] Institut Suprieur de Gestion de Sousse, Tunisie, *benmbarekmhedi@yahoo.fr*

Abstract. Linear discriminant analysis with binary response is considered when the predictor is a functional random variable $X = \{X_t, t \in [0,T]\}, T \in \mathbb{R}$. Motivated by a food industry problem, we develop a methodology to anticipate the prediction by determining the smallest T^*, $T^* \leq T$, such that $X^* = \{X_t, t \in [0,T^*]\}$ and X give similar predictions. The adaptive prediction concerns the observation of a new curve ω on $[0,T^*(\omega)]$ instead of $[0,T]$ and answers to the question "How long should we observe ω ($T^*(\omega) =?$) for having the same prediction as on $[0,T]$?". We answer to this question by defining a conservation measure with respect to the class the new curve is predicted.

Keywords: functional data, discriminant analysis, classification, adaptive prediction

Robust Principal Component Analysis Based on Pairwise Correlation Estimators

Stefan Van Aelst[1], Ellen Vandervieren[2], and Gert Willems[1]

[1] Dept. of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, B-9000 Ghent, Belgium. *stefan.vanaelst@ugent.be ; gertllwillems@gmail.com*

[2] Dept. of Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp, Belgium. *ellen.vandervieren@ua.ac.be*

Abstract. Principal component analysis (PCA) tries to explain and simplify the structure of multivariate data. For standardized variables, these principal components correspond to the eigenvectors of the correlation matrix. Unfortunately, the classical correlations are very sensitive to aberrant observations. Therefore, robust methods for PCA have been developed (see e.g. Hubert et al. (2005)).

To study robustness properties at high dimensional data, Alqallaf et al. (2009) proposed a contamination model, which assumes that each variable is contaminated independently. Hence, each variable is assumed to have a majority of outlier-free values, but there is not necessarily a majority of outlier-free observations.

To obtain a robust PCA that resists better independent contamination, we estimate the correlation matrix componentwise by using robust pairwise correlation estimates. We show that this approach does not need a majority of outlier-free observations which becomes very useful for high dimensional problems. We further demonstrate that the "bivariate trimming" method (see Khan et al. (2007)) especially works well in this setting. Finally, we show how the PCA approach based on pairwise correlation estimators can be used to do a fast and robust principal component regression.

Keywords: principal component analysis, robustness, high dimensional data, trimming, principal component regression.

References

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.*

HUBERT, M., ROUSSEEUW, P.J. and VANDEN BRANDEN, K. (2005): ROBPCA: a New Approach to Robust Principal Component Analysis. *Technometrics 47, 64-79.*

KHAN, J.A., VAN AELST, S. and ZAMAR, R.H. (2007): Robust Linear Model Selection Based on Least Angle Regression. *Journal of the American Statistical Association 102 (12), 1289-1299.*

DetMCD in a Regression Framework

Tim Verdonck[1], Mia Hubert[2], and Peter J. Rousseeuw[3]

[1] Department of Mathematics and Computer Science, University of Antwerp
Middelheimlaan 1, Antwerp, Belgium, *Tim.Verdonck@ua.ac.be*
[2] Department of Mathematics, Katholieke Universiteit Leuven
Celestijnenlaan 200b, Leuven, Belgium, *Mia.Hubert@wis.kuleuven.be*
[3] Department of Mathematics, Katholieke Universiteit Leuven
Celestijnenlaan 200b, Leuven, Belgium, *peter@rousseeuw.net*

Abstract. The minimum covariance determinant (MCD) method is a robust estimator of multivariate location and scatter (Rousseeuw (1984)). The MCD is highly resistant to outliers. It is often applied by itself and as a building block for other robust multivariate methods. Computing the exact MCD is very hard, so in practice one resorts to approximate algorithms. Most often the FASTMCD algorithm of Rousseeuw and Van Driessen (1999) is used. This algorithm starts by drawing many random subsets, followed by so-called concentration steps. The FASTMCD algorithm is affine equivariant but not permutation invariant. Recently we have developed a deterministic algorithm, denoted as DetMCD, which does not use random subsets and which is much faster. It is permutation invariant and very close to affine equivariant. In this paper DetMCD is illustrated in a regression framework. We focus on robust principal component regression and partial least squares regression, two very popular regression techniques for collinear data. We also apply DetMCD on data with missing elements after plugging it into the ER-PCR technique of Serneels and Verdonck (2009).

Keywords: affine equivariance, outliers, robustness, RPCR, RSIMPLS

References

ROUSSEEUW, P.J. (1984): Least median of squares regression. *Journal of the American Statistical Association 79, 871-880.*

ROUSSEEUW, P.J. and VAN DRIESSEN, K. (1999): A fast algorithm for the minimum covariance determinant estimator. *Technometrics 41, 212-223.*

SERNEELS, S. and VERDONCK, T. (2009): Principal component regression for data containing outliers and missing elements. *Computational Statistics and Data Analysis 53(11), 3855-3863.*

Diagnostic Checking of Multivariate Normality Under Contamination

Andrea Cerioli

Dipartimento di Economia, Università di Parma
via Kennedy 6, 43100 Parma, Italy, *andrea.cerioli@unipr.it*

Abstract. The normal distribution has a central place in the analysis of multivariate data. It is also important in the development of robust high-breakdown methodologies, since it is often assumed to describe the genesis of the "good" part of the data (Hubert et al. (2008)). In this paper we describe a simple and effective way to assess multivariate normality which can be used when the data contain outliers. Our proposal is based on accurate distributional results for the squared robust Mahalanobis distances computed from the reweighted Minimum Covariance Determinant estimator (Cerioli (2010)). It and can be seen as a robust version of the classical diagnostic methods usually applied to detect departures from multivariate normality (Gnanadesikan (1997), Small (1978)).

Keywords: outliers, quantile plot, reweighted MCD, robust distances

References

CERIOLI, A. (2010): Multivariate outlier detection with high-breakdown estimators *Journal of the American Statistical Association, 105 (489), 147–156.*

GNANADESIKAN, R. (1997): *Methods for Statistical Data Analysis of Multivariate Observations.* Second Edition. Wiley, New York.

HUBERT, M., ROUSSEEUW, P. J. and VAN AELST, S. (2008): High-breakdown robust multivariate methods *Statistical Science 23 (1), 92–119.*

SMALL, N. J. H. (1978): Plotting squared radii. *Biometrika 65 (3), 657–658.*

Regularized directions of maximal outlyingness

Michiel Debruyne[1]

Department of mathematics and computer science, Universiteit Antwerpen, Middelheimlaan 1G, 2020 Antwerpen, Belgium, *michiel.debruyne@ua.ac.be*

Abstract. In multivariate statistics many robust covariance estimators have been proposed in the literature. If the data contains outlying observations, such estimators detect the outliers and retreive the covariance structure of the regular data. If an outlier is detected, it is quite natural to wonder which variables contribute the most to this outlyingness, especially if the dimension of the data is rather high. A straightforward idea is to check the coefficients of the univariate direction for which the standardized distance between the projected outlier and a projected multivariate location estimate is maximal. However, this strategy comes with a few drawbacks. The coefficients for instance highly depend on the covariance structure of the non outlying observations. Moreover in high dimensions one obtains very unreliable results due to the curse of dimensionality.

A possible solution is to rewrite the direction of maximal outlyingness as the normed solution of a classical least squares regression problem. We propose to add a L_1 penalty term in this expression, thus replacing the classical least squares regression by the LASSO. This yields a path of regularized directions of maximal outlyingness. Based on such a path, an algorithm is proposed to select a subset of variables that are most relevant to the outlyingness of the outlier under consideration.

Keywords: Robust statistics, outlyingness, variable selection, LASSO

Two Kurtosis Measures in a Simulation Study

Anna Maria Fiori

Department of Quantitative Methods for Economics and Business Sciences
University of Milano-Bicocca, Milano, Italy, *anna.fiori@unimib.it*

Abstract. We consider two measures of right/left/overall kurtosis which arise from a recent interpretation of kurtosis as inequality at either side of the median (Zenga (2006), Fiori (2008)). Based on Zenga's kurtosis curve, the measures apply to both symmetric and asymmetric distributions, their interpretation is clear and they presume the existence of either first or second moments. Focusing here on the symmetric case, we derive the symmetric influence functions (Ruppert (1987), Fiori and Zenga (2005)) of these measures and evaluate their sensitivity to contamination in the tails and at the center. Sampling properties of the proposed measures are investigated by a simulation-based approach and bootstrap confidence intervals are constructed for small and medium sample sizes. Compared to the standardized fourth moment coefficient (conventional kurtosis), the two measures are shown to provide both a more reliable and a more sophisticated picture of the kurtosis risk embedded in a dataset.

Keywords: right kurtosis, left kurtosis, inequality, influence function, bootstrap

References

FIORI, A. M. (2008): Measuring kurtosis by right and left inequality orders. *Communications in Statistics: Theory and Methods 37 (17), 2665–2680.*

FIORI, A. M. and ZENGA, M. (2005): The meaning of kurtosis, the influence function and an early intuition by L. Faleschini. *Statistica 65 (2), 135–144.*

RUPPERT, D. (1987): What is kurtosis? An influence function approach. *The American Statistician 41 (1), 1–5.*

ZENGA, M. (2006): Kurtosis. In: S. Kotz, C. B. Read, N. Balakrishnan and B. Vidakovic (Eds.): *Encyclopedia of Statistical Sciences.* Wiley, New York, 2nd online edition.

On Empirical Composite Likelihoods

Nicola Lunardon, Francesco Pauli, and Laura Ventura

Department of Statistics
Via C. Battisti 241, Padova, Italy
lunardon@stat.unipd.it, *fpauli@stat.unipd.it*, *ventura@stat.unipd.it*

Abstract. Composite likelihood functions are convenient surrogates for the ordinary likelihood, when the latter is too difficult or even impractical to compute, and they may be more robust to model misspecification (see e.g. Cox and Reid (2004) and Varin (2008)). One drawback of composite likelihood methods is that the composite likelihood analogue of the likelihood ratio statistic does not have the standard χ^2 asymptotic distribution.

Invoking the theory of unbiased estimating equations, in this paper we ropose and discuss the computation of the empirical likelihood function (Owen, 2001) from the unbiased composite scores. Two Monte Carlo studies are performed in order to assess the finite-sample performance of the proposed empirical composite likelihood procedures.

Keywords: Empirical likelihood, Estimating function, Likelihood methods, Pairwise likelihood, Pseudo-likelihood

References

COX, D.R. and REID, N. (2004): A note on pseudolikelihood constructed from marginal densities. *Biometrika 91, 729–737.*

OWEN, A.B. (2001): *Emprirical likelihood.* Chapman and Hall, London.

VARIN, C. (2008): On composite marginal likelihoods. *Advances in Statistical Analysis 92, 1–28.*

Mixtures of Weighted Distance-Based Models for Ranking Data

Paul H. Lee[1] and Philip L. H. Yu[2]

[1] Department of Statistics and Actuarial Science,
The University of Hong Kong, Hong Kong, *honglee@hku.hk*
[2] Department of Statistics and Actuarial Science,
The University of Hong Kong, Hong Kong, *plhyu@hku.hk*

Abstract. Ranking data has applications in different fields of studies, like marketing, psychology and politics. Over the years, many models for ranking data have been developed. Among them, distance-based ranking models, which originate from the classical rank correlations, postulate that the probability of observing a ranking of items depends on the distance between the observed ranking and a modal ranking. The closer to the modal ranking, the higher the ranking probability is. However, such a model basically assumes a homogeneous population, and the single dispersion parameter may not be able to describe the data very well.

To overcome the limitations, we consider new weighted distance measures which allow different weights for different ranks in formulating more flexible distance-based models. The mixtures of weighted distance-based models are also studied for analyzing heterogeneous data. Simulations results will be included, and we will apply the proposed methodology to analyze a real world ranking dataset.

Keywords: ranking data, distance-based model, mixture model

Clustering with Mixed Type Variables and Determination of Cluster Numbers

Hana Řezanková[1], Dušan Húsek[2], and Tomáš Löster[1]

[1] University of Economics, Prague, nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic, {*hana.rezankova|tomas.loster*}*@vse.cz*
[2] ICS, AS CR, Pod vodárenskou veží 2, 182 07 Praha 8, Czech Republic, *dusan.husek@cs.cas.cz*

Abstract. Cluster analysis is a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. The heart of the method is assignment of a set of objects into subsets so that objects in the same cluster are similar in some sense. For measuring the similarity of objects, these are characterized by variables (features), most often quantitative. However, special techniques for the case of mixed type variables (nominal and quantitative) have been proposed. Logically, also many coefficients for evaluation of clustering and determination of cluster numbers have been proposed. However, these coefficients are proposed mainly for objects characterized by quantitative variables (Gordon, 1999; Gan et al., 2007; Kogan, 2007). We extend the principles of clustering evaluation and determination of cluster numbers for the case of objects with quantitative variables to the objects with mixed types variables. A combination of variance (for quantitative variables) and entropy or Gini's coefficient of mutability (for nominal variables) is applied. We propose the measure of within-cluster variability based on Gini's coefficient, within-cluster variability difference, uncertainty index, tau index, and semi-partial versions of these indices, modified pseudo F indices (based both on entropy and on Gini's coefficient) and analogy of BIC criterion based on Gini's coefficient. Two-step cluster analysis is applied to questionnaire survey data several times, with numbers of clusters as a parameter. Proposed coefficients are used for optimal assignment of respondents to clusters. The results of clustering of respondents characterized by different sets of variables are compared.

Keywords: cluster analysis, entropy, Gini's coefficient of mutability, cluster number determination, Schwarz's Bayesian information criterion

References

GAN, G., MA, C. and WU, J. (2007): *Data Clustering: Theory, Algorithms, and Applications*. ASA-SIAM, Philadelphia.

GORDON, A. D. (1999): *Classification, 2nd Edition*. Chapman & Hall/CRC, Boca Raton.

KOGAN, J. (2007): *Introduction to Clustering Large and High-Dimensional Data*. Cambridge University Press, New York.

Acknowledgement. This work was supported by projects AV0Z10300504, GACR P202/10/0262, 205/09/1079, and IGA VSE F4/3/2010.

Multiblock Method for Categorical Variables. Application to the Study of Antibiotic Resistance

Stéphanie Bougeard[1], El Mostafa Qannari[2] and Claire Chauvin[1]

[1] AFSSA (French Agency for Food Safety), Department of Epidemiology, Zoopole, BP53, 22440 Ploufragan, France, *s.bougeard@afssa.fr, c.chauvin@afssa.fr*

[2] ONIRIS (Nantes-Atlantic National College of Veterinary Medicine, Food Science and Engineering), Department of Sensometrics and Chemometrics, Rue de la Géraudière, 44322 Nantes Cedex, France, *elmostafa.qannari@oniris-nantes.fr*

Abstract. We address the problem of describing several categorical variables with a prediction purpose. We focus on methods in the multiblock modelling framework, each block being formed of the indicator matrix associated with each qualitative variable. We propose a categorical extension of an alternative method (Bougeard et al., 2008) to multiblock PLS (Wold, 1984) and shall refer to it as categorical multiblock Redundancy Analysis. The main idea is that each indicator matrix is summed up with a latent variable which represents the coding of the categorical variable. In addition, the structural model which specifies the relations among latent variables is based on a well-identified global optimization criterion which leads to an eigensolution. Practical uses of the proposed method are illustrated using an empirical example in the field of veterinary epidemiology. The aim is to study of the relationships between antibiotic consumption on farms and antibiotic resistance in healthy slaughtered poultry. The variable of interest is the Nalidixic Acid resistance, studied in the light of 14 potential explanatory variables, related to the chicken production type, the previous antimicrobial treatments and the co-resistances observed. Risk factors are given. Moreover, the method is compared to logistic regression, *Disqual* (Saporta, & Niang, 2006) and a categorical extension of multiblock PLS.

Keywords: Supervised classification, discriminant analysis, multiblock redundancy analysis, multiblock PLS, categorical variables

References

BOUGEARD, S., HANAFI, M., LUPO, C. and QANNARI, E.M. (2008): From multiblock Partial Least Squares to multiblock redundancy analysis. A continuum approach. In: *International Conference on Computational Statistics*. Porto, 607–616

SAPORTA, G. and NIANG, N. (2006): *Correspondence analysis and classification (Chap. 16). Multiple correspondence analysis and related method*. Chapman & Hall, 372-392.

WOLD, S. (1984): Three PLS algorithms according to SW. In: *Symposium MULDAST*, Umea, 26–30.

Boolean Factor Analysis by the Expectation-Maximization Algorithm

Alexander Frolov[1], Pavel Polyakov[2], and Dušan Húsek[3]

[1] Institute of Higher Nervous Activity and Neurophysiology of the Russian Academy of Sciences, Butlerova 5a, Moscow, Russia *aafrolov@mail.ru*
[2] Scientific-Research Institute for System Studies of the Russian Academy of Science, Nakhim. prosp. 36/1, 117 218 Moscow, Russia *pavel.mipt@mail.ru*
[3] ICS, AS CR, Pod vodárenskou veží 2, 182 07 Praha 8, Czech Republic, *dusan.husek@cs.cas.cz*

Abstract. To evaluate the performance of Boolean factor analysis (*BFA*) we suggest, first, a general *BFA* generative model and, second, a general informational measure of *BFA* efficiency.

If hidden factor structure of the data set is unknown its entropy is evaluated as $H_0 = M \sum_j^N h(p_j)$, where M is the number of pattern in the data set, N is pattern dimensionality, p_j is the probability of j-th component to take One, and h is Shannon function. If hidden factor structure of the data set is detected by *BFA* its entropy is evaluated as $H = M \sum_{i=1}^{L} h(\pi_i) + \sum_{m=1}^{M} \sum_{j=1}^{N} h(P_j^m)$, where π_i is the probability of i-th score to take One, L is the total number of factors and P_j^m is the probability of j-th signal component in m-th pattern of the data set to take One when scores are given. The relative information gain $G = (H_0 - H)/H_0$ is suggested to be a general measure of *BFA* efficiency.

Compared are efficiencies of two *BFA* methods. First, based on the expectation-maximization (*EM*) technique (Dempster et al., 1977): called Maximal Causes Analysis (MCA_3) and proposed by Lücke and Sahani (2008). Second, Expectation-Maximization Boolean Factor Analysis (*EMBFA*) we developed as the direct application of *EM* to the suggested general *BFA* generative model. Comparison is based on the so-called bars problem benchmark (Földiak, 1990). It is shown that the efficiency of *EMBFA* is higher than that of MCA_3 in *BFA* generative model parameters entirety.

Keywords: Boolean factor analysis, generative model, information gain, efficiency measure

References

DEMPSTER, A., LAIRD, N. and RUBIN, D. (1977): Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological 39(1), 1–38.*

LÜCKE, J. and SAHNI, M. (2008): Maximal causes for non-linear component extraction. *The Journal of Machine Learning Research 9, 1227–1267.*

FÖLDIAK, P. (1990): Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics 64, 165–170.*

Acknowledgement. This work was partially supported by projects AV0Z10300504, GACR P202/10/0262, GACR 205/09/1079 and 1M10567.

Statistical Inference on Large Contingency Tables: Convergence, Testability, Stability

Marianna Bolla

Institute of Mathematics, Budapest University of Technology and Economics
Egry József u. 1., Budapest, Hungary, *marib@math.bme.hu*

Abstract. Convergence of rectangular arrays with nonnegative, bounded entries is defined together with the limit object and cut distance. A statistic defined on a contingency table is testable if it can be consistently estimated based on a smaller, but still sufficiently large table which is selected randomly from the original one in an appropriate manner. By the above randomization, classical multivariate methods can be carried out on a smaller part of the array. This fact becomes important when our task is to discover the structure of large and evolving arrays, like genetic maps, social, and communication networks. Special block structures behind large tables are also discussed from the point of view of stability and spectra.

Keywords: convergence of contingency tables, testable contingency table parameters, block matrices, spectrum and stability

References

BOLLA, M., FRIEDL, K. and KRÁMLI, A. (2010): Singular value decomposition of large random matrices (for two-way classification of microarrays). *Journal of Multivariate Analysis 101, 434-446.*

DIACONIS, P. and JANSON, S. (2008): Graph limits and exchangeable random graphs. *Rendiconti di Matematica 28 (Serie VII), 33-61.*

ÉRDI, P. and TÓTH, J. (1990): What is and what is not stated by the May-Wigner theorem? *J. Theor. Biol. 145, 137-140.*

FRIEZE, A. and KANNAN, R. (1999): Quick approximation to matrices and applications. *Combinatorica 19, 175-220.*

JUHÁSZ, F. (1996): On the structural eigenvalues of block random matrices. *Linear Algebra and Its Applications 246, 225-231.*

LOVÁSZ, L. and SZEGEDY, B. (2006): Limits of dense graph sequences. *J. Comb. Theory B 96, 933-957.*

MAY, R. M. (1972): Will a large complex system be stable? *Nature 238, 413-414.*

How to Take into Account the Discrete Parameters in the BIC Criterion?

Vincent Vandewalle

Département STID, IUT C Roubaix, Université Lille 2
25-27, rue du Maréchal Foch, 59100 Roubaix, France,
vincent.vandewalle@univ-lille2.fr

Abstract. When using the BIC criterion to select one model among several models, only the continuous parameters are taken into account in the penalization. However, when considering models with discrete parameters to be estimated, this criterion can lead to select too simple models, not taking into account the possible over-fitting caused by the estimation of the discrete parameters. Ideally we would like to integrate the likelihood on every possible value of the discrete parameter. In this article we study how this integral can be approximated from a practical and theoretical point of view in the particular case of the parsimonious multinomial distribution.

Keywords: model selection, discrete parameters, parsimonious models, Bayesian Integration Criterion

Bayesian Flexible Modelling of Mixed Logit Models

Luisa Scaccia[1] and Edoardo Marcucci[2]

[1] Dip. di Istituzioni Economiche e Finanziarie, Università di Macerata, via Crescimbeni 20, 62100 Macerata, Italy, *scaccia@unimc.it*
[2] Dip. di Istituzioni Pubbliche, Economia e Società, Università di Roma Tre, via G. Chiabrera 199, 00145 Roma, Italy, *edoardo.marcucci@uniroma3.it*

Abstract. The widespread use of the Mixed Multinomial Logit model, in the context of discrete choice data, has made the issue of choosing a mixing distribution very important. The choice of a specific distribution may seriously bias results if that distribution is not suitable for the data. We propose a flexible hierarchical Bayesian approach in which the mixing distribution is approximated through a mixture of normal distributions. Numerical results on a real data set are provided to demonstrate the usefulness of the proposed method.

Keywords: Hierarchical Bayes, mixed logit, mixture of distributions, random taste heterogeneity, semi-parametric estimation

Determining the Direction of the Path Using a Bayesian Semiparametric Model

Kei Miyazaki[1], Takahiro Hoshino[1], and Kazuo Shigemasu[2]

[1] Graduate School of Economics, Nagoya University
Furo-cho, Chikusa-Ku, Aichi 464-8601, Japan,
miyazaki.behaviormetrics@gmail.com
[2] Department of Psychology, Teikyo University
Otsuka 359, Hachioji-shi, Tokyo 192-0352, Japan, *kshige@main.teikyo-u.ac.jp*

Abstract. We often face a situation where the observed variables do not follow normal distributions. Thus, the methods that do not require normality for the error variables are required to resolve this situation. Recently, as a solution-oriented approach for the above problem, an estimation method that uses a higher-order moment structure sparked interest among researchers in this field (Shimizu and Kano (2008)). This method makes it possible to determine the direction of path among the models that have the same values of goodness of fit (that is, equivalent models).
On the other hand, in Bayesian estimation, hierarchical models with Dirichlet process prior distributions have been applied to various kinds of models (Ansari and Iyengar (2006)). This method makes it possible to perform MCMC estimation for any assumed shape of distributions for the parameters.
In this study, we consider a simple single regression model, and set two Dirichlet process mixture models wherein the explanatory and dependent variables are alternated with each other under the assumption that the error variables do not follow normal distributions. Then, we decide which model is better by calculating the marginal likelihood (Basu and Chib (2003)) in simulation studies. We can see from the simulation results that originally the direction of the path does not relate to the direction of causation and that it is meaningless to identify the causation by determining the direction of the path using nonnormal error variables.

Keywords: Bayesian semiparametric models, Dirichlet process priors, marginal likelihood, equivalent models

References

ANSARI, A. and IYENGAR, R. (2006): Semiparametric thurstonian models for recurrent choices: A Bayesian analysis. *Psychometrika, 71(4), 631-657.*

BASU, S. and CHIB, S. (2003): Marginal likelihood and Bayes factors for Dirichlet Process mixture models. *Journal of the American Statistical Association, 98, 224-235.*

SHIMIZU, S. and KANO, Y. (2008): Use of non-normality in structural equation modeling: application to direction of causation. *Journal of Statistical Planning and Inference, 138, 3483-3491.*

Metropolis-Hastings Algorithm for Mixture Model and its Weak Convergence

Kengo Kamatani

Graduate School of Mathematical Sciences, The University of Tokyo,
3-8-1 Komaba, Meguro-ku, Tokyo 153-0041, Japan *kengok@ms.u-tokyo.ac.jp*

Abstract. This paper describes an application of the weak convergence framework of the Markov chain Monte Carlo (MCMC) method. It is well known that for the mixture model, when some of the mixture proportion parameters are 0, the Gibbs sampler behaves poorly. In this paper, we propose a simple Metropolis-Hastings (MH) algorithm and study its convergence property. In a usual Harris recurrence framework, both the MH algorithm and the Gibbs sampler are geometrically ergodic in probability 1. However, in the weak convergence framework, the former is consistent in a certain sense, but the latter is not. We present some numerical results.

Keywords: Gibbs sampler, Finite Mixture Model, Geometrical Ergodicity, Local Asymptotic Normality, Bernstein von-Mises's Theorem

A simulation study of the Bayes estimator of parameters in an extension of the exponential distribution

Samira Sadeghi

Depatment of Mathematics, Statistics and computer Sciences, University of Tehran, Tehran, Iran

Abstract. Recently a generalization of the exponential distribution has been introduced by Nadarajah & Haghighi (2009). In this paper, we consider the Bayes estimators of the scale and shape parameters of this family under the assumptions of gamma priors and squared error loss function. We used the idea of Lindley for obtaining the approximate Bayes estimators. Under assumptions of non-informative priors, the approximate Bayes estimators are computed and compared with the corresponding maximum likelihood estimators. One real data set has been analyzed to demonstrate how the proposed method can be used in practice. For this data set, we also computed the approximate Bayes estimators using MCMC technique and compared the results.

Keywords: Bayes estimator, Exponential distribution, Lindley approximation, MCMC, Monte Carlo simulation

Pseudo-Bayes Factors

Stefano Cabras[1], Walter Racugno[1], and Laura Ventura[2]

[1] Department of Mathematics, University of Cagliari
Via Ospedale 72, Cagliari, Italy, *s.cabras@unica.it, racugno@unica.it*
[2] Department of Statistics, University of Padova
Via C. Battisti 241, Padova, Italy, *ventura@stat.unipd.it*

Abstract. The use of Bayes factors (BF) in hypothesis testing may encounter difficulties in the presence of unknown nuisance parameters. Indeed, their elimination requires in general both the computation of multidimensional integrals and the elicitation of prior distributions. In modern frequentist and Bayesian literature, the elimination of nuisance parameters can be carried out by resorting to pseudo-likelihood functions. Here, we propose to substitute in the BF the integrated likelihood with a suitable pseudo-likelihood of the parameter of interest only. A new formulation of the BF is derived, called pseudo-Bayes factor. The properties of the proposed pseudo-Bayes factors are investigated through two examples.

Keywords: Frequentist risk, Hypothesis testing, Marginal posterior distribution, Pseudo- likelihoods

References

BRAZZALE, A.R., DAVISON, A.C. and REID, N. (2007): *Applied Asymptotics*. Cambridge University Press, Cambridge.

CHEN, M., SHAO, Q.M. and IBRAHIM, J.G. (2000): *Monte Carlo Methods in Bayesian Computation*. Springer, New York.

GRECO, L., RACUGNO, W. and VENTURA, L. (2008): Robust likelihood functions in Bayesian inference. *Journal of Statistical Planning and Inference 138 (5), 1258–1270*.

KASS, R.E. and RAFTERY, A.E. (1995): Bayes factors. *Journal of the American Statistical Association 90 (430), 773–795*.

KOTZ, S., LUMELSKII, Y. and PENSKY, M. (2003): *The Stress-Strength Model and its Generalizations*. World Scientific, Singapore.

PACE, L., SALVAN, A. and VENTURA, L. (2006): Likelihood based discrimination between separate scale and regression models. *Journal of Statistical Planning and Inference 136 (10), 3539–3553*.

SEVERINI, T.A. (1999): On the relationship between Bayesian and non-Bayesian elimination of nuisance parameters. *Statistica Sinica 9 (3), 713–724*.

SEVERINI, T.A. (2000): *Likelihood Methods in Statistics*. Oxford University Press.

VENTURA, L., CABRAS, S. and RACUGNO, W. (2009): Prior distributions from pseudo-likelihoods in the presence of nuisance parameters, *Journal of the American Statistical Association 104 (486), 768–777*.

VENTURA, L., CABRAS, S. and RACUGNO, W. (2010): Default prior distributions from quasi- and quasi-profile likelihoods. *Journal of Statistical Planning and Inference* to appear.

Contributions to Bayesian Structural Equation Modeling

Séverine Demeyer[1 2], Nicolas Fischer[1], and Gilbert Saporta[2]

[1] LNE, Laboratoire National de Métrologie et d'Essais
29 avenue Roger Hennequin, 78197 Trappes, France, *severine.demeyer@lne.fr*
[2] Chaire de statistique appliquée & CEDRIC, CNAM
292 rue Saint Martin, Paris, France

Abstract. Structural equation models (SEMs) are multivariate latent variable models used to model causality structures in data. A Bayesian estimation and validation of SEMs is proposed and identifiability of parameters is studied. The latter study shows that latent variables should be standardized in the analysis to ensure identifiability. This heuristics is in fact introduced to deal with complex identifiability constraints. To illustrate the point, identifiability constraints are calculated in a marketing application, in which posterior draws of the constraints are derived from the posterior conditional distributions of parameters.

Keywords: structural equation modeling, Bayesian statistics, Gibbs sampling, latent variables, identifiability

References

BOX, G. E. P. and TIAO G.C. (1973) : *Bayesian Inference in Statistical Analysis (Wiley Classics Library)*. Wiley.

GELMAN, A., MENG, X. L. and STERN, H. (1996) : Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica 6, 733-807*.

GELMAN, A., CARLIN, J.B., STERN, H.S., RUBIN, D.B. (2004) : *Bayesian Data Analysis (Texts in Statistical Science)*. Chapman & Hall/CRC.

LEE, S. Y. (2007) : *Structural Equation Modelling: A Bayesian Approach (Wiley Series in Probability and Statistics)*. Wiley.

PALOMO, J., DUNSON, D. B. and BOLLEN, K. (2007) : Bayesian Structural Equation Modeling. In: S. Y. Lee (Ed): *Handbook of latent variable and related models*. Elsevier, 163–188.

TANNER, M.A., WONG, W.H. (1987) : The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association 82, 528–540*.

Nonlinear Regression Model of Copper Bromide Laser Generation

Snezhana Georgieva Gocheva-Ilieva[1] and Iliycho Petkov Iliev[2]

[1] Department of Applied Mathematics and Modelling, University of Plovdiv
24 Tzar Assen Str., 4000 Plovdiv, Bulgaria, *snow@uni-plovdiv.bg*
[2] Department of Physics, Technical University of Sofia, branch Plovdiv
25 Tzanko Djusstabanov Str., 4000 Plovdiv, Bulgaria, *iliev55@abv.bg*

Abstract. The focus of this study is on the relationship between the output laser power and basic laser input variables in copper bromide vapour laser with wavelengths of 510.6 and 578.2 nm. Based on experimental data, a nonlinear regression model has been constructed. To deal with the multicolinearity the initial predictors were grouped in three PCA factors. The transformation of factors by the Yeo-Johnson transformation was applied (see Yeo and Johnson (2000)). A compact *Mathematica* code for calculating the model parameters is presented. The model has been validated using independent evaluation data sets. The results obtained via the model allow for a more thorough analysis of relationship between the most important laser parameters in order to improve further experiments planning and laser production technology.

Modeling was carried out based on experimental data obtained at the Laboratory of Metal Vapour Lasers with the Georgi Nadjakov Institute of Solid State Physics, Bulgarian Academy of Sciences. The models have been obtained using the statistical package SPSS and *Mathematica* software.

Keywords: Yeo-Johnson transformation, PCA factors, nonlinear regression, output laser power

References

YEO, I. K. and JOHNSON, R. A. (2000): A new family of power transformations to improve normality or symmetry. *Biometrika, Oxford Press 87 (4), 954-959.*

On multiple-case diagnostics in linear subspace method

Kuniyoshi Hayashi[1], Hiroyuki Minami[2] and Masahiro Mizuta[2]

[1] Graduate School of Information Sciences and Technology, Hokkaido University, N14W9, Kita-ku, Sapporo, JAPAN, *k-hayashi@iic.hokudai.ac.jp*
[2] Information Initiative Center, Hokkaido University, N11W5, Kita-ku, Sapporo, JAPAN, *min@iic.hokudai.ac.jp, mizuta@iic.hokudai.ac.jp*

Abstract. In this paper, we discuss sensitivity analyses in linear subspace method,@ especially multiple-case diagnostics.

Linear subspace method proposed by Watanabe (1973) is a useful discriminant method in the field of pattern recognition. We have proposed its sensitivity analyses, with single-case diagnostics and multiple-case diagnostics with PCA.

We propose a modified multiple-case diagnostics using clustering and discuss its effectiveness with numerical simulations.

Keywords: CLAFIC, Sensitivity analysis, Perturbation

References

Watanabe, S. and Pakvasa, N. (1973): Subspace method of pattern recognition, *Proceedings of 1st International Joint Conference of Pattern Recognition, 25-32.*

"Made in Italy" Firms Competitiveness: A Multilevel Longitudinal Model on Export Performance

Matilde Bini[1] and Margherita Velucchi[2]

[1] Università Europea di Roma, Via degli Aldobrandeschi 190 - 00163 Roma, Italy *mbini@unier.it, bini@ds.unifi.it*
[2] European University Institute, Badia Fiesolana, Via dei Roccettini 9, 50014, San Domenico di Fiesole, Firenze, Italy *margherita.velucchi@eui.eu*

Abstract. The competitiveness of the Italian industrial system during the last decade has shown a strong slowdown. To compete in international markets, Italian firms reduced their costs instead of fostering on innovation and investments, being largely influenced by small size. Only the so-called "Made in Italy" sectors succeeded in international markets. To analyze this phenomenon, we investigate, at firm and sector level, factors affecting export competitiveness in "Made in Italy" sectors using a multilevel longitudinal model in the period 1999-2005. We find that "Made in Italy" role in international markets strongly depends on firms' geographical location and sector of activity and on their innovative capacity and productivity.

Keywords: Competitiveness, Made in Italy, multilevel models

A Fast Parsimonious Maximum Likelihood Approach for Predicting Outcome Variables from a Large Number of Predictors

Jay Magidson

Statistical Innovations Inc.
Belmont, Massachusetts, United States *jay@statisticalinnovations.com*

Abstract. A new model with K correlated components is presented for predicting outcome variables where the number of predictors G may exceed the total sample size N. A fast maximum likelihood algorithm provides closed-form expressions for the model parameters and statistical tests for determining the number of components. We also propose a fully integrated step-down variable selection algorithm, at each step eliminating the least important predictor based on a new measure of importance. Validated results from 2 examples suggest that the methods can provide good predictions outside the sample, especially with K = 3 or 4.

Keywords: correlated component regression, variable selection, gene expression, high dimensional data, latent class analysis

Multidimensional Exploratory Analysis of a Structural Model Using a Class of Generalized Covariance Criteria

Xavier Bry[1], Thomas Verron[2], and Patrick Redont[1]

[1] I3M, UM2, Place Eugène Bataillon, 34095 Montpellier, France
[2] SEITA-ITG, SCR, 4 rue André Dessaux, 45404 Fleury les Aubrais, France

Abstract. Our aim is to explore a structural model: several variable groups describing the same observations are assumed to be structured around latent dimensions that are linked through a linear model that may have several equations (Jöreskog and Wold (1982); Lohmöller (1989); Smilde et al. (2000)). This type of model is commonly dealt with by methods assuming that the latent dimension in each group is unique. However, conceptual models generally link concepts which are multidimensional. We propose a general class of criteria suitable to measure the quality of a Structural Equation Model (SEM). This class contains the covariance criteria used in PLS Regression and the Multiple Covariance criterion of the SEER method (Bry et al. (2009)). It also contains quartimax-related criteria. All criteria in the class must be maximized under a unit norm constraint. We give an equivalent unconstrained maximization program, and algorithms to solve it. This maximization is used within a general algorithm named THEME (Thematic Equation Model Exploration), which allows to search the structures of groups for all dimensions useful to the model. THEME extracts locally nested structural component models.

Keywords: Path Modeling, PLS, SEER, SEM, THEME.

References

BRY X., VERRON T., CAZES P. (2009): Exploring a physico-chemical multi-array explanatory model with a new multiple covariance-based technique: Structural equation exploratory regression, *Anal. Chim. Acta, 642 (2009) 45–58.*

JÖRESKOG, K. G. and WOLD, H. (1982): The ML and PLS techniques for modeling with latent variables: historical and competitive aspects, *in Systems under indirect observation, Part 1, 263–270.*

LOHMÖLLER J.-B. (1989): Latent Variables Path Modeling with Partial Least Squares, *Physica-Verlag, Heidelberg.*

SMILDE, A.K., WESTERHUIS, J.A., BOQUÉE, R., (2000): Multiway multiblock component and covariates regression models. *J. Chem. 14, 301–331.*

Boosting a Generalized Poisson Hurdle Model

Vera Hofer[1]

Department of Statistics and Operations Research, University of Graz, Universitaetsstraße 15/E3, 8010 Graz, Austria, *vera.hofer@uni-graz.at*

Abstract. Common boosting techniques are based on estimating one ensemble by means of gradient descent. Count data regressions by means of a generalised Poisson hurdel model consist of three different parameters. Fitting regression functions to all three parameters raises the question how to used boosting techniques. Since a triple of inter-related ensembles ought to be determined, the gradient of the loss is a 3-component vector. A boosting method for this hurdle model using multivariate componentwise least squares is introduced.

Keywords: boosting, count data, triple of ensembles

Quasi-Maximum Likelihood Estimators for Threshold ARMA Models: Theoretical Results and Computational Issues

Marcella Niglio[1] and Cosimo Damiano Vitale[2]

[1] Department of Economics and Statistics
Via Ponte Don Melillo, Fisciano (SA), Italy *mniglio@unisa.it*
[2] Department of Economics and Statistics
Via Ponte Don Melillo, Fisciano (SA), Italy *vitale@unina.it*

Abstract. In this paper we derive quasi-maximum likelihood estimators for the parameters of the threshold autoregressive moving average process (TARMA). After the presentation of the model, we discuss some property that makes this model of interest in most empirical domains. The derivation of the estimators is proposed in details and computational issues are examined in a simulation study.

Keywords: threshold model, Q-ML estimators, parameters initialization

Continuous Wavelet Transform and the Annual Cycle in Temperature and the Number of Deaths*

Milan Bašta[1], Josef Arlt[1], Markéta Arltová[1], and Karel Helman[1,2]

[1] Dept. of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic *milan.basta@vse.cz*
[2] Czech Hydrometeorological Institute

Abstract. The continuous wavelet transform applied to one time series allows the analysis of the temporal evolution and changes of the frequency content of this time series. The application of the cross-wavelet transform to two time series may reveal a complex relationship between the two time series - specifically, a relationship that differs from one frequency range to another and that is transient or evolves in time. As such, the wavelet transform is an intriguing tool for the analysis of demographic time series. In this paper we apply it to the analysis of the daily time series of the number of deaths due to cardiovascular diseases in Prague, Czech Republic and the daily time series of the average temperature in Prague, Czech Republic.

Keywords: wavelets, demography, time series, death rate, temperature

References

GRINSTED, A., MOORE, J. and JEVREJEVA, S. (2004): Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes in Geophysics 11, 561-566*

MEYERS, S., KELLY, B. and O'BRIEN, J. (1993): An introduction to wavelet analysis in oceanography and meteorology: With application to the dispersion of Yanai waves. *Mon. Weather Rev. 121, 2858-2866*

PERCIVAL, D. and WALDEN, A. (2000): *Wavelet Methods for Time Series Analysis. 1 edition.* Cambridge University Press.

TORRENCE, C. and COMPO, G. (1998): A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society 79, 61-78*

TORRENCE, C. and WEBSTER, P. (1999): Interdecadal changes in the ENSO-Monsoon System. *Journal of Climate 12 (8), 2679-2690*

* The paper was written with the support of the Grant Agency of the Czech Republic No. 402/09/0369, Modelling of Demographic Time Series in the Czech Republic

Empirical Mode Decomposition for Trend Extraction: Application to Electrical Data

Farouk Mhamdi[1], Mériem Jaïdane-Saïdane[1], and Jean-Michel Poggi[2,3]

[1] Unité Signaux et Systèmes, ENIT,
Farouk.Mhamdi@enit.rnu.tn, *meriem.jaidane@enit.rnu.tn*
[2] Université Paris-Sud, Mathématiques Bât. 425, 91405 Orsay, France
jean-michel.poggi@math.u-psud.fr
[3] Université Paris Descartes, France

Abstract. This paper presents a method for trend extraction from seasonal time series through the Empirical Mode Decomposition (EMD). Experimental comparison of trend extraction based on EMD and Hodrick Prescott filter are conducted. First results proved the eligibility of EMD trend extraction. Tunisian real peak load is finally used to illustrate the extraction of the intrinsic trend.

Keywords: Empirical Mode Decomposition, Trend extraction, Electrical data

References

ALEXANDROV, T., BIANCONCINI, S., BEE DAGUM, E., MAASS, P. and MCELROY, T. (2009): A Review of Some Modern Approaches to the Problem of Trend Extraction. *Research Report Series, Statistics 2008-3, U.S. Census Bureau, Washington, D.C.*.

FLANDRIN, P., GONCALVES, P. and RILLING, G. (2004): Detrending and Denoising with Empirical Mode Decomposition. *EUSIPCO 2004. September 6-10, Vienna, Austria.*

HUANG, N.E., SHEN, Z., LONG, S.R., WU, M.C., SHIH, H.H., ZHENG, Q., YEN, N., TUNG, C.C., and LIU, H.H. (1998): The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis. *Royal Society London, 903–995.*

OULD MOHAMED MAHMOUD, M., MHAMDI, F., JAIDANE-SAIDANE, M. (2009). Long Term Multi-Scale Analysis of the Daily Peak Load Based on the Empirical Mode Decomposition. *IEEE PowerTech, june 28-july 2, Romania.*

POLLOCK, DSG. (2003). Sharp filters for short sequences. *Journal of Statistical Planning and Inference 113, 663-683.*

SULING, J., YANQIN, G., QIANG, W. and JIAN, Z. (2009): Trend Extraction and Similarity Matching of Financial Time Series Based on EMD Method. *World Congress on Engineering and Computer Science,San Francisco, 20-22 Oct 2009.*

WU, Z., HUANG, N.E., LONG S.R. and PENG, C.K (2007): On the trend, detrending, and variability of nonlinear and nonstationary time series. *PNAS September 18, vol. 104, no. 38, 14889-14894.*

Comparing Two Approaches to Testing Linearity against Markov-switching Type Non-linearity

Jana Lenčuchová, Anna Petričková and Magdaléna Komorníková

Department of Mathematics, Faculty of Civil Engineering, Slovak University of Technology Bratislava
Radlinského 11, 813 68 Bratislava, Slovakia,
lencuchova@math.sk, petrickova@math.sk and *magda@math.sk*

Abstract. In this paper we discuss an alternative approach to testing linearity against Markov-switching type non-linearity. We aim to avoid the classic testing via the likelihood ratio test, which doesn't have a standard distribution. That's why time-consuming simulations must be carried out. We suggest the classical test to be substituted by using Hamilton's dynamic specification test for validity of Markov assumptions. The same idea will be applied to testing the remaining non-linearity to compare 2-regime with 3-regime models. These two approaches will be confronted by being demonstrated on some selected time series, e.g. Slovak macro-economic indicators and some exchange rates.

Keywords: Markov-switching model, Markov assumptions, dynamic specification test, testing non-linearity, testing remaining non-linearity

References

HAMILTON, J. D. (1994): *Time series analysis*. Princeton University Press, Princeton.

HAMILTON, J.D. (1996): Specification testing in Markov-switching time series models. *Journal of Econometrics 70, 127-157.*

HANSEN, B.E. (1992): The likelihood ratio test under nonstandard assumptions: testing the Markov switching model of GNP. *Journal of Applied Econometrics 7, 61-82.*

NEWEY, W.K. (1985): Maximum likelihood specification testing and conditional moment tests. *Econometrica 53, 1047-1070.*

TAUCHEN, G. (1985): Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics 30, 415-443.*

WHITE, H. (1987): Specification testing in dynamic models. In: T. F. Bewley (Eds.): *Advances in econometrics*. Fifth world congress, Cambridge University Press, Cambridge, Vol. 2.

Acknowledgement The support of the grant APVV No. LPP-0111-09 is kindly announced.

Polynomial Methods in Time Series Analysis

Félix Aparicio-Pérez

Instituto Nacional de Estadística
Castellana 183, 28046, Madrid, Spain *fapape@ine.es*

Abstract. Polynomial methods have been extensively used in system theory either as an alternative or in conjunction with state space methods, e.g. Kailath (1980), Chen (1984). However, its use in time series analysis has been very limited. This paper highlights the main results in matrix polynomial algebra, like the solution of matrix polynomial equations or the transformation between a right and a left matrix fraction description and provides some mostly unknown applications in time series analysis.

One of these applications is the evaluation of the autocovariances of a VARMA process by a method that is more efficient than the methods that are used in time series analysis. This method is based on solving a so-called symmetric matrix polynomial equation. The second application allows the automatic computation of the model that follows a filtered VARMA process using a new method. This problem is usually solved in time series analysis by means of lengthy ad-hoc hand computations, while the method proposed in the paper obtains the model by transforming a right matrix fraction description into a left one and then doing a spectral factorization. The third, and also new, technique allows the exact computation of a multivariate Wiener-Kolmogorov filter based on a finite sample for a general VARMA process. To do so it first computes the adjoint of a polynomial matrix and then solves a spectral factorizarion problem to obtain the result in the form of three cascaded filters. The implementation of these filters requires the use of some additional techniques, like the time-reversion or the obtention of echelon realizations of a VARMA process.

All the polynomial techniques that are needed for these (and other) applications can be implemented using numerically reliable and efficient techniques, and the paper provides some references where it is explained how to do so.

Keywords: Polynomial Matrices, Time Series, Wiener-Kolmogorov Filter, Autocovariances

References

CHEN, C.T. (1984): *Linear System Theory and Design*. Holt, Rinehart and Winston.

KAILATH, T., (1980): *Linear Systems*. Prentice-Hall, Englewood Cliffs, N.J.

Analysis of Binary Longitudinal Responses

M.Helena Gonçalves[1] and M.Salomé Cabral[2]

[1] Centro de Estatística e Aplicações da Universidade de Lisboa, Departamento de Matemática, FCT, Universidade do Algarve, Portugal, *mhgoncal@ualg.pt*

[2] Centro de Estatística e Aplicações da Universidade de Lisboa, Departamento de Estatística e Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa, Portugal, *salome@fc.ul.pt*

Abstract. The aim of this work is the analysis of binary longitudinal responses from the point of view of likelihood inference, which requires complete specification of a stochastic model for the individual profile. In the context of binary responses, dependence is more conveniently measured by odds ratios rather than correlations. In our formulation the parameter of interest is the marginal probability of success, which is related to the covariates via a logistic regression model. The dependence structure of the process corresponds to a second order Markov chain. Markov chains provide the simplest stochastic mechanism to introduce serial dependence for discrete random variables. Besides serial dependence, another important source of dependence among data from one given subject is the presence of individual random effects. Random effects are also considered using exact maximum likelihood estimation via numerical integration (Gonçalves and Azzalini (2008)). To illustrate this methodology, we considered the data analyzed by MacDonald and Raubenheimer (1995) about the locomotory behaviour of 24 locusts (*locusta migratoria*) observed at 161 time points. The focus is on the problem of comparison between treatment groups.

Keywords: binary longitudinal data, exact likelihood, serial dependence, random effects.

References

GONÇALVES, M. HELENA and AZZALINI, A. (2008): Using Markov chains for marginal modelling of binary longitudinal data in an exact likelihood approach. *Metron, vol LXVI, 2, 157-181.*

MACDONALD, I. and RAUBENHEIMER, D. (1995): Hidden Markov models and animal behaviour. *Biometrical Journal, 37, 701-712.*

On the Identification of Predictive Biomarkers: Detecting Treatment-by-Gene Interaction in High-Dimensional Data

Wiebke Werft and Axel Benner

Department of Biostatistics, German Cancer Research Center
INF 280, Heidelberg, Germany, *w.werft@dkfz.de, benner@dkfz.de*

Abstract. In recent years, special interest has been placed on the development of biomarkers that are predictive of a patient's response to treatment. Predictive markers can guide the choice of most successful therapies for the benefit of the patient (Simon (2008)).
As outcome variable we will consider a binary endpoint and marker data is supposed to be continuous, e.g. gene expression data. For the identification of predictive markers we therefore apply a logistic regression model with interaction term treatment times marker expression. Adjustment for multiple testing is cruicial when testing several potential markers simultaneously, e.g. in situations with thousands candidate genes (e.g. gene expression microarray data).
By simulation studies the issue of sample-size determination for the identification of predictive biomarkers in high-dimensional situations will be analysed. The performance of different multiple testing procedures for control of the false discovery rate will be compared including resampling-based approaches as described by Dudoit and van der Laan (2007). These resampling-based multiple testing procedures are not limited by the so called subset pivotality assumption which is not fullfilled in case of logisitc regressions. Moreover, the use of fractional polynomials for modelling interactions between treatment and continuous covariates was introduced by Royston and Sauerbrei (2008). This method will be discussed in comparison to the gene-wise logistic regression model.

Keywords: predictive biomarker, treatment by gene interaction, false discovery rate, resampling-based multiple testing procedures

References

DUDOIT, S. and VAN DER LAAN, M. (2007): *Multiple Testing Procedures with Applications to Genomics*. Springer Series in Statistics, Springer, Berlin.

ROYSTON, P. and SAUERBREI, W. (2008): Interactions Between Treatment and Continuous Covariates: A Step Toward Individualizing Therapy. *Journal of Clinical Oncology 26 (9), 1397-1399.*

SIMON, R. (2008): The use of Genomics in Clinical Trial Design. *Clinical Cancer Research 14 (19), 5984-5993.*

Hidden Markov models for DNA sequence segmentation modeling : Change-Point Identification

Darfiana Nur[1] and Kerrie L. Mengersen[2]

[1] School of Mathematical and Physical Sciences
University of Newcastle, Australia, *Darfiana.Nur@newcastle.edu.au*
[2] School of Mathematical Sciences
Queensland University of Technology, Australia, *k.mengersen@qut.edu.au*

Abstract. Many genome sequences display heterogeneity in base composition in the form of segments with similar structure. Early evidence of segmental genomic structure was noticed early on that in the salivary glands of *Drosophila melanogaster* whereas the problem of statistically segmenting DNA sequence has a history about four decades. One approach describes DNA sequence structure by a hidden Markov model (HMM) [1,2]. Change-point detection is an identification of abrupt changes in the generated parameters of sequential data. It has proven to be useful in application such as DNA segmentation modeling. This talk focuses on the various change-point identification of a Bayesian hidden Markov model describing homogeneous segments of DNA sequences. A simulation study will be used to evaluate the change-points followed by the real-life examples.

Keywords: DNA sequence; Bayesian hidden markov model; change point

References

NUR, D., ALLINGHAM, D., ROUSSEAU, J., MENGERSEN, K.L. and McVINISH, R. (2009): Bayesian analysis of DNA sequences segmentation : A prior sensitivity analysis. *Computational Statistics and Data Analysis 53, 1873-1882.*

BOYS,R.J. and HENDERSON,D.A. (2004): A Bayesian approach to DNA sequence segmentation. *Biometrics 60, 573-588.*

Dating Mining for Genomic-Phenomic Correlations

Joyce Niland and Rebecca Nelson

City of Hope, 1500 East Duarte Road, Duarte, CA, USA *jniland@coh.org*

Abstract. Realization of personalized medicine depends on the ability to correlate phenomic (biologic) data with synthesized genomic results. An optimal mode to achieve this goal is through a research data warehouse, blending tissue sample data with information on diagnosis, treatment, and outcomes. We created such a warehouse to document over 12,000 frozen tissue and 150,000 paraffin-embedded samples amassed since 1980 at City of Hope, following several warehousing best practices. The warehouse is jointly managed by the Department of Information Technology Services (ITS), to provide Extract-Transform-Load (ETL) functions for data from our electronic medical record and research systems, and the Department of Information Sciences (DIS), whose analysts provide data validation, mining, and cohort assembly for investigators.

Further, we maintain international coding and quality standards, so that collaborations with investigators worldwide will be facilitated. (Dubois (2002)) Attribute-centric queries to identify sets of patients based on a Boolean combination of parameters is challenging, and requires complete documentation of the metadata ("data about the data"). (Deshpande et al. (2002)) We have adopted a comprehensive metadata repository to store data definitions and code lists, so the warehouse can be accurately mined by users.

Via the warehouse, we have efficiently identified numerous targeted cohorts of patients with available tissue samples and correlated clinical data for analysis. We have developed a robust "honest broker" process, critical for the authorization of data and sample access and correlation. (Dhir et al. (2008)) This presentation describes our warehousing efforts, our cohort identification projects, the honest broker algorithms, and challenges and successes to date.

Keywords: data mining, data warehouse, genomic-phenomic correlations, honest broker process

References

DUBOIS, L. (2002): The ROI of data quality: Six business cases for a data quality solution. *Journal of Data Warehousing 7 (3), 24-33.*

DESHPANDE, A.M., BRANDT, C., and NADKARNI, P.M. (2002): Metadata-driven ad hoc query of patient data. *Journal of the American Medical Informatics Association 9 (4), 369-381.*

DHIR R, PATEL AA, WINDERS S, BISCEGLIA M, SWANSON D, AAMODT R, and BECICH MJ. (2008): A multidisciplinary approach to honest broker services for tissue banks and clinical data: a pragmatic and practical model. *Cancer 113 (7), 1705-1715.*

Over-optimism in biostatistics and bioinformatics

Anne-Laure Boulesteix

Department of Medical Informatics, Biometry and Epidemiology
University of Munich, Marchioninistr. 15, 81377 Munich, Germany
boulesteix@ibe.med.uni-muenchen.de

Abstract. The problem of "over-optimistic" research findings has attracted a lot of attention in medical literature in the last few years. In this talk, I will present and discuss two different examples of over-optimistic results in the context of statistical bioinformatics.

The first example is classification based on high-dimensional data in biomedical studies. Evaluating several classification algorithms via cross-validation in a trial-and-error strategy and reporting only the lowest error rate yields a substantial optimistic bias. Correction is crucial but not trivial. I review some potential solutions from a practical point of view.

The second example concerns methodological biostatistics/bioinformatics research. When developing a new method, researchers naturally tend to optimize their new algorithm to the available data set(s). Based on the concrete example of classification for microarray data using external biological knowledge from KEGG, I show that the new method may consequently "overfit" the data used for its development. I also demonstrate the advantages of a strict validation of methodological research results based on independent data that were not used previously for the development of the method.

References

BOULESTEIX, A.-L. and STROBL, C. (2009): Optimal classifier selection and negative bias in error rate estimation: An empirical study on high-dimensional prediction. *BMC Medical Research Methodology, 9, 85.*

BOULESTEIX, A.-L (2010): Over-optimism in bioinformatics research. *Bioinformatics, 26, 437–439.*

JELIZAROW, M., GUILLEMOT, V., TENENHAUS, A., STRIMMER, K. and BOULESTEIX, A.-L (2010): Over-optimism in bioinformatics: an illustration. *Department of Statistics, Technical Report 81.*

Keywords

Keywords: classification, error rate estimation, bias, benchmarking, good practice, high-dimensional data, cross-validation, validation

μTOSS - An Open Software Solution for Multiple Hypotheses Testing

G. Blanchard[1], T. Dickhaus[2], N. Hack[3], F. Konietschke[4], K. Rohmeyer[5], J. Rosenblatt[6], M. Scheer[7] and W. Werft[8]

[1]Weierstrass Institute for Applied Analysis and Stochastics Berlin, Germany, [2]Humboldt-University Berlin, Germany, [3]Medical University of Vienna, Austria, [4]Georg-August-University, Göttingen, Germany, [5]Leibniz University Hannover, Germany, [6]Tel Aviv University, Israel, [7]German Diabetes Center, Düsseldorf, Germany, [8]German Cancer Research Center, Heidelberg, Germany, *w.werft@dkfz.de*

Abstract. Multiple hypotheses testing (MHT) has emerged as one of the most active research fields in statistics over the last 10-15 years, especially driven by large-scale applications such as genomics or proteomics. μTOSS is an open source, easy-to-extend software solution establishing a unified platform for a broad variety of MHT procedures. The μTOSS software has been realized in a month-long project sponsored by the PASCAL2 European Network of Excellence.
Basically, μTOSS consists of the two R packages `mutoss` and `mutossGUI`, the latter provides a graphical user interface to `mutoss`. It comprises MHT procedures controlling the family-wise error rate (single-step and stepwise rejective methods, resampling-based procedures), and the false discovery rate (FDR) (classical and data-adaptive frequentistic methods, Bayesian approaches, resampling-based techniques). Moreover, novel procedures not yet been implemented in any statistical software package have now been made available in μToss: multiplicity-adjusted simultaneous confidence intervals (Konietschke, 2009), procedures based on the asymptotically optimal rejection curve (Finner et al., 2009) and self-consistency methods for FDR control under arbitrary dependencies (Blanchard and Roquain, 2009).
For researchers, it features a convenient unification of interfaces for MHT procedures and helper functions facilitating the setup of benchmark simulations for comparison of competing methods. For end users, e.g. in clinical practice, a graphical user interface and an online users guide help to identify appropriate adjustment methods for a specified multiple testing problem. Ongoing maintenance and subsequent extensions of novel research developements could establish μTOSS as the state of the art MHT software for the future.

Keywords: Multiple Testing, Multiple Comparisons, open-source software

References

BLANCHARD, G. and ROQUAIN, E. (2009): Adaptive FDR control under independence and dependence. *Journal of Machine Learning Research 10, 2837-2871.*

Data Mining for Population-Based Studies

Stanley P. Azen, Katherine J. Sullivan, Julie K. Tilson, Steven Y. Cen, Jiaxiu He, and Cheryl Vigen

University of Southern California, 1540 Alcazar St., Los Angeles, CA, USA
sazen@usc.edu

Abstract. The Department of Preventive Medicine at the University of Southern California has received substantial NIH grant funding to conduct population-based studies and multicenter clinical trials. As large data repositories are developed to address the specific aims of the studies, this provides opportunities to "mine" the data to identify important new clinical and public health findings. Examples are:

The Los Angeles Latino Eye Study (LALES) is a population-based study in 6,082 Latinos aged 40+ yrs. Prevalence of undiagnosed glaucoma was surprisingly large (75%). Classification and regression tree (CART) analysis created a multivariate algorithm for glaucoma screening. Overall sensitivity and specificity was improved using the CART cutpoints and branch-related diagnostic criteria.

The Locomotor Experience Applied Post Stroke (LEAPS) is a multi-site randomized trial evaluating strategies to improve gait speed in stroke patients. Because the minimal clinical important difference (MCID) in gait speed post-stroke is unknown, we mined the LEAPS database using CART to determine the MCID for improved gait speed. Using a standardized quality of life, we found that optimal gait-speed MCID is 0.16m/s. (Tilson et al (2010))

The Hormonal Regulators of Muscle & Metabolism in Aging (HORMA) study is a randomized trial to test the effect of supplemental testosterone and growth hormone in elderly men. There was individual variability in improved lean body mass (LBM), appendicular skeletal muscle mass (ASMM), muscle strength and physical function. (Sattler et al (2009)) Pathway analyses showed that to enhance muscle strength and physical function, improvements in LBM and ASMM are needed, via testosterone.

Keywords: Data mining, biostatistics, clinical trials, epidemiology

References

SATTLER FR, CASTANEDA-SCEPPA C, BINDER EF, SCHROEDER ET, WANG Y, BHASIN S, KAWAKUBO M, STEWART Y, YARASHESKI KE, ULLOOR J, COLLETTI P, ROUBENOFF R, AZEN SP (2009): Testosterone and growth hormone improve body composition and muscle performance in older men. *Journal of Clinical Endocrinology and Metabolism 94, 1991-2001.*

TILSON JK, SULLIVAN KJ, CEN SY, ROSE DK, KORADIA C, AZEN SP, DUNCAN PW.: Meaningful gait speed improvement during the first 60 days post-stoke: Minimal clinically important difference. *Physical Therapy Journal 90 (2), 1-13.*

Integrating biological knowledge related to co-expression when analysing *Xomic* data

Marie Verbanck[1] and Sébastien Lê[1]

Agrocampus Ouest & IRMAR, UMR 6625 du CNRS
65 rue de St-Brieuc, 35042 Rennes, France
marie.verbanck@orange.fr, *sebastien.le@agrocampus-ouest.fr*

Abstract. Interpreting results provided by multivariate exploratory methods (such as Principal Component Analysis for instance) applied on genomic data is almost impossible at a gene level due to the number of genes. Integrative approaches which involve the incorporation of biological knowledge have become unavoidable. De Tayrac et al. (2009) proposed a strategy which allows to use an *a priori* information, such as Gene Ontology (GO) or Kegg terms to enhance their results. The idea consists in constituting modules of genes according to the *a priori* information and using those modules as a supplementary information in order to interpret results on the basis of the genes' functions.

However, the composition of those modules may be disconnected from the structure of the genomic data to be studied and does not consider the different degrees of specificity of the terms which convey the existence of different levels of regulation. Hence appears the natural idea of improving the way modules are constituted.

The aim of this talk is to propose a new approach combining Canonical Correspondence Analysis with Hierarchical Multiple Factor Analysis (Francoa et al., 2009) to get modules that have two main features: 1) they are constituted of genes that belong to the same biological processes; 2) they are constituted of genes that are co-expressed with respect to the data set of interest.

The interpretation of the biological processes is thus facilitated by the co-expression of the genes within a group, whereas the method highlights a few key-genes whose functions can be easily taken into account to go deeper into the interpretation. An application of this method to a chicken microarray data set has allowed to bring out the well-known mechanisms implemented in reply to fasting, and to come up with new trails.

Keywords: transcriptomic data, integration of biological knowledge, Canonical Correspondence Analysis, Hierarchical Multiple Factor Analysis

References

DE TAYRAC M., LÊ S., AUBRY M., MOSSER J., HUSSON F. (2009): Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics 2009, 10:32.*

FRANCOA J., CROSSAB J., DESPHANDEC S. (2009): Hierarchical Multiple-Factor Analysis for Classifying Genotypes Based on Phenotypic and Genetic Data. *Crop Science 50(1):105-117.*

Additional Hierarchy in the Modelling of Meta-analysis Data

Elizabeth Stojanovski[1] and Kerrie Mengersen[2]

[1] University of Newcastle, Australia, *elizabeth.stojanovski@newcastle.edu.au*
[2] Queensland University of Technology, Australia

Abstract. The random-effects model in the frequentist framework assumes study effects to be randomly sampled from a common distribution. Associated parameters allow further variation between studies compared to fixed-effects models. The quantities of most interest are typically the hyperparameters. In the case that effect variability is small within a population, more borrowing of information occurs across studies.

The variability of an effect is often estimated using a method of moments approximation proposed by DerSimonian and Laird [1986]. This method is assessed in greater detail.

Keywords: meta-analysis, Bayesian, random-effects

References

DerSIMONIAN, R., LAIRD, N. (1986): Meta-analysis in clinical trials. *Controlled Clinical Trials 7, 177-188.*

Bayesian Modelling of Cross-study Discrepancies in Gene Networks

Xiaodan Fan

Department of Statistics, the Chinese University of Hong Kong
Shatin, N.T., Hong Kong, *xfan@sta.cuhk.edu.hk*

Abstract. There are often multiple studies performed to investigate a same biological system from similar or related angles due to its high complexity. Many meta-analyses over these studies suggested that there are an excess of genes showing discordant gene expression across similar studies compared to what would have been predicted by chance alone. Scharpf et al. (2009) introduced a hierarchical Bayesian model for detecting differential gene expression in multiple data sets while allowing for cross-study discrepancy. Fan et al. (2009) and Fan et al. (2010) used a Bayesian approach to integrate cell-cycle microarray data sets and showed that the discrepancy about the cell-cycle regulated genes exists between individual laboratories and across synchronization techniques. In this paper, instead of dealing with the discrepancy at gene level as in Scharpf et al. (2009), we introduced a Bayesian approach to model the discrepancy at gene network level. The fundamental conjecture is that the gene expression discrepancy is resulted from the dynamics of the gene regulatory networks. Starting with different parameter settings, the network dynamics may show multiple steady states. Therefore, a gene can be highly expressed in one phenotype than the other in some studies, while the opposite is observed in other studies. Similarly, in cell-cycle experiments, a gene's expression can be highly periodic in some studies, while aperiodic in other studies. This phenomenon also exists in some stress response studies, where the lists of differential expressed genes for the same stimulus vary significantly across different study. The new Bayesian approach is applied on the time-series microarray data sets from fission yeast cell-cycle experiments. A gene network is inferred from the combined data. Its dynamics is simulated under the inferred parameter setting as well as other settings as an effort to explain the discrepancy observed in the cell cycle studies.

Keywords: gene network, meta-analysis, Bayesian computing, cell cycle

References

FAN, X. and LIU, J.S. (2009): Comment on "A Bayesian Model for Cross-Study Differential Gene Expression" by Scharpf, Tjelmeland, Parmigiani, and Nobel, *Journal of the American Statistical Association 104 (488), 1314-1318.*

FAN, X., PYNE, S. and LIU, J.S. (2010): Bayesian Meta-Analysis for Identifying Periodically Expressed Genes in Fission Yeast Cell Cycle. To appear in *Annals of Applied Statistics.*

SCHARPF, R.B., TJELMELAND, H., PARMIGIANI, G. and NOBEL, A. (2009): A Bayesian model for cross-study differential gene expression. *Journal of the American Statistical Association 104 (488), 1295-1310.*

Mixture models of truncated data for estimating the number of species.

Sebastien Li-Thiao-Té[1], Jean-Jacques Daudin[1], and Stéphane Robin[1]

UMR 518 AgroParisTech / INRA MIA
16 rue Claude Bernard, F-75231 Paris Cedex 05,
sebastien.li-thiao-te@agroparistech.fr

Abstract. Metagenomics goes beyond DNA sequencing by tackling communities of microorganisms in their natural environment. Previously, each microbial strain needed to be cultured before sequencing. Applying DNA sequencing directly to the sample has revealed the great diversity of the microbial populations in soil, sea water or the intestinal flora.

Even though many new species can be studied, many more remain unobserved in the collected data. Estimating the total number of microbial species in the biological sample and their abundance distribution is key to determining the number of sequencing runs needed.

In the standard model introduced by Fisher et al. (1943), each species contribute a Poisson-distributed number of individuals to the dataset, with a species-specific abundance parameter. Unobserved species are those that contribute zero individuals. Mixture models provide flexible models for the distribution of the abundance parameters.

Following Bunge and Barger (2008), we use a truncated mixture model of geometric distributions. We propose to perform parameter estimation in the Bayesian framework with a variational algorithm, Beal and Ghahramani (2003). In this work, the number of components is not selected and we use Bayesian model averaging to combine the estimates from all considered models. In particular, the variational framework provides an efficient way of computing the weights for each model.

Keywords: mixture models, bayesian model averaging, variational methods, truncation, metagenomics

References

BEAL, M. J. and GHAHRAMANI, Z. (2003): The variational Bayesian EM algorithm for incomplete data : with application to scoring graphical model structures. *Bayesian Statistics* 7 (pp. 453–464).

BUNGE, J. and BARGER, K. (2008): Parametric models for estimating the number of classes. *Biometrical Journal*, 50(5).

FISHER, R. A., CORBET, A. S., and WILLIAMS, C. B. (1943): The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population, *Journal of Animal Ecology*, 12, 42–58.

Sequential Monte Carlo techniques for MLE in plant growth modeling

Samis Trevezas[1,2] and Paul-Henry Cournède[2,1]

[1] INRIA Saclay Île-de-France, EPI DIGIPLANTE
91893 Orsay cedex, France, *samis.trevezas@ecp.fr*
[2] Ecole Centrale Paris, Laboratory of Applied Mathematics and Systems
92290, Châtenay-Malabry, France, *paul-henry.cournede@ecp.fr*

Abstract. Parametric identification in plant growth models, those that can be formalized as discrete dynamical systems, is a challenging problem due to specific data acquisition (system observation is supposed to be done with destructive measurements), non-linear dynamics, model uncertainties and high-dimensional parameter space. The general approach for parametric identification in dynamical models involves the use of a stochastic framework for the model and the observation equations. When the dynamical system that describes the state evolution is non-linear with gaussian noise, then Kalman filtering techniques can be used as approximation schemes. Nevertheless, when applied properly, sequential Monte-Carlo (or particle filter) methods offer a better alternative for state and parameter estimation (Doucet and Johansen (2008)). In this talk, we present how sequential Monte-Carlo methods can be used for maximum likelihood estimation via EM-type algorithm in plant growth modeling. In particular, we illustrate this method in a version of the functional-structural plant growth model, called GreenLab (Cournède et al. (2006)). The observed vector consists of organ masses, measured by censoring plant's evolution at a fixed observation time. The model hidden states represent biomasses produced at every growth cycle. Under some assumptions, we show that the estimation problem can be tackled in the framework of hidden (latent variable) models (Cappé et al. (2005)), where an appropriate bivariate stochastic process describes the variables of the system. We use sequential Monte-Carlo in order to approximate the non-explicit E-step in the EM-type algorithm, and parametric bootstrap in order to obtain approximate confidence intervals for the MLE.

Keywords: maximum likelihood estimation; parametric identification; plant growth model; sequential Monte-Carlo

References

DOUCET, A. and JOHANSEN, A.M. (2008): A tutorial on particle filtering and smoothing: fifteen years later. *Technical report, Department of Statistics, University of British Columbia.*

CAPPE, O., MOULINES E. and RYDEN T. (2005): *Inference in hidden Markov models.* Springer.

COURNEDE, P.H., KANG, M.Z., MATHIEU, A., BARCZI, J.F., YAN, H.P., HU, B.G. and DE REFFYE, P. (2006): Structural Factorization of Plants to Compute their Functional and Architectural Growth. *Simulation: 82(7).*

Evaluation of DNA Mixtures Accounting for Sampling Variability

Yuk-Ka Chung[1], Yue-Qing Hu[2], De-Gang Zhu[3], and Wing K. Fung[4]

[1] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China, *yukchung@hku.hk*
[2] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China, *yqhu@hku.hk*
[3] Nanjing Forestry University, Nanjing, China,
[4] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China, *wingfung@hku.hk*

Abstract. In the conventional evaluation of DNA mixtures, the allele frequencies are often taken as constants. But they are in fact estimated from a sample taken from a population and thus the variability of the estimates has to be taken into account. Within a Bayesian framework, the evaluation of DNA mixtures accounting for sampling variability in the population database of allele frequencies are discussed in this paper. The concise and general formulae are provided for calculating the likelihood ratio when the people involved are biologically related. The implementation of the formula is demonstrated on the analysis of a real example. The resulting formulae are shown to be more conservative, which is generally more favorable to the defendant.

Keywords: Bayesian inference, Hardy-Weinberg equilibrium, relatedness coefficient, likelihood ratio, mixed stain, relative

Variable selection and parameter tuning in high-dimensional prediction

Christoph Bernau[1] and Anne-Laure Boulesteix[1,2]

[1] Department of Medical Informatics, Biometry and Epidemiology
University of Munich, Marchioninistr. 15, 81377 Munich, Germany
bernau@ibe.med.uni-muenchen.de, boulesteix@ibe.med.uni-muenchen.de
[2] Department of Statistics, University of Munich, Ludwigstr. 33, 80539 Munich, Germany

Abstract. In the context of classification using high-dimensional data such as microarray gene expression data, it is often useful to perform preliminary variable selection. For example, the k-nearest-neighbors classification procedure yields a much higher accuracy when applied on variables with high discriminatory power. Typical (univariate) variable selection methods for binary classification are, e.g., the two-sample t-statistic or the Mann-Whitney test.

In small sample settings, the classification error rate is often estimated using cross-validation (CV) or related approaches. The variable selection procedure has then to be applied for each considered training set anew, i.e. for each CV iteration successively. Performing variable selection based on the whole sample before the CV procedure would yield a downwardly biased error rate estimate. CV may also be used to tune parameters involved in a classification method. For instance, the penalty parameter in penalized regression or the cost in support vector machines are most often selected using CV. This type of CV is usually denoted as "internal CV" in contrast to the "external CV" performed to estimate the error rate, while the term "nested CV" refers to the whole procedure embedding two CV loops.

While variable selection and parameter tuning have been widely investigated in the context of high-dimensional classification, it is still unclear how they should be combined if a classification method involves both variable selection and parameter tuning. For example, the k-nearest-neighbors method usually requires variable selection and involves a tuning parameter: the number k of neighbors. It is well-known that variable selection should be repeated for each external CV iteration. But should we also repeat variable selection for each it internal CV iteration or rather perform tuning based on fixed subset of variables? While the first variant seems more natural, it implies a huge computational expense and its benefit in terms of error rate remains unknown.

In this paper, we assess both variants quantitatively using real microarray data sets. We focus on two representative examples: k-nearest-neighbors (with k as tuning parameter) and Partial Least Squares dimension reduction followed by linear discriminant analysis (with the number of components as tuning parameter). We conclude that the more natural but computationally expensive variant with repeated variable selection does not necessarily lead to better accuracy and point out the potential pitfalls of both variants.

Keywords: class prediction, variable selection, parameter tuning, nested cross-validation, genomics

Learning Hierarchical Bayesian Networks for Genome-Wide Association Studies

Raphaël Mourad[1], Christine Sinoquet[2], and Philippe Leray[1]

[1] LINA, UMR CNRS 6241, Ecole Polytechnique de l'Université de Nantes, rue Christian Pauc, BP 50609, 44306 Nantes Cedex 3, France, {*raphael.mourad,philippe.leray*}*@univ-nantes.fr*

[2] LINA, UMR CNRS 6241, Université de Nantes, 2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 3, France, *christine.sinoquet@univ-nantes.fr*

Abstract. We describe a novel probabilistic graphical model customized to represent the statistical dependencies between genetic markers, or SNPs, in the Human genome. The motivation is to reduce the dimension of the data to be further submitted to statistical association tests with respect to diseased/non diseased status. Probabilistic graphical models offer an adapted framework for a fine modelling of dependencies between SNPs. Various models have been used for this purpose, including either Markov fields (Verzilli *et al.* (2006)) or Bayesian networks (Nefian (2006); Zhang and Ji (2009)). However, scalability remains a crucial issue. Our proposal generalizes a hierarchy-based framework designed by Hwang and collaborators (Hwang *et al.* (2006)). Our method relies on *forests* of hierarchical latent class models. A generic algorithm, CFHLC, has been designed to tackle the learning of both forest structure and probability distributions. A first implementation has been shown to be tractable on benchmarks describing 10^5 variables for 2000 individuals. Complementary results are further discussed in Mourad *et al.* (2010).

Keywords: Bayesian networks, hierarchical latent class models, data dimensionality reduction, genetic marker dependency modelling

References

HWANG, K.B., KIM, B.-H. and ZHANG, B.-T. (2006): Learning hierarchical Bayesian networks for large-scale data analysis. *ICONIP (1), 670-679.*

MOURAD, R., SINOQUET, C. and LERAY, P. (2010): Learning a forest of Hierarchical Bayesian Networks to model dependencies between genetic markers. *LINA, Research Report, hal-00444087.*

NEFIAN, A.V. (2006): Learning SNP dependencies using embedded Bayesian networks. *Computational Systems Bioinformatics Conference CSB'2006*, Stanford, USA, poster.

VERZILLI, C.J., STALLARD, N. and WHITTAKER, J.C. (2006): Bayesian graphical models for genomewide association studies. *The American Journal of Human Genetics 79(1), 100-112.*

ZHANG, Y. and JI, L. (2009): Clustering of SNPs by a Structural EM Algorithm. *International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, 147-150.*

Differentiation Tests for Three Dimensional Shape Analysis

Stefan Markus Giebel[1], Jens-Peter Schenk[2] and Jang Schiltz[1]

[1] Université du Luxembourg
4, rue Albert Borschette, L-1246 Luxembourg *jang.schiltz@uni.lu*
[2] Uniklinikum Heidelberg
Im Neuenheimer Feld 430, D69120 Heidelberg
jens-peter_schenk@med.uni-heidelberg.deu

Abstract. There are different kinds of tumours in childhood: nephroblastoma, clear cell sarcoma, neuroblastoma etc. The chosen therapy depends upon the diagnosis of the radiologist which is done with the help of MRI (Magnetic resonance images). Our research is the first mathematical approach on MRI of renal tumours (n=80). We are using transversal, frontal and sagittal images and compare their potential for differentiation of the different kind of tumours by use of Statistical Shape Analysis.
Statistical shape analysis is a methodology for analyzing shapes in the presence of randomness. It allows to study two- or more dimensional objects, summarized according to key points called landmarks, with a possible correction of size and position of the object. So objects with different size and/or position can be compared with each other and classified. To get the shape of an object without information about position and size, centralisation and standardisation procedures are used in some metric space. This approach provides an objective methodology for classification whereas even today in many applications the decision for classifying according to the appearance seems at most intuitive.
We determine the key points or three dimensional landmarks of the renal tumours by using the edges of the platonic body (C60). We present a new test for the mean shape based on the variance within tumour groups with the same diagnosis. Unlike the classical test from Ziezold (1994) we do not need any more a mean shape in each case for both groups is necessary for differentiation. Moreover, we apply Logistic regression and Configuration Frequency Analysis for classification on the sample. While Logistic regression handles the data in a continous type, Configuration Frequency Analysis uses discrete variables. Eventually we discuss the consequences of our results for the application in oncology.

Keywords: Statistical shape analysis, Shape differentiation, Renal tumours, Mean shape, Variance

References

GIEBEL, S., SCHENK, J.P. and Schiltz, J.(2010): Analysis on two dimensional objects in medicine: Differentiation of tumours in early childhood. *Proceedings of the the 20th international congress of Jangjeon Mathematical Society.*

ZIEZOLD, H.(1994): Mean Figures and Mean Shapes Applied to Biological Figure and Shape Distributions in the Plane. *Biometrical Journal 36, 491-510.*

On the Correlated Gamma Frailty Model for Bivariate Current Status Data

Niel Hens[1,2] and Andreas Wienke[3]

[1] Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Campus Diepenbeek, Agoralaan 1, 3590 Diepenbeek, Belgium *niel.hens@uhasselt.be*

[2] Centre for Health Economics Research and Modeling Infectious Diseases, Centre for the Evaluation of Vaccination, Vaccine & Infectious Disease Institute, University of Antwerp, Campus Drie Eiken, Universiteitsplein 1, 2610 Antwerpen, Belgium *niel.hens@ua.ac.be*

[3] Institute of Medical Epidemiology, Biostatistics and Informatics, Medical Faculty, Martin Luther University Halle-Wittenberg, Magdeburger Strasse 8, 06097 Halle, Germany, *andreas.wienke@medizin.uni-halle.de*

Abstract. Frailty models are often used to study the individual heterogeneity in multivariate survival analysis. Estimating frailty models is not straightforward due to various types of censoring. In this manuscript focus is on Type II interval censored data commonly known as current status data for which we study the behavior of the correlated gamma frailty model (Hens et al. 2009). We show that misspecification of both the frailty model and/or the baseline hazard leads to biased estimates in the case of current status data. These results shed a first light on the use of the correlated gamma frailty model for bivariate current status data. This situation typically applies in infectious disease epidemiology where multisera data constituting multivariate current status data are studied to quantify the heterogeneity in acquisition of infections using a shared gamma frailty (Farrington et al. 2001). The use of correlated frailty models facilitates the separation of heterogeneity and correlation. Studying this correlation could indicate whether different infections are transmitted via the same routes. This could prove worthwhile for diseases for which the transmission route is unknown. We used data on hepatitis A and B, two infections transmitted through different routes for illustration purposes.

Keywords: correlated frailty, bivariate binary data, Type II interval censored data

References

FARRINGTON, C., KANAAN, M. and GAY, N. (2001): Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Applied Statistics 50, 251-292.*

HENS, N., WIENKE, A., AERTS, M. and MOLENBERGHS, G. (2009): The correlated and shared gamma frailty model for bivariate current status data: An illustration for cross-sectional serological data. *Statistics in Medicine 28, 2785-2800.*

Exact Posterior Distributions over the Segmentation Space and Model Selection for Multiple Change-Point Detection Problems

G. Rigaill[123], E. Lebarbier[12], S. Robin[12]

[1] AgroParisTech, UMR 518, F-75005, Paris, France
[2] INRA, UMR 518, F-75005, Paris, France
[3] Institut Curie, Département de Transfert, F-75005 Paris, France

Abstract. In segmentation problems, inference on change-point position and model selection are two difficult issues due to the discrete nature of change-points. In a Bayesian context, we derive exact, non-asymptotic, explicit and tractable formulae for the posterior distribution of variables such as the number of change-points or their positions. We also derive a new selection criterion that accounts for the reliability of the results. All these results are based on an efficient strategy to explore the whole segmentation space, which can be very large. We illustrate our methodology on both simulated data and a comparative genomic hybridisation profile.

Keywords: change-point detection, posterior distribution of change-points

A Bootstrap Method to Improve Brain Subcortical Network Segregation in Resting-State FMRI Data

Caroline Malherbe[1,2,*], Eric Bardinet[3,4,5], Arnaud Messé[1,2], Vincent Perlbarg[1,2], Guillaume Marrelec[1,2], Mélanie Pélégrini-Issac[1,2], Jérôme Yelnik[3,5], Stéphane Lehéricy[2,3,4,5], Habib Benali[1,2]

[1] Inserm and UPMC Univ Paris 06, UMR_S 678, Laboratoire d'Imagerie Fonctionnelle, 91 boulevard de l'Hôpital, Paris, France, **caroline.malherbe@imed.jussieu.fr*
[2] Inserm, Université de Montréal, UPMC Univ Paris 06, LINeM, Laboratoire International de Neuroimagerie et Modélisation, Paris, France
[3] Inserm and UPMC Univ Paris 06, UMR_S 975, CRICM, Paris, France
[4] UPMC Univ Paris 06, Centre for NeuroImaging Research – CENIR, Pitié-Salpêtrière Hospital, Paris, France
[5] CNRS, UMR 7225, CRICM, Paris, France

Abstract. Brain functional networks are sets of distant cortical, subcortical or cerebellar regions characterized by coherent dynamics. While spatial independent component analysis (sICA) (Perlbarg et al. (2008) reproducibly detects the cortical components of these networks from resting-state functional magnetic resonance imaging (fMRI) data, little is known about their subcortical (basal ganglia, BG) components. We provide a robust method to detect precisely subcortical components. First, we use sICA to extract the cortical components of the networks from resting-state fMRI data in which the subcortical regions are masked out. Second, we detect the BG components corresponding to these cortical regions using a general linear model. Third, we resort to group statistical inference using a bootstrap technique to select BG regions that are robustly found across subjects. The identified subcortical components are validated using a functional atlas of the BG (Yelnik et al. (2007)). Each functional network is finally defined as the union of its cortical and subcortical components.

Keywords: fMRI, functional networks, basal ganglia, sICA, bootstrap

References

PERLBARG, V., et al. (2008): NEDICA: Detection of group functional networks in fMRI using spatial independant component analysis, *Proceedings of the International Symposium on Biomedical Imaging (ISBI'08):1247–1250.*

YELNIK, J., et al. (2007): A three-dimensional, histological and deformable atlas of the human basal ganglia. I. Atlas construction on immunohistochemical and MRI data, *Neuroimage, 34:618–638.*

Clustering of Multiple Dissimilarity Data Tables for Documents Categorization

Yves Lechevallier[1], Francisco de A. T. de Carvalho[2], Thierry Despeyroux[1], and Filipe M. de Melo[2]

[1] INRIA, Paris-Rocquencourt
78153 Le Chesnay cedex, France,
{*Yves.Lechevallier,Thierry.Despeyroux*}*@inria.fr*
[2] Centro de Informatica -CIn/UFPE
Av. Prof. Luiz Freire, s/n -Cidade Universitaria, CEP 50740-540, Recife-PE, Brésil, {*fatc,fmm*}*@cin.ufpe.br*

Abstract. This paper introduces a clustering algorithm that is able to partition objects taking into account simultaneously their relational descriptions given by multiple dissimilarity matrices. These matrices could have been generated using different sets of variables and a fixed dissimilarity function, using a fixed set of variables and different dissimilarity functions or using different sets of variables and dissimilarity functions. This method, which is based on the dynamic hard clustering algorithm for relational data, is designed to provided a partition and a prototype for each cluster as well as to learn a relevance weight for each dissimilarity matrix by optimizing an adequacy criterion that measures the fit between clusters and their representatives. These relevance weights change at each algorithm iteration and are different from one cluster to another. Experiments aiming at obtaining a categorization of a document data base demonstrate the usefulness of this partitional clustering method.

Keywords: Clustering Analysis, Relational Data, Documents Categorization

Improving overlapping clusters obtained by a pyramidal clustering

Edwin Diday[1], Francisco de A. T. de Carvalho[2], and Luciano D.S. Pacifico[2]

[1] LISE-CEREMADE, Université Paris-IX Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris 16 ième, France, *diday@ceremade.dauphine.fr*
[2] Centro de Informatica -CIn/UFPE, Av. Prof. Luiz Freire, s/n -Cidade Universitaria, CEP 50740-540, Recife-PE, Brésil, {*fatc,ldsp*}*@cin.ufpe.br*

Abstract. Indexed standard or spatial hierarchical clustering produce partitions if they are cut at a given level. Such partition can be improved by using a K-means like clustering. In case of a standard or spatial pyramid a cut at a given level produces an overlapping clustering (where some observations can belong to several clusters). In order to improve such overlapping clustering we need an extension of K-means like algorithm to a new kind of algorithm giving at output a better overlapping clustering for a given criterion. The aim of this paper is to provide an algorithm which starts from an overlapping clustering produced by a pyramid and to show that it improves it at each step for a given criterion. Several authors addressed the problem, for example by extending hierarchies to weak hierarchies (Bertrand and Janowitz, (2002)), and more recently Cleuziou (2008) with the OKM algorithm. In this paper , we first present the algorithm , than we show its convergence and finally we give some examples with results and comparisons.

Keywords: Overlapping Clustering, Pyramidal Clustering, Dynamic Clustering

References

BERTRAND, P. and JANOWITZ, M.F. (2002): Pyramids and weak hierarchies in the ordinal model for clustering. *Discrete Applied Mathematics, 122(1–3), 55–81.*

CLEUZIOU, G. (2008): An extended version of the k-means method for overlapping clustering. *In: Proceedings of the Nineteenth International Conference on Pattern Recognition (ICPR 2008): 1–4.*

DIDAY, E. (1986): Orders and Overlapping clusters in pyramids. In: J. De Leeuw, et al., (Eds.): *Multidimensional Data Analysis.* DSWO Press, 201–234.

DIDAY, E. (2008): Spatial classification. *Discrete Applied Mathematics, 156 (8), 1271–1294*

ICHINO, M. and YAGUCHI, H. (1994): Generalized Minkowsky metrics for mixed feature-type data analysis. *IEEE Transactions on System, Man and Cybernetics, 24 (4), 698–708.*

RODRIGUEZ ROJAS, O. (2000): Classification et Modèles linéaires en Analyse des Données Symboliques. *Thèse de Doctorat. Université Paris-IX Dauphine.*

A Cluster-Target Similarity Based Principal Component Analysis for Interval-Valued Data

Mika Sato-Ilic

Faculty of Systems and Information Engineering, University of Tsukuba
Tennodai 1-1-1, Tsukuba, Ibaraki 305-8573, Japan, *mika@sk.tsukuba.ac.jp*

Abstract. This paper proposes a new principal component analysis for interval-valued data. The merit of this analysis is the consideration of dissimilarity of objects in a higher dimensional space when we obtain the projected space by using a covariance matrix involving the contribution degree for the fuzzy classification structure of objects, based on dissimilarity of objects in the higher dimensional space. In order to obtain the adaptable classification structure which is closely related with a selection of an appropriate number of clusters, we propose an alignment criterion which measures similarity between original similarity data and the restored similarity consisting of a fuzzy clustering result under a given number of clusters which we call cluster-target similarity. In addition, we prove the concentration of the alignment criterion which shows that empirical alignment is close to its expectation.

Keywords: Fuzzy clustering, symbolic data, alignment, metric projection

References

BILLARD, L. and DIDAY, E. (2000): Regression analysis for interval-valued data. In: H.A L. Kiers, et al. (Eds.): *Data Analysis, Classification, and Related Methods*. Springer, 369–374.

BOCK, H.H. and DIDAY, E. (Eds.)(2000): *Analysis of Symbolic Data*. Springer.

CRISTIANINI, N., KANDOLA, J., ELISSEEFF, A. and SHQWE-TAYLOR, J. (2006): On kernel target alignment. In: D.E. Holmes and L.C. Jain (Eds.): *Innovations in Machine Learning*. Springer.

KAUFMAN, L. and ROUSSEEUW, P.J. (1990): *Finding Groups in Data*. John Wiley & Sons.

MCDIARMID, C. (1989): *On the Method of Bounded Differences, Surveys in Combinatorics*. Cambridge University Press.

SATO, M. and SATO, Y. (1995): Extended fuzzy clustering models for asymmetric similarity. In: B. Bouchon-Meunier, R.R. Yager, L.A. Zadeh (Eds.): *Fuzzy Logic and Soft Computing*. World Scientific, 228–237.

YOSHIZAWA, G., STIRLING, A. and SUZUKI, T. (2008): Electricity system diversity in the UK and Japan: a multicriteria diversity analysis. *GraSPP Working Paper Series, Graduate School of Public Policy, The University of Tokyo*.

Cutting the Dendrogram through Permutation Tests

Dario Bruzzese[1] and Domenico Vistocco[2]

[1] Dipartimento di Medicina Preventiva, Università di Napoli - Federico II
Via S. Pansini 5, Napoli, Italy, *dario.bruzzese@unina.it*
[2] Dipartimento di Scienze Economiche, Università di Cassino
Via S.Angelo S.N., Cassino, Italy, *vistocco@unicas.it*

Abstract. The output of hierarchical clustering methods is typically displayed as a dendrogram describing a family of partitions indexed by an ultrametric distance. Actually, after the tree structure of the dendrogram has been set up, the most tricky problem is that of cutting the tree with a suitable threshold in order to take out a sub-optimal classification. Several (more or less) objective criteria may be used to achieve this goal, e.g. the deepest step, but most often the partition relies on a subjective choice leaded by interpretation issues. We propose an algorithm, exploiting the methodological framework of permutation test, allowing to find out automatically a sub-optimal partition not necessarily identifiable using a traditional cut approach, as the resulting clusters could correspond to different heights of the tree.

The general working principle of the procedure is as follows. Starting from the root node of the dendrogram, a partial threshold is moved down the tree until a link joining two clusters is encountered. A permutation test is thus performed in order to verify whether the two clusters must be accounted as a unique group (the null hypothesis) or not (the alternative one). If the null cannot be rejected, the corresponding branch will become a cluster of the final partition and none of its sub-branches will be longer processed. Otherwise each of them will be further visited in the course of the procedure. In fact, in both cases, the partial, threshold will continue its path and the next branch of the dendrogram will be processed. The algorithm stops when there are no more branches that stand the test (i.e. the null cannot be rejected any more) .

The proposed procedure can be used regardless of any agglomeration method and distance measure used in the classification process because it relies on the same criteria used for producing it.

Keywords: Hierarchical clustering, Permutation tests, Cluster detection

References

GOOD P. (1994). *Permutation tests: a practical guide to resampling methods for testing hypotheses.* Springer, New York.

RAND W.M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, December 1971, 66, 336, 846–850.

Unsupervised Recall and Precision Measures: a Step towards New Efficient Clustering Quality Indexes

Jean-Charles Lamirel[1], Maha Ghribi[2], and Pascal Cuxac[2]

[1] LORIA - Campus Scientifique BP 239
54506 Vandœuvre-lès-Nancy, France, *lamirel@loria.fr*
[2] INIST-CNRS
2 allée du Parc de Brabois, 54500-Vandœuvre-lès-Nancy, France,
maha.ghribi@inist.fr, pascal.cuxac@inist.fr

Abstract. The use of the methods of classification of information became current to analyze large corpus of data as it is the case in the domain of scientific survey or in that of strategic analyses of research. While carrying out a classification, the aim is to build homogeneous groups of data sharing a certain number of identical characteristics. Furthermore, the clustering, or unsupervised classification, makes it possible to highlight these groups without prior knowledge on the processed data. A central problem that then arises is to qualify these performance in terms of quality: a quality index is a criterion which indeed makes it possible all together to decide which clustering method to use, to fix an optimal number of clusters, and to evaluate or to develop a new method. Traditional quality indexes, that are mainly distance-based indexes relying on the concepts of intra cluster inertia and inter-cluster inertia (Lebart et al. (1982)), do not allow to properly estimate the quality of the clustering in several cases, as in that one of the textual data (Ghribi and al. (2010)). We thus present in this paper an alternative approach for clustering quality evaluation based on unsupervised measures of Recall, Precision exploiting the descriptors of the data associated with the obtained clusters. The Recall makes it possible to measure the exhaustiveness of the contents of the clusters in terms of peculiar descriptors specific to each cluster. The Precision measures the homogeneity of the clusters in terms of proportion of the data containing the associated peculiar descriptors. We finally present an experimental comparison of the behavior of the classical indexes with our new approach on a dataset of bibliographical references issued from the PASCAL database. This comparison clearly highlights that our method is the only one that can distinguish between homogeneous and heterogeneous clustering results.

Keywords: clustering, quality indexes, text mining, heterogeneous data

References

GHRIBI M., CUXAC P., LAMIREL J.C. and LELU A. (2010): Mesures de qualité de clustering de documents : Prise en compte de la distribution des mots-clés. *Atelier EvalECD2010, Hamamet, Tunisie.*

LEBART L., MAURINEAU A. and PIRON M. (1982): *Traitement des données statistiques*, Dunod, Paris.

Two-way Classification of a Table with non-negative entries: Validation of an Approach based on Correspondence Analysis and Information Criteria

Antonio Ciampi[1], Alina Dyachenko[1], and Yves Lechevallier[2]

[1] Department of Epidemiology, Biostatistics and Occupational Health
McGill University, Montreal, Qc., Canada
[2] INRIA, Paris-Rocquencourt
78153 Le Chesnay cedex, France *Yves.Lechevallier@inria.fr*

Abstract. We present a validation of a rule to choose the number of dimension in Correspondence Analysis and of an AIC/BIC based selection approach to block clustering. An example of micro-array analysis is also shown.

Keywords: 2-way clustering, dimension reduction, number of clusters, microarrays

Half-Taxi Metric in Compositional Data Geometry Rcomp

Katarina Košmelj[1] and Vesna Žabkar[2]

[1] Biotechical Faculty, University of Ljubljana
Ljubljana, Slovenia, *katarina.kosmelj@bf.uni-lj.si*
[2] Faculty of Economics, University of Ljubljana
Ljubljana, Slovenia, *vesna.zabkar@ef.uni-lj.si*

Abstract. Miller (2002) presents the half-taxi metric applicable to compositional data without suggesting how it might be applied. We believe that the half-taxi metric is preferable to other metrics in compositional data geometry rcomp because it takes into account the fact that compositions are closed to one and it has a simple geometric representation on the ternary graph. In an application on advertising expenditure components (Electronic, Print and Online) for 17 European countries in the period 2001-2008 we use the half-taxi metric to detect the structural changes in time, in particular in view of the newer Online component. The results are satisfactory and can be explained in the subject-matter context in view of Hofstede's theory.

Keywords: compositional data, R package compositions, online advertising

References

van den BOOGAART, K. G., TOLOSANA-DELGADO, R. (2008). "compositions": A unified R package to analyze compositional data. *Computers and Geosciences, 34(4), 320-338.*

EUROMONITOR (2009): World Marketing Data and Statistics. (www.euromonitor.com/womdas)

HAJDU, L.J. (1981): Graphical Comparison of Resemblance Measures in Phytosociology. *Vegetatio, v. 48, 47-59.*

HOFSTEDE, G.E. (2001): *Culture's consequences: comparing values, behaviors, institutions, and organizations across nations, Sage, London.*

KOŠMELJ, K., ŽABKAR, V. (2008): A Methodology for Identifying Time-Trend Patterns: an Application to the Advertising Expenditure of 28 European Countries in the 1994-2004 Period. *Metodološki zvezki, 5 (2), 161-171.*

MILLER, W. E. (2002): Revisiting the geometry of a ternary diagram with the half-taxi metric. *Mathematical Geology, 34(3), 275-290.*

SEMIPARAMETRIC MODELS WITH FUNCTIONAL RESPONSES IN A MODEL ASSISTED SURVEY SAMPLING SETTING Submitted to COMPSTAT 2010

Hervé Cardot[1], Alain Dessertaine[2], and Etienne Josserand[1]

[1] Institut de Mathématiques de Bourgogne, UMR 5584 CNRS,
Université de Bourgogne, 9, Av. A. Savary - B.P. 47 870, 21078 Dijon, France
herve.cardot@u-bourgogne.fr, etienne.josserand@u-bourgogne.fr
[2] EDF, R&D, ICAME - SOAD
1, Av. du Général de Gaulle, 92141 Clamart , France
alain.dessertaine@edf.fr

Abstract. This work adopts a survey sampling point of view to estimate the mean curve of large databases of functional data. When storage capacities are limited, selecting, with survey techniques, a small fraction of the observations is an interesting alternative to signal compression techniques. We propose here to take account of real or multivariate auxiliary information available at a low cost for the whole population, with semiparametric model assisted approaches, in order to improve the accuracy of Horvitz-Thompson estimators of the mean curve. We first estimate the functional principal components with a design based point of view in order to reduce the dimension of the signals and then propose semiparametric models to get estimations of the curves that are not observed. This technique is shown to be really effective on a real dataset of 18902 electricity meters measuring every half an hour electricity consumption during two weeks.

Keywords: Design-based estimation, Functional Principal Components, Electricity consumption, Horvitz-Thompson estimator

References

CARDOT, H., CHAOUCH, M., GOGA, C. and C. LABRUÈRE (2010). Properties of Design-Based Functional Principal Components Analysis, *J. Statist. Planning and Inference.*, **140**, 75-91.

CARDOT, H., JOSSERAND, E. (2009). Horvitz-Thompson Estimators for Functional Data: Asymptotic Confidence Bands and Optimal Allocation for Stratified Sampling. `http://arxiv.org/abs/0912.3891`.

DESSERTAINE, A. (2006). Sampling and Data-Stream : some ideas to built balanced sampling using auxiliary hilbertian informations. *56th ISI Conference*, Lisboa, Portugal, 22-29 August 2007.

SÄRNDAL, C.E., SWENSSON, B. and J. WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

EOFs for Gap Filling in Multivariate Air Quality data: a FDA Approach

Mariantonietta Ruggieri, Francesca Di Salvo, Antonella Plaia, and Gianna Agró

Department of Statistical and Mathematical Sciences, University of Palermo
viale delle Scienze - building 13, 90128 Palermo, Italy.
ruggieri@dssm.unipa.it, *disalvo@dssm.unipa.it*, *plaia@unipa.it*, *agro@unipa.it*

Abstract. Missing values are a common concern in spatiotemporal data sets. During recent years a great number of methods have been developed for gap filling. One of the emerging approaches is based on the Empirical Orthogonal Function (EOF) methodology, applied mainly on raw and univariate data sets presenting irregular missing patterns. In this paper EOF is carried out on a multivariate space-time data set, related to concentrations of pollutants recorded at different sites, after denoising raw data by FDA approach. Some performance indicators are computed on simulated incomplete data sets with also long gaps in order to show that the EOF reconstruction appears to be an improved procedure especially when long gap sequences occur.

Keywords: FDA, EOF, missing data, gap filling

References

BECKERS, J.M. and RIXEN, M. (2003): EOF Calculations and Data Filling from Incomplete Oceanographic Datasets. *Journal of Atmospheric and Oceanic Technology 20 (12), 1839-1856.*

DI SALVO, F., AGRÓ, G., PLAIA, A. and RUGGIERI, M. (2009): Exploring spatio-temporal patterns in air pollution data by FPCA. *Submitted to Environmetrics.*

KONDRASHOV, D. and GHIL, M. (2006): Spatio-temporal filling of missing points in geophysical data sets. *Nonlinear Processes in Geophysics 13, 151-159.*

LITTLE, R.J.A. and RUBIN, D.B. (1987): *Statistical Analysis with Missing Data.* Wiley, New York.

OTT, W.R. and HUNT, W.F. (1976): A quantitative evaluation of the pollutant standards index. *Journal of the Air Pollution Control Association 26, 1050-1054.*

PLAIA, A. and BONDÍ, M. (2006): Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment 40, 7316-7330.*

RAMSAY, J.O. and SILVERMAN, B.W. (2005): *Functional Data Analysis. Second Edition.* Springer-Verlag.

SORJAMAA, A., LENDASSE, A., CORNET, Y. and DELEERSNIJDER, E. (2009): An improved methodology for filling missing values in spatiotemporal climate data set. *Computational Geosciences 14 (1), 55-64.*

Clustering Functional Data Using Wavelets

Anestis Antoniadis[1], Xavier Brossat[2], Jairo Cugliari[2,3], and Jean-Michel Poggi[3,4]

[1] Université Joseph Fourier, Laboratoire LJK, Tour IRMA, BP53, 38041 Grenoble Cedex 9, France, *anestis.antoniadis@imag.fr*
[2] EDF R&D, 1 avenue du Général de Gaulle, 92141 Clamart Cedex, France, *xavier.brossat@edf.fr*
[3] Université Paris-Sud, Mathématique Bât. 425, 91405 Orsay, France *jairo.cugliari@math.u-psud.fr*, *jean-michel.poggi@math.u-psud.fr*
[4] Université Paris 5 Descartes, France

Abstract. This paper presents a method for effectively detecting patterns and clusters in high dimensional time-dependent functional data. It provides a time-scale decomposition of the signals under which we can visualize and cluster the functional data into homogeneous groups. We consider the contribution of each scale to the global energy, in the orthogonal wavelet transform of each input function to generate a handy number of features that still makes the signals well distinguishable. Our new similarity measure combined with an efficient feature selection technique in the wavelet domain is then used within more or less classical clustering algorithms to effectively differentiate among high dimensional populations.

Keywords: Clustering, Functional Data, Wavelets

References

ANTONIADIS, A., PAPARODITIS, E. and SAPATINAS, T. (2006): A functional wavelet-kernel approach for time series prediction. *Journal Royal Statisticial Society Series B Statistical Methodology, 68(5), 834–857.*

GURLEY, K., KIJEWSKI, T. and KAREEM, A. (2003): First-and higher-order correlation detection using wavelet transforms. *Journal of engineering mechanics, 129(2), 188–201.*

JAMES, G.M. and SUGAR, C.A. (2003): Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach. *Journal of the American Statistical Association, 98(463), 750–764.*

MALLAT, S.G. (1999): A wavelet tour of signal processing. *Academic Press.*

SERBAN, N. and WASSERMAN, L. (2005): Cats: clustering after transformation and smoothing, *J. Am. Statist. Ass., 100, 990–99.*

STEINLEY, D. and BRUSCO, M.J. (2008). A New Variable Weighting and Selection Procedure for K-Means Cluster Analysis. *Multivariate Behavioral Research, 43(1), 77–108.*

TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 63(2), 411–423.*

Forecasting a Compound Cox Process by means of PCP

Paula R. Bouzas[1], Nuria Ruiz-Fuentes[2], and Juan Eloy Ruiz-Castro[1]

[1] Dept. Statistics and Operations Research, University of Granada
Faculty of Pharmacy, Granada, Spain, *paula@ugr.es*
[2] Dept. Statistics and Operations Research, University of Jaén
Campus Las Lagunillas, Jaén, Spain, *nfuentes@ujaen.es*
[3] Dept. Statistics and Operations Research, University of Granada
Faculty of Sciences, Granada, Spain, *jeloy@ugr.es*

Abstract. A compound Cox process (CCP) is a generalization of a Cox process in which the events have an associated mark. The counting statistics of a CCP with marks in a specific subset are presented and the expression of the mode is derived. The representation theorems of the CCP, an ad hoc FPCA estimation of the mean process and principal components prediction models are the basis to forecast the mean and mode of the CCP in the future. Several simulations illustrate it.

Keywords: Cox process, functional principal components analysis, prediction with principal components

References

AGUILERA, A.M., OCAÑA, F.A. and VALDERRAMA, M.J. (1997): An approximated principal component prediction model for continuous-time stochastic processes. *Applied Stochastic Models and Data Analysis. Vol. 13, 61-72.*

BOUZAS, P.R., VALDERRAMA, M.J., AGUILERA, A.M. and RUIZ-FUENTES, N. (2006): Modelling the mean of a doubly stochastic Poisson process by functional data analysis. *Computational Statistics and Data Analysis. Vol. 50 (10), 2655-2667.*

BOUZAS, P.R., RUIZ-FUENTES, N. and OCAÑA, F.M. (2007): Functional approach to the random mean of a compound Cox process. *Computational Statistics. Vol. 22, 467-479.*

BOUZAS, P.R., AGUILERA, M.J. and RUIZ-FUENTES, N. (2010): Functional estimation of the intensity of a Cox process. *Methodology and Computing in Aplied Probability.* Accepted for publication.

BRÈMAUD, P. (1981): *Point processes and queues: Martingale dynamics.* Springer-Verlag, N.Y.

OCAÑA, F.A., AGUILERA, A.M. and VALDERRAMA, M.J. (1999): Functional principal components analysis by choice of norm, *J.Multiv.Analysis, 71:262-276.*

RAMSAY, J.O. and SILVERMAN, B.M. (1997): *Functional Data Analysis.* Springer-Verlag, N.J.

VALDERRAMA, J.M., AGUILERA, A.M. and OCAÑA, F.A. (2000): *Predicción dinámica mediante análisis de datos funcionales.* La Muralla, Madrid.

STOCHASTIC APPROXIMATION TO THE MULTIVARIATE AND THE FUNCTIONAL MEDIAN
Submitted to COMPSTAT 2010

Hervé Cardot[1], Peggy Cénac[1], and Mohamed Chaouch[2]

[1] Institut de Mathématiques de Bourgogne, UMR 5584 CNRS,
Université de Bourgogne, 9, Av. A. Savary - B.P. 47 870, 21078 Dijon, France
herve.cardot@u-bourgogne.fr, peggy.cenac@u-bourgogne.fr
[2] EDF - Recherche et Développement, ICAME-SOAD
1 Av. Général de Gaulle, 92141 Clamart, France
mohamed.chaouch@edf.fr

Abstract. We propose a very simple algorithm in order to estimate the geometric median, also called spatial median, of multivariate (Small, 1990) or functional data (Gervini, 2008) when the sample size is large. A simple and fast iterative approach based on the Robbins-Monro algorithm (Duflo, 1997) as well as its averaged version (Polyak and Juditsky, 1992) are shown to be effective for large samples of high dimension data. They are very fast and only require $O(Nd)$ elementary operations, where N is the sample size and d is the dimension of data. The averaged approach is shown to be more effective and less sensitive to the tuning parameter. The ability of this new estimator to estimate accurately and rapidly (about thirty times faster than the classical estimator) the geometric median is illustrated on a large sample of 18902 electricity consumption curves measured every half an hour during one week.

Keywords: Geometric quantiles, High dimension data, Online estimation algorithm, Robustness, Robbins-Monro, Spatial median, Stochastic gradient averaging

References

DUFLO, M. (1997). *Random Iterative Models*. Springer Verlag, Heidelberg.

GERVINI, D. (2008). Robust functional estimation using the spatial median and spherical principal components. *Biometrika*, **95,** 587-600.

POLYAK, B.T., JUDITSKY, A.B. (1992). Acceleration of Stochastic Approximation. *SIAM J. Control and Optimization*, **30**, 838-855.

SMALL, C.G. (1990). A survey of multidimensional medians. *Int. Statist. Inst. Rev.*, **58**, 263-277.

VARDI, Y., ZHANG, C.H. (2000). The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA*, **97**, 1423-1426.

Different P-spline Approaches for Smoothed Functional Principal Component Analysis

Ana M. Aguilera, M. Carmen Aguilera-Morillo, Manuel Escabias and Mariano J. Valderrama

Department of Statistics and O.R. University of Granada.
Campus de Fuentenueva, 18071-Granada, Spain, *aaguiler@ugr.es*

Abstract. In order to reduce the dimension and to explain the dependence structure of a functional data set in terms of uncorrelated variables, it is usual to use Functional Principal Component Analysis (FPCA). When the sample curves are not smooth enough, the principal component curves have a lot of variability and are difficult to interpret. Regularized FPCA continuously controls the degree of smoothness by introducing a roughness penalty in his own formulation. In this paper we consider two different forms of smoothed FPCA, both of them based on penalized splines (P-splines) smoothing with B-splines basis. They differ in that the first applies the roughness penalty in the construction of principal components whereas the second incorporates it in the approximation of sample curves and then carries out an unsmoothed FPCA.

Keywords: Functional data; principal component analysis; B-spline expansion; P-splines.

References

AGUILERA, A. M., GUTIÉRREZ, R. and VALDERRAMA, M. J. (1996): Approximation of estimators of the PCA of a stochastic process using B-splines. *Communications in Statistics. Simulation and Computation 25 (3), 671-691.*

AGUILERA, A. M., ESCABIAS, M. and VALDERRAMA, M. J. (2008): Discussion of different logistic models with functional data. Application to Systemic Lupus Erythematosus. *Computational Statistics and Data Analysis, 53, 151-163.*

DE BOOR, C. (1977): Package for calculating with B-splines. *Journal of Numerical Analysis 14, 441-472.*

EILERS, P. and MARX, B. (1996): Flexible smoothing with B-splines and penalties. *Statistical Science 11, 89-121.*

ESCABIAS, M., AGUILERA, A.M. and VALDERRAMA, M.J. (2005): Modelling environmental data by functional principal component logistic regression. *Environmetrics 16, 95-107.*

OCAÑA, F.A., AGUILERA, A.M. and ESCABIAS, M. (2007): Computational considerations in functional principal component analysis. *Computational Statistics 22, 449-465*

RAMSAY, J. O. and SILVERMAN, B. W. (1997, 2005): *Functional data analysis.* Springer-Verlag.

SILVERMAN, BW. (1996): Smoothed functional principal component analysis by choice of norm. *Annals of Statistics 24 (1), 1-24*

Score Moment Estimators

Zdeněk Fabián[1]

Institute of Computer Science, Academy of Sciences of the Czech republic
Pod vodárenskou věží 2, 182 00 Prague *zdenek@cs.cas.cz*

Abstract. By the use of newly introduced concept of the scalar score, the score moments are introduced and used for parametric estimation. In contrast to maximum likelihood estimators, the score moment estimators are not efficient, but they are robust for all the parameters of heavy tailed distributions. In contrast to robust methods, the score moment estimators are taking into account the properties of the assumed model. In the present paper we outline shortly the main ideas and derive the estimation equations for some two-parameter distributions and for some distributions with the threshold parameter. In some cases we obtained closed-form solutions. At the end, we compare the score moment and maximum likelihood estimators on the basis of simulation experiments.

Keywords: generalized moments, score moments, robust estimators

References

1.FABIÁN, Z. (2001): Induced cores and their use in robust parametric estimation. *Comm. Statist. Theory Methods 30, 537-556.*
2.FABIÁN, Z. (2007): Estimation of simple characteristics of samples from skewed and heavy-tailed distribution. In C. Skiadas (Ed.): *Recent Advances in Stochastic Modeling and Data Analysis.* World Scientific, Singapore, *43-50.*
3.FABIÁN, Z. (2008): New measures of central tendency and variability of continuous distributions. *Comm. Statist. Theory Methods 37, 159-174.*
4.FABIÁN, Z. (2009): Confidence intervals for a new characteristic of central tendency of distributions. *Comm. Statist. Theory Methods 38, 1804-1814.*

The Set of $3 \times 4 \times 4$ Contingency Tables has 3-Neighborhood Property

Toshio Sumi[1] and Toshio Sakata[2]

[1] Faculty of Design, Kyushu University
4-9-1 Shiobaru, Minami-ku, Fukuoka, 815-8540, Japan,
sumi@design.kyushu-u.ac.jp
[2] Faculty of Design, Kyushu University
4-9-1 Shiobaru, Minami-ku, Fukuoka, 815-8540, Japan,
sakata@design.kyushu-u.ac.jp

Abstract. We consider the sequential conditional test for three-way contingency tables. Conditional tests of no interaction for three-way contingency tables use as the frame of conditional inference the set of all contingency tables with three fixed two-way marginal tables. Lifting between three-way contingency tables means a method of calculating the frame Ω_t of the t-stage from Ω_{t-1} of the $(t-1)$-stage, which makes it easy to perform the sequential conditional test efficiently. In the previous paper, Sakata and Sumi (COMPSTAT'2008), we treated $3 \times 3 \times 3$ tables and 2-neighborhood property. As a continuation, in this paper, we treat $3 \times 4 \times 4$ tables and show that the conditional inference frame Ω_t is obtained from Ω_{t-1} by transformations made by at most three elements of Markov basis for $3 \times 4 \times 4$ contingency tables, that is, 3-neighborhood property.

Keywords: $3 \times 4 \times 4$ contingency tables, sequential conditional test, Markov basis, 3-neighborhood property

References

AGRESTI, A. (2007): An introduction to categorical data analysis. Wiley Series in Probability and Statistics, 2nd edition, *Wiley-Interscience [John Wiley & Sons]*

AOKI, S. (2004): Exact methods and Markov chain Monte Carlo methods of conditional inference for contingency tables. *Doctor Thesis, Tokyo University.*

AOKI, S. and TAKEMURA, A. (2003): Minimal basis for connected Markov chain over $3 \times 3 \times K$ contingency tables with fixed two-dimensional marginals. *Australian and New Zealand Journal of Statistics 45, 229–249.*

SAKATA, T. and SUMI, T. (2008): Lifting between the sets of three-way contingency tables and r-neighborhood property. *Electoric Proceedings of COMPSTAT '2008, Contribulted Papers, Categorical Data Analysis, 87–94.*

STURMFELS, B. (1996): Gröbner bases and convex polytopes. *American Mathematical Society, University Lecture Series 8.*

SUMI, T. and SAKATA, T. (2009a): A proof of 2-neighborhood theorem for $3 \times 3 \times 3$ tables. *preprint.*

SUMI, T. and SAKATA, T. (2009b): The set of $3 \times 3 \times K$ contingency tables for $K \geq 4$ has 3-neighborhood property. *preprint.*

On Aspects of Quality Indexes for Scoring Models

Martin Řezáč[1] and Jan Koláček[1]

Department of Mathematics and Statistics, Masaryk University
Kotlářská 2, 611 37 Brno, Czech Republic, *mrezac@math.muni.cz*

Abstract. Credit scoring models are widely used to predict a probability of an event like client's default. To measure the quality of the scoring models it is possible to use quantitative indexes such as Gini index, K-S statistics, C-statistics and Lift. They are used for comparison of several developed models at the moment of development as well as for monitoring of quality of those models after deployment into real business. The paper deals with mentioned quality indexes, their properties and relationships. The main contribution of the paper is proposition and discussion of indexes and curves based on Lift. Curve of ideal Lift is defined, Lift ratio is proposed as analogy to Gini index. Integrated Relative Lift is defined and discussed.

Keywords: Credit scoring, Quality indexes, Gini index, Lift, Integrated Relative Lift

References

ANDERSON, R. (2007): *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press, Oxford.

CROOK, J.N., EDELMAN, D.B., THOMAS, L.C. (2007): Recent developments in consumer credit risk assessment. *European Journal of Operational Research 183 (3), 1447-1465*.

HAND, D.J. and HENLEY, W.E. (1997): Statistical Classification Methods in Consumer Credit Scoring: a review. *Journal. of the Royal Statistical Society, Series A. 160 (3), 523-541*.

SIDDIQI, N. (2006): *Credit Risk Scorecards: developing and implementing intelligent credit scoring*. Wiley, New Jersey.

THOMAS, L.C. (2000): A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting 16 (2), 149-172*.

THOMAS, L.C. (2009): *Consumer Credit Models: Pricing, Profit, and Portfolio*. Oxford University Press, Oxford.

THOMAS, L.C., EDELMAN, D.B., CROOK, J.N. (2002): *Credit Scoring and Its Applications*. SIAM Monographs on Mathematical Modeling and Computation, Philadelphia.

WILKIE, A.D. (2004): Measures for comparing scoring systems, In: Thomas, L.C., Edelman, D.B., Crook, J.N. (Eds.): *Readings in Credit Scoring*. Oxford University Press, Oxford, 51–62.

A Generative Model for Rank Data Based on Sorting Algorithm

Christophe Biernacki and Julien Jacques

Université Lille I & CNRS, Villeneuve d'Ascq, France,
christophe.biernacki@math.univ-lille1.fr, *julien.jacques@polytech-lille.fr*

Abstract. Rank data arise from a sorting mechanism which is generally unobservable for the statistician. Assuming both that this mechanism relies on paired comparisons and that it aims to minimize their number, the insertion sorting algorithm is one of the best candidates. A Bernoulli event can be naturally introduce in the paired comparison step, leading to an original probabilistic generative model for rank data which depends on the initial presentation order. Its theoretical properties are studied among which unimodality, symmetry and identifiability. In addition, maximum likelihood principle can be easily performed through an EM algorithm thanks to an unobserved latent variables interpretation of the model. Finally, an illustration of adequacy between the proposed model and rank data resulting from a general knowledge quiz suggests the relevance of our proposal.

Keywords: EM algorithm, insertion algorithm, quiz data, rank data, sorting process.

Slimming down a high-dimensional binary datatable: relevant eigen-subspace and substantial content

Alain Lelu

Université de Franche-Comté, LASELDI & LORIA
30 rue Mégevand, 25030 Besançon cedex, France, *alain.lelu@univ-fcomte.fr*

Abstract. Determining the number of relevant dimensions in the eigen-space of a data matrix (Cattell (1966), Bouveyron et al. (2009)) is a central issue in many data-mining applications. We tackle here the sub-problem of finding the "right" dimensionality of a type of data matrices often encountered in the domains of text or usage mining: large, sparse, high-dimensional binary datatables. We present here the application of a randomization test (Cadot (2006)) to this problem. We validate our approach first on an artificial dataset featuring a two-cluster structure and a power-law distribution of the attributes (Newman (2005)), then on a real documentary data collection, i.e. 1900 documents described in a 3600 keywords dataspace, where the actual, intrinsic dimension appears to be 28 times less than the number of keywords - an important information when preparing to cluster or discriminate such data. We also present preliminary results on the problem of clearing non-essential information bits out of the datatable.

Keywords: randomization test, dimensionality reduction, data reconstitution, power-law distribution

References

BOUVEYRON C., CELEUX G. and GIRARD S. (2009): Intrinsic Dimension Estimation by Maximum Likelihood in Probabilistic PCA. In: *PREPRINT - December 10, 2009 1 (HAL 00440372)*

CADOT, M. (2006): *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association*. Ph.D. thesis, Université de Franche-Comté.

CATTELL, R. B. (1966). "The Scree Test for the Number of Factors." *Multivariate Behavioral Research 1(2), 245-276.*

NEWMAN, M. (2005): Power laws, Pareto distributions and Zipf's law. *Contemporary Physics, 323-351.*

Data Mining and Multiple Correspondence Analysis via Polynomial Transformations

Rosaria Lombardo

Economics Faculty, Second University of Naples,
Gran Priorato di Malta, 81043 Capua (CE), Italy, *rosaria.lombardo@unina2.it*

Abstract. In the framework of the Total Quality Management, earlier studies have suggested that enterprises could harness the predictive power of Learning Management System (LMS) data to develop reporting tools that identify at-risk customers/consumers and allow for more timely interventions (Macfadyen and Dawson (2010)). The Customer Interaction System collects different information on customers/consumers by database and data-warehouse dealing with customers handled by the Consumer Affairs and Customer Relations contact center within a company. To support decision making in customer-centric planning tasks, exploratory multivariate data analysis is an important part of corporate data mining. Among different exploratory tools, we focus on Multiple Correspondence Analysis (Greenacre (1984)) via polynomial transformations (OMCA; Lombardo and Meulman (2010), Lombardo and Beh (2010)) to deal with ordered categorical variables and nominal ones too. By OMCA we can easily monitor the overall (dis)satisfaction with respect to the service aspects, where ordered categorical variables or Likert items are involved. By OMCA we automatically obtain clusters of individuals ordered with respect to the ordered categories of responses (no satisfaction, almost satisfaction, good satisfaction, etc). By focusing on the discoveries of actionable patterns in customer data, the marketers or other domain experts make easier to determine actions that should be taken once the customer patterns are discovered.

Keywords: data mining, customer satisfaction, ordered multiple correspondence analysis, customer classification

References

GREENACRE, M. (1984): *Theory and Applications of Correspondence Analysis.* Academic Press, London.

LOMBARDO, R. and BEH, E.J. (2010): Simple and Multiple Correspondence Analysis for Ordinal-scale Variables using Orthogonal Polynomials. *Journal of Applied Statistics, in press.*

LOMBARDO, R. and MEULMAN, J. (2010): Multiple Correspondence Analysis via Polynomial Transformations of Ordered Categorical Variables. *Journal of Classification 10, 32-48.*

MACFADYEN, L. P. and DAWSON, S. (2010): Mining LMS data to develop an early warning system for educators: A proof of concept. *Computers & Education 54 (2), 588-599.*

Structural Modelling of Nonlinear Exposure-Response Relationships for Longitudinal Data

Xiaoshu Lu and Esa-Pekka Takala

Finnish Institute of Occupational Health
Topeliuksenkatu 41 a A, FIN-00250 Helsinki, Finland,*xiaoshu@cc.hut.fi*

Abstract. Exposure-response relationships are of interest in many epidemiological, medical and other applications. Most commonly, linear relationships are examined. However, many longitudinal data show a remarkable dynamic and nonlinear characteristic, which requires a structure-based approach to elucidate the nonlinear exposure-response relationship behind the data. Exposure and response can have strong nonlinear association and no linear correlation. In this paper, we develop a new model for longitudinal data to address these challenges. The methodology includes time series analysis to estimate unobserved components for exposure and response, and to model their dynamic and structural relationship in a fixed-effects form for each subject. An extension of the fixed-effects form to mixed-effects model for all subjects is proposed and the relevant methods for estimating variance-covariance and correlation matrices are presented. The model-building procedure is explained. The performance of the model is demonstrated using the hypothetical data.

Keywords: structural modelling, exposure-response relationship, nonlinear, longitudinal data

Depth Based Procedures for Estimation ARMA and GARCH Models

Daniel Kosiorowski[1]

Department of Statistics, Cracow University of Economics
ul. Rakowicka 27, Cracow, Poland, *daniel.kosiorowski@uek.krakow.pl*

Abstract. Outliers in time series are more complex than in the other situations, where there is no temporal dependence in the data (see Muller and Yohai (2007)). The outliers can have an arbitrarily negative influence on parameter estimates for time series models, and the nature of this influence depends on the type of outlier.

In this paper we propose two strategies for robust estimation of ARMA and GARCH models. The propositions are based on two statistical depth functions, namely regression depth introduced by Rousseeuw and Hubert (for details see e.g. Van Aelst and Rousseeuw (2000)) and general band depth function introduced by Pintado-Lopez and Romo (2006). We study a performance of the propositions on various time series simulated from ARMA(1,1) and GARCH(1,1) models containing additive outliers (AO). The Monte Carlo study shows very good properties of the propositions in terms of robustness to the AO outliers.

The proposed strategy for $ARMA(p,q)$ model estimation is an attractive approach to robust estimation of the real economic processes parameters. Simulation studies show that our approach is not only more robust than conditional least squares but also than proposed recently by Muller et al. (2009) modified M-estimators, and procedures based on robust filters. Note that our user friendly proposition performs well also in the case of time series without outliers.

Our second proposition performs well in the case of the model estimation on the basis of several trajectories generated by the same process $GARCH(p,q)$. The trajectories may concern e. g. several stock exchange companies, districts, and goods of the same kind. The proposition could be also incorporated to a panel data analysis.

Keywords: depth function, robust estimation, ARMA, GARCH

References

LOPEZ-PINTADO, S. and ROMO J. (2006): Depth-based classification for functional data : In: Liu R.Y., Serfling R., Souvaine D. L. (Eds.):*Series in Discrete Mathematics and Theoretical Computer Science*, AMS, vol. 72, 103 - 119.

MULER, N., PENA, D. and YOHAI V. J. (2009) : Robust estimation for ARMA models, *Annals of Statistics 37 (2),816-840.*

MULER, N.,YOHAI, V. J. (2007): Robust estimates for GARCH models, Technical Report Instituto de Calculo Facultaded de Cinecias Exactas y Naturales Universidad de Buenos Aires.

VAN AELST, S., ROUSSEEUW, P. J. (2000):Robustness Properties of Deepest Regression, *J. Multiv. Analysis, 73, 82-106.*

Forecasting by Beanplot Time Series

Carlo Drago[1] and Germana Scepi[1]

University of Naples "Federico II"
Complesso Universitario Monte Sant'Angelo via Cinthia, Naples, Italy,
carlo.drago@unina.it, germana.scepi@unina.it

Abstract. Symbolic data analysis (SDA) proposes an alternative approach to deal with large and complex data sets. It allows the summarization of these datasets into smaller and more manageable ones retaining the key knowledge. In this framework, we propose an approach for the aggregation of complex time series. This approach is based on a peculiar density plot, called beanplot (Kampstra, 2008). In particular, we explicitly take in account the Density of the data in the underlying temporal interval considered. These types of new data can be fruitfully used where there is an overwhelming number of observations, for example in High Frequency Financial Data (Drago, Scepi, 2009). At the same time, they can be useful for analyzing the complex behavior of the markets where we can discover important patterns in the long time (complex patterns of dependency over the time). In general, Beanplots can be used in the Exploratory Data Analysis framework as a data visualization tool at the same way as Stripcharts, Boxplots and Violinplots, where each one, considered singularly, can be a suitable transformation of histograms. Anyway the interest in this paper is not to explore the structures of data, but to consider the possibility of forecasting complex time series by means of these new type of data. In that sense, the first step is to look for a good parameterization of the Complex Objects over the time, by indicators of size, location and shape of our data. In particular we consider a polynomial function as specific measure of the structure of the object. Successively, the Forecasting process can be performed by using the VAR model (Lutkepohl,2005). We show the usefulness of this strategy by means of simulated data.

Keywords: Forecasting, Symbolic Data Analysis, Beanplot

References

DRAGO, C. and SCEPI, G. (2009): Univariate and Multivariate Tools for Visualizing Financial Time Series. In Ingrassia S. and Rocci R. (eds.) *Proceedings of Seventh Meeting of the Classification and Data Analysis Group of the Italian Statistical Society* Cleup editore, Catania *481–485.*

KAMPSTRA, P. (2008): Beanplot: A Boxplot Alternative for Visual Comparison of Distributions *Journal of Statistical Software, 28.*

LUTKEPOHL, H. (2005): *New Introduction to Multiple Time Series Analysis.* Springer.

The Financial Crisis of 2008: Modelling the Transmission Mechanism Between the Markets

M.Pilar Muñoz[1] M.Dolores Márquez[2] and Helena Chuliá[3]

[1] Departament of Statistics and Operations Research, Universitat Politècnica de Catalunya
C/ Jordi Girona 1-3, Campus Nord C5 204, 08034 Barcelona, Spain, *pilar.munyoz@upc.edu*
[2] Departament of Business Economics, Universitat Autònoma de Barcelona
C/ Emprius, 2 Sabadell , Barcelona, Spain, *mariadolores.marquez@uab.cat*
[3] Departament of Economics and Business, Open University of Catalunya
C/ Jordi Girona 1-3, Campus Nord C5 204, 08034 Barcelona, Spain, *hhulia@uoc.edu*

Abstract. During recent years, we have seen how financial crises have extended geographically to the markets around the world. In this work we investigate how the current failures of the United States financial institutions have affected most of the stock markets in the world. First, we apply Time Series Factors Analysis (TSFA) in order to reduce the dimensionality of the number of indexes and obtain a lower number of new factors that can be related to regions. Then we use the dynamic conditional correlation (DCC) model to analyze the linkages between these regions. Our approach allows us to distinguish between contagion and interdependence. The results show evidence of a contagion effect between some regions.

Keywords: Contagion,Multivariate Volatility, Time Series Factor Analysis and Dynamical Conditional Correlation

References

CHIANG, T.C., JEON, B.N., LI H., (2007): Dynamic correlation analysis of financial contagion: Evidence from Asian markets. *Journal of International Money and Finance 26, 1206-1228.*

ENGLE, R.E., (2002): Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business and Economic Statistics 20, 339-350.*

FORBES, K., RIGOBON, R., (2002): No contagion, only interdependence: measuring stock market comovements. *Journal of Finance 57 (5), 2223-2261.*

GILBERT, P. and MEIJER, E., (2005): Time Series Factor Analysis with an Application to Measuring Money*Research Report N 05F10. University of Gronigen, SOM Research School. Available at http://som.rug.nl* .

Modeling and Forecasting Electricity Prices and their Volatilities by Conditionally Heteroskedastic Seasonal Dynamic Factor Analysis

Carolina García-Martos[1], Julio Rodríguez[2] and María Jesús Sánchez[3]

[1] Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid, Spain, *garcia.martos@upm.es*
[2] Facultad de Ciencias Económicas y Empresariales, Universidad Autónoma de Madrid, Spain, *jr.puerta@uam.es*
[3] Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid, Spain, *mjsan@etsii.upm.es*

Abstract. In this work we propose a new model, that allows to extract conditionally heteroskedastic common factors from a vector of series. These common factors, their relationship with the original vector of series, as well as the dynamics affecting both their conditional mean and variance are jointly estimated. Considering that ARCH and GARCH effects can be handled under the state-space formulation (Harvey et al., 1992), the estimation of the model is carried out in this way. The new model proposed is applied to extract seasonal common dynamic factors as well as common volatility factors for electricity prices. Then, the estimation results are used to forecast electricity prices and their volatilities in the Spanish Market.

Keywords: forecasting, dimensionality reduction, electricity prices, conditional heteroskedasticity.

References

HARVEY, A., RUIZ, E. and SENTANA, E. (1992): Unobserved Component Time Series Models with ARCH Disturbances. *Journal of Econometrics, 52, 129-158.*

Estimation and Detection of Outliers and Patches in Nonlinear Time Series Models

Ping Chen

Department of Mathematics, Southeast University, Nanjing, 210096, China,
cp18@263.net.cn

Abstract. In this paper, we propose a Gibbs sampling algorithm to detect additive isolated outliers and patches of outliers in ARMAX and bilinear time series models with Bayesian view. First, we use some methods to delete the influence of input process in ARMAX model, and then mining outliers and patches in ARMAX series based on the former work. Second, we also detect the outliers and patches in bilinear models by analogous method. It is shown that our procedure could reduce possible masking and swamping effects, which is an improvement and extension on ARMA models over the existing detection methods. At last, simulated examples show that we acquire better results.

Keywords: nonlinear time series, ARMAX model, bilinear models, outlier patches, Gibbs sampler

Wavelet-PLS Regression: Application to Oil Production Data

Benammou Saloua[1], Kacem Zied[1], Kortas Hedi[1], and Dhifaoui Zouhaier[1]

[1] Computational Mathematical Laboratory, *saloua.benammou@yahoo.fr*
[2] *ZiedKacem2004@yahoo.fr*
[3] *kortashedi@yahoo.fr*
[4] *dh.zouhaier@yahoo.fr*

Abstract. This paper is devoted to the study of PLS regression in the presence of noise that could affect the quality of the results. To solve this problem, we suggest a hybrid approach which combines PLS regression and wavelet-based thresholding techniques. The proposed method is validated via a simulation study and subsequently applied to petroleum data. Empirical results show the relevance of the selected approach and contribute to a better modelling of the series of study.

Keywords: PLS regression, Thresholding, Minimax, Wavelet-PLS

References

AMINGHAFARI M., CHEZE N. and POGGI J.M. (2006). Multivariate denoising using wavelets and principal component analysis. *Computational Statistics & Data Analysis 50 (9), 2381-2398*

DAUBECHIES I. (1992). *Ten lectures on wavelets*, SIAM, Philadelphia

DONOHO D. (1995). De-noising by soft-thresholding. *IEEE Trans. Inform. theory, 41 (3), 612-627.*

DONOHO D. and JOHNSTONE I. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist, 26 (3), 879-921.*

DONOHO D. and JOHNSTONE I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika, 81, 425-455.*

MALLAT S. (2000). *Une exploration des signaux en ondelettes*. Les éditions de l'école Polytechnique, Ellipses edition.

HAUGH M. (2004). "The Monte Carlo Framework, Examples from Finance and Generating Correlated Random Variables". Course Notes. *www.columbia.edu/mh2078/MCS04/MCS framework FEegs.pdf.*

GLASSERMAN P. (2003). *Monte Carlo methods in financial engineering*. Springer-Verlag.

TENENHAUS M. (1998). *La régression PLS : Théorie et Pratique*. Technip, Paris.

TENENHAUS M. (1995). *Nouvelle Méthodes de Régression PLS*. Les cahiers de recherche, CR540.

TENENHAUS M., GAUCHI J. P. and MENARDO C. (1995). Régression PLS et Applications. *Revue de Statistique Appliquée, (3), 7-63.*

Test of Mean Difference for Longitudinal Data Using Circular Block Bootstrap

Hirohito Sakurai and Masaaki Taguri

National Center for University Entrance Examinations
2–19–23 Komaba, Meguro-ku, Tokyo 153–8501, Japan
{*sakurai, taguri*}*@rd.dnc.ac.jp*

Abstract. This paper proposes a testing method for detecting the difference of two means or mean curves in longitudinal data using the circular block bootstrap. For the detection of mean difference, we consider four types of test statistics. Monte Carlo simulations are carried out in order to examine the sizes and powers of the proposed test.

Keywords: circular block bootstrap, resampling, longitudinal data, comparison of mean curves

Variational Bayesian Inference for Parametric and Non-Parametric Regression with Missing Predictor Data

Christel Faes[1], John T. Ormerod[2], and Matt P. Wand[2]

[1] Center for Statistics, Hasselt University
BE3590 Diepenbeek, Belgium *christel.faes@uhasselt.be*
[2] Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong 2522, Australia

Abstract. Bayesian hierarchical models are attractive structures for conducting regression analyses when the data are subject to missingness. However, the requisite probability calculus is challenging and Monte Carlo methods typically are employed. We develop an alternative approach based on deterministic variational Bayes approximations. Variational Bayes methods are a family of approximate inference techniques based on the notions of minimum Kullback-Leibler divergence and product assumptions on the posterior densities of the model parameters. They are known as mean field approximations in the statistical physics literature (Parisi, 1988, Bishop, 2006, Ormerod and Wand, 2009). Both parametric and nonparametric regression are treated in the case of missing predictor data. We demonstrate that approximate inference with variational Bayes can achieve good accuracy, but with considerably less computational overhead. The main ramification is fast approximate Bayesian inference in parametric and nonparametric regression models with missing data.

Keywords: Bayesian inference; Directed acyclic graphs; Incomplete data; Mean field approximation; Penalized splines; Variational approximations.

References

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Parisi, G. (1988). Statistical Field Theory. Redwood City, California: Addison-Wesley.

Ormerod, J.T., and Wand, M.P. (2009). Explaining variational approximations. Under review for emphThe American Statistician.

Adaptive Histograms from a Randomized Queue that is Prioritized for Statistically Equivalent Blocks

Gloria Teng[1,2], Jennifer Harlow[1,3], and Raazesh Sainudiin[1,4]

[1] Department of Mathematics and Statistics, University of Canterbury
Private Bag 4800, Christchurch, New Zealand
[2] *gat41@student.canterbury.ac.nz*
[3] *jah217@student.canterbury.ac.nz*
[4] *r.sainudiin@math.canterbury.ac.nz*

Abstract. We present a consistent histogram estimator driven by a randomized queue that is prioritised for a generalized statistically equivalent blocks rule. Such data-dependent adaptive histograms are formalized as statistical regular sub-pavings (SRSPs). A regular sub-paving (RSP) or n-tree (Jaulin et al. (2001), Samet (1996)) is an ordered binary tree that recursively bisects a root box $\mathbb{X} \subset \mathbb{R}^d$ along the first longest side. SRSP augments RSP by mutably caching the recursively computable and minimally sufficient statistics of the data. We formalise our histogram estimator as a Markov chain over the space of SRSPs, implement the algorithm and present simulation results.

Keywords: adaptive histograms, data-dependent partitioning, randomized priority queue, statistically equivalent blocks, statistical regular sub-paving

References

JAULIN, L., KIEFFER, M., DIDRIT, O. and WALTER, É. (2001): *Applied Interval Analysis*. Springer-Verlag, London.

SAMET, H. (1990): *The Design and Analysis of Spatial Data Structures*. Addison-Wesley Longman Publishing Co., Inc., Boston.

Application of the generalized jackknife procedure to estimate species richness

Tsung-Jen Shen[1] and Wen-Han Hwang[2]

[1] Department of Applied Mathematics and Institute of Statistics, National Chung Hsing University, Taichung, Taiwan, *tjshen@nchu.edu.tw*
[2] Department of Applied Mathematics and Institute of Statistics, National Chung Hsing University, Taichung, Taiwan, *wenhan@nchu.edu.tw*

Abstract. Many methods are available for estimating the number of species in a community. However, most of them result in considerable negative bias in applications, where field surveys typically represent only a small fraction of sampled communities. The present study develops a new method to estimate species richness based on the generalized jackknife procedure along with a model assumption associated with frequency counts. The proposed estimator is possessed of small bias and fair interval estimation even with small samples. The performance of the proposed estimator is compared with several typical estimators via simulation study.

Keywords: diversity, jackknife procedure, quadrat sampling, species richness

References

GRAY, H. L. and SCHUCANY, W. R. (1972): *The Generalized Jackknife Statistic.* New York, Marcel Dekker.

HWANG, W.-H. and SHEN, T.-J. (2009): Small-sample estimation of species richness applied to forest communities. *Biometrics* (Early View, Dec 2009).

Robust Generalized Additive Models: mean and dispersion function estimation

Christophe Croux[1,3] Irène Gijbels[2,3] and Ilaria Prosdocimi[2,3]

[1] Operations Research and Business Statistics, KU Leuven
Naamsestraat 69, 3000 Leuven, Belgium, *Christophe.Croux@econ.kuleuven.be*
[2] Afdeling Statistiek, KU Leuven
Celestijnenlaan 200B, 3001 Heverlee, Belgium
Irene.Gijbels@wis.kuleuven.be *Ilaria.Prosdocimi@wis.kuleuven.be*
[3] Leuven Statistics Research Centre, KU Leuven
Celestijnenlaan 200B, 3001 Heverlee

Abstract. Generalized additive models (GAM) are since many years a widely used method in statistics which extends the well known GLM framework in order to obtain smooth estimates for the mean function. Different approaches have been proposed to estimate GAM, in particular Marx and Eilers (1998) propose a direct modeling with P-splines (P-GAM). As for other regression models, the performance of GAM can be severely affected by the presence of outliers. We combine the M-estimator for GLM proposed in Cantoni and Ronchetti (2001) and P-GAM to obtain estimates which are both smooth and robust, similarly to what is done in Azadeh and Salibian-Barrera (2009).

One of the assumptions done in GAM is that the dispersion is constant and known. Real data though, show very often a varying dispersion and we propose methods to estimate the dispersion, as well as the mean, as a function of the covariates. Dispersion estimation moreover can be of interest in itself in different applications (e.g. calibration, quality control). The framework for the mean and the dispersion estimation builds further on work done by Gijbels, Prosdocimi and Claeskens (2010).

Keywords: Dispersion estimation, Generalized Additive Models, M-estimation

References

Azadeh, A. and Salibian-Barrera, M. (2009). An outlier-robust fit for Generalised Additive Models with applications to outbreak detection. Manuscript, available at the second author's website.

Cantoni E. and Ronchetti E. (2001). Robust Inference for Generalized Linear Models. *Journal of the American Statistical Association 96, 1022-1030.*

Eilers P. H. C. , Marx B. D. (1996) Flexible Smoothing with B-splines and Penalties. *Statistical Science 11, 89-121.*

Gijbels I. , Prosdocimi I. and Claeskens G. (2010). Nonparametric estimation of mean and dispersion functions in extended generalized linear models. *Test Accepted for publication,* DOI: 10.1007/s11749-010-0187-1.

A Test Statistic for Weighted Runs

Frederik Beaujean and Allen Caldwell

Max Planck Institute for Physics, Munich, Germany

Abstract. A new test statistic based on success runs of weighted deviations is introduced. Its use for observations sampled from independent normal distributions is illustrated with a real life example. The new statistic supplements the classic χ^2 test which ignores the ordering of observations. The exact distribution of the statistic in the non-parametric case is derived and an algorithm to compute p-values is presented. The computational complexity of the algorithm is given in terms of the number of integer partitions.

Keywords: Goodness of fit, Success runs, χ^2, Measurements with Gaussian uncertainty

References

BEAUJEAN, F., CALDWELL, A. (2010): A Test Statistic for Weighted Runs. *arXiv:1005.3233*

Ensembled Multivariate Adaptive Regression Splines with Nonnegative Garrote Estimator

Hiroki Motogaito[1] and Masashi Goto[2]

[1] Division of Mathematical Science, Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama-cho, Toyonaka, Osaka 560-8531, Japan *h-motogt@sigmath.es.osaka-u.ac.jp*

[2] Biostatistical Research Association, NPO. 2-22-10-A411 Kamishinden, Toyonaka, Osaka 560-0085, Japan *gotoo@bra.or.jp*

Abstract. In regression problems, among the most important goals are (i) to obtain a lower prediction error and (ii) to interpret the regression relationships. Friedman's multivariate adaptive regression splines (MARS) method that constructs basis functions with the interaction effects is a very powerful data-driven technique from the viewpoint of (i); further, the single tree-based structure built by MARS is good from the viewpoint of (ii). In terms of (i), however, the naive MARS is often inferior to machine learning methods like tree ensembles. On the other hand, tree ensembles have drawback in terms of (ii) such as variable selection and difficulty to visualize. Further, to address (i) and (ii), better prediction, variable selection and easy to visualize are essential. Shrinkage estimators can help deal with such issues. Recently, especially in the context of linear regression, Breiman's nonnegative garrote estimator and Tibshirani's least absolute shrinkage and selection operator (lasso) estimator have been shown to be stable estimators with variable selection that often outperform the other estimators. In this paper, we focus on nonnegative garrote to incorporate the shrinkage estimators into the tree-based ensembled MARS model and propose a new version of MARS. The proposed method generates diverse basis functions of MARS model with ensembled technique and selects optimal basis functions and nodes from the ensembled trees using nonnegative garrote. The new method is easy to interpret because of a single-tree output format, Following this, we evaluate the performance of the proposed method using a literature example and small simulations.

Keywords: regression trees, prediction, interpretability, pruning

References

Breiman, L. (1995): Better subset regression using the nonnegative garrote. *Technometrics, 37, 373-384.*

Friedman, J. H. (1991): Multivariate adaptive regression spline (with discussion). *Ann. Statist., 19, 1-141.*

Meinshausen, N. (2009): Node Harvest: simple and interpretable regression and classification. *Arxiv preprint arXiv:0910.2145.*

Motogaito, H., Sugimoto, T. & Goto, M. (2007): Multivariate Adaptive Regression Splines with Non-negative Garrote Estimator. *Japanese J. Appl. Statist, 36, 99-118 (in Japanese).*

Comparison of Regression Methods by Employing Bootstrapping Methods

Ayca Yetere Kursun[1] and Inci Batmaz[2]

[1] Department of Scientific Computing, Institute of Applied Mathematics, Middle East Technical University
ODTU, Ankara, Turkey, *e112987@metu.edu.tr*
[2] Department of Statistics, Middle East Technical University
ODTU, Ankara, Turkey, *ibatmaz@metu.edu.tr*

Abstract. There are differences between parametric and nonparametric regression models. This difference can be quantified with the use of numerical techniques such as bootstrapping. Bootstraping method uses numerical computing and variance formula to get an estimate of the standard error and its confidence interval (Efron and Tibshirani (1991), (1993)). There are two basic ways to bootstrap a regression. First the predictors can be treated as random (Random-x Resampling), changing from sample to sample, and as the second method the predictors can be treated as fixed (Fixed-x Resampling)(Fox (2002)). In this study "Fixed-x Resampling" is employed with three different methods for resampling the residuals: 1- From the set of residuals selecting randomly the bootstrap sample of the residuals and attaching the randomly resampled residual to the response variables 2- For the set of residuals calculating the standard deviation σ, generating a random variable from N(0 , σ^2) and attaching this random variable to the response variables 3- From the set of residuals in each bootstrap sample all residuals are randomly multiplied by 1 or -1 and then reattached to its fitted values. In this study as regression models, we employed Least Squares Estimate, Multivariate Adaptive Regression Splines (MARS) (Friedman (1991)), Loess Smoothing, Local Linear Estimator and Nadaraya-Watson Estimator (Martinez and Martinez (2002)). Standard errors and their confidence intervals are calculated for statistics such as MSE and R^2.

Keywords: Bootstrapping, Nonparametric Regression, Loess Smoothing, MARS

References

FRIEDMAN J. H. (1991): Multivariate Adaptive Regression Splines. *The Annals of Statistics 19(1), 1-67.*

EFRON, B. and TIBSHIRANI R. (1991): Statistical data analysis in the computer age. *Science, New Series 253(5018), 390-395.*

EFRON, B. and TIBSHIRANI R. (1993): *An Introduction to the Bootstrap.* Chapman & Hall.

FOX J. (2002): *An R and S-PLUS Companion to Applied Regression, Web Appendix: Bootstrapping Regression Models.* Sage Publications, Inc.

MARTINEZ, W.L. and MARTINEZ, A.R. (2002): *Computational Statistics Handbook with MATLAB®.Chapman&Hall/CRC.*

Statistical inference for Rényi entropy of integer order

David Källberg[1] and Oleg Seleznjev[2]

[1] Department of Mathematics and Mathematical Statistics, Umeå University SE-901 87 Umeå, Sweden, *david.kallberg@math.umu.se*
[2] Department of Mathematics and Mathematical Statistics, Umeå University SE-901 87 Umeå, Sweden, *oleg.seleznjev@math.umu.se*

Abstract. Entropy and its various generalizations are widely used in mathematical statistics, communication theory, physics and computer science. A class of estimators of Rényi entropy of any order are studied in Leonenko et al. (2008). Estimators of quadratic Rényi entropy for discrete and continuous distributions are studied in Leonenko and Seleznjev (2009). We introduce a new class of estimators of integer order Rényi entropy for discrete and continuous distributions. The estimators are based on the inter-point distances in the i.i.d. sample of vectors. We show some properties, e.g., consistency and asymptotic normality. We also study estimators based on m-dependent stationary samples. The proposed estimators can be used in various problems in mathematical statistics and computer science (e.g., distribution identification problems; average case analysis for random databases; the height of digital trees in information theory; approximate pattern matching in bioinformatics; clustering).

Keywords: Entropy estimation, Rényi entropy, U-statistics

References

LEONENKO, N., PRONZATO, L. and SAVANI, V. (2008): A class of Rényi information estimators for multidimensional densities. *Annals of Statistics 36 (5), 2153-2182.*

LEONENKO, N. and SELEZNJEV, O. (2009): Statistical Inference for Quadratic Rényi Entropy. In: L. Sakalauskas, C. Skiadas and E.K. Zavadskas (Eds.): *Proceedings of the XIII International Conference on Applied Stochastic Models and Data Analysis (ASMDA 2009).* Vilnius Gediminas Technical University Press, Vilnius, 223–227.

Modelling of extreme events in linear models and two-step regression quantiles

Jan Dienstbier

Technická Univerzita v Liberci, Katedra aplikované matematiky
Studentská 2, 461 17 Liberec, Czech Republic, *jan.dienstbier@tul.cz*

Abstract. We work with the linear model and two-step regression quantiles introduced by Jurečková and Picek (2005). In the contribution we aim to an approximation of the distribution of the errors in the model. We are interested chiefly in an estimation of the extremal properties of the distribution, i.e. in the extreme value index of the appertaining extreme value distribution, which is a key factor controlling the rate of appearing rare events. Usual methods such as empirical distribution function or kernel estimates are not suitable for such tasks. However, it turns out that all necessary information is stored in the residuals of R-estimates of the slope in the model. Hence two-step regression quantiles and similar approaches based on quantile sensitive methods such as regression quantiles of Koenker and Basset (1978) are appropriate tools for our problem. We compare the two-step regression quantiles approach with other methods already discussed in the literature. The comparison si made on the basis of simulation study as well as on real data cases such as Condroz data introduced in this context by Beirlant et al. (2004).

Keywords: two-step regression quantile, R-estimator, regression quantile, extreme value index

References

BEIRLANT, J., GOEGEBEUR, Y., SEGERS, J. and TEUGELS, J. (2004): *Statistics of Extremes, Theory and Application*. John Wiley & Sons, Chichester.

JUREČKOVÁ, J. and PICEK, J. (2005): Two step regression quantiles, *Sankhyā 67, 227–252*.

KOENKER, R. and BASSET, G. (1978): Regression quantiles *Econometrica 46, 35–50*.

Non Parametric Confidence Intervals for ROC Curves Comparison

Ana Cristina Braga, Lino Costa, and Pedro Oliveira

Department of Production and Systems Engineering, University of Minho
Campus de Gualtar, 4710-057 Braga, Potugal, {*acb,lac,pno*}*@dps.uminho.pt*

Abstract. Performance evaluation of two diagnostic systems is a problem that can be approached through ROC curves, in particular, the Area Under the ROC Curve (AUC) index. When there are crossings between the two ROC curves, the comparison strictly based on the AUC index does not permit the realization of the differences between the two diagnostic systems.

In this work we present a new methodology based on ROC curves for the comparison of two diagnostic systems. The main idea is based on a multiobjective approach in which several Pareto frontiers must be compared (Deb (2001)). Since the points belonging to ROC curves can be perceived as different trade-offs between specificity and sensitivity, a sampling process is used to determine the distribution of areas in ROC space.

The proposed approach defines two performance measures (extension and location) that allow the evaluation of the regions where a curve is superior to another. Based on bootstrap sampling, non parametric confidence intervals are established (Pepe (2003)).

Several simulations were carried out in order to compare our approach with other approaches (Zhang et al. (2002)). Using different seeds, 200 datasets were generated, with disease and non disease samples of equal size (25,50,100) according to normal distributions: $X_{ND} \sim N(50, 25)$ and $X_D \sim N(60, 25)$.

The results show that the proposed methodology provides a performance similar to other approaches, with the advantage of comparing the curves either globally or in specific regions of the ROC space.

Keywords: ROC curves, simulation, bootstrap

References

DEB, K. (2001): *Multi-Objective Optimization Using Evolutionary Algorithms*. Wiley.

PEPE, M. S. (2003): *The Statistical Evaluation of Medical Tests for Classification and Predict*. Oxford Statistical Science Series, Oxford University Press.

ZHANG, D.D., ZHOU, X. H.,FREEMAN, D. H. and FREEMAN,J. L. (2002): A Non-Parametric Method for The Comparison of Partial Areas Under ROC Curves and Its Application to Large Health Care Data Sets. *Statistics In Medicine, 21(5),701-715*.

Non-Parametric Estimation of Forecast Distributions in Non-Gaussian State Space Models

Jason Ng[1], Catherine Forbes[1], Gael Martin[1], and Brendan P.M. McCabe[2]

[1] Department of Econometrics and Business Statistics, Monash University, Australia. Corresponding author: *gael.martin@buseco.monash.edu.au*
[2] Management School, University of Liverpool, UK

Abstract. In the spirit of an evolving literature in which probabilistic forecasting is the focus (see Diebold et al. (1998), Freeland and McCabe (2004), Gneiting et al. (2007) and Czado et al. (2009) for key contributions), we develop a new method for estimating the full forecast distribution of non-Gaussian time series variables. In contrast to the existing literature, in which the focus is almost exclusively on the specification of strict parametric models, a flexible non-parametric approach is adopted here, with a view to producing accurate distributional forecasts, no matter what the nature of the true data generating process. The method is developed within the general framework of non-Gaussian, non-linear state space models, with the distribution for the observed non-Gaussian variable, conditional on the latent state(s), estimated non-parametrically. The requisite recursive filtering and prediction distributions required to evaluate both the likelihood function and the one step-ahead forecast distribution, are estimated via an algorithm that is closed-form up to the solution of one- (or two-) dimensional integrals at each time point, defined only over the standardized support of the measurement error. Standard deterministic integration techniques can then be used to estimate the relevant integrals. The method is illustrated using a variety of financial models. Most notably, it is used to produce sequential, real time estimates of the forecast distribution of realized volatility in the period leading up to the recent financial turmoil.

Keywords: Probabilistic Forecasts; Non-parametric Maximum Likelihood; Realized Volatility

References

CZADO, C., GNEITING, T. and HELD, L. (2009): Predictive model assessment for count data, *Biometrics, 65(4), 1254–1261.*

DIEBOLD, F.X., GUNTHER, T.A. and TAY, A.S. (1998): Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review 39, 863-883.*

FREELAND, R and MCCABE, B. (2004): Forecasting discrete valued low count time series, *International Journal of Forecasting 20, 427-434.*

GNEITING, T., BALABDAOUI, F. and RAFTERY, A. (2007): Probabilistic forecasts, calibration and sharpness, *Journal of the Royal Statistical Society (B) 69, 243–268.*

Part II

Tuesday August 24

Model based clustering and reduction for high dimensional data, Multivariate Data Analysis

Nikolaus Kriegeskorte[1]

Medical Research Council, Cognition and Brain Sciences Unit
15 Chaucer Road, Cambridge, CB2 7EF, UK
nikokriegeskorte@gmail.com

Abstract. Perceptual and cognitive content is thought to be represented in the brain by patterns of activity across populations of neurons. In order to test whether a computational model can explain a given population code and whether corresponding codes in man and monkey convey the same information, we need to quantitatively relate population-code representations. I will give a brief introduction to representational similarity analysis (RSA), a particular approach to this problem.

A population code is characterized by a representational dissimilarity matrix (RDM), which contains a dissimilarity for each pair of activity patterns elicited by a given stimulus set. The RDM encapsulates which distinctions the representation emphasizes and which it deemphasizes. By analyzing correlations between RDMs we can test models and compare different species. Moreover, we can study how representations are transformed across stages of processing and how they relate to behavioral measures of object similarity.

I will use an example from object vision to illustrate the methods potential to bridge major divides that have hampered progress in systems neuroscience.

Keywords: Population code, Similarity, Distance

The NPAIRS Computational Statistics Framework for Data Analysis in Neuroimaging

Stephen Strother[1,2], Anita Oder[1], Robyn Spring[1,2], and Cheryl Grady[1]

[1] Rotman Research Institute, Baycrest
3560 Bathurst Street, Toronto, ON, Canada ∗sstrother@rotman-baycrest.on.ca
[2] Department of Medical Biophysics, University of Toronto

Abstract. We introduce the role of resampling and prediction (p) metrics for flexible discriminant modeling in neuroimaging, and highlight the importance of combining these with measurements of the reproducibility (r) of extracted brain activation patterns. Using the NPAIRS resampling framework we illustrate the use of (p, r) plots as a function of the size of the principal component subspace (Q) for a penalized discriminant analysis (PDA) to: optimize processing pipelines in functional magnetic resonance imaging (fMRI), and measure the global SNR (gSNR) and dimensionality of fMRI data sets. We show that the gSNRs of typical fMRI data sets cause the optimal Q for a PDA to often lie in a phase transition region between gSNR $\simeq 1$ with large optimal Q versus SNR $\gg 1$ with small optimal Q.

Keywords: prediction, reproducibility, penalized discriminant analysis, fMRI

Imaging Genetics: Bio-Informatics and Bio-Statistics Challenges

Jean-Baptiste Poline[1], Christophe Lalanne[1], Arthur Tenenhaus[2], Edouard Duchesnay[1], Bertrand Thirion[3], and Vincent Frouin[1]

[1] Neurospin, Institut d'Imagerie Biomédicale, CEA, 91191 Gif sur Yvette Cedex, France. jbpoline@cea.fr, ch.lalanne@gmail.com, edouard.duchesnay@cea.fr, vincent.frouin@cea.fr

[2] SUPELEC Sciences des Systèmes (E3S)-Department of Signal processing and Electronics systems, 91192 Gif-sur-Yvette Cedex. arthur.tenenhaus@supelec.fr

[3] Neurospin, INRIA-Parietal, 91191 Gif sur Yvette Cedex, France. Bertrand.Thirion@inria.fr

Abstract. The IMAGEN study—a very large European Research Project—seeks to identify and characterize biological and environmental factors that influence teenagers mental health. To this aim, the consortium plans to collect data for more than 2000 subjects at 8 neuroimaging centres. These data comprise neuroimaging data, behavioral tests (for up to 5 hours of testing), and also white blood samples which are collected and processed to obtain 650k single nucleotide polymorphisms (SNP) per subject. Data for more than 1000 subjects have already been collected. We describe the statistical aspects of these data and the challenges, such as the multiple comparison problem, created by such a large imaging genetics study (i.e., 650k for the SNP, 50k data per neuroimage). We also suggest possible strategies, and present some first investigations using uni or multi-variate methods in association with re-sampling techniques. Specifically, because the number of variables is very high, we first reduce the data size and then use multivariate (CCA, PLS) techniques in association with re-sampling techniques.

Keywords: neuroimaging, genome wide analyses, partial least squares

A Numerical Approach to Ruin Models with Excess of Loss Reinsurance and Reinstatements ⋆

Hansjörg Albrecher[1] and Sandra Haas[2]

[1] Department of Actuarial Science, Faculty of Business and Economics, University of Lausanne,
email: hansjoerg.albrecher@unil.ch
[2] Department of Actuarial Science, Faculty of Business and Economics, University of Lausanne,
email: sandra.haas@unil.ch

Abstract. The present paper studies some computational challenges for the determination of the probability of ruin of an insurer, if excess of loss reinsurance with reinstatements is applied. In the setting of classical risk theory, a contractive integral operator is studied whose fixed point is the ruin probability of the cedent. We develop and implement a recursive algorithm involving high-dimensional integration to obtain a numerical approximation of this quantity. Furthermore we analyze the effect of different starting functions and recursion depths on the performance of the algorithm and compare the results with the alternative of stochastic simulation of the risk process.

Keywords: reinsurance, integral operator, ruin probability, high-dimensional integration

⋆ Supported by the Swiss National Science Foundation Project 200021-124635/1.

Computation of the Aggregate Claim Amount Distribution Using R and Actuar

Vincent Goulet

École d'actuariat
Université Laval
Pavillon Alexandre-Vachon
1045, avenue de la Médecine
Québec, Québec G1V 0A6
Canada *vincent.goulet@act.ulaval.ca*

Abstract. actuar is a package providing additional Actuarial Science functionality to the R statistical system. This paper presents the features of the package targeted at risk theory calculations. Risk theory refers to a body of techniques to model and measure the risk associated with a portfolio of insurance contracts. The main quantity of interest for the actuary is the distribution of total claims over a fixed period of time, modeled using the classical collective model of risk theory.

actuar provides functions to discretize continuous distributions and to compute the aggregate claim amount distribution using many techniques, including the recursive method and simulation. The package also provides various plotting and summary methods to ease working with aggregate models.

Keywords: risk theory, aggregate models, compound distribution, R, actuar

Applications of Multilevel Structured Additive Regression Models to Insurance Data

Stefan Lang[1] and Nikolaus Umlauf[1]

University of Innsbruck, Department of Statistics
Universitätsstraße 15, A-6020 Innsbruck, Austria,
stefan.lang@uibk.ac.at and nikolaus.umlauf@uibk.ac.at

Abstract. Models with structured additive predictor provide a very broad and rich framework for complex regression modeling. They can deal simultaneously with nonlinear covariate effects and time trends, unit- or cluster specific heterogeneity, spatial heterogeneity and complex interactions between covariates of different type. In this paper, we discuss a hierarchical version of regression models with structured additive predictor and its applications to insurance data. That is, the regression coefficients of a particular nonlinear term may obey another regression model with structured additive predictor. The proposed model may be regarded as a an extended version of a multilevel model with nonlinear covariate terms in every level of the hierarchy. We describe several highly efficient MCMC sampling schemes that allow to estimate complex models with several hierarchy levels and a large number of observations typically within a couple of minutes. We demonstrate the usefulness of the approach with applications to insurance data.

Keywords: Bayesian hierarchical models, multilevel models, P-splines, spatial heterogeneity

Comprehensive Assessment on Hierarchical Structures of DNA markers Using Echelon Analysis

Makoto Tomita[1] and Koji Kurihara[2]

[1] Clinical Research Center, Tokyo Medical and Dental University Hospital Faculty of Medicine, 1-5-45 Yushima, Bunkyo-ku, Tokyo, 113-8519, Japan. *tomita.crc@tmd.ac.jp*

[2] Faculty of Environmental Science and Technology, Okayama University, 700-8530, Japan

Abstract. A domain where recombination does not occur often, yet maintained linkage disequilibrium exists on DNA sequence is known as a "haplotype block" or "LD block". Many methods are available to identify LD blocks using disequilibrium parameters, such as the well-known Gabriel's method on Haploview, and so on. After identifying LD blocks, we can also select tagging SNPs for these LD blocks such as Tagger on Haploview, etc. We considered that Echelon analysis can be applied to identify LD block and to select tagging SNPs, and report herein that the comprehensive method can be applied according to our new method using Echelon analysis. The results of numerical examples are also provided.

Keywords: spatial data analysis, DNA data, haplotype, tagging SNP

References

Avi-Itzhak H.I., Su X., De La Vega F.M. (2003). Selection of minimum subsets of single nucleotide polymorphism to capture haplotype block diversity. *Pacific Symposium on Biocomputing.* 8, 466-477.

Barrett J. C., Fry B., Maller J., Daly M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* 21(2), 263-265.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., *et al.* (2002). The structure of haplotype blocks in the human genome. *Science.* 296, 2225-2229.

Myers, W. L., Patil, G. P. and Joly, K. (1997). Echelon approach to areas of concern in synoptic regional monitoring. *Enviromental and Ecologocal Statistics.* 4, 131-152.

Tomita, M., Hatsumichi M. and Kurihara K. (2008). Identify LD Blocks Based on Hierarchical Spatial Data. *Computational Statistics and Data Analysis.* 52, 1806-1820.

Zhu, X., Yan, D., Cooper, R.S., Luke, A., Ikeda, M.A., Chang, Y.P., Weder, A. and Chakravarti, A. (2003). Linkage disequilibrium and haplotype diversity in the genes of the renin-angiotensin system: findings from the family blood pressure program. *Genome Research.* 13, 173-181.

Analysis of Breath Alcohol Measurements Using Compartmental and Generalized Linear Models

Chi Ting Yang[1], Wing Kam Fung[2], and Thomas Wai Ming Tam[3]

[1] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China, *yang@graduate.hku.hk*
[2] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China, *wingfung@hku.hk*
[3] Forensic Science Division, Government Laboratory, 88 Chung Hau Street, Ho Man Tin, Kowloon, Hong Kong, China, *wmtam@govtlab.gov.hk*

Abstract. Pharmacokinetic parameters are important for clinical application. Traditionally, compartmental model together with non-compartmental approach are commonly adopted to deal with the analysis of pharmacokinetic data. In this study, we apply the alternative generalized linear model as proposed by Wakefield (2004) to the breath alcohol measurements of Chinese subjects in Hong Kong. Both compartmental and generalized linear models are fitted to each subject involved. Four core pharmacokinetic parameters including the time to maximum blood alcohol concentration (BAC) level, the BAC level attained at peak, the rate of clearance and elimination half-life are examined. The parameter estimates under the two models are then compared. Finally, we also extend the concept from the individual to the population level.

Keywords: Blood alcohol concentration, One-compartment model, Generalized linear model, Nonlinear mixed model, Generalized linear mixed model.

References

WAKEFIELD, J. (2004): Non-linear regression modeling and inference. In *Methods and Models in Statistics*. Adams, N., Crowder, M., Hand, D., Stephens, D. (eds), pp. 119-153. Imperial College Press, London.

A Flexible IRT Model for Health Questionnaire: an Application to HRQoL

Serena Broccoli and Giulia Cavrini

Faculty of Statistics
Via delle Belle Arti 41, Bologna, Italy
serena.broccoli@unibo.it and *giulia.cavrini@unibo.it*

Abstract. Abstract. The aim of this study is to formulate a suitable Item Response Theory (IRT) based model to measure HRQoL (as latent variable) using a mixed responses questionnaire and relaxing the hypothesis of normal distributed latent variable. The new model is a combination of two models already presented in literature, that is a latent trait model for mixed responses and an IRT model for Skew Normal latent variable (Moustaki (1996); Bazan et al. (2004)). It is developed in a Bayesian framework. A Monte Carlo Markov chain procedure is used to generate samples of the posterior distribution of the parameters of interest. The proposed model was tested on a questionnaire composed by 5 discrete items and one continuous to measure HRQoL in children, the EQ-5D-Y questionnaire (Ravens-Sieberer et al. (2010)). A large sample of children collected in the schools was used. The model was formulated as WINBUGS code and the estimates of the parameters were obtained using a Bayesian procedure, more flexible than the likelihood methods and similarly in the results. In comparison with a model for only discrete responses and a model for mixed responses and normal latent variable, the new model has better performances, in term of deviance information criterion (DIC), chain convergences times and precision of the estimates.

Keywords: IRT Model, Skew Normal Distribution, Health-Related Quality of Life.

References

BAZAN, J., BOLFARINE, H., BRANCO, D. M. (2004): *A new family of asymmetric models for item response theory: A skew normal IRT family.* Technical report (RT-MAE-2004-17). Department of Statistics. University of Sao Paulo.

MOUSTAKI, I. (1996): A latent trait and a latent class model for mixed observed variables. *British journal of mathematical and statistical psychology , 49 (2), 313-334.*

RAVENS-SIEBERER, U., WILLE, N., BADIA, X., BONSEL, G., BURSTRM, K. , CAVRINI , G., EGMAR, A., GUSI, N., HERDMAN, H., JELSMA, J., KIND, P., OLIVARES, P., SCALONE, L., GREINER, W. (2010): Feasibility, reliability and validity of the EQ-5D-Y - results from a multinational study. Accepted for publication. *Quality of Life Research [in press].*

Socioeconomic Factors in Circulatory System Mortality in Europe: A Multilevel Analysis of Twenty Countries

Sara Balduzzi, Lucio Balzani, Matteo Di Maso,
Chiara Lambertini, and Elena Toschi

Faculty of Statistics
Via delle Belle Arti 41, Bologna, Italy, E-mail: *lucio.balzani@libero.it*

Abstract. This paper is the result of teamwork carried out during an advanced course of Health Statistics. Our aim was to apply the multilevel models (Hox (2002)) learned on the course to some concrete research problems without using any fiscal resources. We used the WHO health databases to compare Standardized Death Ratios (SDR) due to Circulatory System Diseases in twenty representative European countries between 1992 and 2003 (World Health Organization (2002)) and to explain the role played by socio-economic and lifestyle indicators using a multilevel approach (years nested in countries). The data shows a general decrease in SDR for Circulatory System Diseases in Europe between 1992 and 2003 (Levi et al. (2002); (Nolte et al.(2005)). However there are wide differences in levels and trends of mortality between Southern, Northern and Central European countries (low level and evident decrease) and the Eastern Europe and Former Soviet countries (higher level and more confused trends). An epidemiological transition is currently occurring in Europe. Western countries are leading with a lower and decreasing level of SDR for circulatory system mortality. Some Eastern countries and all Former Soviet Republics show opposite results but a change in their health and economic policies may accelerate the reduction in this gap. This paper underlines the power of research based on free online institutional databases, especially for health policy makers who often require accurate but expensive health information.

Keywords: Multilevel models, Circulatory system mortality, European Health database

References

HOX, J. (2002): *Multilevel Analysis*. LEA.

LEVI, F., LUCCHINI, F., NEGRI, E. and LA VECCHIA, C. (2002): Trends in mortality from cardiovascular and cerebrovascular disease in Europe and other areas of the world. *Heart 88, 119-124*.

NOLTE, E., MCKEE, M. and GILMORE, A. (2005): Morbidity and Mortality in the transition countries of Europe. In: M. Macura, A. L. MacDonald and W. Haug (Eds.): *The New Demographic Regime Population Challenges and Policy Responses. United Nations New York and Geneva, 153-176 (Chapter 9)*.

WORLD HEALTH ORGANISATION (1992): *Health for All database*. Geneva: World Health Organisation.

Time-Varying Coefficient Model with Linear Smoothing Function for Longitudinal Data in Clinical Trial

Masanori Ito[1] , Toshihiro Misumi[2] and Hideki Hirooka[3]

[1] Astellas Pharma Inc.
3-17-1, Hasune, Itabashi-ku, Tokyo, 174-8612, Japan
masanori.ito@jp.astellas.com
[2] Astellas Pharma Inc.
3-17-1, Hasune, Itabashi-ku, Tokyo, 174-8612, Japan
toshihiro.misumi@jp.astellas.com
[3] Astellas Pharma Inc.
3-17-1, Hasune, Itabashi-ku, Tokyo, 174-8612, Japan
hideki.hirooka@jp.astellas.com

Abstract. In clinical trials, more than one visit for efficacy evaluation are scheduled and the analysis for longitudinal data is required. LOCF ANCOVA model is usually chosen. The LOCF approach assumes the missing data MCAR. But since the assumptions are often unrealistic and thus it is not the best choice. TVCM is applied to clinical trial data for evaluation of the drug treatment varying with time. The inference on the model is conducted by a simple linear smoothing function. The knots of the smoothing function are identified according to the scheduled patient's visits. The inference on the model can be conducted in the context of the mixed model methodology and software. From the results of the case study for sample clinical trial data, TVCM was superior to LOCF ANCOVA and MMRM approaches in terms of evaluating the treatment effect coupled with time variation in the early phase of the treatment in particular.

Keywords: time-varying coefficient model, linear smoothing, last observation carried forward, repeated measures mixed-effects model

Selecting Variables in Two-Group Robust Linear Discriminant Analysis

Stefan Van Aelst and Gert Willems

Department of Applied Mathematics and Computer Science, Ghent University
Krijgslaan 281 S9, B-9000 Gent, Belgium.
Stefan.VanAelst@UGent.be, Gert.Willems@UGent.be

Abstract. We consider two-group robust linear discriminant rules that are obtained by replacing the empirical means and covariance in the classical discriminant rules by S or MM-estimates of location and scatter (see e.g. Croux and Dehon (2001)). We consider the problem of selecting the variables that are relevant for separating the two groups. To test which variables contribute significantly to the canonical variate, and thus the discrimination of the classes, we propose a test based on the bootstrap distribution of the canonical variate. The bootstrap is a powerful nonparametric way of obtaining inference. However, classical bootstrap methods are time-consuming when robust estimates are involved and may not be robust. To avoid these problems of the classical bootstrap, Salibian-Barrera and Zamar (2002) introduced the fast and robust bootstrap method to approximate the bootstrap distribution in a consistent way. The fast and robust bootstrap was first developed for robust regression and later extended to multivariate settings such as principal component analysis (Salibian-Barrera et al. (2006,2008)). We will use the fast robust bootstrap to estimate the sampling distribution of the canonical variate. Based on this distribution we can test whether each of the coefficients of the canonical variate differs significantly from zero or not. This test can be used in variable selection procedures in which, for example, the least relevant variables is removed from the model in a stepwise manner. The performance of the procedure will be investigated by simulations and the method will be illustrated with examples. Extensions to settings with more than two groups will also be discussed.

Keywords: bootstrap, linear discriminant analysis, robustness, variable selection

References

CROUX, C. and DEHON, C. (2001): Robust linear discriminant analysis using S-estimators. *The Canadian Journal of Statistics 29, 473-492.*

SALIBIAN-BARRERA, M., VAN AELST, S. and WILLEMS, G. (2006): PCA based on multivariate MM-estimators with fast and robust bootstrap. *Journal of the American Statistical Association 101, 1198–1211.*

SALIBIAN-BARRERA, M., VAN AELST, S. and WILLEMS, G. (2008): Fast and robust bootstrap. *Statistical Methods and Applications 17, 41–71.*

SALIBIAN-BARRERA, M. and ZAMAR, R. H. (2002): Bootstrapping robust estimates of regression. *The Annals of Statistics 30, 556–582.*

Separable Two-Dimensional Linear Discriminant Analysis

Jianhua Zhao[1], Philip L.H. Yu[2], and Shulan Li[3]

[1] School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming, 650221, China. *jhzhao.ynu@gmail.com*
[2] Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong. *plhyu@hku.hk*
[3] Department of Mathematics and Statistics, Yunnan University, Kunming, 650091, China. *lishulan0526@gmail.com*

Abstract. Several two-dimensional linear discriminant analysis LDA (2DLDA) methods have received much attention in recent years. Among them, the 2DLDA, introduced by Ye, Janardan and Li (2005), is an important development. However, it is found that their proposed iterative algorithm does not guarantee convergence. In this paper, we assume a separable covariance matrix of 2D data and propose separable 2DLDA which can provide a neatly analytical solution similar to that for classical LDA. Empirical results on face recognition demonstrate the superiority of our proposed separable 2DLDA over 2DLDA in terms of classification accuracy and computational efficiency.

Keywords: LDA, 2DLDA, two-dimensional data, face recognition

References

YE, J., JANARDAN, R. and LI, Q. (2005). Two-dimensional linear discriminant analysis. In *Advances in Neural Information Processing Systems 17*, pages 1569–1576.

Fast and Robust Classifiers Adjusted for Skewness

Mia Hubert[1] and Stephan Van der Veeken[2]

[1] Department of Mathematics - LStat, Katholieke Universiteit Leuven
Celestijnenlaan 200B, Leuven, Belgium, *Mia.Hubert@wis.kuleuven.be*
[2] Department of Mathematics - LStat, Katholieke Universiteit Leuven
Celestijnenlaan 200B, Leuven, Belgium, *Stephan.Vanderveeken@wis.kuleuven.be*

Abstract. In this paper we propose two new classification rules for skewed distributions. They are based on the adjusted outlyingness (AO), as introduced in Brys et al. (2005) and applied to outlier detection in Hubert and Van der Veeken (2008). The new rules combine ideas of AO with the classification method proposed in Billor et al. (2008). We investigate their performance on simulated data, as well as on a real data example. Throughout we compare the classifiers with the recent approach of Hubert and Van der Veeken (2010) which assigns a new observation to the group to which it attains the minimal adjusted outlyingness. The results show that the new classification rules perform better when the group sizes are unequal.

Keywords: robustness, classification, outlyingness

References

BILLOR, N., ABEBE, A., TURKMEN, A. and NUDURUPATI, S.V. (2008): Classification based on depth transvariations. *Journal of Classification 25, 249-260.*

BRYS, G., HUBERT, M., and ROUSSEEUW, P.J. (2005): A robustification of Independent Component Analysis. *Journal of Chemometrics 19, 364–375.*

HUBERT, M., and VAN DER VEEKEN, S. (2008): Outlier detection for skewed data. *Journal of Chemometrics 22, 235–246.*

HUBERT, M., and VAN DER VEEKEN, S. (2009): Robust classification for skewed data. *Advances in Data Analysis and Classification, in press.*

A New Approach to Robust Clustering in $\mathbb{R}^p$

Catherine Dehon[1] and Kaveh Vakili[1]

European Center for Advanced Research in Economics and Statistics.
50, Avenue Roosevelt CP 114 Brussels, Belgium *kvakili@ulb.ac.be*

Abstract. In this note we present a fresh look at an old problem: that of identifying groups of cohesive observations lying in moderately large spaces when the dataset is potentially contaminated by an unknown number of outliers. The solution we introduce here is invariant to affine transformations, does not place assumptions on the number of clusters, the function governing their distribution or the share of contamination by outliers. Finally, our procedure is supported by a scalable, stable and efficient.

Sparse Bayesian Hierarchical Model for Clustering Problems

Heng Lian[1]

Division of Mathematical Sciences
School of Physical and Mathematical Sciences
Singapore 637371
Singapore, *henglian@ntu.edu.sg*

Abstract. Clustering is one of the most widely used procedures in the analysis of microarray data, with the goal of discovering cancer subtypes based on observed heterogeneity of genetic marks between different tissues. It is well-known that in such high-dimensional settings, the existence of many noise variables can overwhelm the few signals embedded in the high-dimensional space. We propose a novel Bayesian approach based on Dirichlet process with a sparsity prior that simultaneously performs variable selection and clustering, and also discover variables that only distinguish a subset of the cluster components. Unlike previous Bayesian formulations, we use Dirichlet process (DP) for both clustering of samples as well as for regularizing the high-dimensional mean/variance structure. To solve the computational challenge brought by this double usage of DP, we propose to make use of a sequential sampling scheme embedded within Markov chain Monte Carlo (MCMC) updates to improve the naive implementation of existing algorithms for DP mixture models. Our method is demonstrated on a simulation study and illustrated with the leukemia gene expression dataset.

Keywords: Dirichlet process, Markov chain Monte Carlo, Sequential sampling, Sparsity prior

Censored Survival Data: Simulation and Kernel Estimates

Jiří Zelinka

Department of Mathematics and Statistics, Faculty of Science, Masaryk University Kotlářská 2, Brno, Czech Republic, *zelinka@math.muni.cz*

Abstract. Non-parametric estimates of survival and hazard function belongs to the basic instruments in survival analysis. In previous papers methods of kernel estimates involving growth models of cancer cells were designed by author's colleagues. To verify the quality of these methods the tests on the simulated data were suggested.

During the test procedure some theoretical problems appeared. They concerned especially additional requests for distribution of simulated censoring data. The problems were largely resolved and and estimation procedures were successfully tested on simulated data. This paper summarizes the achievements.

Research supported by MŠMT of Czech Republic, no. LC06024.

Keywords: hazard function, censoring, simulation, kernel estimate

References

COLLETT D.: *Modelling Survival Data in Medical Research.* Chapman & Hall/CRC: Boca Raton-London-New York-Washington, D.C., 2003.

HOROVÁ I., ZELINKA J. and BUDíKOVÁ M.: Estimates of Hazard Functions for Carcinoma Data Sets. *Environmetrics*, **17**, 239–255, 2006.

HOROVÁ I. and ZELINKA J.: (2006) Kernel Estimates of Hazard Functions for Biomedical Data Sets. In *Applied Biostatistics: Case studies and Interdisciplinary Methods*, Springer, 2006.

HOROVÁ I., POSPíŠIL Z. and ZELINKA J.: Semiparametric Estimation of Hazard Function for Cancer Patients, *Sankhya*, **69**, 494–513, 2008.

HOROVÁ I., POSPíŠIL Z. and ZELINKA J.: Hazard function for cancer patients and cancer cell dynamics, *Journal of Theoretical Biology*, **258**, 437-443, 2009.

MÜLLER H.G. and WANG J.L.: Nonparametric Analysis of Changes in Hazard Rates for Censored Survival Data: An alternative Change-Point Models. *Biometrika*, **77**(2), 305–314, 1990.

RAMLAU-HANSEN H.: Counting Processes Intensities by Means of Kernel Functions. *The Annals of Statistics*, **11**(2), 453–466, 1983.

TANNER M.A. and WONG W.H.: The Estimation of the Hazard Function from Randomly Censored Data by the Kernel Method. *The Annals of Statistics*, **11**(3), 989–993, 1983.

UZUNOGULLARI U. and WANG J.L.: A comparision of Hazard Rate Estimators for Left Truncated and Right Censored Data. *Biometrika*, **79**(2), 297–310, 1992.

WAND, I.P. and JONES, I.C.: *Kernel smoothing.* Chapman & Hall, London, 1995.

EM-Like Algorithms for Nonparametric Estimation in Multivariate Mixtures

Tatiana Benaglia[1], Didier Chauveau[2], and David R. Hunter[1]

[1] Department of Statistics, Pennsylvania State University, USA
[2] Université d'Orléans, UMR CNRS 6628 - MAPMO - BP 6759
45067 Orléans Cedex 2, France *didier.chauveau@univ-orleans.fr*

Abstract. We propose an iterative algorithm for nonparametric estimation for finite mixtures of multivariate random vectors which has connections with the EM algorithm. The vectors are assumed to have independent coordinates conditionally to their mixture component, but otherwise their density functions may be nonparametric, or may be partially specified (semiparametric). This algorithm is much more easily applicable than existing algorithms in the literature. Several versions of it can be defined, and in particular we discuss here adaptive bandwidth issues for the involved kernel density estimates. An illustration using our implementation in the mixtools package for the R statistical software is given.

Keywords: EM algorithm, kernel density estimation, multivariate mixture, nonparametric mixture.

Longitudinal Data Analysis Based on Ranks and Its Performance

Takashi Nagakubo[1] and Masashi Goto[2]

[1] Asubio Pharma Co., Ltd., Clinical Research & Development Department
Orix Akasaka 2-Chome Building 3F, 2-9-11 Akasaka, Minato-ku, Tokyo 107-8541, Japan
nagakubo.takashi.cw@asubio.co.jp

[2] Biostatistical Research Association, NPO.
2-22-10-A411 Kamishinden, Toyonaka-shi, Osaka 560-0085, Japan
info@bra.or.jp

Abstract. In this study, we examine data measured repeatedly for a single subject over time, which is called longitudinal data. In particular, repeated measures ANOVA(Winer *et al.*, 1991) is commonly used. However, assumption of normality is not always satisfied and validity of parametric approach as repeated measures ANOVA is suspected. As a way to weaken the requirements of such parametric approaches, we would like to consider the use of a method that utilizes rank and does not depend upon distribution.

Brunner & Puri (1996) defined relative effects to describe treatment effects in general nonparametric designs. Relative effects are drawn from empirical distributions and inferred by the rank of observations; accordingly, approximation using relative effects is called the rank empirical distribution (RED) method.

We applied the RED method to an actual case with longitudinal data, and compared the analysis results with those of repeated measures ANOVA. The results showed a case of differing results between repeated measures ANOVA and the RED method. Then, we conducted some simulations in which underlying distribution is supposed to be normal or skewed, and investigated whether for group effect, time effect and interaction the power of two method is different. As a result of the simulations, for group effect, time effect and interaction the power of both methods is almost the same in normally distributed data. And for group effect, time effect and interaction the power of RED method was higher than repeated measures ANOVA in skewed data. So the RED method is suggested to be useful for longitudinal data analysis.

Keywords: repeated measures, cumulative distribution function, relative effect

References

BRUNNER, E. and PURI, M. L. (1996). Nonparametric methods in design and analysis of experiments. *Handbook of Statistics 13, 631-703.*

WINER, B., BROWN, D. and MICHELS, K. (1991). *Statistical Principles in Experimental Design*, 3rd edition. McGraw-Hill.

Computational treatment of the error distribution in nonparametric regression with right-censored and selection-biased data

Géraldine Laurent[1] and Cédric Heuchenne[2]

[1] QuantOM, HEC-Management School of University of Liège
boulevard du Rectorat, 7 Bât.31, B-4000 Liège, Belgium,
G.Laurent@student.ulg.ac.be
[2] QuantOM, HEC-Management School of University of Liège
boulevard du Rectorat, 7 Bât.31, B-4000 Liège, Belgium,
C.Heuchenne@ulg.ac.be

Abstract. Consider the regression model $Y = m(X) + \sigma(X)\varepsilon$, where $m(X) = E[Y|X]$ and $\sigma^2(X) = Var[Y|X]$ are unknown smooth functions and the error ε (with unknown distribution) is independent of X. The pair (X, Y) is subject to generalized selection bias and the response to right censoring. We construct a new estimator for the cumulative distribution function of the error ε, and develop a bootstrap technique to select the smoothing parameter involved in the procedure. The estimator is studied via extended simulations and applied to real unemployment data.

Keywords: Nonparametric regression, selection bias, right censoring, bootstrap, bandwidth selection

Local or Global Smoothing? A Bandwidth Selector for Dependent Data

Francesco Giordano[1] and Maria Lucia Parrella[1]

University of Salerno - Department of Economics and Statistics
Via Ponte Don Melillo, 84084 Fisciano (SA), Italy
giordano@unisa.it, mparrella@unisa.it

Abstract. The selection of the smoothing parameter represents a crucial step in local polynomial regression, due to the implications on the consistency of the non-parametric estimator and to the difficulties in the implementation of the selection procedure. In order to capture the complexity of the unknown regression curve, a local variable bandwidth may be used, but this may increase the variability of the estimates and the computational costs. This paper focuses on the problem of estimating the smoothing parameter adaptively on the support of the function, after evaluating the effective gain in using a local bandwidth rather than a global one.

Keywords: kernel regression, variable bandwidth selection, dependent data.

On Computationally Complex Instances of the *c*-optimal Experimental Design Problem: Breaking *RSA*-based Cryptography via *c*-optimal Designs

Michal Černý,[1] Milan Hladík[2] and Veronika Skočdopolová[1]

[1] University of Economics Prague, Department of Econometrics
Winston Churchill Square 4, 130 67 Prague, Czech Republic, *cernym@vse.cz, veronika.skocdopolova@vse.cz*
[2] Charles University, Department of Applied Mathematics
Malostranské náměstí 25, 118 00 Prague, Czech Republic, *hladik@mff.cuni.cz*

Abstract. We study the computational complexity of the problem to find a ***c***-optimal experimental design over a finite experimental domain. We construct instances of the problem which are computationally very difficult: we show how any algorithm for ***c***-optimality can be used for integer factoring and hence for breaking the RSA cryptographic protocol. These 'hard' instances can also be used as a benchmark for testing algorithms for finding ***c***-optimal designs.

Keywords: c-optimal experimental design, cryptography, RSA, integer factoring

References

ATKINSON, A., DONEV, A. and TOBIAS, R. (2007): *Optimum Experimental Designs with SAS*. Oxford University Press, Oxford.

ČERNÝ, M. and HLADÍK, M. (2010): Complexity of designing a *c*-optimal experiment over a finite experimental domain. Submitted in *Computational Optimization and Applications*. Preprint available at: `http://nb.vse.cz/~cernym/design.pdf`.

GAREY, M. R. and JOHNSON, D. S. (1979): *Computers and Intractability. A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York.

HARMAN, R. and JURÍK, T. (2008): Computing *c*-optimal experimental designs using the simplex method of linear programming. *Computational Statistics and Data Analysis 53, 247–254.*

PAPADIMIRIOU, C. H. (1995): *Computational Complexity*. Addison-Wesley Longman.

PÁZMAN, A. (1986): *Foundations of Optimum Experimental Design*. Reidel Publishing Company, Dordrecht.

PUKELSHEIM, F. and RIEDER, S. (1992): Efficient rounding in approximate designs. *Biometrika 79, 763–770.*

Fourier Analysis and Swarm Intelligence for Stochastic Optimization of Discrete Functions

Jin Rou New and Eldin Wee Chuan Lim

Department of Chemical & Biomolecular Engineering
National University of Singapore
4 Engineering Drive 4, Singapore 117576, *chelwce@nus.edu.sg*

Abstract. A new methodology for solving discrete optimization problems by the continuous approach has been developed in this study. A discrete Fourier series method was derived and used for re-formulation of discrete objective functions as continuous functions. Particle Swarm Optimization (PSO) was then applied to locate the global optimal solutions of the continuous functions derived. The continuous functions generated by the proposed discrete Fourier series method correlated almost exactly with their original model functions. The PSO algorithm was observed to be highly successful in achieving global optimization of all such objective functions considered in this study. The results obtained indicated that the discrete Fourier series method coupled to the PSO algorithm is indeed a promising methodology for solving discrete optimization problems via the continuous approach.

Keywords: Discrete Optimization, Fourier Series, Particle Swarm Optimization, Simulation, Global Optimization

Sub-quadratic Markov tree mixture models for probability density estimation

Sourour Ammar[1], Philippe Leray[1], and Louis Wehenkel[2]

[1] Knowledge and Decision Team
Laboratoire d'Informatique de Nantes Atlantique (LINA) UMR 6241
Ecole Polytechnique de l'Université de Nantes, France,
sourour.ammar@univ-nantes.fr, philippe.leray@univ-nantes.fr
[2] Department of Electrical Engineering and Computer Science & GIGA-Research, University of Liège, Belgium, *L.Wehenkel@ulg.ac.be*

Abstract. To explore the "Perturb and Combine" idea for estimating probability densities, we study mixtures of tree structured Markov networks derived by bagging combined with the Chow and Liu maximum weight spanning tree algorithm and we try to accelerate the research procedure by reducing its computation complexity below the quadratic and keepingg similar accuracy.

We empirically assess the performances of these heuristics in terms of accuracy and computation complexity, with respect to mixtures of bagged Markov trees described in Ammar et al. (2009), and single Markov tree CL built using the Chow and Liu algorithm (Chow and Liu (1968)).

Keywords: density estimation, mixture of trees, Perturb and Combine

References

AMMAR, S., LERAY, Ph., DEFOURNY, B. and WEHENKEL, L. (2009): Probability Density Estimation by Perturbing and Combining Tree Structured Markov Networks. In: ECSQARU '09: *Proceedings of the 10th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Springer-Verlag, 156–167.

CHOW, C.K. and LIU, C.N. (1968): Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory 14 (3), 462-467.*

Evolutionary Stochastic Portfolio Optimization and Probabilistic Constraints

Ronald Hochreiter[1]

Department of Finance, Accounting and Statistics, WU Vienna University of Economics and Business. Augasse 2-6, A-1090 Vienna, Austria.
ronald.hochreiter@wu.ac.at

Abstract. In this paper, we extend an evolutionary stochastic portfolio optimization framework to include probabilistic constraints. Both the stochastic programming-based modeling environment as well as the evolutionary optimization environment are ideally suited for an integration of various types of probabilistic constraints. We show an approach on how to integrate these constraints. Numerical results using recent financial data substantiate the applicability of the presented approach.

Keywords: Probabilistic constraints, portfolio optimization, stochastic optimization, evolutionary algorithms

The Effect of Estimating Parameters on Long-Term Forecasts for Cointegrated Systems

Hiroaki Chigira[1] and Taku Yamamoto[2]

[1] Faculty of Economics, Tohoku University
Kawauchi, Sendai 980-8576, Japan, *hchigira@econ.tohoku.ac.jp*
[2] College of Economics, NihonUniversity
1-3-2 Misaki-cho, Chiyoda-ku, Tokyo 101-8360, Japan,
yamamoto.taku@nihon-u.ac.jp

Abstract. This paper concerns with long-term forecasts for cointegrated processes. First, it considers the case where the parameters of the model are known. The paper analytically shows that neither cointegration nor integration constraint matters in long-term forecasts. It corrects misleading implications of previous influential studies, e.g., those by Engle and Yoo (1987) and Christoffersen and Diebold (1998) among others. The proper Mote Carlo experiment supports our analytical result.

Second, it considers the case where the parameters of the model are estimated. The paper shows that accuracy of the estimation of the drift term is crucial in long-trem forecasts. Namely, the relative accuracy of various long-term forecasts depends upon the relative magnitude of variances of estimators of the drift term. It further experimentally shows that in finite samples the univariate ARIMA forecast, whose drift term is estimated by the simple time average of differenced data, is better than the cointegrated sytem forecast, whose parameters are estimated by the well-known Johansen's maximum likelihood method. Based upon finite sample experiments, it recommends the univariate ARIMA forecast rather than the conventional cointegrated system forecast in finite samples for its practical usefulness and robustness against model misspecifications.

Keywords: time series, cointegrated process, long-term forecasts

References

CHRISTOFFERSEN, P. F. and DIEBOLD, F. X. (1998): Cointegration and long-horizon forecasting. *Journal of Business and Economic Statustucs 16,450-458.*

ENGLE, R. F. and YOO, S. (1987): Forecasting and testing in cointegrated systems. *Journal of Econometrics, 35, 143-159.*

Robustness of the Separating Information Maximum Likelihood Estimation of Realized Volatility with Micro-Market Noise

Naoto Kunitomo[1] and Seisho Sato[2]

[1] Graduate School of Economics, University of Tokyo, Bunkyo-ku, Hongo 7-3-1, Tokyo, Japan, *kunitomo@e.u-tokyo.ac.jp*
[2] Institute of Statistical Mathematics, Tachikawa-shi, Midoricho 10-3, Tokyo, Japan, *sato@ism.ac.jp*

Abstract. For estimating the realized volatility and covariance by using high frequency data, Kunitomo and Sato (2008a,b) have proposed the Separating Information Maximum Likelihood (SIML) method when there are micro-market noises. The SIML estimator, which is different from the class of estimation methods developed by Bandorff-Nielsen et al. (2008), has reasonable asymptotic properties; it is consistent and it has the asymptotic normality (or the stable convergence in the general case) when the sample size is large under general conditions including *non-Gaussian processes* and *volatility models*. We show that the SIML estimator has the asymptotic robustness in the sense that it is consistent and it has the asymptotic normality when there are autocorrelations in the market noise terms and there are endogenous correlations between the signal and noise terms. Some simulation results are given.

Keywords: Realized Volatility with Micro-Market Noise, High-Frequency Data, Separating Information Maximum Likelihood (SIML), Endogenous Noise, Aotocorrelated Noise, Asymptotic Robustness

References

BARNDORFF-NIELSEN, O., P. HANSEN, A. LUNDE and N. SHEPHARD (2008), : Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise, *Econometrica*, Vol.76-6, 1481-1536.

KUNITOMO, N. and S. SATO (2008a) : Separating Information Maximum Likelihood Estimation of Realized Volatility and Covariance with Micro-Market Noise, Discussion Paper CIRJE-F-581, Graduate School of Economics, University of Tokyo, (http://www.e.u-tokyo.ac.jp/cirje/research/dp/2008).

KUNITOMO, N. and S. SATO (2008b) : Realized Volatility, Covariance and Hedging Coefficient of Nikkei-225 Futures with Micro-Market Noise, Discussion Paper CIRJE-F-601, Graduate School of Economics, University of Tokyo.

Augmented Likelihood Estimators for Mixture Models

Markus Haas[1], Jochen Krause[2], and Marc S. Paolella[2]

[1] Financial Econometrics, Department of Statistics
Ludwig-Maximilians-University Munich
Akademiestr. 1/I, 80799 Munich, Germany
[2] Empirical Finance, Swiss Banking Institute
University of Zurich
Plattenstrasse 14, 8032 Zurich, Switzerland

Abstract. The maximum likelihood estimation of mixture models is well-known to suffer from the degeneracy of mixture components usually caused by singularities in the surface of the likelihood function. We present a new solution to this problem based on an augmented maximum likelihood scheme dedicated to mixture models and derive different estimators which avoid degeneracy. For some of them, consistency is ensured. The methodology is general and can straightforwardly be applied to arbitrary mixture distributions as well as mixture models of higher complexity, e.g., mixture GARCH models. Simulation studies show that the new estimators perform well even for relatively small sample sizes, precisely when their need particularly arises.

Keywords: Mixture Distribution, Maximum Likelihood, Degeneracy

References

COHEN, A. C. (1967): Estimation in mixtures of two normal distributions. *Technometrics 9 (1), 15-28.*

DAY, N. E. (1969): Estimating the components of a mixture of normal distributions. *Biometrika 56 (3), 463-474.*

HATHAWAY, R. J. (1985): A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics 13 (2), 795-800.*

KIEFER, N. M. (1978): Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica 46 (2), 427-434.*

REDNER, R. A. AND WALKER, H. F. (1984): Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review 26 (2), 195-239.*

TANAKA, K. (2009): Strong consistency of the maximum likelihood estimator for finite mixtures of locationscale distributions when penalty is imposed on the ratios of the scale parameters. *Scandinavian Journal of Statistics 36 (1), 171-184.*

Using clustering techniques to defining customer churn in a non-contractual setting

Mónica Clemente and Susana San Matías

Departamento de Estadística e Investigación Operativa Aplicadas y Calidad,
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain
mclement@eio.upv.es, *ssanmat@eio.upv.es*

Abstract. One of the most important tasks in customer relationship management (CRM) is probably the detection of a customer churn. Data mining techniques have been used to addressing this issue in contractual settings, as it can be found in the literature (Xie et al. (2009), for example). However, there are comparatively few approaches to churn detection in non-contractual settings, as in Buckinx and Poel (2005). The main reason is the difficulty to establishing a proper definition of churn in such contexts. In this paper, we propose the use of clustering techniques to obtain a set of different definitions, either for total and partial defection of customers. Using real data from a retail company, we compare them numerically from a managerial point of view and we discuss the advisability of their application.

Keywords: Churn, Data Mining, Clustering, CRM.

References

BUCKINX, W. and VAN DER POEL, D. (2005): Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research 164 (1), 252-268.*

XIE, Y., LI, X., NGAI, E.W.T. and YING, W. (2009): Customer churn prediction using improved balanced random forests. *Experts systems with applications 36 (3), 5445-5449.*

Regional Convergence in Japan: A Bayesian Spatial Econometrics Perspective

Kazuhiko Kakamu[1] and Hajime Wago[2]

[1] Faculty of Law and Economics, Chiba University
Yayoi-cho 1-33, Inage-ku, Chiba, 263-8522, Japan *kakamu@le.chiba-u.ac.jp*
[2] Department of Economics, Kyoto Sangyo University
Motoyama, Kamigamo, Kita-ku, Kyoto, 603-8555, Japan *wago@ism.ac.jp*

Abstract. Convergence hypothesis is one of the main themes in neoclassical growth theory. A lot of researches has developed in theoretical and empirical points of view. Temple (1999) gives an excellent survey regarding the problems in convergence hypothesis. In Japanese cases, for example, Barro and Sala-i-Martin (1992) compared the β- and σ-convergences using Japanese prefecture and US state data. On the other hand, although there is no theoretical background, Togo (2002) and Kakamu and Fukushige (2006) examined Markov transition matrices and showed that the Ergodic distributions have two or more peaks, that is, non-normality is observed from the empirical results in Japan.

This paper examines the regional convergence in 1986-2004 in Japan from a Bayesian point of view. To construct the model to examine regional convergence, we take into accounts two features of log per capita income in Japan: skewness and spatial interaction. We combine two-states (high and lower income states) normal mixture and spatial econometric models. From the empirical results, we can find that the σ-convergence is observed until 1997 only in higher income state and is not observed in lower income state. In addition, the source of skewness may be the changes of variance in higher income state and spatial interaction plays a weak but important role in Japan.

Keywords: convergence hypothesis, normal mixture model, Markov chain Monte Carlo (MCMC), spatial autoregressive model

References

BARRO, R.J. and SALA-I-MARTIN, X. (1992): Regional Growth and Migration: A US-Japan Comparison. *Journal of the Japanese and International Economies 6(4), 312-346.*

KAKAMU, K. and FUKUSHIGE, M. (2006): Productivity convergence of manufacturing industries in Japanese MEA. *Applied Economics Letters 13, 649-653.*

TEMPLE, J. (1999): The New Growth Evidence. *Journal of Economic Literature 37, 112-156.*

TOGO, K. (2002): Productivity convergence in Japan 's manufacturing industries. *Economics Letters 75, 61-67.*

A generalized confidence interval for the mean response in log-regression models

Miguel Fonseca[1], Thomas Mathew[2], and João Tiago Mexia[3]

[1] Center of Mathematics and Applications, New University of Lisbon
Faculdade de Ciências e Tecnologia, 2829-516 Caparica, Portugal,
fmig@fct.unl.pt
[2] Department of Mathematics and Statistics, University of Maryland, Baltimore County
1000 Hilltop Circle, Baltimore, MD 21250, USA, *mathew@umbc.edu*
[3] Department of Mathematics, Faculty of Sciences and Technology, New University of Lisbon
Faculdade de Ciências e Tecnologia, 2829-516 Caparica, Portugal, *jtm@fct.unl.pt*

Abstract. The interval estimation of a log-normal mean is investigated when the log-transformed data follows a regression model that also includes a random effect. The generalized confidence interval idea is used to derive the confidence interval. The performance of the interval is numerically investigated, and it is noted that the interval satisfactorily maintains the coverage probability. The proposed methodology is also illustrated using an example.

Keywords: coverage probability, generalized confidence interval, generalized pivotal quantity, random effects, lognormal distribution

References

IYER, H, WANG, J. M. and MATHEW, T. (2004). Models and confidence intervals for true values in interlaboratory trials. *Journal of the American Statistical Association. 99, 1060-1071.*

KRISHNAMOORTHY, K. and GUO, H. (2005). Assessing occupational exposure via the one-way random effects model with unbalanced data. *Journal of Statistical Planning and Inference.128, 219-229.*

KRISHNAMOORTHY, K. and MATHEW, T. (2002). Assessing occupational exposure via the one-way random effects model. *Journal of Agricultural, Biological and Environmental Statistics. 7, 440-451.*

KRISHNAMOORTHY, K. and MATHEW, T. (2003). Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals. *Journal of Statistical Planning and Inference. 115, 103 -121.*

KRISHNAMOORTHY, K. and MATHEW, T. (2009). *Statistical Tolerance Regions: Theory, Applications and Computation.* Wiley, New York.

LYLES, R. H., KUPPER, L. L. and RAPPAPORT, S. M. (1997). Assessing regulatory compliance of occupational exposures via the balanced one-way random effects ANOVA model. *Journal of Agricultural, Biological, and Environmental Statistics. 2, 64-86.*

Copula simulation by means of Adaptive Importance Sampling

Marco Bee[1]

Department of Economics, University of Trento
via Inama, 5, 38100, Trento, Italy, *marco.bee@unitn.it*

Abstract. In recent years there has been a considerable interest in copulas from both the theoretical and practical perspectives. Given p marginal distribution functions $F_1, \ldots, F_p$, copulas allow to construct a p-variate distribution with a given dependence structure and marginals $F_1, \ldots, F_p$. Thus, this way of proceeding separates the marginal distributions and the dependence structure.

General and efficient procedures for sampling copulas are not available, although this is an issue of primary importance. In some cases (elliptical copulas, some Archimedean copulas), the existing methods are satisfactory, but for other copulas (Archimedean with no closed-form Laplace transform, non-Archimedean), the only possibility consists in using the inverse of the conditional distribution function of a variable given the remaining ones, which is quite cumbersome, in particular in large-dimensional problems.

Adaptive Importance Sampling is an extension of standard importance sampling: at each iteration, the method produces a sample simulated from the instrumental density and used to improve the IS density itself. It was first introduced by Cappé *et al.* (2004), who propose to use a mixture distribution as instrumental density. Cappé *et al.* (2008) extend the algorithm to general mixture classes.

Here we employ Adaptive Importance Sampling for sampling various multivariate copulas. Extensive simulation experiments and a real-data application in the field of financial risk management show that the technique works well even when p is large. This is a remarkable result, since the performance of commonly used methods typically deteriorates very quickly as the dimension of the problem increases. Moreover, the experiments allow us to give precise indications about the choice of the number of mixture populations, the parameters of the population densities and the sample size, which are the main inputs of the algorithm. Finally, the method applies to any absolutely continuous distribution, so that it has a considerable potential for the simulation of other (i.e., non copula-based) multivariate distributions.

Keywords: Copulas, Adaptive Importance Sampling, Mixture distributions

References

CAPPE', O., GUILLIN, A., MARIN, J., ROBERT, C.P. (2004): Population Monte Carlo. *Journal of Computational and Graphical Statistics, 13 (4), 907-929.*

CAPPE', O., DOUC, R., GUILLIN, A., ROBERT, C.P. (2008): Adaptive Importance Sampling in General Mixture Classes. *Statistics and Computing, 18 (4), 447-459.*

Approximate Bayesian Computation with Indirect Moment Condition

Alexander Gleim[1] and Christian Pigorsch[2]

[1] Bonn Graduate School of Economics, Department of Economics, University of Bonn
Adenauerallee 24-42, D-53113 Bonn, Germany, *agleim@uni-bonn.de*
[2] Institute of Econometrics and Operations Research, Department of Economics, University of Bonn
Adenauerallee 24-42, D-53113 Bonn, Germany, *christian.pigorsch@uni-bonn.de*

Abstract. In genetics, Approximate Bayesian Computation (ABC) has become a popular estimation method for situations where the likelihood function of a model is unavailable. In contrast to classical, MCMC based Bayesian inference this method does not introduce missing variables to achieve a tractable model specification but instead relies on a distance function which measures the distance between some empirical moments and their population counterparts. An open question is the selection of these moments. In this paper we use an indirect approach with moment conditions based on the score of an auxiliary model as in the Efficient Method of Moments approach. We show that these moment conditions constitute a sufficient summary statistic for the auxiliary model and give conditions under which sufficiency carries over to the structural model of interest. Furthermore, an efficient way of weighting the different moment conditions is presented.

Keywords: Approximate Bayesian Computation (ABC), Efficient Method of Moments (EMM), indirect estimation method, Barndorff-Nielsen-Shephard (BNS) stochastic volatility model, Bayesian inference

Statistical Data Mining for Computational Financial Modeling

Ali Serhan Koyuncugil[1] and Nermin Ozgulbas[2]

[1] Capital Markets Board of Turkey, Department of Research
Eskisehir Yolu 8. Km. No: 156, Balgat, Ankara, Turkey,
askoyuncugil@gmail.com
[2] Baskent University, Department of Healthcare Management
Eskisehir Yolu 20 Km., Ankara, Turkey *ozgulbas@baskent.edu.tr*

Abstract. In this study, a data mining method which calls Chi-Square Automatic Interaction Detector (CHAID) decision tree algorithm has been used for detecting financial and operational risk indicators and financial risk profiles, developing a financial early warning system (FEWS) and obtaining financial road maps for risk mitigation. Therefore, small and medium sized enterprises (SMEs) in Turkey were covered and their financial and operational data was used for mentioned purposes. Financial and operational data of SMEs was obtained from Turkish Central Bank (TCB) after permission. Operational data which couldnt be access by balance sheets and income statements for financial management requirements of SMEs collected via a field study in Ankara. The study covered 7,853 SMEs financial data which was gathered from TCB and 1,876 SMEs operational data in year 2007. According to the financial data, SMEs were categorized into 31 different financial risk profiles, and it was found that 14 financial variables affected financial risk of SMEs. According to the operational data, SMEs were categorized into 28 different financial risk profiles and it was found that 14 operational variables affected financial risk of SMEs. As a result, SMEs in financial distress, operational and financial factors that affected financial risk, financial early warning signals, and road maps were defined automatically via mentioned profiles.

Keywords: statistical data mining, CHAID, finance, computational financial modeling

References

BREALEY, R. A., MYERS, S. C. and MARCUS, A. J. (2004): *Fundamentals of corporate finance*. 4th Edition.USA: Mc Graw Hill Inc..

KOYUNCUGIL, A.S. and OZGULBAS, N. (2010): Social aid fraud detection system and poverty map model suggestion based on data mining for social risk mitigation. In: A. S. Koyuncugil and N. Ozgulbas (Eds.): *Surveillance Technologies and Early Warning Systems: Data Mining Applicatons For Risk Detection*. IGI - Global, USA.

SUBRAMANYAM, K. R. and WILD J. (2009): *Financial statement analysis*. USA: Mc Graw Hill Inc..

Shooting Arrows in the Stock Market

Javier Arroyo[1] and Immanuel Bomze[2]

[1] Departamento de Ingeniería del Software e Inteligencia Artificial, Universidad Complutense, Profesor García-Santesmases s/n, 28040 Madrid, Spain
javier.arroyo@fdi.ucm.es
[2] ISDS, University of Vienna, Bruenner Strasse 72, 1210 Wien, Austria
immanuel.bomze@univie.ac.at

Abstract. In a recent paper, Arroyo (2010) has introduced a method for forecasting financial time series by locally weighted learning methods, based upon candlesticks. Candlesticks contain open/close prices and low/high prices of an equity within a certain trading period, plus the binary information whether the opening price exceeds the closing price or not. We complement this approach by using additional temporal information from the trading periods, namely the occurrence times of high and low, and take into account also the variation between closing price of the previous period, and opening price of the subsequent one. The data thus gathered resemble a zig-zag line like an arrow, and comprise all relevant information. In addition, de-trending can be done in a straightforward way as in Arroyo (2010).

Keywords: forecasting, k-nearest neighbour

References

ARROYO, J. (2010): Forecasting candlesticks time series with locally weighted learning methods. In: H. Locarek-June, C. Weihs (Eds.): *Classification as a Tool for Research*. Springer, DOI 10.1007/978-3-642-10745-0_66.

Smoothly Clipped Absolute Deviation for correlated variables

Mkhadri Abdallah[1], N'guessan Assi[2], and Sidi Zakari Ibrahim[1]

[1] Cadi Ayyad University, B.P. 2390, 40000, Marrakesh, Morocco
mkhadri@ucam.ac.ma i.sidizakari@ucam.ac.ma
[2] Polytech-Lille, University Lille1
59655 VILLENEUVE D'ASCQ CEDEX, Lille, France
Assi.N'Guessan@polytech-lille.fr

Abstract. We consider the problem of variable selection via penalized likelihood using nonconcave penalty fonctions. The Smoothly Clipped Absolute Deviation (SCAD) estimator, proposed by Fan and Li (2001), has promising theoretical properties, including continuity, sparsity, and unbiasedness. To maximize the nondifferential and nonconcave objective function, a new algorithm based on local linear approximation (LLA) was recently proposed by Zou and Li (2008) and which adopts naturally a sparse representation.This sparse representation is justified by the fact that some steps of the LLA are based on the well known LARS algorithm. Although LLA has promising theoretical properties, it inherits some drawbacks of Lasso in high dimensional setting.

To overcome these drawbacks, we propose a new algorithm based on a mixture of the linear and quadratic approximations (MLLQA) for maximizing the penalized likelihood for a large class of concave penalty functions.The importance of the quadratic approximation is that it enables correlated or grouped variables selection. The same idea was used by Zou and Hastie (2005) with the elastic-net in linear regression context. Some simulations and applications to real data sets are considered for comparing the performance of our method to its competitors.

Keywords: Variable Selection, SCAD, LLA algorithm, High dimension.

References

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R.(2004): Least angle regression. *The annals of statistics (32), 407-499*

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.*

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

ZOU, H. and HASTIE, T.(2005): Regularization and variable selection via the elastic-net. *Journal of the Royal Statistical Society (67), 301-320.*

ZOU, H. and LI, R.(2008): One-step sparse estimates in nonconcave penalized likelihood models. *The annals of statistics, volume 36, Number 4, 1509-1533*

Discrete wavelet preconditioning of Krylov spaces and PLS regression

Athanassios Kondylis[1] and Joe Whittaker[2]

[1] Philip Morris International R&D, Philip Morris Products S.A.
Computational Plant Biochemistry
Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland
athanassios.kondylis@pmintl.com

[2] Lancaster University, Department of Mathematics and Statistics
LA1 4YF Lancaster, United Kingdom
joe.whittaker@lancaster.ac.uk

Abstract. In order to predict a response of interest from a large number of interrelated predictors, PLS regression is used. Applying the discrete wavelet transform by means of preconditioning the normal equations the problem is solved on the wavelet domain. The relative importance of the wavelet coefficients is then used to rescale the solution in the transformed coordinates while the final solution is expressed in original terms by means of the inverse wavelet transform. The resulting solution is well adapted for large scale regression problems. It is especially suited for functional data commonly resulting from discretized functions related to modern instrumentation. The computational ease of the wavelet transform combined with its sparse properties is accelerated by the use of the relative importance measured on the wavelet coefficients instead of the original predictors. This allows us to obtain sparse and interpretable solutions in a computationally fast and efficient manner.

Keywords: discrete wavelet transform, preconditioning, Krylov spaces, PLS regression

References

Chui, C.K. (1992): *An Introduction to Wavelets*. Academic Press, London.

Trygg, J. and Wold, S. (1998): PLS regression on wavelet compressed NIR spectra *Chemometrics and Intelligent Laboratory Systems, 42, 209-220.*

Phatak, A. and de Hoog, F. (2002): Exploiting the connection between PLS, Lanczos, and Conjugate Gradients:: alternative proofs of some properties of PLS *Journal of Chemometrics 16, 361-367.*

Parametric and non-parametric multivariate test statistics for high-dimensional fMRI data

Daniela Adolf[1,2], Johannes Bernarding[1], and Siegfried Kropf[1]

[1] Otto-von-Guericke University, Institute for Biometrics and Medical Informatics Leipziger Str. 44, Magdeburg, Germany
[2] *daniela.adolf@med.ovgu.de*

Abstract. Via functional magnetic resonance imaging neuronal activation can be detected in a human brain. The signal response can be measured as a temporal series of three-dimensional values (voxels). These data are high-dimensional, because there are a few hundred scans (sample elements) but hundreds of thousands voxels (variables). Besides the spatial correlation of the voxels, there is also a correlation of temporally adjacent measurements.
Analyses of this high-dimensional fMRI data go beyond the scope of classical multivariate statistics. By default, fMRI data are analyzed voxelwise on the basis of univariate linear models, in which the temporal correlation is taken into account using a Satterthwaite approximation or a prewhitening method. We adapt these strategies in a multivariate context applying so-called stabilized multivariate tests (Läuter et al. (1996, 1998)) to the data, which are designed to cope with high-dimensional data and are based on the theory of left-spherical distributions.
These adapted parametric versions of the procedure need an adequate approximation of the temporal correlation. Usually, a first order autoregressive process is assumed for fMRI measurements and its correlation coefficient is estimated. Based on the stabilized multivariate test procedure, we propose a non-parametric approach of blockwise permutation including a random shift that renders an estimation of the temporal correlation structure unnecessary.
A comparison of the tests using simulated data shows in detail when which version of the tests is advantageous. All methods are finally applied to real fMRI data concerning socioeconomic decisions (Hollmann et al. (2010)).

Keywords: stabilized multivariate tests, block permutation including a random shift, correlated sample elements, fMRI data

References

HOLLMANN, M., MÜLLER, C., BAECKE, S., LÜTZKENDORF, R., ADOLF, D., RIEGER, J. and BERNARDING, J. (2010): Predicting Decisions in Human Social Interactions Using Real-Time fMRI and Pattern Classification. submitted to *Human Brain Mapping*.

LÄUTER, J., GLIMM, E. and KROPF, S. (1996): New Multivariate Tests for Data with an Inherent Structure. *Biometrical Journal 38, 5-23*.

LÄUTER, J., GLIMM, E. and KROPF, S. (1998): Multivariate Tests Based on Left-Spherically Distributed Linear Scores. *Annals of Statistics 26, 1972-1988*.

Robust Mixture Modeling Using Multivariate Skew t Distributions

Tsung-I Lin

Department of Applied Mathematics and Institute of Statistics
National Chung Hsing University, Taichung 402, Taiwan, *tilin@amath.nchu.edu.tw*

Abstract. This paper presents a robust mixture modeling framework using the multivariate skew t distributions, an extension of the multivariate Student's t family with additional shape parameters to regulate skewness. The proposed model results in a very complicated likelihood. Two variants of Monte Carlo EM algorithms are developed to carry out maximum likelihood estimation of mixture parameters. In addition, a general information-based method for obtaining the asymptotic covariance matrix of maximum likelihood estimates is discussed. Some practical issues including the selection of starting values as well as the stopping criterion are also discussed. The proposed methodology is applied to a subset of the Australian Institute of Sport data for illustration.

Keywords: MCEM algorithms, MSN distribution, MST distribution, multivariate truncated t distribution, outliers

References

LIN, T. I. (2009): Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis* 100, 257–265.

LIN, T. I., LEE, J. C., HSIEH, W. J. (2007): Robust mixture modeling using the skew t distribution. *Statistics and Computing* 17, 81–92.

LIN, T. I., LEE, J. C., YEN, S. Y. (2007): Finite mixture modelling using the skew normal distribution. *Statistica Sinica* 17, 909–927.

Robust scatter regularization

Gentiane Haesbroeck[1] and Christophe Croux[2]

[1] Department of Mathematics, University of Liège (B37), Grande Traverse 12, B-4000 Liège, Belgium, *G.Haesbroeck@ulg.ac.be.*
[2] ORSTAT and University Center of Statistics, K. U. Leuven, Naamsestraat 69, B-3000 Leuven, Belgium, *Christophe.Croux@econ.kuleuven.be*

Abstract. Estimating the location and scatter of data is a usual first step in many multivariate analyses. Often, the resulting scatter matrix needs to be regular for the application of the next steps. However, in practice (for example when there are more variables than observations), such a condition might be difficult to achieve. Regularization techniques are then necessary and usually consist of penalizing the likelihood function (e.g. Friedman et al. (2008)), yielding procedures sensitive to contamination in the data.

In this talk, robust regularization will be discussed. The idea is to regularize well known robust estimators, namely the M and MCD estimators, while ensuring that they keep their good breakdown behaviour. Algorithms to compute these regularized estimators will be described and the performance of the three proposals will be compared by a simulation study.

The talk will also consider some applications. First, the methodology will be used to detect outliers in high dimensional data settings. Then, as regularization and robustness are also necessary in graphical modeling (Finegold and Drton (2009)), it will be shown that robust graphical models can be constructed from the regularized robust scatter estimators.

Keywords: MCD estimator, Penalization, Breakdown point

References

FINEGOLD, M.A. and DRTON, M. (2009): Robust graphical modeling with t–distributions. *Proceeding of the 25th Conference on Uncertainty in Artificial Intelligence*, Quebec, 169–176 .

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008): Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9, 432-441.*

On Feature Analysis Methods for Collective Web Data

Ken Nittono

Depertment of Market Business Administration, Hosei University, 2-17-1 Fujimi, Chiyoda-ku, Tokyo, 102-8160, Japan, *nittono@hosei.ac.jp*

Abstract. A variety of methods and applications using data mining approaches for web data has been proposed and achieved success. The approaches are typically by the use of clustering, classification, retrieval methods and so on (Baldi et al. (2003)). On the other hand, in recent years, the programming scheme or libraries for those methods also have been developed energetically (Alag (2008), Segaran(2007)).

The aim of this research are to survey the analyzing methods for the collective web data, including corpus of texts, hyperlinks or retrieval keywords, and discover new approaches for such types of data. We make some comparative evaluations for typical methods via practical programming library and discuss the accumulation of knowledge on the web.

Keywords: web mining, data mining, collective intelligence

References

ALAG, S. (2008): *Collective Intelligence in Action*. Manning Publications Co., Greenwich.

BALDI, P., FRASCONI, P. and SMYTH, P. (2003): *Modeling the Internet and the Web*. John Wiley & Sons Ltd.

SEGARAN, T. (2007): *Programming Collective Intelligence: Building Smart Web 2.0 Applications*. O'Reilly.

Paired comparison or exhaustive classification to explain consumers' preferences

Salwa Benammou[1], Besma Souissi[2], and Abir Abid[3]

[1] Faculté de Droit et des Sciences Economiques et Politiques, Sousse, Tunisie & Computational Mathematics Laboratory, *saloua.benammou@fdseps.rnu.tn*
[2] Institut Supérieur de Gestion, Sousse, Tunisie,& Computational Mathematics Laboratory, *besma.swissi@yahoo.fr*
[3] Faculté de Droit et des Sciences Economiques et Politiques, Sousse, Tunisie & Computational Mathematics Laboratory, *abir.abid007@yahoo.fr*

Abstract. Based on experimental designs, conjoint analysis is a statistical method widely used in data analysis. Its main objective is to explain consumers' preferences for a product according to its attributes.
We are interested by the full profile method in conjoint analysis under its two forms: exhaustive classification and paired comparison which is more realistic. One of the major problems encountered while performing theses methods is the abundance of products presented to the interviewee. The number of pairs is much more important than the number of products which causes a theoretical loss of efficiency.
We show, empirically, through several real cases an equivalence between both forms of full profile in terms of model adjustment, market shares and importance's of utilities.
We tested these results on several examples and several numbers of products. We confirm our findings by simulations.

Keywords: conjoint analysis, paired comparison, exhaustive classification, D-efficiency.

References

Benammou, S. , Saporta, G and Souissi, B.(2007): Une procédure de réduction du nombre de paires en analyse conjointe. *Journal de la Société Française de Statistique, 148, (4).57- 76.*

Droesbeke, J.J. and Saporta, G. (1997): *Plans d'Expriences Applications l'Entreprise.* Editions Technip, Paris.

Green, P.E. and Srinivasan, V. (1990): Conjoint analysis in marketing: new developments with implications for research and practice. *Journal of Marketing, 3-19.*

Selecting an Optimal Mixed Effect Model Based on Information Criteria

Wataru Sakamoto

Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama-cho, Toyonaka, Osaka, Japan
sakamoto@sigmath.es.osaka-u.ac.jp

Abstract. In selecting an optimal random or mixed effect model from some candidate models in hierarchical structure of hypotheses, a true model is often located at the edge of the parameter space of the comprehensive model. In this situation, one of so-called *regularity conditions*, which guarantees asymptotic properties of maximum likelihood estimators, does not hold. Likelihood ratio tests on such hypotheses have been already considered in some balanced designs (e.g. Self and Liang (1987), Stram and Lee (1994), Crainiceanu and Ruppert (2004)). However, in the hypothetical tests, a smaller model, corresponding to a null hypothesis, would be chosen only in passive fashion.

Selecting an optimal mixed effect model with information criteria is considered. The second term in information criteria is obtained through evaluating asymptotical bias in estimating an expected log-likelihood $\mathrm{E}\{l(\theta)\}$ by a maximum log-likelihood $l(\hat{\theta})$. The bias was evaluated by a simulation approach in fitting linear mixed effect models to the longitudinal pig weight data (Diggle *et al.* (2002)). It was illustrated that, if the estimated model is over-parametrized in comparison with the true model, information criteria (such as AIC) built under regularity conditions could overestimate the bias.

Keywords: Longitudinal data, regularity conditions, restricted maximum likelihood

References

SELF, S. G. and LIANG, K. Y. (1987): Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc., 82, 605–610.*

STRAM, D. O. and LEE, J. W. (1994): Variance components testing in the longitudinal mixed effects model. *Biometrics, 50, 1171–1177.*

CRAINICEANU, C. M. and RUPPERT, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *J. Roy. Statist. Soc. Ser. B, 66, 165–185.*

DIGGLE, P. J., HEAGERTY, P., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data, Second Edition.* Oxford University Press.

Implementation of Moment Formula of Unitary Matrix Elements by Statistical Soft R and Its Applications

Toshio Sakata[1] and Kazumitsu Maehara[2]

[1] Department of Human Science & Faculty of Design, Kyushu University
4-9-1, Shiobaru, Minami-ku, Fukuoka, Japan, *sakata@design.kyushu-u.ac.jp*
[2] Graduate School of Design, Kyushu University
4-9-1, Shiobaru, Minami-ku, Fukuoka, Japan, *kazumits@gmail.com*

Abstract. In this paper we consider to implement the moment formula of unitary elements given in the paper of Matsumoto and Novak(2009) by statistical software **R**. The same formula is given by Collins and Sniady(2006), and the latter uses Schur-Weyl duality which are not familiar to statistical world. On the other hand the former expresses the formula only through a double summation on the subset of symmetric group with appropriately defined matrix G as a summand, and this is easy to implement by **R**. As an application we consider multivariate polynomials $p(\mathbf{x}, U)$ and $q(\mathbf{x}, U)$ with coefficients consisting of elements of random unitary matrix U. We ask whether two multivariate polynomial statistics $p(\mathbf{x}, U)$ and $q(\mathbf{x}, U)$ are related to each other by $p(\mathbf{x}, U) = q(\mathbf{x}, UU_0)$ where U_0 is a unknown fixed unitary matrix. Note that under the relation both have the same distribution because U is distributed as the Haar measure on the unitary group. So, for several polynomial pairs we performed, (1) looking of histograms, (2) Kolmogorov-Smirnov test and (3) exact calculations of the moments of the distributions by using the moment formula. Note that a generation of unitary matrix obeying the Haar measure is realized by a simple Gram-Schmidt method (see Ozolos(2009)). The results from three methods are compared.

Keywords: Integration over unitary group, moments formula, distributions of statistics on unitary group

References

Collins, B. and Śniady, P. (2006): Integration with respect to the Haar measure on unitary, orthogonal and sympletic group. *Comm. Math. Physics 264 (3),773-795.*

Matshumoto, S. and Novak, J. (2009): Jucys-Murphy elements and unitary matrix integrals. *ArXiv:0905.2009.*

Maris Ozols M. (2009): How to generate a random unitary matrix. *available on-line at http://home.lu.lv/ sd20008/papers/essays/Random unitary [paper].pdf.*

Spatial sampling design criterion for classification based on plug - in Bayes discriminant function

Kestutis Ducinskas[1] and Lina Dreiziene[2]

[1] Herkaus Manto str. 84, LT-92294 Klaipeda, Lithuania
kestutis.ducinskas@ku.lt
[2] Herkaus Manto str. 84, LT-92294 Klaipeda, Lithuania
lina.spss@gmail.com

Abstract. The problem considered is that of classifying the Gaussian random field observation into one of two populations specified by different regression mean models and common parametric covariance function for given stratified training sample. The ML estimators of unknown means and covariance parameters are plugged in the Bayes discriminant function. The approximation of the expected error rate associated with plug - in Bayes discriminant function is obtained. This is the extension of the previous one to the case of complete parametric uncertainty. Furthermore,the obtained approximation is proposed as the spatial sampling design criterion for classification based on plug - in Bayes discriminant function. The case of stationary Gaussian random field on lattice with exponential covariance function is used for illustrative examples. Numerical comparison of two spatial sampling designs by the proposed criterion is carried out.

Keywords: Bayes discriminant function, Spatial sampling design, Expected error rate

Evaluation of Deformable Image Registration Spatial Accuracy Using a Bayesian Hierarchical Model

Ying Yuan[1], Richard Castillo[2], Thomas Guerrero[2] and Valen E. Johnson[1]

[1] Department of Biostatistics
The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77230
yyuan@mdanderson.org
[2] Department of Radiation Oncology
The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77230

Abstract. Objective evaluation of deformable image registration (DIR) algorithms is necessary for the assessment of new medical imaging modalities; yet statistical methods for completing these evaluations are lacking. We propose a Bayesian hierarchical model to evaluate the spatial accuracy of DIR algorithms that is based on matched pairs of landmarks identified by human experts in source and target images. To fully account for the locations of landmarks in all images, we treat the true locations of landmarks as latent variables and impose a hierarchical structure on the magnitude of registration errors observed across image pairs. DIR registration errors are modeled using Gaussian processes with reference prior densities on prior parameters that determine the associated covariance matrices. We develop a Gibbs sampling algorithm and an approximated two-stage estimation procedure to efficiently fit our models to high-dimensional data.

Keywords: Image processing, latent variable, spatial correlation

Spatial Distribution of Trees

Makiko Oda[1], Fumio Ishioka[2], and Koji Kurihara[3]

[1] Graduate School of Environmental Science, Okayama University, 3-1-1 Tsushima-naka Kita Ward Okayama City Okayama 700-8530, JAPAN, *oda@ems.okayama-u.ac.jp*
[2] School of Law, Okayama University, 3-1-1 Tsushima-naka Kita Ward Okayama City Okayama 700-8530, JAPAN, *fishioka@law.okayama-u.ac.jp*
[3] Graduate School of Environmental Science, Okayama University, 3-1-1 Tsushima-naka Kita Ward Okayama City Okayama 700-8530, JAPAN, *kurihara@ems.okayama-u.ac.jp*

Abstract. Forest monitoring are conducted in the long run in many forests. The monitoring is performed with much time and energy to get data such as kinds of trees, position coordinates and diameters of trunk. The characteristic forests are attributed to many factors such as geographic features, temperature and light. Therefore, various approaches like nonparametric Bayesian inference, and Markov model are performed. One of the methods which decide the tree distribution type is point process (Shimatani (2001)). Tree data have various information not only the position but tree size and species. If the spot significantly concentrating big tree are indicated, the result is used in forest management in the future. We proposed Echelon analysis (Ishioka et al. (2007)) as this analysis method. Echelon analysis can objectively show the forest hierarchy construction. The objective forest is Ogawa in Japan. Data are in Forest Dynamics Database that have been compiled by the Forestry and Forest Products Research Institute. Masaki (2002) shows that shade-intolerant species are replacing shade-tolerant in this forest. We showed the tree distribution and usability of Echelon analysis based on this characteristics. We focused on the species diversity and showed the construction too.

Keywords: Tree distribution, Size distribution, Echelon analysis

References

Forestry and Forest Products Research Institute (2003): "Forest Dynamics Database. http://fddb.ffpri-108.affrc.go.jp/". Viewed August 20, 2009.

ISHIOKA, F., KURIHARA, K., SUITO, H., HORIKAWA, Y. and ONO, Y. (2007): Detection of Hotspots for Three-Dimensional Spatial Data and its Application to Environmental pollution Data. *Journal of Environmental Science for Sustainable Society 1, 15-24.*

MASAKI, T. (2002): Individual-based model of forest dynamics. In Nakashizuka,T. Matsumoto, Y. (eds.). Diversity and interaction in a temperate forest community-Ogawa Forest Reserve of Japan-. *Ecological Studies 158, Springer, Tokyo, 53-65.*

SHIMATANI, K. (2001): Multivariate point processes and spatial variation of species diversity. *Forest Ecology and Management 142, 215-229.*

Application of Local Influence Diagnostics to the Buckley-James Model

Nazrina Aziz[1] and Dong Qian Wang[2]

[1] Universiti Utara Malaysia
O6010, Sintok, Kedah Darul Aman,
Malaysia, *nazrina@uum.edu.my*
[2] Victoria University Of Wellington
New Zealand

Abstract. This article reports the development of local influence diagnostics of Buckley-James model consisting of variance perturbation, response variable perturbation and independent variables perturbation. The proposed diagnostics improves the previous ones by taking into account both censored and uncensored data to have a possibility to become an influential observation. Note that, in the previous diagnostics of Buckley-James model, influential observations merely come from uncensored observations in the data set. An example based on the Stanford heart transplant data is used for illustration. The data set with three covariates is considered in an attempt to show how the proposed diagnostics is able to handle more than one covariate, which is a concern to us as it is more difficult to identify peculiar observations in a multiple covariates.

Keywords: Buckley-James model, censored data, diagnostic analysis, local influence, product-limit estimator

Imputation by Gaussian Copula Model with an Application to Incomplete Customer Satisfaction Data

Meelis Käärik[1] and Ene Käärik[2]

[1] Senior researcher, Institute of Mathematical Statistics
University of Tartu, Estonia, *Meelis.Kaarik@ut.ee*
[2] Researcher, Institute of Mathematical Statistics
University of Tartu, Estonia, *Ene.Kaarik@ut.ee*

Abstract. We propose the idea of imputing missing value based on conditional distributions, which requires the knowledge of the joint distribution of all the data. The Gaussian copula is used to find a joint distribution and to implement the conditional distribution approach (Clemen and Reilly (1999)).

The focus remains on the examination of the appropriateness of an imputation algorithm based on the Gaussian copula.

In the present paper, we generalize and apply the copula model to incomplete correlated data using the imputation algorithm given by Käärik and Käärik (2009).

The empirical context in the current paper is an imputation model using incomplete customer satisfaction data. The results indicate that the proposed algorithm performs well.

Keywords: Gaussian copula, incomplete data, imputation

References

CLEMEN, R.T. and REILLY, T. (1999): Correlations and copulas for decision and risk analysis. *Management Science 45(2), 208-224.*

KÄÄRIK, E. and KÄÄRIK, M. (2009): Modelling dropouts by conditional distribution, a copula-based approach. *Journal of Statistical Planning and Inference, 139(11), 3830 - 3835.*

The Evaluation of Non-centred Orthant Probabilities for Singular Multivariate Normal Distributions

Tetsuhisa Miwa

National Institute for Agro-Environmental Sciences
3-1-3 Kannondai, Tsukuba 305-8604, Japan, *miwa@niaes.affrc.go.jp*

Abstract. We present a method to evaluate non-centred orthant probabilities for any singular multivariate normal distributions. Miwa *et al.* (2003) proposed a procedure for evaluating non-centred orthant probabilities accurately for non-singular multivariate normal distributions. Their procedure makes us enable to calculate any normal distribution functions. However, it was essential in their method that the covariance matrix should be non-singular.

In this paper we consider an m-dimensional normal distribution with any singular covariance matrix of rank n ($n < m$). It can be seen that the m-dimensional orthant probability is the probability volume of a polyhedron in the n-dimensional space. We show that a polyhedron can be expressed as differences between n-dimensional polyhedral cones. And each of these cones can be evaluated accurately by the procedure proposed by Miwa *et al.* (2003). Then we can evaluate any singular orthant probabilities and singular normal distribution functions. This procedure could be applied to many multiple comparison problems.

Keywords: multiple comparisons, normal distribution function, polyhedral cones, polyhedron

References

GENZ, A. and BRETZ, F. (2002): Comparison of methods for the computation of multivariate t probabilities. *Journal of Computational and Graphical Statistics, 11, 950-971.*

HAYTER, A. J. (1990): A one-sided studentized range test for testing against a simple ordered alternative. *Journal of American Statistical Association, 85, 778-785.*

MARCUS, R. (1976): The powers of some tests of equality of normal means against an ordered alternative. *Biometrika, 63, 177-183.*

MIWA, T. (1998): Bartholomew's test as a multiple contrast test and its applications. *Japanese Journal of Biometrics, 19, 1-9.*

MIWA, T., HAYTER, A. J. and KURIKI, S. (2003): The evaluation of general non-centred orthant probabilities. *Journal of the Royal Statistical Society, Ser. B, 65, 223-234.*

On the use of Weighted Regression in Conjoint Analysis

Salwa Benammou[1], Besma Souissi[2], and Gilbert Saporta[3]

[1] Faculté de Droit et des Sciences Economiques et Politiques, Sousse, Tunisie, *saloua.benammou@fdseps.rnu.tn*
[2] *Institut Supérieur de Gestion, Sousse, Tunisie, besma.swissi@yahoo.fr*
[3] *Chaire de statistique appliquée & CEDRIC, CNAM 292 rue Saint Martin, Paris, France, gilbert.saporta@cnam.fr*

Abstract. Conjoint analysis seeks to explain an ordered categorical ordinal variable according to several variables using a multiple regression scheme. A common problem encountered, there, is the presence of missing values in classification-ranks. In this paper, we are interested in the cases where consumers provide a ranking of some products instead of rating these products (i.e. explained variable presents missing values).In order to deal with this problem, we propose a weighted regression scheme. We empirically show (in several cases of weighting) that, if the number of missing values is not too large, the data remain useful, and our results are close to those of the complete order. A simulation study confirms these findings.

Keywords: conjoint analysis, missing values, weighted regression

References

Benammou, S., Harbi, S. and Saporta, G.(2003): Sur l'utilisation de l'analyse conjointe en cas de réponses incompltes ou de non réponses. *Revue de Statistique Appliquée, 51, 31-55.*

Benammou, S. , Saporta, G and Souissi, B.(2007): Une procédure de réduction du nombre de paires en analyse conjointe. *Journal de la Société Française de Statistique, 148, (4).57- 76.*

Cattin, P. and Wittink, D.R. (1989): Commercial Use of Conjoint analysis: An Update. *Journal of Marketing, 53, .91-96.*

Green, P.E. and Srinivasan, V. (1990): Conjoint analysis in marketing: new developments with implications for research and practice. *Journal of Marketing, 3-19.*

Green, P.E. and Rao, V.R.(1971): Conjoint Measurement for Quantifying Judgment Data. *Journal of Consumers Research, 5 September, 103-123.*

Weisberg, S. (1985): *Applied linear regression.* John Wiley and Sons, inc, second edition, New York.

A Case Study of Bank Branch Performance Using Linear Mixed Models

Peggy Ng[1], Claudia Czado[2], Eike Christian Brechmann[2], and Jon Kerr[1]

[1] School of Administrative Studies, York University
4700 Keele Street, Toronto, Canada, *peggyng@yorku.ca*, *jonkerr@yorku.ca*
[2] Center for Mathematical Sciences, Technische Universität München
Boltzmannstr. 3, D-85747 Garching, Germany, *cczado@ma.tum.de*, *eike.brechmann@mytum.de*

Abstract. The assessment of performance and potential is central to decisions pertaining to the location of bank branches. A common method for evaluating branch performance is data envelope analysis in which in-branch variables are typically considered. This paper adopts an alternate methodology that quantifies the influence of local socio-economic variables on bank deposits (a common measure of performance) using linear mixed models (LMM). It also illustrates the potential of using LMM to build a predictive model to support branch location decisions.

Keywords: Bank performance, branching, linear mixed models

References

AVKIRAN, N. K. (1997): Models of retail performance for bank branches: predicting the level of key business drivers. *International Journal of Bank Marketing 15 (6), 224-237.*

BERGER, A. N. and HUMPHREY, D. B. (1997): Efficiency of financial institutions: International survey and directions for future research. *European Journal of Operational Research 98 (2), 175-212.*

BOUFOUNOU, P. V. (1995): Evaluating bank branch location and performance: A case study. *European Journal of Operational Research 87, 389-402.*

CHELST, K. R., SCHULTZ, J. P. and SANGHVI, N. (1988): Issues and decision aids for designing branch networks. *Journal of Retail Banking 10 (2), 5-17.*

DOYLE, P., FENWICK, L. and SAVAGE, G. P. (1979): Management planning and control in multi- branch banking. *Journal of Operational Research Society 30 (2), 105-111.*

GART, A. (1994): *Regulation, Deregulation, Reregulation.* Wiley, New York.

PINHEIRO, J. C. and BATES, D. M. (2000): *Mixed-effects models in S and S-PLUS.* Springer, New York.

PINHEIRO, J. C., BATES, D. M., DEBROY, S., SARKAR, D., and the R Core team (2009): nlme: Linear and Nonlinear Mixed Effects Models. *R package version 3.1-96.*

WEST, B. T., WELCH, K. B. and GALECKI, A. T. (2006): *Linear Mixed Models: A Practical Guide Using Statistical Software.* Chapman & Hall/CRC, Boca Raton.

Visualizing the Sampling Variability of Plots

Rajiv S. Menjoge[1] and Roy E. Welsch[2]

[1] Operations Research Center, M.I.T.
77 Massachussetts Avenue, Cambridge, MA, *menjoge@mit.edu*
[2] Sloan School of Management, M.I.T.
77 Massachussetts Avenue, Cambridge, MA, *rwelsch@mit.edu*

Abstract. A general method for providing a description of the sampling variability of a plot of data is proposed. The motivation behind this development is that a single plot of a sample of data without a description of its sampling variability can be uninformative and misleading in the same way that a sample mean without a confidence interval can be.

The method works by using bootstrap methods to generate several plots that could have arisen from different samples from the population, and then conveying the information given in the collection of plots by methodically selecting a few representative plots in the subset.

The method includes the capacity to incorporate prior knowledge and distributional assumptions and is useful in a broad range of situations. It is illustrated with a scatter plot example and a histogram example.

Keywords: uncertainty visualization, bootstrap, scatterplot

References

CHATEAU, R., LEBART, L. (1996): Assessing sample variability in visualization techniques related to principal component analysis: bootstrap and alternative simulation methods. *Computational Statistics* Prats, A. (ed.), 205-210.

EFRON, B. (1979): Bootstrap methods: another look at the jackknife. *Annals of Statistics 7, 1-26.*

HAERDLE, W. (1990): Applied Non-parametric Regression. *Oxford University Press.*

KORTE, B. and Vygen, J. (2000): *Combinatorial Optimization: Theory and Algorithms.* Springer, NY.

LEVINA, E., BICKEL, P. (2001): The EarthMover's Distance is the Mallows Distance: Some Insights from Statistics. *Proceedings of ICCV, 251-256.* Vancouver, Canada.

MARKOWITZ, H. (1952): Portfolio Selection. *Journal of Finance 7, 77-91.*

PELEG, S., WERMAN, M., and ROM, H." (1989): A Unified Approach to the Change of Resolution: Space and Gray-level. *IEEE Transactions on Pattern Analysis and Machine Intelligence 11, 739-742.*

RUBNER, Y., TOMASI, C., and GUIBAS, L.J. (1998): A metric for distributions with applications to image databases. *Proceedings of IEEE International Conference on Computer Vision, 59-66.* Bombay, India

Visualisation of Large Sized Data Sets: Constraints and Improvements for Graph Design

Jean-Paul Valois[1]

TOTAL Exploration Production, F64018 Pau Cedex, *jean-paul.valois@total.com*

Abstract. Background and history of Data Visualisation are reviewed and actualised taking into account recent workings from several fields. Continuity is outlined since old Playfair's intuitions (Spence and Wainer (2005)) until the recent conclusions in ergonomy or neurophysiology, these are found as largely agreing with classical publications by Bertin (1977) or Tukey (1990) for instance. Then the paper considers the constraints resulting from Large Sized Data Sets. Limitations of usual displays like histogram or scatter-plot can in fact be found even using medium sized Data Sets. So improvements are suitable, like for histograms. Displaying density of points appears as an important challenge. Examples from oil industry show that it can be useful both for 2D scatter plots and in the case of parallel coordinates plots, these last ones used for comparing subpopulations or alarming.

Keywords: Data Visualisation, Exploratory Data Analysis, parallel coordinates, statistical population density

References

BERTIN, J. (1977): *La graphique et le traitement graphique de l'information*, Flammarion.

SPENCE,I and WAINER, H (2005): Introduction to the reprint of *The Commercial and Polical Atlas and statistical breviary*, by Playfair, W. 1786, Cambridge University Press, 1-35.

TUKEY, J.W. (1990): Data-Based Graphics: Visual Display in the decades to come, *Statistical Science,5, 3, 327-339.*

Visualization techniques for the integration of rank data

Michael G. Schimek[1] and Eva Budinská[2]

[1] Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation
Auenbruggerplatz 2/V, 8036 Graz, Austria, *michael.schimek@medunigraz.at*
[2] Swiss Institute of Bioinformatics, Bioinformatics Core Facility
Quartier Sorge - Batiment Genopode, 1015 Lausanne, Switzerland,
eva.budinska@isb-sib.ch

Abstract. In consumer preference studies, in Web-based meta-search (Mamoulis et al., 2007) or in meta-analysis of microarray experiments (DeConde et al., 2006), we are confronted with ranked lists representing the same set of distinct objects. All these applications have in common that one is interested in the top-ranked objects with considerable overlap in their rankings across the lists. Consolidation of such lists requires the estimation of the truncation point beyond which the ordering of the objects is dominated by noise. This point can be obtained from an inference procedure due to Hall and Schimek (2008) which even works for irregular rankings and huge data sets. Before its execution, it is essential to specify the distance parameter δ. In this paper, graphical approaches for the δ-choice, as well as for the integration of the top-ranked objects, are introduced for the first time. A consolidated result based on irregular rankings can never be unique. Our tools aim to assist the user when selecting the most appropriate result for their research purpose. Finally, the new graphical tools are applied to the integration of microarray data from several experiments due to Sørlie et al. (2003). It is demonstrated that a more complete set of top-ranked genes compared to their findings can be obtained.

Keywords: data integration, microarray data, ranked list, statistical graphics, top-k list

References

DECONDE, R.P. et al. (2006): Combined results of microarray experiments: a rank aggregation approach. *Statistical Applications in Genetics and Molecular Biology 5, 1, article 15.*

HALL, P. and SCHIMEK, M.G. (2008): Inference for the top-k rank list problem. In: Brito, M.P. (Ed.): *COMPSTAT 2008. Proceedings in Computational Statistics.* Physica, Heidelberg, 433-444.

MAMOULIS, N. et al. (2007): Efficient top-k aggregation of ranked inputs. *ACM Transactions on Database Systems, 32, 3, article 19.*

SØRLIE, T. et al. (2003): Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA, 100, 8418-8423.*

Dealing with Nonresponse in Survey Sampling: an Item Response Modeling Approach

Alina Matei[1,2]

[1] Institute of Statistics, University of Neuchâtel,
Pierre à Mazel 7, 2000 Neuchâtel, Switzerland, *alina.matei@unine.ch*
[2] Institute for Pedagogical Research and Documentation (IRDP) Neuchâtel,
Fbg. de l'Hôpital 43, cp 556, 2002 Neuchâtel, Switzerland

Abstract. Dealing with nonresponse is a very important topic, since nonresponse is present almost in all surveys, and can cause biased estimation. Nonresponse is defined as the failure to provide the required information by a unit selected in a sample. We distinguish between unit nonresponse and item nonresponse. Unit nonresponse implies that we have no information at all from the sampled unit. Item nonresponse means that the sampled unit does not fill some of the survey items. Each unit selected in the sample has associated a sampling weight and a response probability to answer the questionnaire. The response probability is unknown and should be estimated. The main method to deal with the unit nonresponse is to use reweighting. This method adjusts the initial sampling weights by the inverse of the estimated response probabilities, providing new weights. We focus on unit nonresponse adjustment in survey data and estimate the response probabilities using an item response model called the Rasch model. This model uses a latent parameter. We believe that this latent parameter can explain a part of the unknown behavior of a unit to respond in the survey. No information about the nonrespondents and no auxiliary information are required in the proposed method. Theoretical aspects and simulations are used to support our theory.

Keywords: survey sampling, nonresponse, item response modeling

References

BAKER, F., and KIM, S.H. (2004): *Item Response Theory.* Marcel Dekker, New York, 2nd edition.

RASCH, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

RASCH, G. (1961): On general laws and the meaning of measurement in psychology, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV.* Berkeley: University of Chicago Press, 1961, *321–333.*

Using Auxiliary Information Under a Generic Sampling Design

Giancarlo Diana[1] and Pier Francesco Perri[2]

[1] Department of Statistical Sciences, University of Padova
Via Cesare Battisti, 241, 35121 Padova, Italy, *giancarlo.diana@unipd.it*
[2] Department of Economics and Statistics, University of Calabria
Via P. Bucci, 87036 Arcavacata di Rende, Italy, *pierfrancesco.perri@unical.it*

Abstract. The estimation of population parameters is a persistent issue in sampling from finite population when auxiliary variables are available. Many efforts have been made to estimate the mean (or total) through the ratio, product and regression estimation methods and a great deal of literature has been produced according to simple random sampling without replacement.

Under a generic sampling design we consider a very simple class of asymptotically unbiased estimators for the population mean. The minimum variance bound of the class is obtained and it is found that the best estimator attaining the bound is the regression one. It is theoretically shown that some recent estimators proposed by Bacanli and Kadilar (2008) belong to the class but are not optimum.

The efficiency gain of the regression estimator upon different estimators is numerically evaluated when sampling is performed according to probability proportional to size. In so doing, we consider the problem of estimating the first and second order inclusion probabilities when their exact determination becomes unfeasible. A simple estimation algorithm is implemented in the R environment and its stability ascertained through a number of simulation experiments. The main finding is that the estimated inclusion probabilities allow us to achieve satisfactory results both in terms of accuracy and in the reduction of time and memory required. Finally, the results demonstrate that Bacanli-Kadilar estimators do not work well when compared with the regression estimator.

Keywords: Horvitz-Thompson estimator, estimated inclusion probabilities, pps sampling, simulation

References

BACANLI, S., KADILAR, C. (2008): Ratio estimators with unequal probability designs. *Pakistan Journal of Statistics 24 (3), 167-172.*

Estimating Population Proportions in Presence of Missing Data

Álvarez-Verdejo, E.[1], Arcos, A.[1], González, S.[2], Muñoz, J.F.[1] and Rueda, M.M.[1]

[1] University of Granada, Spain, *encarniav@ugr.es, arcos@ugr.es, jfmunoz@ugr.es, mrueda@ugr.es*
[2] University of Jaén, Spain *sgonza@ujaen.es*

Abstract. This paper discusses the estimation of a population proportion in the presence of missing data and using auxiliary information at the estimation stage. A general class of estimators, which makes an efficient uses of the available information, is proposed. Some theoretical properties of the proposed estimators are analyzed, and they allow us to find the optimum values in each proposed class. Optimum estimators are thus more efficient than the customary estimator. In particular, the estimator based on the difference method is an optimal estimator in the sense it has minimal variance into the class. Results derived from a simulation study show that the proposed optimum estimators give desirable results in comparison to alternative estimators.

Keywords: Auxiliary information, ratio and difference estimators

References

LITTLE, R.J.A. and RUBIN, D.B. (1987): *Statistical analysis with missing data*. John Wiley, New York.

RANDLES, R.H. (1982): On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics 10, 462-474*.

RUEDA, M. and GONZÁLEZ, S. (2004): Missing data and auxiliary information in surveys. *Computational Statistic 19 (4), 555-567*.

RUEDA, M., MUÑOZ, J.F., BERGER, Y.G., ARCOS, A. and MARTÍNEZ, S. (2007): Pseudo empirical likelihood method in the presence of missing data. *Metrika 65, 349-346*.

SRIVASTAVA, S.K. and JHAJJ, H.S. (1981): A class of estimators of the population mean in surve sampling using auxiliary information. *Biometrika 68, 341–343*.

TOUTENBURG, H. and SRIVASTAVA, V.K. (1998): Estimation of ratio of population means in survey sampling when some observations are missing. *Metrika 48, 177-187*.

TRACY, D.S. and OSAHAN, S.S. (1994): Random nonresponse on study variable versus on study as well as auxiliary variables. *Statistica, 54, 163-168*.

Application of a Bayesian Approach for Analysing Disease Mapping Data: Modelling Spatially Correlated Small Area Counts

Mohammadreza Mohebbi[1] and Rory Wolfe[1]

1. Department of Epidemiology and Preventive Medicine, Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, Australia, *Mohammadreza.Mohebbi@med.monash.edu.au*

Abstract. Maps of regional disease rates are potentially useful tools for examining spatial patterns of disease. In recent years, models including both spatially correlated random effects and spatially unstructured random effects have been very popular for this purpose (Banerjee et al., 2004). We used a full Bayesian approach to estimate region-specific geographically-correlated disease rates. We fitted three-stage hierarchical models in which the disease counts were modelled as a function of area-specific relative risks at stage one; the collection of relative risks across the study region were modelled at stage two; and at stage three prior distributions were assigned to parameters of the stage two distribution. This Bayesian hierarchical model framework included simultaneous modelling of iid random effect and spatially correlated random effect. The spatial random effects were modelled via Bayesian prior specifications reflecting spatial heterogeneity globally and relative homogeneity among neighbouring areas. Estimation was implemented using MCMC methods. The MCMC algorithm convergence was examined using a number of convergence diagnostics. To illustrate the procedures, we present an analysis of esophageal cancer incidence in the Caspian region of Iran. The goal of the analysis was to produce smoothed small-area estimates of cancer incidence ratios (Mohebbi et al., 2008). Diagnostic measures were examined and different modelling strategies were implemented. The results suggested that models based on the use of spatial random effects appear to work well and provide a robust basis for inference.

Keywords: Bayesian inference, Disease mapping, Ecologic regression, Poisson regression, Spatial correlation

References

MOHEBBI, M., MAHMOODI, M., WOLFE, R., NOURIJELYANI, K., MOHAMMAD, K., ZERAATI, H. and FOTOUHI, A. (2008): Geographical spread of gastrointestinal tract cancer incidence in the Caspian Sea region of Iran: Spatial analysis of cancer registry data *BMC Cancer 8, 137*.

BANERJEE, S., CARLIN, B. and GELFAND, A. (2004): *Hierarchical Modeling And Analysis For Spatial Data*. Chapman and Hall/CRC, Boca Raton.

A Mann-Whitney spatial scan statistic for continuous data

Lionel Cucala

Université des Sciences et Techniques du Languedoc
Place Eugène Bataillon, Montpellier, France, *lcucala@math.univ-montp2.fr*

Abstract. A new scan statistic is proposed for identifying clusters of high or low values in georeferenced continuous data. On the one hand, it relies on a concentration index which is based on the Mann-Whitney statistic and thus is completely distribution-free. On the other hand, the possible spatial clusters are given by an original graph-based method. This spatial scan test seems to be very powerful against any arbitrarily-distributed cluster alternative. These results have applications in various fields, such as the epidemiological study of rare diseases or the analysis of astrophysical data.

Keywords: Cluster detection, epidemiology, scan statistics, spatial marked point processes.

Detection of Spatial Cluster for Suicide Data using Echelon Analysis

Fumio Ishioka[1], Makoto Tomita[2], and Toshiharu Fujita[3]

[1] School of Law, Okayama University, Okayama 700-8530, Japan, *fishioka@law.okayama-u.ac.jp*
[2] Clinical Research Center, Faculty of Medicine, Tokyo Medical and Dental University, Tokyo 113-8519, Japan, *tomita.crc@tmd.ac.jp*
[3] The Institute of Statistical Mathematics, Research Organization of Information and Systems, Tokyo 106-8569, Japan, *fujita-t@ism.ac.jp*

Abstract. Recently, the number of suicides in Japan increases rapidly. For this problem, it is clear that a statistical implication is important. Our data consists of male suicide data from Kanto area in central part of Japan during 1973-2007. In this paper, we investigate the transition and the tendency of male suicides by detecting geographical spatial cluster. It is performed by echelon scan which we have proposed as one of the cluster detection. Furthermore, the performance of the cluster detection based on echelon analysis is compared to the cluster based on previous study.

Keywords: spatial data, spatial scan statistic, geographical clusters, echelon analysis

References

Cabinet Office. (2008): White Book for Strategy to Prevent Suicide. Saiki Printing Co.

FUJITA, T. (2009): Statistics of Community for the Death from Suicide. National Institute of Mental Health, National Center of Neurology and Psychiatry, Japan.

KULLDORFF, M. (1997): A spatial scan statistics. *Communications in Statistics, Theory and Methods, 26, 1481-1496.*

KULLDORFF, M. (2006): Information Management Services Inc: SaTScan v7.0: Software for the spatial and space time scan statistics, http://www.satscan.org/.

A Comparison between Two Computing Methods for an Empirical Variogram in Geostatistical Data

Takafumi Kubota[1] and Tomoyuki Tarumi[2]

[1] Okayama University, Graduate school of humanities and social sciences
Tsushimanaka 3-1-1 Okayama, Japan, *kubota@law.okayama-u.ac.jp*
[2] Okayama University, Admission Centre
Tsushimanaka 3-1-1 Okayama, Japan, *t2@ems.okayama-u.ac.jp*

Abstract. In this paper, we propose a new calculation method for an empirical variogram, which the range of distance of points are divided to equal number of observation pairs. Then, we do both simulation study and application for Meuse river data set (Burrough and McDonnell(1998)) in order to compare our proposal calculation method with traditional calculation method for an empirical variogram, which the range of points are divided to equal distance.

Keywords: Geostatistics, empirical variogram, kriging, Cross-Validation

References

Burrough, P.A. and McDonnell, R.A. (1998): Principles of Geographical Information Systems. *Oxford University Press.*

Kubota, T. Iizuka, M. Fueda, K. and Tarumi, T.(2005): The Selection of the Cutoff in Estimating Variogram Model. *The 5th IASC Asian Conference on Statistical Computing.* 97–100

Kubota, T. and Tarumi, T.(2008): Using Geometric Anisotropy in Variogram Modeling. *COMPSTAT2008 Proceedings in Computational Statistics.* 793–801

Monotone Graphical Multivariate Markov Chains

Roberto Colombi[1] and Sabrina Giordano[2]

[1] Dept. of Information Technology and Math. Methods, University of Bergamo
viale Marconi, 5, 24044 Dalmine (BG), Italy, *colombi@unibg.it*
[2] Dept. of Economics and Statistics, University of Calabria
via Bucci, 87036 Arcavacata di Rende (CS), Italy, *sabrina.giordano@unical.it*

Abstract. When multivariate categorical data are collected over time, the dynamic character of their association must be taken into account. This aspect plays an important role in modelling discrete time-homogeneous multivariate Markov chains (MMC).

We show that a deeper insight into the relations among marginal processes of an MMC can be gained by testing hypotheses of Granger non-causality, contemporaneous independence and monotone dependence coherent with a stochastic ordering.

The tested hypotheses are associated to a multi edge graph where the nodes represent the univariate components of the MMC and directed and bi-directed edges describe the dependence among them. These hypotheses are proven to be equivalent to equality and inequality constraints on certain interactions of a multivariate logistic model parameterizing the transition probabilities.

As the parameter space under the null hypothesis is specified by inequality constraints, the likelihood ratio statistic has a chi-bar-square asymptotic distribution whose tail probabilities can be computed by Monte Carlo simulations.

The introduced hypotheses are tested on real categorical time series. The procedures here used for the maximum likelihood estimation and hypothesis testing under equality and inequality constraints are implemented in the R-package *hmmm*, available from the authors.

Keywords: graphical models, Granger causality, stochastic orderings, chi-bar-square distribution

References

COLOMBI, R. and FORCINA, A. (2001): Marginal regression models for the analysis of ordinal response variables. *Biometrika, 88, 1007-1019.*

COLOMBI, R. and GIORDANO, S. (2009): Multi edges graphs for multivariate Markov chains. In: J. G. Booth (Eds.): *Proceedings of 24th International Workshop on Statistical Modelling.* Ithaca (NY), 102 – 109.

SEN, P. K. and SILVAPULLE, M. J. (2005): *Constrained Statistical Inference.* Wiley, New-Jersey.

An Exploratory Segmentation Method for Time Series

Christian Derquenne

Electricité de France R&D
1, avenue du Général de Gaulle, Clamart, France, *christian.derquenne@edf.fr*

Abstract. Generally, time series are decomposed in several behaviours: trend, seasonality, volatility and noise. Due to the regularity of the series, it could be more or less easy to decompose them as this scheme. Then it can be interesting to detect behaviour breakpoints when pre or post-treating data. Building of sub-models on each detected segment, achieving stationarity of time series with a segmentation model, building symbolic curves to cluster series, modelling multivariate time series are so many examples in this context. Many methods have been and are developed to answer different issues in economics, finance, human sequence, meteorology, energy management, etc. Several groups of methods exist: exploring the segmentation space for the assessment of multiple change-point models (Guédon (2008)) to inference on the models with multiple breakpoints in multivariate time series, notably to select optimal number of breakpoints (Lavielle et al. (2006)). Most algorithms use dynamic programming to decrease computation complexity of segmentations, because it would be illusory to calculate all segmentations. These methods of breakpoints detection aims at answering three detection problems: change mean with a constant variance, change of variance with a constant mean and change for overall distribution of time series without change of level, in dispersion and on the distribution of errors. The method proposes to segment a time series. It offers an original process with a first step of preparing data which is crucial to build the most adequate structure to initialize the second step of modelling an heteroskedastic linear model including the different trends, levels and variances. The proposed method allows also to reduce the computation complexity in $O(KT)$, where T is length of series and K is generally less than to $\sqrt{T}$. Lastly, this method is rather preliminary and we work to improve some steps of this method, particularly on the detection of volatility in data and on the evaluation and the validation tools of segmentations to obtain a better means to have a hierarchy of these last ones.

Keywords: segmentation, change-point, time series, variance components

References

GUEDON, Y. (2008): Exploring the segmentation space for the assessment of multiple change-point models. Institut National de Recherche en Informatique et en Automatique, *Cahier de recherche 6619*.

LAVIELLE, M. and TEYSSIERRE, G. (2006): Détection de ruptures multiples dans des séries temporelles multivariées. *Lietuvos Matematikos Rinikinys, Vol 46*.

Multiple Change Point Detection by Sparse Parameter Estimation

Jiří Neubauer[1] and Vítězslav Veselý[2]

[1] Department of Econometrics, University of Defence Kounicova 65, Brno, Czech Republic, *Jiri.Neubauer@unob.cz*
[2] Department of Applied Mathematics and Computer Science, Masaryk University, Lipová 41a, Brno, Czech Republic, *vesely@econ.muni.cz*

Abstract. The contribution is focused on multiple change point detection in a one-dimensional stochastic process by sparse parameter estimation from an overparametrized model. Stochastic process with changes in the mean is estimated using dictionary consisting of Heaviside functions. The basis pursuit algorithm is used to get sparse parameter estimates. Some properties of mentioned method are studied by simulations.

Keywords: multiple change point detection, overparametrized model, sparse parameter estimation, basis pursuit algorithm

References

BRUCKSTEIN, A. M. et al. (2009): From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images. *SIAM Review 51 (1), 34–81.*

CHEN, S. S. et al. (1998): Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput. 20 (1), 33-61* (2001 reprinted in *SIAM Review 43 (1), 129–159).*

CHRISTENSEN, O. (2003): *An introduction to frames and Riesz bases.* Birkhuser, Boston-Basel-Berlin.

NEUBAUER, J. and VESELÝ, V. (2009): Change Point Detection by Sparse Parameter Estimation. In: *The XIIIth International conference: Applied Stochastic Models and Data Analysis.* Vilnius, 158–162.

SAUNDERS, M. A. (1997–2001): *pdsco.m: MATLAB code for minimizing convex separable objective functions subject to* $Ax = b, x \geq 0$.

VESELÝ, V. (2001–2008): *framebox: MATLAB toolbox for overcomplete modeling and sparse parameter estimation.*

VESELÝ, V. and TONNER, J. (2005): Sparse parameter estimation in overcomplete time series models. *Austrian Journal of Statistics, 35 (2&3), 371–378.*

VESELÝ, V. et al. (2009): Analysis of PM10 air pollution in Brno based on generalized linear model with strongly rank-deficient design matrix. *Environmetrics, 20 (6), 676–698.*

ZELINKA et al. (2004): Comparative study of two kernel smoothing techniques. In: Horová, I. (ed.) *Proceedings of the summer school DATASTAT'2003, Svratka.* Folia Fac. Sci. Nat. Univ. Masaryk. Brunensis, Mathematica 15: Masaryk University, Brno, Czech Rep., 419–436.

M-estimation in INARCH Models with a Special Focus on Small Means

Hanan El-Saied[1] and Roland Fried[1]

Department of Statistics, TU Dortmund University
Vogelpothsweg 87, 44221 Dortmund, Germany, {*saied, fried*}*@statistik.tu-dortmund.de*

Abstract. We treat robust M-estimation of INARCH-models for count time series. These models assume the observation at each point in time to follow a Poisson distribution conditionally on the past, with the mean being a linear function of previous observations. This simple linear structure allows to transfer M-estimators for autoregressive models to this situation, with some simplifications being possible because the conditional variance given the past equals the conditional mean. The situation of a small mean deserves special attention because of the strong asymmetry of such Poisson distributions. The proposed generalized M-estimators show good performance in simulations.

Keywords: Time series, Outliers, Robustness, Huber M-estimator, Tukey M-estimator

References

BOLLERSLEV, T. (1986): Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31, 307-327.*

CADIGAN, N.G. and CHEN, J. (2001): Properties of robust M-estimators for Poisson and negative binomial data. *Journal of Statistical Computation and Simulation 70, 273-288.*

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

FERLAND, R., LATOUR, A. and ORAICHI, D. (2006): Integer–valued GARCH processes. *Journal of Time Series Analysis 27, 923-942.*

FOKIANOS, K., RAHBEK, A. and TJØSTHEIM, D. (2009): Poisson autoregression. *Journal of the American Statistical Association 104, 1430-1439.*

FOKIANOS, K. and FRIED, R. (2010): Interventions in INGARCH processes. *Journal of Time Series Analysis*, forthcoming.

SIMPSON, D.G., CARROLL, R.J. and RUPPERT, D. (1987): M-estimation for discrete data: asymptotic distribution theory and implications. *Annals of Statistics 15, 657-669.*

Rplugin.Econometrics: R-GUI for teaching Time Series Analysis

Dedi Rosadi[1]

Department of Mathematics, Research Group: Statistics
Gadjah Mada University, Indonesia, *dedirosadi@ugm.ac.id*

Abstract. In this paper, we introduce `RcmdrPlugin.Econometrics` (Rosadi, Marhadi and Rahmatullah, 2009), a R-GUI for time series analysis, as a plug-in for R-Commander. `RcmdrPlugin.Econometrics` has nearly all of the typical models introduced in undergraduate time series courses, such as exponential smoothing, ARIMA/SARIMA, ARIMAX, ARCH/GARCH, etc. We discuss the philosophy of the plug-in design, compare and show its difference with a similar purpose R-Commander plug-in, called as `RcmdrPlugin.epack` (Hodgess and Vobach, 2008) and a commercial econometrics software EViews.

Keywords: R Commander Plug-ins, Open Source, Time Series Analysis

Fourier methods for sequential change point analysis in autoregressive models

Marie Hušková[1], Claudia Kirch[2], and Simos G. Meintanis[3]

[1] Charles University of Prague, Department of Statistics
Sokolovská 83, CZ – 186 75, Praha 8, Czech Republic,
marie.huskova@karlin.mff.cuni.cz
[2] Karlsruhe Institute of Technology, Institute for Stochastics
Kaiserstr. 89, 76133 Karlsruhe, Germany, *claudia.kirch@kit.edu*
[3] National and Kapodistrian University of Athens, Department of Economics,
8 Pesmazoglou Street, 105 59 Athens, Greece, *simosmei@econ.uoa.gr*

Abstract. We develop a procedure for monitoring changes in the error distribution of autoregressive time series. The proposed procedure, unlike standard procedures which are also referred to, utilizes the empirical characteristic function of properly estimated residuals. The limit behavior of the test statistic is investigated under the null hypothesis, while computational and other relevant issues are addressed.

Keywords: empirical characteristic function, change point analysis

Threshold Accepting for Credit Risk Assessment and Validation

Marianna Lyra[1], Akwum Onwunta[2], and Peter Winker[3]

[1] Department of Economics, University of Giessen
Licher Strasse 64, 35392, Giessen, Germany,
Marianna.Lyra@wirtschaft.uni-giessen.de
[2] Deutsche Bank AG
Frankfurt, Germany, *Akwum.Onwunta@db.com*
[3] Department of Economics, University of Giessen & Center for European Economic Research (ZEW), Mannheim
Germany, *Peter.Winker@wirtschaft.uni-giessen.de*

Abstract. According to the latest Basel framework of Banking Supervision, financial institutions should internally assign their borrowers into a number of homogeneous groups. Each group is assigned a probability of default which distinguishes it from other groups. This study aims at determining the optimal number and size of groups that allow for statistical ex post validation of the efficiency of the credit risk assignment system. Our credit risk assignment approach is based on Threshold Accepting, a local search optimization technique, which has recently performed reliably in credit risk clustering especially when considering several realistic constraints. Using a relatively large real-world retail credit portfolio, we propose a new technique to validate ex post the precision of the grading system.

Keywords: credit risk assignment, Threshold Accepting, statistical validation

References

WINKER, P. (2001): *Optimization Heuristics in Econometrics: Applications of Threshold Accepting*. Wiley, New York.

BASEL COMMITTEE ON BANKING SUPERVISION (2006): *International convergence of capital measurement and capital standards a revised framework com- prehensive version*. Technical report. Bank for International Settlements.

LYRA, M., PAHA, J., PATERLINI, S. and WINKER, P. (2010): Optimization heuristics for determining internal rating grading scales. *Computational Statistics & Data Analysis. forthcoming.*

Clustering of 561 French Dwellings into Indoor Air Pollution Profiles

Jean-Baptiste Masson[1,2] and Gérard Govaert[2]
correspondence: *massonje@utc.fr*

[1] INERIS (French National Institute for Industrial Environment and Risks)
Parc Technologique ALATA — BP 2
60550 Verneuil-en-Halatte, France
[2] HEUDIASYC, UMR CNRS 6599
Université de Technologie de Compiègne
BP 20529, 60205 Compiègne Cedex, France

Abstract. INERIS aims to assess the spatial inequalities in the exposure of French people to various chemicals via diverse ways (inhalation, food...). Besides, indoor air is a very specific environment: specific sources (smoking, varnishes, cleaning products, cooking...) can cause the levels of many known chemical pollutants to be much lower outdoors than indoors, where developed countries' people spend most of their time (Spengler and Sexton (1983)). Our specific goal is to define a spatial indicator of the dwellings' indoor air's contribution to the overall exposure. The French Observatory of Indoor Air Quality performed measurements of the indoor air concentrations of 20 volatile organic compounds in 561 dwellings, during one week each, between 2003 and 2005 (Kirchner et al. (2007)). We propose to use clustering methods to define profiles of indoor air pollution, but we have to face two specificities of the data: some concentrations are unknown (due to breakdown of the measurement devices), and some are known only to lie in a certain interval (due to sensitivity thresholds), hence left-censorships. In this communication, we describe adaptations of the classical EM-based clustering methods (Celeux and Govaert (1995)) to this context, in an approach inspired by (Samé et al. (2006)).

Keywords: clustering, mixture model, EM algorithm, censored data

References

CELEUX, G., and GOVAERT, G. (1995): Gaussian parsimonious clustering models. *Pattern Recognition 28 (5), 781–793.*

KIRCHNER, S., ARÈNES, J.-F., COCHET, C., DERBEZ, M., DUBOUDIN, C., ÉLIAS, P., GRÉGOIRE, A., JÉDOR, B., LUCAS, J.-P., PASQUIER, N., PIGNERET, M., and RAMALHO, O. (2007): État de la qualité de l'air dans les logements français, *Environnement, risques et santé 6 (4), 259–269.*

SAMÉ, A., AMBROISE, C., and GOVAERT, G. (2006), A classification EM algorithm for binned data, *Computational Statistics & Data Analysis 51 (2), 466–480.*

SPENGLER, J. D. and SEXTON, K. (1983): Indoor air pollution: a public health perspective, *Science 221 (4605), 9–17.*

Classification Ensemble That Maximizes the Area Under Receiver Operating Characteristic Curve

Eunsik Park[1] and Yuan-Chin I. Chang[2]

[1] Department of Statistics, Chonnam National University, Gwangju 500-757, Korea, *espark02@chonnam.ac.kr*
[2] Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, *ycchang@sinica.edu.tw*

Abstract. It is well known that there is no individual classification method versatile enough to handle all classification problems alone. This can be seen from Lim and Shih (2000), where they compared thirty-three new and old classification methods. Hence, the classification ensemble becomes a possible outlet for improving on the performance of individual classification methods.

We study an ensemble method, targeting at maximizing the area under ROC curve, with non-homogeneous classifiers as its ingredients. Since all classifiers are applied to the same data set, their outputs should be correlated. It is, however, difficult to have information about the correlation among outputs from different classifiers, which makes the ensemble method dependant on such an information less useful here. Hence, the PTIFS method of Wang et al. (2007) is adopted in our paper as the integration method due to its nonparametric character. Each base-classifier will be optimally trained if it has such an option available, and the features selected can be different if the classifier itself has an internal feature selection function. In other words, our method allows each classifier to do its best in all possible senses. Then we take their classification function output values as new features to conduct final ensemble while maximizing AUC as the final objective. That is, our method can integrate nonhomogeneous base-classifiers and each classifier is well-trained before being included into the final ensemble. The proposed method is evaluated with real data examples, and the results show that our method improved on the performances of all base classifiers.

Keywords: classification, ensemble, area under ROC curve

References

LIM, T. and SHIH, Y. (2000): A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning 40, 203 - 229.*

WANG, Z., CHANG, Y., YING, Z., ZHU, L. and YANG, Y. (2007): A parsimonious threshold-independent protein feature selection method through the area under receiver operating characteristic curve. *Bioinformatics 23, 2788 - 2794.*

Constrained latent class models for joint product positioning and market segmentation

Michel Meulders[1,2]

[1] CMS, HUBrussel, Stormstraat 2
1000 Brussels, Belgium, *michel.meulders@hubrussel.be*
[2] Department of Psychology, KUL, Tiensestraat 102
3000 Leuven, Belgium *michel.meulders@psy.kuleuven.be*

Abstract. A key task of strategic marketing is to study the competitive structure of products. This type of analysis is most often based on a spatial configuration of the products or on a categorization of the products. Besides information on the similarity of products, an important goal of competitive structure analysis is to investigate to what extent distinct customer segments prefer a specific group of products, or whether the perception of the products depends on customer segments (DeSarbo et al. (2008)). Candel and Maris (1997) use a probabilistic feature model to categorize products on the basis of binary latent features. However, in this analysis no customer differences are taken into account. To solve this problem we develop a mixture of probabilistic feature models in which the categorization of the products in terms of latent features may differ across customer segments. A related but distinct approach for modelling rater differences was presented by Meulders et al. (2002). A Gibbs sampling algorithm is used to compute a sample of the observed posterior distribution of the model. In addition, posterior predictive simulations are used to evaluate to what extent the model is able to capture observed differences in product perception. As an illustration, the model is used to analyze binary judgments of 191 first-year psychology students who indicated for 25 types of sandwich fillings and 14 fillings characteristics whether or not a certain sandwich filling has a certain characteristic.

Keywords: constrained latent class model, three-way three-mode data, product positioning, market segmentation, Bayesian analysis

References

CANDEL, M. J. J. M. and MARIS, E. (1997): Perceptual analysis of two-way two-mode frequency data: probability matrix decomposition and two alternatives. *I*nternational Journal of Research in Marketing 14, 321-339.

DESARBO, W. S., GREWAL, R. and SCOTT, C. J. (2008): A clusterwise bilinear multidimensional scaling methodology for simultaneous segmentation and positioning analyses. *J*ournal of Marketing Research 280 Vol. XLV, 280-292.

MEULDERS, M., DE BOECK, P., KUPPENS, P. and VAN MECHELEN, I. (2002): Constrained latent class analysis of three-way three-mode data. *J*ournal of Classification 19, 277-302.

A proximity-based discriminant analysis for Random Fuzzy Sets

Gil González-Rodríguez[1], Ana Colubi[2] and M. Ángeles Gil[2]

[1] European Centre for Soft Computing
33600 Mieres, Spain, *gil.gonzalez@softcomputing.es*
[2] Universidad de Oviedo
33006 Oviedo, Spain *colubi@uniovi.es, magil@uniovi.es*

Abstract. The supervised classification problem on the basis of imprecise data (represented by intervals or fuzzy sets) is considered. In Colubi et al. (2010) two discriminat criteria for fuzzy data based on the linkage between fuzzy sets and functional elements in a Hilbert space were proposed. As an alternative, a proximity-based classification criteria for Random Fuzzy Sets (RST, also referred to as fuzzy random variables according to Puri and Ralescu (1986)) inspired by one of the expressions of the optimal criterium for real random variables (in case of symmetric distributions) is introduced. The new approach is based on the estimation of a specific relative proximity of a given point to each class. Let Ω be the population and denote by $d_j(\omega)$ the distance of the RFS $\mathcal{X}$ measured at $\omega \in \Omega$ to the center (represented by the expected value) of the class j. Since the d_j is a real-valued random variable, the relative proximity of ω^* to the class j is proposed to be the conditional probability $P(d_j > d_j(\omega^*))$. The performance of this discriminat rule is tested in an experiment regarding the perception of some people of the relative length of different segments shown in a screen with respect to to a maximum (see Colubi et al. (2010) and also the SMIRE web page http://bellman.ciencias.uniovi.es/SMIRE/Perceptions.html). The relative perception is represented by a trapezoidal fuzzy set over a scale ranging from 0 to 100. The trapezoidal fuzzy set is fixed by the participants according to their perception of each instance, and it will be used to predict the linguistic description of the relative size provided also by the individual (Very Small, Small, Medium, Large or Very Large).

Keywords: supervised classification, random fuzzy sets

References

COLUBI, A., GONZÁLEZ-RODRÍGUEZ, G., GIL, M.A., TRUTSCHNIG, W. (2010): Nonparametric Criteria for Supervised Classification of Fuzzy Data. *Submitted*

PURI, M. L., RALESCU, D. A. (1986): Fuzzy random variables. *Journal of Mathematical Analysis and Applications 114, 409-422*

On Mixtures of Factor Mixture Analyzers

Cinzia Viroli

Department of Statistics, University of Bologna, Italy *cinzia.viroli@unibo.it*

Abstract. Model-based clustering has been successfully applied to many classification problems (Fraley and Raftery, 2000). For quantitative data it is usually based on multivariate Gaussian mixtures, although it is well known that they can lead to over-parameterized solutions in high dimensional data. A possible way to avoid this problem consists of performing model-based clustering in a dimensionally reduced space defined by latent variables. In this perspective, Mixtures of Factor Analyzers (MFA, Gharamani and Hilton, 1997, McLachlan, Peel and Bean, 2003) perform a "local" dimension reduction by assuming that, within each group, the data are generated according to an ordinary factor model, thus reducing the number of parameters on which the covariance matrices depend. In the same perspective, Factor Mixture Analysis (FMA, Montanari and Viroli, 2010) realizes a "global" dimension reduction by assuming that the data are described by a factor model with factors modelled by a multivariate mixture of Gaussians. The main difference between MFA and FMA is that in MFA the data are described by a mixture of several factor models, on the contrary in FMA a unique factor model describes the data, but it assumes factors modelled by a multivariate Gaussian mixture, thus performing clustering in the latent space.

In this work, we propose a two-level dimensionally reduced model-based clustering approach by assuming that the data are generated according to several factor models with a certain prior probability (thus performing a local dimension reduction at the first level), and that in each factor model the factors are described by a multivariate mixture of Gaussians (thus performing a global dimension reduction at the second level). The proposed model generalizes and combines both MFA and FMA and we refer to it as Mixtures of Factor Mixture Analyzers (MFMA).

Keywords: Gaussian mixture models, factor analysis, EM-algorithm

References

FRALEY, C., AND RAFTERY, A.E. (2002): Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association, 97, 611-631.*

GHAHRAMANI, Z., AND HILTON, G.E. (1997): The EM algorithm for mixture of factor analyzers, *Techical Report CRG-TR-96-1*, Department of Computer Science, University of Toronto, Canada.

MCLACHLAN, G.J., PEEL, D., AND BEAN, R.W. (2003): Modelling high-dimensional data by mixtures of factor analyzers, *Computational Statistics and Data Analysis, 41, 379-388.*

MONTANARI, A., AND VIROLI, C.(2010): Heteroscedastic Factor Mixture Analysis, *Statistical Modeling*, forthcoming.

Hybrid Image Classification using Captions and Image Features

Iulian Ilieş[1], Arne Jacobs[2], Otthein Herzog[2] and Adalbert Wilhelm[1]

[1] School of Humanities and Social Sciences, Jacobs University Bremen Campus Ring 1, 28759 Bremen, Germany, *i.ilies|a.wilhelm@jacobs-university.de*
[2] Technologiezentrum Informatik, Universität Bremen, Am Fallturm 1, 28359 Bremen, *jarne|herzog@tzi.de*

Abstract. The continuously increasing quantity of image data available on the Internet necessitates efficient classification and indexing methods for easy access and usage. The application of established information retrieval algorithms such as bag-of-words classifiers (Baeza-Yates & Ribeiro-Neto, 1999) is rather counterintuitive, primarily since images lack words and sentences or similar structures. Consequently, the prevalent approach in current web search engines is to associate images with text, thus allowing access to the image database via text queries. This approach restricts the set of searchable images to those associated with text, and can lead to errors if the associations are incorrect. To enable more successful queries based on visual information, alternative procedures relying on image processing have been proposed: using semantic data generated by image interpretation techniques (e.g. Schober, Hermes, & Herzog, 2005), or via a visual vocabulary constructed from low-level image features (e.g. Sivic & Zisserman, 2003). We developed an integrated procedure, relying on both textual information (keywords extracted from captions) and on descriptors of local image features (SIFT; Lowe, 1999) in the construction of the visual vocabulary. The visual words are prototypes representing groups of similar image features encountered in the training data set, which are further associated to the extracted keywords based on the frequencies of co-occurrences. This approach permits the classification of novel images into text- derived categories using only image-based data, and, correspondingly, the inclusion of images with no caption information in the search space of text queries.

Keywords: image descriptors, image classification, SIFT, text information

References

BAEZA-YATES, R. and RIBEIRO-NETO, B. (1999): *Modern information retrieval.* ACM Press, New York.

LOWE, D.G. (1999): Object recognition from local scale-invariant features. In: *Proceedings, 7th IEEE Conf. Computer Vision.* Kerkyra, 1150-1157.

SCHOBER, J.-P., HERMES, T. and HERZOG, O. (2005): PictureFinder: Description logics for semantic image retrieval. In: *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo.* Amsterdam, 1571- 1574.

SIVIC, J. and ZISSERMAN, A. (2003): Video Google: A text retrieval approach to object matching in videos. In: *Proceedings of the 9th IEEE International Conference on Computer Vision.* Nice, 1470-1477.

An extensive evaluation of the performance of clusterwise regression and its multilevel extension

Eva Vande Gaer[1&2], Eva Ceulemans[1], and Iven Van Mechelen[2]

[1] Centre for Methodology of Educational Research, K.U.Leuven
Andreas Vesaliusstraat 2, Leuven, Belgium,
[2] Research Group of Quant. Psychology & Individual Differences, K.U.Leuven
Tiensestraat 102, Leuven, Belgium

Abstract. In behavioral sciences, many research questions pertain to the prediction of some dependent variable on the basis of a single or several independent variables. Such questions are often answered by performing a regression analysis. However, sometimes the relation between the independent variables and the dependent variable differs across observations. Such differences can be modeled by assigning the observations to different groups with a separate regression model being associated to each group. Such a model already exists, and is called clusterwise linear regression (CR, Späth, 1979). The original CR model is a one-level model, implying that all observations are assumed to be independent from each other. Hence, the method is not applicable to multilevel data where, for instance, observations are nested within persons. Therefore, DeSarbo et al. (1989) extended the clusterwise regression methodology to multilevel data (MLCR), where observations are clustered on the second level. Surprisingly, almost no simulation results are available regarding the performance of clusterwise regression and its multilevel extension. Moreover, on the basis of numerical examples, Brusco et al. (2008) recommend considerable caution when using regular CR. To gain more insight into the performance of both methods, we present the results of an extensive simulation study, which shows that MLCR better recovers the underlying clustering of the persons and the associated regression models than regular CR.

Keywords: clusterwise linear regression, multilevel data, individual differences, clustering

References

BRUSCO, M.J., CRADIT, J.D., STEINLEY, D. AND FOX, G.L. (2006): Cautionary Remarks on the Use of Clusterwise Regression. *Multivariate Behavioral Research 43, 29-49.*

DESARBO, W.S., OLIVER, R.L. AND RANGASWAMY, A. (1989): A simulated annealing methodology for clusterwise linear regression *Psychometrika 54 (4), 707-736.*

SPÄTH, H. (1979): Algorithm 39: Clusterwise linear regression *Computing 22, 367-373.*

Spatial clustering for local analysis

Alessandra Petrucci and Federico Benassi

Department of Statistics "G. Parenti", University of Florence
viale Morgagni 59 - 50134 Firenze, Italy, *alessandra.petrucci@unifi.it, benassi@ds.unifi.it*

Abstract. The need for statistical information at detailed territorial level has greatly increased in recent years. This need is often related to the identification of spatially contiguous and homogeneous areas according to the phenomenon studied.
The aim of the paper lies in a review of methods for the analysis and detection of spatial clusters and in the application of a recently proposed clustering method. In particular, we discuss the nature and the developments of spatial data mining with special emphasis on spatial clustering and regionalization methods and techniques (Guo (2008)).
We present an original application using data from the statistical office of the city of Florence. The first step of the analysis is devoted to describe the structure of the population of Florence's municipality using demographic indicators computed on the 2001 census data at the enumeration district level. Then, we implement a regionalization model in order to get a classification of the Florence's municipality area into a number of homogeneous (with respect to the demographic structure) and spatially contiguous zones.
The empirical application shows that ignoring spatial clustering can lead to misleading inference and that, on the other hand, the use of appropriate methods for the detection of spatial clusters leads to meaningful inference of urban socioeconomic phenomena. The results provide a considerable information to local authorities and policy makers for regional and urban planning: the application of local policies without taking into account spatial dimension can produce a lost in term of efficiency and effectiveness.

Keywords: spatial data mining, spatial statistics, clustering, zoning

References

GUO D. (2008): Regionalization with dynamically constrained agglomerative clustering and partitioning. *International Journal of Geographical Information Science 22 (7), 801-823.*

HASTIE, J., TIBSHIRANI, R. and FRIEDMAN, J. (2001): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer. New York.

LLOYD C.D. (2007): *Local models for spatial analysis.* Boca Renton, Florida.

PETRUCCI A. and BROWNSLEES C. T. (2007): Spatial Clustering Methods for the Detection of Homogenous Areas. In: *Proceedings of CLADAG 2007*, EUM, Macerata.

Semi-supervised Discriminant Analysis for Interval-valued Data

Kenji Toyoda[1], Hiroyuki Minami[2] and Masahiro Mizuta[2]

[1] Graduate School of Information Science and Technology, Hokkaido University, N14W9, Kita-ku, Sapporo, JAPAN, *toyoda@iic.hokudai.ac.jp*
[2] Information Initiative Center, Hokkaido University, N11W5, Kita-ku, Sapporo, JAPAN, *min@iic.hokudai.ac.jp, mizuta@iic.hokudai.ac.jp*

Abstract. We discuss semi-supervised discriminant analysis, focusing on interval-valued data, one of the typical data in Symbolic Data Analysis (SDA). Semi-supervised learning is a method that tries to make a model based on labeled input data and unlabeled ones effectively when both are mixed in the input. Many algorithms have been proposed for classification problems with semi-supervised learning, including transductive support vector machines, graph-based methods.

In the studies of SDA, supervised learning and unsupervised learning have been studied, but semi-supervised learning has not been established. Discriminant analysis for interval-valued data has been proposed by Silva and Brito (2006) which assumes a uniform distribution in each observed data. Cluster analysis based on parametric probabilistic models for interval-valued data were proposed by Bock *et al.* (2009), where the vector of midpoints has normal distribution and midranges has a gamma distribution.

In this paper, we propose a method for discriminant analysis for interval-valued data, which makes use of both labeled and unlabeled data as training data. We give theoretical study of our proposed method. Then, we show effectiveness with simulation data to control parameters of generative model and proportion of labeled and unlabeled data.

Keywords: Semi-supervised learning, Discriminant analysis, Symbolic Data Analysis

References

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining.* Wiley.

Bock, H. H. (2009). Analyzing Symbolic Data: Problems, Methods, and Perspectives. In *Cooperation in Classification and Data Analysis* (editors, Okada, A. *et al.*), Springer, pp. 3-12.

Chapelle, O., Schölkopf,, B. and Zien, A. (2006). *Semi-supervised Learning.* Cambridge, MA: MIT Press.

Duarte Silva, A. P. and Brito, P.(2006). Linear Discriminant Analysis for Interval Data. *Computational Statistics*, 21(2): pp. 289-308, June 2006.

Symbolic Clustering Based on Quantile Representation

Paula Brito[1] and Manabu Ichino[2]

[1] Faculdade de Economia & LIAAD -INESC Porto LA, Universidade do Porto, Porto, Portugal *mpbrito@fep.up.pt*
[2] Department of Information and Arts, Tokyo Denki University, Hatoyama, Saitama 350-0394, Japan, *ichino@ia.dendai.ac.jp*

Abstract. Quantile representation (Ichino, 2008) provides a common framework to represent symbolic data described by variables of different types. The principle is to express the observed variable values by some predefined quantiles of the underlying distribution. In the interval variable case, a distribution is assumed within each observed interval, e.g. uniform (Bertrand and Goupil, 2000) ; for a histogram-valued variable, quantiles of any histogram may be obtained by simply interpolation, assuming a uniform distribution in each class (bid); for categorical multi-valued variables, quantiles are determined from the ranking defined on the categories based on their frequencies. When quartiles are chosen, the representation for each variable is defined by the 5-uple (min, Q_1, Q_2, Q_3, max).

This common representation then allows for a unified analysis of the data set, taking all variables simultaneously into account. In a numerical clustering context, the Ichino-Yaguchi dissimilarity (Ichino and Yaguchi, 1994) is used to compare data units; hierarchical and pyramidal models, with several aggregation indices, may be applied and clusters are formed on the basis of quantile proximity.

In this work, we focus on a conceptual clustering approach. Clusters are represented, for each variable, by a mixture of the quantile-distributions of the merged clusters and then compared on the basis of the current quantile representation. The proposed hierarchical/pyramidal clustering model follows a bottom-up approach; at each step, the algorithm selects the two clusters with closest quantile representation to be merged. The newly formed cluster is then represented according to the same model, i.e., a quantile representation for the new cluster is determined from the uniform mixture cumulative distribution.

Keywords: Symbolic data, Quantile representation, Clustering

References

BERTRAND, P. and GOUPIL, F. (2000): Descriptive Statistics for Symbolic Data. In: H.H.Bock and E. Diday (Eds.): *Analysis of Symbolic Data.* Springer, Heidelberg, 106–124.

ICHINO, M. and YAGUCHI, H. (1994): Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis. *IEEE Tr. Systems, Man and Cybernetics, 24, 4.*

ICHINO, M. (2008): Symbolic PCA for Histogram-Valued Data. In: Proc. IASC 2008, December 5–8, Yokohama, Japan.

High-Dimensional Classification in the Presence of Correlation: A Factor Model Approach

A. Pedro Duarte Silva

Faculdade de Economia e Gestão & CEGE, Univ. Católica Portuguesa at Porto, Rua Diogo Botelho, 1327, 4169-005 Porto, Portugal; *psilva@porto.ucp.pt*

Abstract. A class of linear discrimination rules, designed for problems with many correlated variables, is proposed. These rules try to incorporate the most important patterns revealed by the empirical correlations and accurately approximate the optimal Bayes rule as the number of variables increases without limit. In order to achieve this goal, the rules rely on covariance matrix estimates derived from Gaussian factor models with small intrinsic dimensionality.

Asymptotic results, based on an analysis that allows the number of variables to grow faster than the number of observations, show that the worst possible expected error rate of the proposed rules converges to the error of the optimal Bayes rule when the postulated model is true, and to a slightly larger constant when this model is a close approximation to the data generating process.

Simulation results using real micro-array data and a new variable selection scheme based on Donoho and Jin's Higher Criticism statistic, suggest that under the conditions they were designed for, rules derived from covariance one-factor models perform equally well or better than the most successful extant alternatives.

Keywords: Discriminant Analysis, High Dimensionality, Expected Misclassification Rate, MicroArray Classification.

References

BICKEL, P.J and LEVINA, E. (2004): Some theory for Fisher's Linear Discriminant Function, Naive Bayes and some alternatives when there are many more variables than observations. *Bernoulli 10 (6), 989-1010.*

DONOHO, D. and JIN, J. (2004): Higher Criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics 32 (3), 962-994.*

DONOHO, D. and JIN, J. (2008): Higher criticism thresholding. Optimal feature selection when useful features are rare and weak. In: *Proc. Natl. Acad. Sci.*, 105: 14790-14795.

Symbolic PCA of compositional data

Sun Makosso Kallyth[1] and Edwin Diday[2]

[1] CEREMADE Université Paris Dauphine, Paris, Place du Maréchal de Lattre de Tassigny 75775 PARIS Cedex 16 France, France *makosso@ceremade.dauphine.fr*
[2] CEREMADE Université Paris Dauphine, Paris, Place du Maréchal de Lattre de Tassigny 75775 PARIS Cedex 16 France. *diday@ceremade.dauphine.fr*
CEREMADE Université Paris Dauphine, Paris, Place du Maréchal de Lattre de Tassigny 75775 PARIS Cedex 16 France.

Abstract. This paper deals with Principal Component Analysis (PCA) where the cells of the input data table are not numerical values but histograms. Histograms are compositional data. PCA extended to such data table can be used when histogram variables don't have the same number of bins. In this paper, we propose at first two ways for attributing scores to variables. Afterward, an ordinary PCA of mean of variables is achieved. Representation of dispersion of variable is done in using Tchebychev inequality. This inequality allows transforming histogram to interval. Then we project hypercube associated to each observation on principal axes. We also propose usage of angular transformation for removing drawbacks of histograms which are compositional data.

Keywords: Symbolic histogram variable, hypercube, Tchebychev inequality, angular transformation.

References

AITCHISON J.(1986) The Statistical Analysis of Compositionnal Data. London: *Chapman and Hall.*

CAZES P., CHOUAKRIA A., DIDAY E., SCHEKTMAN Y. (1997): Extension de l'analyse en composantes principales a des donnes de type intervalle, *Rev. Statistique Appliquee, Vol. XLV Num. 3 pag. 5-24, France.*

CAZES, P. (2002). Analyse factorielle d'un tableau de lois de probabilit. *Revue de Statistique Applique, 50 n 3, p. 5-24.*

DIDAY E. (2008) Comment extraire des connaissances partir des concepts de vos bases de donnes? Les deux tapes de l'Analyse des donnes symboliques. *Revue Modulad n38.*

FISHER R. A. (1922), On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser.* A 222 309]368.

ICHINO M.: Symbolic PCA for histogram-valued data. *Proceedings IASC.* December 5-8, Yokohama, Japan, 20085.

NAGABHUSHAN P. , KUMAR P.(2007): Principal Component Analysis of histogram Data. *Springer-Verglag Berlin Heidelber*g. EdsISNN Part II LNCS 4492, 1012-1021

Bivariate Normal Symbolic Regression Model for Interval Data Sets

Eufrásio de A. Lima Neto[1], Gauss M. Cordeiro[2], Francisco de A. T. de Carvalho[3], Ulisses U. dos Anjos[1] and Abner G. da Costa[1]

[1] Departamento de Estatística - UFPB, Cidade Universitária, s/n, 58051-900, João Pessoa (PB), Brazil - *eufrasio@de.ufpb.br*,
[2] Departamento de Estatística e Informática - UFRPE, Rua Dom Manoel de Medeiros, s/n, 52171-900, Recife (PE), Brazil - *gauss@deinfo.ufpb.br*,
[3] Centro de Informática - UFPE, Av. Prof. Luiz Freire, s/n, Cidade Universitária, 50740-540, Recife (PE), Brazil - *fatc@cin.ufpe.br*

Abstract. In Symbolic Data Analysis (Diday and Fraiture-Noirhomme, 2008), the actual regression methods for interval-valued variables do not consider any probabilistic statement for the error of the model. The lack of a probabilistic distribution for the response interval variable has limited the use of inference techniques by these symbolic regression methods. In this paper, we present a symbolic regression method for interval-valued variables based on bivariate normal distribution that is belong to the bivariate exponential family of distributions (Iwasaki and Tsubaki, 2005). Application to a real interval data set, in an cross-validation framework, demonstrated that the new method based on bivariate normal distribution presents a better prediction performance when compared with the non-probabilistic symbolic regression methods proposed by Billard and Diday(2000) and Lima Neto and De Carvalho (2008, 2010). A simulated study is strongly recommended in future works for a more consistent conclusion about this new method.

Keywords: Normal distribution, Symbolic data analysis, Regression Models

References

BILLARD, L. and DIDAY, E. (2000): Regression Analysis for Interval-valued Data. In: *Data Analysis, Classification and Related Methods: Proceedings of the Seventh Conference of the International Federation of Classification Societies*. Springer-Verlag, Belgium, 369-374.

DIDAY, E. and FRAITURE-NOIRHOMME, M. (2008): *Symbolic Data Analysis and the SODAS Software*. Wiley-Interscience.

IWASAKI, M. and TSUBAKI, H. (2005): A New Bivariate Distribution in Natural Exponential Family. *Metrika*, 61, 323–336.

LIMA NETO, E.A. and DE CARVALHO, F.A.T. (2008): Centre and Range Method to Fitting a Linear Regression Model on Symbolic Interval Data. *Computational Statistics and Data Analysis*, 52, 1500–1515.

LIMA NETO, E.A. and DE CARVALHO, F.A.T. (2010): Constrained Linear Regression Models for Symbolic Interval-Valued Variables. *Computational Statistics and Data Analysis*, 54, 333–347.

Statistical Disclosure Control Using the ϵ-uncertainty Intervals and the Grouped Likelihood Method

Jinfang Wang

Department of Mathematics and Informatics, Graduate School of Science, Chiba University, 1-33 Yayoi-cho, Inage-ku, Chiba, 263-8522 Japan,
wang@math.s.chiba-u.ac.jp

Abstract. The global recoding and local recoding are the two most studied measures of statistical disclosure control (see Willenborg and de Waal (1996)). For a continuous variable, *substitution* of a datum by adding a 'noise' to the value of the datum is proposed by Barnes (1995). In this paper we shall propose a systematic method of substitution so that the data $y_1, \ldots, y_n$ are replaced by their respective intervals $I_1, \ldots, I_n$ having some desired properties. The randomly determined intervals I_i have the property that I_i covers y_i with probability ϵ, where ϵ is a predetermined value by the statistical agency. The intervals I_i are called random uncertainty intervals and are proposed by Wang (2002). However, Wang (2002) does not give methods for computing these intervals. In this paper we shall review the random uncertainty intervals and propose a practical method for easily computing the intervals under some mild conditions. Further, we propose to make statistical inference using the interval data $I_1, \ldots, I_n$ based on the idea of grouped likelihood methods of Barnard (1965) and Kempthorne (1966).

Keywords: Box-Cox transformation, disclosure control, ϵ-uncertainty interval, grouped likelihood, linear model

References

BARNARD, G.A. (1965). The use of the likelihood function in statistical practice. *Proceedings of the 5th Berkeley Symposium of Mathematical Statistics and Probability*, eds. by L.M. LeCam and J. Neyman, Vol. I, Berkley: University of California Press, 27–40.

BARNES, G. (1995): Local perturbation. *Report, Department of Statistical Methods, Statistics Netherlands*, Voorburg.

KEMPTHORNE, O. (1966). Some aspects of experimental inference. *Journal of the American Statistical Association* **61**, 11–34.

WANG, J. (2002). Statistical disclosure control based on random uncertainty intervals. *Enabling Society with Information Technology*. Springer-Verlag, Tokyo, 244–255.

WILLENBORG, L. and WAAL, T. D. (1996): *Statistical Disclosure Control in Practice (Lecture Notes in Statistics)*. Springer, New York.

Symbolic Analysis of Hierarchical-Structured Data. Application to Veterinary Epidemiology

Christelle Fablet[1], Edwin Diday[2], Stéphanie Bougeard[1], Carole Toque[3], and Lynne Billard[4]

[1] Afssa-Site de Ploufragan, Unité d'Epidémiologie et de Bien-Etre du Porc, Zoopôle Beaucemaine, 22440 Ploufragan France *c.fablet@AFSSA.FR, s.bougeard@AFSSA.FR*,
[2] CEREMADE, University of Paris 75775 Paris Cedex 16 France *edwin.diday@ceremade.dauphine.fr*,
[3] Syrokko, Aéropôle de Roissy, Bat. Aéronef, 5 rue de Copenhague, 95731 Roissy Charles de Gaulle Cedex France, *toque@syrokko.com*
[4] Department of Statistics, University of Georgia Athens GA 30602 USA *lynne@stat.uga.edu*

Abstract. In veterinary epidemiology, the data usually present a hierarchical structure, e.g., $n \times p$ observations which arise from p animals each within n farms. The dependence between observations which results from this structure, must be taken into account in the statistical process (Dohoo et al., 2003). The main aim is concerned with assessing, among several explanatory variables, the risk factors of a dependent variable, i.e., the disease. Generalized Estimating Equations including a random measurement effect are relevant methods. However, the disease under study is usually described with several variables which must be summed up into an overall dependent variable, manifested as a variable which covers a spectrum of categories ranging from unapparent to fatal. The real epidemiological unit being the farm, the variable information is summed up with animal frequencies (resp., median score) for categorical (resp., continuous) variables. It leads to the building of a large number of variables and to the selection of the most relevant ones, the classification of the farms being processed on the selected variables. Considering the hierarchical-structured data as symbolic data, is a relevant alternative solution. Symbolic data analysis (Billard and Diday, 2006) provides tools to manage complex data while dealing with concepts (i.e., farms), each concept being described with prototypes (i.e., distributional vectors from animals within each farm). The interest of the symbolic analysis is illustrated on the basis of a dataset in the field of veterinary epidemiology. The aim is to categorize 125 farms (30 pig lungs per farm) described with 17 variables related to various respiratory diseases, into classes of disease severity.

Keywords: classification, hierarchical data, symbolic analysis, epidemiology

References

BILLARD, L. and DIDAY, E. (2006): *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley Series in Computational Statistics, London.

Non-linear dimensionality reduction for functional computer code modelling

Benjamin Auder

UPMC Paris 6 - CEA2 Cadarache, DER/SESI/LSMR
Benjamin.Auder@cea.fr

Abstract. Our aim is to develop a model for the computer code CATHARE ("Code Avancé de THermohydraulique pour les Accidents de Réacteurs à Eau"), written $\Phi : \mathbb{R}^p \to \mathbb{R}^{[a,b]}$, $x \mapsto y = \Phi(x)$, which itself model the thermohydraulic behavior of some parts of a nuclear reactor's vessel. We assume n samples $(x_i, y_i = \Phi(x_i))$ are available, where $x_i \in \mathbb{R}^p$ are the initial conditions and $y_i \in \mathbb{R}^{[a,b]}$ the corresponding computed curves. The model built is composed of three main parts:

1. dimensionality reduction: each y_i is represented by $z_i \in \mathbb{R}^d$ with $d \ll D$, satisfying some topological constraints;
2. regression to learn the relation between inputs x_i and representations z_i (Projection Pursuit Regression, by Friedman et al. (1981));
3. build a function $\phi : \mathbb{R}^d \to \mathbb{R}^D$ which maps a vectorial representation $\hat{z}$ to the estimated corresponding curve $\hat{y}$ (manifold learning).

Three dimensionality reduction methods are compared for the first step, Functional Principal Component Analysis (Ramsay and Silverman (2005)) and two nonlinear approaches (Lin et al. (2006), Zhan et al. (2008)). The main assumption is that the output curves lie on a (smooth) manifold.

We use the algorithm of Farahmand et al. (2007) to estimate the manifold dimension d from the samples y_i. The predicted curves $\hat{y}$ through the model are generally more trustful visually in the non-linear case, because the little irregularities are preserved (although not optimally). The curves obtained using the PCA basis look somewhat too much smooth.

Keywords: functional data, dimensionality reduction, surrogate model

References

LIN, T., ZHA, H. and LEE, S. U. (2006): Riemannian Manifold Learning for Nonlinear Dimensionality Reduction. In *9th European Conference on Computer Vision (Graz, Austria), pp. 44–55.*

FARAHMAND, A. M., SZEPESVÁRI, C. and AUDIBERT, J-Y. (2007): Manifold-adaptive dimension estimation. In *24th International Conference on Machine Learning (Corvalis, Oregon), pp. 265–272.*

RAMSAY, J. and SILVERMAN, B. W. (2005): Functional Data Analysis. In Springer Series in Statistics.

ZHAN, Y., YIN, J., ZHANG, G. and ZHU, E. (2008): Incremental Manifold Learning Algorithm Using PCA on Overlapping Local Neighborhoods for Dimensionality Reduction. In *3rd International Symposium on Advances in Computation and Intelligence, pp. 406–415.*

Inference for the difference of two percentile residual life functions

Alba M. Franco-Pereira[1], Rosa E. Lillo[2] and Juan Romo[3]

[1] Universidad Carlos III de Madrid
Calle Madrid 126, 28903 Getafe, Spain *alba.franco@uc3m.es*
[2] Universidad Carlos III de Madrid
Calle Madrid 126, 28903 Getafe, Spain *rosaelvira.lillo@uc3m.es*
[3] Universidad Carlos III de Madrid
Calle Madrid 126, 28903 Getafe, Spain *juan.romo@uc3m.es*

Abstract. In Joe and Proschan (1984) the percentile residual life orders were introduced, but they were extensively studied in Franco-Pereira et al. (2009). In this paper, some interpretations and properties of these stochastic orders were given and some applications in reliability theory and finance were described.

Given the advantages of the percentile residual life orders, specially in practical situations, it is convenient to develop an statistical tool to test whether two independent random samples have underlying random variables which are close with respect to a γ-percentile residual life order.

In this work, we present a nonparametric method for constructing confidence bands for the difference of two percentile residual life functions. This functional data analysis technique incorporates bootstrap resampling and the concept of statistical depth. The confidence bands provide us with evidence of whether two random variables are close with respect to some percentile residual life order. The practical performances of the bands are evaluated through simulation. Some applications with real data are also given.

Keywords: Confidence bands, functional data analysis, bootstrap, percentile residual life, statistical depth

References

FRANCO-PEREIRA, A. M., LILLO, R. E., ROMO, J. and SHAKED, M. (2009): Percentile residual life orders. To appear in *Applied Stochastic Models in Business and Industry*

JOE, H. and PROSCHAN, F. (1984): Percentile residual life functions. *Operations Research 32, 668–678.*

A New Statistical Test for Analyzing Skew Normal Data

Hassan Elsalloukh[1] and Jose Guardiola[2]

[1] Assistant Professor of Statistics, Department of Mathematics and Statistics,University of Arkansas at little Rock
2801 S. University Avenue, Little Rock AR 72211, USA *hxelsalloukh@ualr.edu*
[2] Assistant Professor of Statistics, Department of Mathematics and Statistics, Texas A & M University - Corpus Christi
6300 Ocean Dr., CI 309 Corpus Christi TX, 78412, USA
jose.guardiola@tamucc.edu

Abstract. Symmetric distributions are not practical for modeling skewed data. Several researchers have derived new parametric skew distributions; in particular, Azzalini (1985) derived the Skew Normal Distribution (SN). In many practical situations, the Lagrange Multipliers test or score test is an attractive competitor to the likelihood ratio (LR) and the Wald tests because it requires the model parameters to be estimated under only the null hypothesis. In this talk, a new score statistic is derived for testing the presence of non-normality within a modified SN distribution and applied to real examples. Maximum likelihood estimators are derived and used to fit the data with the modified SN distribution and compared to the normal distribution fit.

Keywords: Score tests, Skew normal distributions, Skewness

References

AZZALINI, A. (1985):A Class of Distributions Which Includes the Normal Ones. *Scand. J. Statist* 12, 171-178.

AZZALINI, A. (1986): Further Results on a Class of Distributions Which Includes the Normal Ones. *Statistica* 46, 199-208.

ELSALLOUKH, H., GUARDIOLA, J. H., and D. M. YOUNG (2005): The Epsilon-Skew Exponential Power Distribution Family. *Far East Journal of Theoretical Statistics* 17(1), 97-112.

ÓHAGAN, A. and LEONARD, T. (1976): Bayes Estimation Subject to Uncertainty about Parameter Constraints. *Biometrika* 63, 201-202.

POIRIER, D. J., TELLO, D., and ZIN, E. (1986): A diagnostic Test for Normality within the Power Exponential Family, *Journal of Business and Economic Statistics* 4, 359-373.

Generalized Linear Factor Models: a local EM estimation algorithm

Xavier Bry[1], Christian Lavergne[1,2], and Mohamed Saidane[3]

[1] Université Montpellier II, I3M UMR-CNRS 5149
[2] Université Paul-Valéry, Montpellier III
[3] Université du 7 Novembre à Carthage, ISCC de Bizerte

The framework is that of factor models (FM): a set of q observed random variables (RV) $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_q\}$ is assumed to be produced by fewer ($k < q$) unobserved (latent) ones, $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_k\}$, called factors. So far, most developments on FM's were limited by the assumption that $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_q\}$ are normally distributed, and used this specific distribution to carry out their estimation, through the EM algorithm. The Generalized Linear Factor Models extend the FM class to any type of distribution belonging to the exponential family: binomial, gamma, Poisson, etc. Therefore, we must also deal with the framework of generalized linear models (GLM), which take observed variables only as predictors, and are estimated using these observed values. In this work, we consider a GLFM, which, conditional to the factors, is modeled as a GLM taking these factors as predictors. For identification purposes, the factors are taken uncorrelated and normally distributed with 0 mean and unit variance. Moreover, $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_q\}$ are assumed to be independent conditional to the factors. Initially, FM's and GLM's have been developed independently. Recently [Moustaki, I., and Knott, M. (2000)] have proposed a maximum likelihood method, in which the Gauss-Hermite quadrature is used to approximate an integral within the likelihood maximization process. This method was developed for the case of a single factor. A Monte Carlo approach was also proposed by [Wedel, M. and Kamakura, W.A. (2001)]. Finally, an iterative estimation method inspired from the indirect inference technique [Gourieroux (1993)] was proposed by [Moustaki, I and Victoria-Feser, M.P.(2006)]. In this work, we suggest an alternative approach to GLFM. The problem with GLFM's estimation is that the EM algorithm - using an explicit expression of the expected completed log likelihood of parameters conditional to observations - does not directly extend to non-normal distributions. To circle this difficulty, we consider the GLM's estimation algorithm that iteratively linearizes the model and performs Generalized Least Squares on it, and we propose to apply the EM procedure "locally" to this linearized GLM. The algorithm is exposed, and its performance is examined on various simulated datasets.

Keywords: Factor Models, Generalized Linear Model, EM Algorithm, Scores Algorithm

The Aggregate Association Index

Eric J. Beh[1]

School of Mathematical & Physical Sciences, University of Newcastle, Callaghan, NSW 2308, Australia *eric.beh@newcastle.edu.au*

Abstract. Consider a single two-way contingency table where both variables are dichotomous in nature. Suppose that n individuals/units are classified into this table such that the number classified into the (1, 1)th cell is denoted by n_{11}. Let the i'th row marginal frequency be denoted by $n_{i\bullet}$, for $i = 1, 2$, and the j'th column marginal frequency by $n_{\bullet j}$, for $j = 1, 2$. Also, denote the i'th row and j'th column marginal proportion by $p_{i\bullet} = n_{i\bullet}/n$ and $p_{\bullet j} = n_{\bullet j}/n$ respectively. Table 1 provides a description of this notation.

Table 1. Notation for a single 2×2 contingency table.

	Column 1	Column 2	Total
Row 1	n_{11}	n_{12}	$n_{1\bullet}$
Row 2	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 1}$	n

Suppose that the cell values in Table 1 are unknown so that only the information in the marginal frequencies is known, and fixed. The problem at hand is to obtain some information concerning the nature of the association between the two dichotomous variables when only this marginal information is provided. When a single 2×2 table is of interest (as is the case here), Fisher (1935) considered this issue and judged there to be very little or no information in the margins for inferring individual (or cellular) level data - such an issue has implications in ecological inference where $G(> 1)$ 2×2 tables of this type are commonly encountered. However, when only the marginal information is known, such information can still provide an indication of the possibility that there exists a statistically significant association between the two categorical variables using the aggregate association index, the AAI. Such an index was originally proposed by Beh (2008) and subsequently elaborated upon by Beh (2010). We shall consider the development and application of this index for exploring the possibility of an association existing for such aggregated information.

Keywords: 2×2 contingency table, correlation, ecological inference

References

BEH, E. J. (2008): Correspondence analysis of aggregate data: The 2×2 table. *Journal of Statistical Planning and Inference 138, 2941 - 2952*

BEH, E. J. (2010): The aggregate association index. *Computational Statistics & Data Analysis 54, 1570 - 1580*

FISHER, R. A. (1935): The logic of inductive inference (with discusion). *Journal of the Royal Statistical Society, Series A 98, 39 - 82*

Functional Estimation in Systems Defined by Differential Equation using Bayesian Smoothing Methods

Jonathan Jaeger[1] and Philippe Lambert[2]

[1] Institut de Statistique, Biostatistique et Sciences Actuarielles, UCL
20 voie du Roman Pays, Louvain-la-Neuve, Belgique
jonathan.jaeger@uclouvain.be
[2] Institut des sciences humaines et sociales, ULg
7 boulevard du Rectorat, Liège, Belgique *p.lambert@ulg.ac.be*

Abstract. Differential equations are frequently used to specify models in chemical engineering, pharmacokinetics and other sciences. Current methods for parameter and functional estimations in such models use minimization techniques and numerical solver. These approaches are computationally intensive and often poorly suited for statistical inference.

Alternative estimation methods of the state function and of the dynamic model parameters were proposed by Ramsay et al. (2007). It accounts for measurements errors and elude numerical integration. That approach involves some basis function expansion of the state function and a penalty term expressed using the set of differential equations.

The methodology above may be viewed as a generalization of the penalized spline theory for which a Bayesian framework was proposed by Berry et al. (2002). We aim at providing a Bayesian framework for the more general approach described by Ramsay et al (2007). First, we present a brief introduction to dynamic models defined by systems of differential equations. We then propose a full Bayesian smoothing approach for the joint estimation of the differential equation parameters and of the state functions and extend it to a hierarchical context. We conclude the presentation by some practical examples.

Keywords: Bayesian smoothing methods, Differential equations, Functional estimation

References

BERRY S.M., CARROLL R.J. and RUPPERT D. (2002): Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association 97, 160-169.*

RAMSAY J.O., HOOKER G., CAMPBELL D. and CAO J. (2007): Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society, Series B 69, 741-796.*

Efficient Analysis of Three-Level Cross-Classified Linear Models with Ignorable Missing Data

Yongyun Shin

Abstract

This article presents an efficient estimation method for a three-level hierarchical cross-classified linear model (HCLM) where explanatory as well as outcome variables may be subject to missingness with a general missing pattern at any of the levels. The key idea is to reexpress the desired HCLM as the joint distribution of the variables, including the outcome variable, that are subject to missingness conditional on all of the variables that are completely observed. Unlike a hierarchical linear model where a lower-level unit (e.g. a child) is nested within a single unit (e.g. a school) at a higher level, however, an HCLM may represent a mobile child that may attend multiple schools. The mobile children produce a complicated network of dependence among schools that may otherwise be assumed independent under a hierarchical linear model. Conventional estimation of the joint model takes the entire sample at once to fully account for the complicated dependence structure among the schools. With large-scale data and multiple variables subject to missingness at multiple levels, the joint model becomes extremely high dimensional and difficult to estimate well. The challenge is to estimate the joint model in a way that does not burden computation with respect to the unit mobility; and in a way that produces efficient analysis of the desired HCLM. This paper presents a maximum likelihood method that overcomes the challenge in the presence of ignorable missing data. Confined to normal linear models, the application is illustrated with real data.

KEY WORDS: Hierarchical; Cross-Classified; Ignorable Missing Data; Maximum Likelihood; Efficient Estimation.

Analysis of Competing Risks in the Pareto Model for Progressive Censoring with binomial removals

R. HASHEMI * and J. AZAR †

Department of Statistics, Faculty of Science, Razi University, Kermanshah , Iran.* E-mail: rhashemi@razi.ac.ir

Department of Statistics, Faculty of Basic Science, University of Mazandaran, Babolsar, Iran.†

Abstract

In medical studies or in reliability and survival analysis, it is quite common that the failure of any individual or any item may be attributable to more than one cause (competing risks). In this paper, we will study the competing risks model when the data is progressively type II censored with random removals. We study the model under the assumption of independent causes of failure and assume that the lifetime of each unit which fails due to a different cause of failure, follows a Pareto distribution with different parameters , where the number of items or individuals removed at each failure time follows binomial distribution. The maximum likelihood estimators of the different parameters and the UMVUE's are obtained. We consider the Bayesian estimation using the Gamma distribution as a prior. In the Bayesian context, we develop credible intervals for the parameters. Asymptotic confidence intervals and two bootstrap confidence intervals are also proposed. We also present a numerical example and a simulation study to illustrate the results.

Keywords: Binomial removals; Competing risks; Pareto model; Progressive censoring.

References

[1] Balakrishnan N., Aggarwala R., *Progressive censoring: Theory, Methods and Applications*, Birkhäuser, Boston, 2000.

[2] Crowder M.J., *Classical Competing Risks*, Chapman and Hall/CRC, 2001.

[3] Efron B., *The Jackknife, the bootstrap and other re-sampling plans*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 38. SIAM, Philadelphia, PA, 1982.

[4] Kundu D., Kannan N., Balakrishnan N., *Analysis of progressively censored competing risk data*, Handbook of statistics, 23, 331-348, 2004.

[5] Soliman A. A., *Estimations for Pareto Model Using General Progressive Censored Data and Asymetric Loss*, Communications in Statistics- Theory and Methods, vol. 37, iss. 9, 1353-1370, Jan. 2008.

[6] Tse S.K., Yang C. and Yuen H.K., *Statistical analysis of Weibull distributed lifetime data under Type II progressive censoring with binomial removals*, Journal of Applied Statistics, Vol. 27, 1033- 1043, 2000.

A Method for Time Series Analysis Using Probability Distribution of Local Standard Fractal Dimension

Kenichi Kamijo[1] and Akiko Yamanouchi[2]

[1] Graduate School of Life Sciences, Toyo University,
1-1-1 Izumino, Itakura, Gunma, 374-0193, Japan, *kamijo@toyonet.toyo.ac.jp*
[2] Izu Oceanics Research Institute,
3-12-23 Nishiochiai, Shinjuku, Tokyo, 161-0031, Japan,
gx0400018@toyonet.toyo.ac.jp

Abstract. The moving local standard fractal dimension (LSFD) on the uniform or standard normal random process belongs to a non-symmetric normal distribution with a long tail towards the right hand side. These results can be applied to the statistical quality control, especially to the so-called control charts. Also the proposed method can be applied to the difference time series of seawater temperatures as a function of depth and the probability distribution of the moving LSFD was shown to generally conform to a power-law distribution. That is, prediction of abnormal phenomena in the global monitoring system may be possible using the moving LSFD and observing when it increases past the upper 5% significance level.

Keywords: local standard fractal dimension, statistical quality control, random process, difference time series of seawater temperatures

References

Ayache, A. et al. (2007): A central limit theorem for the quadratic variations of the step fractional Brownian motion, *Stat. Inference for Stoch. Processes*, 10, 1-27.

Benassi, A. et al. (2000): Identification of the Hurst index of a Step Fractional Brownian Motion, *Stat. Inference for Stoch. Processes*, 3, 101-111.

Kamijo, K. and Yamanouchi, A. (2008): Time Series Analysis Using Local Standard Fractal Dimension -Application to Fluctuations in Seawater Temperature-, *International Conference on Computational Statistics(COMPSTAT2008)*, Porto, Portugal.

Kamijo, K. and Yamanouchi, A. (2009): Numerical and Practical Method for Statistical Quality Control Using Local Fractal Dimension in Discrete Time Series, *European Conference on Numerical Mathematics and Advanced Applications(ENUMATH 2009)*, Uppsala, Sweden.

Bootstrapping Additive Models in Presence of Missing Data

Rocío Raya-Miranda, María Dolores Martínez-Miranda and Andrés González-Carmona

Department of Statistic and O.R.
c/ Severo Ochoa, s/n 18071 Granada, Spain, *rraya@ugr.es*

Abstract. The problem of estimating nonparametric additive models when missing data appear in the response variable is considered. Two estimators are constructed from the Backfitting Local Polynomial estimators by Opsomer (2000). The simplest (SB) is defined by using the available data and another imputed version (IB) is based on a complete sample from imputation techniques. The problem of selecting the smoothing parameter is solved by using a local selector, which is based on a Wild Bootstrap approximation of the Mean Squared Error. Several simulation experiments illustrate the performance of the proposed methods.

Keywords: Missing data; Imputation; Wild Bootstrap; Smoothing Parameter; Backfitting

References

AERTS, M., CLAESKENS, G., HENS, N. and MOLENBERGS, G. (2002): Local multiple imputation. *Biometrika 89 (2), 375-388.*

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989): Linear smoothers and additive models (with discussion). *The Annals of Statistics 17, 453-555.*

GONZÁLEZ-MANTEIGA, W. and PÉREZ-GONZÁLEZ, A. (2004): Nonparametric mean estimation with missing data. *Communications in Statistics, Theory and Methods 33 (2), 277-303.*

LITTLE, R.J.A. and RUBIN, D.B. (2002): *Statistical Analysis with Missing Data.* Willey-Interscience.

MARTÍNEZ-MIRANDA, M.D., RAYA-MIRANDA, R., GONZÁLEZ-MANTEIGA, W. and GONZÁLEZ-CARMONA, A. (2008): A bootstrap local bandwidth selector for additive models. *Journal of Computational and Graphical Statistics 17, 38-55.*

NIELSEN, J.P. and SPERLICH, S. (2005): Smooth backfitting in practice. *Journal of the Royal Statistical Society, Series B 67 (1), 43-61.*

OPSOMER, J.D. (2000): Asymptotic properties of backfitting estimators. *Journal of the Multivariate Analysis 73, 166-179.*

OPSOMER, J.D., and RUPPERT, D. (1997): Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics, 25, 186-293.*

RUBIN, D.B. (1987): *Multiple Imputation for Nonresponse in Surveys.* Willey & Sons.

Global hypothesis test to simultaneously compare the predictive values of two binary diagnostic tests in paired designs: a simulation study

J. A. Roldán Nofuentes,[1] J. D. Luna del Castillo[2] and M. A. Montero Alonso[3]

[1] Biostatistics, School of Medicine, University of Granada, 18071, Spain, *jaroldan@ugr.es*
[2] Biostatistics, School of Medicine, University of Granada, 18071, Spain, *jdluna@ugr.es*
[3] School of Social Sciences, Campus of Melilla, University of Granada, 52071, Spain, *mmontero@ugr.es*

Abstract. The positive and negative predictive values of a binary diagnostic test are measures of the clinical accuracy of the diagnostic test that depend on the sensitivity and the specificity of the binary test and on the disease prevalence. Moreover, the positive predictive value and the negative predictive value are not parameters which are independent of each other. In this article, a global hypothesis test is studied to simultaneously compare the positive and negative predictive values of two binary diagnostic tests in paired designs.

Keywords: Binary diagnostic test, Predictive values, Simultaneously comparison

References

AGRESTI, A. (2002): *Categorical data analysis*. John Wiley & Sons, New York.

BENNETT, B.M. (1972): On comparison of sensitivity, specificity and predictive value of a number of diagnostic procedures. *Biometrics 28, 793-800*.

HOLM, S. (1979): A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistic 6, 65-70*.

HOCHBERG, Y. (1988): A sharper Bonferroni procedure for multiple tests of significance. *Biometrika 75, 800-802*.

LEISENRING, W., Alonzo, T., PEPE, M.S. (2000): Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics 56, 345-351*.

VACEK, P.M. (1985): The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics 41, 959-968*.

WANG, W., DAVIS, C.S., SOONG, S.J. (2006): Comparison of predictive valus of two diagnostic tests from the same sample of subjects using weighted least squares. *Statistics in Medicine 25, 2215-2229*.

Model-Based Nonparametric Variance Estimation for Systematic Sampling. An Application in a Forest Survey

Mario Francisco-Fernández[1], Jean Opsomer[2], and Xiaoxi Li[3]

[1] Universidad de A Coruña. Departamento de Matemáticas, Facultad de Informática, A Coruña, 15071, Spain, *mariofr@udc.es*
[2] Colorado State University. Department of Statistics, Fort Collins, CO 80523, USA, *jopsomer@stat.colostate.edu*
[3] Pfizer, Inc. Groton, CT 06340, USA, *xiaoxi.li@pfizer.com*

Abstract. Systematic sampling is frequently used in natural resource and other surveys, because of its ease of implementation and its design efficiency. An important drawback of systematic sampling, however, is that no direct estimator of the design variance is available. A whole chapter of the recently reissued classic monograph by Wolter (Wolter (2007)) is devoted to this issue, and a number of possible estimation approaches are evaluated there. In particular, it considers a set of eight "model-free" estimators and outlines a model-based estimation approach. On the other hand, in Bartolucci and Montanari (2006), an unbiased model-based variance estimator when the population follows a linear regression model is proposed. In this work, we describe a new estimator of the model-based expectation of the design variance, under a nonparametric model for the population. We prove the model consistency of the estimator for both the anticipated variance and the design variance. We compare the nonparametric variance estimators with several design-based estimators on data from a forestry survey. Full results of a comprehensive simulation study comparing several estimators and of the real data forest application are shown in Opsomer et al. (2009).

Keywords: local polynomial regression, two-per-stratum variance approximation, smoothing

References

BARTOLUCCI, F. and MONTANARI, G. E. (2006): A new class of unbiased estimators for the variance of the systematic sample mean. *Journal of Statistical Planning and Inference 136, 1512-1525.*

OPSOMER, J., FRANCISCO-FERNANDEZ, M. AND LI, X. (2009): Additional results for model-based nonparametric variance estimation for systematic sampling in a forestry survey. Technical report, Department of Statistics, Colorado State University.

WOLTER, K. M. (2007): *Introduction to Variance Estimation (2 ed.).* Springer-Verlag Inc., New York.

Panel Data Models for Productivity Analysis

Luigi Grossi[1] and Giorgio Gozzi[2]

[1] Departimento di Economia, Universitá di Verona,
Via dell'Artigliere 19,
37129, Verona, Italy
(e-mail: luigi.grossi@univr.it)
[2] Departimento di Economia, Universitá di Parma,
Via Kennedy 6,
43100, Parma, Italy
(e-mail: giorgio.gozzi@unipr.it)

Abstract. In the present paper dynamic panel models for productivity analysis will be analyzed. Recent years have seen a relevant increase in studies on productivity. This is partly due to rising availability of longitudinal micro-level data. This paper is an attempt to give an answer to some questions about productivity dynamics and determinants, using the large data base of company accounts constructed by Research Center of Unioncamere. In our study we investigate the distribution of labor productivity in two important Italian manufacturing sectors. A new derivation of dynamic panel model starting from a Cobb-Douglas production function has been applied in this paper to discover the underlying generating process of productivity growth and to estimate the elasticities of productivity to personnel expenditure.

Keywords. Italian manufacturing sector, panel data, productivity growth.

Consensus Analysis Through Modal Symbolic Objects

Jose M Garcia-Santesmases[1] and M. Carmen Bravo[2]

[1] Universidad Complutense de Madrid, Departamento de Estadistica e Investigacion Operativa, Facultad de Ciencias Matematicas, 28040 Madrid, Spain, *josemgar@mat.ucm.es*

[2] Universidad Complutense de Madrid, Servicio Informatico de Apoyo a Docencia e Investigacion, Edificio Real Jardin Botanico Alfonso XIII, 28040 Madrid, Spain, *mcbravo@pas.ucm.es*

Abstract. This paper addresses the problem of analyzing the existence of different patterns of consensus when data come from several observers who separately evaluated several issues on a rating scale of ordered categories. To analyze it, two main steps can be distinguished: a) The use of consensus measures to evaluate the strength of consensus in a class of individuals; b) The evaluation of each individual to propose changes in his opinion in order to increase the strength of the consensus. For step a) we give a consensus measure for a group of individuals and extend this measure to several issues. Also, we define a consensus measure for symbolic objects. For step b) we extend the procedure described in Garcia-Santesmases and Bravo (2008) and we use modal symbolic objects to analyze the consensus on a class of individuals. A clustering based solution is proposed. Symbolic objects built from the obtained clusters are the consensus groups, which extensions of a fixed level satisfying at least a fixed number of issues cover the set of individuals. A graphical representation of the consensus groups is used, based on the zoom star representation. An example is given to illustrate a complete analysis.

Keywords: symbolic objects, consensus analysis, cluster analysis, consensus measure

References

BOCK, H. H. and DIDAY, E. (Eds.) (2000): *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer Verlag, Heidelberg.

GARCIA-SANTESMASES, J.M. and BRAVO,M.C.(2008): Analysis of Consensus Through Symbolic Objects. In: P. Brito (Ed.): *Proceedings in Computational Statistics 2008, vol II.* Physica Verlag, 481-489.

NOIRHOMME-FRAITURE, M. and ROUARD, M. (2000): Visualizing and Editing Symbolic Objects. In: H.H. Bock and E. Diday (Eds.): *Analysis of Symbolic Data.* Springer Verlag, Heidelberg, 125-138.

TASTLE, W. J., WIERMAN, M. J. and DUMDUM, U. R. (2005): Ranking Ordinal Scales Using the Consensus Measure. *Issues in Information Systems VI (2), 96-1.*

Clusters of Gastrointestinal Tract Cancer in the Caspian Region of Iran: A Spatial Scan Analysis

Mohammadreza Mohebbi[1] and Rory Wolfe[1]

1. Department of Epidemiology and Preventive Medicine, Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, Australia, *Mohammadreza.Mohebbi@med.monash.edu.au*

Abstract. Spatial cluster detection is an important tool in cancer surveillance to identify areas of high risk and to generate hypotheses about cancer etiology. Numerous studies in the literature have focused on gastrointestinal (GI) cancer because it is the most common organ system involved. In this paper we explore whether there is statistically significant global clustering of GI tract cancer in the Caspian region of Iran.
Method: A spatial scan statistic was used, which searched for clusters of disease without specifying their size or location ahead of time, and which tested for their statistical significance while adjusting for the multiple testing inherent in such a procedure. The approach combined time series scan statistic and spatial search machine methods. The spatial scan test used a moving circle of varying size to find a set of regions or points that maximised the likelihood ratio test for the null hypothesis of a purely random Poisson.
Results: Demographic data and age-specific GI cancer incidence rates were obtained for all 160 agglomerations in 26 wards of Caspian region for 2001-2006 (Mohebbi et al, 2008). The analysis was performed in male female and both sexes combined. A primary significant cluster of high incidence was identified in the eastern agglomerations for esophageal and stomach cancer in male, female and both sexes combined. There was also evidence of significant non-overlapping secondary clusters in the region.
Conclusions: The analysis identified geographic areas with elevated GI cancer incidence in the Caspian region of Iran. Surveillance findings such as these have the benefit of providing insight to the epidemiologist and might lead to monitoring geographical trends for cancer control activities.

Keywords: Disease clustering, Gastrointestinal cancer, Monte Carlo method, Poisson distribution, Spatial autocorrelation

References

MOHEBBI, M., MAHMOODI, M., WOLFE, R., NOURIJELYANI, K., MOHAMMAD, K., ZERAATI, H. and FOTOUHI, A. (2008): Geographical spread of gastrointestinal tract cancer incidence in the Caspian Sea region of Iran: Spatial analysis of cancer registry data *BMC Cancer 8, 137.*

Design of Least-Squares Quadratic Estimators Based on Covariances from Interrupted Observations Transmitted by Different Sensors

R. Caballero-Águila[1], A. Hermoso-Carazo[2] and J. Linares-Pérez[2]

[1] Dpto. de Estadística e I.O., Universidad de Jaén, 23071 Jaén, Spain, *raguila@ujaen.es*
[2] Dpto. de Estadística e I.O., Universidad de Granada, 18071 Granada, Spain, *ahermoso@ugr.es, jlinares@ugr.es*

Abstract. This paper discusses the least-squares quadratic estimation problem of discrete-time signals from uncertain noisy observations coming from multiple sensors. For each sensor, the uncertainty about the signal being present or missing in the observation is modelled by a set of Bernoulli random variables whose probabilities are not necessarily the same for all the sensors.

It is assumed that only information on the moments (up to the fourth-order ones) of the signal and observation noise is available. The estimators do not require full knowledge of the state-space model generating the signal process, but only the autocovariance and crosscovariance functions of the signal and their second-order powers in a semidegenerate kernel form, and the probability that the signal exists in the observed values.

To address the quadratic estimation problem, augmented signal and observation vectors are introduced by assembling the original vectors with their second-order powers, defined by the Kronecker product. Using an innovation approach, a recursive linear estimation algorithm of the augmented signal based on the augmented observations is obtained, from which the required quadratic estimators are derived.

To illustrate the theoretical results established in this paper, a simulation example is presented, showing the feasibility of the proposed algorithm and the superiority of the quadratic estimators over the linear ones.

Keywords: least-squares quadratic estimation, uncertain observations, multiple sensors

Using Logitboost for Stationary Signals Classification

Pedro Saavedra, Angelo Santana, Carmen Nieves Hernández, Juan Artiles, and Juan-José González

Departamento de Matemáticas. Universidad de Las Palmas de Gran Canaria 35017 Las Palmas de Gran Canaria. Spain *saavedra@dma.ulpgc.es*

Abstract. The use of Boosting in conjunction with decision trees has been shown to be an effective method for classification problems characterized by high dimensionality. We propose using Boosting for classification of signals generated by stationary processes with mixed spectra, taking as features vector the ordinates of the components of the spectral distribution obtained at Fourier frequencies. The proposed method is evaluated by means of two studies with simulated data and a third study with real data from electro-encephalogram (EEG) signals measured on healthy and epileptic subjects in seizure-free intervals. The error rates are compared with those obtained from another proposed classification method in the literature.

Keywords: logitboost, stationary signals classifier provides predictions that are clearly better than the V&P method.

LTPD Plans by Variables when the Remainder of Rejected Lots is Inspected

J. Klufa[1] and L. Marek[2]

[1] University of Economics, Department of Mathematics
W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic, *klufa@vse.cz*
[2] University of Economics, Department of Statistics and Probability
W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic, *marek@vse.cz*

Abstract. In this paper we shall consider two types of LTPD plans - for inspection by variables and for inspection by variables and attributes (all items from the sample are inspected by variables, remainder of rejected lots is inspected by attributes).

For the given parameters N, $\bar{p}$, p_t and c_m we must determine the acceptance plan (n, k), minimizing

$$I_{ms} = n \cdot c_m + (N - n) \cdot [1 - L(\bar{p}; n, k)] \tag{1}$$

under the condition

$$L(p_t; n, c) = 0.10 \tag{2}$$

(LTPD single sampling plans), where N is the number of items in the lot (the given parameter), $\bar{p}$ is the process average fraction defective (the given parameter), p_t is the lot tolerance fraction defective (the given parameter, $P_t = 100p_t$ is the lot tolerance per cent defective – denoted LTPD), $c_m = \frac{c_m^*}{c_s^*}$ (c_s^* is the cost of inspection of one item by attributes, c_m^* is the cost of inspection of one item by variables), n is the number of items in the sample ($n < N$), k is the critical value, $L(p)$ is the operating characteristic (the probability of accepting a submitted lot with fraction defective p), i.e. minimizing the mean inspection cost per lot of process average quality $C_{ms} = I_{ms} \cdot c_s^*$ under the condition (2) - see Klufa (1994).

We shall report on an algorithm allowing the calculation of these plans when the non-central t distribution is used for the operating characteristic. The calculation is considerably difficult, we shall use an original method and software Mathematica - see Klufa (in print). From the results of numerical investigations it follows that under the same protection of consumer the LTPD plans for inspection by variables are in many situations more economical than the corresponding Dodge-Romig attribute sampling plans.

Keywords: Acceptance sampling, LTPD plans, software Mathematica

References

DODGE, H. F. and ROMIG, H. G. (1998): *Sampling Inspection Tables: Single and Double Sampling.* John Wiley.

KLUFA, J. (1994): Acceptance Sampling by Variables when the Remainder of Rejected Lots is Inspected. *Statistical Papers 35, 337 - 349.*

KLUFA, J. (in print): Exact calculation of the Dodge-Romig LTPD single sampling plans for inspection by variables. *Statistical Papers*

Modelling the Andalusian Population by Means of a non-Homogeneous Stochastic Gompertz Process

Huete Morales, M.D.[1] and Abad Montes, F.[2]

[1] Department of Statistics & O.R. University of Granada
C/Fuente Nueva, s/n, Faculty of Sciences, 18071-Granada, Spain,
mdhuete@ugr.es
[2] Department of Statistics & O.R. University of Granada
C/Fuente Nueva, s/n, Faculty of Sciences, 18071-Granada, Spain, *fabad@ugr.es*

Abstract. In this study, we examine the stochastic Gompertz non-homogeneous diffusion process, analysing its transition probability density function and conducting inferences on the process parameters using discrete sampling. All of the results are applied to the population of Andalusia (Spain), disaggregating the data by sex for the period 1981 to 2002, taking as exogenous factors only variables that are purely demographic, i.e. life expectancy at birth, foreign immigration to Andalusia and total fertility rate.

Keywords: Gompertz diffusion process, exogenous factors, demography, population

References

Crow, E.L. and Shimizu, K. (1988): *Lognormal distribution theory and application*. Ed. Marcel Dekker.

Ferrante, L. and Bompadre, S. and Possati, L. and Leone, L. (2000): Parameter estimation in a gompertzian stochastics model for tumor growth. *Biometrics 56, 10761081*.

Gutiérrez, R. and Gutiérrez-Sánchez, R. and Nafidi, A. and Román, P. and Torres, F. (2005): Inference in gompertz-type nonhomogeneous stochastic systems by means of discrete sampling. *Cybernetics and Systems 36, 203-216*.

Huete, M.D.(2006): *El modelo estocástico de Gompertz. Modelización de datos sociodemográficos*. PhD. thesis, University of Granada.

Nafidi A. (1997): *Difusiones Lognormales con factores exógenos. Extensiones a partir proceso de difusión de Gompertz*. PhD thesis, University of Granada.

Ricciardi, L.M. (1977): Diffusion processes and related topics in biology. *Lect. Notes Biomath, 14. Springer Verlag*.

Suddendun, B.(1988): *Stochastic Processes in Demography and Applications*. Ed. Wiley Eastern Limited, New Delhi.

The Moving Average Control Chart Based on the Sequence of Permutation Tests

Grzegorz Konczak[1]

Karol Adamiecki University of Economics in Katowice
40-287 Katowice, Bogucicka 14, Poland, *grzegorz.konczak@ae.katowice.pl*

Abstract. The classical methods for monitoring the process mean in quality control procedures are based on the normality assumption. The Shewhart control charts are graphical representations of the sequence of parametric tests. It is important to remember assumptions such as normality and independence.

Permutation tests were introduced by R.A. Fisher in the early 1930's. These tests are a computer-intensive statistical methods. These tests are free of mathematical assumptions, especially completely removes the normality condition. Permutation tests could be used even if the normality assumption is not fulfill. There is presented in the paper a method for monitoring process mean. The construction of the control chart based on the sequence of permutation tests is presented in the paper.

The properties of the proposed control chart are considered in the Monte Carlo study. The simulation study has shown that the permutation control chart could be used for monitoring process mean in the short production run situation. This control chart is useful for monitoring non-normal processes. The permutation control chart gives accurate probabilities of the incorrect out-of-control signals even for non-normal processes.

Keywords: moving average, control charts, permutation tests, bootstrap, Monte Carlo

References

BERTRAND, P.R. and FLEURY, G. (2008): Detecting Small Shift on the Mean by Finite Moving Average. *International Journal of Statistics and Management System, vol. 3 no.1-2, 56-73.*

CHAKRABORTI, S., van der Laan, P. and van de Wiel, M.A. (2004): A Class of Distribution-free Control Charts. *Applied Statistics 53 part 3, 443-462.*

CRAWLEY, M.J. (2005): *Statistics. An Introduction Using R.* John Wiley & Sons, Ltd., London.

EFRON, B. and TIBSHIRANI, R. (1993): *An Introduction to the Bootstrap.* Chapman & Hall, New York.

MONTGOMERY, D. C. (1996): *Introduction to Statistical Quality Control.* John Wiley & Sons, Inc.

SHESKIN, D. J. (2004): *Handbook of Parametric and Nonparametric Statistical Procedures.* Chapman & Hall/CRC, Boca Raton.

Cointegrated Lee-Carter Mortality Forecasting Method*

Josef Arlt[1], Markéta Arltová[1], Milan Bašta[1], and Jitka Langhamrová[2]

[1] Department of Statistics and Probability, University of Economics Prague, W. Churchill sq. 4, 130 67 Prague, Czech Republic *arlt@vse.cz*, *arltova@vse.cz*, *basta@vse.cz*

[2] Department of Demography, University of Economics Prague, W. Churchill sq. 4, 130 67 Prague, Czech Republic *langhamj@vse.cz*

Abstract. The classical Lee-Carter forecasting method is based on the assumption that the overall mortality index follows the random walk model with drift or the ARIMA model. If the two sexes in one country are forecasted separately, the forecasts of male and female mortalities can diverge increasingly over time. This problem can be solved by the Cointegrated Lee-Carter method. On the example of some European countries it is shown that this method leads to more tied up forecasts of the overall mortality index with smaller standard errors in comparison with the classical Lee-Carter method.

Keywords: Mortality, Lee-Carter Method, Cointegration

References

BROUHNS, N., DENUIT, M. and VERMUNT, K.J. (2002): A Poisson Log-bilinear Regression Approach to the Construction of Projected Lifetables. *Insurance: Mathematics and Economics 31, 373-393.*

DEATON, A. and PAXSON, Ch. (2004): *Mortality, Income, and Income Inequality Over Time in the Britain and the United States.* Technical Report 8534 National Bureau of Economic Research Cambridge, MA.

GIROSI, F. and KING, G. (2007): *Understanding the Lee-Carter Mortality Forecasting Method.* Working paper, Harvard University.

JOHANSEN, S. (1991): Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica 59, 1551-80.*

LEE, R.D. (2000): The Lee-Carter Method for Forecasting Mortality, with Various Extensions and Applications. *North American Actuarial Journal 4 (1), 80-93.*

LEE, R.D. and CARTER, L. (1992): Modeling and Forecasting the Time Series of U. S. Mortality. *Journal of the American Statistical Association 87, 659-671.*

LI, N. and LEE, R. (2005): Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography 42(3): 575-594.*

WILMOTH, J. (1993): *Computational Methods for Fitting and Extrapolating the Lee-Carter Model of Mortality Change.* Technical report Department of Demography, University of California, Berkeley.

* This paper was written with the support of Grant Agency of the Czech Republic No. 402/09/0369 "Modelling of Demographic Time Series in Czech Republic".

A Class of Multivariate Type I Generalized Logistic Distributions

Salvatore Bologna

Department of Statistical and Mathematical Sciences, University of Palermo
Faculty of Economics-Viale delle Scienze, Palermo, Italy, *bologna@unipa.it*

Abstract. The logistic distribution has found important applications in many different fields and several different forms of generalizations have been proposed in the literature. However it seems, with a few exceptions, that there are not in the literature forms of multivariate generalized logistic distributions. In this paper we focus on the type I generalized logistic distribution and, based on a procedure of multivariate transformation of multivariate exponential distributions, we introduce a class of multivariate type I generalized logistic distributions. We provide some examples of bivariate and multivariate distributions of this class.

Keywords: type I generalized logistic distribution, multivariate exponential distributions, multivariate type I generalized logistic distributions

References

GUPTA, R.D. and KUNDU, D.: Generalized logistic distributions. *Journal of Applied Statistical Sciences*. (To appear).

JOHNSON, N.L. KOTZ, S. and BALAKRISHNAN, N. (1995): *Continuous Univariate Distributions, vol.2*. Wiley and Sons, New York.

KOTZ, S., BALAKRISHNAN, N., and JOHNSON, N.L. (2000): *Continuous Multivariate Distributions, vol.1*. Wiley and Sons, New York.

ZELTERMAN, D. and BALAKRISHNAN, N. (1992): Univariate Generalized Distributions. In: N. BALAKRISHNAN (Ed.): *Handbook of the Logistic Distribution*. Marcel Dekker, New York, 209-221.

A General Strategy for Determining First-Passage-Time Densities Based on the First-Passage-Time Location Function

Patricia Román-Román, Juan José Serrano-Pérez and Francisco Torres-Ruiz

Departamento de Estadística e Investigación Operativa (Universidad de Granada) Avda Fuentenueva s/n, 18071 Granada, Spain, {*proman,jjserra,fdeasis*}*@ugr.es*

Abstract. This paper presents a general strategy for the efficient application of numerical schemes for solving Volterra integral equations which have as solution first-passage-time density functions associated to certain stochastic processes. Such strategy is based on the information provided by the First-Passage-Time Location function about the location of the variation range of the first-passage-time variable, and it is valid for general type situations that expand on the particular cases considered in Román et al. (2008). In addition, numerical applications are shown to prove the validity of the strategy as well as its computational advantages.

Keywords: Diffusion processes, First-passage-times, Volterra integral equations, First-Passage-Time Location function

Using Observed Functional Data to Simulate a Stochastic Process via a Random Multiplicative Cascade Model

G. Damiana Costanzo[1], S. De Bartolo[2], F. Dell'Accio[3], and G. Trombetta[3]

[1] Dip. Di Economia e Statistica, UNICAL, Via P. Bucci, 87036 Arcavacata di Rende (CS), Italy, *dm.costanzo@unical.it*
[2] Dip. di Difesa del Suolo *V. Marone*, UNICAL, Via P. Bucci, 87036 Arcavacata di Rende (CS), *samuele.debartolo@unical.it*
[3] Dip. di Matematica, UNICAL, Via P. Bucci, 87036 Arcavacata di Rende (CS), *fdellacc@unical.it*, *trombetta@unical.it*

Abstract. *Functional data* has received in recent years considerable interest from researchers and the classical tools from finite multivariate analysis have been adapted to this kind of data. Preda and Saporta (2005) proposed PLS regression in order to perform LDA on functional data. Following this approach, to address the problem of *anticipated prediction* of the outcome at time T of the process in $[0, T]$, in Costanzo *et al.* (2006) we measured the predictive capacity of a LDA for functional data model on the whole interval $[0, T]$. Then, depending on the quality of prediction, we determined a time $t^* < T$ such that the model considered in $[0, t^*]$ gives similar predictions to that considered in $[0, T]$. We consider a new approach based on the definition of special Random Multiplicative Cascades to model the underlying stochastic process. In particular, we consider a class $\mathcal{S}$ of stochastic processes whose realizations are real continuous piecewise linear functions with a constrain on the increment. Let $\mathcal{R}$ be the family of all binary responses Y associated to a process X in $\mathcal{S}$ and consider data from a continuous phenomenon which can be simulated by a pair $(X, Y) \in \mathcal{S} \times \mathcal{R}$, with the same objective of prediction of the binary outcome earlier than the end of the process, we introduce the *adjustement curve* for the binary response Y of the simulated stochastic process X. Such a tool is a decreasing function which would make it possible to predict Y at each point in time before time T. For real industrial processes this curve can be a useful tool for monitoring and predicting the quality of the outcome before completion.

Keywords: functional data, stochastic process, multiplicative cascade

References

COSTANZO, G. D., PREDA, C., SAPORTA, G. (2006), Anticipated Prediction in Discriminant Analysis on Functional Data for binary response. In: Rizzi, A., Vichi, M. (eds.) *COMPSTAT'2006 Proceedings*, pp. 821-828. Springer, Heidelberg.

PREDA C., SAPORTA, G. (2005), PLS regression on a stochastic process. *Computational Satistics and Data Analysis*, **48**:149-158.

Constructing Economic Summary Indexes via Principal Curves

Mohammad Zayed[1,2] and Jochen Einbeck[1]

[1] Department of Mathematical Sciences, Durham University, Science Laboratories, South Rd., Durham, DH1 3LE, UK, *jochen.einbeck@dur.ac.uk*
[2] Applied Statistics and Insurance Department, Mansoura University, Mansoura, 35516, Egypt, *m.a.zayed@dur.ac.uk*

Abstract. Index number construction is an important and traditional subject in both the statistical and the economical sciences. A novel technique based on *localized principal components* to compose a single summary index from a collection of indexes is proposed, which is implemented by fitting a (local) principal curve to the multivariate index data. We exploit the ability of principal curves to extract robust low-dimensional 'features' (corresponding to the summary index) from high-dimensional data structures, yielding further useful analytic tools to study the behaviour and composition of the summary index over time.

Keywords: summary indexes, feature extraction, principal component analysis, smoothing

References

EINBECK, J., TUTZ, G. and EVERS, L. (2005): Local principal curves. *Statistics and Computing 15 (4), 301-313.*

EINBECK, J., EVERS, L., and HINCHLIFF, K. (2009): Data compression and regression based on local principal curves. In: Fink et al. (Eds): Advances in Data Analysis, Data Handling and Business Intelligence, Springer, Heidelberg, 701–712.

HASTIE, T. and STUETZLE, W. (1989): Principal curves. *Journal of the American Statistical Association 84 (406), 502-516.*

MING-MING, Y., JIAN, L., CHUAN-CAI, L. and JING-YU, Y. (2010): Similarity preserving principal curve: an optimal one-dimensional feature extractor for data representation. *IEEE Transactions on Neural Networks*, to appear.

MOSER, J. W. (1984): A principal component analysis of labor market indicators. *Eastern Economic Journal X (3), 243-257.*

TARPEY, T. and FLURY, B. (1996): Self-consistency: a fundamental concept in statistics. *Statistical Science 11 (3), 229-243.*

THEIL, H. (1960): Best linear index numbers of prices and quantities. *Econometrica 28 (2), 464-480.*

TINTNER, G. (1946): Some applications of multivariate analysis to economic data. *Journal of the American Statistical Association 41 (236), 472-500.*

Regression Diagnostics for Autocorrelated Models with Moving Average Errors

Sugata Sen Roy[1] and Sibnarayan Guria[2]

[1] Department of Statistics,
University of Calcutta, Kolkata, India.
senroy_sugata@hotmail.com
[2] Department of Statistics
West Bengal State University, Kolkata, India.
sng_65@yahoo.co.in

Abstract. Autocorrelation in a regression model is a common phenomenon in most practical studies. The presence of autocorrelated errors require the use of the generalized least-squares technique in estimating the regression coefficients. Using the deletion technique in checking for outliers in such a model leads to the disruption of the error structure. In this paper we take account of this disruption in studying the diagnostics of a regression model whose errors follow a first-order moving average process.

Keywords: Moving average process, outliers, deletion technique

References

Belsley,D.A., Kuh, E. and Welsch, R.E., (1980): *Regression Diagnostics.* John Wiley, New York.

Cook, R.D. (1986): *Assessment of local Analysis, J. Roy. Statist. Soc., Ser.B, 48, 133-169.*

Haslett, J. and Hayes, K. (1998): *Residuals for linear model with general covariance structure, J.R. Statist. Soc. B, 60, part 1, 201-215.*

Sen Roy, S and Guria, S.N. (2004): *Regression Diagnostics in an Autocorrelated Model, Brazilian Journal of Probability and Statistics,18, 103-112.*

Latent Variable Regression Model for Asymmetric Bivariate Ordered Categorical Data: A Bayesian Approach

Rasool Gharaaghaji[1] and Soghrat Faghihzadeh[2]

[1] Department of Biostatistics and Epidemiology, Faculty of Medical Science Urmia Medical Science Unicersity, P.O.Box: 5756115111, Urmia, Iran *rasool1350@yahoo.com*

[2] Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University
Theran, Iran. *faghihz@modares.ac.ir*

Abstract. The analysis of correlated ordinal responses is usually more complex than continuous and binary response data. Particularly, existing methods in modeling ordered data have not been developed to asymmetric bivariate or multivariate ordinal responses. In this area, for estimating model parameters, classical approach usually fits the regression model using maximum likelihood approach and inferences about the model parameters are based on the associated asymptotic theory (see Zayeri et al. (2006)). In this paper, we use a latent variable regression model for analyzing asymmetric bivariate ordinal response data. In this model, a generalization of bivariate Gumbele's distribution (Gumbele (1961)), which developed by Satterthwaite and Hutchinson (1978),

$$G_\nu(x,y) = \left(1 + e^{-x} + e^{-y}\right)^{-\nu},$$

$$f(x,y) = \frac{\nu(\nu+1)e^{-x-y}}{(1+e^{-x}+e^{-y})^{\nu+2}},$$

was utilized as the latent variable. In addition, we used a Bayesian approach and MCMC algorithm (such as Gibbs Sampling and Metropolis- Hasting) for estimating the model parameters $(\beta, \theta, \nu, y^*)$.

Keywords: Asymmetric Ordinal Response, Bayesian Analysis, MCMC, Latent Variable.

References

GUMBEL, E. J. (1961): Bivariate Logestic Distribution. *Journal of the Acoustical Society of America 56, 335–349.*

SATTERTHWAITE, S. P. and HUTCHINSON, T. P. (1978): A generalization of Gumbel's bivariate logistic distribution. *Metrika, 25, 163–170.*

ZAYERI, F. and KAZEMNEJAD, A. (2006): A latent variable regression model for asymmetric bivariate ordered categorical data. *Applied Statistics 33, 743–753.*

Determinants of the Italian Labor Productivity: A Pooled Analysis

Margherita Velucchi[1] and Alessandro Viviani[2]

[1] Department of Statistics "G.Parenti" - Università di Firenze, Viale G.B. Morgagni, 59 - 50134 Firenze, Italy
velucchi@ds.unifi.it
[2] Department of Statistics "G.Parenti" - Università di Firenze, Viale G.B. Morgagni, 59 - 50134 Firenze, Italy
viviani@ds.unifi.it

Abstract. Since the early '90s, the Italian economy has been characterized by a relative decline in its economic growth. Leading Italian economists attribute this to the feature of the Italian productive system (low labor productivity, small firm size and specialization in traditional sectors). This paper investigates how some relevant aspects of firms behavior affected the dynamics of the Italian firms' labor productivity (1998-2004) using an original database from the Italian National Institute of Statistics at a micro level (firm level). In particular, we test how the cost of labor, investments in R&D activities and patents, size of firms as well as the sector of activity (ATECO 2004) influenced the poor performance of Italian firms labor productivity growth in the period considered. We run a pooled regression model on manufacturing (high and low tech sectors) and services, separately. We find evidence of a strong positive relationship between labor productivity growth and size of firms, investments in intangible assets and cost of labor. We also stress the key role of technology in both manufacturing and service sectors.

Keywords: labor productivity, pooled regression, intangible capital

References

KLEINKNECHT, A. and MOHNEN, P. (Eds.),(2002): *Innovation and Firm Performance*, London, Palgrave.

GRIFFITH, R., REDDING, S. and VAN REENEN J. (2004): Mapping the Two Faces of R&D: Productivity Growth in a Panel of OECD Industries, *Review of Economics and Statistics*, 86(4), pp. 883-895.

Calibration through Shuttle Algorithm: Problems and Perpectives

Lucia Buzzigoli[1], Antonio Giusti[1], and Monica Pratesi[2]

[1] Department of Statistics, University of Florence
Viale Morgagni, 59, I 50134 Florence, Italy
buzzigoli@ds.unifi.it, giusti@ds.unifi.it
[2] Department of Statistics and Mathematics Applied to Economics
Via Ridolfi, 10, I 56122 Pisa, Italy *m.pratesi@ec.unipi.it*

Abstract. The main objective of the calibration methodology is to use auxiliary information to obtain estimators that are approximately unbiased with a variance smaller than that of the Horvitz-Thompson estimator (Deville and Särndal (1992)). In many cases the final weights are able to make the estimates be consistent or reproduce the auxiliary information, which consist of known marginal counts or totals in a two or a multi-way table (Deville and Tillé (2004)). When a weighted sum of squares distance is used to measure the distance between the two sets of weights, the estimator obtained through calibration corresponds to a generalized regression estimator.
The execution time of the estimation software used in calibration is proportional to the number of auxiliary variables and to the level of complexity of the multi-way table. As far as the individual weights are concerned, these can be negative with results of no sense in many real life applications. Under more general constraint equations this problem can be overcome, but the execution time is obviously increased by the complexity of the constraint. The problem can be faced efficiently throughout algorithms, which analyze the structure of an unknown n-dimensional array given its marginal distributions.
The authors propose to estimate calibration weights by means of the Shuttle algorithm (Buzzigoli and Giusti (2006)), a simple method to calculate lower and upper bounds of a generic n-way array given all its (n-1) marginals. The algorithm has relevant links with probabilistic and statistical aspects and is particularly easy to implement, has a low storage requirement and is very fast. Some first applications show promising results.

Keywords: Sample Survey, Calibration, Shuttle Algorithm

References

BUZZIGOLI, L. and GIUSTI, A. (2006): From marginals to array structure with the Shuttle algorithm. *Journal of Symbolic Data Analysis 4, 1-14.*

DEVILLE, J.C. and SÄRNDAL, C.E. (1992): Calibration estimators in survey sampling. *Journal of the American Statistical Association 87, 376-382.*

DEVILLE, J.C. and TILLÉ, Y. (2004): Efficient balanced sampling: The cube method. *Biometrika 91 (4), 893-912.*

Statistical Power and Sample Size Requirements in Experimental Studies with Hierarchical Data

Satoshi Usami[1]

Graduate School of Education, the University of Tokyo. *usami_s@p.u-tokyo.ac.jp*

Abstract. A Hierarchical Linear Model (HLM) is a regression model for hierarchical data sets and has attracted interest in various areas of psychological, social, educational and clinical research. Examples of hierarchical relation include patients in hospitals, student in schools (e.g., Raudenbush & Bryk, 2002). HLM allows variance in outcomes to be analyzed at multiple hierarchical levels, whereas in a simple regression model, all effects are modeled to occur at a single level.

Statistical power and sample size requirements are fundamental problems in research design, and are generally evaluated by power analysis. Key decisions in power analysis include setting the number of units and clusters in order to have a desired power for detecting meaningful differences. In past research into HLMs, several researches derived formulae for evaluating statistical precision and optimal sample sizes in two-level hierarchical cluster randomized trials (CRTs: i.e., the cluster is randomized, not the units). Recently, Heo & Leon (2008) derived a closed-form power function and a formula for determining sample size to detect a single intervention effect on outcome in three-level hierarchical CRTs.

In experimental research, research hypotheses are mainly for main effects and interaction effects of interventions. More specifically, experimental research focuses on tests of contrasts for these effects. However, most of the proposed techniques so far focused on some specific situations and cannot be applied to a wide range of experimental designs since the number of factors and groups are restricted.

In the present research, we propose a general method for evaluating statistical power to test intervention effects in experimental research with hierarchical data. This approach enables statistical power to be evaluated for various types of contrasts in regards to main effects and interaction effects within multiparameter tests based on Wald statistics. Additionally, we show how power curves for various contrasts change for various values of effect size, sample size, intraclass correlation.

Keywords: statistical power, sample size, hierarchical linear model, multilevel model, experimental study

References

Heo, M. & Leon, A.C. (2008): Statistical power and sample size requirements for three level hierarchical cluster randomized trials. *Biometrics, 64, 1256-1262.*

Raudenbush, S.W. & Bryk, A.S. (2002): *Hierarchical linear models: Applications and data analysis methods.* (2nd ed.). London: Sage.

Properties of range-based volatility estimators

Peter Molnár[1]

Norwegian School of Economics
Helleveien 30, 5045 Bergen, Norway, *Peter.Molnar@nhh.no*

Abstract. Volatility plays crucial role in many areas of finance and economics. It is not directly observable and must be estimated. If we have only daily closing prices and we need to estimate volatility on a daily basis, the only estimate we have is squared daily return. [1] This estimate is very noisy. However, closing prices are not the only daily data available. For most financial data, open, high and low daily prices are available too. Range, the difference between high and low prices is natural candidate to be used for volatility estimation. Parkinson (1980) introduces range (the difference between high and low prices) as a volatility estimator which is less noisy than squared returns. Garman and Klass (1980) subsequently introduce estimator based on open, high, low and close prices, which is even less noisy.

We study properties of range-based estimators and find that the best estimator is Garman and Klass (1980) estimator. The property we focus the most is the effect of the use of range-based volatility estimators on the distribution of returns scaled by their standard deviations. Using volatility estimated from high frequency data, Andersen et al. (2001) show that normalized returns are indeed Gaussian. Contrary, returns scaled by sigmas estimated from GARCH type of models (based on daily returns) are not Gaussian, they have fat tails. We show that even when returns are normally distributed, returns standardizes by (imprecisely) estimated volatility are not normally distributed. However, approximate normality of standardized returns is obtained for Garman-Klass volatility estimator. We test this estimator empirically and find that we can obtain the same results from daily data as Andersen et al. (2001) obtained from high-frequency (transaction) data.

Keywords: volatility, range, high, low

References

ANDERSEN, T.G., BOLLERSLEV, T., DIEBOLD, F.X., EBENS, H., (2001): The distribution of stock return volatility. *Journal of Financial Economics 61, 43-76.*

GARMAN, M., KLASS, M., 1980: On the estimation of security price volatilities from historical data. *Journal of Business 53, 67-78.*

PARKINSON, M., 1980: The extreme value method for estimating the variance of the rate of return. *Journal of Business 53, 61-65.*

[1] Since mean daily return is much smaller that its standard deviation for most of the financial assets, we assume that mean daily return is zero.

Using mixture of distributions

Mare Vähi

Institute of Mathematical Statistics, University of Tartu
J Liivi 2-516, 50409 Tartu, Estonia, *Mare.Vahi@ut.ee*

Abstract. The aim is to identify the suitable distribution for modeling the future changes in fertility distributions. Different distributions are used: Beta distribution, Gamma distribution, single and mixture Hadwiger function. The results indicate that mixture Hadwiger model is useful in describing fertility curves of the Estonia.

The Hadwiger function is expressed as

$f(x) = \frac{\alpha\beta}{\gamma\sqrt{\pi}} \left(\frac{\gamma}{x}\right)^{\frac{3}{2}} \exp\left(-b^2\left(\frac{\gamma}{x} + \frac{x}{\gamma} - 2\right)\right)$ (1),

where x is the age of mother at birth and α, β, γ are the parameters to be fitted.

If population is not homogeneous but two somewhat different populations it would be convenient to apply a combination of a pair of curves: Chandola et al. (1999).

The mixture function is expressed as $f(x) = mf_1(x) + (1-m)f_2(x)$,

where x is the age of mother at birth, m is the mixture parameter that determines the relative size of two component distributions and $f_1(x), f_2(x)$ are the Hadwiger functions in subpopulations given by (1).

The fit of all models is tested using period data by single year of age of mother. The data were obtained from statistical database of Statistical Office of Estonia.

Keywords: Hadwiger function, mixture of distributions, modeling

References

CHANDOLA, T., COLEMAN, D. A. and HIORNS R. W. (1999): Recent European fertility patterns: Fitting curves to 'distorted' distributions. *Population Studies 53, 317-329.*

ORTEGA, J. A. and KOHLER, H.-P. (2002): Measuring Low Fertility: rethinking demographic methods. *MPIDR Working Paper.* Rostok.

Comparing ORF Length in DNA Code Observed in Sixteen Yeast Chromosomes

Anna Bartkowiak[1 2] and Adam Szustalewicz[1]

[1] Institute of Computer Science, University of Wrocław, Jolliot Curie 15, 50-383, Wrocław, PL *aba@ii.uni.wroc.pl, asz@ii.uni.wroc.pl*
[2] Wrocław High School of Applied Informatics, Wejherowska 28, 54-239, Wrocław, PL

Abstract. Generally, the yeast genome – composed from 16 chromosomes – contains 6686 ORFs (Open Reading Frames) with inscribed genetic information about the functioning of the yeast organism – see Christianini and Hahn (2007). We count the ORF length in molecules named amino-acids. Bartkowiak (2008) has shown that ORF length found in four chromosomes can be described by the Negative Binomial (NB) distribution, characterized by two parameters: r and p.

Now we consider the same random variable (i.e. ORF length, counted in amino-acids) for all the 16 yeast chromosomes containing together $n = 6686$ ORFs. We notice that the estimates of the parameters $\hat{r}_i, \hat{p}_i$ exhibit very similar values. This is shown by displaying confidence ellipses for estimates of the pairs $\hat{r}_i, \hat{p}_i$ for individual chromosomes and a joint estimate for the Grand Total, i.e. for all data taken together. The confidence ellipses were evaluated on the basis of the observed information matrix obtained from the Hessian of the log-Likelihood (Stuart et al., 1999). Using the Likelihood Ratio (LR) test we show that the hypothesis on the equality of the parameters in individual chromosomes can not be rejected. We confirm further this hypothesis by some simulation experiments using NB pseudo-random numbers generator 'nbinrnd' from Matlab Statistical Toolbox (Matlab 2002).

It is really amazing that such a simple probabilistic model like the NB distribution is able to describe such complicated phenomenon as the DNA code working in an environment subjected to cell divisions, cell repairs and mutational pressure.

Keywords: DNA code, amino-acids, ORF length, negative binomial, parameter estimation, confidence ellipses

References

BARTKOWIAK, A. (2008): Orf length is negative binomial – why? In: P. Brito (Ed): *COMPSTAT 2008, Proceedings in Computational Statistics*, Contributed papers. Physica Verlag, a Springer Company, 291–298.

CHRISTIANINI, N. and HAHN, M. W. (2007): *Introduction to Computational Genomics. A Case Studies Approach*. Cambridge University Press, UK.

MATLAB STATS TOOLBOX (2002): *Statistics Toolbox For Use with MATLAB*, Users Guide Version 4., ©1993 - 2002 by The MathWorks, Inc., 6th printing.

STUART, A. et al. (1999): *Kendall's Advanced Theory of Statistics*, Volume 2A, Sixth Edition, Arnold, London.

A New Computational Approach to Calculate Tests Sizes for Unconditional Non-inferiority Tests

Félix Almendra-Arao

UPIITA del Instituto Politécnico Nacional
Av. Instituto Politécnico Nacional 2580, 07340, México, D. F., México.
falmendra@ipn.mx

Abstract. There is a very wide range of spheres in which comparing two groups subjected to different treatments is naturally required. Among the possible ways of comparing two treatments, one has recently gained importance, especially in the area of clinical trials: the so-called non-inferiority tests, these are statistical procedures used to verify whether there is sample evidence that a new treatment is not substantially inferior in terms of its effectiveness than a control treatment.

In the area of non-inferiority tests, a problem of practical interest is the calculation of test sizes. This calculation presents difficulties because a nuisance parameter appears, enormously complicating the calculations and making them a computationally intensive problem. In the literature rarely does the calculation of test sizes for non-inferiority tests appear, apparently due to the computational effort required.

An important result in the direction of achieving more efficiently calculated test sizes is a theorem owed to Röhmel and Mansmann (1999). The merit of this theorem is enormous as it greatly simplifies calculation of test sizes when critical regions fulfill the Barnard convexity condition. In spite of this, however, it can be computationally intensive even if the exhaustive method is used to obtain the maximum required in this theorem, as typically has been done. It is therefore advisable to seek a more practical and computationally less demanding alternative than the exhaustive method.

In search of an alternative to reduce this calculation, it is natural to investigate the possible application of Newton's method because it is a very efficient optimization method. However representation of derivates is extremely complicated to calculate. In this work is obtained a quite manageable closed form for the first two derivatives of the power function because when is assumed that the critical regions of the tests fulfill the Barnard convexity condition, therefore Newton's method can be applied to computation of test sizes reducing computational time considerably.

Keywords: Test size, unconditional test, non-inferiority test

References

RÖHMEL, J. and MANSMANN, U. (1999): Unconditional nonasymptotic one sided tests for independent binomial proportions when the interest lies in showing noninferiority y or superiority. *Biom. J. 2:149-170.*

A functional relationship model for simultaneous data series

Xiaoshu Lu

Finnish Institute of Occupational Health, Topeliuksenkatu 41 a A, FIN-00250 Helsinki, Finland, *xiaoshu@cc.hut.fi*

Abstract. Modelling the relationship between simultaneous data series is important for a wide variety of applications. Despite apparently wide application, methodologies are at present still inadequate. The well-known modelling technique is longitudinal data analysis. Longitudinal data analysis mainly focuses on how to handle the within-person correlations among the repeated measurements. Data are often obtained with few measures per subject and the models are formulated as linear. For two longitudinal data with large-scale measures for subjects, the dimensionality is high, hence there are few robust alternatives that can successfully address the unique features characterised by the data. As such, another technique, referred to as functional data analysis, is often employed. Various smoothing techniques are introduced and applied for analysing functional data sets, such as curves and shapes. A sufficiently large amount of data is needed to adequately approximate the function. However, many data series are short, hence functional data model may not be able to simulate with reasonable accuracy. In addition, a significant characteristic of real life's data is their nonlinear nature. It is thus desirable to devise a method able to discover and identify the nonlinear structure of the relationship between the data series. The purpose of this study is to present a new mathematical methodology for addressing all these issues. We extend the literature to both periodic time series and longitudinal data. The main difference of the proposed model from other methods is its capability for identifying complex nonlinear structure of the relationship behind the simultaneous data series. We use singular value decomposition technique to extract and model the dominant relationship between two data series. The functional relationship can be used to explore complex interplay among the mechanical and physical factors which govern the targeting system. The dataset of measured computer-related workload and health outcome was used to test the proposed model with promising results even though the data suffer from a number of limitations such as collection of time series of the data is short. In addition, computation algorithms are relatively simple which are easily computed by computers with available commercial software.

Keywords: simultaneous data series, periodic time series, longitudinal data, functional relationship, mathematical modelling

Change point Detection in trend of mortality DATA

Firouz Amani[1], Anoshirvan Kazemnejad[1], and Reza Habibi[2]

[1] Department of biostatistics, Ardabil Medical University Faculty of Medical Science
Box. 5615787881, Ardabil, Iran
[2] Department of Statistics, Central Bank of Iran, Tehran, Iran

Abstract. Mortality refers to death that occurs within a population. It is linked to many factors such as age, sex, race, occupation and social class. Chang in the pattern of mortality trend can affect the population standards of living and health care. This event makes a change point is occurred in mortality rates. The aim of this study is to detect change point in Iranian mortality data during 1970 to 2007. We use several frequencies and Bayesian methods to estimate the change point under both Poisson and poisson regression modeling for mortalities. All method show that a change has occurred in mortality rates at 1993. This result corresponds to descriptive result of Statistical Center of Iran.

Keywords: change point, MCMC, mortality, Poisson regression, Bayes information criterion

Choosing variables in cluster analysis based on investigating correlation between variables

Jerzy Korzeniewski

Department of Statistical Methods University of Lodz
POW 3/5 , Lodz , Poland *jurkor@wp.pl*

Abstract. In the paper a novel technique of variable choice in the context of cluster analysis is proposed. The characteristic feature of this technique is that it does not need neither the specification of the number of clusters nor any reference to any particular method of object grouping. The technique is based entirely on the investigation of a kind of variable correlation measure designed specially for the task. This correlation measure is defined in the following way.

$$CORR(A,B) = \frac{cov(d_A, d_B)}{stdev(d_A)\, stdev(d_B)}.$$

The measure defined above is a linear correlation coefficient between distances calculated on two different sets A and B of variables. In order to compute this coefficient we have to draw a number of pairs of objects and to repeat this drawing and computation a number of times to find a relatively stable value. Both of these numbers have to be fixed throughout the whole procedure. This measure is very effective in searching for variables creating data set cluster structure. The idea behind this reasoning is as follows. If there is a distinct cluster structure in a data set then any two sets of variables creating this structure should be positively correlated in the sense of the measure proposed (pairs of objects belonging to the same cluster should have small distances on both sets of variables and pairs of objects belonging to different clusters should have big distances on both sets of variables). Looking for variables important to cluster structure we associate positive notes with sets of variables with positive value of the coefficient and negative notes with one or both sets of variables with negative value of the coefficient. One can use this idea in a number of ways. We are going to present the results of a simple algorithm which minimizes the average value of the coefficient computed for all possible pairs of variables both of which belong to one of two different sets of variables into which the set of all variables is split. We asses the efficiency of this algorithm in a simulation experiment following closely Steinley and Brusco pattern (Steinley & Brusco, 2008). In presence of correlation between masking variables the algorithm almost never goes wrong. When there is no correlation between masking variables the results are fractionally worse.

Keywords: cluster analysis, variable choice

References

STEINLEY, D., BRUSCO, M. (2008): Selection of variables in cluster analysis: an empirical comparison of eight procedures. *Psychometrika 73 (1), 125-144.*

Nonparametric approach for Scores of Department Required Test

Li-Fei Huang[1]

Department of Applied Statistics and Information Science, Ming Chuan University
5 Teh-Ming Rd., Gwei-Shan Taoyuan County 333, Taiwan
lhuang@mail.mcu.edu.tw

Abstract. The data comes from the ROC College Entrance Examination Center in 2009. I would like to analyze the scores of Chinese, English, Scientific Mathematics (abbreviated SciMath) and Social Scientific Mathematics (abbreviated SSciMath), which are the four most important subjects in the Department Required Test. I obtain a random sample of 5000 scores from about 80,000 high school students who attained the Department Required Test. The examination time is 80 minutes and the range of test score is from 0 to 100 for all subjects.

There are three major interests:

(1) Will the gender affects students performance in different subjects?
(2) Is there any difference in study performance for students who live in the city and students who live in the countryside?
(3) Is there an interaction between the gender and the residence?

The confidence interval for the ratio of scores can provide a clear standard to justify the difference in study performance. If the difference in study performance indeed exists, some steps must be taken to shorten the difference.

Keywords: Median, Ratio of scores, Empirical confidence interval, Non-parametric confidence interval

References

HUANG, LI-FEI and JOHNSON, R.A. (2006): Confidence Regions for the Ratio of Percentiles. *Statistics and Probability Letters 76 (2006) 384-392.*

HUANG, LI-FEI and JOHNSON, R.A. (2003): Some Exact and Approximate Confidence Regions for the Ratio of Percentiles from Two Different Distributions In: Lindqvist, B. and Doksum, K.(Eds): *Mathematical and Statistical Methods in Reliability.* World Scientific, Singapore, 455–468.

Changes of Proportions of Overlooked Dementia in the Japanese Elderly during 5.9 Years

Chisako Yamamoto[1] and Tanji Hoshi[2]

[1] School of Nursing, Jobu University
Shimmachi, Takasaki 370-1393, Japan, *chisako-y@k9.dion.ne.jp*
[2] Department of Urban Environment, Tokyo Metropolitan University
Minamiosawa, Hachioji 192-0397, Japan, *star@onyx.dti.ne.jp*

Abstract. With the world longest life expectancy of Japanese, the number of people affected by dementia is estimated to jump up from 2.05 million in 2005 to 4.45 million in 2035. Early detection and treatment of dementia is essential, but early symptoms of dementia are often misdiagnosed as senile memory disorder. The previous study by Boise, Camiciolib and Morgan(1999) suggested diagnosis of dementia hit only 50% in primary care setting, thus it is overlooked and diagnosed one to two years late as reported by Kurz, Scuvee-Moreau, Salmon et al. (2001). Municipalities have been providing programs of early detection and prevention of dementia as one part of health promotion during past decade.

Yamamoto and Hoshi (2010) clarified the proportions of overlooked dementia in 2001, which were 6.3% in men and 9.3% in women, by statistically analysing data of 13,068 elders in association of mortality and longevity. The purpose of this study is to clarify changes of proportions of overlooked dementia after 2001. Analyzed subjects were 13,182 in 2004 and 15,084 in 2007. Using receiver operator characteristic curves, cognitive scores were measured by performance of deposits/withdrawals, filling out forms/documents and reading books/newspapers, whose odds ratios to demented status were shown high by multiple logistic regression. It significantly shows changes of proportions of overlooked dementia during 5.9 years and verifies effectiveness of those programs for quality life of the elderly.

Keywords: proportions of overlooked dementia, multiple logistic regression, receiver operator characteristic curves, 5.9-year follow-up

References

BOISE, L., CAMICIOLIB, R. and MORGAN, D. L. (1999): Diagnosing dementia: perspectives of primary care physicians. *Gerontologist* 39(4), 457-464.

KURZ, X., SCUVEE-MOREAU, J. and SALMMON, E. et al. (2001): Dementia in Belgium: prevalence in aged patients consulting in general practice. *Revue Medical de Liege* 56(12), 835-839.

YAMAMOTO, C. and HOSHI, T. (2010): Proportions of overlooked dementia in the community-dwelling elderly: the relationship between cognitive impairment and 5.9-year survival in an urban population. *Journal of Health and Welfare Statistics* (forthcoming).

A Model-Based Approach to Identify Historical Controls in Clinical Trials

Jessica (Jeongsook) Kim[1] and John Scott[1]

Center for Biologics Evaluation and Research, FDA
1401 Rockville Pike, Rockville, MD 20852-1448, USA *Jessica.Kim@fda.hhs.gov*

Abstract. In certain clinical trials, one may encounter situations where no active controls are available. In designing such studies, both placebo controls and historical controls have been used as comparators. Understanding historical control rates of some event may present a challenge to the drug development process due to innovations in medical treatments or changes in medical practice over time and the comparability of the specific target patient populations. Our goal is to identify a better method of determining appropriate historical control rates for the proper efficacy evaluation of biologic products. In addition to considering standard applications of meta-analysis, we will apply Bayesian meta-analysis techniques to identify appropriate historical control rates using information from literature reviews. We will discuss a model-based approach to obtain a more representative evaluation of possible historical control rates and control populations. We will evaluate these methods by comparing intervals (confidence and credible) and via cross-validation procedures using subsets of the available data. There may be sensitivity to the prior on the between-study variability in a Bayesian meta-analysis, espeically with a small number of studies and/or small numbers of patients in each study(poor prior info). We will explore a variety of plausible priors for the between-study variability, especially in the class of so-called non-informative priors.

Keywords: historical control, Bayesian meta-analysis, prior information

References

BERLIN, J. A., LAIRD, N. M., SACKS, H. S. and CHALMERS, T. C. (1989): A random-effects regression model for meta-analysis. *Statistics in Medicine 14, 395–411.*

CARLIN, B. P. and LOUIS, T. A. (2008): *Bayesian Methods for Data Analysis, Third Edition.* Chapman and Hall / CRC Press, London.

DEEKS, J. J. (2002): Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine 21, 1575–1600.*

MORRIS, C. N. and NORMALAND, S. L. (1992): Hierarchical models for combining information and for meta-analyses. In: J. M. Bernando, et al. (Eds.): *Bayesian Statistics 4.* Clarendon Press, London, 338–341.

SMITH, T. C., SPIEGELHALTER, D. J. and THOMAS, A. (1995): Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine 14, 2685–2699.*

SPIEGELHALTER, D. J., ABRAMS, K. R. and MYLES, J. P. (2004): *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* Wiley, New York.

Accurate Distribution and its Asymptotic Expansion for the Tetrachoric Correlation Coefficient.

Haruhiko Ogasawara

Department of Information and Management Science, Otaru University of Commerce
3-5-21, Midori, Otaru 047-8501 Japan, *hogasa@res.otaru-uc.ac.jp*

Abstract. Accurate distributions of the estimator of the tetrachoric correlation coefficient and, more generally, functions of sample proportions for the 2 by 2 contingency table are derived. The results are obtained given the definitions of the estimators even when some marginal cell(s) are empty. Then, local asymptotic expansions of the distributions of the parameter estimators standardized by the population asymptotic standard errors up to order $O(1/n)$ and those of the studentized ones up to the order next beyond the conventional normal approximation are derived. The asymptotic results can be obtained in a much shorter computation time than the accurate ones. Numerical examples were used to illustrate advantages of the studentized estimator of Fisherfs z transformation of the tetrachoric correlation coefficient.

For the full result of this paper, see Ogasawara (2010).

Keywords: tetrachoric correlation coefficient, Edgeworth expansion, Cornish-Fisher expansion, asymptotic cumulants, studentized estimators, Fisherfs z transformation.

References

OGASAWARA, H. (2010): Accurate distribution and its asymptotic expansion for the tetrachoric correlation coefficient. *Journal of Multivariate Analysis, 101 (4), 936-948).*

Error augmentation for the conditional score in joint modeling

Yih-Huei Huang[1], Wen-Han Hwang[2], and Fei-Yin Chen[1]

[1] Department of Mathematics, Tamkang University, Taipei County, Taiwan. *yhhuang@mail.tku.edu.tw*
[2] Department of Applied Mathematics and Institute of Statistics, National Chung Hsing University, Taichung, Taiwan. *wenhan@nchu.edu.tw*

Abstract. It happens that the times of repeat measurements depends on the outcome variable. In a joint model of time-to-event and longitudinal data, the longitudinal data is available or meaningful only before the time to event. Thus, those who have longer studying time should have more measurements. It follows that the average of measurements or coefficients of estimated regression function have distributions depend on response variable. This is a kind of differential measurement error problem. In this talk, an error augmentation algorithm will be introduced to modify existent conditional score analysis to incorporate such differential measurement error. We found that the modification is easy, robust, consistent and can result more efficient estimators.

Keywords: measurement error, conditional score, joint modeling, error augmentation

References

Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C. M. (2006): *Measurement Errors in Nonlinear Models*. Chapman & Hall, London.

Tsiatis, A. A. and Davidian, M. (2001): A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika 88, 447-458*

On the use of random forests and resampling techniques for predicting the duration of chemotherapy-induced neutropenia in cancer patients

Susana San Matías, Mónica Clemente, and Vicent Giner-Bosch

Departamento de Estadística e Investigación Operativa Aplicadas y Calidad,
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain
ssanmat@eio.upv.es, *mclement@eio.upv.es*, *vigibos@eio.upv.es*

Abstract. Febrile neutropenia (FN) is a frequent side-effect in cancer patients treated with chemotherapy. Determining the duration of the FN episode is essential because high FN durations are associated with a higher probability of suffering serious complications, including death. In recent papers, as Lalami et al. (2006) and San Matías et al. (2009), several scoring models have been proposed in order to classify patients into one of two categories -namely predicted low or high FN duration. San Matías et al. showed also that the predictive ability of scoring models can be outperformed using a combination of statistical techniques. The problem consists on developing such a predictive methodology when the training dataset has a small sample size. In the present work we propose the use of random forests (see Breiman (2001)) for predicting the duration class of a patient. Particularly, numerical results are shown in the case that we have a small dataset. A discussion and comparison with other resampling techniques is presented.

Keywords: Neutropenia, Random Forests, Survival Analysis

References

BREIMAN, L. (2001): Random forests. *Machine Learning 45 (1), 5-32.*

LALAMI, Y., PAESMANS, M., MUANZA, F., BARETTE, M., PLEHIERS, B., DUBREUCQ, L., GEORGALA, A. and KLASTERSKY, J. (2006): Can we predict the duration of chemotherapy-induced neutropenia in febrile neutropenic patients, focusing on regimen-specific risk factors? A retrospective analysis. *Annals of Oncology 17 (3), 507-514.*

SAN MATÍAS, S., CLEMENTE, M., GINER-BOSCH, V. and GINER, V. (2009): Predicting the duration of chemotherapy-induced neutropenia: new scores and validation. *Technical Report DEIOAC-2009-03, Departamento de Estadstica e IO Aplicadas y Calidad, Universidad Politcnica de Valencia.*

On estimation of tree abundance from a presence-absence map

Wen-Han Hwang

Department of Applied Mathematics and Institute of Statistics,
National Chung Hsing University, Taichung, Taiwan, *wenhan@nchu.edu.tw*

Abstract. A presence-absence map is consisted of regular rectangle grids, where each grid flags the occurrence or unoccurrence of a tree species in the rectangle. The scientific question arising from here is how to estimate the abundance of the species using the presence-absence map only. Instead of the negative binomial model which is often used in ecology, we propose a mixed gamma poisson model to describe the species distribution over the grids. A novel reproduced three-map techinque is introduced to estimate the species abundance as well as other parameters. We evalute the performance of the proposed method through a forest data set which contains 817 tree speices and the total tree abundance is over than 300,000. The forest plot is in Malysia, it has an area of 50 hectares and the data was censused during 1985-1987.

Keywords: negative binomial, mixed gamma Poisson, random placement Model

References

HE, F, and GASTON, K.J. (2000): Estimating species abundance from occurrence. *American Naturalist 156 553-559.*

KUNIN, W.E. (1998): Extrapolating species abundance across spatial scales. *Sicence 281 1513-1515.*

Scoring vs. statistical classification methods for predicting the duration of chemotherapy-induced neutropenia

Vicent Giner-Bosch, Susana San Matías, and Mónica Clemente

Departamento de Estadística e Investigación Operativa Aplicadas y Calidad,
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain
vigibos@eio.upv.es, *ssanmat@eio.upv.es*, *mclement@eio.upv.es*

Abstract. Chemotherapy-induced neutropenia (CIN) is the most common side effect associated with the administration of anticancer drugs. Up to 25% of patients treated with chemotherapy are likely to develop a febrile neutropenia (FN) episode (Crawford et al. (2004)). It is now accepted that there is a relationship between the aggressiveness of the chemotherapy regimen and FN duration. Lalami et al. (2006) developed a scoring model which aims to predict CIN duration according to the aggressiveness of the cytotoxic regimen. However, we found contradictory results when applying their proposal using a new sample. For this reason, in this work our goal has been to compare the performance of scoring methods for predicting FN duration vs. new predictive systems based on statistical classification techniques, using as an input only the information on the cytotoxic regimen.

We have developed three new scores for predicting if a patient will belong either to a low or high duration group. Two of them are scoring methods and the third one was built by using advanced statistical tools such as cluster analysis (Hastie et al. (2001)) and classification trees (Breiman (1984)). This methodological approach identifies the patients that will need the longest times to recover from FN. Numerical results of all the predictive methods are presented and discussed.

Keywords: Biostatistics, Classification trees, Cluster, Febrile Neutropenia

References

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, C.J. (1984): *Classification and regression trees*. Chapman & Hall/CRC, New York.

CRAWFORD, J., DALE, D.C. and LYMAN, G.H. (2004): Chemotherapy-induced neutropenia. *Cancer 100 (2), 228-237*.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001): *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer, New York.

LALAMI, Y., PAESMANS, M., MUANZA, F., BARETTE, M., PLEHIERS, B., DUBREUCQ, L., GEORGALA, A. and KLASTERSKY, J. (2006): Can we predict the duration of chemotherapy-induced neutropenia in febrile neutropenic patients, focusing on regimen-specific risk factors? A retrospective analysis. *Annals of Oncology 17 (3), 507-514*.

Robust Model for Pharmacokinetic Data in 2x2 Crossover Designs and its Application to Bioequivalence Test

Yuh-Ing Chen[1] and Chi-Shen Huang[2]

[1] Institute of Statistics, National Central University, Jhongli, Taiwan 32054, R.O.C, *ychen@stat.ncu.edu.tw*
[2] Institute of Statistics, National Central University, Jhongli, Taiwan 32054, R.O.C, *92245001@cc.ncu.edu.tw*

Abstract. To have a robust analysis of pharmacokinetic (PK) data in a 2x2 crossover design where involves both the test and reference drugs, we proposed a semi-parametric model for drug concentrations in blood over time when the kinetic of the drugs under study is uncertain or too complicated. The proposed model primarily includes a smooth mean drug concentration-time curve that can be estimated by using restricted cubic splines and an error variable which is distributed to a generalized gamma distribution (Stacy, 1962). To take into account of the effect of subject's covariates on the drug concentration, we also consider a smooth function of covariates in the model which can be estimated by using tensor product regression spline (Eilers and Marx, 2003). A global bioequivalence test for the two drugs is then suggested which is based on the difference between the two drug concentration-time curves. The robustness of the level and power performances of the suggested test, comparative to the one in Chen and Huang (2009), is then investigated in a simulation study. Finally, we illustrate the proposed model and test based on two datasets for bioequivalence study.

Keywords: semi-parametric model; restricted cubic splines; crossover design; bioequivalence test; pharmacokinetic study

References

Chen, Y. I. and Huang, C. S. (2009): Multiple testing for bioequivalence with pharmacokinetic data in 2 x 2 crossover designs. *Statistics in Medicine 28, 2567-2579.*

Eilers, P. H. C. and Marx, B. D. (2003): Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems, 66, 159-174.*

Stacy, E. W. (1962):A generalization of the gamma distribution. *Annals of Mathematical Statistics 33, 1187-1192.*

Functional data analysis to modelling the behaviour of customers

Mónica Clemente[1], Susana San Matías[1] and Teresa León[2]

[1] Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain
mclement@eio.upv.es, *ssanmat@eio.upv.es*

[2] Departamento de Estadística e Investigación Operativa, Universitat de València
Dr. Moliner, 50, 46100 Burjassot, Valencia, Spain
Teresa.Leon@uv.es

Abstract. In most applications, the available information in the context of Customer Relationship Management (CRM) for the analysis of customers consists of historical data. In the last years, the main questions in CRM have been faced up by the application of a great variety of data mining techniques (Ngai et al. (2009)). Our proposal consists of considering historical data about customers as trajectories evolving over time, in such a way that we can use Functional Data Analysis (FDA) tools to analyse their behaviour. FDA concerns the statistical analysis of data which come in the form of continuous functions, usually smooth curves (see Ramsay and Silverman (1997)).

Particularly, our goal is to define a behaviour pattern that could be associated with *churners* or customers leaving the firm. With this aim, we have focused on the segmentation of customers, or more precisely, the establishment of behaviour patterns useful for marketing or business applications. We propose to apply functional cluster analysis (as in Cerioli et al. (2006)) and functional PCA in order to obtain groups of customers with a similar evolution over time, and then to select the churn pattern. We present some numerical results when applying this methodology to a dataset of real customers.

Keywords: Functional Data Analysis, Marketing, Churn analysis, Data mining

References

CERIOLI, A., LAURINI, F. and CORBELLINI, A. (2006): Functional Cluster Analysis of Financial Time Series. In: H.-H. Bock, W. Gaul and M. Vichi (Eds.): *New Developments in Classification and Data Analysis*. Springer, Berlin Heidelberg, 333–341.

NGAI, E.W.T., XIU, L. and CHAU, D.C.K. (2009): Application of data mining techniques in customer relationship management. *Expert Systems with Applications 36 (2), 2592-2602.*

RAMSAY, J.O. and SILVERMAN, B.W. (1997): *Functional Data Analysis*. Springer, New York.

A New Methodology of Gini Coefficient Decomposition and its application to data in China

Xu Cao

School of Statistics,Southwestern University of Economics and Finance
Guanghuacun Street,Number 55,Chengdu,Sichuan,610074,PR China
caosichuan@swufe.edu.cn

Abstract. Gini coefficient is an important index used by economists to measure the disparity of income within one population group.its decomposition becomes more and more demanding for economists to compute Gini coefficient accurately and economically because the collection of data for a large group of population is costly ans time consuming.In this paper,a new methodology of decomposition has derived,in which the interaction terms betweem the subgroups have more significant economic sense than other decomposition methods.In application to Macro-Economic data in China in which case it has been proven that UMVUE does not exist,an invariant Pitman estimator is obtained .Numerical simulation has been done By SAS.The results turn out to be more accurate and revealing.

Keywords: Gini coefficient, decomposition, Pitman estimator

References

Anand,Sudhir,(1983): Inequality and Poverty in malaysia. *Measurement and Decomposition*. New York,Oxford University Press

Atkinson,Anthony B.,(1970): On the Measurement of Inequality. *Journal of Economic Theory*. 244-263.

Bhattacharya,N. and B.Mahalanobis,(1967: Regional Disparities in Household Consumption in India. *Journal of the American Statistical Association*. 62,143-161.

Biewen, martin,(2002): Bootstrap Inference for Inequality,Mobility and poverty Measurement. *Journal of Econometrics*. 108,317-342.

Blackorby,Charles and David Donaldson,(1978): Measures of Relative Equality and Their meaning in Terms of social Welfare. *Journal of Economic Theory*. 18,59-80.

Cowell,Frank A.,(1998): Inequality Decomposition: Three bad Measures. *Bulletin of Economic Research*. 40(4),309-311.

A Widely Linear Estimation Algorithm*

Rosa M. Fernández-Alcalá, Jesús Navarro-Moreno, Juan C. Ruiz-Molina, Javier Moreno-Kayser, and Antonia Oya-Lechuga

Department of Statistics and Operations Research. University of Jaén
Campus Las Lagunillas, Jaén, Spain. {*rmfernan, jnavarro, jcruiz, aoya*}*@ujaen.es*

Abstract. A general discrete-time estimation algorithm is provided for a type of complex-valued signal which is a second-order stationary process. This type of signal is characterized by having a constant mean function and both the covariance and relation functions only depend on the difference of time instants. In Picinbono and Bondon (1997), a detailed study about their advantages in relation to the widely stationary signals can be found.

In contrast to the conventional treatment of this problem, called strictly linear, which considers that the involved signals are proper (null relation functions), the methodology proposed here is based on a widely linear processing approach characterized by considering the information of both the covariance and relation functions. This type of processing has contributed considerable improvements in relation to the conventional ones in the sense of providing a smaller mean square error (see, for example, Picinbono and Bondon (1997), Schreier et al. (2005), Navarro-Moreno (2008) or Navarro-Moreno et al. (2009)).

The obtained solution is valid for all kind of estimators (filter, predictor and smoother) and includes a great variety of estimation problems.

Keywords: Linear Estimation, Widely Linear Processing, Second Order Stationary Processes.

References

PICINBONO, B. and BONDON, P. (1997): Second-Order Statistics of Complex Signals. *IEEE Transactions on Signal Processing 45 (2), 411-420.*

SCHREIER, P.J., SCHARF, L.L. and CLIFFORD, T.M. (2005): Detection and Estimation of Improper Complex Random Signals. *IEEE Transactions on Information Theory 51 (1), 306-312.*

NAVARRO-MORENO, J. (2008): ARMA Prediction of Widely Linear Systems by Using the Innovations Algorithm. *IEEE Transactions on Signal Processing 56 (7), 3061-3068.*

NAVARRO-MORENO, J., ESTUDILLO, M.D., FERNÁNDEZ-ALCALÁ, J.C. and RUIZ-MOLINA (2009): Estimation of Improper Complex-Valued Random Signals in Colored Noise by Using the Hilbert Space Theory. *IEEE Transactions on Information Theory 55 (6), 2859–2867.*

* This work was supported in part by Project MTM2007-66791 of the Plan Nacional de I+D+I, Ministerio de Educación y Ciencia, Spain, which is financed jointly by the FEDER.

Artificial Neural Network Design for Modeling of Mixed Bivariate Responses in Medical Research Data

Sedehi M[1], Mehrabi Y[2], Kazemnejad A[3], Joharimajd V[4], Hadaegh F[5]

1. PhD Student in Biostatistics, Faculty of medicine, Tarbiat Modares University, Tehran, Iran
2. Professor of Biostatistics, Department of Epidemiology, School of Public Health, Shahid Beheshti University of Medical Sciences.
3. Professor of Biostatistics, Faculty of Medicine, Tarbiat Modares University.
4. Associate Professor of Electricity Control Engineering, Faculty of Electricity Engineering, Tarbiat Modares University.
5. Associate Professor of Endocrinology, Prevention of Metabolic Disorders Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences.
Corresponding Author Mehrabi Y. ymehrabi@gmail.com

Abstract

Introduction and objectives: Mixed outcomes arise when, in a multivariate model, response variables measured on different scales such as binary and continuous. We frequently face these types of mixed outcomes in medical research. Artificial neural networks (ANN) can be used for modeling in situations where classic models have restricted application when some (or all) of their assumptions are not met. In this paper, we proposed a method based on ANNs for modeling and predicting mixed binary and continuous outcomes.

Methods: Univariate and bivariate models were evaluated based on two different sets of simulated data. The scaled conjugate gradient (SCG) algorithm was used for optimization. To end the algorithm and finding optimum number of iteration and learning coefficient, mean squared error (MSE) was computed. Predictive accuracy rate criterion was employed for selection of appropriate model at the final stage. We also used our model in real medical data for joint prediction of metabolic syndrome (binary) and HOMA-IR (continues) in Tehran Lipid and Glucose Study (TLGS). The codes were written in R 2.9.0 and MATLAB 7.6.

Results: The predictive accuracy for univariate and bivariate models based on simulated dataset I where two outcomes associated with a common covariate, were shown to be approximately similar. However, in simulated dataset II in which two outcomes associated with different covariates, predictive accuracy in bivariate models were seen to be larger than that of univariate models. In real dataset the results indicated the highest predictive accuracy, 87.37 and 87% in test and validation data, respectively, in model with 10 nodes in hidden layer,

Conclusion: Results indicated that the predictive accuracy gain is higher in bivariate model, when the outcomes share a different set of covariates with higher level of correlation between the outcomes..

Keywords: Mixed Response, Artificial Neural Network, Bivariate Models, TLGS.

Bayesian Analysis on Accelerated Life Tests of a Series System with Masked Interval Data Under Exponential Lifetime Distributions

Tsai-Hung Fan[1] and Tsung-Ming Hsu[2]

[1] National Central University
300 Jhongda Road, Jhongli, Taiwan, *thfanncu@gmail.com*
[2] National Central University
300 Jhongda Road, Jhongli, Taiwan, *owowo99@cycu.org.tw.com*

Abstract. We will discuss the reliability analysis of a series system under accelerated life tests when interval data are observed while the components are assumed to have independent exponential lifetime distributions. In a series system, the system fails if any of the component fails. It is often to include masked data in which the component that causes failure of the system is not observed. Bayesian approach incorporated with subjective prior distribution with the aid of MCMC method is applied to draw statistical inference on the model parameters as well as the system mean life time and the reliability function. Some simulation study and an illustrative example will be presented to show the appropriateness of the proposed method.

Keywords: accelerated life tests; interval data; series system; masked data; MCMC.

References

Kuo, L. and Yang, T. Y. (2000): Bayesian reliability modeling for masked system lifetime data. *Statistics and Probability Letters 17, 229-241.*

Usher, J. Z. (1996): Weibull component reliability-prediction in presence of masked data. *IEEE Transactions on Reliability 45, 229 - 232.*

Zhao, W. and Elsayed, E. A. (2005)): A general accelerated life model for step-stress testing. *IIE Transactions 37, 1059 - 1069.*

Random forests for the optimization of parameters in experimental designs

Susana San Matías, Adriana Villa and Andrés Carrión

Departamento de Estadística e Investigación Operativa Aplicadas y Calidad,
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain
ssanmat@eio.upv.es, *a_villa43@hotmail.com*, *acarrion@eio.upv.es*

Abstract. In many industrial applications, the parameter design optimization problems play an important role but its resolution presents quite a complexity. The objective of Taguchi method is to solve these problems, and it focus on the design of less sensitive products to random factors that make oscillate the parameters that define its quality. This is known as robust design.

On account of Taguchi´s limitations, several authors have presented new alternatives to the parameter design problems, as for example Su and Chang (2000) and Chang (2008). Starting from a Taguchi´s orthogonal design, Su and Chang proposed a two-phase methodology to improving the effectiveness of the optimization of parameter design in the case of a single response variable. Phase 1 determines the objective function using an artificial neural network (ARN) for predicting the value of the response for a given parameter setting. During phase 2, the optimal parameter combination is obtained by a simulated annealing algorithm.

In this work, we have investigated the use of random forests (see Breiman (2001)) to determining the objective function in phase 1. We present comparative numerical results using different data mining techniques, including ARN. Our results show that random forests is the most inexpensive technique (from a computational point of view) and the most stable (from a statistical point of view) in order to determine the response variable.

Keywords: Taguchi design, Random Forests

References

BREIMAN, L. (2001): Random forests. *Machine Learning 45 (1), 5-32.*

CHANG, H.-H. (2008): A data mining approach to dynamic multiple responses in Taguchi experimental design. *Expert Systems with Applications 35 (3), 1095-1103.*

SU, C.-T. and CHANG, H.-H. (2000): Optimization of parameter design: an intelligent approach using neural network and simulated annealing. *International Journal of Systems Science 31 (12), 1543-1549.*

Asymptotic Results in Partially Non-regular Log-Location-Scale Models

Inmaculada Barranco-Chamorro[1], Dolores Jiménez-Gamero[1], Juan L. Moreno-Rebollo[1], and Ana Muñoz-Reyes[1]

Dpt. Estadística e I.O. & Fac. de Matemáticas, Universidad de Sevilla
C/ Tarfia s.n., 41012 Sevilla, Spain, *chamorro@us.es*

Abstract. In this paper we get approximations to the moments, different possibilities for the limiting distributions and approximate confidence intervals for the maximum likelihood estimator (MLE) of a given parametric function when sampling from a partially non-regular log-location-scale model. Our results are applicable to the two-parameter exponential, Power-Function and Pareto distributions. Numerical simulations have been carried out to illustrate the applicability of our results. Specifically, asymptotic confidence intervals for quantiles in Pareto models have been calculated. Our results are compared to other asymptotic approaches available in the literature. The superiority of intervals we propose is also assessed probabilistically.

Keywords: asymptotic, non-regular, log-location-scale models

References

AKAHIRA, M. and TAKEUCHI, K. (1995): *Non-regular Statistical Estimation. Lecture Notes in Statistics 107.* Springer, New York.

BARRANCO-CHAMORRO, I. et al. (2007): An Overview of Asymptotic Properties of Estimators in Truncated Distributions. *Communication in Statistics-Theory and Methods 36:2351-2366.*

DUBININ, T.M. and VARDEMAN S.B. (2003): Likelihood-Based Inference in Some Continuous Exponential Families with Unknown Threshold Parameters. *JASA 98(463), 741-749.*

Bayesian Methods to Overcome the Winner's Curse in Genetic Studies

Radu V. Craiu[1], Lei Sun[2], and Lizhen Xu[3]

[1] Department of Statistics, University of Toronto
100 St George Street, Toronto, Canada *craiu@utstat.utoronto.ca*
[2] Dalla Lana School of Public Health and
Department of Statistics, University of Toronto
155 College Street, Toronto, Canada *sun@utstat.utoronto.ca*
[3] Department of Statistics, University of Toronto
100 St George Street, Toronto, Canada *lizhen@utstat.utoronto.ca*

Abstract. Parameter estimates for associated genetic variants (e.g. Single-Nucleotide Polymorphisms), reported in the discovery samples are often grossly inflated compared to the values observed in the follow-up studies. This type of bias is a consequence of the sequential procedure since a declared associated variant must first pass a stringent significance threshold. This phenomenon is also known as the Beavis effect (Xu, 2003) or the Winner's curse (Zöllner and Pritchard, 2007) in the biostatistics literature. We propose a hierarchical Bayes method in which a spike-and-slab prior is used to account for the possibility that the significant test result may be due to chance. We investigate the robustness of the method using different priors corresponding to different degrees of confidence in the testing results and propose a Bayesian model averaging procedure to combine estimates produced by different models. The Bayesian estimators yield smaller variance compared to the conditional likelihood estimators of Zöllner and Pritchard (2007) or Zhong and Prentice (2008) and outperform the estimators proposed by Ghosh, Zou and Wright (2008) in studies with low power. We investigate the performance of the method with simulations and four real data examples.

Keywords: Bayesian Model Averaging, Hierarchical Bayes Model, Spike-and-Slab Prior, Winner's Curse

References

GHOSH, A., ZOU, F., and WRIGHT, F. A. (2008): Estimating odds ratios in genome scans: An approximate conditional likelihood approach.*American Journal of Human Genetics 82, 1064–1074.*

XU, S. (2003): Theoretical basis of the Beavis effect. *Genetics 165, 2259–2268.*

ZHONG, H. and PRENTICE, R. L. (2008): Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics 9(4), 621–634.*

ZÖLLNER, S. and PRITCHARD, J. (2007): Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *American Journal of Human Genetics 80, 605–615.*

Robust inference in generalized linear models with missing responses

Ana M. Bianco[1], Graciela Boente[2], and Isabel M. Rodrigues[3]

[1] Facultad de Ciencias Exactas and Naturales, Universidad de Buenos Aires and CONICET, Argentina, *abianco@dm.uba.ar*
[2] Facultad de Ciencias Exactas and Naturales, Universidad de Buenos Aires and CONICET, Argentina, *gboente@dm.uba.ar*
[3] Departamento de Matemática and CEMAT, Instituto Superior Técnico, Technical University of Lisbon (TULisbon), Lisboa, Portugal, *irodrig@math.ist.utl.pt*

Abstract. The generalized linear model (GLM) is a popular technique for modelling a wide variety of data and assumes that the observations $(y_i, \mathbf{x}_i)$, $1 \leq i \leq n$, $\mathbf{x}_i \in \mathbb{R}^p$, are independent with the same distribution as $(y, \mathbf{x}) \in \mathbb{R}^{p+1}$ such that the conditional distribution of $y|\mathbf{x}$ belongs to the canonical exponential family. In this situation, the mean $\mu(\mathbf{x}) = \mathbb{E}(y|(\mathbf{x}))$ is modelled linearly through a known link function, g, i.e., $g(\mu(\mathbf{x})) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$. Robust procedures for GLM have been considered among others; by Bianco and Yohai (1996), Cantoni and Ronchetti (2001), Croux and Haesbroeck (2002) and Bianco *et al.* (2005).

In practice, some response variables may be missing by design (as in two-stage studies) or by happenstance. Problems can arise when methods designed for complete data sets are applied with missing responses and covariates completely observed. In this work, we introduce robust procedure to estimate the parameter $\boldsymbol{\beta}$ under a GLM model in order to build test statistics for this parameter when missing data occur in the responses. We derive the asymptotic behaviour of the robust estimators for the regression parameter under the null hypothesis and under contiguous alternatives in order to study the robust Wald test. The influence function of the test functional is also studied. The finite sample properties of the proposed procedure are investigated through a Monte Carlo study where the robust test is also compared with nonrobust and robust alternatives.

Keywords: Generalized Linear Models, Missing Data, Robust Estimation

References

BIANCO, A., GARCÍA BEN, M. and YOHAI, V. (2005): Robust estimation for linear regression with asymmetric errors. *Canad. J. Statist., 33, 511–528.*

BIANCO, A. and YOHAI, V. (1996): Robust estimation in the logistic regression model. *Lecture Notes in Statistics, 109, 17-34.* Springer-Verlag, New York.

CANTONI, E. and RONCHETTI, E. (2001): Robust inference for generalized linear models. *Journal of the American Statistical Association, 96, 1022–1030.*

CROUX, C and HAESBROECH, G. (2003): Implementing the Bianco and Yohai estimator for logistic regression. *Comp. Statist. Data Anal., 44, 273–295.*

An Efficient Bayesian Binary Regression Model Considering Misclassifications

Jacinto Martín, Carlos Javier Pérez, and María Jesús Rufo

Departamento de Matemáticas, Universidad de Extremadura, Avda. Elvas, s/n, 06006 Badajoz, Spain, *jrmartin@unex.es*, *carper@unex.es*, *mrufo@unex.es*

Abstract. When information is gathered in the real world, the response variable is often misclassified. Misclassified data can produce an important impact on inferences because the effective amount of information can be dramatically reduced. Therefore, misclassification errors should be considered in the statistical models.

During the last years, Bayesian categorical data models have had a significant growth, mainly by the development of Markov chain Monte Carlo methods. At present, there are many error-free Bayesian categorical models that can be extended with a relative effort to address misclassifications.

In this work, the interest is focused on a Bayesian probit regression model considering misclassified data. The proposed method is an extension (to address misclassifications) of the error-free probit regression model proposed by Albert and Chib (1993). Holmes and Held (2006) noticed that the proposed Gibbs sampling algorithm produced a rather slow convergence because there is a strong correlation between the regression parameters and the latent variables. Then, they proposed to update both of them jointly through a suitable factorization.

In the proposed model, two types of auxiliary variables are included. One type is related to misclassifications and the other one is based on Albert and Chib's proposal. The introduction of the auxiliary variables related to misclassifications breaks the correlation between the regression parameters and the other type of auxiliary variables, avoiding the need of seeking another factorization as in Holmes and Held (2006). Although the augmented model increases its dimensionality, the generation process becomes easier. A Gibbs-within-Gibbs algorithm has been designed to generate efficiently from the posterior distribution.

The application of the proposed model is illustrated with a real data problem arisen when studying macrovascular complications (coronary artery disease, peripherical arterial disease and stroke) in patients suffering from type 2 diabetes.

Keywords: Bayesian analysis, probit regression model, Gibbs sampling, misclassified data.

References

ALBERT, J.H. and CHIB, S. (1993): Bayesian analysis of Binary and Polychotomous response data. *Journal of the American Statistical Association 88, 669-679.*

HOLMES, C. C. and HELD, L. (2006): Bayesian Auxiliary Variable Models for Binary and Multinomial Regression. *Bayesian Analysis 1 (1), 145-168.*

Understanding co-expression of co-located genes using a PCA approach

Marion Ouedraogo[1], Frédéric Lecerf[1] and Sébastien Lê[2]

[1] UMR598 GARen, INRA & Agrocampus Ouest
65 rue de Saint-Brieuc, 35042 Rennes, France,
marion.ouedraogo@rennes.inra.fr, lecerf@agrocampus-ouest.fr
[2] UMR6625 IRMAR, CNRS & Agrocampus Ouest
65 rue de Saint-Brieuc, 35042 Rennes, France, *sebastien.le@agrocampus-ouest.fr*

Abstract. Studying the genome structure and its role in the gene function regulation could reveal new insights in the regulation of genes expressions by their chromosomal locations . The genome is distributed on several chromosomes where the genes locations could be interpreted as a spatial organization. Therefore, the main hypothesis is that some co-located genes could be involved in a common regulation. (Madan Babu, et al. (2008)). At present, there is very few genome-wide methods to identify pairs of genes or genomic region with co-expressed genes (Pehkonen, et al. (2010)). We propose here a novel approach to identify such clusters at that scale, to assess the role of the genome structure on the regulation of gene expression.

We introduce the so called "autovariogram", in reference to the autocorrelogram used in time series and the variogram used in spatial analysis, for understanding the seasonal and spatial dependencies respectively. This autovariogram is obtained by means of a sequence of Principal Component Analysis (PCA) performed on sets of consecutive genes (*ie* "co-located" genes). It displays graphically the variance of the first principal component for consecutive sets of consecutives genes.This variance may be interpreted as co-expression.

This graphical representation is enhanced by p-values that corresponds each to the following test for a given set of genes.
H_0: The genes are not co-expressed.
H_1: The genes are co-expressed.
The aim of this talk is to illustrate the interest of the autovariogram and to presents some of its main features.

Keywords: genome-wide analysis, genome organization, gene expression, clustering under constraints

References

MADAN BABU M., CHANDRA JANGA S., DE SANTIAGO I., POMBO A., (2008): Eukaryotic gene regulation in three dimensions and its impact on genome evolution. *Current Opinion in Genetics & Development 18,571-582.*

PEHKONEN P., WONG G., TORONEN P., (2010): Heuristic Bayesian Segmentation for Discovery of Coexpressed Genes within Genomic Regions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 7(1), 37-49.*

Assessing Neural Activity Related to Decision-Making through Flexible Odds Ratio Curves and their Derivatives

Javier Roca-Pardiñas[1], Carmen Cadarso-Suárez[2], Jose L. Pardo-Vazquez[3], Victor Leboran[3], Geert Molenberghs[4], Christel Faes[4], and Carlos Acuña[3]

[1] Dept. of Statistics and Operations Research. University of Vigo, Spain, *roca@uvigo.es* .
[2] Unit of Biostatistics, Dept. of Statistics and Operations Research. University of Santiago, Spain.
[3] Department of Physiology and Complejo Hospitalario Universitario, University of Santiago, Spain.
[4] International Institute for Biostatistics and Statistical Bioinformatics, Universiteit Hasselt, Belgium.

Abstract. It is well established that neural activity is stochastically modulated over time. Therefore, direct comparisons across experimental conditions and determination of change points or maximum firing rates are not straightforward. This study sought to compare temporal firing probability curves that may vary across groups defined by different experimental conditions. Odds ratio (OR) curves were used as a measure of comparison, and the main goal was to detect significance features of such curves through the study of their derivatives. An algorithm is proposed that enables ORs based on generalised additive models, including factor-by-curve type interactions to be flexibly estimated. Bootstrap methods were used to draw inferences from the derivatives curves, and binning techniques were applied to speed up computation in the estimation and testing processes. A simulation study was conducted to assess the validity of these bootstrap-based tests. This methodology was applied to study premotor ventral cortex neural activity associated with decision-making. The proposed statistical procedures proved very useful in revealing the neural activity correlates of decision-making in a visual discrimination task.

Keywords: bootstrap, derivatives, generalised additive models, interactions, neural activity.

References

NÁCHER, V., OJEDA, S., CADARSO-SUÁREZ, C., ROCA-PARDIÑAS, J. and ACUÑA, C. (2006): Neural correlates of memory retrieval in the prefrontal cortex. *European Journal of Neuroscience 24, 925-936.*

ROCA-PARDIÑAS, J., CADARSO-SUÁREZ, C., NÁCHER, V. and ACUÑA, C. (2006): Bootstrap-based methods for testing factor-by-curve interactions in generalized additive models: assessing prefrontal cortex neural activity related to decision-making. *Statistics in Medicine 25, 2483-501.*

Utilization of Bayesian Networks Together with Association Analysis in Knowledge Discovery

Derya Ersel[1], Yasemin Kayhan[2], and Suleyman Gunay[3]

[1] Hacettepe University, Department of Statistics, 06800, Ankara, Turkey, *dtektas@hacettepe.edu.tr*
[2] Hacettepe University, Department of Statistics, 06800, Ankara, Turkey, *ykayhan@hacettepe.edu.tr*
[3] Hacettepe University, Department of Statistics, 06800, Ankara, Turkey, *sgunay@hacettepe.edu.tr*

Abstract. Bayesian Networks are probabilistic graphical models that encode probabilistic relationships among a set of random variables in a database. Since they have both causal and probabilistic aspects, data information and experts knowledge can easily be combined by them. Bayesian Networks can also represent knowledge about uncertain domain and make strong inferences. Association analysis is a useful technique to detect hidden associations and useful rules in large databases, and it extracts previously unknown and surprising patterns from already known information. Association analysis algorithms usually generates many patterns. Hence, suitable interestingnes measures must be performed to eliminate uninteresting patterns. Bayesian Networks can be used to define subjective interestingness measures. In this study, utilization of Bayesian Netwoks together with association analysis in knowledge discovery will be presented. As association rules can be used to create a Bayesian Network, subjective interestingness measures to determine interesting patterns can be established by Bayesian Networks.

Keywords: bayesian networks, association analysis, knowledge discovery

References

DONG-PENG, Y. and JIN-LIN, L. (2008): Research on personal credit evaluation model based on bayesian network and association rules. *Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on , 457-460.*

JAROSZEWICZ, S. and SIMOVICI, D. A. (2004): Interestingness of frequent itemsets using bayesian networks as background knowledge. *In Proceedings of the SIGKDD Conference on Knowledge Discovery and Data Mining, 178-186.*

JENSEN, F. V. (2001): *Bayesian Networks and Decision Graphs.* Springer-Verlag, New York.

SILBERSCHATZ, A. and TUZHILIN, A. (1996): What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering 8(6), 970-974.*

LMS As a Robust Method for Outlier Detection in Multiple Linear Regression Models with No Intercept

Yasemin Kayhan[1] and Suleyman Gunay[2]

[1] Hacettepe University, Department of Statistics, 06800, Ankara, Turkey, *ykayhan@hacettepe.edu.tr*
[2] Hacettepe University, Department of Statistics, 06800, Ankara, Turkey, *sgunay@hacettepe.edu.tr*

Abstract. In practical aplications, many data sets contain outliers that do not go along with the majority of the data. So investigating the residuals to mark these outliers is one of the fundamental task of applied regression analysis. It is well know fact that Least Squares / LS regression is very sensitive to the outliers that is why this method is not sufficient to examine the residuals to make decision. On the contrary Least Median of Squares / LMS, as an high-breakdown estimator, is not influenced outliers like LS. However when the regression model has no intercept LMS fit evaluated via the PROGRESS can fail (Kayhan and Gunay, 2008). So at this study by using another algorithm which can find the exact LMS solution when the muliple regression model through the origin is presented to analyse the residuals. Also the results obtained from LMS with PROGRESS and LMS with new algorithm will be compared.

Keywords: LMS, PROGRESS, robust regression, residuals, outliers detection

References

CASELLA, G. (1983): Leverage and Regression Through the Origin. *The American Statistician 37(2).*

HUBERT, M., ROUSSEEUW, P. J. and AELST, S. V. (2008): High-Breakdown Robust Multivariate Methods. *Statistical Science 23(1), 92-119.*

KAYHAN, Y. and GUNAY, S. (2008): A new approach to Least Median of Squares and Regression Through the Origin. *Communications in Statistics Theory and Methods 37(5).*

LEROY, A. and ROUSSEEUW, P. (1985): Computing Robust Regression Estimation with 'PROGRES' and Some Simulation Result. *Statistics and Decisions, Supplement Issue (2), 321-325.*

VERBOVEN, S. and HUBERT, M. (2005): LIBRA: a MATLAB Library for Robust Analysis. *http://wis.kuleuven.be/stat/robust/LIBRA.html.*

ROUSSEEUW, P. J. and VAN ZOMEREN, B. C. (1990): Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association, Theory and Methods 85(111).*

Weighted Kaplan-Meier test when the population marks are missing

Dipankar Bandyopadhyay[1] and M. Amalia Jácome[2]

[1] Department of Biostatistics, Bioinformatics and Epidemiology
Medical University of South Carolina, USA, *bandyopd@musc.edu*
[2] Facultade de Ciencias, Campus da Zapateira
Universidade da Coruña, A Coruña, Spain, *majacome@udc.es*

Abstract. The problem of comparing survival distributions, very often in biomedical research, it is typically accomplished using the Tarone-Ware family. It encompasses the well-known log-rank test introduced by Mantel (1966), optimal when the hazard rates are proportional to each other. However, it has very low power for some alternative hypothesis. The weighted KaplanMeier (KM) statistic, introduced by Pepe and Fleming (1989), is an important special case of the more general statistics for testing the equality of two survival functions proposed by Gu et al. (1999) as an alternative to rank-based methods. When the hazards are proportional, the power is comparable but slightly less than that of the log-rank test, and substantially higher when the two hazard rates cross. These tests assume that every individual is clearly identified to belong to any of the groups to be compared.

In a similar spirit of Pepe and Fleming (1989), we develop a new test for testing equality of (conditional) survival functions S_j given that all the subjects have failed due to one of the causes (groups) of failure. Clearly, the population membership of the censored subjects are unknown due to right censoring. The idea is to estimate the survival functions S_j using fractional risk sets (FRS) (Bandyopahyay and Datta (2008), Bandyopadhyay and Jácome (2010)), since the classical Kaplan-Meier estimator can not be applied in this context. We study the properties of this test statistic using simulation studies and illustrate its application to a real data.

Keywords: Fractional risk sets, Right censoring

References

BANDYOPADHYAY, D. and JÁCOME, M.A. (2010): Nonparametric estimation of conditional cumulative hazards for missing population marks. *Australian and New Zealand Journal of Statistics 52 (1), 75-91.*

BANDYOPADHYAY, D. and DATTA, S. (2008): Testing equality of survival distributions when the population marks are missing. *Journal of Statistical Planning and Inference 138, 1722-1732.*

GU, M., FOLLMANN, D. and GELLER, N.L. (1999): Monitoring a general class of two-sample survival statistics with applications. *Biometrika 86, 45-57.*

MANTEL, N. (1966): Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports 50, 163-170.*

PEPE, M.S. and FLEMING, T.R. (1989): Weighted Kaplan-Meier statistics: A class of distance tests for censored survival data. *Biometrics 45, 497-507.*

A linear-time inverse of a Covariance Matrix of a special structure

Sarada Velagapudi

New Jersey Institute Of Technology
University Heights, Newark, NJ 07102 *sv88@njit.edu*

Abstract. A statistical model for the problem of one-dimensional Global minimization of an objective function on a continuous interval has been studied in [1]. Data observed is assumed to be corrupted due to presence of noise and the function is modelled as a standard Wiener process with independent Gaussian noise both for adaptive and nonadaptive strategies. Algorothmic implementation of it introduces an $O(n^2)$ technique for inversion of a covariance matrix of a special structure arising in this context. In this presentation, an O(n) technique is introduced to solve this inversion for adaptive strategy. It is based on 3n entries of the inverse of nth step to obtain inverse of (n+1)th step. O(n) operations of multiplications, additions and storage are required for this technique.

Keywords: Global Optimization, Covariance Matrix, Inverse

References

J.M.Calvin and A.Zilinskas(2005): One-Dimensional Global Optimization for Observations with Noise *Computers and Mathematics with Applications 50 (2005) 157-169.*

Development of a Web-based integrated platform for test analysis

Takekatsu Hiramura[1], Tomoya Okubo[2], and Shin-ichi Mayekawa[1]

[1] Graduate School of Decision Science and Technology, Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8550 Japan
t.hiramura@ms.hum.titech.ac.jp, *mayekawa@hum.titech.ac.jp*
[2] The National Center for University Entrance Examinations
2-19-23 Komaba, Meguro-ku, Tokyo 153-8501 Japan *okubo@rd.dnc.ac.jp*

Abstract. In this paper, we discuss the development of an integrated platform to aid researchers and testing agencies in conducting tests analysis. We focus on a system that allows users to analyse test data using multiple item response models conveniently. The system can not only analyse test items and evaluate respondents but can also be used to maintain the item bank. This system allows test data analysts to conduct research within the framework of item response theory.

The system uses one, two and three-parameter logistic models for some response models: the graded response model (Samejima (1969)), the partial credit model (Masters (1982)), the generalized partial credit model (Muraki (1992)), and the order-constrained nominal categories model (Okubo et al. (2009)) for items in a rank-order scale, and the nominal categories model (Bock (1972)) for items in a nominal scale. Further, the system estimates the item parameters of tests containing mixed items.

The system is a Web application based on the client-server model, and it can be used through usual Web browser. Therefore, the statistic estimation is performed on the server, and is not dependent on the client's computer performance. The system will be made available on the Web.

Keywords: Item Response Theory, Educational Measurement, Item Bank, Web Application

References

BOCK, R. D. (1972): Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37: 29-51*

MASTERS, G. N. (1982): A Rasch model for partial credit scoring. *Psychometrika, 47, 149-174*

MURAKI, E. (1992): A generalized partial credit model: Application of an EM algorithm. *Applied Psychologocal Measurement, 16: 159-176*

OKUBO, T., HOSHINO, T. and MAYEKAWA, S. (2009): Partially Ordered Nominal Categories Model, *IMPS 2009*

SAMEJIMA, F. (1969): Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, No.17*

Identifying risk factors for complications after ERCP

Christine Duller

Institute for Applied Statistics
Johannes Kepler University Linz
Altenberger Strasse 69, 4040 Linz, Austria
christine.duller@jku.at

Abstract. Endoscopic retrograde cholangiopancreatography (ERCP) is an important procedure for investigation and management of pancreatic and bile ducts. Quality assessment in gastrointestinal endoscopy aims to improve medical quality in challenging endoscopic procedures and provides patients with the best medical care.

ERCP entails high risk of pancreatitis, cholangitis, perforation and bleeding. Therefore the Austrian Society of Gastroenterology and Hepatology (OeGGH) initiated a nation-wide project for voluntary benchmarking of ERCP in 2006, which is still going on. Success and complication rates for non-selected patients were evaluated especially with respect to centre size and endoscopist-individual case volume.

In this poster some results for identification of risk factors are presented. Twenty-nine sites participated in the project and reported about 5000 cases of ERCP. The data from the participating sites as well as patient data were transmitted pseudonymously via an online questionnaire, the endoscopists remained anonymous. For each ERCP 70 variables, including patient characteristics and variables related to the ERCP procedure were collected. Complications occured in 9,9% of the patients, the most frequent complications being post-ERCP pancreatitis (4,8%), bleeding (3,7%), cholangitis (1,6%) and perforation (0,6%).

The statistical focus of interest are logit models for various risks. The fitting of the models will be done with classical methods implemented in different software surroundings (R, SAS, SPSS). Beside some results the pros and cons of common software packages will be considered.

Keywords: Logistic Regression, R, SAS, SPSS

References

FREEMAN, M. L. (2003): Adverse outcomes of endoscopic retrograde cholangiopancreatography: avoidance and management. *Gastrointestinal Endoscopy Clinics of North America 13 (4), 775-798*

HILBE, J. M. (2009): *Logistic Regression Models*. Chapman & Hall, Boca Raton Fl.

KAPRAL, C., DULLER, C., WEWALKA, F., KERSTAN, E., VOGEL, W. and SCHREIBER, F. (2008): Case volume and outcome of endoscopic retrograde cholangiopancreatography: results of a nationwide Austrian benchmarking project. *Endoscopy, 40(8), 625-630*

Differences in wages for atypical contracts and stable jobs in Italy: A Multilevel Approach for Longitudinal Data

Valentina Tortolini[1] and Davide De March[2]

[1] Dipartimento di Statistica, Università di Firenze
Viale Morgagni 59, Florence, Italy *tortolini@ds.unifi.it*
[2] Dipartimento di Statistica, Università di Firenze
Viale Morgagni 59, Florence, Italy *demarch@ds.unifi.it*

Abstract. In the last years, Italian labor market, as almost all European countries, had to deal with a even growing unemployment rate. As a response to this problem, Governments introduced several labor market reforms; in Italy one important instrument to solve this problem was the introduction of atypical employment contracts forms to get labor market flexibility (Leonbruni (2008)). The aim of this paper is to investigate the evolution of wages (output) in Italy, focusing on the difference between typical and atypical contract forms. We use the 2000-2004 WHIP (Work History Italian Panel) database of individual work histories, based on Inps administrative archive. The multilevel analysis for longitudinal data allows us to relate the output variability to other individual characteristics as gender, geographic areas, skills, age. We fit a Random coefficient model(Rabe-Hesketh and Skrondal (2008)), for the wages. We highlight a strong positive relationship between incomes and atypical contracts, even if some individual characteristics and time-varying variables have a different influence on the rate of increase of the subjects' wages.

Keywords: Labor Market Analysis, Multilevel Models, Longitudinal Data, Random Coefficient Model

References

LEONBRUNI, R. (2008): Le carriere dei lavoratori dipendenti prima e dopo le riforme: un'esplorazione su WHIP. *Working paper in Laboraorio R. Revelli* LABOR, Torino.

RABE-HESKETH, S. and SKRONDAL, A. (2008): *Multilevel and Longitudinal Modeling Using Stata*. STATA Press, Lakeway Drive, Texas.

DOREIAN, P., BATAGELJ, V. and FERLIGOJ, A. (2000): Symmetric-acyclic decompositions of networks. *Journal of Classification 17 (1), 3-28.*

BAT - The Bayesian Analysis Toolkit

Frederik Beaujean[1], Allen Caldwell[1], Daniel Kollár[2], and Kevin Kröninger[3]

[1] Max-Planck-Institute for Physics, Munich, Germany
[2] CERN, Geneva, Switzerland
[3] University of Göttingen, Göttingen, Germany

Abstract. The main goals of data analysis are to infer the free parameters of models from data, to draw conclusions on the models' validity, and to compare their predictions allowing to select the most appropriate model.

The Bayesian Analysis Toolkit, BAT, is a tool developed to evaluate the posterior probability distribution for models and their parameters. It is centered around Bayes' Theorem and is realized with the use of Markov Chain Monte Carlo giving access to the full posterior probability distribution. This enables straightforward parameter estimation, limit setting and uncertainty propagation. Additional algorithms, such as Simulated Annealing, allow to evaluate the global mode of the posterior.

BAT is implemented in C++ and allows for a flexible definition of models. It is interfaced to software packages commonly used in high-energy physics: ROOT, Minuit, RooStats and CUBA. A set of predefined models exists to cover standard statistical cases.

An overview on the software will be presented and as well as the algorithms implemented. A set of physics examples will show the spectrum of applications of BAT. New features and recent developments will be summarized.

Keywords: Data analysis, Bayesian inference, Markov Chain Monte Carlo, C++ library

References

CALDWELL, A., KOLLAR, D., and KROENINGER, K. (2009): BAT - The Bayesian Analysis Toolkit. *Computer Physics Communications 180 (11), 2197-2209.*

Deriving a euro area monthly indicator of employment: a real time comparison of alternative modelbased approaches

Filippo Moauro[1]

Eurostat,
11, rue A. Wecker, L-2721, Luxembourg, *filippo.moauro@ec.europa.eu*

Abstract. The paper presents the results of an extensive real time analysis of alternative model-based approaches to derive a monthly indicator of employment for the euro area. In the experiment the Eurostat quarterly national accounts series of employment is temporally disaggregated using the information coming from the most significant related monthly indicators, among which unemployment and labour input indexes. The strategy benefits of the contribution of the information set of the euro area and its 6 larger member states, as well as the split into the 6 sections of economic activity. The models under comparison include univariate regressions of the Chow and Lin' type where the euro area aggregate is directly and indirectly derived, as well as multivariate structural time series models of small and medium size. The specification in logarithms is also systematically assessed. The largest multivariate setups, up to 58 series, are estimated through the EM algorithm. Main conclusions are the following: mean revision errors of disaggregated estimates of employment are overall small; a gain is obtained when the model strategy takes into account the information by both sector and member state; the largest multivariate setups outperforms those of small size and the strategies based on classical disaggregation methods.

Keywords: temporal disaggregation methods, multivariate structural time series models, mixed-frequency models, EM algorithm, Kalman filter and smoother

References

CHOW, G. and LIN, A. C. (1971): Best linear unbised interpolation, distribution and extrapolation of time series by related series. *The Review of Economics and Statistics 53 (4), 372-375.*

MOAURO, F. and SAVIO, G. (2005): Temporal disaggregation using multivariate structural time series models. *The Econometrics Journal 8, 214-234.*

PROIETTI, T. and MOAURO, F. (2006): Dynamic factor analysis with non-linear temporal aggregation constraints. *Applied Statistics 55 (2), 281-300.*

KOOPMAN, S. J. (1993): Disturbance smoother for state space models. *Biometrika 80, 117-126.*

Enhancing spatial maps by combining difference and equivalence test results

Harald Heinzl[1] and Thomas Waldhoer[2]

[1] Center for Medical Statistics, Informatics, and Intelligent Systems
Medical University of Vienna, Austria, *harald.heinzl@meduniwien.ac.at*
[2] Department of Epidemiology, Center for Public Health
Medical University of Vienna, Austria, *thomas.waldhoer@meduniwien.ac.at*

Abstract. It is common practice in spatial epidemiology that regionally partitioned health indicator values are presented in choropleth maps. State and local health authorities use them among others for health reporting, demand planning, and quality assessment.

Quite often there are concerns whether the health situation in certain areas can be considered different or equivalent to a reference value. The common approach of solely reporting the result of difference tests may intuitively lead to the false impression that spatial units showing non-significant results are close to the reference value.

We suggest a combined graphical representation of statistical difference and equivalence tests in choropleth maps in order to overcome this weakness. We will exemplify with health data of Austrian newborns that integrating both difference and equivalence tests in choropleth maps provides more insight into the spatial distribution than sole difference tests.

Keywords: confidence interval, two one-sided tests, choropleth map, spatial epidemiology

Some measures of multivariate association relating two spectral data sets

Carles M. Cuadras[1] and Silvia Valero[2]

[1] Departament d'Estadistica, Universitat de Barcelona
Diagonal 645, Barcelona, Spain, *ccuadras@ub.edu*
[2] GIPSA-lab, Signal & Image Dept., Grenoble Institute of Technology
Grenoble, France *silvia.valero-valbuena@gipsa-lab.grenoble-inp.fr*

Abstract. We study some measures of association between two data sets to construct hierarchical region-based image representations. A hyperspectral image is a data cube of P spectral bands, each one belonging to a specific wavelength. Each pixel represents the radiance spectrum of the measured material across the spectral range defined by the P adjacent wavelengths. In practice, we have a set M of n spectra belonging to the same material. M can be modelled by P probability density functions, where each density represents the probability of having a specific radiance value in the corresponding wavelength. Any measure of association between two data sets M_1 andM_2 should consider all densities. However, there exists a high correlation between adjacent densities, since they come from contiguous wavelengths. Hence, any measure of association should take into account such redundancies. We propose several distances between densities for each data set, and study some distance-based measures depending on canonical correlations relating principal coordinates. These measures include Wilks, Hotelling, Pillai, Cramer-Nicewander and other measures of multivariate association. We also study measures based on Mahalanobis distances between multinomial distributions.

Keywords: image representations, probability related to wavelenghts, distance-based association

References

CRAMER, E. M. and NICEWANDER, W. A. (1979): Some symmetric, invariant measures of multivariate association. *Psychometrika 44, 43-54.*

CUADRAS, C. M. (2008): Distance-based multisample tests for multivariate data. In: B. C., Arnold, N. Balakrishnan, J. M. Sarabia, R. Mínguez (Eds.): *Advances in Mathematical and Statistical Modeling*. Birkhauser, Boston, 61–71.

VALERO, S. (2010): *Arbre binaire de partitions pour limagerie hyperspectrale*. Working PhD Thesis (supervised by J. Chanussot and P. Salembier), Gipsa-lab, Grenoble.

The Influence of Exchange Rate on the Volume of Japanese Manufacturing Export

Hitomi Okamura[1], Yumi Asahi[2], and Toshikazu Yamaguchi[3]

[1] Graduate School of Indutrial Management, Tokyo University of Sience
1-3 kagurazaka, Shinjuku, Tokyo, Japan, *hito_oka@ms.kagu.tus.ac.jp*
[2] Industrial Management, Tokyo University of Sience
1-3 kagurazaka, Shinjuku, Tokyo, Japan, *asahi@ms.kagu.tus.ac.jp*
[3] Industrial Management, Tokyo University of Sience
1-3 kagurazaka, Shinjuku, Tokyo, Japan, *yama@ms.kagu.tus.ac.jp*

Abstract. This paper investigates the influence of exchange rate volatility on the volume of Japanese manufacturing export. The volatility in yen is showed by conditional variance from EGARCH model, allowing for asymmetric effects that a shock of an appreciation of the yen is different from that of a depreciation of the yen. The export action model including exchange rate volatility is constructed based on VAR model to examine the relationship between exchange rate uncertainty and the volume of export. Tests are performed for typical eight kinds of industry in Japan. Few empirical studies focus on each Japanese industry export. Results indicate significant negative effects of exchange rate volatility on most manufacturing exports. In addition, this paper characterizes Japanese manufacturing industry by the influence of exchange rate. We find equipment industries, occupying 60% or more of total Japanese exports, especially tend to be received negative influence of exchange.

Keywords: time series analysis, EGARCH model, manufacturing export, exchange rate volatility

References

Caporale,T. and Doroodian, K.(1994): Exchange Rate Variability and the Flow of International Trade. *Economics Letters 46, 49-54.*

Choudhry, T. (2008) : Exchange rate volatility and United Kingdom trade: evidence from Canada, Japan and New Zealand. *Empirical Economics 35, 60 7-619.*

Kimura, T. and Nakayama, K.(2000): Foreign exchange volatility and firms export activities (Kawasereito no boratiriti to kigyou no yushutsu koudou). *Japan Monthly Bulletin, March, 83-109.*

Koray, F. and W. D. Lastrapes (1989) : Real Exchange Rate Volatility and U.S. Bilateral Trade A VAR Approach. *The Review of Economics and Statistics 71, 708-712.*

Pozo, S. (1992) : Conditional Exchange-Rate Volatility and the Volume of International Trade : Evidence from Early 1900's. *The Review of Economics and Statistics 74, 325-329.*

Descriptive patterns for multivariate time series based on KPCA-Biplot. A comparison between classical PCA and kernel PCA

Toni Monleón-Getino[1], Esteban Vegas[1], Ferran Reverter[1], and Martín Ríos[1]

[1] Departament of Statistics. University of Barcelona
Avda. Diagonal 645, Barcelona, E-08028 Spain, *amonleong@ub.edu, evegas@ub.edu, freverter@ub.edu, mrios@ub.edu*

Abstract. We present a graphical-exploratory method, called KPCA-Biplot (Reverter and Vegas, 2009) that combines Singular Value Decomposition (SVD)-Biplot and Kernel PCA (KPCA). We compare KPCA-Biplot with PCA for their ability to elucidate relationships between samples and variables as a method to describe patters for multivariate time series in epidemiology. The main properties of KPCA-Biplot are the extraction of nonlinear features, together with the preservation of the input variables in the output display: 1) Perform SVD of the preprocessed data input matrix $\mathbf{X} = \mathbf{GH}'$. 2) Take the rows of $\mathbf{H}$ as a set of observations and compute the corresponding kernel matrix K. 3) Compute KPCA with the kernel matrix $\mathbf{K}$. We can use it to extract the nonlinear features of the data. 4) Project the rows of $\mathbf{G}$ onto the subspace expanded by the leading eigenvectors of $\mathbf{K}$.
We compare the results obtained using KPCA and classical PCA in representing the structure of a multivariate time series, by using a previous analysis. This was a study of tuberculosis incidence (reported cases of TB per 10^5 inhabitants) and trends in the WHOs European region (Ríos and Monleón, 2009), where a graphical output was obtained using classical PCA techniques. Differences in the overall incidence and trends were identified during the 1980–2006 period using KPCA and PCA. The lowest rates were reported in the eastern Mediterranean, Scandinavia and Iceland. As regards development of tuberculosis in Europe, 1992 was a turning point, when the decreasing trend observed since 1980 reached a minimum and began to increase. Our results indicate that KPCA-Biplot is complementary to the classical PCA currently used to describe patterns for multivariate time series.

Keywords: PCA, kernel PCA, multivariate, time series, epidemiology

References

RIOS, M. and MONLEON-GETINO, T. (2009): A graphical study of tuberculosis incidence and trends in the WHOs European region (1980-2006). *European Journal of Epidemiology 24, 381-387.*

REVERTER, F. and VEGAS, E. (2009): A kernel PCA Biplot method applied to gene expression data). *Proceedings of the 17th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 8th European Conference on Computational Biology (ECCB). Stockolm (Sweeden).*

Nonparametric variance function estimation with correlated errors and missing response

Pérez- González, A.[1], Vilar Fernández, J.M.[2] and González-Manteiga, W.[3]

[1] Department of Statistics and Operations Research, University of Vigo
Ourense, Spain, *anapg@uvigo.es*
[2] Department of Mathematics, University of A Coruña
La Coruña, Spain, *eijvilar@udc.es*
[3] Department of Statistics and Operations Research, University of Santiago de Compostela
Santiago de Compostela, Spain *wenceslao.gonzalez@usc.es*

Abstract. In this work we consider a fixed design regression model where the errors follow a strictly stationary process. In this model the conditional mean function and the conditional variance function are unknown curves. The errors are correlated and the response variable has missing observations. In this context we study four nonparametric estimators of the conditional variance function based on local polynomial fitting. Our estimators are based on the estimators for dependent data of Härdle and Tsybakov(1997) and Fan and Yao (1998).

Keywords: Variance Function, Missing response, Correlated errors

References

FAN, J. AND YAO, Q.(1998): Efficient estimation of conditional variance functions in stochastic regression. *Biometrika 85 645-660.*

HÄRDLE, W. and TSYBAKOV, A. (1997):Local polynomial estimators of the volatility function in nonparametric autoregression. *J. Econometrics 81 223-242.*

Widely Linear Simulation Of Complex Random Signals *

Antonia Oya[1], Jesús Navarro-Moreno[1], Juan C. Ruiz-Molina[1], Dirk Blömker[2], and Rosa M. Fernández-Alcalá[1]

[1] Department of Statistics and Operations Research, University of Jaén, Spain
{*aoya, jnavarro, jcruiz, rmfernan*}*@ujaen.es*
[2] Institut für Mathematik, Universität Augsburg
dirk.bloemker@math.uni-augsburg.de

Abstract. The widely linear processing approach, based on a complete second-order description of complex random signals in which both the covariance and the complementary functions are taken into consideration (Picinbono and Bondon (1997)), has been shown to yield significant improvements in most areas of statistical signal processing, in particular, in the classical detection and estimation problems (Schreier et al. (2005), Oya et al. (2009)). In this paper we have applied the widely linear perspective to the simulation of complex-valued random signals. Specifically, a unified and practical methodology for generating second-order complex random processes defined on any interval of the real line has been derived, under the single hypothesis that the correlation matrix of the augmented signal is known. This technique uses an improper version of the Karhunen-Loéve (KL) expansion for complex signals (Schreier et al. (2005)). In contrast with the classical KL expansion of a complex random signal in which the resultant random variables are complex and, in general, correlated that makes it difficult to simulate the complementary function, the main advantage of the improper KL representation is that the associated random variables are real avoiding those difficulties. Furthermore, the assessment of the algorithm performance is illustrated by means of some numerical examples.

Keywords: improper complex random processes, simulation, widely linear processing

References

OYA, A. NAVARRO-MORENO, J. and RUIZ-MOLINA, J.C. (2009): A numerical solution for multichannel detection. *IEEE Trans. On Communications 57(6), 1734-1742.*

PICINBONO, B. and BONDON, P. (1997): Second-order statistics of complex signals. *IEEE Trans. Signal Processing 45(2), 411-420.*

SCHREIER, P.J., SCHARF, L.L. and MULLIS, C.T. (2005): Detection and estimation of improper complex random signals. *IEEE Trans. Inform. Theory 51(1), 306-312.*

* This work was supported in part by Project MTM2007-66791 of the Plan Nacional de I+D+I, Ministerio de Educación y Ciencia, Spain, which is financed jointly by the FEDER.

Model checks for nonparametric regression with missing response: a simulation study.

González-Manteiga, W.[1], Cotos-Yáñez, T. R.[2] and Pérez- González, A.[2]

[1] Department of Statistics and Operations Research, University of Santiago de Compostela
Santiago de Compostela, Spain, *wenceslao.gonzalez@usc.es*
[2] Department of Statistics and Operations Research, University of Vigo
Ourense, Spain, *cotos@uvigo.es, anapg@uvigo.es*

Abstract. In the context of nonparametric regression there are several ways to check the model. We can distinguish mainly two types: tests based on the estimation of the regression function (Härdle and Mammen (1993), González-Manteiga and Cao (1993)) and tests based on the estimation of the integrated regression function (Stute (1997)). In the context of nonparametric regression with missing response, the goodness of fit test for linear regression model has been studied by González-Manteiga and Pérez-González (2006). They considered the statistics based on the L^2 distance between nonparametric estimator of the regression function and a root-n consistent estimator of the regression model under the linear model. Our objective is this paper is to study the behavior of the goodness of fit test for regression model using the test statistics based on empirical processes of the estimated integrated regression function when there are missing observations in the response variable.

Different test can be build using or simplified statistics (only with complete observations) or making imputation. We will compare our methodology based on empirical process with the test based on L^2 distance given in González-Manteiga and Pérez-González (2006).

Keywords: Empirical Process, Goodness of fit tests, Missing response

References

GONZÁLEZ-MANTEIGA, W. and CAO R. (1993): Testing the hypothesis of a general linear model using nonpara- metric regression estimation. *Test, 2, 161-188.*

GONZÁLEZ-MANTEIGA, W. and PÉREZ-GONZÁLEZ, A. (2006): Goodness-of-fit tests for linear regression models with missing response data. *The Canadian Journal of Statistics, 34, 1, 149-170.*

HÄRDLE, W. and MAMMEN E. (1993): Comparing nonparametric versus parametric regression fits. *The Annals of Statistics, 21, 1926-1947.*

STUTE, W. (1997): Nonparametric model checks for regression. *The Annals of Statistics, 25, 613-641.*

Automatic Categorization of Job Postings

Julie Séguéla[1,2] and Gilbert Saporta[2]

[1] Multiposting.fr
33 rue Réaumur, Paris, France, *jseguela@multiposting.fr*
[2] Laboratoire CEDRIC, CNAM
292 rue Saint Martin, Paris, France

Abstract. Since the beginning of the Nineties, the increasing proportion of job vacancies which are published on the internet has led to a multiplication of on-line job search sites (job boards). Consequently, the need to assess job board performance has become a priority for recruiters. But an important issue is that each job board has a specific nomenclature to describe the type of the post. As part of the modelisation of job posting performance, we need to establish a common classification for the "function" criterion. To achieve that goal, we are working on a corpus of manually labelled texts of job offers, and we are proposing a method to categorize the texts into a two-level predefined classification of occupations.

First, a preprocessing adapted to the particularities of job offer texts is performed (stemming, use of a specific dictionnary,...). Then, we are reducing the dimensionality of the problem thanks to a feature selection method (we can see a comparative study in Yang and Pedersen (1997)). The Vector Space Model is used to represent the texts and the terms are weighted with a function depending on the position of the term in the text (title or mission description). Finally, classification of job postings is performed with SVM (e.g. Joachims (1997)). Popular performance measures such as recall and precision are used and adapted to our context with a weighting for errors according to the seriousness of misclassification. In addition, we are exploring the effects on the classification quality of the Probabilistic Latent Semantic Analysis, another dimensionality reduction method which allows to address the issue of synonymy (Hofmann (1999)).

Our method could also be applied to achieve automatic labelling of job postings according to the nomenclature of a particular job board.

Keywords: text categorization, Latent Semantic Analysis, Support Vector Machine, job posting

References

HOFMANN, T. (1999): Probabilistic Latent Semantic Analysis. In: *Proceedings of UAI'99, Uncertainty in Artificial Intelligence*. Stockholm.

JOACHIMS, T. (1997): Text categorization with Support Vector Machines: Learning with many relevant features. In: *Machine Learning: ECML-98*. Springer, Berlin, 137–142.

YANG, Y. and PEDERSEN, J. P. (1997): A comparative study on feature selection in text categorization. In: *Proceedings of ICML'97, International Conference on Machine Learning*. Nashville, US, 412–420.

Statistics and Data Quality Towards more collaboration between these communities

Soumaya Ben Hassine-Guetari[1], Olivier Coppet[2], and Brigitte Laboisse[3]

[1] A.I.D. Company - ERIC Laboratory *sbenhassine@aid.fr*
[2] President of A.I.D. Company - Member of EXQI *ocoppet@aid.fr*
[3] A.I.D. Company - Founder member of EXQI
EXQI Excellence Quality Information (Data Quality association www.exqi.asso.fr) *blaboisse@aid.fr*

Abstract. Summer 1980, during a conference given in the Institute of Statistics of Paris, a very impressive presentation on the FCA analysis that came along with multiple investigation tracks was turned out to be false as it was based on inaccurate data. Thirty years later, data quality is an autonomous discipline with dedicated academic mastering courses (Talburt et al. (2006)), publications (Redman (2001), Wand and Wang (1996)) and software (Gouasdoué et al. (2007)). In fact, a plethora of dimensions, metrics, models and database design techniques (Wang et al. (2001)) are now defined to handle data and their quality in the same flow, helping, then, the statisticians qualify and evaluate their results (Berti-Equille (2007)). In the other hand, statistical models were proposed to define the dimensions' metrics, detect outliers and anomalous data, analyze data heterogeneity, etc. (Batini and Scannapieco (2006))
Let's, then, build a bridge between the two communities and have a track data quality at CompStat 2011!

Keywords: Statistics, data quality, collaboration between these communities

References

Batini, C. and Scannapieco, M. (2006): Data quality : concepts, methodologies and techniques. Springer.

Berti-Equille, L. (2007): Quality Awareness for Managing and Mining Data.

Gouasdou, V., Nugier, S., Duquennoy, D., Laboisse, B. (2007): Une grille pour évaluer la qualité de vos données et choisir votre outil de DQM. *Direct Marketing News 360.*

Talburt, J. , Wu,N., Swaty,J., Adams,J. (2006): Master of Science in Information Quality. *The Journal of Computing Sciences in Colleges.*

Redman, T. (2001): Data quality : the field guide. Digital Press.

Wand, Y. and Wang, R. Y. (1996): Anchoring Data Quality Dimensions in Ontological Foundations.*Communications of the ACM. volume 39.*

Wang, R. Y., Ziad, M., Lee, Y. W. (2001): Data Quality. Kluwer Academic Publishers.

Mobile Learning and e-Book for Teaching Statistics

Tae Rim Lee[1]

Department of Information Statistics, Korea National Open University #169 dongsung dong, jongro ku, Seoul, Korea, *trlee@knou.ac.kr*

Abstract. The Mobile learning (m-Learning) is novel in that it facilitates delivery of learning to the right person, at the right time, in the right place using portable electronic devices. From December 2008 KNOU kick off the mobile learning system under the MOU with Korean Telephone Company KT. In KNOU, m-learning is expanded in almost every department of educational fields.
The Mobile learning and ubiquitous learning system for distance education that anyone who wants to study could study anywhere, anytime with the internet and multimedia system. In the future knowledge based society with rapid change of educational circumstances and paradigm, distance education using ICT technology could satisfy educational desires in various levels of learners. Mobile technologies, like mobile devices and wireless internet services, have the potential to introduce new innovations in the area of education m-learning, a new form of education using mobile internet systems and handheld devices can offer students and teachers the opportunity to interact with and gain access to educational materials, independent of time and spaces. 2009 study suggested some considerable suggestions for preparing the future of distance education based on mobile and one more step advanced ubiquitous learning.
e-Book was produced for supplementary material for teaching statistics with the resources of video files, sound files, Java applet and various kinds of html files. This kinds of new media approach could renovate the new paradigm of teaching statistics.

Keywords: M-Learning, e-Book, ubiquitous learning

References

Junghoon Leem (2007). M-Learning; Its concepts and Implication for the future of education in life-long learning society, *Journal of Lifelong Learning, 3(2): 1-26.*

Joung, Y., R. & Kwak, D. H. (2004). A study of standardization model of learner information in e-learning. *The journal of Korean Association of Computer Education, 7(4), 77–91.*

Keegan, D.(2005), *The Incorporation of Mobile Learning Into Mainstream Education And Training*, Keynote address to world m-Learn congress, South Africa, Oct. 2005

Part III

Wednesday August 25

Bootstrap Prediction in Unobserved Component Models

Alejandro F. Rodríguez[1] and Esther Ruiz[1]

Departamento de Estaística, Universidad Carlos III de Madrid, 28903 Getafe (Madrid), Spain.

Abstract. One advantage of state space models is that they deliver estimates of the unobserved components and predictions of future values of the observed series and their corresponding Prediction Mean Squared Errors (PMSE). However, these PMSE are obtained by running the Kalman filter with the true parameters substituted by consistent estimates and, consequently, they do not incorporate the uncertainty due to parameter estimation. This paper reviews new bootstrap procedures to estimate the PMSEs of the unobserved states and to construct prediction intervals of future observations that incorporate parameter uncertainty and do not rely on particular assumptions of the error distribution. The new bootstrap PMSEs of the unobserved states have smaller biases than those obtained with alternative procedures. Furthermore, the prediction intervals have better coverage properties. The results are illustrate by obtaining prediction intervals of the quarterly mortgages changes and of the unobserved output gap in USA.

Keywords: Backward representation, Random Walk plus noise, NAIRU, output gap, Parameter uncertainty, Prediction Intervals, State Space Models.

A comparison of estimators for regression models with change points

Cathy WS Chen[1], Jennifer SK Chan[2], Richard Gerlach[3], and William Hsieh[1]

[1] Graduate Institute of Statistics & Actuarial Science, Feng Chia University, Taiwan, *chenws@mail.fcu.edu.tw*
[2] School of Mathematics and Statistics, University of Sydney, Australia, *jenniferskchan@gmail.com*
[3] Faculty of Economics and Business, University of Sydney, Australia, *richard.gerlach@sydney.edu.au*

Abstract. We consider two problems concerning locating change points in a linear regression model. One involves jump discontinuities (change-point) in a regression model and the other involves regression lines connected at unknown points. We compare four methods for estimating single or multiple change points in a regression model, when both the error variance and regression coefficients change simultaneously at the unknown point(s): Bayesian, Julious (2001), grid search, and the segmented methods (Muggeo 2008). The proposed methods are evaluated via a simulation study and compared via some standard measures of estimation bias and precision. Finally, the methods are demonstrated and compared using three real data sets. The simulation and empirical results overall favor both the segmented and Bayesian methods of estimation, which simultaneously estimate the change point and the other model parameters, though only the Bayesian method is able to handle both continuous and dis-continuous change point problems successfully. If it is known that regression lines are continuous then the segmented method ranked first among methods.

Keywords: Change point, Jump discontinuities, MCMC, Grid-search, Segmented regression

References

CARLIN, B. P., GELFAND, A. E. and SMITH, A. F. M. (1992): Hierarchical Bayesian analysis of change point problems. *Applied Statistics, 41, 389-405.*

JULIOUS, S. A. (2001): Inference and Estimation in a Changepoint Regression Problem. *The Statistician, 50, 51-61.*

MUGGEO, V. M. R. (2008): Segmented: an R package to fit regression models with broken-line relationships. *the Newsletter of the R project, 8, 20-25.*

An Asymmetric Multivariate Student's t Distribution Endowed with Different Degrees of Freedom

Marc S. Paolella

Chair of Empirical Finance, Swiss Banking Institute
Plattenstrasse 14, 8032 Zurich, Switzerland, *paolella@isb.uzh.ch*

Abstract. An open and active question concerns the construction of a multivariate distribution whose marginals are Student's t but with potentially different degrees of freedom, and the possibility for asymmetry. This is of particular value in empirical finance, where it is well known that the tail indices, or maximally existing moments of the returns, differ markedly across assets, and that stock returns tend to be negatively skewed. While several constructions can be found in the literature, all have various weaknesses. In this paper, we propose a new construction which overcomes many of these drawbacks. The computation of the density via the definition is not just feasible, but numerically reliable, in low dimensions, but too time-consuming in high dimensions, thus prohibiting direct calculation and optimization of the likelihood. To circumvent this, we propose using the method of indirect inference, and demonstrate its efficacy via simulation studies. An example using series comprising the DJIA is illustrated.

Keywords: Multivariate Distribution, Asymmetric Multivariate Student's t Distribution, Indirect Inference

References

GALLANT, A. R. AND TAUCHEN, G. (1996): Which moments to match? *Econometric Theory 12), 657-681.*

GENTON, M. G. (2004): *Skew-Elliptical Distributions and their Applications: A Journey Beyond Normality.* Chapman & Hall/CRC, Boca Raton.

GOURIEROUX, C., MONFORT, A. AND RENAULT, E. (1993): Indirect Inference. *Journal of Applied Econometrics 8, 85-118.*

JONDEAU, E., POON, S. AND ROCKINGER, M. (2007): *Financial Modeling Under Non-Gaussian Distributions.* Springer, London.

JONES, M. C. (2002): A Dependent Bivariate t Distribution with Marginals on Different Degrees of Freedom. *Statistics and Probability Letters 56 (2), 163-170.*

RACHEV, S. T. AND MITTNIK, S. (2000): *Stable Paretian Models in Finance.* John Wiley & Sons, New York.

SHAW, W. T. AND LEE, K. T. A. (2008): Bivariate Student t Distributions with Variable Marginal Degrees of Freedom and Independence. *Journal of Multivariate Analysis 99, 1276-1287.*

Evolutionary Algorithms for Complex Designs of Experiments and Data Analysis

Irene Poli

Department of Statistics at University of Ca Foscari
Cannaregio 873, Venice, Italy, *irenpoli@unive.it*

Abstract. Scientific experimentation on systems behavior is increasingly characterized by high dimensional search spaces. The greater availability of information on the systems, the more powerful technologies for conducting experiments, and deeper experimental questions together pose the problem of involving very large sets of parameters that can affect the experimental results. Classical fractional factorial designs, (Cox and Reid (2005)) do not seem to address properly the problem, since they reduce dimensionality a priori, which may mislead the search and even hide important components or interactions. In this paper we present an approach to experimental design based on the evolutionary paradigm: the design, regarded as a small population of experimental points with different selections of parameters, is evolved in a number of generations according to a set of genetic operators. The construction of this evolutionary design is sequential and interactive: the information collected in one generation is processed to construct the next generation, generating a path of improvement in the space with respect to a defined criterion. At each generation, statistical models are constructed to uncover patterns in the experimental data sets that can make faster and more efficient the search. Some of these Evolutionary Model based Experimental Designs (EMED) have been evaluated in a simulation study exhibiting very successful results (De March et al. (2009); Forlin et al. (2008); Pepelyshev et al.(2009)).

Keywords: combinatorial complexity in experimental design, evolutionary computation, statistical models for evolution.

References

COX D. R and REID N. (2005): *The theory of design of experiments*. Chapman & Hall, London.

DE MARCH, D., FORLIN, M., SLANZI, D., POLI I., (2009): *An evolutionary predictive approach to design high dimensional experiments*. In R. Serra, I. Poli, M. Villani (eds): Artificial Life and Evolutionary Computation: proceedings of WIVACE 2008. World Scientific Publishing Company, Singapore.

FORLIN, M., POLI, I., DE MARCH, D., PACKARD, N., SERRA, R., (2008): *Experiments for self-assembling amphiphilic systems*. Chemometrics and Intellingent Laboratory Systems, 90, 153-160.

PEPELYSHEV, A., POLI, I., MELAS, V., (2009): *Uniform coverage designs for mixture experiments*, Journal of Statistical Planning and Inference, 139, 3442-3452.

Evolutionary Computation for Modelling and Optimization in Finance

Sandra Paterlini[1,2]

[1] Department of Economics, CEFIN and RECent, University of Modena and Reggio E., Viale Berengario 51, Modena, Italy, *sandra.paterlini@unimore.it*
[2] CEQURA, Center for Quantitative Risk Analysis, Akademiestr. 1/I, Munich, Germany

Abstract. In the last decades, there has been a tendency to move away from mathematically tractable, but simplistic models towards more sophisticated and real-world models in finance. However, the consequence of the improved sophistication is that the model specification and analysis is no longer mathematically tractable. Instead solutions need to be numerically approximated. For this task, evolutionary computation heuristics are the appropriate means, because they do not require any rigid mathematical properties of the model. Evolutionary algorithms are search heuristics, usually inspired by Darwinian evolution and Mendelian inheritance, which aim to determine the optimal solution to a given problem by competition and alteration of candidate solutions of a population. In this work, we focus on credit risk modelling and financial portfolio optimization to point out how evolutionary algorithms can easily provide reliable and accurate solutions to challenging financial problems.

Keywords: population-based algorithms, multi-objective optimization, clustering, credit risk modelling, financial portfolio optimization

Heuristic Optimization for Model Selection and Estimation

Dietmar Maringer[1]

Department for Quantitative Methods, Economics Faculty, University of Basel
Peter Merian-Weg 6, CH-4002 Basel, Switzerland, *dietmar.maringer@unibas.ch*

Abstract. Model selection and estimation does not always come with the luxury of closed form analytical solutions. Often, numerical search or optimization procedures have to be used instead. This search process, however, is not always straightforward: local optima, frictions in the search space and model constraints add more complexity than simple methods can deal with. This is one major reason why, as found in several empirical studies, different software packages can differ substantially in their reported results for a given problem.

Heuristic methods are less restricted with regard to the properties of the search space, objective function and constraints. A popular class are evolutionary methods. Inspired by natural principles, these methods typically maintain a number of different candidate solutions which are combined and modified into new solutions and where improved candidates are likely to replace inferior ones. These methods can be shown to find the global optimum with an increasing probability when CPU time is increased and with suitable calibration.

This presentation discusses some basic principles as well as selected applications. Numerical experiments show that these methods can outperform standard approaches.

Keywords: heuristic optimization, model selection, estimation, calibration

Complexity Questions in Non-Uniform Random Variate Generation

Luc Devroye

School of Computer Science
McGill University
Montreal, Canada H3A 2K6
lucdevroye@gmail.com

Abstract. In this short note, we recall the main developments in non-uniform random variate generation, and list some of the challenges ahead.

Keywords: random variate generation, Monte Carlo methods, simulation

Large-Scale Machine Learning with Stochastic Gradient Descent

Léon Bottou

NEC Labs America, Princeton NJ 08542, USA
leon@bottou.org

Abstract. During the last decade, the data sizes have grown faster than the speed of processors. In this context, the capabilities of statistical machine learning methods is limited by the computing time rather than the sample size. A more precise analysis uncovers qualitatively different tradeoffs for the case of small-scale and large-scale learning problems. The large-scale case involves the computational complexity of the underlying optimization algorithm in non-trivial ways. Unlikely optimization algorithms such as stochastic gradient descent show amazing performance for large-scale problems. In particular, second order stochastic gradient and averaged stochastic gradient are asymptotically efficient after a single pass on the training set.

Keywords: stochastic gradient descent, online learning, efficiency

Temporally-Adaptive Linear Classification for Handling Population Drift in Credit Scoring

Niall M. Adams[1], Dimitris K. Tasoulis[1], Christoforos Anagnostopoulos[2], and David J. Hand[1,2]

[1] Department of Mathematics
Imperial College London, UK *[n.adams, d.tasoulis, d.j.hand]@imperial.ac.uk*
[2] The Institute for Mathematical Sciences
Imperial College London, UK *canagnos@imperial.ac.uk*

Abstract. Classification methods have proven effective for predicting the creditworthiness of credit applications. However, the tendency of the underlying populations to change over time, *population drift,* is a fundamental problem for such classifiers. The problem manifests as decreasing performance as the classifier ages and is typically handled by periodic classifier reconstruction. To maintain performance between rebuilds, we propose an adaptive and incremental linear classification rule that is updated on the arrival of new labeled data. We consider adapting this method to suit credit application classification and demonstrate, with real loan data, that the method outperforms static and periodically rebuilt linear classifiers.

Keywords: classification, credit scoring, population drift, forgetting factor

Multivariate Stochastic Volatility Model with Cross Leverage

Tsunehiro Ishihara[1] and Yasuhiro Omori[2]

[1] Graduate School of Economics, University of Tokyo. 7-3-1 Hongo, Bunkyo-Ku, Tokyo 113-0033, Japan.
[2] Faculty of Economics, University of Tokyo. 7-3-1 Hongo, Bunkyo-Ku, Tokyo 113-0033, Japan. Tel: +81-3-5841-5516. *omori@e.u-tokyo.ac.jp*

Abstract. The Bayesian estimation method using Markov chain Monte Carlo is proposed for a multivariate stochastic volatility model that is a natural extension of the univariate stochastic volatility model with leverage, where we further incorporate cross leverage effects among stock returns.

Keywords: asymmetry, Bayesian analysis, leverage effect, Markov chain Monte Carlo, multi-move sampler, multivariate stochastic volatility, stock returns

Estimating Factor Models for Multivariate Volatilities: An Innovation Expansion Method

Jiazhu Pan[1], Wolfgang Polonik[2], and Qiwei Yao[3]

[1] Department of Mathematics and Statistics, University of Strathclyde
26 Richmond Street, Glasgow, G1 1XH, UK, *jiazhu.pan@strath.ac.uk*
[2] Division of Statistics, University of California at Davis
Davis, CA 95616, USA, *wpolonik@ucdavis.edu*
[3] Department of Statistics, London School of Economics
London WC2A 2AE, UK, *q.yao@lse.ac.uk*

Abstract. We introduce an innovation expansion method for estimation of factor models for conditional variance (volatility) of a multivariate time series. We estimate the factor loading space and the number of factors by a stepwise optimization algorithm on expanding the "white noise space". Simulation and a real data example are given for illustration.

Keywords: dimension reduction, factor models, multivariate volatility

Semiparametric Seasonal Cointegrating Rank Selection

Byeongchan Seong[1], Sung K. Ahn[2], and Sinsup Cho[3]

[1] Department of Statistics, Chung-Ang University,
221, Heukseok-dong, Dongjak-gu, Seoul 156-756, Korea, *bcseong@cau.ac.kr*
[2] Department of Management and Operations, Washington State University,
Pullman, WA 99164-4736, USA, *ahn@wsu.edu*
[3] Department of Statistics, Seoul National University,
Seoul 151-747, Korea, *sinsup@snu.ac.kr*

Abstract. This paper considers the issue of seasonal cointegrating rank selection by information criteria as the extension of Cheng and Phillips (The Econometrics Journal, Vol. 12, pp. S83–S104, 2009). The method does not require the specification of lag length in vector autoregression, is convenient in empirical work, and is in a semiparametric context because it allows for a general short memory error component in the model with only lags related to error correction terms. Some limit properties of usual information criteria are given for the rank selection and small Monte Carlo simulations are conducted to evaluate the performances of the criteria.

Keywords: seasonal cointegrating rank, information criteria, nonparametric, model selection

References

CHENG, X. and PHILLIPS, P. C. B. (2009): Semiparametric cointegrating rank selection. *The Econometrics Journal 12, S83–S104.*

Part IV

Thursday August 26

Bayesian space-time modelling of count data using INLA

Leonhard Held[1], Andrea Riebler[1], Håvard Rue[2], and Birgit Schrödle[1]

[1] University of Zurich, IFSPM, Biostatistics Unit
Hirschengraben 84, 8001 Zurich, Switzerland, *held@ifspm.uzh.ch*
[2] Department of Mathematical Sciences, Norwegian University of Science and Technology
N-7491 Trondheim, Norway

Abstract. Integrated nested Laplace approximations (INLA) have been recently proposed for fitting Bayesian hierarchical models. In this talk I will discuss the application of INLA to space-time modelling of count data. Such data often arise in epidemiological applications. We will describe the fitting of several models allowing for space-time interactions, including a novel approach based on correlated random walk priors. An application to space-time counts of selected diseases among cattle in Switzerland is given.

Keywords: INLA, Space-time modelling, count data

Assessing the Association between Environmental Exposures and Human Health

Linda J. Young[1] Carol A. Gotway[2] Kenneth K. Lopiano[1] Greg Kearney[2] and Chris DuClos[3]

[1] Department of Statistics, IFAS, University of Florida, Gainesville FL 32611-0339 USA, *LJYoung@ufl.edu*, *klopiano@ufl.edu*
[2] U.S. Centers for Disease Control & Prevention Atlanta, GA USA *cdg7@cdc.gov*, *irr8cdc.gov*

Abstract. In environmental health studies, health effects, environmental exposures, and potential confounders are seldom collected during the study on the same set of units. Some, if not all of the variables, are often obtained from existing programs and databases. Suppose environmental exposure is measured at points, but health effects are recorded on areal units. Further assume that a regression analysis the explores the association between health and environmental exposure is to be conducted at the areal level. Prior to analysis, the information collected on exposure at points is used to predict exposure at the areal level, introducing uncertainty in exposure for the analysis units. Estimation of the regression coefficient associated with exposure and its standard error is considered here. A simulation study is used to provide insight into the effects of predicting exposure. Open issues are discussed.

Keywords: modified areal unit problem, change of support, errors in variables

Examining the Association between Deprivation Profiles and Air Pollution in Greater London using Bayesian Dirichlet Process Mixture Models

John Molitor, Léa Fortunato, Nuoo-Ting Molitor, Sylvia Richardson

MRC-HPA Centre for Environment and Health and Department of Epidemiology and Biostatistics, Imperial College, London

Abstract. Standard regression analyses are often plagued with problems encountered when one tries to make inference going beyond main effects, using datasets that contain dozens of variables that are potentially correlated. This situation arises, for example, in environmental deprivation studies, where a large number of deprivation scores are used as covariates, yielding a potentially unwieldy set of interrelated data from which teasing out the joint effect of multiple deprivation indices is difficult. We propose a method, based on Dirichlet-process mixture models that addresses these problems by using, as its basic unit of inference, a profile formed from a sequence of continuous deprivation measures. These deprivation profiles are clustered into groups and associated via a regression model to an air pollution outcome. The Bayesian clustering aspect of the proposed modeling framework has a number of advantages over traditional clustering approaches in that it allows the number of groups to vary, uncovers clusters and examines their association with an outcome of interest and fits the model as a unit, allowing a region's outcome potentially to influence cluster membership. The method is demonstrated with an analysis UK Indices of Deprivation and PM10 exposure measures corresponding to super output areas (SOA's) in greater London.

Keywords: Bayesian analysis, Dirichlet processes, mixture models, MCMC, environmental justice

Bag of Pursuits and Neural Gas for Improved Sparse Coding

Kai Labusch, Erhardt Barth, and Thomas Martinetz

University of Lübeck
Institute for Neuro- and Bioinformatics
Ratzeburger Allee 160
23562 Lübeck, Germany {*labusch,barth,martinetz*}*@inb.uni-luebeck.de*

Abstract. Sparse coding employs low-dimensional subspaces in order to encode high-dimensional signals. Finding the optimal subspaces is a difficult optimization task. We show that stochastic gradient descent is superior in finding the optimal subspaces compared to MOD and K-SVD, which are both state-of-the art methods. The improvement is most significant in the difficult setting of highly overlapping subspaces. We introduce the so-called "Bag of Pursuits" that is derived from Orthogonal Matching Pursuit. It provides an improved approximation of the optimal sparse coefficients, which, in turn, significantly improves the performance of the gradient descent approach as well as MOD and K-SVD. In addition, the "Bag of Pursuits" allows to employ a generalized version of the Neural Gas algorithm for sparse coding, which finally leads to an even more powerful method.

Keywords: sparse coding, neural gas, dictionary learning, matching pursuit

On the Role and Impact of the Metaparameters in t-distributed Stochastic Neighbor Embedding

John A. Lee[1] and Michel Verleysen[2]

[1] Imagerie Moléculaire et Radiothérapie Expérimentale
Avenue Hippocrate 54, B-1200 Brussels, Belgium *john.lee@uclouvain.be*
[2] Machine Learning Group - DICE, Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium *michel.verleysen@uclouvain.be*

Abstract. Similarity-based embedding is a paradigm that recently gained interest in the field of nonlinear dimensionality reduction. It provides an elegant framework that naturally emphasizes the preservation of the local structure of the data set. An emblematic method in this trend is t-distributed stochastic neighbor embedding (t-SNE), which is acknowledged to be an efficient method in the recent literature. This paper aims at analyzing the reasons of this success, together with the impact of the two metaparameters embedded in the method. Moreover, the paper shows that t-SNE can be interpreted as a distance-preserving method with a specific distance transformation, making the link with existing methods. Experiments on artificial data support the theoretical discussion.

Keywords: similarity-based embedding, dimensionality reduction, nonlinear projection, manifold learning, t-SNE

An Empirical Study of the Use of Nonparametric Regression Methods for Imputation

I. R. Sánchez-Borrego, M. Rueda and E. Álvarez-Verdejo

Department of Statistics and Operational Research
18071 University of Granada, Spain *ismasb@ugr.es*, *mrueda@ugr.es*, *encarniav@ugr.es*

Abstract. We address the problem of data incompleteness. A new algorithm based on Multivariate Adaptive Regression Splines is proposed to impute missing observations. A comparison with several imputations methods is addressed by considering missing at random (MAR) and missing completely at random (MCAR) missing data mechanisms. Two different ways of adding a disturbance to the imputation estimators are also addressed. A simulation study has been performed and a real-life data have been considered to illustrate the precision of the proposed method.

Keywords: Local polynomial regression, MARS, imputation, survey sampling, nonparametric regression.

References

BREIDT, F.J. and OPSOMER, J.D. (2000) Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4), 1026–1053.

CHENG, P.E. (1994) Nonparametric Estimation of Mean Functionals with Data Missing ar Random. *Journal of the American Statistical Association*, 89, 425, 81–87.

FAN, J. and GIJBELS, I. (1996) *Local Polynomial Modelling and Its Applications*. Ed. Chapman and Hall.

FRIEDMAN, J.H. (1991) Multivariate Adaptive Regression Splines *The Annals of Statistics*, Vol. 19, No. 1 (Mar. 1991), pp. 1–67.

RUPPERT, D. and WAND, M. P. (1994) Multivariate locally weighted least squares regression *The Annals of Statistics*, 22(3), 1346–1370.

SÄRNDAL, C.E. and LUNSTRÖM, S. (2005) *Estimation in Surveys with Nonresponse*. Wiley Series in Survey Methodology.

The Problem of Determining the Calibration Equations to Construct Model-calibration Estimators of the Distribution Function

S. Martínez[1], M. Rueda[2], A. Arcos[2], H. Martínez[1] and J.F. Muñoz[3]

[1] Department of Statistics and Applied Mathematics
04120 University of Almería, Spain *spuertas@ual.es*, *hmartinez@ual.es*
[2] Department of Statistics and Operational Research
18171 University of Granada, Spain *mrueda@ugr.es*, *arcos@ugr.es*
[3] Department of Quantitative Methods in Economics
18171 University of Granada, Spain *jfmunoz@ugr.es.*

Abstract. The calibration approach to estimating the finite population distribution function was proposed by Rueda et al. (2007). The proposed estimator is built by means of constraints that require the use of a set of fixed values $t_1, ..t_p$. Martínez et al. (2010), under the context of a linear regression working model, consider the case of only one point for the calibration and determine the optimum value t_1 in the sense of minimum variance. In the present paper, assuming the use of more complex models, we study the problem of determining the optimal values t_i that gives the best estimation under simple random sampling without replacement for the case $p = 2$

Keywords: distribution function, finite population, model-calibration approach

References

MARTÍNEZ, S., RUEDA, M., ARCOS, A. and MARTÍNEZ, H. (2010): Optimum calibration points estimating distribution functions. *Journal of Computational and Applied Mathematics 233, 2265-2277.*

RUEDA, M., MARTÍNEZ, S., MARTÍNEZ, H. and ARCOS, A. (2007): Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference 137 (2), 435-448.*

Computation of the projection of the inhabitants of the Czech Republic by sex, age and the highest education level*

Tomáš Fiala and Jitka Langhamrová

Department of Demography, Faculty of Informatics and Statistics, University of Economics Prague, W. Churchill sq. 4, 130 67 Prague, Czech Republic, *fiala@vse.cz*, *langhamj@vse.cz*

Abstract. The computation is based on the classical component method of population projections computations. Four education levels: primary education, secondary lower education, secondary higher education and tertiary education are distinguished. The surviving probabilities are supposed to depend not only on the sex and age but also on the education level of the person. The projection has been computed for the population of the Czech Republic since 2001 until 2051.

Keywords: population projection, component method, education level

References

ARLT, J. and ARLTOVÁ, M. (2009): *Ekonomické časové řady*. Prof. Publ., Prague.

BOGUE, D. J., ARRIAGA, E. E. and ANDERTON, D., L. (eds.). (1993): *Readings in Population Research Methodology* Vol. 5. Population Models, Projections and Estimates. UNFPA, Social Development Center, Chicago, Illinois.

ČSÚ (Czech Statistical Office) (2009): *Projekce obyvatelstva České republiky do roku 2065*. http://www.czso.cz/csu/2009edicniplan.nsf/p/4020-09.

ČSÚ (Czech Statistical Office) (2003): *Úroveň vzdělání obyvatelstva podle výsledku sčítání lidu*. http://www.czso.cz/csu/2003edicniplan.nsf/p/4113-03.

FIALA, T. and LANGHAMROVÁ, J. (2009): Human resources in the Czech Republic 50 years ago and 50 years after. In: *IDIMT-2009 System and Humans A Complex Relationship*. Trauner Verlag universitat, Linz.

HULÍK, V. and TESÁRKOVÁ, K. (2009): Vývoj přístupu terciárního vzdělávání v České republice v závislosti na demografickém vývoji. In: *Reprodukce lidského kapitálu Vzájemné vazby a souvislosti* [CD-ROM]. Oeconomica, Praha, 1-21.

KAČEROVÁ, E. (2008): International migration and mobility of the EU citizens in the Visegrad group countries: Comparison and bilateral flows. In: *European Population Conference*. Barcelona. EPC, 142.

KOSCHIN, F. (2005): *Kapitoly z ekonomické demografie*. Oeconomica, Praha.

LANGHAMROVÁ, J. at al. (2009): *Prognóza lidského kapitálu obyvatelstva České republiky do roku 2050*. Oeconomica, Praha.

MAZOUCH, P. and FISCHER, J. (2007): Střední délka života podle nejvyššího ukončeného vzdělání. In: *Firma a konkurenční prostředí*. MSD, Brno, 91-95.

* This paper was written with the support of research project 2D06026, "Reproduction of Human Capital", financed by the MŠMT ČR.

A comparison between Beale test and some heuristic criteria to establish clusters number

Angela Alibrandi[1] and Massimiliano Giacalone[2]

[1] Department of Economical, Financial, Social, Environmental, Statistical and Territorial Sciences (S.E.FI.S.A.S.T.), University of Messina, Via dei Verdi 75, 98122 Messina, *aalibrandi@unime.it*

[2] Department of Public Organization Law, Economy and Society (D.O.P.E.S), University of Catanzaro "Magna Graecia", Campus of Germaneto, 88100 Catanzaro, *maxgiacit@yahoo.it*

Abstract. Cluster analysis represents an ideal data-mining tool because the classes or groups that the data form are unknown, especially as the state definition is expanded to include an increasing number of variables. Cluster analysis uncovers these underlying patterns in the data and assigns each case to a cluster. As it is known, unlike the discriminant analysis, in the cluster analysis there isn't information about the number of cluster and the characteristics of the groups in the population. Therefore, the individualization of the grouping structure constitutes a fundamental decision to be taken, in order to correctly assign the units to not-previously defined groups of observations. With reference to the individualization of adequate number of clusters a lot of criteria have been proposed in literature (Xu Rui et al., 2008).
Purpose of the paper is to examine the theoretical bases of some most common criteria: Beale test (Gordon, 1999), based on the significance logic and two heuristic methods as Pseudo T^2 Hotelling (Halkidi, 2002) and Cubic Clustering Criterion (Sarle, 1983). Moreover, we want to compare them in terms of flexibility and applicability, taking in account the assumptions on which they are based; finally we apply all these criteria on real data and we compare the obtained results.

Keywords: Clusters number, Grouping structure, Beale test, T^2 Hotelling, Cubic Clustering Criterion

References

GORDON A.D.(1999): *Classification*, 2nd edition, Chapman and Hall.

HALKIDI M., BATISTAKIS Y., VAZIRGIANNIS M.(2002): *Cluster Validity Methods*, Part I. SIGMOD Record, Vol. 31 **2**

SARLE W.S. (1983): *Cubic Clustering Criterion*. SAS Technical Report. **1**, 108.

XU RUI, II DONALD C. WUNSCH (2008): Recent advances in cluster analysis, *International Journal of Intelligent Computing and Cybernetics*, Emerald Group Publishing Limited.

Variable Selection for Semi-Functional Partial Linear Regression Models

Germán Aneiros[1], Frédéric Ferraty[2] and Philippe Vieu[2]

[1] Departamento de Matemáticas, Universidade da Coruña
Campus de Elviña s/n, 15071 A Coruña, Spain, *ganeiros@udc.es*
[2] Institut de Mathématiques, Université Paul Sabatier
31062 Toulouse cedex, France, *ferraty@cict.fr, vieu@cict.fr*

Abstract. We consider a regression model where the regression function is the sum of a linear and a nonparametric component (that is, a partial linear regression model). More specifically, we focus on the case where the covariate that enters in a nonparametric way is of functional nature (see Aneiros-Pérez and Vieu (2006) for a first paper), the number of covariates in the linear part is divergent, and the corresponding vector of regression coefficients is sparse. The aim of this work is variable selection and estimation in such a model.

A penalized-least-squares based procedure to simultaneously select variables and estimate coefficients of variables is proposed, and a guideline is given for indicating how to select the various tuning parameters corresponding to our estimator. Finally, in order to illustrate the practical interest of the proposed procedure, a simulation study is reported

This work is related with those of Liang and Li (2009), Ni et al. (2009) and Xie and Huang (2009), who studied estimation and variable selection in partial linear regression models where all the covariates were scalar. Our main contribution is the introduction of a functional covariate in the model.

The research of Germán Aneiros was supported by Xunta de Galicia Grant PGIDIT07PXIB105259PR, and by the research group MODES.

Keywords: functional data, semiparametric regression, variable selection

References

ANEIROS-PÉREZ, G. and VIEU, P. (2006): Semi-functional partial linear regression. *Statistics and Probability Letters 76, 1102-1110.*

LIANG, H. and LI, L. (2009): Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association 104, 234-248.*

NI, X., ZHANG, H. H. and ZHANG, D. (2009): Automatic model selection for partially linear models. *Journal of Multivariate Analysis 100, 2100-2111.*

XIE, H. and HUANG, J. (2009): SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics 37, 673-696.*

Analysis of Baseball Data for Evaluating the Sacrifice Bunt Strategy Using the Decision Tree

Kazunori Yamaguchi[1] and Michiko Watanabe[2]

[1] College of Business, Rikkyo University
Nishi-Ikebukuro Toshima-kuTokyo, 171-8501 Japan, *kyamagu@rikkyo.ac.jp*
[2] Faculty of Economics, Toyo University
Hakusan Bunkyo-ku Tokyo, 112-8606 Japan, *watanabe_michiko@nifty.com*

Abstract. In this paper, we explore the situation that the sacrifice bunts strategy in baseball games in Japan is effective using the decision tree.

The baseball is one of the favorite sports in US and Japan. Tango et al.(2007) said that few strategies elicit as much emotion and controversy as the sacrifice bunt. In lparticular, the sacrifice bunt is so common in Japan. The strategy has been used much more frequently in Japan than MLB in US.

Many researches have been done for evaluations for such strategy (e.g. see Albert and Bennet, 2003, Thorn and Palmer 1985). They discussed about overall effect of the strategy using MLB data in US and concluded that the sacrifice bunt is generally an ineffective and archaic strategy.

We use the decision tree to find the situation the sacrifice bunts strategy is effective. We use data from all games of Japan professional baseball in two years. We use the first year data to find the situations and use the second year data to check the effectiveness of the situations.

Keywords: baseball strategy, decision tree, sports statistics

References

ALBERT, J. and BENNETT, J.(2003) *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game* Springer.

TANGO, T., LICHTMAN, M. and DOLPHIN, A.(2007): *The Book: Playing the Percentages in Baseball.* Potomac Books.

THORN, J. and PALMER, P.(1985): *The Hidden Game of Baseball* Doubleday, New York.

A Transient Analysis of a Complex Discrete k-out-of-n:G System with Multi-state Components

Ruiz-Castro, Juan Eloy[1] and Paula R. Bouzas[2]

[1] Department of Statistic and Operations Research.
University of Granada, 18071-Granada, Spain. *jeloy@ugr.es*
[2] Department of Statistic and Operations Research.
University of Granada, 18071-Granada, Spain. *paula@ugr.es*

Abstract. A system with n components that works if and only if at least k of them does it is called a k-out-of-n:G system. A discrete k-out-of-n:G system is modelled by considering multi-state components. Phase-type distributions for the lifetime of the units are considered. The units can undergo repairable and non-repairable failures from any state. We assume a general number of repairpersons. The repair time for each repairperson is general distributed and the phase type representation is considered. The system is modelled and some performance measures of interest are built. All results have been implemented computationally with $Matlab$. A numerical application shows the versatility of the model.

Keywords: discrete k-out-of-n:G system, Phase-type distribution, Reliability

References

KRISHNAMOORTHY, A. and USHAKUMARI, P.V. (2001): k-out-of-n: G system with repair: the D-policy. *Computers & Operations Research 28, 973-981.*

MOUSTAFA, M.S. (2001): Availability of K-out-of-N:G Systems with Exponential Failures and General Repairs. *Economic Quality Control 16 (1), 75-82.*

NEUTS, M. F. (1981): *Matrix-geometric solutions in stochastic models: an algorithmic approach.* Johns Hopkins University Press, Baltimore.

PHAM, HOANG (2008): *Recent Advances in Reliability and Quality in Design.* Springer.

RUIZ-CASTRO, J.E., FERNÁNDEZ-VILLODRE, G. and PÉREZ-OCÓN, R. (2009a): A level-dependent general discrete system involving phase-type distributions. *IIE Transactions 41 (1), 45-56.*

RUIZ-CASTRO, J.E., FERNÁNDEZ-VILLODRE, G. and PÉREZ-OCÓN, R. (2009b): A Multi-Component General Discrete System Subject to Different Types of Failures with Loss of Units. *Discrete Event and Dynamic Systems 19 (1), 31-65.*

TIAN, Z., ZUO, M.J. and YAM, R.C.M. (2009): Multi-state k-out-of-n systems and their performance evaluation. *IIE Transactions 49, 32-44.*

Empirical analysis of the climatic and social-economic factors influence on the suicide development in the Czech Republic*

Markéta Arltová[1], Jitka Langhamrová[2], and Jana Langhamrová[3]

[1] Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics Prague, W. Churchill sq. 4, 130 67 Prague, Czech Republic, *arltova@vse.cz*
[2] Depart. of Demography, Fac. of Informatics and Statistics, Univ. of Economics Prague, W. Churchill sq. 4, 130 67 Prague, Czech Rep., *langhamj@vse.cz*
[3] Student of Statistics, University of Economics Prague, W. Churchill sq. 4, 130 67 Prague, Czech Republic, *xlanj18@vse.cz*

Abstract. The suicide rate is closely linked with social-economic and climatic factors. For the empirical analysis of this influence we made use of time series of the level of unemployment, average temperatures, the average duration of sunlight in hours and total precipitation. The analysis was carried out on monthly time series in the period from Jan. 1999 to Dec. 2007. The estimated model shows the influence of the duration of sunlight and the level of unemployment on the development of the number of suicides. From the viewpoint of the influence of weather on the suicide rate it was demonstrated that with the increase in hours of sunlight there is also an increase in the number of suicides. Weather clearly acts as a "trigger mechanism". Further mediating factors, such as, for example, psychic aspects and mental illness, must also be taken into account.

Keywords: Suicide, climatic factors, social-economic factors, time series

References

ARLT, J. and ARLTOVÁ, M. (2009): *Ekonomické časové řady*. Prof. Publ., Prague.

KREJČÍKOVÁ, J. (2009): *Analýza počtu sebevražd v České republice*. Thesis, University of Economics, Prague.

POLÁŠEK, V. (2006): *Sebevraždy v České republice - 2001 až 2005*. Czech Statistical Office. http://www.czso.cz/csu/2006edicniplan.nsf/p/4012-06.

YIP, P.S.F., CHAO, A. and CHIU, C.W.F. (2000): Seasonal variation in suicides: diminished or vanisshed. Experience from England and Wales, 1982–1996. *British Journal of Psychiatry, 177, 366-369.*

ZOLLNER, L., MOLLER, S. and JENSEN B.F. (2003): Meteorological factors and seasonality in suicidal behaviour in Denmark 1970-2000. *Working papers*, Centre for Suicide Research, Odense, Denmark.

* This paper was written with the support of Grant Agency of the Czech Republic No. 402/09/0369 "Modelling of Demographic Time Series in Czech Republic".

Thresholding-Wavelet-Based Functional Estimation of Spatiotemporal Strong-Dependence in the Spectral Domain

María Pilar Frías[1] and María Dolores Ruiz-Medina[2]

[1] University of Jaén
Campus Las Lagunillas
23071 Jaén, Spain
(e-mail: mpfrias@ujaen.es)
[2] University of Granada
Campus Fuente Nueva
18071 Granada, Spain
(e-mail: mruiz@ugr.es)

Abstract. Four functional parameter estimation algorithms are proposed for the statistical analysis of temporal and spatial long-range dependence models. Specifically, the class of strong-dependence spatiotemporal random fields studied in Frías et al. (2006a, 2008, 2009) is considered. The functional sample information is assumed to be collected in the spectral domain, and affected by additive measurement noise. In the estimation methodology proposed, a wavelet analysis of the spectral functional data is first performed. Compactly supported wavelet functions are considered in this analysis. Thresholding techniques are applied for removing the observation noise. The parameter estimators are then computed by applying linear regression in the log-thresholding wavelet domain. The performance of the estimation algorithms proposed is illustrated from simulated data.

Keywords: Fractal spectral processes, long-range dependence parameters, spatiotemporal parametric models, wavelet transform.

References

FRIAS, M. P., RUIZ-MEDINA, M. D., ALONSO, F. J. and ANGULO, J. M. (2006a): Spatiotemporal generation of long-range dependence models and estimation. *Environmetrics 17, 139–146.*

FRIAS, M. P., RUIZ-MEDINA, M. D., ALONSO, F. J. and ANGULO, J. M. (2008): Parameter estimation of self-similar spatial covariogram models. *Commputation Statatistics - Theory and Methods 37, 1011–1023.*

FRIAS, M. P., RUIZ-MEDINA, M. D., ALONSO, F. J. and ANGULO, J. M. (2009): Spectral-marginal-based estimation of spatiotemporal long-range dependence. *Commputation Statatistics - Theory and Methods 38, 103–114.*

On Composite Pareto Models

Sandra Teodorescu[1] and Raluca Vernic[2]

[1] Faculty of Economic Sciences, Ecological University of Bucharest
1G Vasile Milea Blvd., Bucharest, Romania, *cezarina_teodorescu@yahoo.com*
[2] Faculty of Mathematics and Computer Science, Ovidius University of Constanta
124 Mamaia Blvd., Constanta, Romania, and
Institute for Mathematical Statistics and Applied Mathematics, Casa Academiei
13 Calea 13 Septembrie, Bucharest, Romania, *rvernic@univ-ovidius.ro*

Abstract. To model statistical data generated by two different distributions, Cooray and Ananda (2005) introduced a composite Lognormal-Pareto model, further developed by Scollnik (2007). In this paper, we consider a more general composite Pareto model by replacing the Lognormal distribution with an arbitrary continuous one. The main characteristics of this model, as well as some statistical inference are presented. We will also provide comprehensive and numerical details to illustrate the particular case of the composite Gamma-Pareto model.

Keywords: composite distributions, Pareto distribution, Gamma distribution, statistical inference

References

COORAY, K. and ANANDA, M.A. (2005): Modeling actuarial data with a composite Lognormal-Pareto model. *Scandinavian Actuarial Journal 5, 321–334.*

SCOLLNIK, D.P.M. (2007): On composite Lognormal-Pareto models. *Scandinavian Actuarial Journal 1, 20–33.*

Data Visualization and Aggregation

Junji Nakano[1] and Yoshikazu Yamamoto[2]

[1] The Institute of Statistical Mathematics
Tachikawa, Tokyo, Japan, *nakanoj@ism.ac.jp*
[2] Tokushima Bunri University
Takamatsu, Kagawa, Japan *yamamoto@is.bunri-u.ac.jp*

Abstract. Visualizing data using interactive and dynamic graphics is a useful first step of statistical data analysis, especially when the data are new to the analyst and the amount of them is very large. Recently, several data are collected by automatic data acquisition systems over networks, and become so huge that even high-speed computers require considerable time to draw interactive graphics that show all the observations. Therefore, we sometimes "aggregate" data by grouping them appropriately to reduce the amount of data without losing the information of the original data too much.

There exist several data aggregation techniques. Symbolic data analysis expresses a group of data as a "concept", a second level data described by variables which take complicated values such as intervals and histograms. In relational database techniques, online analytical processing (OLAP) is extensively used to calculate the summation of variable values by interactively grouping the data.

Data usually contain both categorical and real valued variables. It is not easy to express the structure of such data clearly by traditional statistical graphics.

We propose an interactive and flexible aggregation of groups of data which are induced mainly by the values of categorical variables. The aggregation result is expressed by several graphics components, such as a dot, a boxplot and a histogram on usual graphics. We demonstrate an aggregation and visualization system which include extended parallel coordinate plot and scatter diagram. Simple example shows that an aggregation is a powerful visualization tool to reveal the structure of complex data.

Keywords: OLAP, Parallel coordinate plot, Symbolic Data Analysis

Clustering of Czech Household Incomes Over Very Short Time Period

Marie Forbelská[1] and Jitka Bartošová[2]

[1] Masaryk University, Department of Mathematics and Statistics of the Faculty of Science, Kotlářská 2, Brno, Czech Republic, *forbel@math.muni.cz*
[2] University of Economics Prague, Department of Management of Information of the Faculty of Management, Jarošovská 1117/II, Jindřichův Hradec, Czech Republic, *bartosov@fm.vse.cz*

Abstract. The article deals with cluster analysis of household income dynamics based on the results of statistical survey EU SILC 2005, 2006 and 2007. We handle the problem of clustering many short time series. Mixed effects models offer a flexible framework for appropriate modeling of among trial correlations and individual trial variance heterogeneity. Consequently, we assume that random parameters are distributed according to a finite normal mixture and we use this mixture model for clustering short time series. The R environment (R Development Core Team, 2008) is used for both mixed model analysis and cluster analysis .

Keywords: household income, finite mixture model, clustering, mixed effects models

Testing the Number of Components in Poisson Mixture Regression Models

Susana Faria[1] and Fátima Gonçalves[2]

[1] Department of Mathematics and Applications, Mathematical Research Centre , University of Minho, 4800-058 Guimarães, Portugal *sfaria@math.uminho.pt*
[2] University of Minho, 4800-058 Guimarães, Portugal *fat.rod.goncalves@sapo.pt*

Abstract. Estimating the number of mixture components is one of the major difficulties in the application of finite mixture models. The likelihood ratio test is a general statistical procedure to use. Unfortunately, a number of specific problems arise and the classical theory fails to hold. In this paper we investigate the testing of hypotheses concerning the number of components in Poisson regression models (PMR) via parametric and nonparametric bootstrap. We also compare the performance of these procedures with criteria AIC and BIC in testing the number of components in these models.

Keywords: EM algorithm, Mixture Poisson Regression Models, Likelihood ratio test, Resampling.

References

AITKIN, M. (1996): A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing, 6: 251-262.*

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977): Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc., Ser. B 39: 1-38.*

FRUHWIRTH-SCHNATTER (2006): *Finite Mixture and Markov Switching Models*, Springer, Heidelberg.

HURN, M., JUSTEL, A. and ROBERT, C.P.(2003): Estimating Mixtures of Regressions. *Journal of Computational and Graphical Statistics, 12: 55-79.*

KARLIS, D. and XEKALAKI, E. (1999): On testing for the number of components in finite Poisson mixtures. *Ann. Inst. of Stat. Math., 51: 149-161.*

MCLACHLAN, G.J. and PEEL, D. (2000): *Finite Mixture Models*, Wiley, New York.

SCHLATTMANN, P. (2005): On bootstrapping the number of components in finite mixtures of Poisson distributions. *Statistics and Computing 15(3): 179-188* .

TURNER, T. (2000): Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *Applied Statistics 49 (3): 371-384* .

WANG, P., PUTERMAN, M.L., COCKBURN, I.M. and LE, N. (1996): Mixed Poisson Regression Models with Covariate Dependent Rates. *Biometrics 52 (2): 381-400.*

A Statistical Survival Model Based on Counting Processes

Jose-Manuel Quesada-Rubio, Julia Garcia-Leal,
Maria-Jose Del-Moral-Avila, Esteban Navarrete-Alvarez
and Maria-Jesus Rosales-Moreno

Dpto. Estadistica e I.O. - Facultad de Ciencias
Campus de Fuentenueva s/n, Granada, Spain,
quesada@ugr.es, juliagl@ugr.es, delmoral@ugr.es,
estebang@ugr.es, mrosales@ugr.es

Abstract. We discuss some survival models with the intensity process of the counting process having a multiplicative structure. The most commonly used model is the Cox multiplicative hazard model. This model can be extended in different ways. We propose an additive-multiplicative model, where some of the covariates act multiplicatively on the risk function and others do so additively.

Keywords: Survival analysis, counting process, martingales

References

AALEN, O.O. (1978): Nonparametric inference for a family of counting processes. *Ann. Statist. 6, 701-726.*

ANDERSEN, P.K., BORGAN, Ø, GILL, R.D. and KEIDING, N. (1993): *Statistical Models Based on Counting Processes.* Springer-Verlag New York.

COX, D.R. (1972): Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B 34, 187-220.*

KRAUS, D. (2004): Goodness-of-fit inference for the Cox-Aalen additive-multiplicative regression model. *Statistics & Probability Letters 70, 285-298.*

LIN, D.Y. and YING, Z. (1995): Semiparametric analysis of general additive-multiplicative hazard models for counting processes. *Ann. Statist. 23, 1712-1734.*

MARTINUSSEN, T. and SCHEIKE, T.H. (2006): *Dynamic Regression Models for Survival Data.* Springer. New York.

QUESADA-RUBIO, J.M., GARCIA-LEAL, J., LARA-PORRAS, A.M. and NAVARRETE-ALVAREZ, E. (2001): An Additive Intensity Model in a Multivariate Process Counting. *Revista de Estatística. Portugal. Volume II, 2 Quadrimestre, 329-330.*

QUESADA-RUBIO, J.M. (2002): *Aportaciones en Análisis de Supervivencia.* PhD thesis, Univ. of Granada.

SCHEIKE, T.H. and ZHANG, M. (2002): An additive-multiplicative Cox-Aalen regression model. *Scand. J. Statist. 29, 75-88.*

SCHEIKE, T.H. and ZHANG, M. (2003): Extensions and Applicationss of the Cox-Aalen Survival Model. *Biometrics 59, 1036-1045.*

Assessment of Scoring Models Using Information Value

Jan Koláček[1] and Martin Řezáč[1]

Department of Mathematics and Statistics, Masaryk University
Kotlářská 2, 611 37 Brno, Czech Republic, *kolacek@math.muni.cz*

Abstract. It is impossible to use a scoring model effectively without knowing how good it is. Quality indexes like Gini, Kolmogorov-Smirnov statistics and Information value are therefore used to assess quality of given scoring model.

The paper deals mainly with Information value. Commonly it is computed by discretisation of data into bins using deciles. One constraint is required to be met in this case. Number of cases have to be nonzero for all bins. If this constraint is not fulfilled there are numerous practical procedures for preserving finite results. As an alternative method to empirical estimates we can use the kernel smoothing theory.

Keywords: Credit scoring, Quality indexes, Information value, Quantiles, Kernel smoothing

References

ANDERSON, R. (2007): *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press, Oxford.

SIDDIQI, N. (2006): *Credit Risk Scorecards: developing and implementing intelligent credit scoring*. Wiley, New Jersey.

TERRELL, G.R. (1990): The Maximal Smoothing Principle in Density Estimation. *Journal of the American Statistical Association 85, 470-477*.

THOMAS, L.C. (2009): *Consumer Credit Models: Pricing, Profit, and Portfolio*. Oxford University Press, Oxford.

THOMAS, L.C., EDELMAN, D.B., CROOK, J.N. (2002): *Credit Scoring and Its Applications*. SIAM Monographs on Mathematical Modeling and Computation, Philadelphia.

WAND, M.P. and JONES, M.C. (1995): *Kernel smoothing*. Chapman and Hall, London.

A stochastic Gamma diffusion model with threshold parameter. Computational statistical aspects and application

R. Gutiérrez[1], R. Gutiérrez-Sánchez[1], A. Nafidi[2], and E. Ramos-Ábalos[1]

[1] Department of Statistics and Operational Research, University of Granada, Faculty of Sciences, Campus de Fuentenueva
18071 Granada, Spain, *ramosa@ugr.es*

[2] Ecole Supérieure de Techonologie de Berrechid, Université Hassan 1er, Quartier Tagadom, Passage d'Alger
B.P: 218, Berrechid, Maroc

Abstract. In this paper, we propose a new study of a stochastic gamma diffusion process, with threshold parameter, which can be considered as an extension of the gamma diffusion process. The estimation of the threshold parameter requires the solution of a nonlinear equation. To do so, we propose the classical Newton-Raphson method. This methodology is applied to an example with simulated data.

Keywords: Discrete sampling, Statistical inference in diffusion process, Application

References

BIBBY, B.M. and SORENSEN, M. (1995): Martingale estimation functions for discretely observed diffusion processes. *Bernoulli 1(1/2), 17–39.*

EUGENE, M.C. (2000), Maximum likelihood estimations of a class of one-dimensional stochastic differential equation models from discrete data. *Journal of Time Series Analysis 22(5), 505–515.*

GUTIÉRREZ, R., GUTIÉRREZ-SÁNCHEZ, R., NAFIDI, A. and RAMOS, E. (2006): A new stochastic Gompertz diffusion process with threshold parameter: Computational aspects and applications. *Appl. Math. Comput. 183, 738–747.*

GUTIÉRREZ, R., GUTIÉRREZ-SÁNCHEZ, R., NAFIDI, A. and RAMOS, E. (2009): Three-parameter stochastic lognormal diffusion model: statistical computation and simulating annealing. Application to real case. *J. Stat. Comput. Simul. 79(1), 25–38.*

GUTIÉRREZ, R., GUTIÉRREZ-SÁNCHEZ, R. and NAFIDI, A. (2009): The trend of the total stocks of the private car-petrol in Spain: Stochastic modelling using a new Gamma Diffusion Process. *Appl. Energy 86, 18–24.*

LIPTER, RS. and SHIRYAYEV, AN (1978): *Statistics of Random Processes II. Applications.* Springer-Verlag, New York.

PRAKASA RAO, B.S.L. (1999): *Statistical inference for diffusion type process.* Arnold, London and Oxford University Press, New York, 1999.

WONG, E. and HAJEK, B. (2008): *Stochastics processes in engineering systems.* New York, 1985.

ZEHNA, P.W. (1966): Invariance of maximum likelihood estimators. *Ann. Math. Stat. 37, 744.*

Clustering of Waveforms-Data Based on FPCA Direction

Giada Adelfio[1], Marcello Chiodi[1], Antonino D'Alessandro[2] and Dario Luzio[3]

[1] Dip. di Scienze Statistiche e Matematiche "S. Vianelli", University of Palermo, Italy, *adelfio@unipa.it, chiodi@unipa.it*
[2] Centro Nazionale Terremoti OBS Lab. Gibilmanna, INGV, Italy, *antonino.dalessandro@ingv.it*
[3] Dip. di Chimica e Fisica della Terra, University of Palermo, Italy, *luzio@unipa.it*

Abstract. The necessity of finding similar features of waveforms data recorded for earthquakes at different time instants is here considered, since eventual similarity between these functions could suggest similar behavior of the source process of the corresponding earthquakes. In this paper we develop a clustering algorithm for curves based on directions defined by an application of PCA to functional data.

Keywords: FPCA, clustering of curves, waveforms

Maximum Margin Learning of Gaussian Mixture Models with Application to Multipitch Tracking

Franz Pernkopf and Michael Wohlmayr

Signal Processing and Speech Communication Laboratory (SPSC)
Graz University of Technology, Austria.
pernkopf@tugraz.at
michael.wohlmayr@tugraz.at

Abstract. We present a maximum-margin based learning algorithm for Gaussian mixture models. In contrast to existing methods, our approach includes the sum-to-one constraint of probabilistic models. Model parameters are optimized by maximizing the margin between training samples of distinct classes. Optimization is based on the extended Baum-Welch procedure, which attains a local maximum of the proposed optimization criterion. We apply the proposed algorithm to the task of multipitch tracking given single-channel recordings of two simultaneously speaking subjects. Using the mixture-maximization interaction model, we are able to combine classifiers trained on single speakers to classify the mixture of both speakers. We demonstrate the superior performance over generative training based on the expectation maximization algorithm under low-noise conditions.

Keywords: Gaussian mixture model, discriminative classifiers, extended Baum Welch, maximum margin learning.

Estimation of the Bivariate Distribution Function for Censored Gap Times

Luís Meira-Machado[1] and Ana Moreira[1]

Department of Mathematics and Applications, University of Minho
4800-058 Azurém, Guimarães, Portugal, *lmachado@math.uminho.pt*

Abstract. In many medical studies, patients may experience several events. The times between consecutive events (gap times) are often of interest and lead to problems that have received much attention recently. In this work we consider the estimation of the bivariate distribution function for censored gap times. Some related problems such as the estimation of the marginal distribution of the second gap time is also discussed. These issues were investigated, among others, by Lin et al. (1999) and de Uña-Álvarez, Meira-Machado (2008) and de Uña-Álvarez and Amorim (2009). In this paper we introduce a nonparametric estimator of the bivariate distribution function based on Bayes' theorem and Kaplan-Meier survival function. In addition we explore the behavior of the estimators through simulations.

Keywords: bivariate censoring, Kaplan-Meier, nonparametric estimation

References

UÑA-ÁLVAREZ, J. and MEIRA-MACHADO, L. (2008): A simple estimator of the bivariate distribution function for censored gap times. *Statistics and Probability Letters 78, 2440-2445.*

UÑA-ÁLVAREZ, J. and AMORIM, A.P. (2009) A semiparametric estimator of the bivariate distribution function for censored gap times. Discussion Papers in Stats OR, Report 09/03. Dept. Estadstica e IO, U. Vigo, Spain.

LIN, D.Y., SUN, W. and YING, Z. (1999): Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika 86, 59-70.*

Two Measures of Dissimilarity for the Dendrogram Multi-Class SVM Model

Rafael Pino Mejías[1] and María Dolores Cubiles de la Vega[1]

[1] Departamento de Estadística e I.O. Avda. Reina Mercedes s/n, Sevilla, Spain, *rafaelp@us.es, cubiles@us.es*

Abstract. Several schemes for multi-class problems have been proposed. One of these approaches, the Dendrogram-based SVM model, builds a set of binary SVM models, arising from a hierarchical cluster analysis of the set of classes, where the matrix of dissimilarities between the classes is obtained by calculating the distances between the gravity centers. However, these vectors are not good representatives of their associated samples and other measures of dissimilarity could be more appropriate. We propose two measures, the first is based on the distance between matrices containing a set of sample quantiles, while the second one computes a distance between the empiric characteristic functions of the samples associated to the considered classes.

Parcellation Schemes and Statistical Tests to Detect Active Regions on the Cortical Surface

Bertrand Thirion[1,2], Alan Tucholka[1,2], and Jean-Baptiste Poline[1,2]

[1] Parietal team, INRIA Saclay-le-de-France, Saclay, France
CEA Saclay, Bâtiment 145, 91191, Gif-sur-Yvette, France
bertrand.thirion@inria.fr,
[2] CEA, DSV, I^2BM, Neurospin,
CEA Saclay, Bâtiment 145, 91191, Gif-sur-Yvette, France

Abstract. Activation detection in functional Magnetic Resonance Imaging (fMRI) datasets is usually performed by thresholding activation maps in the brain volume or, better, on the cortical surface. However, basing the analysis on a site-by-site statistical decision may be detrimental both to the interpretation of the results and to the sensitivity of the analysis, because a perfect point-to-point correspondence of brain surfaces from multiple subjects cannot be guaranteed in practice. In this paper, we propose a new approach that first defines anatomical regions such as cortical gyri outlined on the cortical surface, and then segments these regions into functionally homogeneous structures using a parcellation procedure that includes an explicit between-subject variability model, i.e. random effects. We show that random effects inference can be performed in this framework. Our procedure allows an exact control of the specificity using permutation techniques, and we show that the sensitivity of this approach is higher than the sensitivity of voxel- or cluster-level random effects tests performed on the cortical surface.

Keywords: statistical testing, EM algorithm, spatial models, neuroimaging

Nonparametric Functional Methods for Electricity Demand and Price Forecasting

Juan Vilar[1], Germán Aneiros[2] and Ricardo Cao[3]

[1] Facultad de Informática, University of A Coruña
Campus de Elviña, A Coruña, Spain, *eijvilar@udc.es*
[2] University of A Coruña, Spain, *ganeiros@udc.es*
[3] University of A Coruña, Spain, *rcao@udc.es*

Abstract. Nowadays, the production and sale of electricity is traded under competitive rules in free markets. The agents involved in this market have a great interest in the study of electricity load and price. Since electricity cannot be stored, the demand must be satisfied instantaneously and system operators need to anticipate to future demands to avoid overproduction. Good forecasting of electricity demand is then very important. In the past, demand was predicted in centralized markets but competition has opened a new field of study. On the other hand, if system operators and consumers have reliable predictions of electricity price, they can develop their bidding strategies and establish a pool bidding technique to achieve a maximum benefit. Consequently, prediction of electricity demand and price are significant problems in this sector.

This work focuses on next day forecasting of electricity demand and price. Therefore, for each day of the week, 24 forecasts (of demand or price) need to be computed. For this, we propose to use functional nonparametric and semi-functional partial linear models to forecast electricity demand and price. The approach in this work uses functional data methods to take into account the daily seasonality of the electricity demand and price series. Nonparametric regression estimation methods are used to forecast these series (see Aneiros-Pérez *et al.* (2010)). The idea is to cut the observed time series into a sample of functional trajectories and to incorporate in the model just a single past functional observation (last day trajectory) rather than multiple past time series values. A functional version of the partial linear model has been considered in order to incorporate vector covariates in the forecasting procedure. As a consequence, this approach allows both to consider additional explanatory covariates and to use continuous past paths to predict future values. These two new forecasting functional methods are applied in short-term forecasting of electricity demand and price in the market of mainland Spain, in the period 2008-2009, and they are compared with a naive method and with seasonal ARIMA forecasts.

Keywords: electricity markets, functional data, time series forecasting

References

ANEIROS, G., CAO, R. and VILAR, J.M. (2010): Functional methods for time series prediction: a nonparametric approach, *Journal of Forecasting, to be published (DOI: 10.1002/for.1169).*

Functional ANOVA Starting from Discrete Data: An Application to Air Quality Data

Graciela Estévez-Pérez and Jose A. Vilar

Departamento de Matemáticas, Universidad de A Coruña
Campus da Zapateira, 15071 A Coruña, Spain
graci@udc.es, *jose.vilarf@udc.es*

Abstract. A nonparametric functional approach is proposed to compare the mean functions of k samples of curves. In practice, functional data are usually collected in a discrete form and hence they must be pre-processed to obtain smooth curves that will be analyzed with functional data techniques. However, in the context of k-sample tests, the pre-processing step can have effects in terms of power reduction. This problem was studied by Hall and Van Keilegom (2007) in the particular case of testing whether two independent samples of functional data were generated from the same distribution. In the present work, a hypothesis test to check if k samples of curves come from populations with identical mean functions is developed following the ideas by Hall and Van Keilegom (2007).

Specifically, the procedure consists of the following steps: (i) the raw data are previously smoothed using local polynomial regression, (ii) estimates of the mean curves are obtained by averaging the smoothers in each group, (iii) a test statistic of Cramér-von Mises type based on the estimated mean curves is constructed, and (iv) the distribution of the statistic under the null is approximated using conventional residual-based bootstrap. The procedure is analyzed in detail and its asymptotic validity is established. Finally, the proposed method is applied to air quality data collected from several monitoring stations placed at different geographical locations in the Community of Madrid (Spain).

Keywords: Functional data, local-linear regression, bootstrap, equality test of functional means

References

HALL, P. and VAN KEILEGOM, I. (2007): Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica 17, 1511-1531.*

Presmoothed log-rank test

M. Amalia Jácome[1] and Ignacio López-de-Ullibarri[2]

[1] Facultade de Ciencias, Campus da Zapateira
Universidade da Coruña, A Coruña, Spain, *majacome@udc.es*
[2] Escuela Universitaria Politécnica, Campus de Serantes
Universidade da Coruña, Ferrol, Spain, *ilu@udc.es*

Abstract. The log-rank test is probably the most popular test for comparing two or more samples of right censored survival lifetimes, available in almost all statistical software packages. Proposed by Mantel (1966) and studied by many authors, it is based on the comparison of the Nelson-Aalen estimator (Nelson (1972) and Aalen (1978)) of the cumulative hazard functions $\Lambda_j, j = 1, \ldots, J$ of J populations.

The power of the test depends on the efficiency of the estimation of Λ_j. Recently, presmoothing techniques have been shown to give good results in estimation under censoring (see, v.g., Jácome and Cao (2007)). Presmoothing estimators are more efficient than the classical ones if the presmoothing bandwidth is suitably chosen. The proposed test, with the same asymptotic distribution as the classical one, compares the presmoothed estimator of Λ_j (Cao et al. (2005)) of the samples with each other.

The log-rank test emphasizes the tail of the distributions and it is optimal to detect differences when the J hazard rates are proportional to each other. However, when the curves differ in the earlier or latter part of the follow-up period, or even cross, a *weighted* log-rank test should be used. Presmoothing can also be applied, with promising results, in any of the great variety of weighted tests proposed in the literature.

Keywords: Cumulative hazard function, Log-rank test, Presmoothing

References

AALEN, O.O. (1978): Nonparametric inference for a family of counting processes. *The Annals of Statistics 6, 701-726.*

CAO, R., LOPEZ-DE-ULLIBARRI, I., JANSSEN, P. and VERAVERBEKE, N. (2005): Presmoothed Kaplan-Meier and Nelson-Aalen estimators. *Journal of Nonparametric Statistics 17, 31-56.*

JACOME, M.A. and CAO, R. (2007): Almost sure asymptotic representation for the presmoothed distribution and density estimators for censored data. *Statistics 41, 517-534.*

MANTEL, N. (1966): Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports 50, 163-170.*

NELSON, W. (1972): Theory and applications of hazard plotting for censored failure data. *Technometrics 14, 945-965.*

On the estimation in misspecified models using minimun ϕ divergence

M.D. Jiménez-Gamero[1], V. Alba-Fernández[2], R. Pino-Mejías[1], and J.L. Moreno-Rebollo[1]

[1] Dpt. Statistics and O.R., University of Sevilla, c/Tarfia, s/n, 41012 Sevilla, Spain. {*dolores,rafaelp,jlmoreno*}*@us.es*
[2] Dpt. Statistics and O.R., University of Jaén, Paraje Las Lagunillas s/n., 23071 Jaén, Spain. *mvalba@ujaen.es*

Abstract. This work studies the consequences of model misspecification for multinomial data when using minimum ϕ divergence or minimum disparity estimators to estimate the model parameters. These estimators are shown to converge to a well-defined limit. As an application of the results obtained, we consider the problem of testing goodness-of-fit to a given parametric family for multinomial data, using as test statistic a divergence between the observed frequencies and a estimation of the null model cell probabilities. In some previous simulation studies, it has been observed that the asymptotic approximation to the null distribution of the test statistics in this class is rather poor. As an alternative way to approximate this null distribution, we prove that the bootstrap consistently estimates it. We present a numerical example illustrating the convenience of the bootstrap approximation which, in spite of demanding more computing time, it is more accurate than the approximation yielded by the asymptotic null distribution.

Keywords: Minimum phi-divergence estimator; consistency; asymptotic normality; goodness-of-fit; bootstrap distribution estimator.
Acknowledgements
The research in this paper has been partially supported by grant MTM2008-00018 (Spain).

References

MORALES, D., PARDO, L. and VAJDA, I. (1995): Asymptotic divergence of estimates of discrete distributions, *J. Statist. Plann. Inference* 48, 347–369.
LINDSAY, B.G. (1994): Efficiency versus robustness: the case for minimum Helllinger distance and related methods. *Ann. Statist.* 22, 1081–1114.
SERFLING, R. (1980): *Approximation Theorems of Mathematical Statistics*. Wiley.
WHITE, H. (1982): Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1–25.

Implementation of Regression Models for Longitudinal Count Data through SAS

Gul Inan and Ozlem Ilk

Department of Statistics, Middle East Technical University, Ankara, Turkey, ginan@metu.edu.tr, oilk@metu.edu.tr

Abstract. The longitudinal feature of measurements and counting process of responses motivate the regression models to be developed for longitudinal count data (LCD) which take into account the phenomenon such as within-subject association and overdispersion. In this study, firstly, we restrict ourselves to the marginal model and generalized linear mixed model (GLMM) classes for LCD and review the Log-Log-Gamma marginalized multilevel model (MMM) (Griswold and Zeger (2004)), which combines the features of marginal models and GLMMs. After giving information about the considerable characteristics of these models, we reintroduce the popular epileptic seizures data. Due to the special features of these models, implementation of them requires more special attention. As a consequence, this leads us to use SAS GENMOD procedure for the marginal model, SAS GLIMMIX procedure for the GLMM, and SAS NLMIXED procedure, which has a high degree of model specification, for the Log-Log-Gamma MMM by the method of Nelson et al. (2006) and that of Liu and Yu (2008). Besides showing how these models are implemented through the epileptic seizure data via SAS procedures, a comprehensive comparison of the SAS procedures is also presented. Finally, we conclude the study with the discussion of the results obtained from the implementation of the models through epileptic seizures data.

Keywords: SAS GENMOD, SAS GLIMMIX, SAS NLMIXED

References

GRISWOLD, M.E. and ZEGER, S.L. (2004): On Marginalized Multilevel Models and their Computation. The Johns Hopkins University, Department of Biostatistics Working Papers.

LIU, L. and YU, Z. (2008): A likelihood reformulation method in non-normal random effects models. *Statistics in Medicine 27, 3105-3124.*

NELSON, K.P., LIPSITZ, S.R., FITZMAURICE, G.M., IBRAHIM, J., PARZEN, M.A and STRAWDERMAN, R. (2006): Use of the Probability Integral Transformation to Fit Nonlinear Mixed-Effects Models With Nonnormal Random Effects. *Journal of Computational & Graphical Statistics 15 (1), 39-57.*

Bayesian tomographic restoration of Ionospheric electron density using Markov Chain Monte Carlo techniques

Eman Khorsheed[1], Merrilee Hurn[2], and Chris Jennison[3]

[1] Department of Mathematics, University of Bahrain
P.O.Box 32038, Kingdom of Bahrain, *ekhurshid@sci.uob.bh*
[2] Department of Mathematical Sciences, University of Bath
Bath, BA2 7AY, UK, *M.A.Hurn@bath.ac.uk*
[3] Department of Mathematical Sciences, University of Bath
Bath, BA2 7AY, UK, *C.Jennison@bath.ac.uk*

Abstract. The ionized region of the Earth's atmosphere is known as the ionosphere. The ionosphere is not static because electron concentrations vary considerably with time, season and intensity of solar radiation. Due to the large concentration of free electrons the ionosphere affects all radio waves including those transmitted by the Navy Navigation Satellite System (NNSS) and the Global Positioning System (GPS), causing errors. Thus one of the most important physical parameters in the ionosphere is the electron density, and accurate knowledge of its spatial distribution is essential. In Ionospheric tomography the data are integrals of total electron density along many intersecting paths, and are usually collected from satellite-to-ground based receivers, Spencer et al. (2001). These data are inverted to reconstruct an image of electron density in the ionospheric plane under study. We propose a Bayesian approach to the inversion problem using spatial priors. To obtain inferences the Bayesian approach is accompanied with a special Markov Chain Monte Carlo algorithm that we developed. The algorithm is based on a principle components analysis of initial output.

Keywords: Bayesian modeling, Ionospheric Tomography, Markov Chain Monte Carlo, Principle Components, Inversion

References

SPENCER, P.S.J. and MITCHELL, C.N., (2001): Multi-instrument data analysis system. In: Proceedings of the International Beacon satellite Symposium. Boston, MA, 4-6.

Consistent biclustering by sparse singular value decomposition incorporating stability selection

Martin Sill and Axel Benner

Division of Biostatistics, German Cancer Research Center
Im Neuenheimer Feld 280, Heidelberg, Germany
m.sill@dkfz.de
benner@dkfz.de

Abstract. High-dimensional gene expression data arises in all fields of life science and is usually stored as two-way two-mode data matrix. Often interest lies in finding a set of genes that show a correlated gene expression within a subset of the samples. In order to find solutions to this two-way clustering problem a large number of so called biclustering approaches have been proposed. In an idealized case, e.g. assuming low noise and assuming that the gene expression matrix has a block-diagonal structure, each block displays a bicluster. Decomposing this matrix by singular value decomposition (SVD) results in matrix factorization where each singular vector pair is associated with a bicluster. Therefore many biclustering methods are strongly related to SVD.

Applying SVD to real data sets will result in singular vectors with many non zero elements. Recently, a sparse SVD method has been proposed and successfully applied to find reasonable biclusters in gene expression data. This regularized form of the SVD alternately fits penalized regression models to the singular vectors to obtain a sparse matrix decomposition. The sparsity of the singular vectors depends on the choice of the penalization parameters.

We propose to choose the right amount of penalization by incorporating a stability selection. The stability selection is a subsampling procedure that can be applied to penalized regression models to find stable variables and additionally offers the possibility of an error control. The performance of this sparse SVD method will be compared with other biclustering methods related to SVD.

Keywords: microarrays, biclustering, sparse SVD, penalized regression, stability selection

References

BUSYGIN, S., PROKOPYEV, O. and PARDALOS P.M. (2008): Biclustering in data mining *Computers & Operations Research (35) ,2964-2987*

LEE, M., SHEN, H., HUANG, J.Z. and MARRON, J.S. (2010): Biclustering via Sparse Singular Value Decomposition *Biometrics, DOI 10.1111/j.1541-0420.2010.01392.x*

MEINSHAUSEN, N. and BÜHLMANN, P. (2009): Stability Selection *Preprints of Journal of the Royal Statistical Society, To be published in Series B*

Comparison of Dimensionality Reduction Methods Used in Case of Ordinal Variables

Lukáš Sobíšek, Hana Řezanková, and Vanda Vilhanová

University of Economics, Prague, nám. W. Churchilla 4, 130 67 Praha 3. Czech Republic, *lukas.sobisek@yahoo.com* | *hana.rezankova@vse.cz* | *vandav@email.cz*

Abstract. Prior to application of some classification methods, it is desirable to reduce the number of variables characterizing individual objects (Lee and Verleysen, 2007). It is possible either to use latent variables created on the basis of original variables or to select variables which characterize the certain groups of similar variables. In the contribution we compare dimensionality reduction methods and methods for clustering variables. Our aim is to analyze data files with ordinal variables created on the basis of questionnaire surveys. We used a proximity matrix as an input for some analyses (multidimensional scaling and hierarchical cluster analysis). In this matrix, different association coefficients for ordinal variables are used (Spearman correlation coefficient, Kendall's tau-b, Kendall's tau-c, gamma, symmetric Somers' d). To identify groups of similar variables (including determination of cluster number) on the basis of results of dimensionality reduction methods, we interpret these results by fuzzy cluster analysis. The soft version of CSPA (cluster-based similarity partitioning algorithm) is applied for ensembles of fuzzy clustering results obtained on the basis of different techniques. For illustration, we analyze the real data files based on questionnaires surveys, e.g. perception of policemen by young people (survey from 2006, 24 variables characterizing a typical policeman and the same number of variables characterizing an ideal policeman, 356 respondents), active lifestyle of university students (survey from 2008, 15 variables expressing a satisfaction with different aspects of students' life, 1,453 respondents), males and females with university diploma (survey from 1998, 13 variables expressing a satisfaction with different aspects of job, 1,908 respondents). Respondents answers are coded from 1 to 7 (the first two files) and from 1 to 4 (the third file). For the analyses, SPSS, STATISTICA, S-PLUS and Latent GOLD systems are used.

Keywords: dimensionality reduction, ordinal variables, categorical principal component analysis, latent class modeling, cluster analysis

References

LEE, J. A. and VERLEYSEN, M. (2007): *Nonlinear Dimensionality Reduction*. Springer, New York.

Acknowledgement. This work was supported by projects GACR P202/10/0262 and IGA VSE F4/3/2010.

Implications of primary endpoint definitions in randomized clinical trials with time-to-event outcome

Martina Mittlböck and Harald Heinzl

Center for Medical Statistics, Informatics, and Intelligent Systems
Medical University of Vienna, Spitalgasse 23, Vienna, Austria
martina.mittlboeck@meduniwien.ac.at, *harald.heinzl@meduniwien.ac.at*

Abstract. In clinical studies with survival time as primary endpoint, its definition is not always straightforward and clear. E.g. in cancer studies, overall-survival, event-free survival or recurrence free survival may be suitable candidates among others. Definitions with a short expected study duration may be favored by medical researchers.
Consequences of the definition of the primary endpoint for randomized clinical trials with time-to-event outcome will be compared and discussed. Effects on the planning phase, analysis and interpretation of results are investigated. Pitfalls when comparing treatment effects between different studies are mentioned. Special attention will be given to the comparison of hazard ratios from studies with unequal proportions of non-disease related events.
Different choices of endpoints are not interchangeable, and sample size calculations and interpretations of resulting hazard ratios should be handled with care. Furthermore, hazard ratios among studies with unequal proportions of non-disease related events cannot be compared in a straightforward manner. Systematic reviews or meta-analyses of treatment comparisons become more complicated when either the primary endpoints differ across studies or when they are identical, but the hazards for the non-disease related events differ.

Keywords: time-to-event, survival, sample size calculation, systematic review

Estimation of Abilities by the Weighted Total Scores in IRT Models using R

Sayaka Arai[1] and Shin-ichi Mayekawa[2]

[1] The National Center for University Entrance Examinations
2-19-23, Komaba, Meguro-ku, Tokyo, Japan, *sayarai@rd.dnc.ac.jp*
[2] Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, Japan, *mayekawa@hum.titech.ac.jp*

Abstract. Under item response theory (IRT), the ability parameter for each individual is estimated, in most cases, by either the MLE or EAP method from the response patterns to the test items. However, since these estimation methods are complicated, it is almost impossible for a layman to comprehend how the score is calculated.

On the other hand, it is possible to estimate the IRT ability on the basis of the observed weighted total score (e.g. Thissen (2001)). The weighted total score is easy to calculate, but there are two major problems associated with this estimation method, namely, the choice of the weights and the calculation of the conditional distribution of the weighted total score given the ability. For the first problem, Mayekawa (2008) developed the globally optimal weights, which maximize the expected test information, and showed that the globally optimal weights reduced the posterior variance when evaluating the posterior distribution of the ability given the weighted total score. For the second problem, Mayekawa and Arai (2008) proposed an efficient algorithm for calculating the conditional distribution of the weighted total score of polytomous items.

In this study, we combine these two methods and propose a quick scoring method for estimation of IRT abilities based on the weighted total scores using R program. Given the set of item parameters and the weights, the program produces the correspondence table between the weighted total scores and the EAP ability estimates with the associated posterior standard deviation.

Keywords: item response theory, globally optimal weights, weighted total scores

References

MAYEKAWA, S. (2008): Estimation of ability using the globally optimal scoring weights. Paper presented at the International Meeting of the Psychometric Society, IMPS2008.

MAYEKAWA, S. and ARAI, S. (2008): Distribution of the Sum of Scored Multinomial Random Variables and Its Application to the Item Response Theory. In K. Shigemasu, A. Okada, T. Imaizumi, and T. Hoshino (Eds.): *New Trends in Psychometrics*. Tokyo: Universal Academic Press, 263–272.

THISSEN, D. and WAINER, H. (2001): *Test Scoring*. Lawrence Erlbaum Assoc, Inc., Publishers. Mahwah, New Jersey.

Association Rules Extraction from the Otolaryngology Discharge Notes

Basak Oguz[1], Ugur Bilge[1], M. Kemal Samur[1], and Filiz Isleyen[1]

The Department of Biostatistics and Medical Informatics, Akdeniz University Antalya, TURKEY, *basakoguz@akdeniz.edu.tr*

Abstract. The explosive growth of databases in almost every area of human activity has created a great demand for new, powerful tools for turning data into useful knowledge. Text mining is a tool that saw an increasing interest in the 2000s, for enabling people to find unknown information and facts from the free-text data (Konchady (2006)). Recently, the number of text mining applications in medical sciences has grown with an increasing rate. Unstructured free-text data, such as patient discharge notes and reports, doctor's notes, clinical trials and studies, research reports, web pages and hospital records are some of the important data sources for physicians. To analyze and access this kind of data by human efforts is difficult and time consuming. Considering the time it takes for decision making, and accesing accurate and required information about patients, this kind of systems have become necessary.

In this study, we developed a domain based software system to transform 600 discharge notes, from the Department of Otolaryngology of Akdeniz University, to a structured form, extracting clinical data from the discharge notes, and analyzing extracted data. First of all, discharge notes which are kept as Microsoft Office Word documents have been transformed into a data table after preprocessing (tokenization, correcting spelling errors, eliminating stop words,). In order to identify common section in the discharge notes, including patient history, age, problems, and diagnosis etc., several word lists have been constituted. Using the terms co-Occurrences within discharge notes, the keyword lists have been created. To identify the significant content words within each section keyword lists have been used and content words have been converted into a predefined coded structure. Association Rules, that is one of the methods of the traditional data mining, has been applied to extracted data in order to discover the patterns and relations between entities/concepts. There are two important basic measures for association rules, i.e. support and confidence. In addition to these measures, lift ratio has also been calculated to identify interesting rules. The system has been designed to visualize the extracted association rules in table format.

Keywords: otolaryngology, text mining, data mining, association rules

References

KONCHADY, M. (2006): *Text Mining Application Programming.* Charles River Media, Boston.

Bayesian nonparametric analysis of GARCH models

M. Concepcion Ausin[1], Pedro Galeano[2], and Pulak Ghosh[3]

[1] Department of Statistics, Universidad Carlos III de Madrid
Calle Madrid, 126 - 28903 Getafe, Madrid, Spain *causin@est-econ.uc3m.es*
[2] Department of Statistics, Universidad Carlos III de Madrid
Calle Madrid, 126 - 28903 Getafe, Madrid, Spain *pgaleano@est-econ.uc3m.es*
[3] Indian Institute of Management Bangalore
Bannerghatta Road, Bangalore, India *pulak.ghosh@iimb.ernet.in*

Abstract. Financial time series analysis deals with the understanding of data collected on financial markets. Investors and financial managers need to understand the behavior of asset prices to have good expectations about future prices and the risks to which they will be exposed. For that, the usual approach is to derive probability distributions of the future values which are especially useful because also provide with measures of investment risk. Several parametric models have been entertained for describing, estimating and predicting the dynamics of financial returns. Alternatively, this article considers a Bayesian semiparametric analysis of financial time series. In particular, the usual parametric distributional assumptions of the GARCH-type models are relaxed by entertaining the class to location-scale mixtures of Gaussian distributions with Dirichlet process mixture models on the mixing distribution. Although in different settings than considered here, Dirichlet process mixture models, by now, have an extensive literature in Bayesian analysis and provide a broad and flexible class of distributions. The proposed specification allows for a greater flexibility in capturing both the skewness and kurtosis frequently observed in financial returns. Also, the Bayesian methodology offers a natural way to introduce parameter uncertainty in estimation of in-sample volatilities and to obtain predictive distributions of future returns and volatilities. Value at Risk (VaR) has become the most widely used measure of market risk for practitioners. Statistically speaking, the VaR is the negative value of a quantile of the conditional distribution of the return series. The developed methodology offers a convenient specification of the return distribution, which is crucial to give an accurate estimate of the VaR. Moreover, it allows to obtain a reliable predictive distribution of the VaR, which also provides with a measure of precision for VaR estimates via predictive intervals.

Keywords: Bayesian estimation; Dirichlet process mixture; Financial time series; Location-scale Gaussian mixtures; Markov chain Monte Carlo.

Electricity Consumption and Economic Growth in Turkey: Time Series Analysis by Break Function Regression

Cherkez Agayeva[1], Goknur Yapakci[2], and Sel Ozcan[3]

[1] Yasar University, Department of Statistics
Bornova, Izmir, Turkey *agayeva.cherkez@yasar.edu.tr*
[2] Yasar University, Department of Statistics
Bornova, Izmir, Turkey *goknur.yapakci@yasar.edu.tr*
[3] Yasar University, Department of Industrial Engineering
Bornova, Izmir, Turkey *sel.ozcan@yasar.edu.tr*

Abstract. While determining the changes of trend, break function regression is the easiest model that can be used. Break function is defined as continuous and it consists of two linear segments. It is firstly introduced by Hinckley (1970) and Hinkley (1971) as "two-phase regression".Besides, Mudelsee (2000) offered a three-phase regression ("ramp") as a model for climate transitions. This paper is useful since "break" methodology and "ramp" methodology has several similarities. On the other hand, various studies identifying the relationship between two factors were carried out in the literature; for example, Yuan et al. (2007) investigated the relationship between electricity consumption and GDP in China. Furthermore, Mozumder and Marathe (2007), analyzed the relationship between electricity consumption and GDP in Bangladesh. In our study, we aimed to examine the relationship between electricity consumptions and economic growth using data from the Turkish market, and different than the previous studies, the trend was estimated by using the break function regression defined in Mudelsee (2009).

Keywords: Break function regression, time series, economic growth

References

HINKLEY, D., V. (1970): Inference about the change points in a sequence of random variables. *Biometrika 57, 1-17.*

HINKLEY, D., V. (1971): Inference about the change-point from cumulative sum tests. *Biometrika 58, 509-523.*

MOZUMDER, P., MARATHE, A. (2007): Causality relationship between electricity consumption and GDP in Bangladesh. *Energy Policy 35, 395-402.*

MUDELSEE, M. (2000): Ramp function regression: a tool for quantifying climate transitions. *Computers and Geosciences 26, 293-307.*

MUDELSEE, M. (2009): Break function regression: A tool for quantifying trend changes in climate time series. *The European Physical Journal, Special Topics 174, 49-63.*

YUAN, J. et. al. (2007): Electricity consumption and economic growth in China: cointegration and co-feature analysis. *Energy Economics 29, 1179-1191.*

Tests for Abnormal Returns under Weak Dependence

Niklas Ahlgren[1] and Jan Antell[2]

[1] Hanken School of Economics
PO Box 479 (Arkadiagatan 22), 00101 Helsingfors, Finland,
niklas.ahlgren@hanken.fi
[2] Hanken School of Economics
PO Box 479 (Arkadiagatan 22), 00101 Helsingfors, Finland, *jan.antell@hanken.fi*

Abstract. In event studies (Brown and Warner (1985)), abnormal returns are assumed to be cross-sectionally independent. If the event day is common, and if the firms are from the same industry, returns are usually positively correlated. We propose to use a spatial error model (see e.g. LeSage and Pace (2009))

$$u = (I_n - \rho W)^{-1}\varepsilon, \quad |\rho| < 1, \quad \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)', \quad \varepsilon \sim IID(0, \sigma_\varepsilon^2 I_n),$$

where ρ measures the strength of dependence and W is a spatial weights matrix, to model cross-sectional dependence in returns. Returns of firms belonging to the same sector or industry are correlated, but returns of firms belonging to different sectors or industries are uncorrelated. The spatial error model formalises weak dependence.

The performance of tests for event effects under cross-sectional dependence are evaluated by simulation. We find that moderate spatial autocorrelation causes overrejection of the null hypothesis of no event effect.

We derive a correction to tests for abnormal returns, which takes into account cross-sectional dependence in returns. We apply the corrected tests to US stock returns.

Keywords: abnormal returns, cross-sectional dependence, event studies, spatial error model

References

BROWN, S. J. and WARNER, J. B. (1985): Using daily stock returns: The case of event studies. *Journal of Financial Economics 14 (1), 3-31.*

LESAGE, J. and PACE, R. K. (2009): *Introduction to Spatial Econometrics.* CRC Press, Boca Raton.

Penalized Splines and Fractional Polynomials for Flexible Modelling the Effect of Continuous Predictor Variables: A Systematic, Simulation-Based Comparison

Alexander M. Strasak*, Nikolaus Umlauf**, Ruth M. Pfeiffer***, Stefan Lang**

* *Innsbruck Medical University, Schöpfstr. 41, A-6020 Innsbruck, Austria*
** *University of Innsbruck, Universitätsstr. 15, A-6020 Innsbruck, Austria*
*** *National Cancer Institute, 6120 Executive Blvd, Bethesda, MD, 20892-7244, USA*

Abstract

P(enalized)-splines and fractional polynomials (FP) have emerged as powerful smoothing techniques with increasing popularity in several fields of applied research. Both approaches allow for considerable flexibility, but not much is known about how these methods compare to each other in a given setting and it may not be obvious for the practitioner which one to use best. A systematic mainly simulation-based comparison of P-splines and FP is not available to date. We performed extensive simulation analyses using standard implementations in STATA, R and BayesX to compare FP´s of degree 2 and 4 (FP2, FP4) and P-splines, using generalized cross validation (GCV) and restricted maximum likelihood (REML) for smoothing parameter selection. Overall, we found that spline-based estimators (REML, GCV) and FP4 perform equally well in most simulation settings. However, for more curved functions FP2, as current default implementation in standard software, showed considerably bias and consistently higher mean squared error (MSE) compared to all other estimators. Moreover, FP's in general are prone to artefacts due to the specific choice of the origin, while P-splines based on GCV reveal sometimes wiggly estimates in particular for small sample sizes. Finally an application on undernutrition in India highlights the specific features of the approaches. The application shows that FP's, in contrary to P-splines, are not able to capture all features of very complex functions.

Keywords: generalized additive models, fractional polynomial, penalized spline, simulation, smoothing

Which Objective Measure can Mimic Experts Opinion for Quality of Dermatologic Images?

Filiz Isleyen[1], Ayse Akman[2], Kemal H. Gulkesen[1], Yilmaz K. Yuce[1], Anil A. Samur[1], and Erkan Alpsoy[2]

[1] Departments of Biostatistics and Medical Informatics, Akdeniz University Antalya, Turkey, *ifiliz@akdeniz.edu.tr*
[2] Departments of Dermatology, Akdeniz University, Antalya, Turkey

Abstract. Since the image quality (IQ) has critical importance in medical imaging, the reliability of image compression algorithms is a major issue to address. Compressed IQ can be evaluated objectively and subjectively. In subjective studies, experts evaluate the quality of images. There are many objective measures, which are defined by mathematical definitions, such as Mean Square Error (MSE), Mean Average Error (MAE), Peak Signal-to-Noise Ratio (PSNR), Maximum Difference (MD), Structural Content (SC), Normalized Absolute Error (NAE) and Image Fidelity (IF) (Przelaskowski (2004), Grgic et al (2002)). However, presence of several objective measures for IQ in compressed images raises the problem of selecting the appropriate measure. Our aim is to try to compare the objective methods for their ability to mimic human experts. The image set is taken from the study done by Gulkesen et al (2009). In this study, the objective evaluation results were compared with that studys subjective results using Receiver Operating Characteristic (ROC) analysis. For objective evaluation, the images were compared with each other for all objective measures and at four different Compression Ratio (CR) (1:50, 1:40, 1:30 and 1:20). A statistically significant difference was found that JPEG2000 (JP2) provided better results for all evaluation criteria at each CR ($p<0.001$). NAE and SC showed a lower performance relative to other measures for both JPEG and JP2. MD is the closest to experts opinion for JPEG, MSE is the closest to experts opinion for JP2, but the difference between the methods were statistically insignificant. NAE and SC appear to be less successful in reflecting experts opinion.

Keywords: Image compression, JPEG, JPEG2000, Quality measures

References

GRGIC, S., GIRGIC, M., MRAK, M. (2002): Reliability of objective picture quality measures. *Journal of Electrical Engineering 55 (1-2), 3-10.*

GULKESEN, K.H., AKMAN, A., YUCE, Y.K., YILMAZ, E., SAMUR, A.A., ISLEYEN, F., CAKCAK, D., ALPSOY, E. (2009): Evaluation of JPEG and JPEG2000 compression algorithms for dermatological images. *Journal of the European Academy of Dermatology and Venereology, DOI: 10.1111/j.1468-3083.2009.03538.x.*

PRZELASKOWSKI, A. (2004): Vector quality measure of lossy compressed medical images. *Computers in Biology and Medicine 34, 193-207.*

An Application of the Poisson Regression on Infertility Treatment Data

Anil Aktas Samur[1], Osman Saka[1], and Murat Inel[2]

[1] Department of Biostatistics and Medical Informatics, Akdeniz University Antalya, Turkey, *anilaktas@akdeniz.edu.tr*
[2] Department of Obstetrics and Gynecology, Akdeniz University Antalya, Turkey, *muratinel@hotmail.com.tr*

Abstract. Regression models are the most popular tools for modeling the relationship between a response variable and a set of predictors. In many applications, the response variable of interest is a count, i.e. takes on non-negative integer values. For count data, the most widely used regression model is Poisson regression while the logistic (or probit) regression is often applied for binary data (Sellers (2008)). Many outcomes in clinical medicine are a finite set of non-negative integer values that are not normally distributed; hence, they may be more appropriately analyzed as discrete rather than as continuous measures (Byers et al. (2003)). In the analysis of categorical or count data, transformation techniques those are used to ensure the assumptions of normality may be inadequate in most cases. Hence, Poisson regression which is based on exponential family can be used. Poisson regression is accounted to understand the correlation between explanatory variables and count data type response variables. In this regression model, the link function which associates the linear structure of the explanatory variable(s) to the expected values of the response variable(s) is found by the logarithmic transformation (Frome (1983)). In our study, we applied a Poisson regression model to an infertility data set that consists of 1,401 women who enrolled in infertility treatment between 2000 and 2008 at Department of Obstetrics and Gynecology, Akdeniz University, Turkey. The number of pregnancies was assigned as dependent variable while age (grouped in five categories) and the duration of treatment (grouped in two categories) were the independent variables. The outcomes show that both age and duration of treatment had significant effect on number of pregnancies.

Keywords: Count data, Poisson regression, infertility

References

BYERS, AL., ALLORE, H., GILL, TM., PEDUZZI, PN. (2003): Application of negative binomial modeling for discrete outcomes: A case study in aging research. *Journal of Clinical Epidemiology 56(6), 559-564.*

FROME, EL., (1983): The Analysis of Rates Using Poisson Regression Models. *Biometrics 39(3), 665-674.*

SELLERS, KF., (2008): A Flexible Regression Model for Count Data. *Working Paper No. RHS-06-060, Robert H. Smith School of Business University of Maryland College Park, MD.*

Application of Particle Swarm Approach to Copula Models Involving Large Numbers of Parameters

Enrico Foscolo[1] and Matteo Borrotti[12]

[1] Department of Statistical Science "Paolo Fortunati", University of Bologna
Via delle Belle Arti 41, 40126 Bologna, Italy
{*enrico.foscolo2, matteo.borrotti*}*@unibo.it*
[2] European Centre for Living Technology (ECLT), University of Ca' Foscari
S. Marco 2940, 30129 Venice, Italy

Abstract. One-parameter multidimensional copulas represent a standard tool for non-linear dependence structures between random variables and joint extreme events. Nevertheless, non-linear systems require flexible models with an increasing numbers of parameters and different margins. Therefore, the issue of parameters estimation plays a central role in the study of joint behavior of the variables.

The celebrated pseudo-maximum likelihood estimator of the copula parameters investigated by Genest et al. (1995) requires an initial vector of values in the optimization procedure. Estimates based on Kendall's tau or Spearman's rho are usually considered to this aim. Unfortunately, closed-form relations between the copula parameters and the measures of association are not always available. In order to avoid additional numerical computation, we propose the Particle Swarm Optimization (PSO, Kennedy and Eberhart (1995)), based on the social behavior reflected in swarm of birds, called particles.

Starting from a finite randomly chosen set of values of the parameters, we evaluate the pseudo-likelihood function associated with the copula model and we use the PSO to find the region where the pseudo-likelihood function is maximized. PSO is compared with other optimization algorithms, *i.e.* gradient-based algorithm.

Preliminary studies carried out with a three-parameters three-dimensional copulas show that our approach maximizes the pseudo-likelihood function with reasonable computational efforts. Moreover the proposed method provides a more robust alternatives to the procedure involving Kendall's tau or Spearman's rho.

Keywords: particle swarm optimization, copula models, pseudo-likelihood

References

GENEST, C., GHOUDI, K. and RIVEST, L.-P. (1995): A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika 82 (3), 543-552.*

KENNEDY, J. and EBERHART, R. (1995): Particle swarm optimization. In: *IEEE international conference on neural networks.* Piscataway, New York, 1942–1948.

Algorithm of Sequential Assimilation of Observational Data in Problem of Kalman Filtration

Yuri N. Skiba

Centro de Ciencias de la Atmósfera, UNAM University
Av. Universidad 3000, CU, México D.F., México, *skiba@servidor.unam.mx*

Abstract. A new recursion method of assimilation of measurements in the problem of Kalman filtration is proposed. The method assimilates sequentially (one by one) all the measurements which enter into the linear stochastic dynamic system at a given discrete moment of time. It is proved that if the dimension of the vector of observations is equal to r, then for r steps method exactly realizes the standard algorithm of the optimum evaluation of the state of physical system in Kalman filter (Brammer and Siffling (1975), Leondes (1978)). The merit of new algorithm is the direct method of realization, which does not require the application of approximate iterative procedures for the inversion of matrix MPM^T+W, entering the equations of Kalman filter.

The method is economical and convenient in practice. In the case when the dimension of the vector of observations depends on time, and hence the matrix order varies in time, the application of the recursion method makes it possible to avoid the use of dynamic arrays in computer programs and thus to prevent undesirable memory fragmentation and excessive use of computer resources. The proposed algorithm can also be applied in problems of optimal interpolation.

Keywords: Kalman filtering, sequential data assimilation

References

BRAMMER, K. and SIFFLING, G. (1975): *Kalman-Bucy-Filter, Deterministische Beobachtung und stochastische Filterung*. R. Oldenbourg Verlag, München.

LEONDES, K. (1978): *Filtering and Stochastic Control in Dynamic Systems*. McGraw-Hill, New York.

Computational Methods for Fitting the Lee-Carter Model of Turkish Mortality Change

Banu Ozgurel

Banu Ozgurel , Yasar University
Bornova, Izmir, Turkey, *banu.ozgurel@yasar.edu.tr*

Abstract. Lee-Carter method is used to forecasting and modeling mortality. A non-linear model for fitting and forecasting age-specific mortality rates at age x and time t is proposed in 1992 by Lee and Carter. This method included weighted least squares (WLS), maksimum likelihood estimation (MLE), and singular value decomposition (SVD) method and time series. Time series methods are used to make long-run forecasts, with confidence intervals, of age-specific mortality. In this study, the Lee-Carter model is developed for Turkey. This study describes the technical procedures to fit and extrapolate these models and includes comparisons of results of these models. It is used to fit mortality data in Turkey from 1960-2006 and to forecast mortality change from 2007 to 2030. Mortality data are obtained by TUIK (Turkish Statistical Institute).

Keywords: Lee-Carter Method, Weighted least squares (WLS), Maksimum likelihood estimation (MLE), Singular value decomposition (SVD), Time series

References

BENJAMIN B., POLLARD J.H. (1993): The Analysis of Mortality and Other Actuarial Statistics., Institute of Actuaries and Faculty of Actuaries.

LEE R. (2000): The Lee-Carter Method For Forecasting Mortality, With Various Extensions and Applications. *North American Actuarial Journal, Vol. 4, No. 1.*

WANG D., LU P. (2005): Modelling and Forecasting Mortality Distributions in England and Wales Using the Lee-Carter Model. *Journal of Applied Statistics Vol. 32, No. 9, p. 873-885.*

Application of Artificial Neural Network and Logistic Regressions on the Data Obtained from Pediatric Endocrinology Information System to Predict Familial Short Stature

Mehmet Kemal Samur[1], Ugur Bilge[1], Anil Aktas Samur[1], and Ozgur Tosun[1]

The Department of Biostatistics and Medical Informatics, Akdeniz University Antalya, Turkey, *samur@akdeniz.edu.tr*

Abstract. Pediatric Endocrinology is a scientific discipline specializing in endocrinology disorders those affect childrens physical and sexual development, such as growth disorder. As Evidence Based Medicine (EBM) is increasingly gaining importance, recent studies show that EBM can provide a higher quality of care with reduced costs, using new technologies, with an ethical and a person-centered approach (LEWIS (2004), MALUF-FILHO(2009)). Two of the most commonly used computer based EBM methods are artificial neural network (ANN) and logistic regression (LR). In this study, we used a dataset captured and stored by the Pediatric Endocrinology Clinical Information System. We compared LR and ANN results to automatically identify familial short stature disorder. Whole data is divided into training and test datasets; from the training dataset we obtained models from LR and ANN. Afterwards, LR and ANN were applied to the test dataset. In training dataset, LR achieves a sensitivity of 0.795, specificity of 0.662, and ROC AUC of 0.813. Using the cut-off points obtained from this analysis; a sensitivity of 0.771, specificity of 0.625, and ROC AUC of 0.766 were found with the test dataset. On the other hand, ANN achieves a sensitivity of 0.832, specificity of 0.683, and ROC AUC of 0.811 on training dataset. With the obtained cut-off points, the performance in the test data was: sensitivity of 0.705, specificity of 0.826, and ROC AUC of 0.832.When results from LR and ANN on the training and test datasets are compared, both techniques showed similar performances on both sets. LR showed similar results with its own training dataset performance while ANNs sensitivity decreased and specificity increased compared to its performance in training set. None of the differences were statistically significant.

Keywords: Artificial Neural Network, Clinical Information System, Familial Short Stature, Logistic Regression, Pediatric Endocrinology

References

LEWIS, S.J. and ORLAND B.I. (2004): The importance and impact of evidence-based medicine. *J Manag Care Pharm 10(5 Suppl A), 3-5.*

MALUF-FILHO, F. (2009): The importance of evidence-based medicine concepts for the clinical practitioner. *Arq Gastroentero 46(2), 87-89.*

On estimation and influence diagnostics for Student-t semiparametric linear models

Germán Ibacache-Pulgar[1] and Gilberto A. Paula[2]

[1] Instituto de Matemática e Estatística, USP, Brazil, *germanp@ime.usp.br*
[2] Instituto de Matemática e Estatística, USP, Caixa Postal 66281 (Agência Cidade de São Paulo), 05314-970 São Paulo, Brazil, *giapaula@ime.usp.br*

Abstract. Semiparametric or partial linear models (PLMs) have become an important tool in modeling economic and biometric data, and are considered a flexible generalization of the linear model by including a nonparametric component of some covariates. Such models assume that the relationship between the response variable and the explanatory variables can be represented as (see, for instance, Green (1987))

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + f(\mathrm{t}_i) + \epsilon_i, \qquad i = 1, \ldots, n \ , \tag{1}$$

where y_i denotes the response from the experiment, $\mathbf{x}_i$ is a $(p \times 1)$ vector of explanatory variables, $\boldsymbol{\beta}$ is a $(p \times 1)$ parameter vector, t_i is a scalar, f is a smooth function, and $\epsilon_i \sim N(0, \phi)$. In this work we extend partial linear models with normal errors to Student-t errors. Penalized likelihood equations are applied to derive the maximum likelihood estimates which appear to be robust against outlying observations in the sense of the Mahalanobis distance. An iterative process based on the back-fitting algorithm to adjust the model is proposed. The iterative process that we present is analytically simple and easy to implement computationally, and can be generalized to the class of elliptical models. We present some simulations to evaluate the performance of the iterative process. Finally, in order to study the sensitivity of the penalized estimates under some usual perturbation schemes in the model or data, the local influence curvatures are derived and some diagnostic graphics are proposed; see, for instance, Cook (1986) and Zhu et al. (2003). A motivating example preliminary analyzed under normal errors is reanalyzed under Student-t errors. The local influence approach is used to compare the sensitivity of the model estimates.

Keywords: Student-t distribution, Semiparametric models, Penalized maximum likelihood estimates, Robust estimates, Sensitivity analysis

References

COOK, R. D. (1986): Assessment of local influence (with discussion). *Journal of the Royal Statistical Society B 48, 133-169.*

GREEN, P. J. (1987): Penalized likelihood for general semi-parametric regression models. *International Statistical Review 55, 245-259.*

ZHU, Z., HE, X. and FUNG, W. (2003): Local influence analysis for penalized gaussian likelihood estimators in partially linear models. *Scandinavian Journal of Statistics 30, 767-780.*

Dependence Analysis of Gas Flow at Nodes within Gas Transportation Networks

Radoslava Mirkov[1], Herwig Friedl[2], Isabel-Wegner Specht[1], and Werner Römisch[1]

[1] Humboldt Universität zu Berlin, Department of Mathematics, Unter den Linden 6, 10999 Berlin, Germany, *mirkov@math.hu-berlin.de*
[2] Graz University of Technology, Institute of Statistics, Münzgrabenstraße 11, 8010 Graz, Austria, *HFriedl@TUGraz.at*

Abstract. The flow of natural gas within a gas transmission network is studied with the aim to optimize such networks. The analysis of real data provides a deeper insight into the behavior of gas in- and outflow. Different nonlinear regression models are fitted to describe dependence between the maximal daily gas flow and the temperature on network exits. Based on the residuals of the fitted model we study the dependence structure of the gas flow at nodes within the network, and use the results for the optimization of the network capacity.

Keywords: dependence analysis, correlation, gas flow, optimization

References

CERBE, G. (2008): *Grundlagen der Gastechnik*. Hanser Verlag.

Cooperation Agreement (2008): *Vereinbarung über die Kooperation gemäß* §20 *Absatz* 1 *b) EnWG zwischen den Betreibern von in Deutschland gelegenen Gasversorgungsnetzen*. BMJ Deutschland.

LEISCH, F. (2004): FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *J*ournal of Statistical Software, 11, 1-18.

RITZ, C., and STREIBIG, J.C. (2008): *N*onlinear Regression with R. Springer.

SEBER, G.A.F., and WILD, C.J. (2003): *N*onlinear Regression. John Wiley & Sons.

Right-censored Survival Analysis of Data with an Indefinite Initial Time Point

Tatsunami S.[1], Ueno T.[1], Kuwabara R[1]., Mimaya J.[2], Shirahata A.[2] and Taki M.[2].

[1] Unit of Medical Statistics and Institute of Radioisotope Research, St. Ma-rianna University School of Medicine, 2-16-1 Sugao, Miyamae-ku, Kawa-saki, Japan, *s2tatsu@marianna-u.ac.jp*

[2] The Research Committee for the National Surveillance on Coagulation Disorders in Japan

Abstract. Several types of censoring methods have been used in medical research as described by Harezlak and Tu (2006). For example, the last seronegative date or the first seropositive date for a specific antibody, or the midpoint between them is sometimes used. We deal with similar but different data in order to estimate the onset rate of critical hepatic disease after infection with hepatitis C virus (HCV). Among adult people with HCV infection, the first seropotivite date does not reflect the time point of infection, because testing for HCV antibody became available only after 1989. However, in the case of hemophiliac patients, the most likely cause of infection is via injection of coagulation factors which is a therapeutic treatment. The treatment is started just after birth of the patient as described by Taki and Shirahata (2009). Therefore, we used the data of 318 HCV-infected hemophiliac patients using the time of birth as an important time point. The time of the infection with HCV was simulated by random numbers distributed between 0 and 5 years after the time of birth. The onset of critical hepatic disease such as liver cirrhosis, liver failure or hepatocellular carcinoma was treated as the event. Data were censored at the end of May 2008 or at the time of lost-to-follow-up. Random numbers were repeatedly generated 1000 times, and Kaplan-Meier survival fractions were obtained each time. The fraction of patients with critical hepatic disease after 30 years from infection was 5.2% when we used the time of birth as the initial point. However, it was 7.6±0.8%(mean±SD) and 9.3±1.0% from 1000 runs using linear and uniform distribution, respectively. If we can find the appropriate distribution pattern for the random numbers, a more precise estimation may be possible.

Keywords: left-censored, survival, Kaplan-Meier method, random number

References

HAREZLAK, J. and TU, W. (2006): Estimation of survival functions in interval and right censored data using STD behavioural diaries. *Statistics in Medicine 25(23), 4053-4064.*

TAKI, M., SHIRAHATA, A. (2009): Current situation of regular replacement therapy (prophylaxis) for haemophilia in Japan. *Haemophilia 15(1), 78-82.*

A Comparison of Some Functional Data Depth Approaches

Stanislav Nagy

Charles University of Prague, Department of Probability and Mathematical Statistics,
Sokolovská 83, 186 75 Praha 8, Czech Republic *s.nagy@volny.cz*

Abstract. The concept of depth for functional data is presented as a useful tool for detection of typical and outlier functions in a functional data random sample.

Several definitions of functional data depth are compared using both simulated and real data sets, namely band and generalized band depths introduced in López-Pintado and Romo (2009), Fraiman's simplical and halfspace depths defined in Fraiman and Muniz (2001) and a functional induced by an isomorphism of arbitrary depth defined on a finite-dimensional Euclidean space of coefficients of functions with respect to the basis in the case of finite-dimensional function spaces.

In order to compensate for an arising disadvantage, we generalize the last-named. We exceed the induced depth by defining functional depth on an arbitrary functional space as induced by a mapping of depth on a finite-dimensional Euclidean space of coefficients of functions with respect to their first few functional principal component. The functional PC's were defined in Ramsay and Silverman (2005). Although some properties are lost as a result of reduced dimension, the reasonability of depth-approximating method is shown for particular cases.

Keywords: functional data, data depth, band depth, functional principal component

References

FRAIMAN, R. ,MUNIZ, G. (2001): Trimmed means for functional data. *Test 10 (2), 419-440.*

LÓPEZ-PINTADO, S., ROMO, J. (2009): On the concept of depth for functional data. *Journal of the American Statistical Association 104 (486), 718-734.*

RAMSAY, J. O., SILVERMAN, B. W. (2005): *Functional data analysis. 2nd ed.*. Springer Series in Statistics, New York.

Beta-κ distribution, an application to extreme hydrologic events

Md. Sharwar Murshed[1] and Jeong Soo Park[2]

[1] Graduate Student of Department of Statistics, Chonnam National University. Gwangju 500-757, Korea. *ruposbd@hotmail.com*
[2] Professor, Department of Statistics, Chonnam National University. Gwangju 500-757, Korea. *jspark@jnu.ac.kr*

Abstract. The beta-κ distribution is one of the distinct special cases of the generalized beta distribution of the second kind. In this study, we have introduced method of moments and method of L-moments to estimate the parameters from the beta-κ distribution. Also assessing the performance of the model by three estimating methods including maximum likelihood estimation method, a simulation study is employed in addition to some real extreme events data sets.

Keywords: Beta-κ distribution, Maximum likelihood estimation, Method of moments, Method of L-moments, Simulation

Nonparametric hypothesis testing for non increasing density family on $\mathbb{R}^+$

S. Khazaei

CEREMADE, Université Paris Dauphine

May 22, 2010

Abstract

In this paper we study nonparametric bayesian inference on the family of non-increasing density function on $\mathbb{R}^+$. One interesting question is to consider a goodness of fit test in such a context. In other words, given observations $X^n = (X_1, ..., X_n)$ supposed to be independent and identically distributed from some given decreasing density, say f, our aim is to answer the following testing problem :

$$H_0 : f = f_0, \quad \text{against} H_1 : f \text{ is decreasing}$$

where f_0 is a given decreasing density.

To do this we define a Nonparametric hypothesis testing. We compare two different approaches. One is based on the Bayes factor, which can be written in this case

$$B_n = \frac{f_0(X^n)}{m_n(X^n)}, \quad m_n(X^n) = \int_{\mathcal{F}} f(X^n) d\pi(f)$$

and $\mathcal{F}$ is the set of decreasing densities. This approach is the most commonly known Bayesian approach for testing, although its computation is still an open problem, given its difficulty.

The second approach is driven by decision theoretic considerations. Consider the loss function $L(f, \delta) = \delta(\epsilon - d(f, f_0))\mathbb{1}_{d(f,f_0)<\epsilon} + (1 - \delta)(d(f, f_0) - \epsilon)\mathbb{1}_{d(f,f_0>\epsilon}$ for a given distance, d, which will be the L_1 distance in the present work. The Bayesian solution to this loss function is given by :

$$\delta^\pi = 1 \quad \text{iff} \quad E^\pi \left(d(f, f_0)|X^n\right) > \epsilon.$$

This second approach has the advantage of taking into account the distance to the null hypothesis, but needs the definition of a threshold ϵ. When no such threshold is known a priori a possibility is to consider a p-value. The method then becomes more complicated to compute. We propose a hybrd algorithm following Ross et al. 2009 to accelerate the computation f the p-value.

The comparison of both approaches is based on a simulation study.

Keywords: Nonparametric Bayesian inference, k-monotone density, kernel mixture, Bayes factor, goodness of fit, p-value.

Parameter Sensitivity analysis for α-stable claim process Submitted to COMPSTAT 2010

Amel Louar[1] and Kamel Boukhetala[2]

[1] ENSSMAL, Dély Ibrahim, Ageria
Coop El Amel G3 N6, Algeria, *amel.laouar@gmail.com*
[2] USTHB, Algeria
Bp. 32,El Alia, Ager *kboukhrtala@usthb.dz*

Abstract. In this paper, we study the behavior of the wealth for an insurance company, with a claims process described by a stochastic differential equation governed by α-stable motion. The additional non classical feature that the company is also allowed to invest in stock market; A part of the wealth is invest in non risky asset, and the rest in a risky asset, modelled by geometric Brownian motion, when we suggest that the market rate (diffusions parameter) of return on financial risky asset can be stochastic, we propose for this latter the Vasicek and CIR models, we make a comparison of investment process for these models. Finally we study the sensitivity of the risk model and wealth with respect to changes in parameters by simulation.

Keywords: α-stable process, interest rate models,α-stable Lévy motion , simulation.

References

Brigo, D. and Mercurio, F. (2007): Interest Rate Models - Theory and Practice, With Smile, Inflation and Credit.ISBN 978-3-540-22149-4 2nd ed.Springer-Verlag Berlin Heidelberg.

Janiko, A. and Weron, A. (1994): Simulation and chaotic behavior of α-stable process. Marcel Dekker, INC, New York .

Zolotarev.V.M and Uchaikin.v.v.(1999):Chance and stability. Stable distribution and their application. *Modern probability and statistics*.Netherlands, Utrecht, VSP.Tokyo, Japan.

New spatial statistics procedures suggested by a critical comparison between geoestatistical packages ArcGIS and R

Carlos Eduardo Melo Martínez[1], Jordi Ocaña Rebull[2] and Antonio Monleón Getino[2]

[1] Engineering Faculty, Distrital University "Francisco José de Caldas" Sede Administrativa-Facultad de Ingeniera, Cr 7 No 40 - 53, Bogotá, Colombia,*cmelo@udistrital.edu.co.*

[2] Departament of Statistics, University of Barcelona Avda Diagonal 645, 08028 Barcelona, Spain *jocana@ub.edu* and *amonleong@ub.edu*

Abstract. This research is focused on the design of programs in R to development geostatistical procedures. Also, some programs on existing theories and an alternative method to nest semivariance models are proposed. In addition, a comparison between programs geostatistics R and ArcGIS is made. In this paper, we present a brief introduction to spatial statistics, the main areas (geostatistics, spatial patterns and lattices), and a short presentation of ArcGIS and R programs in their components of geostatistics. Moreover, the statistical and mathematical aspects of geostatistics are summarized, doing emphasis on the semivariogram, on both probabilistic interpolation methods kriging and deterministic, and the goodness of fit interpolation methods (cross-validation).

We propose a series of functions which are designed in the package R. These allow a more complete geostatistical analysis together with the help of packets previously designed in R such as: geoR, gstat, geostat and akima, among others. In this way, these contributions are: a function for the construction of the trimmed mean semivariance, a nesting function semivariance functions from functions displaced semivariance theoretical models (spherical, exponential and Gaussian), a function for the construction of pocketplot (useful for the analysis of local stationarity), a spline interpolation function from radial basis functions (multiquadric and inverse multiquadric), and a cross-validation function to validate the interpolation methods based in the errors. Also, a comparison of the geostatistical modules between ArcGIS and R is made. Thereby, we analyze its benefits, limitations and overall behavior for this type of statistical analysis.

Keywords: nested models, interpolation methods, R and ArcGIS, geostatistical comparison, pocket plot.

References

R Development Core Team (2009), R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria, URL http://www.Rproject.org/.*

A computational approach to dissect skin pathologies based on gene expression barcodes

Mayte Suárez-Fariñas[1,2], Erika Billick[1], Hiroshi Mitsui[1], Fuentes-Duculan, J.[1], Fujita, H.[1], Lowes, M.[1], Nograles, K.E.[1], and James G. Krueger[1]

[1] Laboratory for Investigative Dermatology and
[2] Center for Clinical and Translational Science, The Rockefeller University. 1230 York Avenue, New York, NY 10065.

Abstract. Zilliox and Irrizarry (2007) developed a gene expression barcode based on thousands of chips from different human tissues. The barcode methodology assumes that each gene is expressed in some cell types and not expressed in others, so the distribution of the $\log_2$ intensities across cell types is multimodal. Thresholds derived from the distributions are used to assign discrete gene expression states, called a *gene expression barcode*. As their classifier, based on general tissues, does not make fine discriminations among skin cell types we applied the method to a database of 120 skin samples to create a barcode specific for skin cell types. We present our skin-specific barcode classifiers plus three applications:

Skin-Cell Map: Using the binary data for each gene across samples of different cell types, we identify genes that are unique to a cell type and those that are shared. Almost 70% of the probes are noninformative while only 11% of the genes are uniquely expressed by a specific cell type.

Hierarchical Classifier: The skin-cell barcode was used to create a classifier, taking advantage of the hierarchical lineage of skin cell types using the *pamr* method in three stages. In each stage, the classifier finds the set of genes that discriminates amongst a reduced number of cell types. These classifiers are able to predict the cell type accurately from a single hybridization, with 0% errors in leave-one-out cross-validation, and 1% in the testing samples.

Deconvolution of a complex tissue: Biopsies contain variable amounts of healthy and diseased tissues each with multiple cell types. mRNA extracted from a biopsy sample results in expression profiles which are averages of the expression profiles of the individual cell types, weighted by the prevalence of that cell type in that specific sample. Skin diseases having subtly different cell-type compositions are hard to discriminate based on raw expression profiles; reconstructing such composition from a single hybridization is called deconvolution. We present a computational deconvolution algorithm based on our Skin-Cell Barcode, and demonstrate its power by deconvolving Dermis and Epidermis of Normal and Psoriasis skin samples obtained by Laser Capture Microdissection.

Keywords: microarrays, gene-expression barcode, classification

Assessing DNA copy numbers in large-scale studies using genomic arrays

Robert B Scharpf[1] and Ingo Ruczinski[2]

[1] Division of Oncology Biostatistics, The Johns Hopkins Sidney Kimmel Comprehensive Cancer Center, Baltimore, MD 21205, USA *rscharpf@jhsph.edu*
[2] Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA *ingo@jhu.edu*

Abstract. Oligonucleotide-based arrays are commonly used to assess genotype-phenotype association on a genomic scale, and studies involving DNA copy numbers are becoming more common. While genotyping algorithms are largely concordant when for example assessed on HapMap samples, methods to assess DNA copy number alterations often yield discordant results. One explanation for the discordance is that DNA copy number estimates are particularly susceptible to systematic differences that arise across batches of samples that were processed at different times or by different labs. Analysis algorithms that do not account for such systematic biases are more prone to spurious findings that will not replicate in other studies.

This presentation describes statistical methods and software for the locus-level estimation of DNA copy number that specifically adjust for batch effects (Scharpf et al. (2009)). We illustrate a workflow for copy number analysis using HapMap Phase 3 data, including normalization, genotyping, adjusting for batch effects in copy number estimation, visualizations to inform downstream processing, and smoothing point estimates as a function of physical position via segmentation and hidden Markov models (Scharpf et al. (2008)). All analyses are performed in the statistical environment R, using S4 class definitions for high-throughput SNP arrays that extend the Bioconductor eSet class (Scharpf and Ruczinski (2002)). Extending the eSet class promotes code reuse through inheritance as well as interoperability with other R packages and is less error-prone, and facilitates reproducible research.

Keywords: DNA copy number, genomic arrays, S4 classes and methods, reproducible research

References

SCHARPF R.B., PARMIGIANI G., PEVSNER J., RUCZINSKI I. (2008): Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. *The Annals of Applied Statistics, 2(2): 687-713.*

SCHARPF R.B., RUCZINSKI I. (2010): R classes and methods for SNP array data. *Methods in Molecular Biology 593: 67-79.*

SCHARPF R.B., RUCZINKI I., CARVALHO B., DOAN B., CHAKRAVARTI A., IRIZARRY R.A. (2009): A multi-level model to address batch effects in copy numbers using SNP arrays. *Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 197.*

The Weighted Halfspace Depth – a Generalization of the Halfspace Depth

Lukáš Kotík[1,2]

[1] Katedra pravděpodobnosti a matematické statistiky, Matematicko–fyzikální fakulta Univetzity Karlovy v Praze
Sokolovská 83, 186 75 Praha 8, Czech Republic, *lukaskotik@gmail.com*
[2] Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 4, 182 08 Praha 8, Czech Republic

Abstract. The statistical depth can be thought as a generalization of the univariate quantiles. It has become quite popular tool in nonparametric inference for multivariate data and several definitions of the depth has been introduced in recent years. The halfspace (Tukey) depth and the simplex depth are now probably the most known and the most used depth functions. We propose a generalization of the halfspace depth – the weighted halfspace depth. It is based on using weighted probabilities of a (half)space instead of plain probability of a halfspace. The proposed depth can be considered to be more appropriate if our data aren't symmetrically distributed or in a case of mixtures of distributions. The definition of the weighted halfspace depth, its basic properties and examples are shown in the poster.

Keywords: data depth, nonparametric multivariate analysis

References

HLUBINKA, D., KOTÍK, L. and VENCÁLEK, O. (2010): Weighted Halfspace Depth. *Kybernetika 46 no. 1, 125-148.*

LIU, R.Y., SERFLING, R. and SOUVAINE D.L. (2006): *DIMACS: Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications.* American Mathematical Society DIMACS Book Series, Volume 72.

Indices of Nonlinearity and Predictability for Time Series Models

Norio Watanabe[1] and Yusuke Yokoyama[2]

[1] Industrial and Systems Engineering, Chuo University
Kasuga 1-13-27, Bunkyo-ku, Tokyo, Japan, *watanabe@indsys.chuo-u.ac.jp*
[2] Graduate School of Science and Engineering, Chuo University
Kasuga 1-13-27, Bunkyo-ku, Tokyo, Japan

Abstract. In this study we consider the problem on characterization of the nonlinearity for stochastic dynamical systems which generate time series.

The Lyapunov exponent is a well-known index for deterministic systems and it can measure the degree of nonlinearity and imply the information of predictability. The maximum Lyapunov exponent can be defined for stochastic systems (McCaffrey et al. (1992)). However, it is not clear what the Lyapunov exponent measures in the case of stochastic systems. Other indices on nonlinearity and predictability of stochastic systems have been proposed by Watanabe (2007). Though the meaning of their indices is understandable, the computation is not easy. In this study we modify their indices and improve the computational aspect.

We consider the discrete-time stochastic dynamical system:

$$x_n = f(x_{n-1}, x_{n-2}, ..., x_{n-p}) + e_n, \tag{1}$$

where $\{e_n\}$ is a white noise with the variance σ^2 ($\sigma > 0$), and p is a positive integer. We assume that f and σ are known here.

We introduce two predictors of x_{n+k} based on $x_n, x_{n-1}, ..., x_{n-p+1}$ for definition of indices. First the conditional expectation is the best predictor with respect to the mean squared error. Usually it is very difficult to calculate the conditional expectation not only theoretically but also numerically. Then we adopt the approximation of the best predictor by multilayered neural networks. Second we consider the recursive predictor. Indices of nonlinearity and predictability are defined by using the mean prediction errors of two predictors.

The validity of indices is demonstrated by simulation studies. And we consider the relationship among our indices and the Lyapunov exponent.

Keywords: Nonlinear model, Prediction, Lyapunov exponent

References

MCCAFFREY, D. F. et al. (1992): Estimating the Lyapunov exponent of a chaotic system with nonparametric regression. *J. of American Statistical Association, 87, 682-695.*

WATANABE, N. (2007): Computational indices of nonlinearity and predictability for stochastic nonlinear systems by neural networks. In: *Bulletin of the International Statistical Institute 56th Session, Proceedings.*

Solution Tuning - an attempt to bridge existing methods and to open new ways

Tatjana Lange[1]

Professor, University of Applied Science Merseburg, Germany
Geusaer Straße, D-06217 Merseburg, *tatjana.lange@hs-merseburg.de*

Abstract. When we use measured data for searching a solution or making decisions sometimes the results may be unstable, biased or "distorted".

Different domains of science and engineering, such as

- Regularisation Theory (e.g. Tichonow and Arsenin (1979), Lange (1994))
- Robust Estimations
- Data Analysis (e.g. Mosler et al. (2009))
- Regression Analysis and Modelling (e.g. Lange (1995))
- Minimization of Empirical Risk (e.g. Vapnik (1998))
- Learning Theory and Neuronal Networks (e.g. Wasiljew and Lange (1998))
- Psi Optimization

investigate these problems independently and often do not perceive the achievements of the neighbouring disciplines.

This presentation tries to find out what is common between the different methodologies. It deals with the connections between the methods that are typical for the scientific domains listed above. Finally it tries to make some generalizations and to open new ways to quantify the influence of outliers of different nature on the **solution tuning**.

Keywords: Data analysis, regularization, optimization, estimation, solution tuning

References

LANGE, T. (1994): New Structure Criteria in GMDH. *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modelling*, Kluwer Academic Publishers.

LANGE, T. (1995): Structure Criteria for Automatic Model Selection in Multilayered GMDH Algorithms in Case of Uncertainty of Data. *'Systems Analysis - Modelling - Simulation (SAMS)'*, Berlin.

MOSLER, K., LANGE, T., BAZOVKIN, P. (2009): Computing zonoid trimmed regions of dimension $d > 2$. *Computational Statistics & Data Analysis*. Elsevier Science Publishers B.V, Amsterdam.

TICHONOW A.N., ARSENIN W.J. (1979) - *in Russian*: *Metody rescheniya nekorrektnyh zadatsch*. Nauka, Moscow.

VAPNIK, V. (1998): *Statistical Learning Theory*. John Wiley, New York.

WASILJEW W.I., LANGE, T. (1998) - *in Russian*: Prinzip dual'nosti v probleme obutscheniya raspoznavaniya obrazov. *Kibernetika i vytschislitel'naya technika*, Kiev.

Comparison of inference for eigenvalues of covariance matrix with missing data

Shin-ichi Tsukada[1] , Yuichi Takeda[2] and Takakazu Sugiyama[3]

[1] Department of Education, Meisei University
2-1-1 Hodokubo, Hino, Tokyo, 191-8506, Japan, *tsukada@ge.meisei-u.ac.jp*
[2] Kanagawa Institute of Technology
1030 Shimo-Ogino, Atsugi, Kanagawa, 243-0292, Japan,
y-takeda@ctr.kanagawa-it.ac.jp
[3] Department of Mathematics, Chuo University

Abstract. In real multivariate statistical analysis, there is often a case with missing data. The inference of the eigenvalue is not evaluated for each method though some methods for missing data are proposed for principal component analysis (PCA).

We assume that the population distribution is 10-dimensional normal distribution and contaminated normal distribution. Since the larger eigenvalues are interested on PCA, we treat everything from the first largest eigenvalue to the third largest eigenvalue. The missing value is made at random and let a missing rate τ be 10% and 20%. The methods we use are PCA for decreased $(1-\tau)$% sample of all sample (DS), PCA for a complete sample (CS), PCA for the covariance matrix using pairwise sample (PS), KNNimpute (KNN), MLE for covariance matrix with missing data (MLE), Probabilistic PCA (PPCA) and Nipals PCA (NIPALS). KNN, MLE, PPCA and NIPALS methods are easily carried out on R using the EMV package, the mvnmle package and the pcaMethods package, respectively. MLE routine estimates the mean and covariance matrix of multivariate normal distribution from missing data. If data values are missing, the routine implements the ECM algorithm of Meng and Rubin (1993) with enhancements by Sexton and Swensen (2000). The setting of package is default.

By numerical simulation, it is found that MLE method is best as a whole. This method is assumed that the population distribution is normal, but effective under contaminated normal distribution as a symmetric distribution. Pairwise sample method, which is more simple, has a relative error of about 1%. We recommend MLE method as an inference for eigenvalues of covariance matrix.

Keywords: missing data, eigenvalue, principal component analysis

References

Meng, XL. and Rubin, D.B. (1993). Maximum Likelihood Estimation via the ECM Algorithm, *Biometrika*, **80**, 267–278.

Sexton, J. and Swensen, A.R. (2000). ECM Algorithms that Converge at the Rate of EM, *Biometrika*, **87**, 651–662.

Spatial modeling of extreme values : A case of highest daily temperature in Korea

SangHoo Yoon[1] YoungSaeng Lee[2] and JeongSoo Park[3]

[1] Graduate student, Department of Statistics, Chonnam National University 300 Yongbong-dong Bukgu, Gwangju, South Korea, *statstar96@gmail.com*
[2] Graduate student, Department of Statistics, Chonnam National University 300 Yongbong-dong Bukgu, Gwangju, South Korea, *hellight@naver.com*
[3] Professor, Department of Statistics, Chonnam National University 300 Yongbong-dong Bukgu, Gwangju, South Korea, *jspark@jnu.ac.kr*

Abstract. Generalized Extreme Value models are found to be approximate limits of univariate extremes. However, models for multivariate and spatial extremes are not so straightforward. We showed the usefulness of generalized extreme value distribution in applying extreme daily temperature for each location (Kim and Park(2008)). For establishing a spatial model of extreme value, geographical aspects could be used because a spatial dependence exists between locations. To perform statistical analysis on extreme daily temperature, we should build models which are capable of detecting and determining this spatial distance in the extremal level. Several authors have suggested methods to handle spatial extremes (Buishad et al.(2008), Smith(1990) and Cooley et al.(2007)). We suggest a spatial model of extreme daily temperature using suitable max-stable models together with composite likelihood to perform likelihood based statistical inference on spatial extremes. The data is collected from 59 stations in Korea between 1973 and 2009. 4 spatial modes representing 4 regions are proposed compare to one global spatial model. The administrative and geographic senses are considered in dividing.

Keywords: Spatial modeling, Max-Stable process, Composite likelihood

References

BUISHAND, T. A., HAAN, L. D. and ZHOU C. (2008): On spatial extreme: With application to a rainfall problem. *Annals of Applied Statistics 2 (2), 624-642.*

COOLEY, D., NYCHKA, D. and NAVEAU, P. (2007): Bayesian spatial modeling of extreme precipitation return levels, *Journal of the American Statistical Association 102, 824-840.*

KIM, H. M. W. and PARK, J. S. (2008): Statistical tend analysis for extreme of daily highest temperature. *Journal of the Korean Data analysis Society 10 (3), 1591-1601.*

SMITH, R. L. (1990): Max-stable processes and spatial extremes. *Unpublished, available from http://www.stat.unc.edu/postscript/rs/spatex.pdf.*

indexGosinska, E.@*Gosińska, E.*

Cointegration analysis of models with structural breaks

Emilia Gosińska[1]

Chair of Econometric Models and Forecasts, University of Lodz
Rewolucji 1905 r. No. 41, 90-214 Lodz, Poland, *emfemj@uni.lodz.pl*

Abstract. The Vector Error Correction Model with structural breaks is considered. It is assumed that the data generation process consists of stochastic as well as deterministic component, Saikkonen et al. (2004). Structural breaks can be implemented as the appropriate adjustments of deterministic variables or by time-varying parameters in stochastic component. In this study the stochastic component is represented by VAR model, time-varying VAR model or threshold VAR model, Hansen and Seo (2002). In the paper it is proved that the form of VECM model is dependent on the way the structural change is explained in data generation process. The presence of deterministic components in DGP leads to a model with deterministic components in cointegration space. The investigation concerns the model of Polish inflation in the presence of structural breaks, assuming different forms of DGP.

Keywords: cointegration, vector error correction model, structural breaks

References

Saikkonen P., Lütkepohl H., Trenkler C. (2004): Break Date Estimation and Cointegration Testing in VAR Processes with Level Shift. *EUI Working Paper ECO, No.2004/21.*

Hansen B., Seo B. (2002): Testing for two-regime threshold cointegration in vector error-correction models. *Journal of Econometrics 110 (2002) 293-318.*

Application Of Regression-Based Distance Matrix Analysis To Multivariate Behavioral Profile Data

Ozgur Tosun[1], William G. Iacono[2], Matthew McGue[2], and Nicholas J. Schork[3]

[1] The Department of Biostatistics and Medical Informatics, Akdeniz University Antalya, Turkey, *otosun@hotmail.com*
[2] Department of Psychology, University of Minnesota Minneapolis, MN, 55455, United States. *wiacono@tfs.psych.umn.edu*
[3] Scripps Translational Science Institute La Jolla, CA, 92037, United States. *nschork@scripps.edu*

Abstract. Multivariate distance matrix regression (MDMR) is a statistical method that can be used to analyze high dimensional data sets. The technique works by testing the relationship between similarity profiles computed across a large number of variables and an additional set of (predictor) variables (Zapala et al. (2006)). MDMR can be framed in a multivariate multiple regression setting so that the impact of each of the set of predictor variables can be related to variation in the similarity of the multivariate profiles (Wessel et al. (2006)). We considered the application of MDMR to a twin data set in which a large number of behavioural measures were collected in addition genotype information. We found a statistically significant association between a single nucleotide polymorphism on chromosome 15 in the GABRG3 gene (rs10852211) and a set of substance abuse and behavioural disinhibition measures. The association was replicated in two separate subsets of monozygotic twin data. Dick and colleagues has previously shown that GABRG3 gene is significantly involved in the risk for alcohol dependence. They advocated that nearly all SNPs across this gene yielded evidence of association (Dick et al. (2004)).

Keywords: MDMR, genetics, substance abuse, alcohol dependence

References

DICK, D.M., EDENBERG, H.J., XUEI, X., GOATE, A., KUPERMAN, S., SCHUCKIT, M., CROWE, R., SMITH, T.L., PORJESZ, B., BEGLEITER, H. and FOROUD, T. (2004): Association of GABRG3 with alcohol dependence. *Alcohol Clin Exp Res. 28(1),4-9.*

WESSEL, J. and SCHORK, N.J. (2007): Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet. 79(5),792-806.*

ZAPALA, M.A. and SCHORK, N.J. (2006): Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proc Natl Acad Sci U S A. 103(51),19430-5.* .

Simulating multi-self-similar spatiotemporal models with CUDA

Francisco Martínez[1], María Pilar Frías[2], and María Dolores Ruiz-Medina[3]

[1] Department of Computer Science, University of Jaén
Campus Las Lagunillas s/n, 23071 Jaén (Spain), *fmartin@ujaen.es*
[2] Department of Statistics and Operations Research, University of Jaén
Campus Las Lagunillas s/n, 23071 Jaén (Spain), *mpfrias@ujaen.es*
[3] Department of Statistics and Operations Research, University of Granada
Campus Fuentenueva s/n, 18071 Granada (Spain), *mruiz@ugr.es*

Abstract. In Frías et al. (2009) multi-self-similar spatiotemporal models displaying long-range dependence, in space and time, are introduced as the mean-square convolution of an input process Y with the spatiotemporal Riesz kernel. Suitable conditions, that ensure a second-order-output process X, are assumed on process Y. In addition, an estimation method based on the marginal spectral densities is implemented to approximate the long-memory and/or strong-spatial-dependence parameters.

In order to experiment with the models the input process Y is approximated using a Monte Carlo method. This approximation is a very CPU-demanding task, so that computing the approximation of an input process takes more than one day.

Fortunately, the algorithm used to simulate the input process Y can be easily parallelized. In this work we have studied two ways of getting a parallel algorithm.

The first way is using the computational power of current multi-core processors. We have created a computing thread for every core getting a "perfect speedup"—a 4 speedup on a processor with 4 cores.

The second way is developing a SIMD—Single Instruction Multiple Data—algorithm so that it can be executed on current graphics cards (GPUs) supporting the CUDA architecture, see Kirk and Hwu (2010). This second algorithm gets a 30 speedup, but it is far from achieving a perfect speedup. We are currently trying to improve this SIMD algorithm so that a better performance can be achieved.

Keywords: spatiotemporal models, CUDA architecture, parallel programming

References

KIRK, D. B. and HWU, W. W. (2010): *Programming Massively Parallel Processors*. Morgan Kaufmann.

FRÍAS, M.P., RUIZ-MEDINA, M.D., ALONSO, F.J. and ANGULO, J.M., (2009): Spectral-marginal-based estimation of spatiotemporal long-range dependence. *Communications in Statistics - Theory and Methods 38 (1), 103-114*.

Acknowledgements: This work has been partially granted by the Ministerio de Ciencia y Tecnología of Spain and the European Union by means of the ERDF funds, under the research project TIN2007-67474-CO3-03, and by the Conserjería de Innovación, Ciencia y Empresa of the Junta de Andalucía and the European Union by means of the ERDF funds, under the research projects P06-TIC-01403 and P07-TIC-02773.

Application of autocopulas for analysing residuals of Markow –Switching models Submitted to COMPSTAT 2010

Jozef Komorník[1] and Magda Komorníková[2]

[1] Faculty of management, Comenius University, Odbojárov 10, P.O.BOX 95, 820 05 Bratislava, Slovakia, *Jozef.Komornik@fm.uniba.sk*

[2] Faculty of Civil Engineering, Slovak University of Technology, Radlinského 11, 813 68 Bratislava, Slovakia, *magda@math.sk*

Abstract. The topics of this investigation was motivated by modeling of a large number of economic and financial time series from the emerging Central - European economics using Markov – Switching (MSW) models (see Frances and van Dijk (2000), Hamilton (1989)).

We based the selection of the models (optimizing the number of hidden states and the order of the local autoregressive models) on the BIC criterion.

Recall that the residuals of these models are supposed to be independent (not only serially non–correlated). This property can be tested e.g. by the BDS test (see Brock et al. (1996)).

Inspired by the approach of Rakonczai (2009) we applied autocopulas to the time series of the above mentioned residuals in order to gauge how much they violate the assumptions of independence. We arrived at an interesting conclusion concerning the residuals of the models that were selected as optimal on the basis of the BIC criterion. We observed that the autocopulas for the residuals of the optimal models were mostly substantially closer to the (independence indicating) product form (especially for lags $k \geq 2$) than those for competing non–optimal models.

Ackonowledgement: The research was partly supported by the Grants APVV-0012-07 and LPP-0111-09.

Keywords: time series, Markov–Switching models, autocopulas

References

BROCK, W. A., DECHERT, W. D., SCHEINKMAN, J. A. and LE BARON, B. (1996): A test for independence based on the correlation dimension. *Econometric Reviews 15, 197–235.*

FRANCES, P. H. and VAN DIJK, D. (2000): *Non-linear time series models in empirical finance.* Cambridge University Press.

HAMILTON, J.D. (1989): A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica 57, 357–384.*

RAKONCZAI, P. (2009): On modeling and prediction of multivariate extremes with applications to environmental data. In: Licentiate Theses in Mathematical Sciences 2009:3, ISSN 1404–028X, 83–109.

Comparison of Robust Estimators in One-Way-Classification Experimental Design Model Submitted to COMPSTAT 2010

Assoc.Prof.Dr. Inci Batmaz[1] and Ibrahim Erkan[2]

[1] The Middle East Technical University, Department of Statistics METU, Ankara, TURKEY *ibatmaz@metu.edu.tr*
[2] The Middle East Technical University, Department of Statistics METU, Ankara, TURKEY *ierkantr@yahoo.com*

Abstract. In this study, one-way-classification experimental design model $y_{ij} = \mu + \tau_i + e_{ij}$ for $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n$ and $\sum_{i=1}^{k} \tau_i = 0$ (without loss of generality) is considered. Traditionally,e_{ij} have been assumed to be normally distributed with mean zero and variance σ_2. In recent years, however, there has been a realization that the normality assumption is often not valid. In fact, the exact distribution of e_{ij} is not known. The strategy is to locate the most plausible distribution for e_{ij}, and develop techniques which are efficient under an assumed distribution and retain their efficiency under reasonable alternatives. Such techniques are called robust procedures. Several such procedures are known and are based on the following estimators: Tukey's trimmed sample mean and the matching variance, Tiku's modified maximum likelihood estimators based on censored samples, Huber's M-estimators based on randomly censored samples, Tiku's robust estimators based on complete samples, distribution free procedures based on ranks (the relative positions of sample observations). e.g. R-Estimates, Hodges-Lehmann Estimates.

In this study, assuming that e_{ij}'s have a Long Tailed Symmetric Distribution (shape parameter, p = 3.5) we compared the above procedures and evaluate their relative performances. We also compared their robustness properties under Dixon's outlier model, mixture model, contamination model, autocorrelation model, misclassification model and skew alternative models such as Generalized Logistic (b=2).

Keywords: ANOVA, robustness, simulation

References

MCKEAN, J. W., HETTMANSPERGER, T. P. (1978): A Robust Analysis of The General Linear Model Based on One Step R-Estimates. *Biometrika*, 65 (3), 571 - 579.

DUNNET, C. W. (1982): Robust Multiple Comparisons. *Communications in Statistics - Theory and Methods*, 11 (22), 2611 - 2629.

GROSS, A. M. (1976): Confidence Interval Robustness with Long-Tailed Symmetric Distributions. *Journal of the American Statistical Association*, 71 (354), 409 - 416.

Heterocedasticity in the *SEM* using Robust Estimation

Manuela Souto de Miranda[1], João Branco[2], and Anabela Rocha[3]

[1] Department of Mathematics, University of Aveiro
Campus de Santiago, 3810-193 Aveiro, Portugal, *manuela.souto@ua.pt*
[2] Department of Mathematics, I.S.T., Technical University of Lisbon
Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal *jbranco@math.ist.utl.pt*
[3] I.S.C.A., University of Aveiro
Apartado 58, EC AVEIRO, 3811-902 Aveiro, Portugal, *anabela.rocha@ua.pt*

Abstract. A Simultaneous Equations Model (SEM) is formalized by a system of equations having correlated error terms. In its more general form the model is heterocedastic, which means that different observations of each equation may have different error variances. The most popular estimators for the SEM are based on least squares and they are derived for the particular case of homocedasticity. For the heterocedastic system the Generalized Method of Moments estimator (GMM) is recommended. Unfortunately none of those estimators are robust. Robust estimation for the SEM has been studied, among other authors, by Amemiya (1982) and Maronna et al. (1997), following the least squares approach. A robust version of the GMM was developed in Rocha (2010). The robust version of the GMM includes a resampling step which allows the estimation of each element of the covariance matrix of the errors. The computation of the robust estimates becomes very simplified for the particular case of homocedasticity, since this case does not require a resampling step.
A simulation study was conducted for observing the performance of the simplified robust version of the GMM (no resampling step) when the data were generated from an heterocedastic model. The results pointed out that the no resampling version performs very well under Gaussian errors, either with or without contamination and its computational time decreases drastically when compared with the general estimation process. Thus, unless there exist strong reasons for assuming the heterocedastic SEM, we conclude that it may be preferable to act as if the model was homocedastic.

Keywords: *SEM*, robust estimation, *GMM*.

References

AMEMIYA, T. (1982): Robust estimation in simultaneous equations models. *Journal of Statistical Planning and Inference 57, 233–244.*

MARONNA, R. and YOHAI, V. (1997): Robust estimation in simultaneous equations models. *Journal of Statistical Planning and Inference 57, 233–244.*

ROCHA, A. (2010): *Estimação Robusta em Modelos Lineares de Equações Simultâneas.* PhD Thesis (in Portuguese), University of Aveiro, Portugal.

Assessing environmental performance using Data Envelopment Analysis combined with cluster analysis

Eugenia Nissi[1] and Agnese Rapposelli[1]

Department of Quantitative Methods and Economic Theory Viale Pindaro,42 -65127 Pescara Italy *nissi@unich.it - a.rapposelli@unich.it*

Abstract. There has been increasing recognition in developed nations of the importance of good environmental performance, in terms of reducing environmental disamenities such as pollutants emissions and waste. It is well known that national institutions have various macroeconomic objectives, for instance a high level of real GDP per capita or a low rate of inflation. Recently, another objective that needs to receive considerable attention is a reduction in the amounts of environmental disamenities generated in the air or water as outputs of the production of goods and services. Hence their performance (efficiency) needs to be evaluated in terms of their ability to maximise macroeconomic objectives while minimising environmental disamenities. The aim of the present paper is to measure the environmental efficiency of Italian provinces for 2004 and our objective is to adapt the techniques of efficiency measurement, such as Data Envelopment Analysis, to the problem at hand, where outputs do not refer only to goods, but we have also undesirable outputs. Data Envelopment Analysis (DEA) is a linear-programming-based method for evaluating the relative performance of a homogeneous set of Decision Making Units (DMUs). Since the original DEA study by Charnes, Cooper and Rhodes (1978) this methodology has been widely studied in literature and there has been rapid growth in the field. Hence, the idea of this work is firstly to cluster the operating DMUs (Decision Making Units) or "provinces" by input/output mix and then to estimate efficiency for each cluster. Then, for each cluster DEA models are carried out.

Keywords: Data Envelopment Analysis, undesiderable outputs, environmental filed

References

BANKER, R.D., CHARNES R.F, COOPER W.W. (1984): Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis, Management Science,30, 1078-1092.

CHARNES A., COOPER W., RHODES E., (1978): Measuring the efficiency of decision-making units, European Journal of Operational Research,2, 429-444.

COLI M., NISSI E., RAPPOSELLI A. (2008): Performance Measurement by means of data Envelopment Analysis: a new perspective for undesiderable output.. In: Jibendu Kumar Mantri(Eds.): *Research Methodology on Data Envelopment Analysis*. Boca Raton, 217–226.

Stochastic Newmark Schemes for the Discretization of Hysteretic Models

Pedro Vieira[1,2], Paula M. Oliveira[2], and Álvaro Cunha[2]

[1] University of Trás-os-Montes e Alto Douro
Quinta de Prados, Vila Real, Portugal, *pmfvieira@hotmail.com*
[2] Faculty of Engineering, University of Porto
Rua Dr. Roberto Frias, Porto, Portugal *poliv@fe.up.pt acunha@fe.up.pt*

Abstract. The need to study and to obtain digital solutions of stochastic non-linear differential equations is a common situation in Seismic Engineering. This is the case for the hysteretic models. These models do not have an exact solution and can only be approximated by numerical methods. We discretize the solutions using the stochastic improved Euler scheme and the three parameter implicit stochastic Newmark schemes: a higher order and a lower order Newmark scheme. In the case of hysteretic models subjected to gaussian white noises, we were able to reduce the problem of approximating the solution to that of a linear system in each time step avoiding the Newton–Raphson method in the same time steps. This allowed us to save computational effort in the approximation of the response of the hysteretic system and was achieved by giving explicitly the value of one of the parameters in the equation of the Newmark scheme that corresponds to the hysteretic variable while keeping the equations of the displacement and velocity implicit. We compare the performance of these two implicit Newmark schemes. In the simulation study for the Bouc-Wen model, we compare the solutions produced for the specific choice of the parameters ($\alpha = 0.5$, $\beta = 0.5$) which are the values used by Roy and Dash(2005) in the case of linear systems. We conclude that the standard deviation of the displacement obtained from the proposed higher order Newmark scheme is larger than that obtained from the proposed lower order Newmark scheme. The proposed lower order Newmark scheme is computationally atractive to compete with the improved Euler scheme.

Keywords: Stochastic Differential Equations, Newmark schemes, Hysteretic models

References

Roy, D. and M.K. Dash (2005): Explorations of a family of stochastic Newmark methods in engineering dynamics. *Comput. Methods in Applied Mechanics and Engineering, 194, 4758–4796.*

Tocino, A. and J. Vigo-Aguiar (2005): Weak second order conditions for stochastic Runge-Kutta methods. *SIAM, J. Sci. Comput., 24(2), 507–523.*

Comparing the Central Venous Pressures Measured via Catheters Inserted in Abdominal Vena Cava Inferior and Vena Cava Superior in Intensive Care Patients with Bland Altman Analysis

Deniz Ozel[1] and Melike Cengiz[2]

[1] Biostatistics and Medical Informatics, Akdeniz University, Medical Faculty Antalya, Turkey, *denizozel@akdeniz.edu.tr*
[2] Anaesthesiolgy and ICU, Akdeniz University, Medical Faculty Antalya, Turkey, *melikecengiz@yahoo.com*

Abstract. Clinicians often need to know whether a new method of measurement is equivalent to an established one already in clinical use: Hanneman (2008). An appropriate comparison of the two measures needs to highlight such differences hence the Bland-Altman plot, which explicitly shows differences between the two measures (on the Y axis) over their range (on the X axis): Hopkins (2004). In our study, Bland-Altman analyses were used to determine the level of agreement between the central venous pressures (CVP) measured via catheters inserted in abdominal vena cava inferior (aVCI) and vena cava superior (VCS) in intensive care patients. We performed 148 simultaneous measurements in 49 patients. Analyses were performed by MedCalc version 11.2.1.0. The 95% limits of agreement, estimated by mean difference +/- 1.96 standard deviation of the differences for CVP were 2.33 and +4.16 mmHg. The bias (mean difference between aVCI and VCS measurements) was 0.92 mmHg. The computed upper and lower levels of agreement We also found that 0.047 (7/148) measurements are out of the predefined limits. According to our data aVCI and VCS measurements have 0.92 bias but we found some factors (such as mechanical ventilation, PEEP, PAP, MAP and IAP) were effective on the difference between CVP values obtained simultaneously via two different routes. So the clinical importance of this difference may not be important.

Keywords: statistical agreement, BlandAltman, limits of agreement, intensive care

References

HANNEMAN, S.K. (2008): Design, analysis, and interpretation of method-comparison studies. *AACN Adv Crit Care 19(2), 223-34.*

HOPKNS W. G.(2004): Bias in Bland-Altman but not Regression Validity Analyses. *Sportscience 8, 42-46.*

Ratio Type Statistics for Detection of Changes in Mean and the Block Bootstrap Method

Barbora Madurkayova[1]

Faculty of Mathematics and Physics, Charles University in Prague
Sokolovska 83, 186 75 Prague 8, Czech Rebublic, *madurka@karlin.mff.cuni.cz*

Abstract. Procedures for detection of changes in mean in models with dependent random error terms are considered. In particular test procedures based on ratio type test statistics that are functionals of partial sums of residuals are studied.

Ratio type statistics are interesting for the fact that in order to compute such statistics there is no requirement to estimate the variance of the underlying model. Therefore they represent a suitable alternative for classical (non-ratio) statistics, most of all in cases when it is difficult to find a variance estimate with satisfactory properties.

We assume to have data obtained in ordered time points and study the null hypothesis of no change against the alternative of a change occurring at some unknown time point. We explore the possibility of applying the block bootstrap method for obtaining critical values of the proposed test statistics in a model with α-mixing random errors. We follow the ideas described in Horváth et al. (2008) and Hušková, M. and Marušiaková, M. (2009).

Keywords: ratio type test statistics, block bootstrap, α-mixing random errors

References

HORVÁTH, L., HORVÁTH, Z., and HUŠKOVÁ, M. (2008): Ratio tests for change point detection. In: *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor P. K. Sen* ***1****:293 – 304.*

HUŠKOVÁ, M. and MARUŠIAKOVÁ, M. (2009): M-procedures for detection of changes for dependent observations In: *Proceedings of the 6th St. Petersburg Workshop on Simulation,* ***2****:673 – 678.*

Rank scores tests in measurement error models – computational aspects

Jan Picek[1]

Department of Applied Mathematics, Technical University of Liberec
Studentská 2, 461 17 Liberec, Czech Republic, *jan.picek@tul.cz*

Abstract. Consider the linear regression model

$$Y_i = \beta_0 + \mathbf{x}'_{ni}\boldsymbol{\beta} + \mathbf{z}'_{ni}\boldsymbol{\delta} + e_i, \quad i = 1, \ldots, n \tag{1}$$

with observations $Y_1, \ldots, Y_n$, independent errors $e_1, \ldots, e_n$, identically distributed according to an unknown distribution function F; $\mathbf{x}_{ni} = (x_{i1}, \ldots, x_{ip})'$, $\mathbf{z}_{ni} = (z_{i1}, \ldots, z_{iq})'$ are the rows of regression matrices, β_0, $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\delta} \in \mathbb{R}^q$, are unknown parameters.

We want to test the hypothesis

$$\mathbf{H}: \ \boldsymbol{\delta} = \mathbf{0}, \tag{2}$$

considering β_0 and $\boldsymbol{\beta}$ as a nuisance parameters.

In the model without measurement errors, the hypothesis (2) can be tested by the regression rank score test. If the regressors are affected by random measurement errors then Jurečková, Picek and Saleh (2008) considered the tests based on the regression rank scores. No estimation of the nuisance parameters is necessary. If the $\mathbf{x}_{ni}$ or both the $\mathbf{x}_{ni}$ and $\mathbf{z}_{ni}$ are affected by random errors then we cannot use a regression rank scores test. Jurečková, Picek and Saleh (2008) suggested the aligned rank test with estimated nuisance parameters.

The present paper deals with the computational aspects of both test procedures. The efficiency changes caused by the measurement errors is also illustrated by a simulation study.

Keywords: Linear regression, Measurement errors, Rank test

References

CHENG, C. L. and VAN NESS, J. W. (1999): *Statistical Regression with Measurement Error*. Arnold, London.

GUTENBRUNNER, C. and JUREČKOVá, J. (1992): Regression rank scores and regression quantiles. *Ann. Statist. 20, 305-330.*

GUTENBRUNNER, C., JUREČKOVá, KOENKER, R. and PORTNOY, S. (1993): Tests of linear hypotheses based on regression rank scores. *J. Nonpar. Statist. 2, 307-331.*

HÁJEK, J. and ŠIDÁK, Z. and SEN, P. K. (1999): *Theory of Rank Tests*. Second Edition. Academic Press, New York.

JUREČKOVÁ, J., PICEK, J. and SALEH, A.K.MD.E. (2009): Rank tests and regression rank scores tests in measurement error models. *Computational Statistics and Data Analysis, doi:10.1016/j.csda.2009.08.020.*

Classification based on data depth

Ondrej Vencalek[1]

Charles University in Prague
KPMS MFF UK, Sokolovska 83, 186 75 Praha 8, Czech Republic,
vencalek@karlin.mff.cuni.cz

Abstract. Classification problem has been studied many times. We wish to find a rule to classify a new observation into one of $k \geq 2$ populations based on the random sample from these populations (with known classification).

During last ten years quite a lot of effort has been put into use of data depth for solving the classification problem. Data depth is a nonparametric concept which enables dealing with multidimensional data. Roughly speaking, depth function is any function providing an ordering of multivariate data. Formal definition was stated by Zuo and Serfling. Several classification rules based on data depth was developed in recent years.

Our contribution deals with some aspects of depth-based classification. We concentrate on problem of distributions with nonconvex density levelsets. Use of weighted halfspace depth, proposed in Hlubinka et al. (2010), can bring an interesting solution. We will present some results of simulation study.

Keywords: data depth, classification, nonparametric, simulation

References

HLUBINKA, D., KOTIK, L. and VENCALEK, O. (2010): Weighted halfspace depth. *Kybernetika 46 (1), 125-148.*

A Monte Carlo Simulation Study to Assess Performances of Frequentist and Bayesian Methods for Polytomous Logistic Regression

Tugba Erdem[1] and Zeynep Kalaylioglu[2]

[1] Department of Statistics Middle East Technical University, Ankara, Turkey, *terdem@metu.edu.tr*

[2] Department of Statistics Middle East Technical University, Ankara, Turkey, *kzeynep@metu.edu.tr*

Abstract. Polytomous Logistic Regression method is used to clarify the effects of covariates on categorical responses. In our study, we investigated the performances of three methods to estimate the regression parameters of polytomous logistic regression model: Maximum Likelihood Estimatation Method, Bayesian Estimatation Method and Pseudo-Conditional -Likelihood Estimation Method (PCL) for Two Stage Polytomous Logistic Regression (Chatterjee (2004)) by constructing an extensive Monte Carlo Simulation Study design. Two Stage Logistic Regression modelling is proposed for the data where the response is defined with cross-classification of disease characteristics that cause very large number of response categories. Bayesian estimation is carried out using Gibbs sampling through WinBUGS. In order to assess the performances of these three methods, we tried different scenarios, by changing the number of categories of response variable, or changing the type and number of the covariates. We noticed that Pseudo Likelihood Estimation Method is the most efficient method that the estimates obtained by PCL estimation have the smallest mean square error and bias values when the number of response categories are large. Another superiority of the PCL is the ease of interpretation of response categories when the number of levels are large.

Keywords: polytomous logistic regression, Bayesian estimation, maximum likelihood, pseudo-conditional-likelihood, WinBUGS

References

CHATTERJEE N. (2004): A Two-Stage Regression Model for Epidemiological Studies With Multivarite Disease Classification Data. *Journal of American Statistical Association, 99 (465), 127-138.*

CONGDON, P. (2005): *Bayesian Models for Categorical Data.* Wiley Series in Probability and Statistics, John Wiley and Sons Inc.

Transformed Gaussian model for joint modelling of longitudinal measurements and time-to-even under R

Inês Sousa

Department of Mathematics and Applications
Minho University, Portugal
isousa@math.uminho.pt

Abstract. We propose a model for the joint distribution of a longitudinal variable and a single time-to-event. This model, which we will call joint transformed Gaussian model, assumes a multivariate Gaussian distribution for the vector of repeated measurements expanded with the natural logarithm of failure time. The marginal distribution of time-to-event is log-Normal distributed, as well as the conditional distribution the repeated measurements. The variance covariance structure of this model is of main interest, so we consider different parametric structures for the covariance matrix of the proposed model. We will present an analysis of a data set as an example. Finally, we will discuss the possible extensions of this model in a multivariate setting. This model is proposed under a set of functions for R, which is being prepared to be submitted to CRAN.

Semi-Automated K-means Clustering

Sung-Soo Kim

Dept. of Information Statistics, Korea National Open Univ. Seoul, Korea, 110-791. *sskim@knou.ac.kr*

Abstract. The crucial problems of K-means clustering are deciding the number of clusters and initial centroids of clusters. Hence, the steps of K-means clustering are generally consisted of two-stage clustering procedure. The first stage is to run hierarchical clusters to obtain the number of clusters and cluster centroids and second stage is to run nonhierarchical K-means clustering using the results of first stage. Here we provide automated K-means clustering procedure to be useful to obtain initial centroids of clusters which can also be useful for large data sets, and provide software program implemented using R.

Keywords: K-means clustering, Ward's method, Mojena's stopping rule, Model-based clustering, BIC(Bayesian Information Criteria), Automated K-means clustering.

References

Brusco, M.J. and Cradit, J.D.(2001): A variable-selection heuristic for K-means clustering, *Psychometrika 66 (2), 249-270.*

Chen, J. Ching, R. KH and Lin, Y.(2004): An extended study of the K-means algorithm for data clustering and its applications, *Journal of the Operational Research Society, 55, 976-987.*

Fraley, C. and Raftery, A.E.(2006): MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering, *Technical Report No. 504, Dept. of Statistics Univ. of Washington.*

Mojena, R. Wishart, D. and Andrews, GB.(1980): Stopping rules for Wards' clustering method, *COMPSTAT, 426-432.*

Kim, S., Kwon, S. and Cook, D. (2000): Interactive visualization of hierarchical clusters using MDS and MST, *Metrika, Volume 51 Issue 1, 39-51.*

Application of some multivariate statistical methods to data from winter oilseed rape experiments

Zygmunt Kaczmarek[1], Elżbieta Adamska[1], Stanisław Mejza[2], Teresa Cegielska-Taras[3] and Laura Szała[3]

[1] Institute of Plant Genetics Polish Academy of Sciences
Strzeszyńska 34, PL-60-479 Poznań, Poland, *zkac@igr.poznan.pl, eada@igr.poznan.pl*
[2] Poznań University of Life Sciences
ul. Wojska Polskiego 28, 60-637 Poznań, Poland, *smejza@up.poznan.pl*
[3] Plant Breeding and Acclimatization Institute
Strzeszyńska 36, PL-60-479 Poznań, Poland, *postbox@nico.ihar.poznan.pl*

Abstract. Biotechnological production of doubled haploid (DH) lines is valuable technology in modern plant breeding of oilseed rape. It offers various advantages for plant breeders including the possibility to obtain homozygous lines rapidly. Doubled haploids with desirable agronomic characters and proper fatty acid composition of the *Brassica* oil can be chosen and introduced into breeding program as a result of selection for a combination of optimum characters. The paper presents a multivariate approach to the investigation of general and specific combining ability of doubled haploid lines of winter rape based on the analysis of their line x tester hybrids. For the yield and some interesting combination of unsaturated fatty acids: oleic (C18:1), linoleic (C18:2) and linolenic (C18:3), played the important role in biochemical and biodiesel industries, multidimensional effects of general combining ability (GCA) and specific combining ability (SCA) of parental DH lines were estimated and tested. The Mahalanobis distance has been proposed as a measure of similarity among parental DH lines due to the multi-character nature of GCA and SCA effects. The results of statistical analysis for series of 3 line x tester experiments with hybrids were also presented for individual traits (SERGEN, Caliński et al. 1998). Estimates and results of testing hypotheses concerning the interactions of GCA and SCA effects of DH lines with environments were given. Also representations of GCA effects of DH lines in the space of first two dual principal components, determined by the analysis of GCA effects by environments interactions, for seed yield and for the difference of oleic and linolenic fatty acids were illustrated.

Keywords: combining ability effects, MANOVA, series of experiments, rape

References

CALIŃSKI, T., CZAJKA, S., KACZMAREK, Z., KRAJEWSKI, P. AND SIATKOWSKI, I.(1998). *SERGEN - Analysis of series of plant genetic and breeding experiments*. Computer program for IBM-PC, Version 3. IGR PAN, Poznań.

Genotype-by-environment interaction of healthy and infected barley lines

Tadeusz Adamski[1], Zygmunt Kaczmarek[1], Iwona Mejza[2] and Maria Surma[1]

[1] Institute of Plant Genetics Polish Academy of Sciences
Strzeszyńska 34, PL-60-479 Poznań, Poland, *tada@igr.poznan.pl, zkac@igr.poznan.pl, msur@igr.poznan.pl*
[2] Poznań University of Life Sciences
ul. Wojska Polskiego 28, PL-60-637 Poznań, Poland, *imejza@up.poznan.pl*

Abstract. Barley is an important cereal crop and ranks fourth among other cereals. Many studies have revealed a significant influence of environment on phenotypic performance of agronomically important traits. Environmental factors affecting yield and its structure may be of biotic and abiotic origin. Reaction of genotypes on varying abiotic factors can be evaluated by conducting experiments over several years and/or in different localities, whereas the reaction on biotic factors (e.g. fungal pathogens) is mainly investigated in experiments in which artificial infection is applied. The aim of the studies was: (1) to assess to how the degree of infection with pathogen (*Fusarium culmorum*) may change stability and adaptability of barley genotypes, and (2) to select barley lines stable over a range of environments and more resistant to Fusarium head blight. Material for the studies covered 34 spring barley doubled haploids (DH) examined in 6-year experiment. Each line was artificially inoculated by *F. culmorum*. Yield structure traits were examined in control and inoculated plants. During each year, the experiment with k = 34 genotypes and t = 68 treatments was carried out in a group balanced block design according to Gomez and Gomez (1984). The genotypes were divided into two groups: A - inoculated plants, B - control plants. At the first step one-way analysis of variance (ANOVA) was performed for A and B groups of genotypes examined in, particular year. In the next step, statistical analysis of a series of six experiments was performed for each group independently using the computer program SERGEN (Caliński et al. 1998).

Keywords: ANOVA, genotype by environment interaction, group balanced block design

References

CALIŃSKI, T., CZAJKA, S., KACZMAREK, Z., KRAJEWSKI, P. AND SIATKOWSKI, I.(1998). *SERGEN - Analysis of series of plant genetic and breeding experiments*. Computer program for IBM-PC, Version 3. IGR PAN, Poznań.

GOMEZ, K.A. and GOMEZ, A.A. (1984): *Statistical procedures for agricultural research*. John Wiley and Sons, NY.

Selection of Summary Statistics for Approximate Bayesian Computation

David J. Balding and Matthew A. Nunes

Institute of Genetics, University College London
5 Gower Place, London WC1E 6BT, United Kingdom, *d.balding@ucl.ac.uk*

Abstract. How best to summarise high-dimensional datasets is a problem that arises in many areas of science, and in particular it plays a central role in the efficient implementation of Approximate Bayesian Computation (ABC). In ABC (Beaumont et al. (2002)), many parameter values and corresponding datasets are simulated under the chosen model. The parameter values that generate datasets similar to the one observed are used to approximate the posterior distribution of interest. This can be done for complex datasets under models with intractable likelihoods, provided only that simulation is feasible. To identify the simulated datasets most similar to that observed, datasets are summarised by statistics, often in practice chosen on the basis of the investigator's intuition. Ideally the statistics would be sufficient for the parameters of interest, but this is rarely achievable in practice. Joyce and Marjoram (2008) propose a method for choosing summary statistics based on a notion of approximate sufficiency, and Wegmann et al. (2009) suggest using partial least squares (PLS) to construct several informative statistics as orthogonal linear combinations of a larger pool of available statistics. Here, we propose minimum entropy of the resulting posterior approximation as a heuristic for selecting summary statistics. Moreover, we introduce a two-stage procedure in which summary statistics selected using the minimum-entropy criterion are used to identify an initial set of "similar" datasets, which are each successively regarded as the observed datasets in a second stage in which the average squared error of the ABC posterior approximation is minimised to identify a suitable set of summary statistics. We illustrate the performances of the minimum entropy and two-stage algorithms, and compare them with the approximate sufficiency and PLS approaches, for both univariate and bivariate posterior inferences of the scaled mutation and recombination parameters from a population sample of DNA sequences.

Keywords: Entropy, sufficiency, data reduction, population genetics

References

BEAUMONT, M. A., ZHANG, W., and BALDING, D. J. (2002) Approximate Bayesian computation in population genetics. *Genetics, 162, 2025–2035.*

JOYCE, P. and MARJORAM, P. (2008) Approximately sufficient statistics and Bayesian computation. *Stat. Appl. Gen. Mol. Biol. 7, 1–16.*

WEGMANN, D., LEUENBERGER, C., and EXCOFFIER, L. (2009) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics 182, 1207–1218.*

Choosing the Summary Statistics and the Acceptance Rate in Approximate Bayesian Computation

Michael G.B. Blum[1]

Laboratoire TIMC-IMAG, CNRS, UJF Grenoble
Faculté de Mdecine, 38706 La Tronche, France, *michael.blum@imag.fr*

Abstract. Approximate Bayesian Computation encompasses a family of likelihood-free algorithms for performing Bayesian inference in models defined in terms of a generating mechanism. The different algorithms rely on simulations of some summary statistics under the generative model and a rejection criterion that determines if a simulation is rejected or not. In this paper, I incorporate Approximate Bayesian Computation into a local Bayesian regression framework. Using an empirical Bayes approach, we provide a simple criterion for 1) choosing the threshold above which a simulation should be rejected, 2) choosing the subset of informative summary statistics, and 3) choosing if a summary statistic should be log-transformed or not.

Keywords: approximate Bayesian computation, evidence approximation, empirical Bayes, Bayesian local regression

Integrating Approximate Bayesian Computation with Complex Agent-Based Models for Cancer Research

Andrea Sottoriva[1] and Simon Tavaré[2]

[1] Department of Oncology, University of Cambridge, CRUK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK, *as949@cam.ac.uk*

[2] Department of Oncology and DAMTP, University of Cambridge, CRUK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK, *st321@cam.ac.uk*

Abstract. Multi-scale agent-based models such as hybrid cellular automata and cellular Potts models are now being used to study mechanisms involved in cancer formation and progression, including cell proliferation, differentiation, migration, invasion and cell signaling. Due to their complexity, statistical inference for such models is a challenge. Here we show how approximate Bayesian computation can be exploited to provide a useful tool for inferring posterior distributions. We illustrate our approach in the context of a cellular Potts model for a human colon crypt, and show how molecular markers can be used to infer aspects of stem cell dynamics in the crypt.

Keywords: ABC, cellular Potts model, colon crypt dynamics, stem cell modeling

Clustering Discrete Choice Data

Donatella Vicari,[1] and Marco Alf[1]

Dipartimento di Statistica, Probabilit e Statistiche Applicate
Sapienza Universit di Roma, Italy, *donatella.vicari@uniroma1.it*

Abstract. When clustering discrete choice (e.g. customers by products) data, we may be interested in partitioning individuals in disjoint classes which are homogeneous with respect to product choices and, given the availability of individual- or outcome-specific covariates, in investigating on how these affect the likelihood to be in certain categories (i.e. to choose certain products). Here, a model for joint clustering of statistical units (e.g. consumers) and variables (e.g. products) is proposed in a mixture modeling framework. A similar purpose can be found when looking for a joint partition of genes and tissues (or experimental conditions) in microarray data analysis (see e.g. Martella et al., 2008), of words and documents in web data analysis (see e.g. Li and Zha, 2006), or, in general, when latent block-based clustering is pursued (see e.g. Govaert and Nadif, 2007). Further interesting links can be established with multi-layer mixture, see e.g. Li (2005), and with hierarchical mixture of experts models, see e.g. Titsias and Likas (2002). The proposal has been sketched in the field of consumers' behavior, but can be easily extended to other research contexts, where a partition of objects and features is of interest. Further extensions of this model can be proposed by looking at different combinations for individual and product-specific covariates, and by adopting different representations for component-specific distribution parameters.

References

GOVAERT G. and NADIF M. (2003): Clustering with block mixture models. *Pattern Recognition 36(2), 463-473.*

LI J. (2005): Clustering based on a multi-layer mixture model. *Journal of Computational and Graphical Statistics 14(3), 547-568*

LI J. and ZHA H. (2006): Two-way Poisson mixture models for simultaneous document classification and word clustering. *Computational Statistics and Data Analysis 50(1), 163-180.*

MARTELLA, F., ALFO', M., VICHI, M. (2008): Biclustering of gene expression data by an extension of mixtures of factor analyzers. *The International Journal of Biostatistics 4(1), art. 3.*

TITTERINGTON, D.M., SMITH, A.F.M. and MAKOV, U.E. (1985): *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, Chichester.

TITSIAS K. and LIKAS A. (2002): Mixture of Experts Classification Using a Hierarchical Mixture Model. *Neural Computation 14(9), 2221-2244.*

VERMUNT, J.K. (2007): A hierarchical mixture model for clustering three-way data sets. *Computational Statistics and Data Analysis 51, 5368-5376.*

VERMUNT, J.K. (2008): Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research 17, 33-51.*

Multiple Nested Reductions of Single Data Modes as a Tool to Deal with Large Data Sets

Iven Van Mechelen and Katrijn Van Deun

Centre for Computational Systems Biology (SymBioSys), KULeuven
Tiensestraat 102 - box 3713, 3000 Leuven, Belgium,
Iven.VanMechelen@psy.kuleuven.be

Abstract. The increased accessibility and concerted use of novel measurement technologies give rise to a data tsunami with matrices that comprise both a high number of variables and a high number of objects. As an example, one may think of transcriptomics data pertaining to the expression of a large number of genes in a large number of samples or tissues (as included in various compendia). The analysis of such data typically implies ill-conditioned optimization problems, as well as major challenges on both a computational and an interpretational level.

In the present paper, we develop a generic method to deal with these problems. This method was originally briefly proposed by Van Mechelen and Schepers (2007). It implies that single data modes (i.e., the set of objects or the set of variables under study) are subjected to multiple (discrete and/or dimensional) nested reductions.

We first formally introduce the generic multiple nested reductions method. Next, we show how a few recently proposed modeling approaches fit within the framework of this method. Subsequently, we briefly introduce a novel instantiation of the generic method, which simultaneously includes a two-mode partitioning of the objects and variables under study (Van Mechelen et al. (2004)) and a low-dimensional, principal component-type dimensional reduction of the two-mode cluster centroids. We illustrate this novel instantiation with an application on transcriptomics data for normal and tumourous colon tissues.

In the discussion, we highlight multiple nested mode reductions as a key feature of the novel method. Furthermore, we contrast the novel method with other approaches that imply different reductions for different modes, and approaches that imply a hybrid dimensional/discrete reduction of a single mode. Finally, we show in which way the multiple reductions method allows a researcher to deal with the challenges implied by the analyis of large data sets as outlined above.

Keywords: high dimensional data, clustering, dimension reduction

References

VAN MECHELEN, I., BOCK, H.-H. and DE BOECK, P. (2004): Two-mode clustering methods: A structural overview. *Statistical Methods in Medical Research 13, 363-394.*

VAN MECHELEN, I. and SCHEPERS, J. (2007): A unifying model involving a categorical and/or dimensional reduction for multimode data. *Computational Statistics and Data Analysis 52, 537-549.*

The Generic Subspace Clustering Model

Marieke E. Timmerman[1] and Eva Ceulemans[2]

[1] Heymans Institute for Psychology, University of Groningen
Grote Kruisstraat 2/1, 9712TS Groningen, the Netherlands,
m.e.timmerman@rug.nl
[2] Centre for Methodology of Educational Research, Catholic University of Leuven
A. Vesaliusstraat 2, BE-3000 Leuven, Belgium, *Eva.Ceulemans@ped.kuleuven.be*

Abstract. In this paper we present an overview of methods for clustering high dimensional data in which the objects are assigned to mutually exclusive classes in low dimensional spaces. To this end, we will introduce the generic subspace clustering model. This model will be shown to encompass a range of existing clustering techniques as special cases. As such, further insight is obtained into the characteristics of these techniques and into their mutual relationships. This knowledge facilitates selecting the most appropriate model variant in empirical practice.

Keywords: reduced k-means, common and distinctive cluster model, mixture model

Yield Curve Predictability, Regimes, and Macroeconomic Information: A Data-Driven Approach

Francesco Audrino[1] and Kameliya Filipova[2]

[1] Institute of Mathematics and Statistics, University of St. Gallen, Bodanstrasse 6, CH-9000 St. Gallen, Switzerland. *francesco.audrino@unisg.ch*
[2] Institute of Mathematics and Statistics, University of St. Gallen, Bodanstrasse 6, CH-9000 St. Gallen, Switzerland. *kameliya.filipova@unisg.ch*

Abstract. We propose an empirical approach to determine the various economic sources driving the US yield curve. We allow the conditional dynamics of the yield at different maturities to change in reaction to past information coming from several relevant predictor variables. We consider both endogenous, yield curve factors and exogenous, macroeconomic factors as predictors in our model, letting the data themselves choose the most important variables. We find clear, different economic patterns in the local dynamics and regime specification of the yields depending on the maturity. Moreover, we present strong empirical evidence for the accuracy of the model in fitting in-sample and predicting out-of-sample the yield curve.

Keywords: Yield curve modeling and forecasting; Macroeconomic variables; Tree-structured models; Threshold regimes; Bagging.

Performance Assessment of Optimal Allocation for Large Portfolios

Fabrizio Laurini[1] and Luigi Grossi[2]

[1] Università di Parma, Dipartimento di Economia
Via Kennedy 6, 43100, Parma, Italy, *fabrizio.laurini@unipr.it*
[2] Università di Verona, Dipartimento di Economia
Via Dell'Artigliere 19, 37129, Verona, Italy, *luigi.grossi@univr.it*

Abstract. We consider the problem of optimal asset allocation for portfolio with a large number of shares. The numerical solution relies on the estimation of the covariance matrix between the assets. Such estimation, typically obtained with maximum likelihood, is affected be the so-called "maximization estimation error", which grows with the dimension of the covariance matrix. The use of a robust estimator of the covariance matrix can reduce such estimation error considerably, even when data are outlier free and outperform the standard approaches when data have marked heavy tails or affected by the presence of outliers. The performance of our new robust estimator is studied with simulations, and real data.

Keywords: Financial asset allocation, influential data, robust estimators

Some Examples of Statistical Computing in France During the 19th Century

Antoine de Falguerolles

Université de Toulouse, IMT, Laboratoire de Statistique et Probabilités
118 route de Narbonne, F-31068 Toulouse, France,
Antoine.Falguerolles@math.univ-toulouse.fr

Abstract. Statistical computing emerged as a recognised topic in the seventies. Remember the first COMPSTAT symposium held in Vienna (1974)! But the need for proper computations in statistics arose much earlier. Indeed, the contributions by Laplace (1749-1829) and Legendre (1752-1833) to statistical estimation in linear models are well known. But further works of computational interest originated in the structuring of the concept of regression during the 19th century. While some were fully innovative, some appear now unsuccessful but nevertheless informative. The paper discusses, from a French perspective, the computational aspects of selected examples.

Section 2 combines approaches in covariance analysis developed in two different areas, namely the econometry of road building (Georges Müntz, 1834) and the metrology for cartographic triangulation (Aimé Laussedat, 1860). Section 3 details Augustin Cauchy's heuristic for simple and multiple regression (1837). An example of use of Cauchy's method can be found in an article by Vilfredo Pareto (1897). Section 4 reviews an outline, in a particular situation, of what is called now the iteratively weighted least squares algorithm which Pareto introduces in the same paper. Finally, section 5 exemplifies two cases of modest treatment with genuine intuitions. One is by an anonymous subscriber (1821) and addresses the computing of a weighted mean with data driven weights; the second is the famous introduction of the Gamma distribution for the modeling of the distribution of wages by Lucien March (1898).

It turns out that most papers mentioned in this article are authored by former graduates and/or professors from the *École Polytechnique*. This institution is certainly the common denominator for mathematical statistics and statistical computing in the 19th century in France. But the *Conservatoire National des Arts et Métiers* is not too far behind.

Keywords: history of statistics, regression, statistical computing

Modeling Operational Risk: Estimation and Effects of Dependencies

Stefan Mittnik, Sandra Paterlini, and Tina Yener

Center for Quantitative Risk Analysis (CEQURA), Department of Statistics, LMU Munich, Germany

Abstract. Being still in its early stages, operational risk modeling has, so far, mainly been concentrated on the marginal distributions of frequencies and severities within the context of the Loss Distribution Approach (LDA). In this study, drawing on a fairly large real–world data set, we analyze the effects of competing strategies for dependence modeling. In particular, we estimate tail dependence both via copulas as well as nonparametrically, and analyze its effect on aggregate risk–capital estimates.

Keywords: operational risk, risk capital, value–at–risk, correlation, tail dependence

Influence of the Calibration Weights on Results Obtained from Czech SILC Data

Jitka Bartošová and Vladislav Bína

University of Economics in Prague, Faculty of Management, Jarošovská 1117/II, 37701 Jindřichův Hradec, Czech Republic, {*bartosov, bina*}*@fm.vse.cz*

Abstract. The purpose of income sample survey is to obtain a representative data concerning level and structure of incomes and fundamental social-demographic characteristics of households and their members. The survey results of inhabitant's income in the Czech Republic (SILC and Mikrocensus) are generalized on the whole population using the calibration weights created by Czech Statistical Office. The important question in the process of generalizing conclusions arising from the analysis of data files is the influence of calibration weighting on the results. The contribution concerns the effect of calibration process on the measuring of monetary poverty in the Czech Republic (see BARTOŠOVÁ, J. (2009a), BARTOŠOVÁ, J. (2009b) or BARTOŠOVÁ, J., BÍNA, V. (2009a)). The paper presents significant changes of the results concerning the threshold of monetary poverty which appear in dependence on the definition of consuming unit (see NICODEMO, C., LONGFORD, N.,T. (2009)).

Keywords: Calibration weights, EU-SILC, household incomes, poverty.

References

BARTOŠOVÁ, J. (2009a): Analysis and Modelling of Financial Power of Czech Households. *Aplimat – Journal of Applied Mathematics 2 (3), STU Bratislava, 31–36.*

BARTOŠOVÁ, J. (2009b): Výběrové šetření příjmů domácností v České republice. *Forum Statisticum Slovacum 2 (7), SŠDS, Bratislava, 4–9.*

BARTOŠOVÁ, J. and BÍNA, V. (2009a): Modelling of income distribution of Czech households in years 1996 - 2005. *Acta Oeconomica Pragensia 17 (4), Oeconomica, Prague, 3–18.*

NICODEMO, C., LONGFORD, N.,T. (2009): A sensitivity analysis of poverty definitions used with EU-SILC. In: *Finanční potenciál domácností 2009 (Proceedings of workshop of GAČR 402/09/0515). University of Economics in Prague, CDROM.*

A Markov Switching Re-evaluation of Event-Study Methodology

Rosella Castellano[1] and Luisa Scaccia[2]

[1] Dip. di Istituzioni Economiche e Finanziarie, Università di Macerata via Crescimbeni 20, 62100 Macerata, Italy, *castellano@unimc.it*
[2] Dip. di Istituzioni Economiche e Finanziarie, Università di Macerata via Crescimbeni 20, 62100 Macerata, Italy, *scaccia@unimc.it*

Abstract. This paper reconsiders event-study methodology in light of evidences showing that Cumulative Abnormal Return (CAR) can result in misleading inferences about financial market efficiency and pre(post)-event behavior. In particular, CAR can be biased downward, due to the increased volatility on the event day and within event windows. We propose the use of Markov Switching Models to capture the effect of an event on security prices. The proposed methodology is applied to a set of 45 historical series on Credit Default Swap (CDS) quotes subject to multiple credit events, such as reviews for downgrading. Since CDSs provide insurance against the default of a particular company or sovereign entity, this study checks if market anticipates reviews for downgrading and evaluates the time period the announcements lag behind the market.

Keywords: Hierarchical Bayes, Markov switching models, credit default swaps, event-study

Neural Network Approach for Histopathological Diagnosis of Breast Diseases with Images

Yuichi Ishibashi[1], Atsuko Hara[2], Isao Okayasu[2], and Koji Kurihara[1]

[1] Graduate School of Environmental Science, Okayama University, Okayama 700-8530, Japan, *ishibashi@ems.okayama-u.ac.jp*
[2] Kitasato University School of Medicine, Sagamihara, Kanagawa, 228-8555, Japan

Abstract. Diagnosis of breast diseases relies on recognizing diseased tissue in histopathological images. The tissues studied will contain both diseased and normal areas. To insure a correct diagnosis a method is described here that is made up of three steps. The 1st step is to subdivide the histopathological image into sections. These subdivisions will then all be digitized (step 2). Several methods were tested and Wavelet transformation was found to be the best. The final step was evaluation by neural network analysis. The collective evaluation of subdivisions will increase the accuracy of diagnosis and help to avoid missing cancerous or inflamed tissue. In some studies (ref) malignancy of cancer was measured by support vector machine etc. in histopathology, but identification of the kind of cancer by pattern recognition is new. Our study attempts to digitize the features of tissue pattern for each kind of disease and to recognize the kind of disease by neural network.

Keywords: Neural network,Wavelet transformation,Learning Vector Quantization,histopathological diagnosis,breast disease

References

ARAI, K. (2000): *Fundamental Theory on Wavelet Analysis.* Morikita Shuppan (in Japanese),70-71.

ISHIBASHI, Y. et al. (2010): Statistical analysis of histopathological diagnosis reports with text mining, *Joint meeting of Japan-Korea Special Conference of Statistics,227-230.*

JIN, M. (2007): *Data Science with R.* Morikita Shuppan (in Japanese),247-255.

KANAMORI, T. et al. (2009): *Pattern Recognition.* Kyoritsu Shuppan (in Japanese),100-106.

KOHONEN, T. et al. (1995): LVQ PAK:The Learning Vector Quantization Program Package Technical Report. *Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, FINLAND,5-10.*

KUMAGAI, J. and ITO, T. (2006): Histopathological diagnosis by image processing. *Pathology and Clinic, 24(4),387-391.*

MUKAI, K. (1999): Computer Application in Pathology -Possibilities and Problems- *Utilization of computer and internet in the pathological field, Saitama, Japan.*

Variable Inclusion and Shrinkage Algorithm in High Dimension

Mkhadri Abdallah[1] and Ouhourane Mohamed[2]

[1] Department of Mathematics, Faculty of Sciences-Semlalia, B.P. 2390, Marrakech, Morocco. *mkhadri@ucam.ac.ma*
[2] Department of Mathematics, Faculty of Sciences-Semlalia, B.P. 2390, Marrakech, Morocco. *hourali@hotmail.com*

Abstract. We propose a new method to simultaneously select variables and encourage a grouping effect where strongly correlated predictors tend to be in or out of the model together. Moreover, our method is capable of selecting sparse models while avoiding over shrinkage of Lasso. It combines the idea of VISA algorithm which avoides over shrinkage of regression coefficients and those of the Elastic Net which overcomes the limitation of Lasso in high dimension. Our method is based on a modified VISA algorithm, so is also computationally efficient. A detailed simulation study in small and high dimensional settings is performed, which illustrates the advantages of our approach in relation to several other possible methods.

Keywords: Variable selection, VISA algorithm, LARS, Linear Regression.

References

EFRON, B. HASTIE, T. JOHNSTONE, I. and TIBSHIRANI, R. (2004): Least angle regression. *Annals of Statistics*, **32**, 407-499.

MEINSHAUSEN, N. (2007): Relaxed lasso. *Computational Statistics & Data Analysis*, **52**, 374-393.

RADCHENKO, P. and JAMES, G. M.(2008): Variable inclusion and shrinkage algorithms. *Journal of the Americain Statistical Association*, **103**, n 483, 1304-1315.

TIBSHIRANI, R.(1996): Regression shrinkage and selection via the Lasso. *journal of the Royal Statistical Society, series B*, **58**, 267-288.

ZOU, H. and HASTIE, T. (2005): Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, series B*, **67**, 301-320.

Support Vector Machines for Large Scale Text Mining in R

Ingo Feinerer[1] and Alexandros Karatzoglou[2]

[1] Database and Artificial Intelligence Group
Institute of Information Systems
Vienna University of Technology
Austria *Ingo.Feinerer@tuwien.ac.at*
[2] LITIS, INSA de Rouen
Avenue de Universite
76801 Saint-Etienne du Rouvray
France *alexis@ci.tuwien.ac.at*

Abstract. SVM are an established tool in machine learning and data analysis. Though many implementations of SVM exist often specific applications require tailor made algorithms. In text mining in particular the data often comes in large sparse data matrices. Typical SVM algorithms like SMO do not take advantage of the sparsity, and do not scale well to data sets with millions of entries. In this paper we present an implementation of linear SVM's for R that address both of these issues.

Keywords: SVM, text mining, large scale

Random Forests Based Feature Selection for Decoding fMRI Data

Robin Genuer[1,2], Vincent Michel[1,2,3,5], Evelyn Eger[4,5], and Bertrand Thirion[3,5]

[1] Université Paris-Sud 11, Mathématiques, Orsay, France
[2] Select team, INRIA Saclay-Île-de-France, France
[3] Parietal team, INRIA Saclay-Île-de-France, France
[4] INSERM U562, Gif/Yvette, France
[5] CEA, DSV, I2BM, NeuroSpin, Gif/Yvette, France

Abstract. In this paper we present a new approach for the prediction of a behavioral variable from Functional Magnetic Resonance Imaging (fMRI) data. The difficulty in this problem comes from the huge number of image voxels that may provide relevant information with respect to the limited number of available images. A very common solution consists in using feature selection techniques, i.e. to evaluate the significance of each individual brain region with respect to the target information, and then to use the best ranked features as input to a classifier, such as linear Support Vector Machines (SVM; we take this as the *reference method*). However, this kind of scheme ignores the correlations between features, so that it is potentially suboptimal, and it does not generally provide an interpretable pattern of predictive voxels. Based on Random Forests, our approach provides an accurate auto-calibrated framework for selecting a set of very few jointly informative regions. Comparisons with the reference method on real data show that our approach yields a little bit higher classification performance, but the real gain comes from the sparsity of our variable selection.

Keywords: Feature selection, Variable Importance, Random forests, Classification, fMRI

References

BREIMAN, L. (2001): Random Forests. *Machine Learning 45 ,5-32.*

COX, D.D. and SAVOY, R.L. (2003): Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage 19 (2),261–270.*

DAYAN, P. and ABBOTT, L.F. (2001): *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems.* The MIT Press.

EGER, E., KELL, C. and KLEINSCHMIDT, A. (2008): Graded size sensitivity of object exemplar evoked activity patterns in human LOC subregions. *Journal of Neurophysiology 100 (4) , 2038-47.*

GENUER, R., POGGI, J.-M. and TULEAU, C. (2010): Variable selection using random forests. *Pattern Recognition Lett.* doi:10.1016/j.patrec.2010.03.014

Peak Detection in Mass Spectrometry Data Using Sparse Coding

Theodore Alexandrov[1], Klaus Steinhorst[1], Oliver Keszöcze[1], and Stefan Schiffler[1]

University of Bremen, Center for Industrial Mathematics
Bibliothekstr. 1, D-28334 Bremen, Germany *theodore@math.uni-bremen.de*

Abstract. Mass spectrometry is an important tool in the analysis of chemical compounds. A crucial step of mass spectrometry data processing is the peak detection which selects peaks corresponding to molecules with high concentrations. We present a new procedure of the peak detection based on a sparse coding algorithm, for which we propose an elastic-net modification in Alexandrov et al. (2009). The evaluation with simulated data shows that using the sparse coding prototype spectra gives improvement over using per-class mean spectra, although the former ones are extracted in an unsupervised manner. Finally, we apply the procedure to a colorectal cancer of de Noo et al. (2006) and to a liver diseases of Ressom et al. (2007) mass spectrometry datasets.

Interestingly, the prototype spectra are similar to per-class mean spectra, although are obtained in an unsupervised manner. Hence, our peak detection procedure can be applied when assignments of spectra to classes are unknown. We have detected peaks in prototype spectra with a simple method to demonstrate the potential of our approach and expect that application of a more advanced peak detection method can improve the results.

Keywords: mass spectrometry, peak picking, sparse coding

References

ALEXANDROV, T., KESZÖCZE, O., LORENZ, D. A., SCHIFFLER, S. and STEINHORST, K. (2009): An active set approach to the elastic-net and its applications in mass spectrometry. In *Proc. Signal Processing with Adaptive Sparse Structured Representations (SPARS'09)*. Available at http://hal.inria.fr/inria-00369397.

DE NOO, M., MERTENS, B., OZALP, A., BLADERGROEN, M., VAN DER WERFF, M., VAN DE VELDE, C., DEELDER, A. and TOLLENAAR, R. (2006): Detection of colorectal cancer using MALDI-TOF serum protein profiling. *Eur. J. Cancer 42(8):1068-76.*

LEE, H., BATTLE, A., RAINA, R. and NG A. Y.. (2006): Efficient sparse coding algorithms. In *Proc. Neural Information Processing Systems (NIPS'06)*, 801–8.

RESSOM, H. W., VARGHESE, R. S., DRAKE, S., HORTIN, G. L., ABDEL-HAMID, M., LOFFREDO, C. A. and GOLDMAN, R.(2007): Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics 23(5):619-26.*

Adaptive mixture discriminant analysis for supervised learning with unobserved classes

Charles Bouveyron

Laboratoire SAMM, Université Paris 1 Panthéon–Sorbonne
90 rue de Tolbiac, 75013 Paris, France, *charles.bouveyron@univ-paris1.fr*

Abstract. In supervised learning, an important issue usually not taken into account by classical methods is the possibility of having in the test set individuals belonging to a class which has not been observed during the learning phase. Classical supervised algorithms will automatically label such observations as belonging to one of the known classes in the training set and will not be able to detect new classes. This work introduces a model-based discriminant analysis method, called adaptive mixture discriminant analysis (AMDA), which is able to detect unobserved groups of points and to adapt the learned classifier to the new situation. Two EM-based procedures are proposed for the parameter estimation in an inductive and a transductive way respectively. Experimental studies will demonstrate the ability of the proposed method to deal with complex and real word problems.

Keywords: supervised classification, unobserved classes, adaptive learning, novelty detection, model selection.

Improvement of acceleration of the ALS algorithm using the vector ε algorithm

Masahiro Kuroda[1], Yuchi Mori[2], Masaya Iizuka[3], and Michio Sakakihara[4]

[1] Department of Socio-Science, Okayama University of Science
1-1 Ridaicho, Kita-ku, Okayama, Japan, *kuroda@soci.ous.ac.jp*
[2] Department of Socio-Science, Okayama University of Science
1-1 Ridaicho, Kita-ku, Okayama, Japan, *mori@soci.ous.ac.jp*
[3] Graduate School of Environmental Science, Okayama University
1-1-1 Tsushima-naka, Kita-ku, Okayama, Japan, *iizuka@ems.okayama-u.ac.jp*
[4] Department of Information Science, Okayama University of Science
1-1 Ridaicho, Kita-ku, Okayama, Japan, *sakaki@mis.ous.ac.jp*

Abstract. Principal components analysis (PCA) is a popular descriptive multivariate method for handling quantitative data and can be extended to deal with qualitative data and mixed measurement levels data. For these extended PCA methods, the alternating least squares (ALS) algorithm is utilized. This type of algorithm based on least squares estimation may require many iterations in its application to very large data sets and variable selection problems and thus take a long time to converge. Kuroda et al. (2008) proposed an iterative algorithm for speeding up the convergence of the ALS algorithm using the vector ε algorithm of Wynn (1962) that enables the acceleration of convergence of a slowly convergent vector sequence and is very effective for linearly converging sequences. In this paper, we derive a new version of the proposed algorithm which is not modified the original acceleration algorithm but includes additional re-starting criteria for reducing both of the number of iterations and the computational time. Numerical experiments examine the performance of the new algorithm in comparison with the original acceleration algorithm.

Keywords: Principal components analysis, the alternating least squares algorithm, the vector ε algorithm, restarting criteria, acceleration

References

KURODA, M., MORI, Y., IIZUKA, M. and SAKAKIHARA, M. (2008): Accerelation of convergence of the alternating least squares algorithm for principal component analysis. *Program & Abstracts IASC 2008, 172.*

WYNN, P. (1962): Acceleration techniques for iterated vector and matrix problems. *Mathematics of Computation 16, 301-322.*

Numerical methods for some classes of matrices with applications to Statistics and Optimization

J. M. Peña[1]

Departamento de Matemática Aplicada
Universidad de Zaragoza, 50009 Zaragoza, Spain, *jmpena@unizar.es*

Abstract. Recent advances on numerical methods for matrices with applications to Statistics and Optimization are presented. They include results on the localization of eigenvalues, optimally conditioned matrices and error bounds for the linear complementarity problem (see Cortés and Peña (2008), Delgado and Peña (2009), García-Esnaola and Peña (2009, 1 and 2), Peña (2009)). We consider sign-regular matrices as well as some subclasses of these matrices. The interest of these matrices comes from their characterization as variation diminishing linear maps, which leads to a unified presentation of hypothesis tests (see Brown et al. (1981)). They can also present interesting probabilistic interpretations (see Goldman (1985)). We also consider the class of H-matrices, which plays an important role in linear complementarity problems, and some classes of P-matrices (see Chen and Xiang (2006)).

Keywords: Numerical algorithms, statistical computing, sign-regular matrices, H-matrices, error bounds, conditioning

References

BROWN, L. D., JOHNSTONE, I. M. and MacGIBBON, K. B. (1981): Variation Diminishing transformations: a direct approach to total positivity and its statistical applications. *J. Amer. Statist. Assoc. 76, 824-832.*

CHEN, X. and XIANG, S. (2006): Computation of error bounds for P-matrix linear complementarity problems. *Math. Program., Ser. A, 106, 513-525.*

CORTES, V. and PEÑA J. M. (2008): A stable test for strict sign regularity. *Math. Comp. 77, 2155-2171.*

DELGADO, J. and PEÑA, J. M. (2009): Optimal conditioning of Bernstein collocation matrices. *SIAM J. Matrix Anal. Appl. 31, 990-996.*

GARCIA-ESNAOLA, M. and PEÑA, J. M. (2009, 1): Sign consistent linear programing problems. *Optimization 58, 935-946.*

GARCIA-ESNAOLA, M. and PEÑA, J. M. (2009, 2): Error bounds for linear complementarity problems for B-matrices. *Appl. Math. Lett. 22, 1071-1075.*

GOLDMAN, R. N. (1985): Pólya's urn model and computer aided geometric design. *SIAM J. Algebraic Discrete Methods 6, 1-28.*

PEÑA, J. M. (2009): Eigenvalue bounds for some classes of P-matrices. *Numerical Linear Algebra with Applications 16, 871-882.*

Fisher Scoring for Some Univariate Discrete Distributions

Thomas W. Yee

Department of Statistics, University of Auckland, Private Bag 92019,
Auckland Mail Centre, Auckland 1142, New Zealand, *t.yee@auckland.ac.nz*

Abstract. The classes of vector generalized linear and additive models (VGLMs and VGAMs; Yee and Wild (1996), Yee and Hastie (2003)) enables maximum likelihood estimation of many models and distributions including categorical data analysis, survival analysis, time series, data, nonlinear least-squares models, and scores of standard and nonstandard univariate and continuous distributions. Usually Fisher scoring is used for these. This paper focusses on univariate discrete distributions, e.g., the negative binomial, zero-inflated and zero-altered Poisson and negative binomial distributions, the zeta and Zipf distributions, etc. A selection of topics will be chosen, e.g., the choice of initial values that are robust to outliers is often as much an art as it is a science. The author's VGAM package for R is used for illustration.

Keywords: maximum likelihood estimation, Fisher scoring, univariate discrete distributions, vector generalized linear and additive models, VGAM package for R.

References

YEE, T. W., HASTIE, T. J. (2003): Reduced-rank vector generalized linear models. *Statistical Modelling 3(1), 15–41.*

YEE, T. W., WILD, C. J. (1996): Vector generalized additive models. *Journal of the Royal Statistical Society B, 58(3), 481–493.*

Numerical Error Analysis for Statistical Software on Multi-Core Systems

Wenbin Li[1] and Sven Simon[2]

[1] SimTech & IPVS, Stuttgart University
Universitatsstr. 38, Stuttgart, Germany, *liwn@ipvs.uni-stuttgart.de*
[2] SimTech & IPVS, Stuttgart University
Universitatsstr. 38, Stuttgart, Germany, *simon@ipvs.uni-stuttgart.de*

Abstract. In statistical software packages, usually no information about the numerical accuracy of the computed result is available. This leads to a risk of misinterpretation of inaccurate results (Keeling and Pavur (2007), McCullough (1998), McCullough (1999)). The Discrete Stochastic Arithmetic (DSA: Vignes (1988)) provides estimation of numerical accuracy with respect to rounding error propagation. In this paper, the DSA is applied to a statistical package, benchmark results are presented to illustrate its effectiveness. With the DSA, it is possible to estimate the accuracy of any computed result of user's application without the requirement of reference solutions. However, along with its effectiveness and reliability in numerical error analysis, the DSA suffers from computational bottlenecks due to multiple runs of the code with random rounding. For acceleration of the DSA, parallelization approaches on multi-core systems are also investigated. The proposed parallelization approach takes benefit from the increasing parallel computational power of multi-core CPUs, and shows an almost linear scalability.

Keywords: numerical accuracy, DSA, multi-core, parallelization

References

Keeling, K. B. and Pavur, R. J. (2007): A comparative study of the reliability of nine statistical software packages. *Computational Statistics & Data Analysis 51(8), 3811-3831.*

McCullough, B. D. (1998): Assessing the reliability of statistical foftware: part I. *The American Statistician 52(4), 358-366.*

McCullough, B. D. (1999): Assessing the reliability of statistical software: part II. *The American Statistician 53(2), 149-159.*

Vignes, J. (1988): Review on stochastic approach to round-off error analysis and its applications. *Mathematics and Computers in Simulation 30(6), 481-491.*

Computational Statistics: the Symbolic Approach

Colin Rose

Theoretical Research Institute, Sydney
66 Drumalbyn Road, Bellevue Hill, NSW 2023, Australia *colin@tri.org.au*

Abstract. There is a somewhat old-fashioned tendency to think of computational statistical software as a numerical tool for working with data or models. By contrast, in this paper, we illustrate the current state of the symbolic approach to computational statistics, providing examples using mathStatica (2010) which is based on top of the *Mathematica* computer algebra system. Symbolic toolkits enable one to derive exact arbitrary new theoretical results, essentially in real-time. Of course, new tools bring new problems ... and we comment briefly on the changing nature of proof and epistemology in such a world.

Keywords: computer algebra systems, symbolic methods, mathStatica

Quantile Regression for Group Effect Analysis

Cristina Davino[1] and Domenico Vistocco[2]

[1] Dipartimento di Studi sullo Sviluppo Economico, Università di Macerata
Piazza Oberdan 3, Macerata, Italy, *cdavino@unimc.it*
[2] Dipartimento di Scienze Economiche, Università di Cassino
Via S.Angelo S.N., Cassino, Italy, *vistocco@unicas.it*

Abstract. The aim of the paper is to propose an innovative approach based on quantile regression to identify a typology. The detection of a typology could derive either from the clustering of units into groups or from the analysis of the differences among a priori defined groups. In spite of this double potential meaning, the present paper focuses only on the analysis of the differences among groups using an available stratification variable.

The methodological framework is represented by quantile regression, as introduced by Koenker and Basset (1978). This method may be considered as an extension of classical least squares estimation of conditional mean models to the estimation of a set of conditional quantile functions. The use of quantile regression offers a more complete view of the relationships among variables, providing a method for modelling the rates of changes in the response variable at multiple points of its conditional distribution. As the independent variables could affect the response variable in different ways at different locations of its conditional distribution, useful insights derive from extracting information at other places other than the expected value.

It is a matter of fact that if two units have similar features/behaviours or belong to the same group of a stratification variable, the dependence structure of a regression model is more alike. The approach proposed in this paper aims to estimate group effects in a regression model taking into account the impact of the regressors on the entire conditional distribution of the dependent variable.

An empirical analysis is also provided to measure the changes on job satisfaction owing to modification of the evaluation of different job features and taking into account the type of job (self-employed, private employee or public employee). This latter variable is used to estimate the group effects.

Keywords: Quantile Regression, Group Effects

References

DAVINO, C., VISTOCCO, D. (2007): The evaluation of university educational processes: a quantile regression approach. *STATISTICA*, Bologna, n.3, pp. 267-278.

GELMAN, A. HILL, J. (2006): *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

KOENKER, R., BASSET, G.W. (1978): Regression Quantiles, *Econometrica 46, 33-50*.

Ordinary Least Squares for Histogram Data Based on Wasserstein Distance

Rosanna Verde and Antonio Irpino[1]

Dipartimento di Studi Europei e Mediterranei,
Seconda Universitá degli Studi di Napoli
Via del Setificio, Caserta, Italy, {*rosanna.verde, antonio.irpino*}*@unina2.it*

Abstract. Histogram data is a kind of symbolic representation which allows to describe an individual by an empirical frequency distribution. In this paper we introduce a linear regression model for histogram variables. We present a new Ordinary Least Squares approach for a linear model estimation, using the Wasserstein metric between histograms. In this paper we suppose that the regression coefficient are scalar values. After having illustrated the concurrent approaches, we corroborate the proposed estimation method by an application on a real dataset.

Keywords: probability distribution function, histogram data, ordinary least squares, Wasserstein distance

A Clusterwise Center and Range Regression Model for Interval-Valued Data

Francisco de A. T. de Carvalho[1], Gilbert Saporta[2], and Danilo N. Queiroz[1]

[1] Centro de Informática - CIn/UFPE
Av. Prof. Luiz Freire, s/n - Cidade Universitária, CEP 50740-540, Recife-PE, Brazil {*fatc,dnq*}*@cin.ufpe.br*
[2] Chaire de statistique appliquée & CEDRIC, CNAM
292 rue Saint Martin, Paris, France, *gilbert.saporta@cnam.fr*

Abstract. Symbolic interval-valued data occur in two contexts: either when one has uncertainty on individual values, or when one has variation like eg in medical data such as blood pressure, pulse rate observed on a daily time period. We will consider here only the second case. Several methods have been proposed to deal with the case where the response y as well as the predictors are interval-valued variables. We will use the centre and range method proposed by Lima Neto and De Carvalho (2008). Assuming that data are homogeneous (ie there is only one regression model for the whole data set) can be misleading. Clusterwise regression has been proposed long ago, as a way to identify both the partition of the data and the relevant regression models, one for each class. This paper aims to adapt clusterwise regression to interval-valued data. The proposed approach combines the dynamic clustering algorithm with the center and range regression method for interval-valued data in order to identify both the partition of the data and the relevant regression models, one for each cluster. Experiments with a car interval-valued data set show the usefulness of combining both approaches.

Keywords: Clusterwise Regression, Interval-Valued Data, Symbolic Data Analysis

References

BILLARD, L. and DIDAY, E. (2007): *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley-Interscience, San Francisco.

DIDAY, E. and SIMON, J.C. (1976): Clustering analysis. In: K.S. Fu (Eds.): *Digital Pattern Classification*. Springer, Berlin, 47–94.

D'URSO, P. and SANTORO, A. (2006): Fuzzy clusterwise linear regression analysis with symmetrical fuzzy output variable. *Computational Statistics and Data Analysis, 51 (1): 287–313*.

HENNIG, C. (2000): Identifiability of models for clusterwise linear regression. *J. Classification 17 (2), 273-296*.

LIMA NETO, E. A. and DE CARVALHO, F.A.T. (2008): Centre and Range method for fitting a linear regression model to symbolic interval data. *Computational Statistics and Data Analysis, 52 (3): 1500–1515*.

SPAETH, H. (1979): Clusterwise Linear Regression. *Computing 22 (4), 367–373*.

A Decision Tree for Interval-valued Data with Modal Dependent Variable

Djamal Seck[1], Lynne Billard[2], Edwin Diday[3] and Filipe Afonso[4]

[1] Departement de Mathematiques et Informatique,
Universit*é* Cheikh Anta Diop de Dakar, Senegal *djamal.seck@ucad.edu.sn*,
[2] Department of Statistics, University of Georgia
Athens GA 30602 USA *lynne@stat.uga.edu*
[3] CEREMADE, University of Paris Dauphine
75775 Paris Cedex 16 France *edwin.diday@ceremade.dauphine.fr*
[4] Syrokko, A*é*ropôle de Roissy, Bat. A*é*ronef, 5 rue de Copenhague, 95731 Roissy Charles de Gaulle Cedex France, *afonso@syrokko.com*

Abstract. The CART (Breiman et al., 1984) methodology for classical data is extended to symbolic data in Seck (2010). This new methodology, called STREE, is capable of building a pure CART tree, a pure divisive hierarchy, or a weighted combination of both. This paper presents an application of STREE and compares its results with the traditional CART analysis.

Keywords: Classification regression tree, divisive hierarchy tree, Fisher's iris data.

Data Management in Symbolic Data Analysis

Teh Amouh[1], Monique Noirhomme-Fraiture[1], and Benoit Macq[2]

[1] Faculté d'informatique, FUNDP
21 rue Grandgagnage, 5000 Namur, Belgique, {*tam, mno*}*@info.fundp.ac.be*
[2] Université catholique de Louvain, UCL
2 Place Stevin, 1348 Louvain-la-neuve, Belgique, *benoit.macq@uclouvain.be*

Abstract. Relational databases are now ubiquitious in industrial companies and public administrations. They contain first level units described by numerical or categorical data. Higher level statistical units, described by high level data types (called symbolic data types) can be derived from the huge amount of data stored in these databases. These high level data types include interval data, distributions, functional data, etc... While the symbolic data processing research field provides different technics for the visualization and analysis of these high level data types (see Diday (2008) for a detailed state of the art), there is no publication dealing with the persistency of symbolic data. We mean by "persistency" the storage and secure retrieval of data via a DBMS (database management system). The relational database model, originally designed to deal with numerical and categorical data, can hardly cope with symbolic data types. A more appropriate database model for symbolic data management seems to be the object-relational model (see Melton (2003) for a detailed discourse about this model). An elaborated symbolic data management approach based on the object-relational model is proposed in this paper in order to take advantage of effective and efficient data services (such as search facilities, changes tracking, etc.) provided by database management systems. Using our approach, symbolic data analysis and visualization algorithms will benefit interesting functionalities like the drilldown process (Noirhomme-Fraiture et al. (2008)) which constructs a new symbolic description, using a database search, through interaction with a visual representation. Our approach will also help in applications like time series analysis.

Keywords: symbolic data management, database

References

DIDAY, E. (2008): The state of the art in symbolic data analysis: overview and future. In: E. Diday and M. Noirhomme-Fraiture (Eds.): *Symbolic Data Analysis and the SODAS Software*. Wiley, New York, 3–41.

MELTON, J. (2003): *Advanced SQL:1999, Understanding Object-Relational and Other Advanced Features*. Elsevier Science, San Francisco.

NOIRHOMME-FRAITURE, M., BRITO, P., de BAENST-VANDENBROUCK, A. and NAHIMANA, A. (2008): Editing symbolic data. In: E. Diday and M. Noirhomme-Fraiture (Eds.): *Symbolic Data Analysis and the SODAS Software*. Wiley, New York, 81–92.

Non-Hierarchical Clustering for Distribution-Valued Data

Yoshikazu Terada[1] and Hiroshi Yadohisa[2]

[1] Graduate School of Culture and Information Science, Doshisha University
Kyoto 610-0394, Japan, *dij0030@mail4.doshisha.ac.jp*
[2] Department of Culture and Information Science, Doshisha University
Kyoto 610-0394, Japan, *hyadohis@mail.doshisha.ac.jp*

Abstract. A lot of clustering algorithms and methods for symbolic data has been proposed (e.g. Verde, 2004). However, the clustering methods for distribution-valued data, that can consider relationships between variables in each object, have been not really proposed. In this paper, we present a non-hierarchical clustering method for the distribution-valued data involving the histogram-valued data as an empirical frequency distribution function and joint distributions.

Several clustering methods for histogram-data have been proposed. For example, Irpino et al. (2006) presents the dynamic clustering method for histogram-data using a new distance based on Wasserstein metric. Verde and Irpino (2008) presents the Mahalanobis-Wasserstein distance for comparing histograms and apply it to the dynamic clustering. These methods assume a case that only the marginal distributions of the multivariate distribution for each object are obtained and then it is described by independent distributions.

In our method, we can consider relationships (e.g. dependency) among variables in each object by using joint distributions. We define the "centroid distribution" of probability distributions. By using the centroid distributions, we propose a new non-hierarchical clustering method for symbolic objects described by probability distributions.

Keywords: symbolic data analysis, k-means clustering, large and complex data

References

IRPINO, A., VERDE, R. and LECHEVALLIER, Y. (2006): Dynamic clustering of histograms using Wasserstein metric. In: Rizzi, A. and Vichi, M. (Eds.): *COMPSTAT 2006 Proceedings in Computational Statistics*. Physica-Verlag, Berlin, 869–876.

VERDE, R. (2004): Clustering Methods in Symbolic Data Analysis. In: H. H. Bock, W. Gaul and M. Vichi (Eds.): *Classification, Clustering, and Data Mining Applications*. Springer, Berlin, 299–317.

VERDE, R. and IRPINO, A. (2008): Comparing Histogram Data Using a Mahalanobis-Wasserstein Distance. In: Brito, P. (Eds.): *COMPSTAT 2008 Proceedings in Computational Statistics*. Physica-Verlag, Berlin, 77–89.

Symbolic Data Analysis of Complex Data: Application to nuclear power plant

Filipe Afonso[1], Edwin Diday[2], Norbert Badez[3], and Yves Genest[3]

[1] SYROKKO
5 rue de Copenhague BP13918, 95731 Roissy CDG, France, *afonso@syrokko.com*
[2] CEREMADE, Universite Paris Dauphine
Pce du M. Lattre de Tassigny 75775 Paris, France *diday@ceremade.dauphine.fr*
[3] EDF DTG-CEAN
12 rue saint Sidoine 69003 Lyon, France *norbert.badez,yves.genest@edf.fr*

Abstract. Complex data are here composed by several data tables describing different kinds of observations. The fusion of these data in order to get new knowledge requires to use symbolic data which are an extension of standard numerical or categorical data in order to loose less information than means by using intervals, distributions, sets of categories and the like. Symbolic Data Analysis (SDA) have been studied in recent books as Billard and Diday (2006), Diday and Noirhomme (2008), and enables to build and analyze such data. SDA of complex data is illustrated by the study of the degradation problems occurring on nuclear power plant cooling towers. Different kinds of measures have been collected by the French energy company EDF since the construction of each cooling tower. Several data tables describe cracks, corrosions, subsidence taking care of the shape. The fusion of these heterogeneous measures results in a symbolic data table containing in each cell histograms and intervals. SDA has shown to be suitable in order to study data on buildings degradation, performing the combination of the measures, discovering the correlations between them, highlighting and analyzing different problems of degradation, ordering and classifying the deteriorations of the towers.

Keywords: Symbolic data analysis, Complex data, Data visualization, Industrial application

References

BILLARD, L. and DIDAY, E. (2006): *Symbolic Data Analysis: conceptual statistics and data Mining. Wiley, 330 p., ISBN 0-470-09016-2*

BRITO, P. (2002): *Hierarchical and Pyramidal Clustering for Symbolic Data, Journal of the Japanese Society of Computational Statistics, Vol. 15, 2, 231-244.*

BRITO, P., BERTRAND, P., CUCUMEL, G. and DE CARVALHO, F. (2007): *Selected contributions in Data Analysis and Classification. Springer, 634 p.*

DIDAY, E. and NOIRHOMME, M. (eds) (2008): *Symbolic Data Analysis and the SODAS software. Wiley, 457 p., ISBN 9780-470-01883-5.*

KAVANAGH, B.F. (2009): *Surveying with Construction Applications. 7th Edition,Prentice Hall, 704p., ISBN 978-0135000519*

NAG, P. K. (2007): *Power Plant Engineering. 3rd Edition, McGraw-Hill, 992p., ISBN 978-0070648159*

Interactive graphics interfacing statistical modelling and data exploration

Adalbert Wilhelm[1]

School of Humanities and Social Sciences, Jacobs University Bremen
Campus Ring 1, 28759 Bremen, Germany, *a.wilhelm@jacobs-university.de*

Abstract. Graphics have been used in applied statistics for many decades, mostly to visualize results and to present conclusions. In the last twenty years interaction with graphical displays could be put into practice and made readily available for almost everyone. As a consequence plots changed their character from formerly being a final product to now being a temporary tool that can be modified and adapted according to the situation by simple mouse clicks or keyboard commands. Information overload that would prevent perception can be hidden at the first stage and made available on demand by responding to interactive user queries. In the applied fields, on the other hand, statistics is still done using the traditional mantra "only look at your data after you have analysed it". Hence, little work has been published on the interface between data exploration and statistical modelling, Doing statistical modelling without a proper graphical representation of data and model is risky and problematic. Exploring the data graphically without the attempt to model them properly, usually falls short and leaves the analyst with isolated insights and anecdotes. The systematic approach of modelling combined with the flexible use of exploratory graphics combines the strengths of both fields and constitutes a powerful research tool. This paper will illustrate this by providing an eclectic tour through the modelling process and illustrating the potential applications of exploratory graphics in the various steps. We will focus on three main stages of modelling: visualization prior to the modelling to check data quality and model adequacy, during the modelling process to check for model assumptions and model quality and after the modelling process to enhance interpretation of the modelling parameters as well as comparing between competing models.

Keywords: data and model exploration, interactive statistical graphics, linear models, model quality

References

BOWERS, J. and DRAKE, K.W. (2005): EDA for HLM: Visualization when Probabilistic Inference Fails. *Political Analysis, 13, 301-326.*

GELMAN, A. (2004): Exploratory Data Analysis for Complex Models. *Journal of Computational and Graphical Statistics 13, 755-787.*

GROENEN, P. and KONING, A.J. (2004): A new model for visualizing interactions in analysis of variance. *Econometric Institute Report, No EI 2004-06, EUR.*

TUKEY, J.W. (1969): Analyzing data: Sanctification or detective work? *American Psychologist, 24, 83-91.*

Contextual Factors of the External Effectiveness of the University Education: a Multilevel Approach

Matilde Bini[1], Leonardo Grilli[2], and Carla Rampichini[2]

[1] European University of Rome, Via Aldobrandeschi 190, 00163 Rome, Italy, *mbini@unier.it*

[2] Department of Statistics, University of Florence, Viale Morgagni, 59, 50134 Florence, Italy *grilli@ds.unifi.it, carla@ds.unifi.it*

Abstract. The increasing unemployment rates of young graduates recently observed in many European countries revives interest in studying the transition from university to work. An important aspect of this phenomenon is verifying if the degree of education acquired from university is adequate to the needs of the labour market (Biggeri et al. (2001)). This study deals with having a job about one year after graduation and it analyses the external effectiveness considering: 1) the study of the influence of the individuals' factors that affect their probabilities to get job; 2) the effects of the differences among economics and social territories (Barbieri et al. (2008)) in the probability to get a job and in the choices of people in searching for a job; 3) the evaluation of the differences among course programs of universities with respect to the probability to get a job. The use of context characteristics improve results in the analysis of ranking of institutions but there is a need to improve this measure including characteristics of institutions. In our study, the data have a hierarchical structure with two levels, represented by graduates (level-one units) and by groups of course programs combined with universities (group/university: level-two units). The observed response y_{ij} is a binary indicator which is equal to 1 if the graduate is employed at the date of the interview. The analysis is performed via a two-level logistic model (Goldstein (1995)) on a dataset coming from a survey on job opportunities of the Italian graduates in 2004, conducted by Istat in 2007 (Istat (2007)).

Keywords: contextual effects, effectiveness, multilevel data analysis

References

BIGGERI, L., BINI M. and GRILLI, L. (2001): The transition from university to work. A multilevel approach to the analysis of the time to obtain the first job. *Journal of the Royal Statistical Society, series A, 164, 293-305.*

GOLDSTEIN, H. (1995): *Multilevel Statistical Models.* Edward Arnold, London (1995).

ISTAT (2007): *Inserimento professionale dei laureati. Indagine 2007.* File Standard, Manuale utente e tracciato record. ISTAT, Roma.

BARBIERI, A.G., CRUCIANI, S. and FERRARA, A. (a cura di) (2008): *100 Statistiche per il Paese. Indicatori per conoscere e valutare.* ISTAT, Stampa CSR.

Multivariate Value at Risk Based on Extremality Measures

Henry Laniado[1], Rosa E. Lillo[2] and Juan Romo[3]

[1] Department of Statistics, Universidad Carlos III de Madrid, 28911, Leganés, Madrid, Spain *hlaniado@est-econ.uc3m.es*
[2] Department of Statistics, Universidad Carlos III de Madrid, 28903, Getafe, Madrid, Spain *lillo@est-econ.uc3m.es*
[3] Department of Statistics, Universidad Carlos III de Madrid, 28903, Getafe, Madrid, Spain *romo@est-econ.uc3m.es*

Abstract. We propose a new multivariate order based on a concept that we will call "extremality". Given a unit vector, the extremality allows to measure the "farness" of a point in $\Re^n$ with respect to a data cloud or to a distribution in the vector direction $\boldsymbol{u}$. We establish the most relevant properties of this measure and provide the theoretical basis for its nonparametric estimation. We propose a multivariate Value at Risk (VaR) with level sets constructed through extremality. Specifically, if $\boldsymbol{u} = \frac{1}{\sqrt{2}}[\pm 1, \pm 1]'$, our VaR coincides with the VaR defined in Tibiletti (2001). Besides, if $\boldsymbol{u} = \frac{-1}{\sqrt{n}}[1, \ldots, 1]'$ or if $\boldsymbol{u} = \frac{1}{\sqrt{n}}[1, \ldots, 1]'$, the oriented VaR proposed in this work, coincides respectively with the multivariate lower orthant VaR or the multivariate upper orthant VaR that were discussed in Embrechts and Pucceti (2006). However, fixing other directions, more conservative VaR can be obtained.

Keywords: extremality, value at risk, multivariate order.

References

BARNETT, V. (1976): The ordering of multivariate data. *Journal of the Royal Statistical Society, Ser. A, 139, 318-354.*

EMBRECHTS, P., PUCCETI, G. (2006): Bounds for functions of multivariate risks. *Journal of Multivariate Analysis 97, 526-547.*

HALLIN, M., PAINDAVEINE, D. and ŠIMAN, M. (2010): Multivariate quantiles and multiple-output regression quantiles: from L_1 optimization to halfspace depth. *Annals of Statistics 38, 635-669.*

JORION, P. (1997): *Value at risk: the new benchmark for controlloing markets risk.* The McGraw-Hill Companies, Inc., New York.

TIBILETTI, L. (2001): Incremental value at risk: traps and misinterpretations. In *trends in Mathematics,* (M. Kohlmann ed.) Birkhauser Verlag, Basel (Switzerland), 355–364.

ZANI, S., RIANI, M., and CORBELLINI, A. (1999): New methods for ordering multivariate data: an application to the performance of investment funds. *Applied Stochastic Models and Data Analysis 15, 485-493.*

Modifications of BIC: Asymptotic optimality properties under sparsity and applications in genome wide association studies

Florian Frommlet[1], Piotr Twaróg[2], and Małgorzata Bogdan[2]

[1] Department of Statistics, University Vienna
Brünnerstraße 72, 1210 Vienna
Florian.Frommlet@univie.ac.at
[2] Department of Statistics, TU Wroclaw
Wybrzeze Wyspianskiego 27, 50370 Wroclaw
Piotr.Twarog@pwr.wroc.pl, Malgorzata.Bogdan@pwr.wroc.pl

Abstract. In many statistical applications today one is confronted with regression models where the number of independent factors is larger than the number of observations on which statistical analysis is based. Bogdan et al. (2004) have pointed out in the context of QTL mapping that in this situation classical model selection criteria like AIC or even BIC tend to select too large models. Thus a modification of BIC (mBIC) taking into account the potential number of models of a certain size was introduced. The original version of mBIC corresponds to controlling family wise error in multiple testing, whereas modifications in the spirit of Abramovich et al. (2006) correspond to controlling FDR.

Recently Bogdan et al. (2010) have analysed asymptotic optimality under sparsity for multiple testing rules using the framework of Bayesian Decision Theory. Here we will present similar results concerning asymptotic optimality properties of different versions of mBIC for the choice of multiple regression models under sparsity. These procedures are then applied to analyse data from genome wide association studies. Comparison with multiple testing procedures shows that power can be increased, although intelligent search strategies for models need to be developed.

Keywords: model selection, asymptotic optimality, sparsity, genome wide association studies

References

ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. L. and JOHNSTONE, I. M. (2006): Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist. 34, 584–653.*

BOGDAN, M., GHOSH, J. K. and DOERGE, R. W. (2004): Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci. *Genetics 167, 989-999.*

BOGDAN, M., CHAKRABATI, A., FROMMLET, F. and GHOSH, J. K. (2010): The Bayes oracle and asymptotic optimality of multiple testing procedures under sparsity. *S*ubmitted.

A New Post-processing Method to Deal with the Rotational Indeterminacy Problem in MCMC Estimation

Kensuke Okada[1] and Shin-ichi Mayekawa[2]

[1] Senshu University
2-1-1, Higashimita, Tama-ku, Kawasaki-shi, Kanagawa, Japan,
ken@psy.senshu-u.ac.jp
[2] Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, Japan, *mayekawa@hum.titech.ac.jp*

Abstract. Some spatial models of psychometrics such as MDS or Factor Analysis are known to be "under-identified," which means that infinite number of solutions give the same likelihood in these models. This is attributed to the fact that orthogonal transformations of configuration matrix do not alter the likelihood. Therefore, when applying Markov chain Monte Carlo (MCMC) estimation to these models, the indeterminacy problem must be addressed.

In former studies, there have been three major approaches to deal with this problem: (1) introduction of additional parameter constraints, (2) use of strong informative priors on the configuration matrix, and (3) post-processing (Park et al. (2008)). Post-processing approach is more flexible than the others because it does not require a priori information on the configuration matrix.

In this study, we propose a new post-processing method that iteratively maximizes the congruence between each of the MCMC output and the average configuration matrix. That is, each of the configuration matrix is post-processed by the generalized Procrustes rotation (Schönemann and Carroll,1970) toward the average configuration matrix.

A numerical experiment showed that the proposed method performed better than previously proposed methods, both in terms of the means and of the variances of the mean squared errors (MSE). The applicability of the proposed method to other multivariate analysis is discussed.

Keywords: Markov chain Monte Carlo, post-processing, generalized Procrustes rotation

References

PARK, J., DESARBO, W. S. and LIECHTY, J. (2008): A hierarchical Bayesian multidimensional scaling methodology for accommodating both structural and preference heterogeneity. *Psychometrika 73 (3), 451–472.*

SCHÖNEMANN, P. H. and CARROLL, R. (1970): Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika 35 (2), 245–255.*

A Constrained Condition-Number LS Algorithm with Its Applications to Reverse Component Analysis and Generalized Oblique Procrustes Rotation

Kohei Adachi[1]

[1] Graduate School of Human Sciences, Osaka University, Suita, Osaka 565-0871, Japan. *k-adachi@lt.ritsumei.ac.jp*

Abstract. I propose an algorithm for minimizing $f = f(\mathbf{H}) = \|\mathbf{Y} - \mathbf{H}\|^2$ over an $n \times p$ matrix $\mathbf{H}$ subject to its condition number $CN(\mathbf{H})$ not exceeding prescribed positive constant u with $\mathbf{Y}$ ($n \times p$) fixed and $p \leq n$. This rank preserving problem has not been found in the existing literature of matrix computation (e.g., Golub & van Loan, 1996). In the proposed algorithm, $\mathbf{H}$ is reparameterized using its singular value decomposition (SVD) as $\mathbf{H} = a\mathbf{K}\mathbf{\Lambda}\mathbf{L}'$, with $a \neq 0$, $\mathbf{K}'\mathbf{K} = \mathbf{L}'\mathbf{L}$ the $p \times p$ identity matrix, and $\mathbf{\Lambda}$ the diagonal matrix whose lth diagonal element is $\lambda_l > 0$. Then, the above problem is reformulated as minimizing $f = \|\mathbf{Y} - a\mathbf{K}\mathbf{\Lambda}\mathbf{L}'\|^2 = \|\mathbf{Y}\|^2 - 2a\mathrm{tr}\mathbf{K}'\mathbf{Y}\mathbf{L}\mathbf{\Lambda} + a^2\mathrm{tr}\mathbf{\Lambda}^2$ over a, $\mathbf{\Lambda}$, $\mathbf{K}$, and $\mathbf{L}$ subject to $1 \leq \lambda_l \leq u$.

We can thus find the solution by the iteration of minimizing f over each of a, $\mathbf{\Lambda}$, $\mathbf{K}$, and $\mathbf{L}$ with the other parameters kept fixed. It is attained with the following four update formulas: [1] $a = \mathrm{tr}\mathbf{K}'\mathbf{Y}\mathbf{L}\mathbf{\Lambda}/\mathrm{tr}\mathbf{\Lambda}^2$; [2] $\lambda_l = d_{1l}/(ad_{2l})$ if $1 \leq d_{1l}/(ad_{2l}) \leq u$, and otherwise $\lambda_l = \arg\min_{\lambda_l = 1, u} \theta_l(\lambda_l)$ for $l = 1, \ldots, p$. [3] $\mathbf{K} = \mathbf{P}\mathbf{Q}'$; [4] $\mathbf{L} = \mathbf{U}\mathbf{V}'$. Here, $\theta_l(\lambda_l) = a^2 d_l \lambda_l^2 - 2a\lambda_l$ with d_l the lth diagonal element of $\mathbf{K}'\mathbf{Y}\mathbf{L}$, and $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{U}$, and $\mathbf{V}$ are obtained the SVDs $a\mathbf{Y}\mathbf{L}\mathbf{\Lambda} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}'$ and $a\mathbf{Y}'\mathbf{K}\mathbf{\Lambda} = \mathbf{U}\mathbf{\Omega}\mathbf{V}'$ with $\mathbf{\Delta}$ and $\mathbf{\Omega}$ diagonal matrices.

For assessing the algorithm, I consider *reverse* component analysis (RCA) viewed as a reverse PCA problem in which $f = \|\mathbf{H}_{\mathrm{LOW}}' - \mathbf{H}\|^2$ is minimized over $\mathbf{H}$ subject to $CN(\mathbf{H}) \leq u$, where $\mathbf{H}_{\mathrm{LOW}}$ is the lower-rank approximation of $\mathbf{H}_{\mathrm{TRUE}}$. A numerical experiment showed the good recovery of $\mathbf{H}_{\mathrm{TRUE}}$ by $\mathbf{H}$.

The proposed algorithm can be applied to the generalized oblique Procrustes problem. This is formulated as minimizing $h(\mathbf{T}, \mathbf{H}) = \Sigma_k \|\mathbf{A}_k\mathbf{T}_k'^{-1} - \mathbf{H}\|^2 = g(\mathbf{T}) + Kf(\mathbf{H})$ over $\mathbf{H}$ and $\mathbf{T} = [\mathbf{T}_1'^{-1}, \ldots, \mathbf{T}_K'^{-1}]'$ subject to the diagonal elements of $\mathbf{T}_k'\mathbf{T}_k$ being ones, where $g(\mathbf{T}) = \Sigma_k \|\mathbf{A}_k\mathbf{T}_k'^{-1} - K^{-1}\mathbf{A}\mathbf{T}\|^2$ and $f(\mathbf{H}) = \|K^{-1}\mathbf{A}\mathbf{T} - \mathbf{H}\|^2$ with $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_K]$. But, the solution for the problem is known to degenerate with $\mathbf{A}_k\mathbf{T}_k'^{-1}$ and $\mathbf{H}$ of rank one. For avoiding the degeneration, we can iterate the minimization of $g(\mathbf{T})$ over $\mathbf{T}$ with the existing procedure and that of $f(\mathbf{H})$ over $\mathbf{H}$ subject to $CN(\mathbf{H}) \leq u$ with the propose algorithm.

Keywords: Constrained least squares, Condition number, Singular value decomposition, Reverse component analysis, Oblique Procrustes rotation

Reference

GOLUB, G. H. and VAN LOAN, C. F. (1996): *Matrix Computations* (*Third Edition*). Johns Hopkins University Press, Baltimore, MD.

Matrix Visualization for MANOVA Modeling

Yin-Jing Tien[1] Han-Ming Wu[2] and Chun-houh Chen[3]

[1] Institute of Statistics, National Central University
Taoyuan 32001, Taiwan, *gary@stat.sinica.edu.tw*
[2] Department of Mathematics, Tamkang University
Tamsui 25137, Taiwan, *hmwu@mail.tku.edu.tw*
[3] Institute of Statistical Science, Academia Sinica
Taipei 11529, Taiwan *cchen@stat.sinica.edu.tw*

Abstract. Generalized association plots (GAP) introduced by Chen (2002) and Wu et al. (2010) is an environment for general purposes matrix visualization (MV, Chen et al. (2004)). In this study a comprehensive matrix visualization procedure for analyzing multivariate analysis of variance (MANOVA) models is proposed as an extension of GAP.

Existing matrix visualization methods are not suitable for clustering and visualizing data and information structure with a MANOVA setting because they regard individual samples as the base analysis unit without taking into considerations the model effects. In order to have a comprehensive visualization on data and information structure for MANOVA modeling, it is necessary to simultaneously explore related information structures at the model and the residual levels. In our proposed method, we visualize not only the decomposition of covariance matrix into model and residual components but also decomposition of the data matrix. We further convert statistical testing results (p-values from MANOVA and ANOVA for individual variables) into MV format for obtaining a more powerful and complete visualization for understanding MANOVA modeling at both the descriptive and inference aspects. With a covariate adjusted MV adopted before the MANOVA MV procedure this method can be extended to visualization of MANCOVA modeling.

Keywords: generalized association plots(GAP), matrix visualization, MANOVA, p-value

References

Chen, C. H. (2002): Generalized Association Plots for Information Visualization: The applications of the convergence of iteratively formed correlation matrices. *Statistica Sinica 12, 1-23.*

Chen, C. H., Hwu, H. G., Jang, W. J., Kao, C. H., Tien, Y. J., Tzeng, S., and Wu, H. M. (2004): Matrix Visualization and Information Mining. In: *Proceedings in Computational Statistics 2004 (Compstat 2004).* Physika Verlag, Heidelberg, 85–100.

Wu, H. M., Tien, Y. J., and Chen, C. H. (2010): GAP: A graphical environment for matrix visualization and cluster analysis.. *Computational Statistics and Data Analysis 54 (3), 767–778.*

Orthogonal grey simultaneous component analysis to distinguish common and distinctive information in coupled data

Martijn Schouteden[1], Katrijn Van Deun[1], and Iven Van Mechelen[1]

Quantitative Psychology and Individual Differences, Katholieke Universiteit Leuven, Tiensestraat 102 - bus 3713, 3000 Leuven, Belgium
martijn.schouteden@psy.kuleuven.be

Abstract. Often, data are collected consisting of different blocks that all contain information about the same entities (e.g., items, persons, situations). A major challenge is to reveal the mechanisms underlying each of those coupled data blocks and to disentangle in this regard common mechanisms underlying all data blocks and distinctive mechanisms underlying a single data block or a few blocks only. An interesting class of methods for such an approach is the family of simultaneous component methods. These methods yield components that maximally account for the variance in all data blocks. Unfortunately, however, they usually contain a mix of common and distinctive information. Recently, DISCO-SCA has been proposed as a method to tackle this problem by orthogonally rotating a simultaneous component solution towards a target matrix with a clear common and distinctive structure (Schouteden et al. (2010)). Yet, in quite a few cases, to improve the interpretability of the components, it may be needed to impose the target structure on the component solution to a stronger degree, while maintaining the orthogonality restriction. For this purpose, we developed a novel method, starting from the framework of grey component analysis as proposed by Westerhuis et al. (2007). After describing this method, we will illustrate it with data stemming from systems biology.

Keywords: Multiset data, Multiblock data, Component analysis, Data fusion

References

SCHOUTEDEN, M., VAN DEUN, K., VAN MECHELEN, I. and PATTYN, S. (2010): *SCA and Rotation to distinguish common and distinctive information in coupled data*. Manuscript submitted for publication.

WESTERHUIS, J. A., DERKS, E. P. P. A., HOEFSLOOT, H. C. J. and SMILDE, A. K. (2007): Grey component analysis, *J. Chemometrics 21, 474-485.*

Clusterwise SCA-P

Kim De Roover[1], Eva Ceulemans[1], Marieke Timmerman[2], and Patrick Onghena[1]

[1] Centre for methodology of educational research, K.U.Leuven
Andreas Vesaliusstraat 2, Leuven, Belgium,
[2] Heymans Institute of Psychology, University of Groningen
Grote Kruisstraat 2/1, The Netherlands

Abstract. Numerous research questions in educational sciences and psychology concern the structure of a set of variables. To study the structure of a set of variables, one typically relies on scores from a group of persons on those variables. One may, however, wonder whether the same structure would have been retrieved if another group of persons had been studied. To trace such structural differences, one will have to gather data from different groups. Formally, the resulting data then constitute multivariate multiblock data, with persons being nested in groups. Obviously, the crucial question is how such data have to be analyzed to find out whether and in what way the structure of the variables differs across the groups of persons. A number of principal component analysis techniques exist to study such structural differences, for instance, simultaneous component analysis (SCA). However, these techniques suffer from some important limitations. Therefore, in this presentation, we propose a novel generic modeling strategy, called Clusterwise SCA-P, which solves these limitations and which encompasses several existing techniques as special cases. Clusterwise SCA-P generalizes the earlier proposed Clusterwise SCA-ECP model (De Roover et al., 2010). Like Clusterwise SCA-ECP, Clusterwise SCA-P partitions the groups into a number of clusters and specifies a simultaneous component model for each cluster. Unlike Clusterwise SCA-ECP, Clusterwise SCA-P allows for between group differences in the variances and the correlations of the cluster specific components. As such, the Clusterwise SCA-P model offers a more detailed insight into group differences. An algorithm for fitting Clusterwise SCA-P solutions is presented and its performance is evaluated by means of a simulation study. The value of the model for empirical research is illustrated with data from psychiatric diagnosis research.

Keywords: multiblock data, multivariate data, principal component analysis, simultaneous component analysis, clustering

References

DE ROOVER, K., CEULEMANS, E., TIMMERMAN, M.E., VANSTEELANDT, K., STOUTEN, J. and ONGHENA, P. (2010): Clusterwise SCA-ECP for the analysis of structural differences in multivariate multiblock data. *Manuscript submitted for publication.*

A numerical convex hull based procedure for selecting among multilevel component solutions

Eva Ceulemans[1], Marieke E. Timmerman[2], and Henk A.L. Kiers[3]

[1] Centre for Methodology of Educational Research, K.U.Leuven
Andreas Vesaliusstraat 2, Leuven, Belgium, *eva.ceulemans@ped.kuleuven.be*
[2] Heymans Institute of Psychology, University of Groningen
Grote Kruisstraat 2/1, The Netherlands, *m.e.timmerman@rug.nl*
[3] Heymans Institute of Psychology, University of Groningen
Grote Kruisstraat 2/1, The Netherlands, *h.a.l.kiers@rug.nl*

Abstract. Recently, Timmerman (2006) proposed a class of multilevel component models for the analysis of two-level multivariate data. These models consist of a separate component model for each level in the data. Specifically, the between differences are captured by a between component model and the within differences by a within component model. Within the class of multilevel component models a number of variants can be distinguished. These variants differ with respect to the within component model only, in that different sets of restrictions are imposed on the within component loadings and on the variances and correlations of the within component scores. The following question then may be raised: given a specific two-level data set, which of the multilevel component model variants should be selected, and with how many between and within components? We address this question by proposing a model selection procedure that builds on the numerical convex hull based heuristic of Ceulemans and Kiers (2006, 2009). The results of an extensive simulation study show that the proposed hull heuristic succeeds very well in assessing the number of between and within components. Tracing the underlying multilevel component model variant is more difficult: Whereas differences in within loading matrices and differences in variances are very easy to detect, the precise correlational structure of the within components is much harder to capture.

Keywords: model selection, multilevel component analysis

References

CEULEMANS, E. and KIERS, H.A.L. (2006): Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. *British Journal of Mathematical and Statistical Psychology 59, 133-150.*

CEULEMANS, E. and KIERS, H.A.L. (2009): Discriminating between strong and weak structures in three-mode principal component analysis. *British Journal of Mathematical and Statistical Psychology 62, 601-620.*

TIMMERMAN, M.E. (2006): Multilevel component analysis. *British Journal of Mathematical and Statistical Psychology 59, 301-320.*

Treatment Interaction Trees (TINT): A tool to identify disordinal treatment-subgroup interactions

Elise Dusseldorp[1,2] and Iven Van Mechelen[2]

[1] TNO Quality of Life, Department of Statistics
Wassenaarseweg 56, Leiden, The Netherlands, *elise.dusseldorp@tno.nl*
[2] Department of Psychology, Katholieke Universiteit Leuven
Tiensestraat 102, Leuven, Belgium, *Iven.VanMechelen@psy.kuleuven.be*

Abstract. When two competitive treatments, A and B, are available, some subgroup of patients may display a better outcome with treatment A than with B, whereas for another subgroup the reverse may be true. If this is the case, a disordinal (i.e., a qualitative) treatment-subgroup interaction is present. Such interactions imply that some subgroups of patients should be treated differently, and are therefore most relevant for clinical practice. In case of data from randomized clinical trials with many patient characteristics that could interact with treatment in a complex way, a suitable statistical approach to detect disordinal treatment-subgroup interactions is not yet available. In this presentation, we introduce a new method for this purpose, called Treatment INteraction Trees (TINT). TINT results in a binary tree that subdivides the patients into terminal nodes on the basis of patient characteristics; these nodes are further assigned to one of three classes: a first one for which A is better than B, a second one for which B is better than A, and an optional third one for which type of treatment makes no difference. The method will be compared to STIMA (Dusseldorp, Conversano, and Van Os (in press)). Results of a pilot test of TINT on artificial data will be shown, as well as results of an application to real data from the Breast Cancer Recovery Project (Scheier et al. (2007)).

Keywords: Treatment-subgroup interaction, Recursive partitioning, Disordinal interaction, Subgroup analysis, Moderator

References

DUSSELDORP, E., CONVERSANO, C., and VAN OS, B.J. (in press): Combining an additive and tree-based regression model simultaneously: STIMA. *Journal of Computational and Graphical Statistics*.

SCHEIER, M.F., HELGESON, V.S., SCHULZ, R., COLVIN, S., BERGA, S.L., KNAPP, J., GERSZTEN, K. (2007): Moderators of interventions designed to enhance physical and psychological functioning among younger women with early-stage breast cancer. *Journal of Clinical Oncology 25, 5710-5714*.

Risk reduction using Wavelets-PCR models: Application to market data

Nabiha Haouas[1], Saloua benammou[2], and Zied Kacem[3]

[1] Computational Mathematics Laboratory, *n.haouas@yahoo.fr*
[2] *saloua.benammou@yahoo.fr*
[3] *Ziedkacem2004@yahoo.fr*

Abstract. In this paper, we set out a hybrid data analysis method based on the combination of wavelet techniques and Principal Components Regression (PCR). Our purpose is to study the dynamics of the stock returns within the French Stock Market. Wavelet-based thresholding techniques are applied to the stock price series in order to obtain a set of explanatory variables that are practically noise-free. The PCR is then carried out on the new set of regressors. The empirical results show that the suggested method allows extraction and interpretation of the factors that influence the stock price changes. Moreover, the Wavelet-PCR improves the explanatory power of the regression model as well as its forecasting quality.

Keywords: Wavelets, Thresholding, PCR, PCA, French stock exchange, Returns.

References

DAUBECHIES, I. (1992). *Ten lectures on wavelets*. SIAM. Philadelphia. USA.

DONOHO, D. (1992). *Wavelet shrinkage and W.V.D: a 10 minute tour*. Progress in Wavelet Analysis and Applications. Edition Frontieres. Dreuk, 109-128.

MALLAT, S. G. (1989). *A theory for multiresolution signal decomposition: the wavelet representation*. IEEE Transactions on Pattern Analysis and Machine Intelligence. (11), 74-693.

SAPORTA, G. (2006). *Probabilits, analyse des donnes et Statistique*. 2ed Edition, Technip, Paris.

TSAY, R. S. (2005). *Analysis of Financial Time Series*. 2ed Edition, Wiley-Interscience.

Cepstral-based Fuzzy Clustering of Time Series

Elizabeth Ann Maharaj[1] and Pierpaolo D'Urso[2]

[1] Department of Econometrics and Business Statistics
Monash University, Melbourne, Australia *ann.maharaj@buseco.monash.edu.au*
[2] Dipartimento di Teoria Economica e Metodi Quantitativi per le Scelte Politiche
Sapienza Universita' di Roma, Rome, Italy *pierpaolo.durso@uniroma1.it*

Abstract. We introduce a fuzzy clustering approach for time series based on cepstral coefficients. The cepstrum of a time series is the spectrum of the logarithm of the spectrum. Kalpakis et al. (2001) used cepstral coefficients to cluster time series that were first fitted with autoregressive models. Savvides et al. (2008) used cepstral coefficients which were estimated from a semiparmetric model to cluster biological time series. Boets et al. (2005) proposed a clustering process based on a cepstral distance between stochastic models. In all cases, hierarchical methods were used, resulting in crisp clusters. In employing fuzzy clustering, we are ensuring that useful information about cluster membership of time series of a changing nature is not lost, which would otherwise be the case if only crisp clustering was used. In our simulation studies we show that the fuzzy clustering based on the cepstral coefficients generally performs better than that based on the normalized periodogram, and on the logarithm of the normalized periodogram. We also compare the performances of the cepstral-based fuzzy clustering with the autocorrelation-based fuzzy clustering proposed by D'Urso and Maharaj (2009). We apply this approach to classify time series of annual changes in CO2 emissions of a number of countries.

Keywords: Cepstral coefficients, Normalized periodogram, Log normalized periodogram, Fuzzy clustering

References

BOETS, J., DE COCK, K., ESPINOZA, M. and DE MOOR, B.(2005): Clustering time series, subspace identification and cepstral distances. *Communications Information Systems 5 (1)69-96.*

D'URSO, P. and MAHARAJ, E. A. (2009): Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems 160 , 3565-3589.*

KALPAKIS, K., GADA, D. AND PUTTAGUNTA, V. (2000): Distance measures for effective clustering of ARIMA time series: *Proceedings of the IEEE International Conference on Data Mining.* San Jose, 273–280.

SAVVIDES, A., PROMPONAS, V.J. and FOKIANOS, K. (2008): Clustering of biological time series by cepstral coefficients based distances. *Pattern Recognition 41, 2398-2412.*

On two SSA-based methods for imputation of missing time-series data

Marina Zhukova[1] and Nina Golyandina[2]

[1] Math. Department, St.Petersburg State University, Universitetsky av. 28, 198504, St.Peterburg, Petrodvorets, Russia *aniram.zhukova@gmail.com*
[2] Math. Department, St.Petersburg State University, Universitetsky av. 28, 198504, St.Peterburg, Petrodvorets, Russia *nina@gistatgroup.com*

Abstract. In this paper we consider the problem of filling in missing data in time series. Two methods based on Singular Spectrum Analysis (Golyandina et al. (2001)), shortly SSA, are studied. These methods are applied to the series of the form $f_n = s_n + \varepsilon_n$, $n = 0, \dots, N-1$, where ε_n is noise, s_n is a signal that satisfies the relation $s_{i+d} = \sum_{k=1}^{d} a_k s_{i+d-k}$, $0 \le i \le N-d-1$. The minimum possible value of d is called the rank of the signal.

The approach to missing data imputation suggested in Golyandina and Osipov (2007) uses estimation of the signal subspace, in particular, of the coefficients a_k, $n = 1, \dots, d$. Beckers and Rixen (2003) propose the ideas for the iteration method of filling in with imputation of missing data by zeroes at the first iteration of the algorithm. In this paper we compare the above methods of filling in. The problems of applicability, time consumption and accuracy of imputation are considered.

In both algorithms it is assumed that the rank of the signal is known in advance. Many methods have been proposed for estimating the rank of signals with no missing data. Presence of missing data complicates the problem. We study different methods of the rank estimation in presence of missing data. Most attention is paid to the approach described in Beckers and Rixen (2003). In that approach the rank is estimated by the use of a test set consisting of artificial missing data. In this paper we consider the problem of choosing the proper test set and then compare the performance of the constructed method with other rank-detection methods.

Keywords: time series, missing data, Singular Spectrum Analysis

References

BECKERS, J.M. and RIXEN, M. (2003): EOF calculations and data filling from incomplete oceanographic datasets. *Journal of Atmospheric and Oceanic Technology 20 (12), 1839–1856.*

GOLYANDINA, N., NEKRUTKIN, V., and ZHIGLJAVSKY, A., (2001): *Analysis of Time Series Structure: SSA and Related Techniques.* Chapman & Hall/CRC, London.

GOLYANDINA, N. and OSIPOV. E. (2007): The "Caterpillar"-SSA method for analysis of time series with missing values. *Journal of Statistical Planning and Inference 137 (8), 2642–2653.*

Data-driven window width adaption for robust online moving window regression

Matthias Borowski

Fakultät Statistik, Technische Universität Dortmund
44221 Dortmund, Germany *borowski@statistik.tu-dortmund.de*

Abstract. In many fields, like intensive care online-monitoring, industrial process control, or financial markets, observations are made at a high sampling rate, typically leading to time series which are corrupted by outliers and noise. In this context, estimating the underlying noise- and outlier-free signal *online*, i.e. for every new incoming observation, is a challenging task. Since these time series typically exhibit enduring trends, estimating the signal by robust linear regression in a moving time window is more appropriate than using location-based methods like a running median (Davies et al. (2004)). The central or alternatively right-end level of the regression line can be used as a signal estimation.

Having a fixed window width n which is suitable at all time points is impossible because of trend changes and level shifts or other aspects of non-stationarity. In periods without structural data changes, a large n is requested in order to get smooth signal estimation time series; when a structural change like a level shift occurs, n should be small so that the data change is traced well.

We propose techniques for a data-driven window width adaption which follow the ideas by Borowski et al. (2009) and Schettlinger et al. (2010): As long as the data structure remains unchanged, the window width gradually grows with each incoming observation; but when a structural change occurs, it is set to a predetermined minimum value.

A good window width adaption technique must detect data changes reliably but also be 'stable' in periods without changes. The introduced techniques are compared with respect to these requirements.

Keywords: online signal estimation, robust regression, window width adaption, monitoring time series

References

BOROWSKI, M., SCHETTLINGER, K. and GATHER, U. (2009): Multivariate real time signal processing by a robust adaptive regression filter. *Communications in Statistics – Simulation and Computation 38 (2), 426-440.*

DAVIES, P.L., FRIED, R. and GATHER, U. (2004): Robust signal extraction for on-line monitoring data. *Journal of Statistical Planning and Inference 122, 65-78.*

SCHETTLINGER, K., FRIED, R. and GATHER, U. (2010): Real time signal processing by adaptive repeated median filters. *International Journal of Adaptive Control and Signal Processing 24, 346-362.*

Robust forecasting of non-stationary time series

Koen Mahieu

ORSTAT and University Center of Statistics, K. U. Leuven
Naamsestraat 69, B-3000 Leuven, Belgium, *Koen.Mahieu@econ.kuleuven.be*
in cooperation with: Christophe Croux, Irène Gijbels and Roland Fried

Abstract. This paper (Croux et al. (2010)) presents a *flexible* and *robust* forecasting technique for *non-stationary* (typically heteroscedastic) time series. In real data sets, there is a simultaneous need for flexibility, robustness against outliers and ability of coping with heteroscedasticity in the time series. First, flexible modelling is typically suitable for time series for which the signal does not lie in a prespecified family of parametric functions. Second, robustness should certainly be involved to prevent outliers in the data from having an adverse effect on the predictions. Finally, the ability to handle heteroscedastic time series is of major importance, since the restriction of homoscedasticity is often too stringent in real data examples.

Most forecasting techniques in the literature lack at least one of the aforementioned properties. Local polynomial regression, for instance, is a flexible technique that is not robust; neither can it cope with heteroscedasticity. Furthermore, the local polynomial M-regression of Grillenzoni (2009) and the weighted repeated median of Fried et al. (2007) are robust and flexible, though they do not take into account a possible change of the variance over time.

A simulation study has shown that, in the presence outliers and heteroscedasticity, our proposed method outperforms benchmark methods such as Local Polynomial regression, Weighted Repeated Median forecasting and local M-estimation, while it still achieves comparable performance results in an uncontaminated setting.

Since the estimation procedure involves a local scale of the one-step-ahead forecast errors, one could use this scale estimate for the construction of local prediction intervals. A drawback of these scale estimates is that they suffer from a finite sample bias. This problem may be an interesting topic for future research.

Keywords: Flexibility, Forecasting, Local MM-regression, Non-stationarity, Robustness

References

CROUX, C., FRIED, R., GIJBELS, I. and MAHIEU K. (2010): Robust forecasting of non-stationary time series *manuscript*

FRIED, R., EINBECK, J. and GATHER, U. (2007): Weighted Repeated Median Smoothing and Filtering *Journal of the American Statistical Association 102 (480), 1300-1308.*

GRILLENZONI, C. (2009): Robust non-parametric smoothing of non-stationary time series *Journal of Statistical Computation and Simulation 79 (4), 379-393.*

Spline approximation of a random process with singularity

Konrad Abramowicz[1] and Oleg Seleznjev[2]

[1] Department of Mathematics and Mathematical Statistics, Umeå University SE-901 87 Umeå, Sweden, *konrad.abramowicz@math.umu.se*
[2] Department of Mathematics and Mathematical Statistics, Umeå University SE-901 87 Umeå, Sweden, *oleg.seleznjev@math.umu.se*

Abstract. The close relationship between the smoothness properties of a function and the best rate of its linear approximation is one of the basic ideas of conventional (deterministic) approximation theory. We study similar properties for random signals (processes). In particular, we consider Hermite spline approximation of a continuous (in quadratic mean, q.m.) random process $X(t)$, $t \in [0,1]$, based on n observations. The performance of the approximation is measured by mean errors (integrated or maximal q.m. errors). Let l-th q.m. derivative of X satisfy a Hölder condition with exponent $0 < \alpha < 1$, say, $X \in C^{l,\alpha}([0,1])$, and have a continuous q.m. m-derivative, $m > l$, for all points $t > 0$. It is known that the approximation rate $n^{l+\alpha}$ is optimal for linear approximation methods in a certain sense for the Hölder class $C^{l,\alpha}([0,1])$ (Buslaev and Seleznjev (1999), Seleznjev (2000)). But for such smooth process with singularity at one point (or a finite number of points) and a certain local stationarity property, we investigate the sequence of quasi regular designs (observation locations) and find the sequence of sampling designs with approximation rate at least n^m and asymptotically optimal properties as $n \to \infty$, e.g., for integrated q.m. norm. These results can be used in various problems in numerical analysis of random functions, for archiving telecommunication, multimedia, environmental data in databases, and in simulation studies.

Keywords: Approximation, Random process, Hermite spline

References

BUSLAEV, A.P. and SELEZNJEV,O. (1999): On certain extremal problems in theory of approximation of random processes. *East Journal on Approximations 5, 467-481.*

SELEZNJEV, O. (2000): Spline approximation of random processes and design problems. *Journal of Statistical Planning and Inference 84, 249-262..*

Monitoring time between events in an exponentially distributed process by using Optimal Pre-control

Vicent Giner-Bosch and Susana San Matías

Departamento de Estadística e Investigación Operativa Aplicadas y Calidad,
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain
vigibos@eio.upv.es, *ssanmat@eio.upv.es*

Abstract. Several quality control tools such as control charts have been designed or adapted for the task of supervising time between events (e.g., time between failures, in a reliability context), as in Xie et al. (2002). One of the most recent proposals consists of applying Optimal Pre-control (OPC) to monitoring exponentially distributed observations (San Matías and Giner-Bosch (2008)). OPC (San Matías and Giner-Bosch (2010)) is an extension of the classical quality technique known as Pre-control, in which some parameters associated to this technique are determined by means of optimization methods.

OPC provides a simple and effective way to detect changes in the event occurrence rate that does not require continuously monitoring the process, but only checking for events having occurred at or before the instants determined by the OPC plan. One limitation of this approach is the assumption of a kind of symmetry of chart limits (here, Pre-control time instants) in terms of probability. In our present contribution we further develop the theoretical OPC model for the exponential case and eliminate some assumptions about the way Pre-control time instants are distributed. This leads to more flexible models that are able to meet a wider range of user specifications. We illustrate our proposal by means of some numerical examples and discuss its usefulness in reliability and maintenance applications.

Keywords: Pre-control, Time between events, Reliability

References

XIE, M., GOH, T. and RANJAN, P. (2002): Some effective control chart procedures for reliability monitoring. *Reliability Engineering and System Safety 77 (2), 143-150.*

SAN MATÍAS, S. and GINER-BOSCH, V. (2009): Optimal Pre-Control as a tool to monitor the reliability of a manufacturing system. In: Martorell, S., Guedes Soares, C. and Barnett, J. (Eds.): *Safety, Reliability and Risk Analysis: Theory, Methods and Applications.* CRC Press, London, 2735–2741.

SAN MATÍAS, S. and GINER-BOSCH, V. (2010): Selection of best pre-control technique by optimization tools. *Technical Report DEIOAC-2010-06, Departamento de Estadstica e IO Aplicadas y Calidad, Universidad Politcnica de Valencia.*

Calibration of hitting probabilities via adaptive multilevel splitting

Ioannis Phinikettos[1] and Axel Gandy[2]

[1] Imperial College, London, United Kingdom, *ioannis.phinikettos@imperial.ac.uk*
[2] Imperial College, London, United Kingdom, *a.gandy@imperial.ac.uk*

Abstract. We are considering a cadlag Markov process on a finite time interval that takes values in the real numbers. We are interested in the event that the process hits a threshold. More precisely, we want to find the threshold that results in a given low hitting probability. The particular example we have in mind is cumulative sum (CUSUM) control charts in discrete or continuous time. For those we want to choose the threshold to yield a fixed low false alarm probability.

Low desired false alarm probabilities will result in hitting the threshold becoming a rare event, making direct Monte Carlo simulation unreliable. We present an adaptive multilevel splitting algorithm to overcome this problem. We will show that, under certain regularity conditions, the suggested method is consistent. We will present simulation results.

Keywords: multilevel splitting, rare event simulation, CUSUM charts

On Calculation of Blaker's Binomial Confidence Limits

Jan Klaschka

Institute of Computer Science, Academy of Sciences
Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic, *klaschka@cs.cas.cz*

Abstract. Exact confidence intervals for the parameters of discrete distributions are often constructed by inverting equal-tailed tests (the Clopper-Pearson interval for a binomial proportion being an example), but several less conservative alternatives have been proposed since the 1950's. Among these, the intervals by Blaker (2000) have attracted much attention due to desirable properties (e.g. monotonicity w.r.t. confidence level) and (at least in the binomial case) easy calculation.

The Blaker's $1-\alpha$ confidence interval is the convex hull of set $C = \{p; f(p) \geq \alpha\}$ where f is the so-called *confidence function* (defined in Blaker (2000)). Set C may be an interval but frequently it is a union of disjoint intervals.

Beyond the theory, the paper by Blaker (2000) includes a simple program in R (later corrected in Blaker (2001)) for confidence limits calculation in the binomial case. The search for the lower (upper) confidence bound starts in the lower (upper) Clopper-Pearson confidence limit an proceeds up (down) with a constant step as long as the confidence function values keep below α. Less known than the algorithm itself are, perhaps, its drawbacks: It is prone, when C is discontinuous, to skipping short segments and finding an incorrect solution (which results in coverage below $1-\alpha$). The risk of such failures may be reduced (though not eliminated) by setting the step very short. This, however, leads to a drastic slow-down of the calculations.

In the present work, a new algorithm, free of the drawbacks of the original Blaker's algorithm, is proposed. It is based on several lemmas on the confidence function properties. These allow for determining effectively such interval I that the (lower or upper) Blaker's confidence bound being searched either coincides with one of the limits of I, or can be safely searched for in the interior of I by interval halving (or similar standard procedures). The algorithm finds the correct confidence limits reliably and remains fast even when a high accuracy is required.

Keywords: Blaker's confidence limits, binomial distribution, algorithm

References

BLAKER, H. (2000): Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics 28 (4), 783-798.*

BLAKER, H. (2001): Corrigenda: Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics 29 (4), 681.*

Acknowledgement. This work was supported by projects AV0Z10300504 and GA CR 205/09/1079.

Asymptotics and Bootstrapping in Errors-in-variables Model with Dependent Errors

Michal Pešta

Charles University in Prague, Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics
Sokolovská 83, 18675 Prague, Czech Republic, *pesta@karlin.mff.cuni.cz*

Abstract. An *errors-in-variables* (EIV) regression model with dependent errors is considered and a *total least squares* (TLS, see Golub and Van Loan (1980)) estimate is constructed. Its consistency and asymptotic normality for *weak dependent* observations (α- and ϕ-mixing, Bradley (2005)) are proved. TLS estimate is highly nonlinear and, moreover, the asymptotic variance depends on unknown quantities, which cannot be estimated. Because of this, many statistical procedures for constructing confidence intervals and testing hypotheses cannot be applied. One possible solution to this dilemma is bootstrapping. Justification for use of the *moving block bootstrap* (MBB, Lahiri (2003)) technique is given. The results are illustrated through a simulation study. An application of this approach to real data is presented.

Keywords: errors-in-variables, dependent errors, mixing, bootstrap

References

BRADLEY, R. C. (2005): Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys 2, 107-144.*

GOLUB, G. H. and VAN LOAN, C. F. (1980): An analysis of the total least squares problem *SIAM Journal on Numerical Analysis 17 (6), 883-893.*

LAHIRI, S. N. (2003): *Resampling Methods for Dependent Data.* Springer-Verlag, New York.

A Power Comparison for Testing Normality

Shigekazu Nakagawa[1], Hiroki Hashiguchi[3], and Naoto Niki[2]

[1] Kurashiki University of Science and the Arts
Kurashiki, 712-8505, Japan, *nakagawa@cs.kusa.ac.jp*
[2] Saitama University
Saitama, 338-8570, Japan, *hiro@ms.ics.saitama-u.ac.jp*
[3] Tokyo University of Science
Tokyo, 162-8601, Japan, *niki@ms.kagu.tus.ac.jp*

Abstract. This paper is intended to report a power comparison of the newly developed omnibus test statistic for normality described in Nakagawa et al. (2008). It is such an improvement of Jarque–Bera (1987) test statistic that its cumulants hold the Cornish–Fisher assumption in normal sampling. A normalizing transformation based on the Wilson–Hilferty transformation is also given.

In this paper, three competitors are considered: the original test of Jarque–Bera (1987), the extension test of Urzúa (1996), and the test of Shapiro–Wilk (1965). According to the numerical examples in Thadewald and Buning (2007), the power comparison is conducted via Monte Carlo simulation under alternatives are contaminated normal distributions with varying mean and variance parameters and different proportions of contamination.

We show that the power of the improved Jarque–Bera test is slightly superior to that of the original Jarque–Bera for symmetric distributions with medium up to long tails.

Keywords: contaminated normal, Jarque–Bera test, Monte Carlo simulation

References

JARQUE, C. M. and BERA, A. K. (1987): A test for normality of observations and regression residuals. *Internat. Statist. Rev., 55(2), 163–172.*

NAKAGAWA, S., NIKI, N. and HASHIGUCHI, H. (2008): Numerical comparisons of power of omnibus test for normality. *Proceeding of COMPSTAT'2008*, 769–774.

SHAPIRO, S. S. and WILK, M. B. (1965): An analysis of variance test for normality: Complete samples. *Biometrika, 52, 591–611.*

THADEWALD, T. and BUNING, H. (2007): Jarque–Bera Test and its Competitors for Testing Normality–A Power Comparison. *Journal of Applied Statistics, 34(1), 87–105* .

URZÚA, C. M.(1996): On the correct use of omnibus tests for normality. *Econom. Lett., 90(3), 304–309.*

Genetics and/of basket options

Wolfgang K. Härdle[1] and Elena Silyakova[2]

[1] Ladislaus von Bortkiewicz Chair of Statistics, Humboldt-Universitäjt zu Berlin, Germany, *stat@wiwi.hu-berlin.de*
[2] Ladislaus von Bortkiewicz Chair of Statistics, Humboldt-Universitäjt zu Berlin, Germany *silyakoe@cms.hu-berlin.de*

Abstract. Basket options are over-the-counter derivatives. There are various types of such products traded on the market either separately or as a part of more sophisticated structures. The buyers of such derivatives benefit from the portfolio effect that makes possible to buy volatility cheaper in comparison with single stock options. However to price and hedge such derivatives one needs to account for correlation structure of a basket. One of the ways, usually used in practice, is to estimate correlation from historical data of stock prices. However, intuitively, for derivatives pricing the forward-looking implied values would be of more use.

In this paper we are concerned about modeling implied correlations. This is a challenging task both in terms of computational burden and estimation error. First it cannot be observed directly and must be recovered from option prices. Second, since it is obtained from implied volatilities, it is is not constant over maturities and strikes. We also expect this object to change over time. To analyze structure and dynamics of implied correlation surfaces we consider the dynamic semiparametric factor model (DSFM), which assumes nonparametric loading functions and low-dimensional time series of factors. Factors and factor loadings are estimated by semiparametric methods. In such way we study the dynamics of the system in its low-dimensional representation. We make the inference of the whole system based on low-dimensional time series analysis.

The empirical analysis is made on the dynamics of implied correlation structure of the 30 DAX stocks.

Keywords: implied correlation, basket options, dynamic semiparametric models

References

ALEXANDER, C. (2001): *Market Models: A Guide to Financial Data Analysis.* Wiley.

FENGLER, M. R., HÄRDLE, W. K. and MAMMEN, E. (2007): A semiparametric factor model for implied volatility surface dynamics. *Journal of Empirical Finance, 10, 603-621.*

FENGLER, M. R., PILZ, K. F. and SCHWENDNER, P., (2007): Basket Volatility and Correlation In: *Volatility as an Asset Class.* Risk Publications, 95-131.

PARK, B. U., MAMMEN, E., HÄRDLE, W. K. and BORAK, S. (2009): Time Series Modelling With Semiparametric Factor Dynamics. *Journal of the American Statistical Association, 284–298.*

Part V

Friday August 27

Indefinite Kernel Discriminant Analysis

Bernard Haasdonk[1] and Elżbieta Pękalska[2]

[1] Institute of Applied Analysis and Numerical Simulation
University of Stuttgart, Germany, *haasdonk@mathematik.uni-stuttgart.de*
[2] School of Computer Science
University of Manchester, United Kingdom, *pekalska@cs.man.ac.uk*

Abstract. Kernel methods for data analysis are frequently considered to be restricted to positive definite kernels. In practice, however, indefinite kernels arise e.g. from problem-specific kernel construction or optimized similarity measures. We, therefore, present formal extensions of some kernel discriminant analysis methods which can be used with indefinite kernels. In particular these are the multi-class kernel Fisher discriminant and the kernel Mahalanobis distance. The approaches are empirically evaluated in classification scenarios on indefinite multi-class datasets.

Keywords: kernel methods, indefinite kernels, Mahalanobis distance, Fisher Discriminant Analysis

Data Dependent Priors in PAC-Bayes Bounds

John Shawe-Taylor[1], Emilio Parrado-Hernández[2], and Amiran Ambroladze

[1] Dept. of Computer Science & CSML, University College London
London, WC1E 6BT, UK, *jst@cs.ucl.ac.uk*
[2] Dept. of Signal Processing and Communications, University Carlos III of Madrid
Leganés, 28911, Spain, *emipar@tsc.uc3m.es*

Abstract. One of the central aims of Statistical Learning Theory is the bounding of the test set performance of classifiers trained with i.i.d. data. For Support Vector Machines the tightest technique for assessing this so-called generalisation error is known as the PAC-Bayes theorem. The bound holds independently of the choice of prior, but better priors lead to sharper bounds. The priors leading to the tightest bounds to date are spherical Gaussian distributions whose means are determined from a separate subset of data. This paper gives another turn of the screw by introducing a further data dependence on the shape of the prior: the separate data set determines a direction along which the covariance matrix of the prior is stretched in order to sharpen the bound. In addition, we present a classification algorithm that aims at minimizing the bound as a design criterion and whose generalisation can be easily analysed in terms of the new bound.

The experimental work includes a set of classification tasks preceded by a bound-driven model selection. These experiments illustrate how the new bound acting on the new classifier can be much tighter than the original PAC-Bayes Bound applied to an SVM, and lead to more accurate classifiers.

Keywords: PAC Bayes Bound, Support Vector Machines, generalization prediction, model selection

Use of Monte Carlo when estimating reliability of complex systems

Jaromír Antoch[1], Julie Berthon[2], and Yves Dutuit[3]

[1] Charles University, Dept. of Statistique, Sokolovská 83, CZ – 185 75 Prague 8, Czech Republic; *jaromir.antoch@mff.cuni.cz*
[2] University Bordeaux 1, Bordeaux, France
[3] Thales Avionics, Le Haillan, France

Abstract. The behavior of components of complex systems and their interactions such as sequence- and functional-dependent failures, spares and dynamic redundancy management, and priority of failure events cannot be always adequately captured by traditional statistical approaches when trying to estimate reliability of large complex systems.

Among approaches suggested for coping with this type of problems an important role is played by fault trees, Petri nets, scan statistics and decision trees. Basically, most of the algorithms consist in looking on the data through a (sliding) window. However, in the general case counting arguments do not suffice because one has to consider individually each component, and this can easily result in a combinatorial explosion. Therefore, most of the methods were designed for computing lower and upper bounds of reliability rather than for computing exact values. Moreover, most of the methods assume that all components have the same reliability, being not the case of complex systems.

In the lecture we will focus especially on the k-out-of-n modeling. More precisely, we will concentrate on three approaches. Firstly, on dynamic modification of classical fault trees, which extend traditional approach by defining additional "dynamic gates" enabling to model complex interactions. Because state space becomes almost immediately too large for calculation with Markov models when the number of gate inputs increases, Monte Carlo simulation-based approach call to be used. Secondly, in the area of scan statistics, which can be applied here as well, most of the authors concentrated either on asymptotical or combinatorial approach. Conversely, we will show how Monte Carlo methods with suitably chosen Markov chain can be not only comparable but can increase the accuracy of results. Finally, we will compare Monte Carlo with combinatorial methods when solving k-out-of-n models and demonstrate their applicability and competitivity.

Respective approaches will be illustrated on two examples, i.e., a system with non-repairable components, and simplified scheme of a complex repairable system having both tested and maintained spares. The results obtained through the Monte Carlo will be compared with those obtained using the analytical approach. In addition, resulting estimates of reliability, failure and repair time distributions will be considered.

Keywords: Monte Carlo, k-out-of-n model, scan statistics, decision trees, fault tree analysis, Petri nets.

Some Algorithms to Fit some Reliability Mixture Models under Censoring

Laurent Bordes[1] and Didier Chauveau[2]

[1] Université de Pau et des Pays de l'Adour
Laboratoire de Mathématiques et de leurs Applications
UMR CNRS 5142 - Avenue de l'Université - BP 1155
64013 Pau Cedex, France *laurent.bordes@univ-pau.fr*
[2] Université d'Orléans
UMR CNRS 6628 - MAPMO - BP 6759
45067 Orléans Cedex 2, France *didier.chauveau@univ-orleans.fr*

Abstract. Estimating the unknown parameters of a reliability mixture model may be a more or less intricate problem, especially if durations are censored. We present several iterative methods based on Monte Carlo simulation that allow to fit parametric or semiparametric mixture models provided they are identifiable. We show for example that the well-known data augmentation algorithm may be used successfully to fit semiparametric mixture models under right censoring. Our methods are illustrated by a reliability example.

Keywords: reliability, mixture models, stochastic EM algorithm, censored data

Computational and Monte-Carlo Aspects of Systems for Monitoring Reliability Data

Emmanuel Yashchin[1]

IBM, Thomas J. Watson Research Ctr., Box 218,
Yorktown Heights, NY 10598, USA, *yashchi@us.ibm.com*

Abstract. Monitoring plays a key role in today's business environment, as large volumes of data are collected and processed on a regular basis. Ability to detect onset of new data regimes and patterns quickly is considered an important competitive advantage. Of special importance is the area of monitoring product reliability, where timely detection of unfavorable trends typically offers considerable opportunities of cost avoidance. We will discuss detection systems for reliability issues built by combining Monte-Carlo techniques with modern statistical methods rooted in the theory of Sequential Analysis, Change-point theory and Likelihood Ratio tests. We will illustrate applications of these methods in computer industry.

Keywords: SPC, lifetime data, wearout, warranty

Computational Statistics Solutions for Molecular Biomedical Research: A Challenge and Chance for Both

Lutz Edler, Christina Wunder, Wiebke Werft, and Axel Benner

Department of Biostatistics-C060, German Cancer Research Center
Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany,
edler@dkfz.de, c.wunder@dkfz.de, w.werft@dkfz.de, benner@dkfz.de

Abstract. Computational statistics, supported by computing power and availability of efficient methodology, techniques and algorithms on the statistical side and by the perception on the need of valid data analysis and data interpretation on the biomedical side, has invaded in a very short time many cutting edge research areas of molecular biomedicine. Two salient cutting edge biomedical research questions demonstrate the increasing role and decisive impact of computational statistics. The role of well designed and well communicated simulation studies is emphasized and computational statistics is put into the framework of the International Association of Statistical Computing (IASC) and special issues on Computational Statistics within Clinical Research launched by the journal Computational Statistics and Data Analysis (CSDA).

Keywords: computational statistics, molecular biomedical research, simulations, International Association of Statistical Computing, computational statistics and data analysis

Index