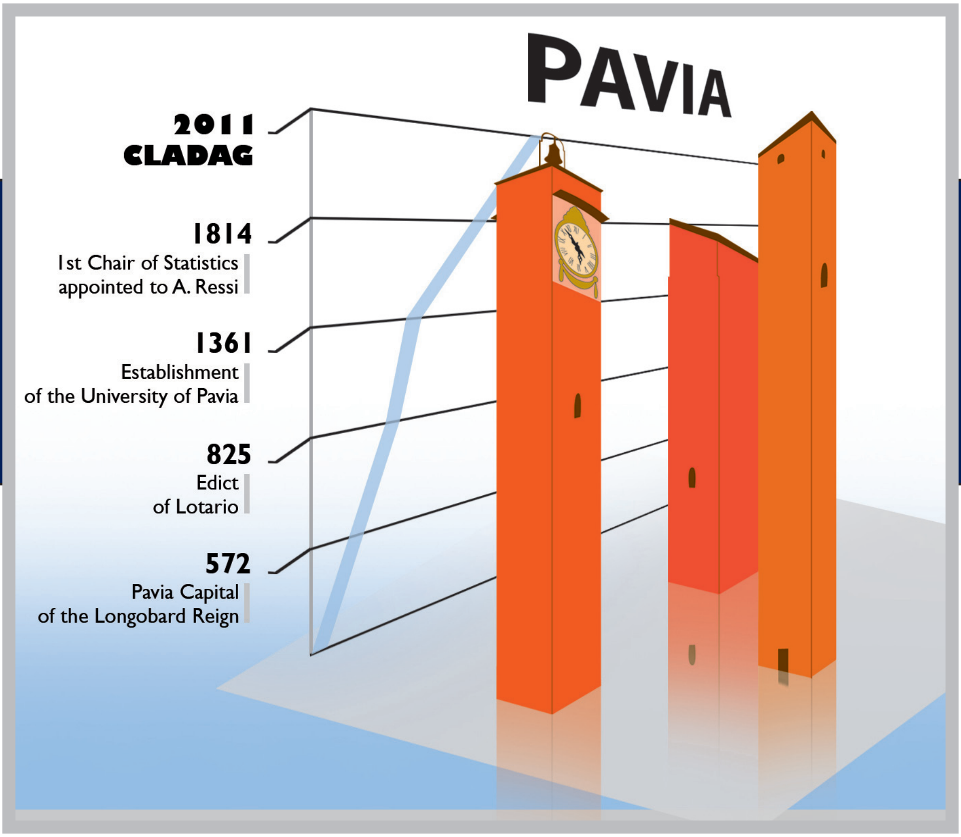

CLADAG 2011
Book of Abstracts

8th Scientific Meeting of the CLAssification and Data Analysis Group of the Italian Statistical Society

University of Pavia, September 7-9, 2011

Edited by

Paola Cerchiello – Claudia Tarantola

Pavia University Press
Editoria scientifica

CLADAG 2011

CLAssification and Data Analysis Group
of the Italian Statistical Society

CLADAG 2011
Book of Abstracts

8th Scientific Meeting

University of Pavia, September 7-9, 2011

Edited by

Paola Cerchiello – Claudia Tarantola

PaviaUniversityPress

Società italiana di statistica : Gruppo di classificazione e analisi dei dati

Cladag 2011 : book of abstracts : 8th scientific meeting : University of Pavia, September 7-9, 2011 / edited by Paola Cerchiello, Claudia Tarantola. - Pavia : Pavia University Press, 2011. - VIII, 56 p. ; 24 cm.

In testa al front.: Cladag 2011, [Università degli Studi di Pavia], Classification and data analysis group of the Italian statistical society, Società italiana di statistica, AIFIRM.

ISBN 9788896764220

I. CLADAG II. Cerchiello, Paola III. Tarantola, Claudia

1. Analisi Multivariata - Statistica - Congressi - Pavia - 2011

519.535 CDD-22 - Analisi dei dati. Analisi multivariata

ISBN: 978-88-96764-22-0

Texts published by Pavia University Press in the series “Editoria scientifica” have been peer-reviewed prior to acceptance by the Editorial Board.

Editore / Publisher: Pavia University Press
Edizioni dell'Università degli Studi di Pavia
Via Ferrata, 1 – 27100 Pavia – Italia
<http://www.paviauniversitypress.it>

Stampato da / Printed by: Print Service – Strada Nuova, 67 – 27100 Pavia – Italia

Table of Contents

CLADAG 2011

ised
INGEGNERIA DEI SISTEMI

GRUPPO BANCARIO
Credito
Valtellinese

Cladag 2011 - Program Schedule

Wednesday, September 7, 2011

09:00 - 09:30	**Opening Ceremony (Aula Magna): A. Stella (Rector of the University of Pavia); A. Cattaneo (Mayor of Pavia); S. Ingrassia (SC President); P. Giudici (OC President)**			
09:30 - 10:20	**Keynote Lecture - M. Titterington: "A review of some hybrid generative-discriminative methods"** (*Chair: A. Cerioli* , Aula Magna)			
10:20 - 10:50	***Coffee Break***			
Solicited Session	**Quality indexes for sensory and consumer science** *Chair: S. Salini* (Room Volta)	**Item response theory** *Chair: T. Di Battista (Room Foscolo)*	**Latent variable model, causal inference and evaluation** *Chair: I. Morlini* (Room Scarpa)	
10:50 - 11:10	E. Brentari, P. Zuccolotto	L.Bertoli Barsotti, A. Punzo	A. Forcina, S. Modica	
11:10 - 11:30	E. Brentari, R. Levaggi	T. Di Battista, S. Di Zio, C. Ceccatelli, R. Marianacci	L. Grilli R. Varriale	
11:30 - 11:50	P. Amenta, R. Lombardo	M. Matteucci, S. Mignani, B. Veldkamp	T. Agasisti, F. Pennoni, G. Vittadini	
Specialized Session	**Advances in Measurement Error Models and Methods** *Chair: M. Carpita* (Room Volta)	**Robust Clustering** *Chair: O. Papaspiliopoulos* (Room Foscolo)	**Advances in Bayesian structural learning for data analysis** *Chair: G. Consonni* (Room Scarpa)	
11:55 - 12:15	M. Battauz, R. Bellio	D. Perrotta, A. Cerioli, F. Torti	L. La Rocca	
12:15 - 12:35	A. Sarra, L. Fontanella, T. Di Battista, R. Di Nisio	A. Gordaliza , C. Ruwet, L. A. García-Escuder, A. Mayo-Isc	D. Rossell, D. Telesca, V. E. Johnson	
12:35 - 12:55	P. Lovaglio, G. Vittadini	M. Bini, M. Velucchi	R. Silva	
12:55 - 13:10	*Discussant: M. Carpita*	*Discussant: M. Riani*	*Discussant: G. Consonni*	
13:10 - 14:10	***Lunch***			
Contributed Session	**Multivariate Data Analysis** *Chair:M. D'Esposito* (Room Volta)	**Classification and Clustering (1)** *Chair: P. Amenta* (Room Magna)	**Data Analysis in Economic and Finance (1)** *Chair: G.Vittadini (Room Scarpa)*	**Data Analysis in Environmental Sciences** *Chair: G. Diana* (Room Foscolo)
14.10 - 14:25	R. Bellio, N. Soriani	L. Bagnato, F. Greselin	T. Bellini	E. Nissi, A. Rapposelli
14:25 - 14.40	L. Cutillo, I. De Feis, C. Nikolaidou, T. Sapa	I. Morlini, S.Zani	L. Dalla Valle, M.E. De Giuli, C. Manelli, C. T	R. Rotondi, E. Varini, G. Zonno
14:40 - 14:55	G. Fonseca, F. Giummolè, P. Vidoni	R. O'Reilly, S. Wilson, C. Carlow, P. McNich	P. Dellaportas, L. Bottolo	M. Tsuji, K. Horinouchi, Y. Izumoto, T. Sh
14:55 - 15:10	H. Hruschka	I. Vrbik, P. McNicholas	G. Schoier, F. Marsich	
15:10 - 15:25	A. Mazza, A. Punzo	R. Winkler, F. Klawonn, R. Kruse	R. Stecking, K. B. Schebesch	
Specialized Session	**New trends in incremental and semi-supervised classification : theoretical issues and applications** *Chair: P. Kuntz* (Room Volta)	**Web mining and Web surveys** *Chair: S. Biffignandi* (Room Foscolo)	**Scaling procedures for ordinal data and open issues in the measurement of latent variables** *Chair: G. Boari* (Room Scarpa)	
15:30 - 15:50	N. Greffard , F. Picarougne	J. Bethlehem	L. Cappellari	
15:50 - 16:10	R. Lefort, R. Fablet, J.-M. Boucher	A.Bianchi, S. Biffignandi	A. Zanella, G.Boari, A. Bonanomi, G. Cantaluppi	
16:10 - 16:30	C. Salperwyck, V. Lemaire	E. Lorenzini	B. Zumbo	
16:30 - 16:45	*Discussant: P.Kuntz*	*Discussant: S. Biffignandi*	*Discussant: G. Boari*	
14:45 - 17:10	***Coffee Break***			
Solicited Session	**Evaluation of Academic Research Performance: Measurement and data analysis** *Chair: G. Nicolini (Room Volta)*	**Multivariate statistics** *Chair: R. Rocci* (Room Foscolo)	**Clustering of multivariate functional data: techniques and applications** *Chair: M. Chiodi (Room Scarpa)*	
17:10 - 17:30	F. De Battisti, S. Salini	M. Alfò, I. Rocchetti	M. Chiodi, G. Adelfio, A. D'Alessandro, D. Luzio	
17:30 - 17:50	A. Costantini, M. Franceschet	G. Calò, A. Montanari	A. Guglielmi, F.Leva, A.M. Paganoni, F. Ruggeri, J. Sorian	
17:50 - 18:10	C. Davino, R. Romano	A. Okada, H. Tsurumi	E. Romano, A. Balzanella	
18:15 - 19:30	***CLADAG Assembly*** (AulaFoscolo)			

Cladag 2011 - Program Schedule

Thursday, September 8, 2011

Specialized Session	**Advances in Student-t based statistical modeling** *Chair: F. Greselin* (Room Foscolo)	**Analysis of asymmetric relationships** *Chair: A. Okada* (Room Volta)	**Analysis and classification of non standard data** *Chair: G. Ragozini* (Room Scarpa)
08:30 - 08:50	A. Kleefeld, V. Brazauskas	G. Bove	A. M. Alonso - D. Casado - S. López-Pintado - J. Romo
08:50 - 09:10	P. McNicholas	M. de Rooij	A. Balzanella, L. Rivoli, R. Verde
09:10 - 09:30	R. Fucci, A.C. Monti	M. Nakai	P. Secchi, S. Vantini, V. Vitelli
09:30- 09:45	*Discussant: F. Greselin*	*Discussant: A. Okada*	*Discussant: G. Ragozini*

Solicited Session	**Classification problems in finance** *Chair: D. Zappa* (Room Foscolo)	**Modern computational methods for classification of** *Chair: F. Laurini (Room Volta)*	**Regression models for spatial data** *Chair: M. Mezzetti* (Room Scarpa)
09:50 - 10:10	A. Bramante, A. Cipollini, A. Manzini	M. Cattelan, C. Varin	R. Chambers, E. Dreassi, N. Salvati
10:10 - 10:30	G. Gabbi	N. Chopin, P. Jacob, O. Papaspiliopoulos	M. Bevilacqua, C. Gaetan, E. Porcu
10:30 - 10:50	E. Otranto	P. Pauli	S. Leorato, M. Mezzetti
10.50 - 11:15	***Coffee Break***		

Solicited Session	**Statistical methods for the assessment of universities** *Chair: L. Grilli* (Room Foscolo)	**Statistical issues and inferential results in CUB models** *Chair: D. Piccolo* (Room Volta)	**Symbolic data analysis** *Chair: A. Irpino* (Room Scarpa)
11:15 - 11:35	M. Attanasio, G. Boscaino, V. Capursi, A. Plaia	S. Bonnini, L. Salmaso, F. Solmi	E. Diday
11:35 - 11:55	M. Bini, L. Grilli	L. Deldossi, R. Paroli	C. Drago, C.Lauro, G. Scepi
11:55 - 12:15	P. Cerchiello	M. Iannario	R. Verde, A. Irpino

Contributed Session	**Classification and Data Analysis (1)** *Chair: S. Zaccarin (Room Foscolo)*	**Data Analysis in Economic and Finance** *Chair: S. Ingrassia (Room Magna)*	**Latent Class Model, Classification and** *Chair: D. Calò (Room Scarpa)*	**Non Parametric statistics** *Chair: P. Cerchiello (Room Volta)*
12.20 - 12:35	J. Andrews, P. D. McNicholas	A. Dridi, M. El Ghourabi, M. Limam	S. Bacci, F. Bartolucci, M. Gnaldi	L. Dalla Valle, G. Nicolini
12:35 - 12.50	S. Bolasco, P. Pavone	D. F. Iezzi, M. Mastrangelo, S. Sarlo	F. Bartolucci, G. E. Montanari, S. Pandolfi	D. De Stefano
12.50 - 13.05	F. Campobasso, A. Fanizzi	A. Lourme, C. Biernacki	F. Bassi, M. Croon, A. Pittarello	M. Manisera
13:05 - 13.20	F. Caruso, G. Giuffrida, C. Zarba	F. Telmoudi, M. El Ghourabi, M. Limam	M. Gnaldi, F. Bartolucci, S. Bacci	E. Raffinetti
13:20 - 13.35	A. Unlu, A. Sargin		I. Sulis	F. Ferraty, A. Goia, E. Salinelli, P. Vieu
13:35 - 14:40	***Lunch***			

Specialized Session	**Genetic algorithms and surveys** *Chair: M. Gasparini (Room Foscolo)*	**Hidden Markov models: theory developments and** *Chair: A. Maruotti* (Room Volta)	**Statistics for Shape, Image and Functional data** *Chair: R. Verde* (Room Scarpa)
14:40 - 15:00	M.Ballin, G. Barcaroli	J. Bulla	L. Fontanella, L. Ippoliti, P. Valentini, F. Festa
15:00 - 15:20	R. Di Manno, M. Scanu, P. Vicard	F. Lagona, M.Picone	S.A. Gattone
15:20 - 15:40	D. Zardetto, M. Scannapieco	R. Langrock, I. L. MacDonald, Walter Zucchini	C. Glasbey
15:40 - 15:55	*Discussant: M. Gasparini*	*Discussant: A. Maruotti*	*Discussant: R. Verde*
16:00 - 16:40	**Keynote Lecture - F. Palumbo "Exploratory analysis for interval-valued data"** (*Chair: O. Papaspiliopoulos* , Aula Magna)		
16:40 - 17.05	***Coffee Break***		

Contributed Session	**Classification and Data Analysis (2)** *Chair: P.Bove (Room Foscolo)*	**Classification and data Analysis (2)** *Chair: S.Figini (Room Volta)*	**Service evaluation and customer** *Chair: M.Ballin (Room Scarpa)*	**Classification models** *Chair: P.Giudici (Room Magna)*
17:05 - 17:20	L. Bocci, M. Vichi	F. Crippa, M. Marelli	C. Capuano, D. De Stefano, A. Del Monte, I	A. Attanasio M. Maravalle, C. Marziliano
17:20 - 17.35	B. Franczak, R. Browne, P. McNicholas	R. D'Agata, V. Tomaselli	C. Liberati, P. Mariani	V. Calzati
17.35 - 17:50	F. Martella, D. Vicari, M. Vichi	L. Dancelli, M. Manisera, M. Vezzoli	F. Musella, P. Vicard	S. Fontanella, C. Fusilli and L. Ippoliti
17:50 - 18.05	E. Sironi	F. De Natale, L. Fattorini, S. Franceschin,P.	M. Scagliarini, S. Evangelisti	B. Hagen, A. Zucchella, P. Cerchiello, N. De
18:05 - 18.20	A. Tarsitano, M. Falcone	N. Solaro	C. Tarantola, P. Vicard, I. Ntzoufras	
20.00	***Social Dinner***			

Friday, September 9, 2011

Specialized Session	**Misura integrata dei rischi (in collaborazione con** *Chair: M. Anolli (Room Foscolo)*	**Classification methods for time dependent data** *Chair: C. Perna (Room Scarpa)*	**Statistical methods for high-dimensional biological** *Chair: A. Montanari (Room Volta)*
09:00 - 09:20	M. Bignami	G. Cubadda, B. Guardabascio, A. Hecq	M.R. Oelker, A. L. Boulesteix
09:20 - 09:40	S. Figini	M. Gerolimetto, I. Procidano	D. Causeur
09:40 - 10:00	D. Mignacca	M. Niglio, G. Storti, C. Vitale	L. Trinchera, E. Le Floch, A. Tenenhaus
10:00- 10:15	*Discussant: M. Anolli*	*Discussant: C. Perna*	*Discussant: A.Montanari*
10:20 - 11:10	**Keynote Lecture - V. Batagelj "Clustering of large data sets of mixed units"** *(Chair: M. Vichi, Aula Magna)*		
11:10 - 11:30	***Coffee Break***		
Solicited Session	**Integrating administrative data and large scale surveys** *Chair: M. Scanu* (Room Foscolo)	**New trends in complexity reduction in multivariate** *Chair: G.D. Costanzo (Room Scarpa)*	**Advanced methods in network analysis** *Chair: M. P. Vitale (Room Volta)*
11:30 - 11:50	P.L. Conti, D. Marella	C. Cappelli, F. Di Iorio, P. D'Urso	V. Amati
11:50 - 12:10	M. D'Orazio	F. Domma, F. Condino	D. De Stefano
12:10 - 12:30	B. Vantaggi, A. Capotorti	D. Vicari, M. Alfò	M. van Duijn
12:30 - 13:30	***Lunch***		

Contributed Session	**Data Analysis in Economic and Finance** *Chair: M. Costa (Room Foscolo)*	**Data Analysis (1)** *Chair: D. Vicari (Room Magna)*	**Classification and Clustering (3)** *Chair: P. Giudici (Room Volta)*	**Data Analysis (2)** *Chair: C. Tarantola* (Room Scarpa)
13.45 - 14:00	R. Castellano, G. Punzo	S. Calligaris, F. Mecatti	C. Agostinelli, M. Romanazzi	A. Costa
14:00 - 14:15	F. Della Ratta Rinaldi, F Gallo, B. Lorè	T. Di Battista, S. Di Zio, C. Ceccatelli, R. Ma	B. Arpino, F. Billari, M. Cannas	P. Mariani, E. Zavarrone
14:15 - 14:30	S. Lombardi, F. Verrecchia	G. D'Epifanio	A. Guglielmi, F. Ieva, A. M. Paganoni, F. Ru	G. Morandi
14:30 - 14:45	S. Minotti, G. Spedicato	G. Giuliani, R. Lima, R. Ranaldi	L. Scrucca	I. Sulis, M. Porcu
14.45 - 15:00		A. Simonetto, M. Carpita	S. Subedi, P. McNicholas	M. Vezzoli, E. Zavarrone
15:15 - 17:30	***Symposium: E. Giovannini (ISTAT), N. Pagnoncelli (IPSOS), E. Rocca (CREVAL), M. Degli Esposti (VIA-Academy), P. Giudici (UNIPV) B. Marchione (AICUN)***			
17.30	***Closing ceremony: S. Ingrassia, P. Giudici***			

Presentation

The 8th Biennial International Meeting of the CLAssification and Data Analysis Group (CLADAG) of the Italian Statistical Society was hosted by the University of Pavia within the celebration of its 650th anniversary, from September 7th to September 9th, 2011.

The present book contains the abstract of the papers presented during the meeting. The four pages version of the papers is contained in the USB pen, with an ISBN code as well. All papers were reviewed.

CLADAG promotes advanced methodological research in multivariate statistics with a special interest in Data Analysis and Classification. It supports the interchange of ideas in these fields of research, including the dissemination of concepts, numerical methods, algorithms, computational and applied results.

CLADAG is a member of the International Federation of Classification Societies (IFCS). Among its activities, CLADAG organizes a biennial international scientific meeting, schools related to classification and data analysis, publishes a newsletter and cooperates with other member societies of the IFCS in the organization of their conferences.

The scientific program of the meeting covered the following topics:

- Classification theory
- Multivariate data analysis
- Proximity structure analysis
- Software developments
- Applied classification and data analysis

Previous CLADAG meetings were held in Pescara (1997), Roma (1999), Palermo (2001), Bologna (2003), Parma (2005), Macerata (2007) and Catania (2009).

Scientific Program Committee

Salvatore Ingrassia (University of Catania, Chairman)
Pietro Amenta (Sannio University)
Marco Ballin (ISTAT)
Michele Costa (University of Bologna)
Damiana Costanzo (University of Calabria)
Maria Rosaria D'Esposito (University of Salerno)
Giancarlo Diana (University of Padova)
Tonio Di Battista (University "D'Annunzio" of Chieti-Pescara)
Paolo Giudici (University of Pavia)
Marco Riani (University of Parma)
Roberto Rocci (University of Roma "Tor Vergata")
Susanna Zaccarin (University of Trieste)
Diego Zappa (University Catholic of Sacred Heart, Milan)

Local Organizing Committee

Paolo Giudici (University of Pavia, Chairman)
Paola Cerchiello (University of Pavia)
Silvia Figini (University of Pavia)
Claudia Tarantola (University of Pavia)

ABSTRACTS

Extending Value-Added Models of Educational Production: Stochastic Processes and Clustering

Tommaso Agasisti – Fulvia Pennoni – Giorgio Vittadini

Politecnico of Milan, Italy, e-mail: tommaso.agasisti@polimi.it
University of Milano-Bicocca, Italy, e-mail: {fulvia.pennoni; giorgio.vittadini}@unimib.it

In the economics of education, educational production functions are widely used to evaluate the impact of different factors such as socio-economic background and school's characteristics on students' achievement. We discuss the main features of traditional value added models used in this context. We suggest the use of a multilevel latent Markov Rasch model, recently proposed in the literature, to improve the understanding of educational processes. An empirical illustration is reported in the last section.

Ordering Curves by Data Depth

Claudio Agostinelli – Mario Romanazzi

University "Ca' Foscari", Venice, Italy, e-mail: {claudio; romanaz}@unive.it

Application of depth methods to functional data provides new tools of analysis, in particular an ordering of curves from the center outwards. Two specific depth definitions are band depth and half-region. Another research area is local depth aimed to identify multiple centers and dense subsets of the space. In this work we suggest local versions for both band and half-region depth and illustrate an application with real data.

Mixed Effect Models for Multivariate Mixed Responses

Marco Alfò – Irene Rocchetti

University "La Sapienza" of Rome, Italy, e-mail: {marco.alfo; irene.rocchetti}@uniroma1.it

We describe a regression model for mixed bivariate responses, where association between outcomes is modeled through latent effects, accounting for both heterogeneity and dependence. In a Finite Mixture (FM) context a relevant question arises when dependence should be tested vs independence. Discrete bivariate mixing distributions lead to a (almost) perfect dependence between the two random effects: each location in a margin is associated to only one location in the other one. We relax the undimensionality of the standard FM model, where the joint and the marginal distributions have the same number of components and the same masses, and proceed to define a multidimensional latent class structure, with a possibly different number of locations in each margin.

Time Series Classification

Andrés M. Alonso – David Casado – Sara López-Pintado – Juan Romo

Universidad Carlos III de Madrid, Spain, e-mail: {andres.alonso; david.casado; juan.romo@uc3m.es}
Columbia University, NewYork, USA, e-mail: sl2929@columbia.edu

We propose to classify time series by using functional data techniques. We transform the time series classification into a curves classification problem by considering the integrated periodograms. A new time series is assigned to the group minimizing the distance from its integrated periodogram to the average of the group integrated periodograms. We also extend the technique to non stationary series by considering the periodogram locally. Considering ideas from functional data depth, we make the classification robust. The method shows a good behaviour, both with simulated and real data.

Estimating the Parameter of the Stochastic Actor-Oriented Model: New Statistics and the Generalized Method of Moments

Viviana Amati

University of Konstanz, Germany, e-mail: viviana.amati@uni-konstanz.de

In the stochastic actor-oriented model (SAOM) for longitudinal network data, the most often used procedure for parameter estimation is the Method of Moments (MoM), which estimates the parameters using one observed statistic for each estimated parameter. Here a different estimation method is proposed, which can use more statistics per parameter. Having more than one statistic for a single parameter leads to an over-identified system of equations, so that the ordinary MoM cannot be applied. A suitable method then is the Generalized Method of Moments (GMM), which involves the minimization of a quadratic function of the differences between the expected values of the statistics and their sample counterparts. An optimization simulation algorithm based on the Newton-Raphson step is used to approximate the solution of the minimization problem and to compare the GMM estimator with the estimators deriving from the MoM and the Maximum Likelihood procedures.

Semi-Supervised Variable Selection for Model-Based Clustering and Classification

Jeffrey L. Andrews – Paul D. McNicholas

University of Guelph, Ontario, Canada, e-mail: andrewsj@uoguelph.ca

As data sets continue to grow in size and complexity, efficient methods are needed to simplify and target important features in the variable space. Many of the variable selection techniques that are used alongside clustering algorithms are based upon determining the best variable subspace according to model fitting in a stepwise manner. These techniques are often computationally intensive and can require extended periods of time to run. In this paper, a novel variable selection technique is introduced for use in cluster and classification analyses that is both intuitive and highly computationally efficient.

Machine Learning Techniques for Propensity Score Matching with Clustered Data. A Simulation Study and an Application to Birth Register Data

Bruno Arpino – Francesco Billari – Massimo Cannas

Bocconi University, Milan, Italy,
e-mail: {bruno.arpino; francesco.billari; massimo.cannas} @unibocconi.it

Causal inference in the medical sciences is often complicated by data having a hierarchical structure. In typical observational data patients are clustered in different practitioners and/or hospitals. An increasing number of studies shows that cluster-level variables can have a considerable effect both on the treatment intake and the outcome, i.e., they are potential confounder variables which can bias causal estimates if not controlled for. Via Monte-Carlo simulations, we assess the performance of alternative strategies for the propensity score estimation in the context of hierarchical data. We consider fixed and random effects models, as recently proposed in the treatment effects literature, and Machine Learning algorithms, which have recently been applied in the single-level context. In simulation scenarios we vary the cluster size, the number of clusters and the role of the cluster-level covariate on treatment assignment. Finally, we present an empirical analysis focusing on a widespread obstetric treatment i.e. labor induction, aimed at evaluating its impact on maternal outcomes. We use birth-register data on 8565 patients clustered in 24 hospitals in the Italian region of Sardinia.

Indicators and Measures for the Assessment of University Students' Careers

Massimo Attanasio – Giovanni Boscaino – Vincenza Capursi – Antonella Plaia

University degli Studi di Palermo, Italy, e-mail: {attana;gioboscaino;capursi;plaia}@unipa.it

In the Italian University System the problem of student failure and of delaying the degree is causing increasing concern both for universities and for stake-holders. In this paper we compare the teaching performances of the Italian universities, and individual cohort data of three Faculties of the University of Palermo, to extract some information useful to policy makers.

Different Criteria for the Optimal Number of Clusters. An Application to Extra Virgin Olive Oils Data with R

Alessandro Attanasio – Maurizio Maravalle – Ciro Marziliano

University of Aquila, Italy, e-mail: alessandro_attanasio@yahoo.it, {mm;ciro.marziliano}@cc.univaq.it

The cluster analysis is used very frequently in several fields such as statistics, data mining, marketing and machine learning. One of the most important problems of clustering is to define the number of classes. In fact it is not easy to find an appropriate method to measure whether the cluster configuration is acceptable or not. Another problem can be the selection of the best variables on which to cluster. Clustering algorithms are very sensitive to their input parameters, so it is important to evaluate and compare their results. In this paper we are going to analyze some criteria of clustering in order to differentiate and characterize Italian oils. Three algorithms are used in this study: K-means partitioning methods, Partitioning Around Medoids and the Heuristic Identification of Noisy Variables. Principal component analysis (PCA) is also applied for finding projection of maximal variability.

Multidimensional Latent Class IRT Models: An Extension to Items with Ordinal Responses

Silvia Bacci – Francesco Bartolucci – Michela Gnaldi

University of Perugia, Italy, e-mail: {silvia.bacci; bart; michela.gnaldi}@stat.unipg.it

We formulate a class of Item Response Theory models which may be used to assess the dimensionality of a set of items with ordinal responses. This class of models extends an existing one for items with dichotomous responses, which being multidimensional, represents the latent traits by a random vector. This vector has a discrete distribution with support points corresponding to different latent classes. Several parameterisations may be adopted in the extended class of models for the conditional distribution of the response variables given the abilities, such as those adopted in the Rating Scale model, in the Partial Credit model, and in the Graded Response model. In order to illustrate the proposed models, we analyze data coming from a study about levels of anxiety and depression in oncological patients, for which a Graded Response parameterisation is adopted.

Model-Based Clustering and Classification via Patterned Covariance Analysis

Luca Bagnato – Francesca Greselin

University Milano-Bicocca, Italy, e-mail: {luca.bagnato; francesca.greselin }@unimib.it

This work deals with the classification problem in the case that groups are known and both labeled and unlabeled data are available. The classification rule is derived using Gaussian mixtures, with covariance matrices fixed according to a multiple testing procedure, which allows to choose among four alternatives: heteroscedasticity, homometroscedasticity, homotroposcedasticity, and homoscedasticity. The mixture models are then fitted using all available data (labeled and unlabeled) and adopting the EM and the CEM algorithms. Applications on real data are provided in order to show the classification performance of the proposed procedure.

The Genetic Algorithm: Applications for the Optimisation of the Sampling Strategy

Marco Ballin – Giulio Barcaroli

Istituto Nazionale di Statistica (ISTAT), Rome, Italy, e-mail: {ballin;barcarol}@istat.it

Sampling design and estimation are the two components that lead to the definition of the sampling strategy. It is quite common that both components are based on a partition of the population and sample units in homogeneous classes: sampling strata and reference group for the calibration model. In this paper it is shown as the genetic algorithm can help in the search of the optimal solution in both steps: for sampling design, in finding the best stratification that ensures efficiency; in estimation, by determining the best reference group, the one that guarantees the minimum sampling variance and/or minimum bias due to non response. While this approach has already been studied and applied in ISTAT real surveys for sampling design, the case of estimation still needs a practical verification.

Data Stream Summarization by Histograms Clustering

Antonio Balzanella – Lidia Rivoli – Rosanna Verde

Second University of Naples, Italy, e-mail: {antonio.balzanella; rosanna.verde}@gmail.com,
University of Naples "Federico II", Italy, e-mail: lidia.rivoli@unina.it

In this paper we introduce a new strategy which allows to discover the concepts in an evolving data stream and to represent them through appropriate histograms. It is an on-line clustering algorithm where the prototype of each cluster is a histogram and data are allocated to clusters through a Wasserstein derived distance for histogram data.

Item Selection via Latent Class Based Clustering Methods

Francesco Bartolucci – Giorgio E. Montanari – Silvia Pandolfi

University of Perugia, Italy, e-mail: {bart; giorgio; pandolfi} @stat.unipg.it

Given a set of items used to measure a latent trait, we introduce a method for finding the smallest subset of these items which provides an amount of information close to that of the initial set. The method is based on the latent class (LC) model and proceeds by sequentially eliminating the items that do not significantly change the classification of the subjects in the sample with respect to that based on the full set of items. As usual, this classification is based on the posterior probabilities of belonging to each latent class. The approach is illustrated through an application concerning the evaluation of the quality-of-life of elderly people hosted in nursing homes, which is based on a dataset collected within the "Ulisse" project. For this dataset, we adopt an LC model for polytomous items, which also accounts for missing responses. To deal with multimodality of the model likelihood, we rely on a hierarchical clustering procedure to find sensible starting values for the EM algorithm used for parameter estimation.

A Latent Class Approach for Estimating Labour Market Mobility in the Presence of Multiple Indicators and Retrospective Interrogation

Francesca Bassi – Marcel Croon – Arianna Pittarello

University of Padova, Italy, email: bassi@stat.unipd.it, aripitta@libero.it
University of Tilburg, The Netherlands, email: M.A.Croon@uvt.nl

With panel data analysts can estimate labour force gross flows. Measurement errors in the observed state can induce bias in the estimation of transitions, leading to erroneous conclusions about labour market dynamics. A large body of literature on gross flows estimation is based on the assumption that errors are uncorrelated over time. This assumption is not realistic in many contexts, because of survey design and data collection strategies. We use a model-based approach to adjusting observed gross flows for classification errors, eventually correlated. A convenient framework is provided by latent class analysis, specifically by latent class Markov models. We apply our approach to data collected on the Italian labour market with the Continuous Labour Force Survey, which is cross-sectional, quarterly, with a 2-2-2 rotating design. The questionnaire allows to dis-

pose of multiple indicators of labour force condition for each quarter: two collected in the same interview and a third one collected after one year.

Clustering of Large Data Sets of Mixed Units

Vladimir Batagelj

University of Ljubljana, Republic of Slovenia, e-mail: vladimir.batagelj@fmf.uni-lj.si

In the paper we present an approach to clustering of (very) large data sets of mixed units – units measured in different scales. The approach is based on representation of units by symbolic objects (SOs) (Billard and Diday, 2006). The SOs can describe either single units or groups of initial units condensed into SOs in a pre-processing step. For clustering of SOs we adapted two classical clustering methods: leaders method (a generalization of k-means method) (Hartigan, 1975); Ward's hierarchical clustering method (Ward, 1963). Both adapted methods are compatible; they are based on the same criterion function; they are solving the same clustering problem. With the leaders method the size of the sets of units is reduced to a manageable number of leaders that can be further clustered with the compatible agglomerative hierarchical clustering method to reveal relations among them and (using the dendrogram) also to decide upon the right number of clusters. The proposed approach was successfully applied on different data sets: population pyramids, TIMSS, cars, foods, citation patterns of patents, and others.

Simple Fitting of Semiparametric Regression Models with Covariate Measurement Error

Michela Battauz – Ruggero Bellio

University of Udine, Italy, e-mail: {michela.battauz; ruggero.bellio}@uniud.it

Fitting semiparametric regression models in the presence of covariate measurement error is a challenging problem. We focus on scatter-plot smoothing and consider the linear mixed model representation of penalized splines. A structural approach is followed for measurement error modelling, and a normal distribution is assumed for the true unobserved covariate. The solution proposed is based on one-dimension numerical integration of the variable measured with error and the Laplace approximation for integrating out the random effects. An example illustrates the performance of the method, and a comparison with the approach based on Bayesian splines is provided.

Robust Integrated Banking Economic Capital

Tiziano Bellini

University di Parma, Italy, e-mail: tiziano.bellini@unipr.it

Economic capital models are potentially powerful tools for banking risk management and for the supervisory review process. In order to fulfil this potential, there is the need to go beyond the modular approach that dominates current regulatory capital requirements. For this reason we propose a fully integrated approach on which, exploiting a Monte Carlo simulation framework, we develop a model where both banking assets and liabilities are considered and risks are linked to macroeconomic variables. Concentrating on the integration between credit and interest rates, but considering the potential extension to other risk sources, we apply the forward search to estimate robust parameters. Focusing on outlier detection, this framework can be directly applied for stress testing purposes.

A Robust Algorithm for Maximum Likelihood Estimation of Exponential Random Graph Models

Ruggero Bellio – Nicola Soriani

University of Udine, Italy, e-mail: ruggero.bellio@uniud.it
University of Padua, Italy, e-mail: soria@stat.unipd.it

Maximum likelihood estimation for exponential random graph models (ERGMs) is computationally challenging, due to numerical difficulties to approximate the likelihood function. Markov Chain Monte Carlo methods can be used to obtain sampled networks for a given parameter value, and they are efficiently implemented in publicly available software. However, simulated maximum likelihood methods at times fail to converge as the likelihood approximation may degrade, especially for certain choices of the sufficient statistics of interest. A Monte Carlo quasi-Newton algorithm for computing the maximum likelihood estimate is introduced, borrowing some ideas from the method of maximization by parts. Two crucial parts of the proposed method are the steplength determination based on a simulated likelihood function, and a suitable backtracking mechanism to deal with model near degeneracy. The resulting algorithm is rather robust, and it is capable of estimating a broad array of ERGMs. A numerical example is provided.

Rasch-Rasch Modelization of Missing Data

Lucio Bertoli-Barsotti – Antonio Punzo

University of Bergamo, Italy, e-mail: lucio.bertoli-barsotti@unibg.it
University of Catania, Italy, e-mail: antonio.punzo@unict.it

An item response theory model for dichotomies, called Rasch-Rasch model, is introduced to modelize missing data. Its parameterization has the advantage to satisfy the following conditions: firstly, the missing-data process depends on a person latent trait – say response propensity – that is distinct from the latent ability (a similar bidimensional parameterization holds for the items); secondly, the model belongs to the Rasch family of models. Various Maximum likelihood approaches for the estimation of the RRM-parameters are described and relations with others existing models are finally provided.

Selection Problems of Web Surveys

Jelke Bethlehem

Statistics Netherlands, The Netherlands, e-mail: jbtm@cbs.nl

Web surveys are becoming a more and more a popular instrument for survey data collection. This is not surprising as they allow for collecting large amounts of data cheap and fast. However, there are methodological challenges. Examples are under-coverage and self-selection. These may cause survey results to be misleading. These selection problems are discussed and also possible correction techniques.

Covariance Tapering for Gaussian Multivariate Random Fields

Moreno Bevilacqua – Carlo Gaetan – Emilio Porcu

University of Bergamo, Italy, e-mail: moreno.bevilacqua@unibg.it
University "Ca' Foscari", Venice, Italy, e-mail: gaetan@unive.it
Universidad de Castilla la Mancha, Ronda de Toledo SN, Ciudad Real, Spain, e-mail: eporcu@uni-goettingen.de

In the recent literature there has been a growing interest in the construction cross-covariance functions for spatial random fields. However, effective estimation methods for these functions are somehow unexplored. The maximum likelihood method has attractive features but when we deal with large data set this solution becomes impractical, so computationally efficient solutions have to be devised. In this talk we present an ongoing work about a tapering solution to the estimating problem of the cross-covariance functions for multivariate Gaussian random fields. The effectiveness of our proposal is illustrated through a numerical example.

Web Panel Representativeness and Uses

Annamaria Bianchi – Silvia Biffignandi

University of Bergamo, Italy, e-mail: {annamaria.bianchi; silvia.biffignandi}@unibg.it

Reweighting a sample allows to adjust for known or expected discrepancies between sample and population, thus gaining representativeness. This paper is using a generalization of the maximum entropy weighting (MaxEnt) technique to reweight a sample to match observed population moments. Starting from this approach and combining it with other methods (such as Propensity Score Matching), different weighting schemes are computed on representative samples drawn from a data set which is assumed as population (called panel in the following). Survey data are merged to the population information. Then, different weighting schemes based on MaxEnt approach and Horvitz-Thompson method are applied for estimation purposes. Variables, whose value are known at panel (proxy of population) level are chosen for estimation, so that the panel value is used as benchmark for quality assessments of the obtained estimates. Results are compared and discussed.

Fitting Mixtures of Logit Regressions with the Forward Search

Matilde Bini – Margherita Velucchi

European University of Rome, Italy, e-mail: mbini@unier.it.
University of Florence, Italy, e-mail: velucchi@ds.unifi.it.

The forward search is a general method to detect multiple outliers and to determine their effect on inference about models fitted to data. From the monitoring of a series of statistics based on subsets of data of increasing size we obtain multiple views of any hidden structure. Sometimes, some features emerge unexpectedly during the progression of the forward search only when a specific combination of forward plots is inspected at the same time. These features have to be harmonized and linked together in order to give an exhaustive description of a complex problem. In this paper, we use a set of new robust graphical tools on a mixture of logit regressions. We use simulated data and we show the dynamic interaction with different "robust plots" to highlight the presence of groups of outliers and regression mixtures in the context of logit regression and highlight the effect that these hidden groups provide on the fitted model.

Partitioning Three-Way Dissimilarity Data

Laura Bocci – Maurizio Vichi

University "La Sapienza" of Rome, Italy, e-mail: {laura.bocci; maurizio.vichi}@uniroma1.it

This paper presents a methodology for partitioning two modes (objects and occasions) of three-way dissimilarity data based on the statistical modelling approach of fitting an expected clustering model, expressed in terms of dissimilarities and specified by a classification matrix, to the observed three-way two-mode data. Specifically, occasions are partitioned into homogeneous classes of dissimilarity matrices, and, within each class, a classification matrix, specifying a consensus partition of the objects, is identified. The parameters of the model are estimated in a least-squares fitting context and an efficient coordinate descent algorithm is given.

A Two-Level Fuzzy Classification between Different Profiles of the Gastronomic Lexicon

Sergio Bolasco – Pasquale Pavone

University "La Sapienza" of Rome, Italy, e-mail: sergio.bolasco@uniroma1.it
Scuola Superiore Sant'Anna di Pisa, Italy, e-mail: pasquale.pavone@sssup.it

This paper provides an analysis of the specialist lexicon used by gastronomic critics, based not only on food descriptions but also on menu entries. The information extraction process relies on the individuation of a local grammar, constituted by combinations of multiword expressions. This twofold categorisation allows one to capture the latest trends in the gastronomic offer, where the traditional dimensions of popular cookery intertwine with those of creative high-level cuisine. A fuzzy classification of restaurants is consequently preferred.

Nonparametric Inference via Permutation Tests for CUB Models

Stefano Bonnini – Luigi Salmaso – Francesca Solmi

University of Ferrara, Italy, e-mail: stefano.bonnini@unife.it
University of Padova, Italy, e-mail: {salmaso; solmi}@gest.unipd.it

In statistical surveys, respondents are often asked to express evaluations on several topics. The rating problem can be often faced in many fields. A new approach is represented by a class of mixture models with covariates (CUB models). Together with parametric inference, a permutation solution to test for covariates effects, when an univariate response is considered, has been discussed in [1], where the preference for a permutation test as compared to asymptotic ones when the sample size is moderate or even small has been justified through a simulation study. We propose an extension of this nonparametric inference to deal with the multivariate case. The method is applied to a real data set.

Asymmetric Multidimensional Scaling Models

Giuseppe Bove

University of Rome 3, Italy, e-mail: bove@uniroma3.it

Singular value decomposition (SVD) of skew-symmetric matrices was proposed by Gower (1977) to represent asymmetry of proximity data. Some authors considered the plane (*bimension* or *hedron*) determined by the first two singular vectors to detect orderings (*seriation*) for preference or dominance data. Following this approaches, in this paper some procedures of asymmetric multidimensional scaling useful for seriation are proposed focalizing on a model that is a particular case of *rank-2* SVD model. An application to Thurstone’s paired comparison data on the relative seriousness of crime is also presented.

Mutual Funds Ranking: An Application to the Italian Hedge Funds Industry

Riccardo Bramante – Alessandro Cipollini – Antonio Manzini

Catholic University, Mediobanca London, UBS AG

Due to the complexity and heterogeneity of hedge fund strategies, assessing their performance and risk is a challenging task. Reminiscent of the mutual fund industry, the literature has evolved in the direction of refining traditional measures (e.g. the Sharpe Ratio) or introducing new ones. This paper develops an approach, based on the Principal Component Analysis, to uncover the relevant information for performance measurement, quantify and combine it into a unique rank.

Pricing Strategy for Italian Wine

Eugenio Brentari – Rosella Levaggi

University of Brescia, Italy, e-mail: {brentari; levaggi}@eco.unibs.it

We use a unique dataset to estimate the hedonic price function for Italian red wine sold on the Italian market in the period 2006-2008. For each bottle considered, the dataset allows us to know several characteristic such as the price by retail channel (price in supermarkets and in wine shops), label characteristics, chemical analysis, sensorial characteristics and experts' evaluations. The objective of the analysis is to estimate price formation in the large distribution and in wine shop. In particular we want to explore the relative importance in each channel of characteristics that can be inferred from the label and other characteristics that require tasting (chemical and sensory characteristics). For wine sold using large distribution and wine shops we will also study which are the main determinants of the price difference. The results presented in the paper have been obtained using this database. We will shortly receive data for 2009-2010 and we are planning to estimate the model again.

A Statistical Analysis of Italian Wines in Large Distributions and Wine Shop

Eugenio Brentari – Paola Zuccolotto

University of Brescia, Italy, e-mail: {brentari; zuk} @eco.unibs.it

Italian wine market's supply side is very fragmented: the first 100 producers represent only about 30% of the total production. The wine is sold mainly through two channels: the large-scale retail trade and wine shops. The aim of this paper is to inspect the differences among wines sold in the two distribution channels, highlighting the most important features which make a wine suited for a channel or the other. Also, focusing on wines sold both in the large distribution and in wine shops, we analyze which wine features determine a higher difference of price in the two channels.

On Dynamic Hurdle Models for Longitudinal Zero-Inflated Count Data

Jan Bulla

Université de Caen, Caen Cedex, France, e-mail: bulla@math.unicaen.fr

The models presented deal with the analysis of a longitudinal dataset of buying behaviour, i.e. we record a set of costumer information over several time periods. Such data structure shows a number of characteristics which need to be described as the dependence of dependent variables on covariates, serial dependence and heterogeneity among the customers. In our empirical study, customers may be subject to several factors affecting their buying behaviour such as advertisements, or promotional offers. Over time, the influence of these factors on the costumer's buying behaviour may vary in an unknown (unobserved) way. In this setting, the hurdle-Poisson Hidden Markov Model (HMM), a dynamic extension of the classical hurdle model, can be applied. We analyze the performance of this model and show how to extend the latent process from a Markov to a semi-Markov chain utilizing a computationally convenient and easily deductible approach. As a by-

product of the model estimation algorithm, it is possible to classify customers into several relationship states.

Gender Gap Measurement: Methodological Perspective and Empyrical Research

Silvia Caligaris – Fulvia Mecatti

University of Milan-Bicocca, Italy, e-mail: silviacaligaris85@gmail.com, fulvia.mecatti@unimb.it

Gender equality is a pre-requisite for achieving sustainable and people-centred development. The production and dissemination of gender sensitive statistics are in fact essential to implement gender driven politics whose effects would improve the entire society for impacting upon both women and man. In this paper we focus on the Global Gender Gap Index among the several measures of gender equity provided by international agencies. With the aim of improving upon some evident drawbacks, the addition of a further social dimension and the use of a classical multivariate technique to produce the ordering is proposed. Potential advantages as well as significant changes in the final ranking will be empirically explored by designing and performing a simulation study. Official data restricted to Italy and a subset of European countries comparable with respect to cultural roots and economic and social development will be used.

Model-Based Clustering of Probability Density Functions

Daniela G. Calò – Angela Montanari

University of Bologna, Italy, e-mail: {danielagiovanna.calo; angela.montanari}@unibo.it

In this paper, we propose a method to group a set of probability density functions (pdfs) into homogeneous clusters, provided that the pdfs have to be estimated non parametrically from the data. Since elements belonging to the same cluster are naturally thought of as samples from the same probability model, the idea is to tackle the clustering problem by defining and estimating a proper mixture model on the space of probability densities. The issue of model building is challenging, because of the infinite-dimensionality and the Riemannian geometry of the domain space. By adopting a proper representation for the elements in the space, the task is accomplished using mixture models for hyper-spherical data. The proposed solution is illustrated on a real data set.

Slow Territories and Quality: Paths for Local Sustainable Development for Minor Destinations. The EDEN Project in Lombardy

Viviana Calzati

University of Perugia, Italy, e-mail: viviana.calzati@progetti.unipg.it

This paper aims to analyze the features of slow territories in a new vision of local sustainable development. Slow territories are low demographic density areas, mainly rural, where agriculture is still an important activity. These areas boast a notable, though not well known, historical and artistic heritage, and cultural activities are geared towards enhancing popular traditions, as well as local history and identity. Such territories, by way of a path mixing various forms of cultural, social, participation-oriented and environmental innovation, and with a shared, endogenous and commu-

nity vision, may get out of a marginal condition and place themselves as "distinct areas" geared towards promoting sustainable development. This paper analyzes the most significant results of the project called EDEN (European Destinations of Excellence) in Lombardy, in order to highlight that some areas of this region should not be identified as marginal territories, but, on the contrary, should be seen as 'territories' for the elite, capable of starting pathways towards economic growth and top-quality tourist development, aiming at environmental, cultural, and social sustainability.

Some Results on the Fuzzy Least Squares Regression Model with an Asymmetric Intercept

Francesco Campobasso – Annarita Fanizzi

University of Bari, Italy, e-mail: { fracampo;a.fanizzi }@dss.uniba.it

Fuzzy regression techniques can be used to fit fuzzy data into a regression model. Diamond treated the case of a simple model introducing a metrics into the space of triangular fuzzy numbers. In previous works we provided some theoretical results about the estimates of a multiple regression model; specifically we showed that only in the case of a fuzzy asymmetric intercept the components of the sum of squares of the dependent variable are reduced to the regression sum of squares and the residual one, like in the OLS estimation procedure. Such a decomposition allows us to introduce a stepwise procedure which simplifies, in terms of computational, the identification of the most significant independent variables in the model.

Multiple Structural-Change Model Analysis via Theoretical Regression Trees

Carmela Cappelli – Francesca Di Iorio – Pierpaolo D'Urso

University Federico II di Napoli, Italy, e-mail: {carcappe; fdiiorio}@unina.it;
University "La Sapienza" of Rome, Italy, e-mail: pierpaolo.durso@uniroma1.it

The analysis of structural-change models is nowadays a popular subject of research both in econometric and statistical literature. The most challenging task is to identify multiple breaks occurring at unknown dates. In case of multiple shifts in mean Cappelli *et al.* have proposed a method based on regression trees. In this paper we propose an extension of this method to address the problem of estimating the break dates and their number in the general framework of the linear model with multiple structural changes. We present simulation results pertaining to the behavior of the proposed approach.

A Strategy to Analyze Network Additionality for Territorial Innovation: The Case of Italian Technological Districts

Carlo Capuano – Domenico De Stefano – Alfredo Del Monte – Maria P. Vitale

University of Naples Federico II, Italy, e-mail: {carlo.capuano; delmonte} @unina.it
University of Trieste, Italy, e-mail: domenico.destefano@econ.units.it
University of Salerno, Italy, e-mail: mvitale@unisa.it

Empirical evidence from economic literature suggests that innovative activities based on extensive interactions between industry, university and local government have shown high developing performances. In many countries, institutional arrangements have been designed to start regional technological districts. In this paper we aim to analyze a nationwide systems of public grants to joint R&D projects realized by organizations involved in seven Italian Technological Districts. In this framework the main purpose is to explore how network structures influence network efficiency in terms of knowledge exchange and innovation diffusion by means of Social Network Analysis tools, graph comparison methods and suitable models.

Real-Time Behavioral Targeting of Banner Advertising

Fabrizio Caruso – Giovanni Giuffrida – Calogero Zarba

Neodata, Catania, Italy, e-mail: {fabrizio.caruso; calogero.zarba} @neodatagroup.com
University of Catania, Italy, e-mail: ggiuffrida@dmi.unict.it

We present a new algorithm for behavioural targeting of banner advertisements. We record different user's actions such as clicks, search queries and page views. We use the collected information to estimate in real time the probability of a click on a banner. Each click on a banner generates a profit. Our goal is to maximize the overall profit. We use a naive Bayesian model. We keep track of the click frequencies of the different banners under the additional information provided by the actions that each user has performed. Moreover we use a heuristics to avoid displaying the same banner to the same user too many times.

A Multinomial Model to Deal with Heterogeneity of Self-Employment in Eastern Europe

Rosalia Castellano – Gennaro Punzo

University of Naples "Parthenope", Italy, e-mail: {lia.castellano; gennaro.punzo} @uniparthenope.it

Aim of the work is to shed light on how some determinants, especially in the spheres of family background, differently affect the heterogeneous category of self-employment across a set of transition economies of Eastern Europe, where more or less restrictive policies and dissimilar liberalization processes have been adopted over time. At this end, three-stage structural multinomial logit models as discrete choice models are estimated on 2005 EU-SILC data. Country-specific differentials are sketched and it emerges how, in some countries, employment choice is devised in a dualist perspective.

Hybrid Pairwise Likelihood Analysis of Thurstone-Mosteller Models

Manuela Cattelan – Cristiano Varin

University of Padova, Italy, e-mail: manuela.cattelan@unipd.it
University “Ca’ Foscari”, Venice, Italy, e-mail: sammy@unive.it

Paired comparison data arise when objects are compared in couples. This type of data occurs in many applications. Traditional models developed for the analysis of paired comparison data assume independence among all observations, but this seems unrealistic because comparisons with a common object are naturally correlated. A model that introduces correlation between comparisons with at least a common object is discussed. The likelihood function of the proposed model involves the approximation of a high dimensional integral. This problem is overcome by means of a hybrid pairwise likelihood method.

Feature Selection Stability in High-Dimensional Heterogeneous Data

David Causeur

Rennes cedex, France, e-mail: david.causeur@agrocampus-ouest.fr

The high dimension and large heterogeneity of data generated by highthroughput technologies has markedly renewed the statistical methodology for multiple testing and feature selection in regression or classification issues. Some recent papers (Leek and Storey 2007 and 2008; Friguet *et al.* 2009) have focused on the negative impact of data heterogeneity on the consistency of the ranking which results from multiple testing procedures. Because componentwise strategies such as multiple testing do not account for the system-wide interaction structure, model selection is often used to identify relevant subsets of components. The present paper aims at showing that data heterogeneity also affects the stability of feature selection. It is first shown that selected subsets using well-known procedures such as forward stepwise selection or LASSO are subject to a high variability. As suggested in Friguet *et al.* (2009), a supervised factor model is proposed to identify a low-dimensional linear kernel which captures data dependence and new strategies for model selection are deduced.

Statistical Models to Measure Academic Reputation

Paola Cerchiello

University of Pavia, Italy, e-mail: paola.cerchiello@unipv.it

The aim of this paper is to present a new proposal for the classification of academic institutions in terms of quality of teaching. Our methodological proposal focuses on assessing University performances on the basis of the perceived quality by using ranking indexes. We propose to summarize students’ opinion data using two new non parametric indexes able to exploit efficiently the ordinal nature of the analyzed variables. Moreover, we present the most recent advances in terms of gathering and analyzing the perceived quality of academic teaching from students’ point of view. In particular, we show how the web survey methods can improve the quality and robustness of collected data. Empirical evidence is finally given on the basis of real data from the University of Pavia.

Poisson M-Quantile Models on Disease Mapping

Ray Chambers – Emanuela Dreassi – Nicola Salvati

University of Wollongong, Australia, e-mail: ray@uow.edu.au
University of Florence, Italy, e-mail: dreassi@ds.unifi.it;
Pisa University, Italy, e-mail: salvati@ec.unipi.it

A new approach to ecological analysis on disease mapping is introduced: a semi-parametric approach based on M-quantile models. We define a Poisson M-quantile spatially structured model. The proposed approach is easily made robust against outlying data values for covariates. Robust ecological disease mapping is desirable since covariates at area level usually present measure-type error. We easily consider a spatial structure in the model introducing suitable weights at the estimation step, in order to extend the M-quantile approach to account for spatial correlation between areas. Differences between M-quantile and usual random effects models are discussed and the alternative approaches are compared using a real example and a simulation experiment.

Multidimensional Clustering and Registration of Seismic Waveform Data

Marcello Chiodi – Giada Adelfio – Antonino D'Alessandro – D. Luzio

University of Palermo, Italy, e-mail: {adelfio; chiodi} @unipa.it,
INGV – National Institute of Geophysics and Volcanology Geophysical Observatory of
Gibilmanna OBS Lab – CNT, Italy, University of Palermo, Italy

In order to find similar features between multidimensional curves, we consider the application of a procedure that provides a simultaneous clustering and alignment of such functions. In particular we look for clusters of multivariate seismic waveforms based on EM-type procedure and functional data analysis tools. Application to 4-dimensional seismic waves recorded in Southern Tyrrhenian gave very encouraging results.

SMC^2: A Sequential Monte Carlo Algorithm with Particle Markov Chain Monte Carlo Updates

Nicolas Chopin – Pierre Jacob – Omiros Papaspiliopoulos

ENSAE-CREST, France, e-mail: nicolas.chopin@ensae.fr
CREST & Universite Paris Dauphine, France, e-mail: pierre.jacob@ensae.fr
Universitat Pompeu Fabra, e-mail: omiros.papaspiliopoulos@upf.edu

We consider the generic problem of performing sequential Bayesian inference in a state-space model with observation process *y*, state process *x* and fixed parameter θ. An idealized approach would be to apply the iterated batch importance sampling (IBIS) algorithm of Chopin (2002). This is a sequential Monte Carlo algorithm in the θ-dimension, that samples values of θ, reweights iteratively these values using the likelihood increments $p(y_t|y_{\{1:t-1\}},\theta)$, and rejuvenates the θ-particles through a resampling step and a MCMC update step. In state-space models these likelihood increments are intractable in most cases, but they may be unbiasedly estimated by a particle filter in the *x*-dimension, for any fixed θ. This motivates the SMC^2 algorithm proposed in this article: a sequential Monte Carlo algorithm, defined in the θ dimension, which propagates and resamples many particle filters in the x-dimension. The filters in the x dimension are an example of the random weight particle filter as in Fearnhead *et al.* (2010). On the other hand, the particle Markov chain Monte Carlo (PMCMC) framework developed in Andrieu *et al.* (2010) allows us to design appropriate MCMC rejuvenation steps. Thus, the θ-particles target the correct posterior distribution at each iteration τ, despite the intractability of the likelihood increments. We explore the applicability of our algorithm in both sequential and non-sequential applications and consider various degrees of freedom, as for example increasing dynamically the number of ξ-particles. We contrast our approach to various competing methods, both conceptually and empirically through a detailed simulation study on particularly challenging examples.

Measuring Uncertainty in Statistical Matching

Pier Luigi Conti – Daniela Marella

University "La Sapienza" of Rome, Italy, e-mail: pierluigi.conti@uniroma1.it
University Roma Tre, Italy, e-mail: dmarella@uniroma3.it

An important feature of statistical matching is that the underlying joint distribution of the variables of interest is not identifiable. This produces a form on "uncertainty" on the statistical model. A measure to evaluate such an uncertainty is proposed. The effect of prior information in the form of constraints is exploited. Finally, the estimation of the proposed measure of uncertainty is studied.

International University Rankings: A Critical Approach

Andrea Costa

Bocconi University, Italy, e-mail: andrea.costa@unibocconi.it

International rankings of universities have become popular in the last ten years or so. They can provide some useful advice to prospective students, especially those wishing to study abroad, and other parties. However, none of them is free from methodological flaws which may make them unfit for the purpose. Indeed, they can also have perverse system effects as universities may be tempted to privilege short-term decisions which can make them rise in established rankings to the detriment of long-term strategies.

Bibliometrics: Lessons Learned

Antonio Costantini – Massimo Franceschet

University of Udine, Italy, e-mail: {antonio.costantini; massimo.franceschet} @uniud.it

In the last few years we have been continuously involved in bibliometrics: we have actively participated in the first Italian research evaluation exercise, we have read the relevant literature and successfully written our own papers, and we have discussed and contrasted our opinions on the topic with those of other scholars from our and other fields. Substantially, we have learned some lessons about bibliometrics. We tried to condense some of these learned lessons in this short paper.

Mixed Effects Models in Neurolinguistics: Considering a Shift from Univariate to Bivariate Applications

Franca Crippa – Marco Marelli

University Milano-Bicocca, Italy, e-mail: {franca.crippa; m.marelli1}@unimib

Relatively recent developments in computational statistics have allowed the inclusion of subjects and items as crossed, reciprocally nested random effects in mixed models. Preserving this data structure in psychometrical results poses great advantages over traditional investigations, based on quasi-F tests, by-subjects analyses, combined by-subjects and by-items analyses. A major flaw of the traditional approach for crossed classifications consists of the use of inappropriate techniques, for instance treating the count (or the percentage) of the total number of correct responses for each subject as a continuous response in a linear model, at most with square-root (or arcsin-square-root) transformations. Moreover, issues as the statistical power, heteroscedasticity and non-spherical error variance for either participants or items are not adequately addressed. Even in the mixed modelling frame, though, the two typical responses of psychometrics experiments, accuracy and latency, are fitted in two independent equations, whereas evidence of their interrelation arises in various cases. The prospect of fitting a bivariate hierarchical model to the case of completely cross-classified data in the specific field of neurolinguistic studies is pondered.

A General to Specific Approach for Selecting the Best Business Cycle Indicators

Gianluca Cubadda – Barbara Guardabascio – Alain Hecq

University of Rome-Tor Vergata, Italy, e-mail: {gianluca.cubadda; barbara.guardabascio}@uniroma2.it
Maastricht University, the Netherlands, e-mail: a.hecq@maastrichtuniversity.nl

Combining economic time series in order to obtain an indicator for business cycle analysis is an important issue for policy makers. In this paper we propose tools to select the relevant business cycle indicators in a medium N framework, where N is the number of series. An example is provided by our empirical application, in which we study jointly the short-run co-movements of 24 European countries for the period 1997Q1 to 2010Q3. Given the extremely poor data framework, we are not able to use multivariate regression models. However, we show, under not too restrictive conditions, that parsimonious single-equation models can be used to split a set of N series in three groups of countries that share: i) a synchronous common cycle, ii) a non synchronous common cycle iii) an idiosyncratic cycles.

Wavelet Density Estimation for Weighted Data

Luisa Cutillo – Italia De Feis – Christina Nikolaidou – Theofanis Sapatinas

University Parthenope of Naples, Italy, e-mail: luisa.cutillo@uniparthenope.it
CNR, IAC, Italy, e-mail: i.defeis@iac.cnr.it

We consider the estimation of a density function on the basis of a random sample from a weighted distribution. We propose linear and nonlinear wavelet density estimators, and provide their asymptotic formulae for mean integrated square error. In particular, we derive an analogue of the asymp-

totic formula of the mean integrated square error in the context of kernel density estimators for weighted data, admitting an expansion with distinct squared bias and variance components. Comparisons with two other methods proposed in the literature are also given.

Classifying Tourism Destinations: An Application of Network Analysis

Rosario D'Agata – Venera Tomaselli

University of Catania, Italy, e-mail: {rodagata; tomavene}@unict.it

Tourism is basically a spatial phenomenon, which implies moving consumption within space. Starting from the assumption that the destinations are nodes of a network, we were able to reconstruct a spatial grid where each locality showed different grades and types of centrality. The analysis, paying particular attention to the spatial dimension, showed clusters of locations. Employing traditional network analysis measures, the paper attempts to classify destinations considering the routes of a self-organized tourists sample that visited more than one site in Sicily.

Default Probability Estimation: Bayesian Pair Copula Model

Luciana Dalla Valle – Maria E. De Giuli – Claudio Manelli – Claudia Tarantola

University of Milan, Milan, Italy, e-mail: luciana.dallavalle@unimi.it
University of Pavia, Italy, e-mail: {elena.degiuli; claudia.tarantola} @unipv.it
FMR Consulting SpA, Voghera (PV), Italy, e-mail: claudio.manelli@fmrcons.com

We present a novel Bayesian methodology for default probability estimation based on multivariate contingent claim analysis and pair copula theory. In order to compute the default probability of a firm, we use balance sheet data as a proxy of the equity value. A pair copula approach is applied to obtain the firm pricing function, and Monte Carlo simulations are then used to calculate the distribution of the default probability. The methodology will be applied to real data analysis.

Representativity Indicators: New Proposals to Define Response Propensities

Luciana Dalla Valle – Giovanna Nicolini

University of Milan, Italy, e-mail: {luciana.dallavalle; giovanna.nicolini}@unimi.it

Traditionally the percentage of unit non responses was considered as the most important quality indicator for a survey. However, not always the higher the response rate the higher the quality of the survey and the accuracy of the estimates. Quality and accuracy are guaranteed by small differences between respondents and non respondents. In this case the observed sample is representative of the planned sample. Therefore, we consider indexes introduced by Shlomo *et al.* in 2008, called R-indicators, which are able to measure the contrast between respondents and non respondents using a set of auxiliary variables. However, the methodology normally used to calculate R-indicators for big samples is not feasible for subsamples like the domains. Therefore, we propose two alternatives for calculating R-indicators and we assess their performance compared to the traditional method in a stratified sample when some strata samples are of size n=50.

Weighting the Spearman's Rank Correlation Index

Livia Dancelli – Marica Manisera – Marika Vezzoli

University of Brescia, Italy, e-mail: {dancelli; manisera; vezzoli} @eco.unibs.it

Weighted Rank Correlation indices are useful for measuring the agreement of two rankings when the top ranks are considered more important than the lower ones. In this paper, we investigate the behaviour of i) five existing indices that introduce suitable weights in the simplified formula of the Spearman's ρ and ii) other five indices we derived using the same weights in the Pearson's product-moment correlation index between ranks. Results suggest to avoid linear weights.

Exploring the Sensitivity of Scientific Research Activity Evaluation Models by Multivariate Data Analysis

Cristina Davino – Rosaria Romano

University of Macerata, Italy, e-mail: cdavino@unimc.it, romano.rosaria@gmail.com

The aim of the paper is to introduce an innovative approach based both on confirmative and exploratory statistical methods aiming to assess the sensitivity of research activity evaluation models. The paper centers on one single component of the funding model, the scientific research activity evaluation, for two reasons: it represents the primary component on which the Italian universities are called to invest in the future; the proposed approach has revealed for this component the highest sensitivity to the governmental model as compared to the others. The proposal of the present contribution is to present a new approach to the CI Sensitivity Analysis based on a mixture of explorative and confirmative analysis aiming to investigate the impact of the different subjective choices on the CI variability and the related individual differences among the statistical units as well.

Robust Analysis of Bibliometric Data

Francesca De Battisti – Silvia Salini

University of Milan, Italy, e-mail: { francesca.debattisti; silvia.salini} @unimi.it

In our paper presented in the last edition of CLADAG we tried to depict the research profile of Italian statisticians which can be deduced from multiple bibliometric databases. We highlighted the need for multiple sources in order to have a truer picture of statistical research in Italy and investigated the best ways to combine data to obtain a reliable classification or overall indicators for research productivity, taking into account all possible metrics. The weakness of this exercise lies in the type of data. The resulting matrix containing the set of metrics from a variety of databases for each author is a sparse matrix (i.e. many zero values are present). Furthermore, the variables are leptokurtic and characterized by positive asymmetry. In order to apply the classical techniques of multivariate analysis, the data must be previously transformed or, alternatively, robust analysis techniques have to be used. In this paper we will focus on this type of

bibliometric data, describing their main characteristics and problems. In addition, a robust approach to the analysis of these data will be presented.

Some Notes on Bayesian Inference for CUB Model

Laura Deldossi – Roberta Paroli

Catholic University, Milan, Italy, e-mail: {laura.deldossi; roberta.paroli} @unicatt.it

In this paper we consider a special finite mixture model for ordinal data expressing the preferences of raters with regards to items or services, named CUB (Covariate Uniform Binomial), recently introduced in statistical literature. Our aim is to develop a variable subset selection procedure for this model, to identify the best covariates using a Bayesian approach. Bayesian methods for variable selection and model choice have become increasingly popular in recent years, due to advances in Markov chain Monte Carlo (MCMC) computational algorithms. Several methods have been proposed in the case of linear and generalized linear models. In this paper we adapt to the CUB model the algorithm proposed by Kuo and Mallick. The performance of the new algorithm is evaluated by a simulation study.

Improving ARMA-GARCH Forecasting via Partial Exchangeability

Petros Dellaportas – Leonardo Bottolo

AUEB, Athens, Greece, e-mail: petros@aueb.gr, Imperial College, London, United Kingdom

We exploit the partial exchangeability structure of the parameters of many R-GARCH models to borrow strength for univariate variance forecasting. e adopt a challenging reversible jump MCMC scheme which models the parameters as a finite mixture of normals with unknown number of components. We generalise existing modelling structures by assuming that the component means follow a multivariate normal density. We discuss in detail the careful choice of prior parameters, the construction of the reversible jump algorithm that requires jumps between covariance matrices of different dimensions and the use of an adaptive regional MCMC algorithm. We test our methodology with stock returns from an S&P100 dataset and we find that our forecasts are more robust and offer better forecasting power when compared with those of standard AR-GARCH models.

How do you name your occupation? A Text Mining Application on the Language Used by Workers and by the Standard Occupational Classification

Francesca della Ratta Rinaldi – Francesca Gallo – Barbara Lorè

Istat, Italy, e-mail: {dellarat; gallo; lore} @istat.it

As part of the preparation of the new Classification of occupations (CP2011), particular attention has been devoted to the phase of updating occupational titles. The Labour Force Survey (LFS) has been an interesting source of information to enrich the classification dictionary of all those occupations that have emerged in the recent past. This paper shows the results of a text

mining analysis conducted on answers to the question 'How do you name your occupation?' provided by respondents of the last two years LFS (2009-2010). The analysis aims to assess the overlap between the language spoken by interviewees and the one used by the classification to name occupations. Moreover, the analysis and the categorization of the non-overlapping words throw new light to identify the need for special measures of improvement.

Design-Based non Response Treatment in Forest Surveys

Flora De Natale – Lorenzo Fattorini – Sara Franceschi – Patrizia Gasparini – Daniela Maffei

CRA, Italy, e-mail: { flora.denatale; patrizia.gasparini}@entecra.it
University of Siena, Italy, e-mail: {fattorini; franceschi2}@unisi.it
University of Florence, Italy, e-mail: maffei@ds.unifi.it

The objective of this work is to obtain reduced bias and to estimate the variance of biophysical attribute estimators achieved in forest inventories in presence of item non response due to the unrecorded values of points located in roughly or difficult terrain. In most situations, these points cannot be reached by foresters or, even if reached, the recording activities cannot be performed. Since the points can be reached or not there is no random mechanism generating non response. Thus non response can be handled in a complete design-based framework. A three-phase sampling strategy for forest survey is implemented and the properties of the resulting estimator are investigated by means of a simulation study. Finally the application of non response treatment to the Italian National Forest Inventory for the data regarding the administrative district of Trentino (north Italy) is considered.

Model Based Specifying of Performance Indexes

Giulio D'Epifanio

University of Study of Perugia, Italy, e-mail: ggiulio@unipg.it

The question of interest concerns benchmarking the performance of agents (in particular, individuals) so that assessments, provided by proper indexes, are fully standardized on a given reference evaluative framework F. Framework F is designed by the decision-maker(DM) taking into account both "artificial intentions" (about goals, requirements and constraints) and "prior knowledge" (about processes and conditions which govern them). Assessments are provided by indexes which are referred to well-identified standards, in recognizable and acceptable manner. Due to the difficulty of specifying explicit standards, an indirect approach is necessary which uses some method for making intrinsic standards, provided by a set of requirements specifications on F, explicit. In section 2, we recall standardized indexes on graduated performance scales, conditional on characteristics, normed on a reference standard population. In section 3, an interpretative reference model on latent performance scales is specified, within F, using a probabilistic (prior-feedback like, see Casella and Robert (2002) pseudo-Bayesian setup. We delineate the method of CFP (see, D'Epifanio 1996 and 2004) for automatic eliciting of standardized indexes from reference data.

Spectral Embedding Procedure for Social Network Comparison

Domenico De Stefano

University of Trieste, Italy, e-mail: domenico.destefano@econ.units.it

A key issue in social network analysis is related to the comparison between several observed networks on n actors. To this end, a special graph embedding procedure derived from the spectral properties of the networks is proposed. The procedure consists of two steps: i) define an appropriate metric among the observed networks based on the properties of the eigenvalues/eigenvectors of the so-called Laplacian matrix; ii) compare the corresponding distance matrices among the n actors within each network. The purpose is twofold: on the one hand we aim to define a matrix of actor distances and consequently to use the actors embedding for network comparison; on the other hand, we will be also able to measure the distances among the global network structures, considering them as points in a multivariate space. We will show applications in both exploratory social network analysis and network statistical modelling.

Career Paths in a Gender Perspective: An Application of Multilevel IRT Models

Tonio Di Battista – Simone Di Zio – Cristiana Ceccatelli – Raffaella Marianacci

University of Chieti, Italy, e-mail: {dibattista; s.dizio; c.ceccatelli ; r.marianacci }@.unich.it

This paper shows the results of an application of multilevel IRT models on the data of a survey conducted by the Italian National Institute of Statistics about the criticality of career paths from a gender perspective. Data, referred to 2007, concern a sample of about 10,000 individuals between 18 and 64 years old. The dataset contains groups of items from which we are able, throughout the multilevel IRT models, to extract two particular latent dimensions. In particular, the aim of the present paper is to evaluate/measure the independence of women in the familiar context and the positive attitude of people towards their personal and professional future situation. Moreover, given the social-economic differences which characterize the Italian peninsula, it is interesting to check for geographical diversities, by grouping the individuals in macro-regions. In this hierarchical data structures framework, we fit a series of mixed-effects models, in order to choose the best model to interpret the dataset.

Principal Component Analysis of Metabins for Complex Data Mining

Edwin Diday

Paris Dauphine University, France, e-mail: diday@ceremade.dauphine.fr

In recent years, the Symbolic Data Analysis (i.e. SDA) framework, where the units are categories, classes or concepts described by intervals, distributions, sets of categories and the like, becomes a challenging task since many application fields generate massive amounts of “Complex data” which come from different sources, or live in high dimensional spaces. In practice, “Complex Data” refers to complex objects like images, video, audio or text documents. “Symbolic Data Analysis” is a new paradigm in which theory, tools and practice have shown its ability to extract new knowledge from such complex data, that are difficult to store and to analyze with traditional techniques. In this paper, we propose a strategy for extending standard Principal Component

Analysis (PCA) to such complex data. This leads to variables which values are “bar chart” (i.e., a set of categories called bins with their relative frequency). Metabins are ordered sets of such bins, which mix together bins of the different bar charts and enhance interpretability. Some open questions are advocated and an example of PCA representing trajectories of metabins is given.

Estimating Variable Association in Contingency Tables through Probabilistic Expert Systems when Samples are Drawn according to Stratified Sampling Designs

Roberto Di Manno – Mauro Scanu – Paola Vicard

Ministero dello Sviluppo Economico, Rome, Italy, e-mail: roberto.dimanno@tesoro.it
Istat, Italy, e-mail: scanu@istat.it
University Roma Tre, Italy, e-mail: vicard@uniroma3.it

This article focuses on the learning algorithms of a PES structure, given a sample of observations from a finite population, drawn according to a stratified sample design. The structural learning algorithms can be of two typologies: the score+search and the constraint based algorithms. Here we focus on the score+search case, for which the most likely structure, given the observed data, will be identified optimizing an objective function (typically a penalized likelihood), by means of optimization methods that, in this case, will be given by the greedy search and the genetic algorithms ones. Structural learning will be tackled considering the necessary networks for one class of estimators (named E-PES estimator) built over a PES that includes the variables of interest and the design variable.

A Two Step Non Parametric Procedure for Statistical Matching

Marcello D’Orazio

Istat, Italy, e-mail: madorazi@istat.it

This paper introduces two step nonparametric procedures for statistical matching, i.e. for matching two datasets A and B in order to produce a synthetic data set. These procedures resemble the mixed approach to statistical matching based on the fitting of linear regression models. Instead of considering linear models, regression trees are fitted. The features of these two steps nonparametric statistical matching procedure are assed in a simulation study. The results obtained are encouraging.

Beanplot Data Analysis in a Temporal Framework

Carlo Drago – Carlo Lauro – Germana Scepi

University of Naples “Federico II”, Italy, e-mail: {carlo.drago; clauro; germana.scepi}@unina.it

In this paper, we propose a new approach for the modelling, clustering and forecasting financial time series. The aim is modeling the variability both intra period and related to more temporal intervals. In particular, this approach is based on a peculiar density plot, called beanplot. These types of new aggregated time series can be fruitfully used when there is an overwhelming number of observations, for example in High Frequency financial data.

A New Financial Stress Index Framework Based on RST-SVR-CBR

Amira Dridi – Mohamed El Ghourabi – Mohamed Limam

ESSEC, University of Tunis, Tunis, e-mail: amiradridi@laposte.net
ISGT, University of Tunis, Tunis, e-mail: mohamed.elghourabi @gmail.com, mohamed.limam@isg.rnu.tn

Financial stress index (FSI) is a key technique for quantifying financial vulnerabilities. It is an important risk management tool. The aim of this paper is to propose a new framework based on a combined classifier model where we integrate Rough Set Theory (RST), Support Vector Regression (SVR) and Case Based Reasoning (CBR) clustering in order to identify stress periods. First, RST method is applied for data pre-processing, outputs are used as input data for the FSI-SVR computation. Then, we proceed with the CBR clustering process by choosing a value at risk estimated with extreme value theory as a centroid in order to identify two clusters namely fair stressed cluster and high stressed cluster. Empirical analysis is performed using monthly FSI for Indonesian financial market from January 1995 to March 2007. Our FSI framework is able to detect different stress levels including Asian crisis episodes, as identified by IMF report.

Mutual Information Decomposition for Vines Models

Filippo Domma – Francesca Condino

University of Calabria, Italy, e-mail: f.domma@unical.it
Institute of Neurological Science, CNR, Cosenza, Italy, e-mail: f.condino@isn.cnr.it

In order to specify multivariate distributions Bedford and Cooke proposed a new graphical approach called *vines*. This model specification, graphically represented by a set of trees, is based on a factorization of multivariate density and on the copula approach to describe dependence structure. Such an approach leads to specify the joint density using a cascade of blocks identified by pair-copulae. Among all possible specifications, we consider the so-called *D-vines* and *C-vines* models. Mutual Information (MI) is obtained for these two models. We prove that MI can be decomposed into the sum of conditional and unconditional pair-copula entropies. This result allows us to quantify the amount of the total divergence between the joint model and the model under the hypothesis of complete independence explained by the single edge of the *vine* tree.

Recent Advances on Functional Additive Regression

Frédéric Ferraty – Aldo Goia – Ernesto Salinelli – Philippe Vieu

Université Paul Sabatier, Toulouse Cedex, France, e-mail: {ferraty; vieu}@cict.fr
University del Piemonte Orientale "A. Avogadro", Novara, Italy, e-mail: {aldo.goia; ernesto.salinelli}@eco.unipmn.it

We introduce a flexible approach to approximate the regression function in the case of a functional predictor and a scalar response. Following the Projection Pursuit Regression principle, we derive an additive decomposition which exploits the most interesting projections of the prediction variable to explain the response. The goodness of our procedure is illustrated from theoretical and practical points of view.

A Note on Additively Decomposable Inequalities for Risk Measurement

Silvia Figini

University of Pavia, Italy, e-mail: silvia.figini@unipv.it

In this paper our objective is to construct a class of decomposable additive measures for risk assessment and integration. To reach this objective, our approach is based on the extension of non parametric measures, typically used to measure inequality, which are decomposable. In particular, starting from the generalised entropy we have derived the Theil, the Herfindal Hirschman and the Bourguignon indexes. Using such inequality measures we are able to underline the risk relevance and the relative magnitude with respect to both qualitative and quantitative risk variables, thereby allowing integrated risk measures. In order to show how our proposal works, we give empirical evidences on the basis of a data set that concerns risk estimation.

Supervised Classification of Facial Expression

S. Fontanella – Caterina Fusilli – Luigi Ippoliti

University "G. D'Annunzio" Chieti-Pescara, Italy, e-mail: {s.fontanella; c.fusilli; ippoliti}@unich.it

Over the last decade, the statistical analysis of facial expressions has become an active research topic that finds potential applications in many areas. As expression plays remarkable social interaction, the development of a system that accomplishes the task of automatic classification is challenging. In this work, we thus consider the problem of supervised classification of facial expressions through shape variables represented by log-transformed Euclidean distances computed among a set of anatomical landmarks.

Multilevel Functional Data Analysis of Mandibular Condyles

Lara Fontanella – Luigi Ippoliti – Pasquale Valentini – Felice Festa

University "G. D'Annunzio" Chieti-Pescara, Italy, e-mail: {s.fontanella; ippoliti; pvalent; f.festa}@unich.it

A specific anatomic component of the temporomandibular joint is the mandibular condyle which articulates with the temporal bone in the mandibular fossa. During the growth, and in response to orthodontic treatment, the condyle develops in many directions relative to individual variations. Deviations in the growth, if not detected early, may lead to bone destruction and osseous deformation of the mandibular condyle resulting in growth disturbances and dysmorphic facial features. In this paper, the mandibular condyles are summarized by a continuous outline so that the information about the object comes from the boundary. A functional data analysis is thus proposed in order to detect abnormalities of their shape and size.

Structural Latent Class Models and Causal Inference: An Application to Education Transmission

Antonio Forcina – Salvatore Modica

University of Perugia, Italy, e-mail: forcina@stat.unipg.it, University of Palermo, Italy, e-mail: modica@unipa.it

Non linear structural equation models with latent variables is a conceptual tool which can handle both observed and latent confounders; the causal pathways that it implies can be represented by a causal DAG (Directed acyclic graph). We describe an approach for fitting identifiable structural latent class models by the EM algorithm; once a model has been fitted, it can be used to answer questions concerning causal effects of interest. The method is applied to measure the direct effect of parent's education on that of their children, an issue rather popular in Econometrics.

Prediction in a Multidimensional Setting

Giovanni Fonseca – Federica Giummolè – Paolo Vidoni

University of Udine, Italy, e-mail: {giovanni.fonseca; paolo.vidoni}@uniud.it
University "Ca' Foscari", Venice, Italy, e-mail: giummole@unive.it

This paper concerns the problem of prediction in a multidimensional setting. Generalizing a result presented in Ueki and Fueda (2007), we propose a method for correcting estimative predictive regions to reduce their coverage error to third order accuracy. The improved prediction regions are easy to calculate using a suitable bootstrap procedure. Furthermore, the associated predictive distribution function is explicitly derived. Finally, an example concerning the exponential distribution shows the good performance of the proposed method.

Model Based Clustering using Mixtures of Asymmetric Laplace Distributions

Brian Franczak – Ryan P. Browne – Paul D. McNicholas

University of Guelph, Canada, e-mail: {bfrancza; rbrowne; pmichno}@uoguelph.ca

A model-based clustering scheme using mixtures of asymmetric Laplace distributions is introduced. Maximum likelihood estimates for the parameters are calculated using an expectation-maximization algorithm, where Aitken's acceleration is used to determine convergence. The BIC is used to select the number of mixture components and the parametric structure. Our approach is demonstrated using both simulated and real data; its performance is compared to other model-based clustering methods.

Skew-Symmetric Distributions and Model Choice

Rosa Fucci – Anna Clara Monti

University of Sannio, Italy, e-mail: {rosafucci; acmonti}@gmail.com

Skew-symmetric distributions provide flexible models suitable to fit the distribution of data affected by departures from normality, such as skewness and/or heavy tails. However, when skew-

symmetric models happen to be over-parameterized with respect to the actual distribution, because either only one or none of the deviations occur, their adoption can lead to remarkable losses of efficiency in the estimation of the parameters. Consequently the paper proposes a strategy to identify whether the actual distribution of the data is a sub-model of a skew-symmetric distribution.

Time Series Turning Points Discrimination

Giampaolo Gabbi

*University of Siena, Banking and Finance SDA Bocconi, Italy, e-mail: giampaolo.gabbi@sdabocconi.*it

Financial Crises are a typical stressed case where time series experience turning points. The purpose to detect them often appears a challenging issue both for financial models and for quantitative methods. We suggest a model capturing agents' decisions who tend to act simultaneously during crashes. We apply a multivariate model based on time varying indicators to evaluate periods with high positive correlation between volatility and correlations and to extract probabilities of a coming market crash.

Flexible Shapes Clustering

Stefano A. Gattone

University of Rome-Tor Vergata, Italy, e-mail: gattone@economia.uniroma2.it

Shape analysis deals with all the information which is invariant under translation, rotation and scaling of a given object under study. Standard statistical methods have to be properly adapted in order to be used in shape analysis. We propose a new method for clustering shapes based on the conjoint use of Relative warp analysis and Reduced K-means. We restrict ourselves to landmark-based analysis in two dimensions, where shapes are represented by points (landmarks) located on the object contours.

A Different Perspective on Clustering Time Series

Margherita Gerolimetto – Isabella Procidano

University "Ca' Foscari", Venice, Italy, e-mail: {margherita.gerolimetto; iabella}@unive.it

In this paper we intend to shed further light on time series clustering. Firstly, we aim at clarifying, via Monte Carlo simulations, to which extent the choice of the measure of dissimilarity can affect the results of time series cluster analysis. Then we move to a different point of view and tackle the issue of classifying time series using the Self Organizing Maps (Kohonen 2001), typically employed in pattern recognition for cross-sectional data.

Dynamic Programming versus Graph Cut Algorithms for Fitting Non-Parametric Models to Image Data

Chris A. Glasbey

Biomathematics & Statistics Scotland, Edinburgh, Scotland, e-mail: chris@bioss.ac.uk

Image restoration, segmentation and template matching are generic problems in image processing that can often be formulated as non-parametric model fitting: maximising a penalised likelihood or Bayesian posterior probability for an I-dimensional array of B-dimensional vectors. The global optimum can be found by dynamic programming provided I=1, with no restrictions on B, whereas graph cut algorithms require B =1 and a convex smoothness penalty, but place no restrictions on I. In this talk I compare conditions and results for the two algorithms, illustrated by restoration of a synthetic aperture radar (SAR) image.

Classification Using Kernel Density Estimation: An Application to the ISTAT's Labour Force Survey

Gianluca Giuliani – Rita Lima – Rita Ranaldi

ISTAT, Italy, e-mail: {gigiulia; lima; ranaldi }@istat.it

The primary aim of statistical process performance monitoring is to identify deviations from normal situation within the data production process. In the ISTAT Labour Force Survey, the basis of monitoring process is an indicators system that allows a continuous quality check at local level (different geographical areas, regions and sample municipalities). Although this quality system is in line with the international standards for official surveys, no standard definitions of confidence bounds to detect the onset of process deviations are proposed. This paper proposes an approach for classification of the performance of CAPI interviewers based on the Kernel approach.

Testing Unidimensionality and Clustering Items: An Application to the INVALSI Students'Assessment Data

Michela Gnaldi – Francesco Bartolucci – Silvia Bacci

University of Perugia, Italy, e-mail: {michela.gnaldi; bart; silvia.bacci }@stat.unipg.it

We aim at studying if the assumption of unidimensionality is met for the data collected on middle school students by the National Institute for the Evaluation of the Education System (INVALSI). The applied methodology relies on a class of multidimensional latent class Item Response Theory models based on: (i) a two-parameter logistic parametrisation for the conditional probability of a correct response, (ii) latent traits represented through a random vector with a discrete distribution, and (iii) the inclusion of differential item functioning (DIF) effects due to students' gender and geographical status. On the basis of this model, a hierarchical clustering algorithm is also proposed for dividing items into unidimensional groups referred to different abilities. The resulting classification of the items is represented by a dendrogram.

Robustness Properties of the TCLUST Methodology through the Influence Function

Alfonso Gordaliza – Christel Ruwet – Luis A. García-Escuder – Agustín Mayo-Iscar

University of Valladolid, Spain, e-mail: alfonsog@eio.uva.es, University of Liege, Belgium, e-mail: cruwet@ulg.ac.be

The TCLUST procedure, introduced by García-Escudero *et al.* (2008), is aimed to perform robust clustering to find groups with different scatter structures and proportions. Moreover, as most real data sets contain outliers and background noise, TCLUST is designed to allow for trimming off a fixed proportion of data which will not be assigned to the groups. This subset of non assigned data points is self-determined by the data following the principle of "impartial trimming". TCLUST methodology is supposed to have good robustness properties, inherited somehow from trimmed k-means methodology with which it shares the same principles of impartial trimming. Although some examples have been presented and discussed in García-Escudero *et al.* (2008) and García-Escudero *et al.* (2010) to show the robustness of the TCLUST methodology, until now no comprehensive formal study of its robustness properties has been carried out. In this work, we study the robustness properties of the TCLUST procedure by means of the influence function. Some advances are also given with respect to the dissolution point and isolation robustness. It turns out that the robustness behaviour of TCLUST procedure is close to that of the trimmed k-means.

Visual Detection of Communities in Social Networks

Nicolas Greffard – Fabien Picarougne

Team KOD – Laboratoire d'Informatique de Nantes-Atlantique, France

Community detection is a task of great importance in the analysis of social networks. Most often, communities are first identified with a clustering approach for which various algorithms have been proposed. But, the community detection suffers from a major problem: in many real life situations, communities do not form an non ambiguous partition of the graph and several overlappings are present. In this communication, we investigate a 3D stereoscopic visual network representation which highlights communities and links between them. And, we show that for some network families this representation overcomes the detection performances of more classical 2D or 3D perspective layouts.

Latent Growth Models with Multiple Indicators: A Longitudinal Analysis of Student Ratings

Leonardo Grilli – Roberta Varriale

University of Florence, Italy, e-mail: {grilli; roberta.varriale}@ds.unifi.it

In this paper we focus on a multi-item Latent Growth Curve (LGC) model for modelling change across time of a latent variable measured by multiple items at different occasions. We give guidelines on the specification of the variance-covariance structure of measurement errors. Then we investigate the empirical implications of different model specifications through an analysis of student ratings collected in four academic years about courses of the University of Florence. In the application we compare the compound symmetry correlation structure with the independence

structure. In particular, we discuss the implications of the two specifications in terms of interpretability of the results.

Hospital Clustering in the Treatment of Acute Myocardial Infarction Patients via a Bayesian Nonparametric Approach

Alessandra Guglielmi – Francesca Ieva – Anna M. Paganoni – Fabrizio Ruggeri – Jacopo Soriano

Politecnico of Milan, Italy, e-mail: alessandra.guglielmi@polimi.it, CNR-IMATI, Milan

In this paper we focus on a multi-item Latent Growth Curve (LGC) model for modelling change across time of a latent variable measured by multiple items at different occasions. We give guidelines on the specification of the variance-covariance structure of measurement errors. Then we investigate the empirical implications of different model specifications through an analysis of student ratings collected in four academic years about courses of the University of Florence. In the application we compare the compound symmetry correlation structure with the independence structure. In particular, we discuss the implications of the two specifications in terms of interpretability of the results.

International Strategy and Performance-Clustering Strategic Types of SMEs

Birgit Hagen – Antonella Zucchella – Paola Cerchiello – Nicolò De Giovanni

University of Pavia, Italy, e-mail: bhagen@eco.unipv.it; {paola.cerchiello; antonella.zucchella}@unipv.it

This contribution identifies different strategic types of internationalised SMEs, in so doing providing managers and entrepreneurs with a much better understanding of the main strategic options and their relationship with the international performance of firms. We provide a theoretical analysis of strategic orientations and strategic behaviour in international SMEs, followed by an empirical investigation based on a sample of Italian SMEs. The SMEs are grouped into strategic types using cluster analysis, and the link between strategic type and international performance is subsequently analysed using logistic regression. The empirical data suggest that there are four broad strategic types, namely an entrepreneurial/growth-oriented group of firms, a customer-oriented group, a product/inward-oriented cluster, and a further group of firms that lacks strategic orientation. The characteristics of the strategic clusters are discussed, and the regression results show that a clear and proactive strategic orientation and its consistency with business strategy leads to improved international performance. This confirms the positive and highly significant role of strategic types.

A Finite Mixture Multivariate Tobit Model for Market Basket Analysis

Harald Hruschka

University of Regensburg, Germany, e-mail: harald.hruschka@wiwi.uni-regensburg.de

We deal with multicategory decisions which households take during a shopping trip to a store and focus on interdependence between categories. To this end we introduce a finite mixture of multivariate Tobit-2 models with two response variables, purchase incidence and expenditure. Among several model variants the three segment model with segment specific cross category dependence performs best. Correlations for purchases of different categories are much more important than correlations among

expenditures and correlations among purchase and expenditures of different categories both in terms of frequency and absolute size. About 18% of all pairwise purchase correlations turn out to be significant. Segments differ with respect to number and size of significant purchase correlations.

Methodological Issues for a Clustered Ordinal Response Framework

Maria Iannario

University of Naples "Federico II", Italy, e-mail: maria.iannario@unina.it

In this paper, we present alternative frameworks for clustered ordinal data concerning a specific class of models denoted as CUB. Specifically, we analyze models that contain variables measured at different levels of the hierarchy by integrating them in a multi-context or multilevel structure. After a brief review on CUB models, we describe the two approaches and introduce the methodological issues for interpreting and discussing of the results.

Multivariate Functional Clustering for the Analysis of ECG Curves Morphology

Francesca Ieva – Anna Maria Paganoni – Davide Pigoli – Valeria Vitelli

Politecnico of Milan, Italy, e-mail: {francesca.ieva; anna.paganoni; davide.pigoli; valeria.vitelli }@mail.polimi.it

In this work a statistical analysis of a dataset of electrocardiographic (ECG) traces is proposed, concerning patients whose 12-leads pre-hospital ECG has been sent by life supports to 118 Dispatch Center of Milan. A statistical analysis on ECG curves morphology is proposed. First the signal is reconstructed, denoising measurements with a Daubechies wavelet basis. Then biological variability is removed via landmark registration. Finally, a multivariate functional k-means clustering based on H^1 distance is performed on the 8 leads which represent the QT-segments, extracted from smoothed and registered ECG signals.

A Text Classification Method to Measure Distance between Graduate Profiles and Labour Market Offers

Domenica Fioredistella Iezzi – Mario Mastrangelo – Scipione Sarlo

University of Rome-Tor Vergata, Rome, Italy, email: stella.iezzi@uniroma2.it.
University "La Sapienza" of Rome, Italy, email: {mario.mastrangelo; scipione.sarlo }@uniroma1.it.

In the last years, Universities have created an office of placement to facilitate the employability of graduates. University offices of placement select for companies, which offer a job and/or training position, a large number of graduates only based on degree and grades. We propose a method to measure the distance between a job announcement and Cvs of graduates taking into account characteristics of University courses of candidate. We analyse 1,650 job announcements collected in DB SOUL since January 1st, 2010 to April 5th, 2011.

Folded- and Log-Folded-*t* Distributions as Models for Insurance Loss Data

Andreas Kleefeld – Vytaras Brazauskas

Brandenburg University of Technology Cottbus, Germany, e-mail: kleefeld@tu-cottbus.de
University of Wisconsin, Milwaukee, USA, e-mail: vytaras@uwm.edu

A variety of probability distributions has been proposed in the actuarial literature for fitting of insurance loss data. We supplement the literature by adding the log-folded-normal and log-folded-*t* families. We present three methods for the estimation of parameters: method of maximum likelihood, method of moments, and method of trimmed moments. We fit the newly proposed distributions to data which represent the total damage done by 827 fires in Norway for the year 1988. The fitted models are then employed in a few quantitative risk management examples, where point and interval estimates for several value-at-risk measures are calculated.

Using Hidden Markov Models for Clustering Multivariate Linear-Circular Time Series

Francesco Lagona – Marco Picone

University Roma Tre, Italy, e-mail: {lagona; marco.picone} @uniroma3.it

We present a hidden Markov model for the analysis of time series of incomplete data profiles with two linear and two circular components, by integrating bivariate circular and bivariate skew-normal densities to describe latent regimes. Maximum likelihood estimation is facilitated by an EM algorithm that treats unknown class membership and missing values as different sources of incomplete information. The model is exploited on multivariate marine time series to identify transitions between wintertime sea regimes in the Adriatic sea.

Some New Stochastic Volatility Models, Fitted as Hidden Markov Models

Roland Langrock – Iain L. MacDonald – Walter Zucchini

Georg-August-Universität Göttingen, e-mail: {rsoleck; walter.zucchini}@uni-goettingen.de
University of Cape Town, e-mail: Iain.Macdonald@uct.ac.za

Objective Bayesian Comparison of Linear Regression Models

Luca La Rocca

University of Modena and Reggio Emilia, Italy, e-mail: luca.larocca@unimore.it

In this short paper, I consider the variable selection problem in linear regression models and review two objective Bayesian methods for which I have been developing R code. These two methods, namely, fractional Bayes factors and intrinsic priors, are useful when models are to be compared in lack of substantive prior information. In particular, they are useful when many variables are available for selection, and thus exponentially many models are to be compared, so that subjective prior elicitation under each model is virtually impossible. A case of special interest,

which ultimately motivates my work on this topic, is when the structure of an acyclic directed graph is to be learned from data; in this case the model space is even larger, because each graph corresponds to a family of linear regression models.

Weakly Supervised Learning

Riwal Lefort – Ronan Fablet – M. Jean-Marc Boucher

Telecom Bretagne/LabSTICC, Technopôle Brest-Iroise – France,
e-mail: {riwal.lefort; ronan.fablet; jm.boucherg}@telecom-bretagne.eu

In supervised learning, a label is associated to each data of the training set. In weakly supervised learning, a prior probability distribution vector is the label and indicates the probabilities for instances to belong to each class. Supervised, unsupervised, semi-supervised or presence/absence learning methods are also included in this model. In this context, we investigated generative models, discriminative models and classification models based on random forest for object recognition. Furthermore, an iterative procedure is proposed for modifying low prior values to higher values. The considered models are evaluated on standard datasets from UCI.

Panel Data Models with Spatial and Temporal Autocorrelation: A Bayesian Approach

Samantha Leorato – Maura Mezzetti

University of Rome-Tor Vergata, Italy, e-mail: {samantha.leorato; maura.mezzetti}@uniroma2.it

A hierarchical Bayesian model for spatial panel data is proposed. The idea behind the method is to analyze panel data taking into account a possible dependence within observations, besides temporal pattern, due to a geographical structure. This is realized by the introduction of a separable covariance matrix for the observation vector.

Evolutionary Customer Evaluation: A Dynamic Approach to a Banking Case

Caterina Liberati – Paolo Mariani

University Milano-Bicocca, Italy, e-mail: {caterina.liberati; paolo.mariani}@unimib.it

Today, the most important asset for a bank is its customer and therefore, the main targets to achieve by management are: knowledge of his needs, anticipation of his concerns and to distinguish itself in his eyes. The awareness that a satisfied customer is a highly profitable asset effort to provide a satisfactory service to the customer by diversifying its services. This paper aims to analyze the customer evaluation evolution of the main attributes of banking services to catch differences among the clusters and time lags through a dynamic factorial model. We propose a new system of weights by which assessing the dynamic factor reduction that is not optimal for all the instances considered across different waves. An empirical study will be illustrated: it is based on customer satisfaction data coming from a national bank with a spread network throughout Italy which wanted to analyse its reduced competitiveness in retail services, probably due to low customer satisfaction.

Workers Classifications in Business Census: Firms Organization Chart and the Origin- Destination Approach

Silvia Lombardi – Favio Verrecchia

ISTAT, Italy, e-mail: {lombardi; verrecchia}@istat.it

The 2011 Italian Business Census refers necessarily to definitions and classification legacies of the past, given its comparative purposes and the necessity to maintain an historical perspective. The definition of local units as statistical units of the 8th Business Census in 2001 according to CEE Regulation n. 696 and CE Regulation n. 2223, as well as the presence of enterprises among economic and legal units of analysis, partly influence classification systems of surveyed characters. The aim of the paper is to describe and evaluate the impact of theoretical approach underlying statistical definition of 2001 Business Census in the occupational field. In particular, employment will be detected within a theoretical and classification framework based on the Organization Chart approach and the Origin-Destination matrix.

Web Mining: An Application to a "Web District"

Eleonora Lorenzini

University of Pavia, Italy, e-mail: eleonora.lorenzini@unipv.it

The aim of this paper is to analyse the start up of an e-commerce experience, Store Valtellina, a "web district" which proposes in an integrated portal the quality produce of the Valtellina region. Odds ratios and social network analysis methods are employed over a database of the online sales, in order to understand the extent of the cross-selling of the different products and brands and the degree of integration of the network. Furthermore, logistic regression is carried out on a survey on the customer satisfaction of the visitors of the websites to evaluate how the different attributes of the site impact on the total satisfaction of the visitors.

Simultaneous *t*-Model-Based Clustering Applied to Company Bankrupt Prediction

Alexandre Lourme – Christophe Biernacki

Université de Pau et des Pays de l'Adour, France, e-mail: alexandre.lourme@univ-pau.fr
Université Lille 1 & CNRS, Villeneuve d'Ascq, France, e-mail: biernack@math.univlille1.fr

In many clustering situations, several samples arising from different populations, share identical statistical units described by the same features. Whereas such samples are usually classified independently we propose, here in a context of *t*-mixtures, a so-called simultaneous clustering method based on a stochastic link between the conditional populations. This link is justified, parametrical, parsimonious and estimable by Maximum Likelihood thanks to a GEM algorithm. We apply our new models on two financial company data sets differing over the years, which mix both healthy and bankrupt firms. Our simultaneous method points out that the hidden structure is more complex than generally expected. It distinguishes three groups: The first two ones correspond to healthy and bankrupt companies and a third one may represent firms of which failure cannot be predicted.

Component Analysis for Structural Equation Models with Concomitant Indicators

Pier Giorgio Lovaglio – Giorgio Vittadini

University Bicocca-Milan, Italy, e-mail: {piergiorgio.lovaglio; giorgio.vittadini}@unimib.it

In this paper, we extend Extended Redundancy Analysis (ERA) in such a way that it enables to specify and fit a variety of relationships among latent variables and endogenous indicators. Specifically, we extend this new class of models to allow for covariate effects both on the endogenous indicators and on the latent variables. In particular, covariates are allowed to affect endogenous indicators indirectly through the latent variables and/or directly. The method proposed herein is called Generalized Extended Redundancy Analysis (GERA).

Model-based Clustering and Classification via Mixtures of Multivariate t-Distributions

Paul D. McNicholas

University of Guelph, Ontario, Canada, e-mail: pmcnicho@uoguelph.ca

The use of mixture models for clustering and classification has received renewed attention within the literature since the mid-1990s. The multivariate Gaussian distribution has been at the heart of this body of work but approaches that utilize the multivariate t-distribution are burgeoning into viable and effective alternatives. In this paper, recent work on classification and clustering using the multivariate t-distribution is reviewed. Real and simulated data are used to illustrate the efficacy of these approaches relative to their Gaussian counterparts. The results are discussed and the paper concludes with discussion and suggestions for future work.

Use of Latent Trajectories and Learning Capital

Paolo Mariani – Emma Zavarrone

University Milano-Bicocca, Italy, e-mail: {paolo.mariani; emma.zavarrone} @unimib.it

The evolution of the concept of human capital has resulted in numerous quantification models. This paper, focusing on a specific dimension of human capital, which refers to knowledge accumulation through the attendance of University courses (Learning Capital or Academic Human Capital, shortly AHC), proposes to verify AHC latent trajectories amongst students' cohorts within the same University. Once provided a AHC definition and a measurement in terms of knowledge accumulation, the research aims at classifying latent trajectory by using dynamic principal component analysis.

Assessing Stability in Nonlinear PCA with Hierarchical Data

Marica Manisera

University of Brescia, Italy, e-mail: manisera@eco.unibs.it

Composite indicators of latent variables can be constructed by Non-Linear Principal Components Analysis when data are collected by multiple-item scales. The aim of this paper is to propose a resampling-based procedure able to establish the stability of the contribution of each item to the composite indicator and, simultaneously, take into account the hierarchical structure that there is often in the data, when individuals are nested in groups. The proposed procedure was applied to real data referring to job satisfaction and coming from the most extensive survey on the Italian social cooperatives.

Multivariate Regression Model Based on an Optimal Partition of Predictors

Francesca Martella – Donatella Vicari – Maurizio Vichi

University "La Sapienza" of Rome, Italy, e-mail: {francesca.martella; donatella.vicari ; maurizio.vichi}@uniroma1.it

A multivariate regression model based on an optimal partition of predictors (MRBOP) is presented. The proposed model attempts to identify latent predictors, defined as linear combinations of groups of explanatory variables which similarly predict the responses. In MRBOP the explanatory variables are partitioned into groups and the important groups with high prediction power are identified. The model is formalized in a least squares framework optimizing a quadratic objective function subject to a set of constraints due to the required clustering structure for the explanatory variables. An alternating least-squares (ALS) algorithm for fitting the MRBOP model is developed, and the performance of the new methodology is evaluated on simulation studies and on a real data set.

The Use of Background Information in Item Response Theory Models in the Context of Computer Adaptive Testing

Mariagiulia Matteucci – Stefania Mignani – Bernard Veldkamp

University of Bologna, Italy, e-mail: {m.matteucci; stefania.mignani}@unibo.it
University of Twente, The Netherlands, e-mail: b.p.veldkamp@utwente.nl

In this work, the use of background information both at person and item level is discussed within the framework of item response theory (IRT). Typically, the estimation of IRT models involves two different phases: the estimation of item parameters (calibration) and the estimation of individual abilities (scoring). In both phases, the measurement precision depends on testing conditions and on the availability of a) a large sample of individuals for test calibration, b) a large test length for scoring. Here, it is shown how covariates can be used effectively in order to improve accuracy of estimation. These issues are discussed also in a computerized adaptive testing environment.

Using the Variation Coefficient for Adaptive Discrete Beta Kernel Graduation

Angelo Mazza – Antonio Punzo

University of Catania, Italy, e-mail: {a.mazza; antonio.punzo}@unict.it

Various approaches have been proposed in literature for the kernel graduation of mortality rates. This paper focuses on the discrete beta kernel estimator, proposed in Mazza and Punzo (2011) which pragmatically considers the age as a discrete variable and which is conceived to naturally reduce boundary bias. Here, an attempt to improve its performance is provided. Firstly, we suggest a preliminary transformation of the data that helps to stabilize the variance and to reduce the curvature. Secondly, we allow the smoothing parameter to vary with age according to the reliability of the data measured via the reciprocal of the variation coefficient which is function of both the amount of exposure and the observed mortality rate. A formulation suggested in Gavin *et al.* (1995) is used for the local smoothing parameter.

Risk Attribution, Risk Budgeting and Implied Portfolio Views: A Factor Approach

Domenico Mignacca

Eurizon Capital Management, e-mail: Domenico.Mignacca@eurizoncapital.com

The aim of this presentation is to define a methodological framework that enable us to value, using a factor approach, which are the implied views in a relative or absolute portfolio. In the first part, after introducing factor models, we focus on risk attribution decomposing sensitivities and concentration of risk on both dimensions: asset classes and factors separating the idiosyncratic risk. In the second part, we show that the marginal contribution can be thought as implied expected returns (scaled using expected Info ration or sharpe ratio depending on the absolute or relative “nature” of the portfolio) under the assumption of portfolio’s efficiency (relative or absolute). In the final part of the presentation, we outline an example of portfolio implied views monitoring.

An Actuarial Model for Assessing General Practitioners’ Prescribing Costs

Simona C. Minotti – Giorgio Spedicato

University Milano-Bicocca, Italy, e-mail: simona.minotti@unimib.it
University “La Sapienza” of Rome, Italy, e-mail: spedicato.giorgio@yahoo.it

Monitoring general practitioners’ prescribing costs is an important issue in order to efficiently allocate health care resources. At this aim we propose a methodology based on collective risk theory, where frequency and costs of prescription drugs are modelled by means of Generalized Additive Models for Location, Scale and Shape (GAMLSS). An example based on a quasi-real dataset is discussed.

From the Citation Index to Bibliometric Indicators

Gabriella Morandi

University of Pavia, Italy, e-mail: bibliod@unipv.it

Bibliometric indicators originates in the librarian world and the need for an efficient bibliographic research plays the most important role. The change of focus from the citation index to the need for a measure of literature impact is well documented by E. Garfield and J.D. de Solla Price: the impact factor begins. Ever since many other indicators have been proposed and now a lot of attention is dedicated to their use in the evaluation of Universities and researchers. Nevertheless there is a constant misuse of such indicators and a gap between the increasing amount of scientometric literature and their applications. This may be attributed to the primary goal of bibliometric indicators being disregarded.

Simulation Experiments for an Overall Similarity Index between Two Hierarchical Clusterings

Isabella Morlini – Sergio Zani

University of Modena and Reggio Emilia, Italy, e-mail: isabella.morlini@unimore.it, University of Parma, Italy

We propose a new dissimilarity index for comparing two hierarchical clusterings, on the basis of the whole dendrograms. We present and discuss its basic properties and we show that the index can be decomposed into contributions pertaining to each stage of the hierarchies. We show the relation of each component of the index with the currently used criteria for comparing two partitions, namely the Rand index and the simple matching coefficient. We obtain a similarity index S as the complement to one of the suggested distance measure and we show that its single components Sk obtained at each stage k of the hierarchies can be related to the measure Bk suggested by Fowlkes and Mallows (1983). We report results of a series of Monte Carlo experiments aimed at comparing the behaviour of S, Sk and other similarity measures over different experimental conditions. The first set of simulations is aimed at determining the behaviour of the indexes when the clusterings being compared are unrelated. The second set of simulations tries to investigate the robustness of the indexes with respect to different level of noise.

Kernel Functions in GWR Model

Massimo Mucciardi – Pietro Bertuccelli

University of Messina, Italy, e-mail: {massimo.mucciardi; pbertuccelli }@unime.it

In an innovative way we apply some multivariate kernels to GWR models to better understand the effect of spatial dimensions on spatial weights. Estimated models, in the proposed application, show a strong spatial variability of the parameter which highlight the importance of local models for this kind of studies.

Object-Oriented Bayesian Networks for Combining Several Features of the Quality

Flaminia Musella – Paola Vicard

University Roma Tre, Italy, e-mail: {fmusella;vicard}@uniroma3.it

Customer orientation can be a strategic tool to support management decisions. Private companies and public authorities carry out customer satisfaction surveys to measure the perceived quality. The quality is a dynamic feature often interpreted as a mix of satisfaction items that can be analyzed both separately and jointly. Moreover, they can contribute to produce an index that synthesizes the hidden global quality. Here, we propose to combine several aspects of satisfaction using Object-Oriented Bayesian Networks. We present an application where each satisfaction area is modelled by a Bayesian network learnt from data; an Object-Oriented Bayesian network is built to handle the domain as a whole. The tool can then be used to evaluate improvement actions developed in one or more areas.

Class and Gender Differences in Cultural Participation: Asymmetric Multidimensional Scaling of Cultural Consumption

Miki Nakai

College of Social Sciences, Ritsumeikan University, Kyoto, Japan, e-mail: mnakai@ss.ritsumei.ac.jp

It has been considered that cultural consumption and lifestyles represent a person’s cultural taste and preference, which establish hierarchies among social groups. However, another hypothesis postulates that omnivorous cultural taste patterns has been becoming prevailing. The omnivorous taste pattern shows up in numerous countries, including Western Europe, North America and Japan. However, characteristics of omnivore taste, or repertoires or combinatorial taste patterns have received less consideration. The aim of this paper is to increase the understanding of how one’s omnivorous taste pattern in terms of combinations of cultural engagement is composed, and what is the relation between the omnivore taste patterns and people’s social position. Asymmetric multidimensional scaling is used to develop better understanding of the structure of asymmetric relations between cultural activities and relationship between peoples cultural pattern of practice and their social position. The analysis based upon a nationally representative sample in Japan collected in 2005 (N=2915) reveals that there are some notable dissimilarities in cultural participation practices between males and females.

Classification of Financial Assets on the Basis of their Risk Profile

Marcella Niglio – Giuseppe Storti – Cosimo Damiano Vitale

University of Salerno, Italy, e-mail: {mniglio; storti; vitale}@unisa.it

This paper illustrates a procedure for classifying financial assets on the basis of their risk profile taking into account their risk level, measured in terms of Value at Risk, as well as an approximate measure of risk variability. Returns on each asset are modelled as regime GARCH models with a flexible error distribution. Risk variability is then proxied by the distance between the risk values associated to each model regime. Finally, the results of an application to the full set of S&P 100 stocks are presented.

Performance Measurement of Italian Provinces in the Presence of Environmental Goals

Eugenia Nissi – Agnese Rapposelli

University "G. D'Annunzio" Chieti-Pescara, Italy, e-mail: {nissi; a.rapposelli }@unich.it

The emergence of the idea of sustainable development intimates a vision of an ecologically balanced society, where it is necessary to preserve environmental resources and integrate economics and environment in decision-making. Consequently, there has been increasing recognition in developed nations of the importance of good environmental performance, in terms of reducing environmental disamenities generated as outputs of the production processes. In this context, the aim of the present work is to evaluate the environmental efficiency of Italian provinces for 2008 by using the non-parametric approach to efficiency measurement, represented by Data Envelopment Analysis (DEA) technique. To this purpose, we propose a two-step methodology allowing for increasing the discriminatory power of DEA in the presence of the heterogeneity of the sample. In the first phase, provinces are classified into groups of similar characteristics. Then, efficiency measures are computed for each cluster.

On the Simultaneous Analysis of Clinical and Omics Data A Comparison of Globalboosttest and Pre-validation Techniques

Margret-Ruth Oelker – Anne-Laure Boulesteix

University of Munich, Germany, e-mail: {margret.oelker; boulesteix}@campus.lmu.de

In medical research biostatisticians are often confronted with supervised learning problems involving different kinds of predictors including, e.g., classical clinical predictors and high-dimensional "omics" data. The question of the added predictive value of high-dimensional omics data given that classical predictors are already available has long been under-considered in the biostatistics and bioinformatics literature. This issue is characterized by a lack of guidelines and a huge amount of conceivable approaches. Two existing methods addressing this important issue are systematically compared in the present paper. The globalboosttest procedure (Boulesteix and Hothorn 2010) examines the additional predictive value of high-dimensional molecular data via boosting regression including a clinical offset, while the pre-validation method sums up omics data in form of a new crossvalidated predictor that is finally assessed in a standard generalized linear model (Tibshirani and Efron 2002). Globalboosttest and pre-validation are shortly introduced and discussed, then assessed based on a simulation study with binomial and survival data and finally applied to breast cancer microarray data for illustration.

External Analysis of Asymmetric Multidimensional Scaling Based on Singular Value Decomposition

Akinori Okada – Hiroyuki Tsurumi

Tama University, Japan, e-mail: okada@rikkyo.ac.jp
Yokohama National University, Japan, e-mail: tsurumi@ynu.ac.jp

An asymmetric similarity matrix among objects, for example, a brand switching matrix of consumers, can be analyzed by asymmetric multidimensional scaling. Suppose that *n* brands exist, and that *m* new brands are introduced. While the brand switching from existing to new brands can be observed, the brand switching from new to existing brands nor that among new brands cannot be observed. The present study analyzed the $n \times n$ similarity matrix by the asymmetric multidimensional scaling based on the singular value decomposition. This gives outward and inward tendencies of existing brands. Using the obtained outward tendency, the inward tendency of *m* new brands is derived. An application to the brand switching among margarine brands is presented.

Variable Selection for Microarray Data

Rachel O'Reilly – S. Wilson – C. Carlow – Paul D. McNicholas

University of Guelph, Ontario, Canada, e-mail: {oreillyr; swilso05; ccarlow; pmcnicho}@uoguelph.ca

When working with genetic microarray data, gene selection is a required step in nearly all types of analysis. The selection of genes clearly has an impact on the performance and accuracy of further methods. The algorithm proposed herein aims to retain genes which best differentiate between tissue classes. The proposed algorithm sequentially eliminates genes, and begins by fitting a mixture distribution to the full data set (i.e. with all of the genes). Then, sets of genes are selected by clustering similar genes and the fit of the mixture distribution is assessed with each potential group deleted. This process is repeated until the mixture distribution appears to give the 'best' fit. The method is illustrated using the well known Alon data.

Classification of Volatility in Presence of Time Varying Parameters

Edoardo Otranto

University of Messina, Italy, e-mail: otranto@unime.it

The classification of the volatility of financial time series has recently received a lot of contributions, in particular in the model-based framework. Recent works have evidenced as the volatility structure can vary along the time, with gradual or abrupt changes in the coefficients of the model. We wonder if these changes can affect the classification of the series in terms of similar volatility level and if the existing clustering algorithms can be adapted to this case. We propose a new class of Multiplicative Error Models to represent the realized volatility with the possibility of changes in the parameters of the model in terms of change of regime or time varying smoothed coefficients. The classification of the unconditional levels of volatility, derived from the models, is performed within each regime and in average, to take into account the information deriving from the time varying coefficients. The results are compared with the classification derived from constant coefficient models.

Imprecision and Vagueness: Interval-Valued Exploratory Data Analysis Approaches

Francesco Palumbo – Rosaria Romano

University of Naples Federico II, Italy, e-mail: francesco.palumbo@unina.it
University of Calabria, Italy, e-mail: romano.rosaria@gmail.com

Uncertainty plays a critical role in the whole human everyday knowledge process. Scholars with different scientific backgrounds, have attempted to consider the uncertainty not only in terms of randomness but an unified view has not yet reached. Formalisation and definitions we will consider are most likely been influenced by the statistical approach in data analysis. In the discrete calculus field Interval Arithmetic appeared in the late 60's. The first complete formalisation is marked by the appearance of the book entitled Interval Analysis by R.E. Moore in 1966. In 1965 and in another context, L.A. Zadeh, motivated by the same need, published his famous paper entitled Fuzzy Sets. Interval variables and fuzzy sets have been defined to capture two new and different sources of uncertainty that can be referred to the Imprecision and to the Vagueness, respectively. It appears clear that any effort to treat in one paper the practical and theoretical different implications of uncertainty in the statistical analysis will turn out boastful. This paper focuses on the special case of quantitative variables and it considers the uncertainty as imprecision and vagueness. In this framework some exploratory statistical approaches methods will be illustrated.

Functional Clustering and Temperature Effect on Health

Francesco Pauli

University of Trieste, Italy, e-mail: francesco.pauli@econ.units.it

It is a known fact that either low and high temperature can worsen health condition of the population, in particular among the elderly. In this work we will consider the issue of how temperature can explain particular health phenomena: we consider, for the city of Milan, the daily number of admissions to hospital due to all nonincidental causes for people aged 75 and more during summer periods (June, July and August) in the years 1993 to 2003 and the daily number of deaths in the same age class and year period but for the years 1993 to 2002.

Robust Clustering of Trade Data with Thinning Processes

Domenico Perrotta – Andrea Cerioli – Francesca Torti

EC Joint Research Centre, Ispra site, Italy, e-mail: domenico.perrotta@ec.europa.eu
University of Parma, Italy, e-mail: andrea.cerioli@unipr.it
University of Milano Bicocca and University of Parma, Italy, e-mail: francesca.torti@unimib.it

We address clustering issues in presence of densely populated data points, exhibiting linear structures with high degree of overlapping. To avoid the disturbing effects of high dense areas, we

retain (and then cluster) a sample of data with a process preserving the general structure of the data. The problem is approached as a spatial point process. An intensity function and a thinning process to select the sub-sample for the clustering are derived for the analysis of the EU trade data.

Multivariate Concordance Measures: A Proposal

Emanuela Raffinetti

University of Pavia, Italy, e-mail: emanuela.raffinetti@unipv.it

We propose a novel multivariate measure of goodness of fit for a multiple linear regression model when the relevant involved explanatory variables assume mostly categorical nature. The proposed measure is based on the response variable Lorenz curve and its dual construction. Once the linear regression estimates are obtained, one can proceed by defining the concordance curve represented by the response variable original values ordered according to the ranks assigned to the corresponding estimated values. The concordance curve position explains the goodness of fit of the multiple linear regression model to the data. We also provide an extension of the Lorenz curve definition for ordinal variables in order to develop a further concordance index in this case.

A Two-Phase Clustering Based Strategy for Outliers Detection in Geostatistical Functional Data

Elvira Romano – Antonio Balzanella

Second University of Naples, Italy, e-mail: {elvira.romano; antonio.balzanella }@unina2.it

In this paper we focus on the analysis of functional data spatially correlated. Especially we introduce a two-phase clustering method for outliers detection. It is such that the results obtained by a modified Dynamic Clustering process on which the strategy is based on, are opportunely analyzed thought a minimum spanning tree, in order to detect outliers.

An Application of the Mixed Effect Trend Vector Models to the Analysis of Asymmetric Square Contingency Tables with Auxiliary

Mark de Rooij

Leiden University, The Netherlands, e-mail: rooijm@fsw.leidenuniv.nl

We propose to use the mixed effect trend vector model for modelling of repeated multinomial choice data in the form of a square contingency table. Such data often shows asymmetries where more people change from category a to b than the other way around. In many cases an investigator has, besides the actual choices of the participants, auxiliary variables that pertain to the subjects under study. Most methodologies for asymmetric data do not take into account such variables. We will show how to incorporate these auxiliary variables into the mixed effects trend vector model and how they can be used to study differential change. The models are illustrated in detail with data from the Dutch parliamentary election studies 2006.

High-Dimensional Bayesian Classifiers Using Non-Local Priors

David Rossell – Donatello Telesca – Valen E. Johnson

Institute for Research in Biomedicine of Barcelona, Spain, e-mail: david.rossell@irbbarcelona.org
University of California, USA, e-mail: dtelesca@ucla.edu
University of Texas MD Anderson Cancer Center, USA, e-mail: vejohnson@mdanderson.org

Common goals in classification problems are (i) obtaining predictions and (ii) identifying subsets of highly predictive variables. Bayesian classifiers quantify the uncertainty in all steps of the prediction. However, common Bayesian procedures can be slow in excluding features with no predictive power (Johnson & Rossell 2010). In certain high-dimensional setups the posterior probability assigned to the correct set of predictors converges to 0 (Johnson & Rossell 2011). We study the use of non-local priors (NLP), which overcome the above mentioned limitations. We introduce a new family of NLP and derive efficient MCMC schemes.

Clustering and Classification of Macroseismic Fields in Hazard Evaluation

Renata Rotondi – Elisa Varini – Gaetano Zonno

C.N.R., Istituto di Matematica Applicata e Tecnologie Informatiche, Milan, Italy, e-mail: {reni; elisa}@mi.imati.cnr.it
Istituto Nazionale di Geofisica e Vulcanologia, Milan, Italy, e-mail: zonno@mi.ingv.it

This study is aimed at characterising the attenuation of earthquakes in Italy by exploiting the information provided by the macroseismic fields of the DBMI04 database. The analysis was carried out for the most damaging earthquakes (epicentral intensity of at least VII), which were subdivided into a learning set, composed of earthquakes with a considerable number of macroseismic data points, and a classification set, composed of earthquakes with less rich macroseismic information. The learning set was partitioned into classes of events with similar macroseismic behaviour using agglomerative hierachical clustering; the good quality of the earthquakes of the learning set guaranteed sharp partitions into these classes. Then, the remaining events were assigned to the classes obtained through recursive partitioning. The probability distribution of the intensity at sites for each class was chosen to be Binomial, and the unknown parameters were estimated via the Bayesian method. The models obtained can be used to forecast damage scenarios of future earthquakes.

A Two Layers Incremental Discretization Based on Order Statistics

Christophe Salperwyck – Vincent Lemaire

Orange Labs, Lannion, France – LIFL, Université de Lille 3, France,
e-mail: {christophe.salperwyck; vincent.lemaire }@orange-ftgroup.com

Large amounts of data are produced today: network logs, web data, social network data. The data amount and their arrival speed make them impossible to be stored. Such data are called streaming data. The stream specificities are: (i) data are just visible once and (ii) are ordered by arrival time. As these data can not be kept in memory and read afterwards, usual data mining techniques can not apply. Therefore to build a classifier in that context requires to do it incrementally and/or to keep a subset of the information seen and then build the classifier. This paper

focuses on the second option and proposed a two layers approach based on order statistics. The first layer uses the Greenwald *et al.* quantiles summary and the second layer a supervised method such as MODL.

On the Use of Principal Component Analysis for Assessing Multivariate Process Capability

Michele Scagliarini – Stefania Evangelisti

University of Bologna, Italy, e-mail: {michele.scagliarini; stefania.evangelisti}@unibo.it

This paper studies the effects of multivariate measurement errors on multivariate capability indices computed using the principal components analysis. This study shows that measurement errors influence the results of a multivariate process capability analysis, resulting in either a decrease or an increase in the capability of the process. To avoid unreliable conclusions a method is proposed for overcoming the effects of measurement errors. Furthermore, a statistical test that allows one to determine whether measurement errors alter the process covariance structure is discussed.

A Variance Reduction Method for Credit Risk Models

Gabriella Schoier – Federico Marsich

University of Trieste, Italy, e-mail: gabriella.schoier@econ.units.it

The problem of the asymmetric behaviour and fat tails of portfolios of credit risky corporate assets such as bonds has become very important, not only as regards the defaults but also the migration in different classes. This paper regards the use of a variable reduction method i.e. the Importance Sampling in order to reduce the variability of the tail of the Profit & Loss distribution of a portfolio of bonds. This provides speed up for computing economic capital in the rare event quantile of the loss distribution that must be held in reserve by a lending institution for solvency. An application to a real portfolio of bonds of an insurance company ends the paper.

Interpreting Error Measurement: A Case Study Based on Rasch Tree Approach

Annalisa Sarra – Lara Fontanella – Tonio Di Battista – Riccardo Di Nisio

University of Chieti-Pescara, Italy, email: {asarra; lfontan; dibattista dinisio} @dmqte.unich.it

This paper describes the appropriateness of Differential Item Functioning (DIF) analysis performed via mixed-effects Rasch models. Groups of subjects with homogeneous Rasch item parameters are found automatically by a model-based partitioning (Rasch tree model). The unifying framework offers the advantage to include the terminal nodes of Rasch tree as item random effects in the multilevel formulation of Rasch models. In such a way we are able to handle different measurement issues. The approach is illustrated with a cross-national survey on attitude towards female stereotypes. Evidence of groups DIF was detected and presented as well as the estimates of model parameters.

Classification on a Dimension Reduced Subspace

Luca Scrucca

University of Perugia, Italy, e-mail: luca@stat.unipg.it

Sufficient dimension reduction (SDR) methods aim at replacing a p-dimensional vector X of predictors with a lower-dimensional function R(X), with no loss of information on the dependence of the response variable Y on X. If Y is a categorical variable, this implies that no discriminatory information will be lost if classifiers are restricted to R(X). We present a proposal for deriving a classifier defined on a suitable lower-dimensional projection subspace. Such a subspace is estimated through a mixture-based sliced inverse regression (MSIR) method. This is an extension of sliced inverse regression (SIR) that uses finite mixtures of Gaussian densities to approximate the conditional distribution of the predictors. The proposed approach allows to easily obtain classification of observations projected onto the estimated subspace. An example based on Italian olive oils data is discussed.

Clustering Spatially Dependent Functional Data

Pier Cesare Secchi – Simone Vantini – Valeria Vitelli

Politecnico of Milan, Italy, e-mail: {piercesare.secchi; simone.vantini; valeria.vitelli}@polimi.it

We propose a new algorithm for unsupervised classification of functional data – indexed by the sites of a spatial lattice – that exploits spatial dependence by repeatedly generating random connectivity maps and by clustering, at each iteration, local representatives of neighboring functional data. The algorithm output is the frequency distribution of cluster assignment for each site of the lattice; a final classification map can be obtained via a majority vote on cluster assignments in each site. The frequency distribution of cluster assignment can also be used to define an a-posteriori criterion to choose the most suitable grouping structure. The algorithm is very flexible with respect to its implementation, allowing a context dependent treatment of spatial dependence and supporting diverse clustering schemes. Moreover, it is computationally more efficient than other standard clustering techniques, being suitable for the treatment of large and complex datasets. We illustrate different implementations of the algorithm by analyzing synthetic functional data.

A MCMC Approach for Learning the Structure of Gaussian Acyclic Directed Mixed Graphs

Ricardo Silva

University College London, e-mail: ricardo@stats.ucl.ac.uk

Graphical models are widely used to encode conditional independence constraints and causal assumptions, the directed acyclic graph (DAG) being one of the most common types of models. However, DAGs are not closed under marginalization: that is, a chosen marginal of a distribution Markov to a DAG might not be representable with another DAG, unless one discards some of the structural independencies. Acyclic directed mixed graphs (ADMGs) generalize DAGs so that clo-

sure under marginalization is possible. In a previous work, we showed how to perform Bayesian inference to infer the posterior distribution of the parameters of a given Gaussian ADMG model, where the graph is fixed. In this paper, we extend this procedure to allow for priors over graph structures.

On the Use of Item Response Models in the SEM Perspective

Anna Simonetto – Maurizio Carpita

University of Brescia, Italy, e-mail: {simonett,carpita}@eco.unibs.it

For the analysis of complex models for latent constructs measured with several items, the Structural Equation Models (SEM) are being widely disseminated. In this study, our intent is to show how to include in a SEM framework an Item Response Model (IRM), in order to preserve the important characteristics that distinguish this type of approach, such as the possibility of calculate the measurement scales and the reduction of the complexity of the model. We compare these results with a standard SEM.

Union Formation and Value Adaptation: Empirical Evidence from Bulgaria

Emiliano Sironi

Catholic University, Milan, Italy, e-mail: emiliano.sironi@unicatt.it

Family formation such as other life course events is consequence of socio-economic constraints and ideational factors. At the same time, family formation also causes changes in value orientations. This paper implements propensity score matching in order to prove that union formation affects individuals' value orientations, emphasizing opposite patterns for marriage and non marital cohabitation.

The Forward Search for Proximity Data

Nadia Solaro

University of Milano-Bicocca, Italy, e-mail: nadia.solaro@unimib.it

The Forward Search (FS) is a general methodology developed to disclose hidden patterns or potential anomalies in data. Its standard application requires the knowledge of a starting data matrix, but in several situations it could be not available. This problem arises particularly when proximity data are involved. In this work, we propose a method to extend the FS to a specific kind of proximities, i.e. dissimilarity measures which can be organized in square symmetric matrices.

Classifying Large Credit Data with Symbolic Cluster Representations: Evaluating SVM Based Approaches

Ralf Stecking – Klaus B. Schebesch

Carl von Ossietzky University Oldenburg, Germany, e-mail: ralf.w.stecking@uni-oldenburg.de
Western University Arad, Romania, e-mail: kbschebesch@uvvg.ro

Credit client scoring on medium sized data sets can be accomplished by means of Support Vector Machines (SVM), a powerful and robust machine learning method. However, real life credit client data sets are usually huge, containing up to hundred thousands of records, with good credit clients vastly outnumbering the defaulting ones. Such data pose severe computational barriers for SVM and other kernel methods, especially if all pairwise data point similarities are requested. Hence, methods which avoid extensive training on the complete data are in high demand. A possible solution may be a combined cluster and classification approach. Computationally efficient clustering can compress information from the large data set in a robust way, especially in conjunction with a symbolic cluster representation. Credit client data clustered with this procedure will be subjected to statistical learning methods in order to estimate classification models.

Variational Bayes Approximations for Model-Based Clustering Using Mixtures of Normal Inverse Gaussian Distributions

Sanjeena Subedi – Paul D. McNicholas

University of Guelph, Guelph, ON, Canada, e-mail: {ssubedi; pmcnicho}@uoguelph.ca

Model-based clustering using a finite mixture of normal inverse Gaussian (NIG) distributions is explored. NIG model arises from a mean-variance mixture of a univariate normal distribution with the inverse Gaussian distribution. Deviating from the traditional EM approach, a variational Bayes approximation is utilized for the parameter estimation. Use of the variational Bayes approach alleviates the computational complexities and uncertainties associated with the EM approach. Since NIG comes from an exponential family, upon choosing a conjugate prior, the appropriate hyper-parameters for the approximating density for the variational Bayes algorithm could easily be obtained. Application on simulated data sets with symmetric and skewed subpopulations as well as a real data set is discussed.

A Further Proposal to Multiple Impute Missing Categorical Values Using Latent Class Analysis: A simulation Study in IRT Framework

Isabella Sulis

University of Cagliari, Italy, e-mail: isulis@unica.it

In the framework of multi-item scales it is not unusual to observe uncompleted records, thus the solution of proceeding with a imputation method for incomplete units is often chosen. The aims of this work is to advance a further model-based procedure that uses information provided by the Latent Class Analysis in order to multiple impute missing observations in a set of *J* categorical items. A simulation study in Item Response Theory framework is presented using a data set in which units are simulated missing at random. The simulation study explores the accuracy in estimation in terms of bias and efficiency of the proposed method with regard to other widely validated missing data recovering methods for categorical data. This further proposal to multiple impute missing data using LCA has been implemented in the function miLCApol written in the R language.

Evaluating the University System Controlling for Potential Confounding Factors

Isabella Sulis – Mariano Porcu

University of Cagliari, Italy, e-mail: {isulis; mrporcu}@unica.it

The aim of this research is to advance a statistical methodology suitable for building up adjusted indicators of efficiency of the degree programs taking into account of the internal sources of heterogeneity across the objects under comparison (for instance the heterogeneity across students' socio-cultural characteristics). This aim will be pursued throughout the use of direct statistical standardization methods which allow to calibrate some of the indicators of efficiency currently used by the government with respect to the composition/structure of a standard population. Zero-Augmented models will be use to assess the influence of the Potential Confounding Factors on the credits accumulation process. The estimates of coefficient parameters which characterize the influence of individual factors on the value of the efficiency indicators will be used to adjust the outcome measures with respect to a standard population.

Mystery Shopping at Greek Banks: A Bayesian Network Analysis

Claudia Tarantola – Paola Vicard – Ioannis Ntzoufras

University of Pavia, Italy, e-mail: claudia.tarantola@unipv.it
University Roma Tre, Italy, e-mail: vicard@uniroma3.it
Athens University of Economics and Business, Greece, e-mail: ntzoufras@aueb.gr

The banking industry is highly competitive, and customer satisfaction plays an essential rule. Hence, methods to evaluate customer satisfaction are important for bank managers. In this work, two instruments for customer satisfaction analysis are combined: Mystery shopping methodology and Bayesian Networks. Mystery shoppers are used to survey and monitor the quality of customer service and to identify areas requiring enhancement. After each visit they complete a report prepared in advance on their service experience. Bayesian Networks are then used to provide a pictorial representation of the dependence structure between the variables of interest, and are used to study the effect of different improvement strategies. We present a real data analysis concerning customer satisfaction in Greek banks.

Metrics for Compositional Data

Agostino Tarsitano – Marianna Falcone

University of Calabria, Italy, e-mail: agotar@unical.it, e-mail: maryfalcon@libero.it

Some aspects of the use of compositional data in exploratory multivariate analysis, such as cluster analysis, are reviewed. In particular, results obtained using log-ratio transformed data are contrasted with those using ratio-based comparisons deeply rooted in the Italian statistical tradition. Specifically, we seek to understand which approach gives the most useful interpretation, and why. The points made are illustrated using simulated and real data sets.

VaR Estimation Based on Stock Price Movement Direction Forecasting Using SVM

Fedya Telmoudi – Mohamed El Ghourabi – Mohamed Limam

ISGT, University of Tunis, Tunis, email: {telmoudifedya; Mohamed.limam }@gmail.com
ESSEC, University of Tunis, Tunis, email: mohamed.elghourabi@gmail.com

In order to maximize their profits, the prediction of the Value at Risk (VaR) and the identification of risk periods for stock price index is a challenging task for market dealers. Support vector machine either for classification (SVC) or regression (SVR) has been proved performant in financial areas. In this paper, the SVR and SVC are discussed and applied for the following purposes. We develop a new method for VaR estimation based on GARCH-SVR-EVT model, where GARCH model is compatible with stylized facts of asymmetric volatility. Based on estimated VaR a threshold is set to help identify risk periods for stock price index. Further, we investigate the predictability of financial movement direction using SVC by forecasting the movement direction of the CAC40 index. Experimental results show that GARCH-SVR-EVT model for VaR estimation generates efficient results. The classification model based on SVC for movement direction forecasting is proved performant. The combined system model performs well in estimating risk based forecasted movement.

A Review of Some Hybrid Generative-Discriminative Methods for Classification

D. Mike Titterington – Jinghao H. Xue

University of Glasgow, Scotland, U.K., e-mail: michael.titterington@gla.ac.uk
University College London, England, U.K., e-mail: jinghao@stats.ucl.ac.uk

The generative and discriminative approaches to classification are introduced and their characteristics are compared. The two approaches are both model based, with the latter being more flexible. Each paradigm has its advantages and disadvantages and hybrid methods have been developed with a view to exploiting the best of both worlds. Some of these methods are described and compared.

Variable Selection via Correlated Component PLS-Type Regression

Laura Trinchera – Edith Le Floch – Arthur Tenenhaus

SUPELEC, Gif-sur-Yvette, France, email: {laura.trinchera; arthur.tenenhaus}@supelec.fr
CEA, Neurospin, LNAO, Gif-sur-Yvette, France, e-mail: edith.lefloch@gmail.com

In many applications we are faced with the analysis of landscape matrices having more columns (variables) than rows (observations). Feature selection and feature extraction methods are commonly used to analyze such data but they are rarely used in a conjoint approach. Partial Least Squares (PLS) methods are classical feature extraction tools that work in the case of high-dimensional data sets. However, in such cases, the interpretation of the results is still a hard problem. That is why interest is increasing in developing new PLS methods able to be, at the same time, a feature extraction tool and a feature selection method. In this paper a new PLS-type algorithm which combines feature extraction, by means of correlated components, and variable selection is presented. To conclude, an application of this new method on a real biological data set will be discussed.

Two-Way Oriented Data Approach and its Application

Mitsuhiro Tsuji – Kosuke Horinouchi – Yuki Izumoto – Toshio Shimokawa – Shigenori Tanaka

Kansai University, Takatsuki City, Osaka, Japan, e-mail: tsuji@kansai-u.ac.jp

We investigated some statistical applications by the dynamic graphics of matrix-type presentation. The validity of integration of MDS and Cluster Analysis was observed as follows; (i) The result of MDS was verified by Cluster Analysis, (ii) The result of Cluster Analysis was visualized on the map from MDS. We adopted INDSCAL model and INDCLUS model. Then we created some kinds of maps on the same plane which would show the latent structure. GIS (Geographical Information System) is one of very interesting technologies. Moreover, spatiotemporal GIS includes some latent structure between space-time. The dynamic graphics of matrix type presentation has some targets; (i) We can look at several results at the same time. (ii) We can move several results at the same time into their appropriate place. (iii) We can zoom in/out several results at the same time into their suitable size. We developed the two-way oriented data approach which is the effective application to execute the dynamic graphics, where we can make zoom-up, zoom-down and moving by the simple way like the smart phone (Android, Windows-phone). By making a full use of good interface technology, we could easily derive the latent structure.

Mosaic Displays for Combinatorial Psychometric Models

Ali Ünlü – Anatol Sargin

Dortmund Technical University, Germany, e-mail: uenlue@statistik.tu-dortmund.de
Generali Versicherungen, Munich, Germany, e-mail: anatol@sargin.info

This paper presents an application of mosaic plots to psychometric data. Mosaic displays are used for exploring knowledge structures and in combination with numeric data analysis methods. The usefulness of this graphing method in knowledge space theory is illustrated with empirical data.

Modelling Three-Way Social Network Data: A Cross-Nested Random Effects Model for Gossip in the Workplace

Marijtje A.J. van Duijn

ICS/Department of Sociology, University of Groningen, The Netherlands, e-mail: m.a.j.van.duijn@rug.nl

Social networks are usually collected as dyadic data: the relations between pairs of actors are recorded, directed or undirected, complete or personal networks. The actors act either as sender or receiver of a tie. Three-way social network data are rare and occur when relations are recorded involving three actors. An example that will be analyzed in this paper is gossip. In a closed group setting it is recorded, by means of selfreport, who gossips with whom about whom. In three-way data, actors have three roles: as sender, receiver, and as object (of gossip). A random effects model for binary three-way social network data is developed relating the probability of a gossip tie to individual properties and roles of the actors, network relations that may exist between any pair of them, possibly available three-way characteristics of them as a triplet. The random effects are used to account for the dependence between the observations, caused by each actor having multiple roles, and caused by each actor being involved in multiple ties (within and across roles). The resulting model, a combination of a logistic regression model with a trivariate normal distribution, is estimated using MCMC, as implemented in WinBUGS.

Correction of Incoherence in Statistical Matching

Barbara Vantaggi – Andrea Capotorti

University "La Sapienza" of Rome, Italy, e-mail: vantaggi@dmmm.uniroma1.it
University di Perugia, Italy, e-mail: capot@dmi.unipg.it

We deal with the statistical matching problem, with particular emphasis on the managing of inconsistencies. In fact, when structural zeros among variables are present, incoherence on the probability evaluations can arise. The aim of this paper is to remove such incoherences by using different methods based on distances minimization or least commitment imprecise probabilities extensions. We compare these methods through an exemplifying practical example that carries out to the light peculiarities of the statistical matching problem.

Linear Regression for Modal Symbolic Variables: A Proposal

Rosanna Verde – Antonio Irpino

Second University of Naples, Italy, e-mail: {rosanna.verde; antonio.irpino}@unina2.it

In this paper we present a novel method for the regression of modal symbolic data. The method is based on the Ordinary Least Squares and allows to define a linear relationship between a set of modal variables considered as predictors and response modal variable. The method is based on the Wasserstein distance between quantile functions. For the sake of brevity, we show the main methodological aspects.

Visualizing Reputational Maps

Marika Vezzoli – Emma Zavarrone

University of Brescia, Italy, e-mail: vezzoli@eco.unibs.it
IULM University, Milan, Italy, e-mail: emma.zavarrone@iulm.it

In this paper we introduce the Reputational Academic Map conceived as a synthetic measure of the different facets of the reputation. First, we extract the dominant reputation dimensions, then constituting a mosaic map based on the single latent dimensions. In doing this we offer a new way to better inspect the inner structure of the reputation, in particular relative to the Italian University system.

Model Based Clustering of Consumer Choice Data

Donatella Vicari – Marco Alfò

University "La Sapienza"of Rome, Italy, e-mail: {donatella.vicari; marco.alfo}@uniroma1.it

In some empirical contexts, 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, we may also investigate on how they affect the likelihood 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 modelling framework, and the corresponding (modified) EM algorithm is sketched.

Model-Based Classification of Time Course Data Using the Skew t Distribution

Irene Vrbik – Paul D. McNicholas

University of Guelph, Ontario, e-mail: {ivrbik; pmcnicho }@uoguelph.ca

The analysis of gene expression time course data is arguably the area of modern scientific endeavour with the greatest need for effective clustering techniques. Yet, the prevalence of such approaches within the literature is sparse. In recent work, a robust mixture modelling approach using the skew t distribution has been explored to regulated skewness. In addition, a family of Gaussian mixture models has been successfully utilized to deal with time course data by using a Cholesky decomposition of the covariance matrix. Combining these two adaptations results in a new family of skew t mixture models for the analysis of gene expression time course data. Parameter estimates for these models are found using a Monte Carlo expectation-maximization (MCEM) algorithm and models are selected using the BIC.

A New Distance Function for Prototype Based Clustering Algorithms in High Dimensional Spaces

Roland Winkler – Frank Klawonn – Rudolf Kruse

German Aerospace Center Braunschweig, Germany, e-mail: roland.winkler@dlr.de
Ostfalia, University of Applied Sciences, Germany, e-mail: f.klawonn@ostfalia.de
Otto-von-Guericke University Magdeburg, Germany, e-mail: kruse@iws.cs.uni-magdeburg.de

High dimensional data analysis poses some interesting and counter intuitive problems. One of this problems is, that some clustering algorithms do not work or work only very poorly if the dimensionality is high enough. The reason for this is an effect called distance concentration. In this paper, we show that the effect can be countered for prototype based clustering algorithms by using a clever alteration of the distance function. We show the success of this process by applying (but not restricting) it on FCM. A useful side effect is, that our method can also be used to estimate the number of clusters in a data set.

A Simplified Latent Variable Structural Equation Model with Observable Variables Assessed on Ordinal Scales

Angelo Zanella – Giuseppe Boari – Andrea Bonanomi – Gabriele Cantaluppi

Catholic University, Milan, Italy,
e-mail:{fangelo.zanella, giuseppe.boari, andrea.bonanomi, gabriele.cantaluppi}@unicatt.it

The communication is related to a wide empirical research promoted by the Catholic University of Milan (CUM) aimed at acquiring an insight into the real work possibilities of its students graduated in the last seven years and of the appreciation and satisfaction on the side of the firms, which offered them a job position. The group of 1164 firms which have a special connection with CUM, regarding new job appointments, was considered and they were sent a questionnaire, using web for sending and answering. The analysis of the 203 complete answers was conducted by having recourse to a structural equation model with latent variables.

Optimal Decision Rules for Constrained Record Linkage: An Evolutionary Approach

Diego Zardetto – Monica Scannapieco

Istat, Italy, e-mail: {zardetto; scannapi}@istat.it

Record Linkage (RL) aims at identifying pairs of records coming from different sources and representing the same real world entity. Probabilistic RL methods assume that the pairwise distances computed in the record comparison process obey a well defined statistical model, and exploit the statistical inference machinery to draw conclusion on the unknown Match/Unmatch status of each pair. Once model parameters have been estimated, classical Decision Theory results (e.g. the MAP rule) can generally be used to obtain a probabilistic clustering of the pairs into Matches and Unmatches. Constrained RL tasks (arising whenever one knows in advance that either or both the datasets to be linked do not contain duplicates) represent a relevant excep-

tion. In this paper we propose an Evolutionary Algorithm to find optimal decision rules according to arbitrary objectives (e.g. Maximum complete-Likelihood) while fulfilling 1:1, 1:N and N:1 matching constraints. We also present some experiments on real-world constrained RL instances, showing the accuracy and efficiency of our approach.

An Analysis of the Careers of Italian PhD Graduates: Are they Over-Educated?

M. Bini – L. Grilli

European University of Rome, Italy, e-mail: mbini@unier.it
University of Florence, Italy, e-mail: grilli@ds.unifi.it

The recent debate on the issue of over-education calls for quantitative methods for assessing the effectiveness of PhD programs. The aim of this study is to investigate a relevant aspect of external effectiveness, namely how the PhD holders evaluate the usefulness of their education for the current job. The analysis relies on a multilevel model taking into account the clustering of the graduates into PhD programs, which allows us to determine how the self-assessed usefulness of the education depends on the features of the studies program and on a set of individual characteristics. A key finding is that the usefulness of the PhD education is strongly related to the employment history of the PhD holders and the motivation underlying the choice to enrol in a PhD course.

The 8th Biennial International Meeting of the CLAssification and Data Analysis Group (CLADAG) of the Italian Statistical Society was hosted by the University of Pavia within the celebration of its 650th anniversary, from September 7th to September 9th, 2011.

The present book contains the abstract of the papers presented during the meeting. The four pages version of the papers is contained in the USB pen, with an ISBN code as well. All papers were reviewed.

CLADAG promotes advanced methodological research in multivariate statistics with a special interest in Data Analysis and Classification. It supports the interchange of ideas in these fields of research, including the dissemination of concepts, numerical methods, algorithms, computational and applied results.

CLADAG is a member of the International Federation of Classification Societies (IFCS). Among its activities, CLADAG organizes a biennial international scientific meeting, schools related to classification and data analysis, publishes a newsletter and cooperates with other member societies of the IFCS in the organization of their conferences.

The scientific program of the meeting covered the following topics:

Classification theory
Multivariate data analysis
Proximity structure analysis
Software developments
Applied classification and data analysis

Previous CLADAG meetings were held in Pescara (1997), Roma (1999), Palermo (2001), Bologna (2003), Parma (2005), Macerata (2007) and Catania (2009).

Texts published by Pavia University Press in the series “Editoria scientifica” have been peer-reviewed prior to acceptance by the Editorial Board. www.paviauniversitypress.it/scientifica

ISBN 978-88-96764-22-0

9 788896 764220

€ 9,80

Pavia University Press

Edizioni dell’Università degli Studi di Pavia