PROGRAMME AND ABSTRACTS

20th International Conference on

Computational Statistics (COMPSTAT 2012)

http://www.compstat2012.org

Amathus Beach Hotel, Limassol, Cyprus

27-31 August 2012

http://iasc-isi.org

Scientific Program Committee:

Ex-officio:
COMPSTAT 2012 organiser and Chairperson of the SPC: Erricos John Kontoghiorghes.
Past COMPSTAT organiser: Gilbert Saporta.
Next COMPSTAT organiser: Manfred Gilli.
Incoming IASC-ERS Chairman: Vincenzo Esposito Vinzi.

Members:
Ana Colubi, Dick van Dijk, Peter Filzmoser, Roland Fried, Cristian Gatu, Mia Hubert, Domingo Morales and Tommaso Proietti.

Consultative Members:
Representative of the IFCS: Geoff McLachlan.
Representative of the ARS of IASC: Cathy W.S. Chen.

COMPSTAT2012 Proceedings Management Committee:
Ana Colubi, Konstantinos Fokianos, Erricos John Kontoghiorghes and Gil González-Rodríguez.

Local Organizing Committee:
Erricos John Kontoghiorghes, Constantinos Chappas, Konstantinos Fokianos, George I. Kassinis, Nicos Middleton, Andreas Savvides, Klea Panayidou and Yannis Yatracos.

Local Organization Support Committee:
Evgenia Tsinti, Maria Protopapa and Elizabeth Price.

Dear Friends and Colleagues,

We warmly welcome you to Cyprus, for the 20th International Conference on Computational Statistics (COMPSTAT 2012). It is locally organized by members of the Cyprus University of Technology and University of Cyprus. The COMPSTAT is an initiative of the European Regional Section of the International Association for Statistical Computing (IASC-ERS), a section of the International Statistical Institute (ISI). The first COMPSTAT conference took place in Vienna in 1974, and the last two editions took place in Porto in 2010 and Paris in 2012.

COMPSTAT is one of the most prestigious world conferences in Computational Statistics, regularly attracting hundreds of researchers and practitioners, and has gained a reputation as an ideal forum for presenting top quality theoretical and applied work, promoting interdisciplinary research and establishing contacts amongst researchers with common interests.

Keynote lectures are addressed by Elvezio Ronchetti, University of Geneva, Switzerland, Stanley P. Azen, University of Southern California, USA, and Trevor Hastie, Stanford University, USA. A homage to Prof. Stanley P. Azen for his contribution to Computational Statistics takes place before his keynote talk. From 360 submissions 275 have been retained for presentation in the conference. The conference programme has 40 contributed sessions, 8 invited sessions, 3 Keynote talks, 16 organized sessions and 4 tutorials. There are approximately 350 participants.

The Proceedings have been published in an electronic Book comprising 77 papers and over 900 pages. All the papers submitted have been evaluated through a rigorous peer review process. Those papers that have been accepted for publication in the Proceedings have been evaluated thoroughly by at least 2 referees. This ensures the high quality proceedings volume in the main areas of computational statistics.

The organization would like to thank the editors, authors, referees and all participants of COMPSTAT 2012 who contributed to the success of the conference. Our gratitude to sponsors, scientific programme committee, session organizers, local hosting universities and many volunteers who have contributed substantially to the conference. We acknowledge their work and the support of our hosts, particularly the Cyprus University of Technology.

The COMPSTAT 2012 organizers invite you to the next edition of the COMPSTAT which will take place in Geneva, Switzerland in 2014 and celebrate its 40th anniversary. We wish the best success to Manfred Gilli the Chairman of the 21st COMPSTAT.

Erricos John Kontoghiorghes, Organiser and Chairperson of the SPC.

SCHEDULE

Sunday, 26th August 2012

19:30 - 20:30 Registration and Ice Breaker

Monday, 27th August 2012

08:45 - 17:00 Registration
10:00 - 10:10 Opening (R1: Demetra)
10:10 - 11:10 KEYNOTE TALK (Elvezio Ronchetti, University of Geneva, Switzerland)
11:10 - 11:40 Coffee Break
11:40 - 13:20 PARALLEL SESSIONS B & INVITED SESSION Computational intensive methods in statistics
13:20 - 14:50 Lunch Break
14:50 - 16:30 PARALLEL SESSIONS C & INVITED SESSION Imperfect data
16:30 - 17:00 Coffee Break
17:00 - 18:15 PARALLEL SESSIONS D
20:00 - 21:30 Reception (Mediterranean Beach Hotel - Aegean Terrace)

Tuesday, 28th August 2012

07:30 - 08:55 ERS BoD Meeting (R8: Era)
08:30 - 17:00 Registration
09:00 - 10:40 PARALLEL SESSIONS E & INVITED SESSION Signal extraction and filtering &
TUTORIAL Knowledge extraction through predictive path modeling
11:10 - 11:40 Coffee Break
11:10 - 12:50 PARALLEL SESSIONS F & INVITED SESSION Statistical software in R with applications
12:50 - 14:20 Lunch Break
13:50 - 14:50 Hong Kong ISI Committee (R8: Era)
15:00 - 19:00 Excursion
19:00 - 20:10 Welcome drink in honour of Stan Azen, Ancient Curium
20:30 - 22:30 Concert at the Ancient Curium amphitheatre

Wednesday, 29th August 2012

08:30 - 16:00 Registration
08:40 - 10:45 PARALLEL SESSIONS G & INVITED SESSION Time series modelling and computation &
TUTORIAL Numerical methods and optimization in statistical finance
10:45 - 11:15 Coffee Break
11:15 - 12:55 PARALLEL SESSIONS H & INVITED SESSION Robust multivariate statistical methods &
TUTORIAL Bayesian computing and applications
12:55 - 14:25 Lunch Break
13:55 - 14:55 ERS General Assembly (R8: Era)
15:00 - 20:00 Excursion

Thursday, 30th August 2012

08:30 - 19:00 Registration
08:40 - 10:20 PARALLEL SESSIONS I & INVITED SESSION Small area estimation &
TUTORIAL Mixture models for high-dimensional data
10:20 - 10:50 Coffee Break
10:50 - 11:10 HOMAGE J (Homage to Stanley P. Azen, President of the IASC 2003-2005)
11:10 - 11:25 30 years of CSDA (Stanley P. Azen, University of Southern California, USA)
11:25 - 12:10 KEYNOTE TALK (Stanley P. Azen, University of Southern California, USA)
12:10 - 13:40 Lunch Break and CSDA Editorial Board meeting (R8: Era)
13:40 - 15:45 PARALLEL SESSIONS L & INVITED SESSION New developments in computational econometrics
15:45 - 16:15 Coffee Break
16:15 - 17:30 PARALLEL SESSIONS M
17:30 - 19:00 IASC Council Meeting (R8: Era)
20:30 - Conference Dinner

Friday, 31th August 2012

09:30 - 10:30 KEYNOTE TALK (Trevor Hastie, Stanford University, USA)
10:30 - 11:00 Awards for the best papers and Closing
11:10 - 11:30 ERCIM WG on Computing & Statistics Meeting

TUTORIALS, MEETINGS AND SOCIAL EVENTS

TUTORIALS

The tutorials will take place at room R8 (Era) during the conference and in parallel with the invited, organized and contributed sessions. The first is given by Vincenzo Esposito (Knowledge extraction through predictive path modeling) on Tuesday 28.08.2012, 09:00-10:40. The second tutorial is given by Manfred Gilli (Numerical methods and optimization in statistical finance) on Wednesday 29.08.2012, 08:40-10:45. The third tutorial is given by Cathy W.S. Chen (Bayesian computing and applications) on Wednesday 29.08.2012, 11:15-12:55. Finally, the fourth tutorial is given by Geoff McLachlan (Mixture models for high-dimensional data) on Thursday 30.08.2012, 08:40-10:20.

SPECIAL MEETINGS by invitation to group members

- ERS BoD Meeting, *R8: Era*, Tuesday 28th August 2012, 07:30 - 08:55.
- Hong Kong ISI Committee, *R8: Era*, Tuesday 28th August 2012, 13:50 - 14:50.
- ERS General Assembly, *R8: Era*, Wednesday 29th August 2012, 13:55 - 14:55.
- Lunch Break and CSDA Editorial Board meeting, *R8: Era*, Thursday 30th August 2012, 12:10 - 13:40.
- IASC Council Meeting, *R8: Era*, Thursday 30th August 2012, 17:30 - 19:00.
- ERCIM WG on Computing & Statistics Meeting, *R8: Era*, Friday 31th August 2012, 11:10 - 11:30.

SOCIAL EVENTS

- *The coffee breaks* will last one hour each, which adds fifteen minutes before and after to the times that are indicated in the program. There will be two different locations for each coffee break: in the Lobby area (Ground floor) and The Athenaeum Terrace (Mezzani) which is next to the lecture rooms.
- *Light Lunch* will be served at the Amathus Beach Hotel at a designated area for the conference participants. You must have your Lunch ticket of the appropriate day in order to attend the lunch.
- *Welcome Reception, Monday 27th of August, 20:00-21:30.* The Welcome Reception is open to all registrants (for free) who have reserved a place and non-registered accompanying persons who have purchased a reception ticket. It will take place in the adjacent to the venue Mediterranean Beach Hotel, Ouzeri tis Myrtos - Aegean Terrace. Conference registrants and any accompanying persons must bring their reception tickets in order to attend the reception.
- *Excursion to the Ancient cities and castles (Curium site) including Concert, Tuesday 28th of August 2012, 15:00.* The excursion is open to all registrants and non-registered accompanying persons who have purchased an excursion to the Ancient cities and castles ticket. A welcome drink will be offered to the participants before the Concert.
- *Concert at the Ancient Curium amphitheatre, Tuesday 28th of August 2012, 20:20.* For the registered conference delegates the buses are leaving from the Amathus Beach Hotel at 17:45. The concert will take place at the Ancient Curium Amphitheatre and be performed by The Cyprus Youth Symphony Orchestra and the 60-member World Youth Choir. A welcome drink will be offered to the participants before the Concert.

  The concert is open to all registrants (for free) who haved reserved a place and non-registered accompanying persons who have purchased a concert ticket. Please note that the Excursion tickets cannot be used for the buses that leave directly for the concert.
- *Excursion to the Mountain Villages, Wednesday 29th of August 2012, 15:00.* The excursion is open to all registrants and non-registered accompanying persons who have purchased an excursion to the Mountain Villages ticket.
- *Conference Dinner, Thursday 30th of August 2012, 20:30.* The conference dinner will take place at the Amathus Beach Hotel. The conference dinner is optional and registration is required. You must have your conference dinner ticket in order to attend the conference dinner.
- *Closing Dinner, Friday 31st of August 2012, 20:30.* The closing dinner celebrating the 30 years of the journal Computational Statistics & Data Analysis will take place at the Karatello Restaurant in the center of the town by the castle. Public buses passes through the area. The closing dinner is optional and registration is required. There is a limited number of places. You must have your closing dinner ticket in order to attend the closing dinner.

Address of venue:

The Conference venue is the Amathus Beach Hotel, Amathus Avenue, 3606 Limassol, Cyprus.

Registration and exhibitors

The registration will be open from Sunday late afternoon 26th August 2012 and will take place at the main lobby of the Amathus Beach Hotel. Exhibitors will be based at the Foyer in front of Rooms R1 - R4.

Lecture rooms

The paper presentations will take place at the mezzanine and ground floor of the Amathus Beach Hotel. There will be signs with indications to the various Room. Rooms R1 - R7 are in the mezzanine (first floor) while room R8 is in the ground floor. The opening, hommage, keynote and closing talks will take place at room R1 (Demetra). The poster presentations will take place adjacent to the rooms R5-R6 at the (mezzanine) Athenaeum terrace where the coffee breaks will take place. The room abbreviations are (capacity in brackets):

R1: Demetra (450)	R2: Ares (120)	R3: Hermes (30)
R4: Aphrodite+Poseidon (70)	R5: Athenaeum 1-2 (75)	R6: Athenaeum 3 (40)
R7: Athenaeum 4 (50)	R8: Era (70)	

Presentation instructions

The lecture rooms will be equipped with a PC and a computer projector. The session chairs should obtain copies of the talks on a USB stick before the session starts (use the lecture room as the meeting place), or obtain the talks by email prior to the start of the conference. Presenters must provide to the session chair with the files for the presentation in PDF (Acrobat) or PPT (Powerpoint) format on a USB memory stick. This must be done ten minutes before each session. The PC in the lecture rooms should be used for presentations. The session chairs are kindly requested to have a laptop for backup. Please note that Cyprus has identical plugs /power outlets to the UK, and thus differ from those in the rest of Europe and beyond. We cannot provide adapters, so please do not forget to take your adapters if needed.
IT technicians will be available during the conference and should be contacted in case of problems. The posters should be displayed only during their assigned session. The authors will be responsible for placing the posters in the poster panel displays and removing them after the session. The maximum size of the poster is A0.

Internet

Throughout the hotel there will be wireless Internet connection. The username is: *COMPSTAT2012* and password: *TEPAK*. Furthermore, the Room R3 (Hermes) can be used for Internet access. There will be a small number of laptops and a printer connected to the Internet and Ethernet cables in order to connect your own laptop.

Information and messages

You may leave messages for each other on the bulletin board by the registration desks. General information about restaurants, useful numbers, etc. can be obtained from the hospitality desk of the conference agency. The CPC Events Ltd. conference agency is responsible for the functioning of the venue, accommodation and the social events during the conference.

SPONSORS

International Association for Statistical Computing (http://www.iasc-isi.org)

ELSEVIER (http://www.elsevier.com)

Cyprus University of Technology (http://www.cut.ac.cy)

Central Bank of Cyprus (http://www.centralbank.gov.cy)

Cyprus Tourist Organization (http://www.cyprustourism.org)

SPE Mesa Geitonias (http://www.mesagitonia.coop.com.cy/)

Anorthosis Famagusta FC (http://www.anorthosis.com and http://www.famagusta.org.cy)

EXHIBITORS

Elsevier (http://www.elsevier.com)

John Wiley & Sons Ltd (http://www.wiley.com)

SAS Software Ltd (http://www.sas.com/)

Springer (http://www.springer.org/)

Contents

General Information **I**
Committees . . . II
Welcome . . . III
Scientific and Social Programme Schedule, Meetings and Social Events Information . . . V
Venue, lecture rooms, presentation instructions and internet access . . . VII
Sponsors & Exhibitors . . . VIII

Keynote Talks **1**
Keynote Talk 1 (Elvezio Ronchetti, University of Geneva, Switzerland) Monday 27.08.2012 at 10:10-11:10
Accurate robust inference . . . 1
Keynote Talk 2 (Stanley Azen, University of Southern California, United States) Thursday 30.08.2012 at 11:25-12:10
Computational statistics in support of translational research . . . 1
Keynote Talk 3 (Trevor Hastie, Stanford University, United States) Friday 31.08.2012 at 9:30-10:30
Matrix completion and large-scale SVD computations . . . 1

Parallel Sessions **2**

Parallel Session B (Monday 27.08.2012 at 11:40 - 13:20) **2**
IS04: COMPUTATIONAL INTENSIVE METHODS IN STATISTICS (Room: R1: Demetra) . . . 2
CS03: COMPUTATIONAL BAYESIAN METHODS I (Room: R4: Aph.+Pos.) . . . 2
CS24: TIME SERIES ANALYSIS I (Room: R5: Ath.1+2) . . . 3
CS35: COMPUTATIONAL ECONOMETRICS I (Room: R6: Ath. 3) . . . 3
CS09: CLUSTERING AND CLASSIFICATION I (Room: R7: Ath. 4) . . . 4

Parallel Session C (Monday 27.08.2012 at 14:50 - 16:30) **5**
IS01: IMPERFECT DATA (Room: R1: Demetra) . . . 5
OS11: ADVANCES IN SPARSE PCA AND APPLICATIONS (Room: R2: Ares) . . . 5
CS05: SPATIAL STATISTICS (Room: R7: Ath. 4) . . . 6
CS32: COMPUTATIONAL ECONOMETRICS II (Room: R5: Ath.1+2) . . . 6
CS36: METHODS FOR APPLIED STATISTICS I (Room: R4: Aph.+Pos.) . . . 7
CS31: TIME SERIES ANALYSIS II (Room: R6: Ath. 3) . . . 8

Parallel Session D (Monday 27.08.2012 at 17:00 - 18:15) **9**
OS05: MODELLING THROUGH BIPLOTS (Room: R2: Ares) . . . 9
OS03: INFERENCE FOR TIME SERIES (Room: R1: Demetra) . . . 9
CS02: CATEGORICAL DATA ANALYSIS (Room: R7: Ath. 4) . . . 10
CS13: SAMPLING METHODS (Room: R6: Ath. 3) . . . 10
CS17: HIGH-DIMENSIONAL DATA ANALYSIS I (Room: R4: Aph.+Pos.) . . . 11
CS14: COMPUTATIONAL BAYESIAN METHODS II (Room: R5: Ath.1+2) . . . 11

Parallel Session E (Tuesday 28.08.2012 at 09:00 - 10:40) **13**
IS07: SIGNAL EXTRACTION AND FILTERING (Room: R1: Demetra) . . . 13
TS02: TUTORIAL: KNOWLEDGE EXTRACTION THROUGH PREDICTIVE PATH MODELING (Room: R8: Era) . . . 13
OS15: NEW METHODS FOR ANALYZING MULTISET DATA (Room: R2: Ares) . . . 13
CS07: ROBUST STATISTICS I (Room: R4: Aph.+Pos.) . . . 14
CS10: CLUSTERING AND CLASSIFICATION II (Room: R7: Ath. 4) . . . 14
CS22: HIGH-DIMENSIONAL DATA ANALYSIS II (Room: R5: Ath.1+2) . . . 15
CS37: COMPUTATIONAL ECONOMETRICS III (Room: R6: Ath. 3) . . . 16
PS01: POSTER SESSION I (Room: Athenaeum Terrace) . . . 16
PS02: POSTER SESSION II (Room: Athenaeum Terrace) . . . 18

Parallel Session F (Tuesday 28.08.2012 at 11:10 - 12:50) **19**
IS02: STATISTICAL SOFTWARE IN R WITH APPLICATIONS (Room: R1: Demetra) . . . 19
OS12: VARIABLE SELECTION AND FEATURE EXTRACTION IN PREDICTIVE MODELING (Room: R5: Ath.1+2) . . . 19
OS07: ADVANCES IN COMPUTATIONAL ECONOMETRICS (Room: R2: Ares) . . . 20
CS04: ADVANCES IN DATA ANALYSIS (Room: R4: Aph.+Pos.) . . . 20
CS26: CLUSTERING AND CLASSIFICATION III (Room: R7: Ath. 4) . . . 21
CS27: MULTIVARIATE DATA ANALYSIS I (Room: R6: Ath. 3) . . . 22
PS03: POSTER SESSION III (Room: Athenaeum Terrace) . . . 22
PS04: POSTER SESSION IV (Room: Athenaeum Terrace) . . . 24

Parallel Session G (Wednesday 29.08.2012 at 08:40 - 10:45) **25**
IS03: TIME SERIES MODELLING AND COMPUTATION (Room: R1: Demetra) . . . 25
TS03: TUTORIAL: NUMERICAL METHODS AND OPTIMIZATION IN STATISTICAL FINANCE (Room: R8: Era) . . . 25
OS06: ISBIS SESSION ON INFORMATION MEASURES AND TECHNOLOGY (Room: R7: Ath. 4) . . . 25
OS09: IFCS SESSION ON FINITE MIXTURE MODELS (Room: R2: Ares) . . . 26
CS01: ROBUST STATISTICS II (Room: R4: Aph.+Pos.) . . . 27
CS08: BIOSTATISTICS AND BIOCOMPUTING (Room: R5: Ath.1+2) . . . 27
CS23: MULTIVARIATE DATA ANALYSIS II (Room: R6: Ath. 3) . . . 28
PS05: POSTER SESSION V (Room: Athenaeum Terrace) . . . 29
PS06: POSTER SESSION VI (Room: Athenaeum Terrace) . . . 30

Parallel Session H (Wednesday 29.08.2012 at 11:15 - 12:55) **32**
IS05: ROBUST MULTIVARIATE STATISTICAL METHODS (Room: R1: Demetra) . . . 32
TS01: TUTORIAL BY ARS OF IASC: BAYESIAN COMPUTING AND APPLICATIONS (Room: R8: Era) . . . 32
OS14: COMPONENT-BASED METHODS FOR SEM AND MULTI-BLOCK DATA ANALYSIS (Room: R2: Ares) . . . 32
CS06: TIME SERIES ANALYSIS III (Room: R4: Aph.+Pos.) . . . 33
CS12: NONPARAMETRIC STATISTICS I (Room: R6: Ath. 3) . . . 33
CS16: COMPUTATIONAL ECONOMETRICS IV (Room: R5: Ath.1+2) . . . 34
CS19: STATISTICS FOR INTERVAL DATA (Room: R7: Ath. 4) . . . 35

Parallel Session I (Thursday 30.08.2012 at 08:40 - 10:20) **36**
IS06: SMALL AREA ESTIMATION (Room: R1: Demetra) . . . 36
TS04: TUTORIAL BY IFCS: MIXTURE MODELS FOR HIGH-DIMENSIONAL DATA (Room: R8: Era) . . . 36
OS08: FUZZY CLUSTERING (Room: R4: Aph.+Pos.) . . . 36
OS17: ADVANCES IN COMPUTATIONAL STATISTICS AND DATA ANALYSIS (Room: R2: Ares) . . . 37
CS15: TIME SERIES ANALYSIS IV (Room: R5: Ath.1+2) . . . 37
CS34: COMPUTATIONAL ECONOMETRICS V (Room: R6: Ath. 3) . . . 38
CS30: SURVIVAL ANALYSIS (Room: R7: Ath. 4) . . . 38

Parallel Session L (Thursday 30.08.2012 at 13:40 - 15:45) **40**
IS08: NEW DEVELOPMENTS IN COMPUTATIONAL ECONOMETRICS (Room: R1: Demetra) . . . 40
OS02: ADVANCES IN THE ANALYSIS OF COMPLEX DATA (Room: R8: Era) . . . 40
OS16: ERCIM SESSION ON COMPUTATIONAL AND NUMERICAL METHODS IN STATISTICS (Room: R2: Ares) . . . 41
OS18: SFDS SESSION ON CO-CLUSTERING METHODS AND THEIR APPLICATIONS (Room: R5: Ath.1+2) . . . 41
OS19: BRANCHING MODELS, DERIVED MODELS, AND THEIR APPLICATIONS (Room: R4: Aph.+Pos.) . . . 42
CS18: STATISTICAL SOFTWARE (Room: R6: Ath. 3) . . . 43
CS29: CONTRIBUTIONS IN COMPUTATIONAL STATISTICS (Room: R7: Ath. 4) . . . 44

Parallel Session M (Thursday 30.08.2012 at 16:15 - 17:30) **45**
OS13: GENERALIZED CANONICAL CORRELATION ANALYSIS (Room: R1: Demetra) . . . 45
CS11: FUNTCIONAL DATA ANALYSIS (Room: R2: Ares) . . . 45
CS21: MONTE CARLO METHODS (Room: R5: Ath.1+2) . . . 46
CS25: PARAMETRIC MODELS (Room: R7: Ath. 4) . . . 46
CS28: METHODS FOR APPLIED STATISTICS II (Room: R4: Aph.+Pos.) . . . 46
CS33: NONPARAMETRIC STATISTICS II (Room: R6: Ath. 3) . . . 47

Authors Index **49**

Monday 27.08.2012 10:10-11:10 Room: R1: Demetra Chair: Ana Colubi Keynote Talk 1

Accurate robust inference

Speaker: **Elvezio Ronchetti, University of Geneva, Switzerland**

Classical statistics and econometrics typically rely on assumptions on the structural and the stochastic parts of the model and on optimal procedures derived under these assumptions. Standard examples are least squares estimators in linear models and their extensions, maximum likelihood estimators and the corresponding likelihood-based tests, and GMM techniques in econometrics. Inference is typically based on approximations obtained by standard first-order asymptotic theory. However, in the presence of small deviations from the assumed model, this can lead to inaccurate p-values and confidence intervals. Moreover, when the sample size is moderate to small or even in large samples when probabilities in the tails are required, first-order asymptotic analysis is often too inaccurate. We review a class of techniques which combine robustness and good accuracy in finite samples. They are derived using saddlepoint methods and provide robust tests for testing hypotheses on the parameters and for overidentification which are second-order correct in terms of relative error. Their nonparametric versions are particularly appealing as they are linked to empirical likelihood methods, but exhibit better accuracy than the latter in finite samples even in the presence of model misspecifications. The theory is illustrated in several important classes of models, including linear and generalized linear models, quantile regression, composite likelihood, functional measurement error models, and indirect inference in diffusion models.

Thursday 30.08.2012 11:25-12:10 Room: R1: Demetra Chair: Rand Wilcox Keynote Talk 2

Computational statistics in support of translational research

Speaker: **Stanley Azen, University of Southern California, United States**

New biomedical discoveries require the collaboration of biostatisticians and informaticists with multi-disciplinary investigators in conducting translational research. Multi-disciplinary collaborations lead not only to creating new knowledge that has biomedical, clinical and public health importance, but also to developing new biostatistical methodology. Examples that will be presented include: 1) translational research in cardiovascular disease leading to drug development; 2) population based studies in cardiovascular disease leading to the development of improved screening strategies; 3) population-based studies in ocular disease in multi-ethnic cohorts and its impact on public health; and 4) health promotion and its improvement on quality of life in the aging population The presentation also includes a discussion of 1) the challenges associated with training the next generation of statisticians in translational research; 2) developing quality clinical databases using informatics technology; and 3) examples of data mining databases which provide opportunities for junior faculty, post-doctoral fellows and graduate students to identify interesting findings leading to new screening tools, and identification of treatment-induced biological markers impacting clinical outcomes. Robust procedures for optimizing outcomes will also be discussed.

Friday 31.08.2012 9:30-10:30 Room: R1: Demetra Chair: Patrick Groenen Keynote Talk 3

Matrix completion and large-scale SVD computations

Speaker: **Trevor Hastie, Stanford University, United States**

The Singular Value Decomposition (SVD) is a fundamental tool in all branches of data analysis - arguably one of the most widely used numerical tools. Over the last few years, partly inspired by the Netflix problem, the SVD has again come into focus as a solution to the matrix completion problem. One partially observes a very large matrix, and would like to impute the values not observed. By assuming a low-rank structure, the SVD is one approach to the problem - a SVD with large amounts of missing data. We discuss an approach for building a path of solutions of increasing rank via nuclear-norm regularization. An integral part of this algorithm involves repeatedly computing low-rank SVDs of imputed matrices. We show how these tasks can be efficiently handled by parallel computational algorithms, allowing the method to scale to very high-dimensional problems.

Monday 27.08.2012 11:40 - 13:20 Parallel Session B

IS04 Room R1: Demetra COMPUTATIONAL INTENSIVE METHODS IN STATISTICS Chair: Cristian Gatu

C379: **Bootstrapping of short time-series multivariate gene-expression data**
Presenter: **Roy Welsch**, Massuchusetts Institute of Technology, United States
Co-authors: Piyushkumar Mundra, Jagath Rajapakse

Gene-expression time-series gathered from microarrays play an important role in understanding the functions of genes and their involvement in many biological processes. However, gene expressions at only a few time points are gathered from thousands of genes (variables) in these experiments. Statistical analysis of such data is difficult due to the curse of dimensionality, but could, perhaps, be improved with bootstrapping. However, the standard time-series bootstrapping techniques such as sieve or block bootstrapping are inapplicable due to the small number of time samples involved. In order to improve the predictions of these gene regulatory networks, we propose two approaches to bootstrapping. First, we use penalty methods such as ridge regression to build robust gene regulatory networks with significance testing in the sieve bootstrap formulation. Statistically significant ridge coefficients are used to build the robust predictive structure of the network (as observed in sparse biological networks) by bootstrapping the residuals. Second, we use standard random bootstrapping of the entire time series and introduce missing time points to make the bootstrapping a model-free and efficient approach to build networks with short time-series. Both methods are empirically demonstrated on a number of synthetic datasets derived using biologically relevant synthetic networks. Finally, the application of these bootstrapping techniques in deriving stable gene regulatory networks is demonstrated.

C304: **Fast semi-parallel regression computations for genetic association studies**
Presenter: **Paul Eilers**, Erasmus University medical Centre, Netherlands
Co-authors: Karolina Sikorska, Emmanuel Lesaffre, Patrick Groenen

The advent of non-expensive microarrays has made large scale genotyping of hundreds of thousands of SNPs (single nucleotide polymorphisms) feasible. This has led to a lot of activity in genome wide association studies (GWAS). The computations are simple: a regression model is estimated for each SNP, including covariates like age and gender and possibly corrections for population stratification. One to five million SNPs are common, a large proportion of them imputed, which means fitting one to five million regression models. Two ideas allow dramatic speed-ups. One applies rank-one updating of regression equations, for changes of genotypes from one SNP to the other. The second idea organizes computations as simple matrix operations that handle thousands of SNPS at the same time. Faster computation is even more important for mixed models, because there is growing interest in studying association between genotypes and longitudinal phenotypes. A mixed model is a natural candidate, but it typically takes a second or more to estimate it, for only one SNP. With millions of SNPs that would take a month on one computer. The semi-parallel approach is several orders of magnitude faster.

C292: **Computational strategies for non-negativity model selection**
Presenter: **Cristian Gatu**, Alexandru Ioan Cuza University of Iasi, Romania
Co-authors: Erricos John Kontoghiorghes

The problem of regression subset selection under the condition of non-negative coefficients is considered. The straight-forward solution would be to estimate the corresponding non-negative least squares of all possible submodels and select the best one. A new computationally efficient procedure which computes only unconstrained least squares is proposed. It is based on an alternative approach to quadratic programming that derives the non-negative least squares by solving the normal equations for a number of unrestricted least squares subproblems. The algorithm generates a combinatorial tree structure that embeds all possible submodels. This innovative approach is computationally superior to the straight-forward method. Specifically, it reduces the double exponential complexity to a a single traversal of a tree structure. The computational efficiency of the new selection strategy is further improved by adopting a branch-and-bound device that prunes non-optimal subtrees while searching for the best submodels. The branch-and-bound algorithm is illustrated with a real dataset. Experimental results on artificial random datasets confirm the computational efficacy of the new strategy and demonstrates its ability to solve large model selection problems that are subject to non-negativity constrains.

CS03 Room R4: Aph.+Pos. COMPUTATIONAL BAYESIAN METHODS I Chair: Cathy Chen

C065: **Approximate Bayesian computation based on the signed root log-likelihood ratio**
Presenter: **Samer A Kharroubi**, University of York, United Kingdom
Co-authors: Trevor Sweeting

We explore the use of importance sampling based on signed root log-likelihood ratios for Bayesian computation. Approximations based on signed root log-likelihood ratios are used in two distinct ways; firstly, to define an importance function and, secondly, to define suitable control variates for variance reduction. These considerations give rise to alternative simulation-consistent schemes to MCMC for Bayesian computation in moderately parameterized regular problems. The schemes based on control variates can also be viewed as usefully supplementing computations based on asymptotic approximations by supplying external estimates of error. The methods are illustrated by a genetic linkage model and a censored regression model.

C097: **BAT: The Bayesian analysis toolkit**
Presenter: **Allen Caldwell**, Max Planck Institute for Physics, Germany
Co-authors: Kevin Kroeninger, Daniel Kollar, Shabnaz Pashapour, Frederik Beaujean, Daniel Greenwald

The Bayesian Analysis Toolkit (BAT) is a software package developed in the C++ framework that facilitates the statistical analysis of the data using Bayes' Theorem. The tool evaluates the posterior probability distributions for models and their parameters using a Markov Chain Monte Carlo, which in turn provides straightforward parameter estimation, limit setting and uncertainty propagation. BAT provides a well-tested environment for flexible model definition and also includes a set of predefined models for standard statistical problems. The package is interfaced to other software packages commonly used in high energy physics, such as ROOT, Minuit, RooStats and CUBA. We present a general overview of BAT and its algorithms. Examples where BAT has been used in particle physics analyses are shown to introduce the spectrum of its applications. In addition, foreseen developments, such as parallelization and the extraction of Bayes Factors using a novel Markov Chain technique are summarized.

C390: **Bayesian semiparametric log-linear models for sample disclosure risk estimation**
Presenter: **Cinzia Carota**, University of Turin, Italy
Co-authors: Maurizio Filippone, Roberto Leombruni, Polettini Silvia

The number of categorical observations that are unique in a sample and also unique (or rare) in the population is usually taken as the measure of the overall risk of disclosure in the sample data. Attempts have been made in order to estimate this number in cross classifications of the k̈eyv̈ariables, i.e. multi-way contingency tables of those categorical variables with a key role in the identification of individuals in the sample. Methods based on parametric assumptions predominate. On the one hand, assuming the exchangeability of cells, elaborations of the Poisson model (Poisson-gamma,

Poisson–lognormal, multinomial-Dirichlet) have been extensively applied. Relaxing the exchangeability assumption, logistic or log-linear models have been used to capture the underlying probability structure of the contingency table. Our Bayesian semiparametric approach considers a Poisson model with rates explained by a log-linear model with normal fixed effects and Dirichlet process random effects. Suitable specifications of the base measure of the Dirichlet process allow us to extend many parametric models for disclosure risk estimation. The value of these extended models is discussed in an application to real data.

CS24 Room R5: Ath.1+2 TIME SERIES ANALYSIS I Chair: Konstantinos Fokianos

C150: **Meta-segmentation of time series for searching a better segmentation**
Presenter: **Christian Derquenne**, EDF Research and Development, France

A methodology to build a meta-segmentation of a time series is proposed. We propose first a method to segment a time series in several linear segments based on an exploratory approach and a heteroscedastic Gaussian linear model estimated by the REML estimator. Then we improve this method with an a priori step to better estimate the dispersion of the time series. Each one of both methods allows us to obtain several segmentations of the same time series. The choice of a segmentation can then be done by taking the REML, AIC, BIC or MAPE criterions, etc. However, these indicators allow us only to judge the overall quality of a segmentation. Indeed, a segmentation $t(j)$ can be very successful compared to others on a time interval, while the segmentation $t(k)$ has a very good quality on another time interval, etc. Under these conditions, if one happens to select the best parts of different segmentations proposed associated successively in the time, then we can hope that this new segmentation (or meta-segmentation) will be more optimal than the best segmentations individually. We compare the results for both methods. A simulated example illustrates the approaches. We propose future developments and potential applications.

C229: **Wild bootstrap tests for autocorrelation in vector autoregressive models**
Presenter: **Paul Catani**, Hanken School of Economics, Finland
Co-authors: Niklas Ahlgren

Conditional heteroskedasticity is a common feature of many macroeconomic and financial time series. Standard tests for error autocorrelation are derived under the assumption of IID errors and are unreliable in the presence of conditional heteroskedasticity. We propose wild bootstrap tests for autocorrelation in vector autoregressive (VAR) models when the errors are conditionally heteroskedastic. The bootstrap method is a residual-based recursive wild bootstrap procedure. In particular, we investigate the properties of Lagrange multiplier (LM) and F-type tests. Monte Carlo simulations show that the wild bootstrap tests have satisfactory size properties in models with constant conditional correlation generalised autoregressive conditional heteroskedastic (CCC-GARCH) errors. In contrast, standard asymptotic and residual-based bootstrap tests are shown to be oversized. Some simulation evidence on the power of the tests is given. The tests are applied to credit default swap prices, Euribor interest rates and international stock prices. The results show that there are significant ARCH effects in the residuals from the estimated VAR models. The empirical examples demonstrate that wild bootstrap tests for error autocorrelation should be preferred over standard asymptotic and residual-based bootstrap tests.

C162: **Forecast combination based on multiple encompassing tests in a macroeconomic DSGE-VAR system**
Presenter: **Robert Kunst**, Institute for Advanced Studies, Austria
Co-authors: Mauro Costantini, Ulrich Gunter

We study the benefits of forecast combinations based on forecast-encompassing tests relative to simple uniformly weighted forecast averages across rival models. For a realistic simulation design, we generate data by a macroeconomic DSGE-VAR model. Assumed rival models are four linear autoregressive specifications, one of them a more sophisticated factor-augmented vector autoregression (FAVAR). The forecaster is assumed not to know the true data-generating model. The results critically depend on the prediction horizon. While one-step prediction offers little support to test-based combinations, the test-based procedure clearly dominates at prediction horizons greater than two.

C330: **Multiscale correlations of volatility patterns across the stock market**
Presenter: **Milan Basta**, University of Economics - Prague - Faculty of Informatics and Statistics, Czech Republic

Volatility is an important variable in financial markets. We study to what extent patterns in volatility changes are shared across different sectors of the U.S. stock market as a function of the time horizon. Wavelets are used to handle the multiscale aspect of the problem. The log Garman-Klass estimator is used as a proxy to the unknown historical log volatility. Dissimilarities are calculated from correlation coefficients. Classical multidimensional scaling allows for the visualization of results. These results suggest that the multiscale aspect of the problem is a very crucial one as the proximity pattern changes as a function of the time scale. This supports the use of wavelets in the analysis of the characteristics of the stock market and shows that wavelets might be practically useful for understanding uncertainty in the market.

CS35 Room R6: Ath. 3 COMPUTATIONAL ECONOMETRICS I Chair: Alessandra Amendola

C280: **Some further results of an efficient algorithm for likelihood based biased estimating equations in the general linear model**
Presenter: **Munir Mahmood**, Gulf University for Science and Technology, Kuwait

It is well established that the problem of nuisance parameter in statistical inference is a longstanding one. An efficient algorithm for the likelihood function is delivered based on biased estimating equations following a general theory for estimation problems. The algorithm corrects the bias of a certain class of estimating equations and provides the marginal likelihood estimates when applied to the classical likelihood estimating equations in the context of the general linear model. We note that the classical likelihood provides maximum likelihood estimators which are biased but the algorithm, in contrast, yields marginal likelihood estimators which are unbiased. The striking result is, when the algorithm is applied to the least squares estimating equations, it provides the marginal likelihood estimates. Hence the algorithm unifies the estimates of the least squares method to that of marginal and classical likelihood based methods. We show that the algorithm provides the updating equation of the marginal likelihood when applied to the residual likelihood based estimating equations. Finally we compute the efficiency of the algorithm in terms of central processing unit (CPU) time, commonly known as execution time. This is demonstrated numerically in a comparison with the widely used Newton-Raphson algorithm. It is observed that the proposed algorithm takes less CPU time against the Newton-Raphson algorithm.

C337: **A new mathematical approach for an inverse problem in financial markets**
Presenter: **Masaki Mitsuhiro**, Graduate School of Doshisha University, Japan
Co-authors: Yasushi Ota, Hiroshi Yadohisa

When the Black-Scholes model is applied to financial derivatives, one of most interesting problems is reconciling the deviation between the expected and observed values. We derive the extension of the Black-Scholes model and recover binary call options' real drift from market prices. For space-dependent real drift, we obtain stable linearization and an integral equation. We also find that using market prices with different strike prices enables us to identify the term structure of the real drift. Results demonstrate that our new approach can confirm the existence of arbitrage in the market with a binary option transaction.

C380: **Time-varying analysis of dynamic stochastic general equilibrium models based on sequential Monte Carlo methods**
Presenter: **Koiti Yano**, Komazawa University, Japan

A new method is proposed to estimate parameters, natural rates, and unknown states of dynamic stochastic general equilibrium models simultaneously, based on the particle filter and a self-organizing state space model. We estimate the parameters and the natural rates using the time-varying-parameter approach, which is often used to infer invariant parameters practically. In most previous works on DSGE models, structural parameters of them are assumed to be deep (invariant). However, our method analyzes how stable structural parameters are. Adopting the TVP approach creates the great advantage that the structural changes of parameters are detected naturally. Moreover, we estimate time-varying natural rates of macroeconomic data: real output, inflation rate, and real interest rate. The fit of a DSGE model is evaluated using the log-likelihood of it. Thus, we are able to compare the fits of DSGE models. In empirical analysis, we estimate a new Keynesian DSGE model using the US data.

CS09 Room R7: Ath. 4 CLUSTERING AND CLASSIFICATION I Chair: Antonio Ciampi

C334: **Statistical registration and modeling of frontal view gait data with application to the human recognition**
Presenter: **Kosuke Okusa**, Chuo University, Japan
Co-authors: Toshinari Kamakura

We study the problem of analyzing and classifying frontal view human gait data by registration and modeling on a video data. We assume frontal view gait data as a mixing of scale changing, human movements and speed changing parameter. Our gait model is based on human gait structure and temporal-spatial relations between camera and subject. We estimate the parameters of the human gait using multistep algorithms based on the method of nonlinear least squares. The proposed algorithm is very stable to estimate each parameter. Finally, we apply a k-nearest-neighbor classifier, using the estimated parameters, to perform human recognition, and present results from an experiment involving 120 subjects. Our method shows high recognition rate, that has a better performance compared to other methods.

C309: **Model-based clustering in networks with stochastic community finding**
Presenter: **Aaron McDaid**, University College Dublin, Ireland, Ireland
Co-authors: Thomas Brendan Murphy, Nial Friel, Neil J. Hurley

In the model-based clustering of networks, blockmodelling may be used to identify roles in the network. We identify a special case of the Stochastic Block Model (SBM) where we constrain the cluster-cluster interactions such that the density inside the clusters of nodes is expected to be greater than the density between clusters. This corresponds to the intuition behind community-finding methods, where nodes tend to clustered together if they link to each other. We call this model Stochastic Community Finding (SCF) and present an efficient MCMC algorithm which can cluster the nodes, given the network. The algorithm is evaluated on synthetic data and is applied to a social network of interactions at a karate club and at a monastery, demonstrating how the SCF finds the 'ground truth' clustering where sometimes the SBM does not. The SCF is only one possible form of constraint or specialization that may be applied to the SBM. In a more supervised context, it may be appropriate to use other specializations to guide the SBM.

C352: **Three-way asymmetric hierarchical clustering based on regularized similarity models**
Presenter: **Kensuke Tanioka**, Graduate School of Doshisha University, Japan
Co-authors: Hiroshi Yadohisa

Three-way two-mode asymmetric data are observed in various situations such as brand switching, psychological research, and web mining.When clustering algorithms are applied to such data, several problems occur. One problem involves dealing with asymmetries. For two-way asymmetric data, there are two approaches to deal with asymmetries when using clustering algorithms. The first approach is to convert asymmetric similarities to symmetric similarities.The other approach is to introduce objective functions that consider internal variations of each cluster. However, for these clustering algorithms, it is difficult to understand the asymmetric features of the clustering results.The second problem involves determining the effects of occasions. A fuzzy clustering for three-way two-mode asymmetric data has been previously introduced and the effects of occasions have been considered. We propose two types of regularized similarity models and three-way asymmetric hierarchical clustering using entropy regularization. One regularized similarity model can provide us with factors of the direction of asymmetries, while the other model can provide us with factors comprising symmetric and asymmetric parts of asymmetric data. In addition, we introduce the factors of occasions using entropy regularization. Therefore, an advantage of the proposed algorithm is that researchers can easily interpret the clustering results.

C329: **Modified EM algorithms for model-based clustering of longitudinal data**
Presenter: **Antonio Ciampi**, McGill University, Canada
Co-authors: Yunqi Ji, Vicky Tagalakis

Clinical and public health studies often produce data in the form of measures repeated in time of a univariate disease index. To obtain insight in the nature of the disease, it is useful to describe these data as a finite mixture of a few typical courses, and to model each such course as a linear regression model which takes into account correlations. We present two new algorithms for model based clustering of longitudinal data. The algorithms are based on the extended linear mixed model. We present comparative evaluations of the new algorithms and a real data analysis.

Monday 27.08.2012 14:50 - 16:30 Parallel Session C

IS01 Room R1: Demetra IMPERFECT DATA **Chair: Ana Colubi**

C110: **Identifiability, estimation and inference for copula models based on censored data**
Presenter: **Ingrid Van Keilegom**, Universite catholique de Louvain, Belgium
Co-authors: Maik Schwarz, Geurt Jongbloed

A random variable (survival time) that is subject to random right censoring is considered. Instead of assuming that the censoring variable C is independent of the survival time T, we assume that the pair (T,C) is distributed according to a (bivariate) copula. We study under which conditions the model is identified when the marginals of T and/or C are unknown, and when the copula is either known or unknown but belonging to a parametric family. We also consider the problem of estimating the copula and the survival function.

C111: **Inclusion tests for the Aumann expectation of a random interval**
Presenter: **Ana Colubi**, University of Oviedo, Spain
Co-authors: Ana Belen Ramos-Guajardo, Gil Gonzalez-Rodriguez

Interval data appear frequently in experimental studies involving fluctuations, ranges, censoring times or grouped data. Random intervals (RIs) are suitable to handle such incomplete data in different settings. One and $k-$sample tests for the Aumann expectation for random intervals have been previously developed. On the other hand, the inclusion of the expected value of a normal random variable in a given interval has been studied by considering a multiple hypothesis test. The aim is to extend such a test to the context of random intervals. Procedures to test the inclusion of the Aumann expectation of a random interval in a given interval will be developed. The partial inclusion will also be considered by defining an index measuring the degree of inclusion. Asymptotic techniques will be developed taking advantage of the good properties of the consistent and asymptotically normal estimators. Some methods based on bootstrapping will be developed in order to get better empirical results for moderate sample sizes. A case-study regarding the blood pressure classification in adults is considered.

C172: **Analysis of fuzzy statistical data: A discussion of different methodological approaches**
Presenter: **Renato Coppi**, Sapienza University of Rome, Italy

The analysis of fuzzy statistical data depends crucially on the assumed nature of the related uncertainty, represented in the form of a fuzzy set. In the "ontic" perspective fuzziness is thought of as an intrinsic property of the datum, whereas the "epistemic" view looks at it as a way of formalizing the subjective ignorance about an underlying unknown crisp value. A coherent development of statistical procedures for fuzzy data analysis leads inevitably to different methods and models according to the adopted approach concerning the above mentioned nature. Unfortunately this coherence is not always respected in the huge literature devoted to this methodological area. It is argued that an explicit choice between the "ontic" and "epistemic" interpretations should be made before selecting and constructing an appropriate strategy for analyzing fuzzy statistical data. This involves the various steps of the analysis: from the descriptive indices, through exploratory tools, up to statistical models. The necessity of clarifying the basic issue of the nature of fuzziness becomes yet more stringent when the (fuzzy) data generation process, usually managed by means of probability, is taken into account. The notions of random fuzzy variable (RFV) or fuzzy random variable (FRV) seem appropriate to distinguish the two opposite perspectives. Some examples of this difference are discussed with reference to regression analysis, and a comparison of the obtained results is carried out by applying the different inferential tools, associated with the two approaches, to the same set of data. Several open problems are finally pointed out.

OS11 Room R2: Ares ADVANCES IN SPARSE PCA AND APPLICATIONS **Chair: Gilbert Saporta**

C122: **Sparse simultaneous component analysis**
Presenter: **Katrijn Van Deun**, KU Leuven, Belgium
Co-authors: Tom Wilderjans, Robert van den Berg, Anestis Antoniadis, Iven Van Mechelen

High throughput data are complex and methods that reveal structure underlying the data are most useful. Principal component analysis is a popular technique in this respect. Nowadays the challenge is often to reveal structure in several sources of information that are available for the same biological entities under study. Simultaneous component methods are most promising. However, the interpretation of the components is often daunting because the contributions of each of the thousands of variables have to be taken into account. We propose a sparse simultaneous component method that makes many of the parameters redundant. The method is flexible both with respect to the component model and with respect to the sparse structure imposed: Sparsity can be imposed either on the component weights or loadings, and can be imposed either within data blocks, across data blocks, or both within and across data blocks. A penalty based approach is used that includes the lasso, ridge penalty, group lasso, and elitist lasso. Estimation of the model relies on an alternating least squares and majorization minimization procedure. We will illustrate the method using empirical data and compare the sparse component weight and loading based models using simulated data.

C165: **Three-dimensional tomographic reconstruction with L1-minimization strategy for rotational angiography**
Presenter: **Helene Langet**, Supelec, France
Co-authors: Cyril Riddell, Arthur Tenenhaus, Yves Trousset, Elisabeth Lahalle, Gilles Fleury, Nikos Paragios

In X-Ray image-guided interventional angiography procedures, the patient's vessels are injected with contrast dye and two-dimensional (2D) projections are acquired with a C-arm system. The rotational acquisition of a series of projections enables the 3D reconstruction of the angiography data, but the technical limitations of C-arm systems or possible improper injection yields both spatial and temporal subsampling. Standard reconstruction methods such as the filtered backprojection (FBP) result in a volume that is deteriorated by streak artifacts, which potentially hampers medical interpretation. The recent developments of compressed sensing (CS) have demonstrated that it is possible to significantly improve the reconstruction of subsampled datasets by generating sparse approximations through ℓ_1-penalized minimization. A temporal regularization strategy and a spatial continuation strategy are used to exploit the sparsity of angiography data. A CS-based reconstruction that relies on an iterative derivation of FBP and where sparse constraints are applied via proximal operators is proposed. The relevance is evaluated in parallel geometry on synthetic data and on real angiographic data with the additional challenges of 3D cone-beam geometry, short-scan acquisition and truncated data. This strategy is shown to provide significant sampling artifact reduction and thus, improved image quality and better medical interpretation.

C343: **Sparse principal component analysis for multiblock data and its extension to sparse multiple correspondence analysis**
Presenter: **Anne Bernard**, CERIES, France
Co-authors: Christiane Guinot, Gilbert Saporta

Two new methods to select groups of variables have been developed for multiblock data: Group Sparse Principal Component Analysis (GSPCA) for continuous variables and Sparse Multiple Correspondence Analysis (SMCA) for categorical variables. GSPCA is a compromise between Sparse PCA method and the method Group Lasso. PCA is formulated as a regression-type optimization problem and uses the constraints of the group Lasso on regression coefficients to produce modified principal components with sparse loadings. It leads to reduce the number of nonzero coefficients, i.e. the number of selected groups. SMCA is a straightforward extension of GSPCA to groups of indicator variables, with the chi-square metric.

Two real examples will be used to illustrate each method. The first one is a data set on 25 trace elements measured in three tissues of 48 crabs (25 blocks of 3 variables). The second one is a data set of 502 women aimed at the identification of genes affecting skin aging with more than 370.000 blocks, each block corresponding to SNPs (Single Nucleotide Polymorphisms) coded into 3 categories.

C262: **Modeling resting-state brain activity with sparse decompositions**
Presenter: **Gael Varoquaux**, INRIA, France
Co-authors: Rodolphe Jenatton, Alexandre Gramfort, Guillaume Obozinski, Francis Bach, Bertrand Thirion

Brain imaging can be used to give a view on its functional organization. Starting from spatially-resolved recordings of brain activations at rest, exploratory analysis such as independent component analysis (ICA) can separate regions with different functional characteristics. With this application in mind, we review and compare recent progress in unsupervised models akin to PCA and sparse PCA applied to brain imaging. We introduce a generative model for brain resting-state time series, in which the signals are represented as linear combinations of latent spatial maps. Going beyond exploratory data analysis, our model provides a natural framework to guide parameter selection and impose a prior information on the structure of the signal. In particular, we explore approaches to imposing spatial structure or to modeling multiple subjects in this framework. We show that using well-designed penalizations injecting sparsity in a PCA context can yield better brain parcellations than the current ICA-based approach.

CS05 Room R7: Ath. 4 SPATIAL STATISTICS Chair: Gil Gonzalez-Rodriguez

C094: **Robust estimation and prediction for geostatistical data**
Presenter: **Andreas Papritz**, ETH Zurich, Switzerland
Co-authors: Hans Rudolf Kuensch, Cornelia Schwierz, Werner A. Stahel

Most geostatistical methods rely on non-robust algorithms. This is unfortunate, because outlying observations are rather the rule than the exception in environmental spatial data sets. Outliers affect the modelling of trend (external drift), of spatial dependence of residuals and the kriging predictions. Former studies on robust geostatistics focused on robust estimation of the sample variogram and ordinary kriging without external drift. A novel method for robustly estimating the parameters of a linear external drift and of the variogram of a Gaussian random field, possibly contaminated by independent errors from a long-tailed distribution, will be presented. It is based on robustification of the estimating equations for Gaussian REML estimation. Besides robust estimates of the parameters of the external drift and of the variogram, the method also provides standard errors for the estimated parameters, robustified kriging predictions at both sampled and non-sampled locations and kriging variances. Apart from the modelling framework, some simulation results, by which the properties of the new method were explored, and a case study about heavy metal contamination of soils around a metal smelter will be presented.

C217: **Statistical estimation of the accurate location based on the indoor positioning systems**
Presenter: **Toshinari Kamakura**, Chuo University, Japan
Co-authors: Kosuke Okusa

We daily use the global positioning systems for obtaining the location for car navigation. These systems are very convenient, but we sometimes need more accurate methods for obtaining the location of the specified objects in some special fields that sport athletes play, and we may demand for the location estimation in the indoor environments for obtaining the nursing care information in a care institution. We propose a statistical method for estimating the location in a room where we cannot receive any satellite information on the location. We use the TOA data based on the ultra-wideband (UWB) tag system. The main problem of TOA-based range measurements in indoor environments is that it is very difficult to model the errors by multipath and Non-Line-of-Sight. The proposed method is based on the repeatedly calculated centroid among the weighted equidistance from the several anchor radio sites distributed in the room. We shall compare the proposed statistical method and other previous methods and conclude that our iterative method is promising for practical use.

C252: **Epidemiological figures to show in spatial maps**
Presenter: **Thomas Waldhoer**, Medical University Vienna, Austria
Co-authors: Harald Heinzl

In the last decade epidemiological maps have found wide dissemination due to the availability of geographic information systems as well as sophisticated statistical approaches. A main objective of the latter is the estimation of shrinked epidemiological rates in order to reduce random variation in the observed crude rates. Though shrinking is an essential part of the description of the spatial distribution of rates, recent studies show that additional information could be added to spatial maps in order to improve their interpretability. We suggest the mapping of equivalence test results in combination with difference test results as well as the mapping of the directed power and the Type III error for well-chosen alternatives.

C247: **Detection of spatial clusteres using echelon scanning method**
Presenter: **Fumio Ishioka**, Okayama University, Japan
Co-authors: Koji Kurihara

Several approaches are used to detect spatial clusters for various kinds of spatial data, including a spatial scan statistic for finding spatial clusters based on likelihood ratios, which has been widely used. However, there remains the problem of how to find the regions with maximum likelihood ratio. Several useful techniques have so far been proposed. The authors have proposed a technique using an echelon analysis as the scanning method. Echelon analysis is a useful technique for systematically and objectively investigating the phase-structure of spatial regional data. The authors illustrate a feature of the echelon scanning method, and compare it with other scanning methods by applying it to sudden infant death syndrome (SIDS) data in North Carolina. In addition, the power of true cluster detection for the echelon scanning method are evaluated.

CS32 Room R5: Ath.1+2 COMPUTATIONAL ECONOMETRICS II Chair: Simon Broda

C059: **Alternative methodology for turning-point detection in business cycle: A wavelet approach**
Presenter: **Peter Martey Addo**, University of Paris 1, France, France
Co-authors: Dominique Guegan, Monica Billio

We provide a signal modality analysis to characterise and detect non-linearity schemes in the US Industrial Production Index time series. The analysis is achieved by using the recently proposed 'delay vector variance' (DVV) method, which examines local predictability of a signal in the phase space to detect the presence of determinism and non-linearity in a time series. Optimal embedding parameters used in the DVV analysis are obtained via a differential entropy based method using wavelet-based surrogates. A complex Morlet wavelet is employed to detect and characterise the US business cycle. A comprehensive analysis of the feasibility of this approach is provided. Our results coincide with the business cycles peaks and troughs dates published by the National Bureau of Economic Research (NBER).

C314: **Uncovering the HAR model with the LASSO**
Presenter: **Simon Knaus**, University of St. Gallen, Switzerland
Co-authors: Francesco Audrino

The heterogeneous autoregressive (HAR) model enjoys great popularity in the field of financial econometrics. The HAR model has become the apparent benchmark model for realized volatility not only for its strong predictive power, but also for its ease of implementation and estimation induced by its linear nature. However, the true structure of the underlying volatility process is not known. We show that the least absolute shrinkage and selection operator (LASSO) is well suited to partly answer the question about the true nature of the underlying process. To this end it is shown that the model selection consistency of the LASSO is satisfied under the assumption of the HAR model and the LASSO estimates should thus coincide with those of the HAR model. The poor agreement of the HAR model with the LASSO, together with non-substantial differences in out-of-sample forecasting, lead to the conclusion that the HAR model may not be the true model, however, it is equally well suited as the LASSO regression to capture the linear footprint of the volatility dynamics.

C323: **Estimation of implied intraday periodicity: Expectation-maximization framework**
Presenter: **Carlin Chun-Fai Chu**, The Chinese University of Hong Kong, China
Co-authors: Kai Pui Lam

Macroeconomic news announcement is commonly known to have a significant influence on the volatility of financial markets. Current practices of intraday periodicity estimation are based on sequential estimation methods to approximate the magnitude of the regular periodicity component. (i.e. averaging of the absolute/squared returns and Andresen's FFF approximation method). It can be shown that these methods mix the periodicity with the impact of the news announcement and, as a result, the estimated magnitudes are exaggerated, especially for those intervals subsequent to the news announcement time. Our proposed framework models the periodicity and the news impact through an Expectation-Maximization (EM) framework. The periodicity is modeled in the Expectation step while the news impact is handled in the Maximization step. The periodicity is estimated with respect to the news announcements, which are treated as the exogenous variables of the heteroskedastic structure of the deseasonalized series. The performance of the proposed method is examined through simulations and empirical experiments. The results indicate that the EM method outperforms the heuristic methods.

C351: **Tail probabilities and partial moments for quadratic forms in generalized hyperbolic vectors**
Presenter: **Simon Broda**, University of Amsterdam, Netherlands

Countless test statistics can be written as quadratic forms in certain random vectors, or ratios thereof. Consequently, their distribution has received considerable attention in the literature. Except for a few special cases, no closed-form expression for the cdf exists, and one resorts to numerical methods; under the assumption of Gaussianity, the relevant algorithms are well-known. These results are generalized to the case of multivariate generalized hyperbolic (MGHyp) random vectors. The MGHyp is a very flexible distribution which nests, among others, the multivariate t, Laplace, and variance gamma distributions. An expression for the first partial moment is also obtained, which plays a vital role in financial risk management. The proof involves a generalization of a classic inversion formula which relates the cdf to the characteristic function. Two applications are considered: first, the finite-sample distribution of the 2SLS estimator of a structural parameter. Second, the Value at Risk and Expected Shortfall of a quadratic portfolio with heavy-tailed risk factors.

CS36 Room R4: Aph.+Pos. METHODS FOR APPLIED STATISTICS I Chair: Niels Richard Hansen

C365: **Improvement of extreme temperatures probabilistic short-term forecasting**
Presenter: **Adriana Gogonel**, Electricite de France RD Division, France
Co-authors: Avner Bar-Hen, Jerome Collet

Temperature is a major risk factor for an electricity utility such as Électricité de France: it leads to increase in demand when temperature is lower than 18 Celsius for heating, and larger than 18 Celsius for cooling. To fulfill the risk management needs, one uses Ensemble Prediction Systems (EPS), provided by weather forecasting institutes, such as ECMWF. Nevertheless, the probabilistic representation of the future temperatures provided by EPS is too inaccurate regarding extreme quantiles, which are of particular interest for risk management. Our base modeling is the Best Member Method. The main part of this method is the modeling of the difference between the best member (the closest to the realization) of the EPS and the realization. This method is convenient for the central quantiles, but it is much less convenient for the tails. Investigating this, we state that, when the best member is the smallest or the largest value, the difference with realization has a different behavior than in other cases: it is much larger, and asymmetric. So we have to fit more accurately the dispersion of the error, and its distribution. We experimented some solutions, resulting in a significant improvement of tail representation.

C293: **Analysis of spatio-temporal data using the example of the English Channel ammonium concentration**
Presenter: **Helmut Waldl**, Johannes Kepler University Linz, Austria
Co-authors: Petra Vogl

The spatio-temporal analysis is based on data that have been made accessible by the Belgian Institute *Management Unit of the North Sea Mathematical Models* (MUMM) and are the output of the MUMM ecosystem model of the North Sea. The data set covers six years of model outputs, where realistic boundary and initial conditions have been set. The model outputs were stored at approximately a weekly rate, and for each sampling day several parameters (salinity, nitrates, ammonium, dissolved silicates, chlorophyll etc.) have been recorded for a grid of 3491 survey points. Here the concentration of ammonium is analyzed. Using geostatistical methods like kriging based upon diverse variogram estimators or parameter estimation in dynamic spatial models it was tried to detect and describe important physical phenomena which help to understand the mechanisms that govern the way the marine ecosystem works. Based upon these models and by means of the above mentioned estimates optimal experimental designs should be developed that allow an efficient monitoring of quality of the marine environment.

C071: **J-divergence estimator for scoring models**
Presenter: **Martin Rezac**, Masaryk University, Czech Republic

J-divergence is widely used to describe the difference between two probability distributions. It is also called the Information value for the purpose of scoring models, e.g. credit scoring models used to predict a probability of client's default. Empirical estimate using deciles of scores is the common way how to compute it. However, it may lead to strongly biased results. Moreover, there are some computational issues to solve. To avoid these issues and to lower the bias, the empirical estimate with supervised interval selection (ESIS) can be used. It is based on idea of constructing such intervals of scores which ensure to have sufficiently enough observations in each interval. The aim of the paper is to give an alternative procedure to estimate J-divergence between two probability distributions. It is called ESIS1 and leads to further reduction of the bias and the MSE, which are crucial for correct assessment of scoring models. By means of Monte Carlo simulations, the performance of the proposed estimator under various distribution parameters settings is compared with that of other standard estimators. The results are impressive and the proposed estimator, almost always, has higher performance than that of the other estimators considered.

C386: **Multivariate point process models in R with applications to neuron spike time modeling**
Presenter: **Niels Richard Hansen**, University of Copenhagen, Denmark

One important challenge from neuron science is the joint modeling of spike time patterns of multiple neurons and how stimuli affects not only the distribution of the spike pattern but also the inferred topology of the neuron connectivity. To this end we have developed the R packages processdata and ppstat. These are general purpose packages implementing data structures and the framework of generalized linear point process models, respectively, for dynamic models of recurrent discrete event times of multiple interacting processes. The generalized linear point process models are characterized by the predictable intensity being a non-linear function of a causal filter of observed processes that is linear in the unknown parameters. A flexible formula specification of the models is implemented and standard methods in R for basis expansions, e.g. splines, can be used to expand the filters. A smoother and a kernel component are implemented to support non-parametric modeling of filters and combinations of lasso-type penalties are implemented for selection of filter components as well as filter function support. The usage of the models and packages will be illustrated with examples from neuron spike time modeling.

CS31 Room R6: Ath. 3 TIME SERIES ANALYSIS II Chair: Roland Fried

C113: **Modeling categorical time series with strong autocorrelation**
Presenter: **Theodoros Moysiadis**, University of Cyprus, Cyprus

The multinomial logit model, widely used in the analysis of categorical time series, is applied for categorical data where strong autocorrelation is present. More specifically, the log-odds ratios relative to the conditional expectation of the response vector, given all the necessary information available to the observer until a specific time, are modeled through the link function, as a linear function of their own past values and of the corresponding past values of the multinomial responses, which together represent the time-dependent covariate process. Conditional likelihood inference is used for estimating the unknown parameters. The theory is illustrated by a data example.

C384: **Exploring relationships between financial and economic time series with wavelet and traditional methods**
Presenter: **Milan Basta**, University of Economics - Prague - Faculty of Informatics and Statistics, Czech Republic

The traditional analysis of relationships between time series usually assumes that one privileged time scale governs the dynamics of the time series. This privileged time scale is then often associated with the rate at which the time series is sampled. We compare these traditional approaches (such as regression methods, cross-correlation function etc.) with approaches based on wavelet analysis (such as regression with wavelet coefficients, wavelet cross-covariance and wavelet cross-correlation). We show examples where traditional approaches of one privileged time scale are definitely not appropriate and lead to a substantial loss in understanding of the true coupling between the time series. The results are illustrated on real life data sets as well as on artificial time series whose analysis is evaluated using Monte Carlo simulations.

C176: **Model selection in Bayesian structural breaks models with an application to OECD unemployment rates**
Presenter: **Alexander Vosseler**, Institute for employment research IAB, Germany

A fully Bayesian approach to stochastic model selection in autoregressive unit root test regressions with multiple structural breaks is presented. For this purpose a hybrid Markov Chain Monte Carlo (MCMC) algorithm is proposed, which can be used to draw samples from parameter spaces of varying dimension. In this context the number of breaks, the corresponding break dates as well as the number of autoregressive lags are treated as model indicators, whose marginal posterior distributions are estimated and then utilized for further inference. The performance of this algorithm is demonstrated on the basis of some Monte Carlo experiments. Having identified the most likely model in terms of posterior probabilities this parameterization is then used to test for a nonseasonal unit root. This is conducted by computing the posterior probability of the unit root null hypothesis. To check the prior sensitivity of the test decisions various prior densities are used here. In an empirical application the unemployment rates of 17 OECD countries, including Germany, Greece, France, Italy, Spain, UK and the US, are analyzed to answer the question if there is persistent behavior after a labor market shock. As a unit root implies trending behavior and thus persistence of the underlying stochastic process the proposed Bayesian unit root test with multiple structural breaks is applied to analyze the trending behavior of each country. To investigate further the country specific convergence properties I also compute the respective half lives of a labor market shock. Finally in order to control for uncertainty in the model selection step the joint posterior distributions of the number of structural breaks and the number of autoregressive lags are utilized to compute model averaged half lives for each country as well as the mean half life over all considered OECD countries.

C377: **The t-copula with multiple parameters of degrees of freedom: simulation, calibration and model selection**
Presenter: **Pavel Shevchenko**, CSIRO Australia, Australia
Co-authors: Xiaolin Luo

The t copula is a popular dependence structure often used in risk management as it allows for modeling the tail dependence between risks and it is simple to simulate and calibrate. The use of a standard t copula is often criticized due to its restriction of having a single parameter for the degrees of freedom (dof) that may limit its capability to model the tail dependence structure in a multivariate case. To overcome this problem, the grouped t copula was proposed in the literature, where risks are grouped a priori in such a way that each group has a standard t copula with its specific dof parameter. To avoid a priori grouping, which is often difficult in practice, recently we proposed a generalized grouped t copula, where each group consists of one risk factor. We present characteristics, simulation and calibration procedures for the generalized t-copula, including Markov chain Monte Carlo method for estimation and Bayesian model selection.

Monday 27.08.2012 17:00 - 18:15 Parallel Session D

OS05 Room R2: Ares MODELLING THROUGH BIPLOTS Chair: Patrick Groenen

C119: **Biplots: The palette**
Presenter: **Niel Le Roux**, Stellenbosch University, South Africa
Co-authors: John Gower, Sugnet Lubbe

The elements of biplots are either two sets of points or one set of points and a reference system. A reference system may be conventional linear axes, nonlinear trajectories, a set of points representing category levels or linear category level points representing constraints on category levels. Reference systems give exact representations of given data, possibly after transformations (e.g. normalisation, transformation to canonical variables, scaling to chi-squared and other metrics). For practical use, the exact representation has to be approximated, often in two dimensions. Interpretation of biplots is usually through distance or by orthogonal projections, notions that are underpinned by least-squares and the singular value decomposition. Other types of approximation are potentially acceptable. Several devices may be used to enhance visualisations (e.g. calibrated axes, shifting and rotating axes, use of colour, prediction regions). Biplots may be generalised to triplots that visualise three-mode data, and associated models, geometrically. Little modification is required when there is a preferred mode, such as K-groups or K-occasions; and this includes canonical variate analysis, INDSCAL and forms of Procrustes analysis, all of which relate to some form of group-average. Also, models expressed in one- and two-dimensional terms are readily visualised in biplots. When there is no preferred mode and data are modelled in terms of triple products, visualisation is cumbersome and possibly beyond the bounds of useful application. Two triple-product terms may be shown as three sets of points on the faces of a cube. The model terms are depicted as areas of triangles with vertices chosen from points on the three faces. This is at the limit of interpretability. As befits a visualisation technique, the geometry of biplots is quite elegant but its algebraic expression can be tedious.

C120: **Extending the theory of nonlinear biplots: Categorical canonical variate analysis biplots**
Presenter: **Niel Le Roux**, University of Stellenbosch, South Africa, South Africa
Co-authors: Sugnet Lubbe, John Gower

The theory of nonlinear biplots is used to develop the important special case when all variables are categorical with samples falling into K recognised groups. This technique is termed categorical canonical variate analysis because it has similar characteristics to Rao's canonical variate analysis, especially its visual aspects. It allows group means to be exhibited in increasing numbers of dimensions, together with information on within-group sample variation. Variables are represented by category level points, a counterpart of numerically calibrated biplot axes used for quantitative variables. Mechanisms are provided for relating samples to their category levels, for giving convex regions to help predict categories, and for adding new samples. Inter-sample distance may be measured by any Euclidean embeddable distance. Computation is minimised by working in the K-dimensional space containing the group means. An analysis of distance table is derived for exhibiting the contributions between and within groups. This can be broken down further into contributions arising from different dimensions and sets of dimensions, especially the fitted and the remaining residual dimensions. The latter may be further subdivided into the dimensions holding the group means and the distances orthogonal to them. An R package, aodbipl , for constructing various analysis of distance biplots is introduced. It is shown how provision can be made for various additive Euclidean embeddable distances; incorporating possible group structure; displaying weighted group means or means unweighted by group sizes together with their surrounding sample variability; zooming; interpolation of new samples; adding trajectories or category level points; predicting values for group means or samples and convex regions for predicting category levels.

C179: **Spline-based nonlinear biplots**
Presenter: **Patrick Groenen**, Erasmus University Rotterdam, Netherlands
Co-authors: Sugnet Lubbe, Niel Le Roux, Anthony La Grange

Biplots are helpful tools to establish the relations between samples and variables in a single plot. Most biplots use a projection interpretation of sample points onto linear lines representing variables. These lines can have marker points to make the reconstructed value of the sample point on that variable easy. For classical multivariate techniques such as principal components analysis, such linear biplots are well established. Other visualization techniques for dimension reduction, such as multidimensional scaling, focus on a possibly nonlinear mapping in a low dimensional space with emphasis on the representation of the samples. In such cases, the linear biplot can be too restrictive to properly describe the relations between samples and the variables. We propose a simple nonlinear biplot that represents the marker points of a variable on a curved line that is governed by splines. Its main attraction is its simplicity of interpretation: the reconstructed value of a sample point on a variable is the value of closest marker point on the smooth curved line representing the variable. The proposed spline-based biplot can never lead to a worse in sample fit of the variable as it contains the linear biplot as a special case.

OS03 Room R1: Demetra INFERENCE FOR TIME SERIES Chair: Konstantinos Fokianos

C091: **Saddlepoint approximations for INAR(p) processes**
Presenter: **Xanthi Pedeli**, University of Cyprus, Cyprus
Co-authors: Anthony C. Davison, Konstantinos Fokianos

Saddlepoint techniques have been introduced into statistics in a seminal paper and since then, they have been used successfully in many applications. Their usefulness lies on the high accuracy with which they can approximate densities and tail probabilities that are intractable. We are studying the application of saddlepoint techniques for the estimation of high-order integer-valued autoregressive (INAR(p)) processes. Traditionally, the method of maximum likelihood (ML) has been used for the estimation of the INAR(p) model. However, it requires distributional assumptions and, as the order p increases, it can become extremely complicated to be implemented in practice. As an alternative, we put forward an approximation to the log-likelihood by the saddlepoint method which, whilst simple in its application it performs well even in the tails of the distribution and under very complicated INAR models. We consider the case of Poisson innovations. The performance of the saddlepoint approximation is assessed through a series of simulation experiments which demonstrate its high accuracy even when maximization of the likelihood function is too hard to be feasible.

C109: **On modeling sea surface dynamics**
Presenter: **Anastassia Baxevani**, University of Cyprus, Cyprus
Co-authors: Pierre Ailliot, Anne Cuzol, Valerie Monbet, Nicolas Raillard

The surface of the ocean and hence quantities such as the significant wave height, H_s a parameter related to the energy of the sea-state, may be thought of as random quantities in space which evolve with time. We explore methods for constructing models for H_s based on fitting random field models to data collected from different sources, here hindcast and satellite altimeter observations. An important feature of the H_s fields, non-compatible with the popular and overly used separable models, is their motion. We propose a new approach by subordinating a static field with a dynamically changing velocity. First we define a velocity field through a flow of diffeomorphisms that are the solution to the transport equation and discuss ways to incorporate these velocity fields into the covariance structure. Then, we formulate and solve the motion estimation problem sequentially within a state-space model framework. The hidden state is the velocity field, which is supposed to be a Markovian process with a

transition kernel that is parameterized using a simple physical model. The hidden state is related to the observations through a conservation of the characteristics of the moving sea-states between successive times. Then, the velocity fields are estimated using a particle filter which permits to compute approximations of the distribution of the hidden state.

C396: **Reduced rank vector autoregression with volatility induced stationarity**
Presenter: **Anders Rahbek**, Copenhagen & CREATES, Denmark
Co-authors: Heino Bohn Nielsen

A new discrete-time multivariate model is suggested where lagged levels of the process enter both the conditional mean, as in a vector autoregression, and the conditional variance, as in a multivariate ARCH-type model where lagged residuals are replaced by lagged levels. The chosen formulation may imply volatility induced stationarity, and conditions are shown under which the multivariate process is strictly stationary and geometrically ergodic. Interestingly, these conditions include the case of unit roots and a reduced rank structure in the conditional mean, known from linear co-integration. The limiting distribution of the Gaussian maximum likelihood estimators is derived for a particular structured case, and show $\sqrt{T}$-convergence to Gaussian distributions–despite unit roots and the absence of even first order moments. A small Monte Carlo simulation confirms the usefulness of the asymptotics in finite samples. Finally, an empirical illustration using recent US term structure data is given. This analysis shows how the proposed model allows individual interest rates to have unit roots and no finite first-order moments while at the same time being strictly stationary. The interest rate spread, on the other hand, is autoregressive without a unit root.

CS02 Room R7: Ath. 4 CATEGORICAL DATA ANALYSIS Chair: Eva Ceulemans

C196: **Optimal scaling in multi-component PLS-based regression and principal component analysis**
Presenter: **Giorgio Russolillo**, CNAM, France

Non-Metric NIPALS and PLS-R algorithms extend the Partial Least Squares (PLS) approaches to Principal Component Analysis and regression to the treatment of variables measured on different measurement scales. These algorithms are optimal scaling procedures: a numerical (scaling) value is assigned to each modality of a non-metric variable by means of a quantification procedure. However, the optimality of these algorithms is referred to the first component, while for multi-component models it is not assured. This drawback is due to the fact that standard PLS algorithms extract components sequentially and the extraction of each successive component is conditioned to the knowledge of the previous ones. This peculiarity prevents from estimating optimal scaling parameters for multi-component models, as the solutions are not nested. To overcome this issue, two alternative PLS-R and NIPALS algorithms are proposed. These algorithms compute the whole set of components by means of a unique iterative loop which includes both the estimation and the deflation procedures. They are empirically showed to converge to the same solution than classical PLS-R and NIPALS. Moreover, a non-metric version of these algorithms is proposed, which estimates the scaling parameters as functions of the whole set of components. Finally, the optimality properties of such algorithms are discussed.

C305: **Constrained dual scaling of successive categories for detecting response styles**
Presenter: **Pieter Schoonees**, Erasmus University Rotterdam, Netherlands
Co-authors: Michel Van de Velden, Patrick Groenen

A constrained dual scaling method for detecting response styles is proposed. Response styles arise in questionnaire research when respondents tend to use rating scales in a manner unrelated to the actual content of the survey question. Dual scaling for successive categories is a technique related to correspondence analysis (CA) for analyzing categorical data. However, there are important differences, with one important aspect of dual scaling for successive categories data being that it also provides optimal scores for the rating scale. This property is used here together with the interpretation of a response style as a nonlinear mapping of a group of respondents' latent preferences to a rating scale. It is shown through simulation that the curvature properties of four well-known response styles makes it possible to use dual scaling to detect them. Also, the relationship between dual scaling and CA in conjunction with nonnegative least squares is used to restrict the detected mappings to conform to quadratic monotone splines. This gives rise to simple diagnostic maps which can help researchers to determine both the type of response style and the extent to which it is manifested in the data.

C373: **Simultaneous equating of separately calibrated item parameters under the common item design**
Presenter: **Shin-ichi Mayekawa**, Tokyo Institute of Technology, Japan
Co-authors: Sayaka Arai

When several test forms are calibrated, the item parameters must be equated using the information contained in a set of common (anchor) items. In this study, we developed a new equating method which can place all the item parameters on a common scale simultaneously. The method can handle as many test forms as required as long as they are connected through the common items, and it can handle the mixture of binary and polytomous items including two and three parameter logistic model, generalized partial credit model, graded response model and nominal response model. An alternating least squares (ALS) algorithm is used in which the conditional estimation of a set of item parameter values on the common scale and the conditional estimation of a set of linear transformation coefficients from the common scale to test-form dependent scales are repeated until convergence.

CS13 Room R6: Ath. 3 SAMPLING METHODS Chair: Lola Ugarte

C296: **Economic characteristics of LTPD sampling plans for inspection by variables**
Presenter: **Nikola Kasprikova**, University of Economics in Prague, Czech Republic
Co-authors: Jindrich Klufa

LTPD sampling plans minimizing mean inspection cost per lot of process average quality when the remainder of rejected lots is inspected were originally designed by Dodge and Romig for inspection by attributes. Plans for inspection by variables and for inspection by variables and attributes (all items from the sample are inspected by variables, remainder of rejected lots is inspected by attributes) were then proposed. Under the same protection of consumer the LTPD plans for inspection by variables are in many situations more economical than the corresponding Dodge-Romig sampling plans for inspection by attributes. Economic characteristics of LTPD plans for inspection by variables and for inspection by variables and attributes are discussed and impact of input parameters values on resulting sampling plan is shown.

C339: **Confidence intervals for relative and absolute frequency**
Presenter: **Michal Vrabec**, University of Economics Prague, Czech Republic
Co-authors: Lubos Marek

The very often task is the computation of confidence interval bounds for relative frequency when the sampling is without replacement. In other words, we build the confidence interval of the parameter value M in the parent population of size N (N may be finite) on the basis of a random sample of size n. There are many ways how to build this interval. We can use some one of normal approximations or binomial approximation. More accurate values can be looked up in tables. We consider one more method, based on the MS Excel calculation. We compare these different methods in our paper for specific values of M and we discuss when the methods are suitable.

C332: **Reinforcement-learning for sampling design in Markov random fields**
Presenter: **Mathieu Bonneau**, INRA, France
Co-authors: Nathalie Peyrard, Regis Sabbadin

Optimal sampling in spatial random fields is a complex problem, which mobilizes several research fields in spatial statistics and artificial intelligence. We consider the case where observations are discrete-valued and modelled by a Markov Random Field. Then we encode the sampling problem into the Markov Decision Process (MDP) framework. After exploring existing heuristic solutions as well as classical algorithms from the field of Reinforcement Learning (RL), we design an original algorithm, LSDP (Least Square Dynamic Programming), which uses simulated trajectories to solve approximately any finite-horizon MDP problem. Based on an empirical study of the behaviour of these different approaches on binary models, we derive the following conclusions: i) a naive heuristic, consisting in sampling sites where marginals are the most uncertain, is already an efficient sampling approach. ii) LSDP outperforms all the classical RL approaches we have tested. iii) LSDP outperforms the heuristic in cases when reconstruction errors have a high cost, or sampling actions are constrained. In addition, LSDP readily handles action costs in the optimisation problem, as well as cases when some sites of the MRF can not be observed.

CS17 Room R4: Aph.+Pos. HIGH-DIMENSIONAL DATA ANALYSIS I Chair: Luigi Augugliaro

C055: **Using random subspace method for prediction and variable importance assessment in linear regression**
Presenter: **Pawel Teisseyre**, Polish Academy of Sciences, Poland
Co-authors: Jan Mielniczuk

A prediction problem in regression with a high dimensional feature space is addressed. An approach based on Random Subspace Method (RSM) with a new weighting scheme is proposed. Weights of variables are defined as averages of squared values of pertaining t-statistics over fitted models with randomly chosen features. It is argued that such weighting is advisable as it incorporates two factors: a measure of importance of the variable within the considered model and a measure of goodness-of-fit of the model itself. Some formal properties of the proposed scheme are established, namely the form of asymptotic ordering of the variables for the fixed subset is determined. Moreover, the asymptotic form of weights assigned by the RSM is established. Numerical experiments performed on artificial and real datasets used in QSAR analysis indicate that the proposed method behaves promisingly when its prediction errors are compared with errors of penalty-based methods and it has smaller false discovery rate than the other methods considered. The pertaining method of weight assignment is compared with Breiman's measures proposed for Regression Random Forests and another previous approach.

C263: **Generic sparse group lasso and high dimensional multinomial classification**
Presenter: **Martin Vincent**, University of Copenhagen, Denmark
Co-authors: Niels Hansen

We present a general algorithm for solving the sparse group lasso optimization problem with a broad class of convex objective functions. Convergence of the algorithm is established, and we use it to investigate the performance of the multinomial sparse group lasso classifier. On three different real data examples we find that multinomial sparse group lasso clearly outperforms multinomial lasso in terms of achieved classification error rate and in terms of including fewer features for the classification. For the current implementation the time to compute the sparse group lasso solution is of the same order of magnitude as for the multinomial lasso algorithm as implemented in the R-package glmnet, and the implementation scales well with the problem size. One of the examples considered is a 50 class classification problem with 10k features, which amounts to estimating 500k parameters. The implementation is provided as an R package.

C321: **Differential geometric LARS via cyclic coordinate descent method**
Presenter: **Luigi Augugliaro**, University of Palermo, Italy
Co-authors: Angelo M. Mineo, Ernst C. Wit

The problem of how to compute the coefficient path implicitly defined by the differential geometric LARS (dgLARS) method in a high-dimensional setting is addressed. Although the geometrical theory developed to define the dgLARS method does not need of the definition of a penalty function, we show that it is possible to develop a cyclic coordinate descent algorithm to compute the solution curve in a high-dimensional setting. Simulation studies show that the proposed algorithm is significantly faster than the prediction-corrector algorithm originally developed to compute the dgLARS solution curve.

CS14 Room R5: Ath.1+2 COMPUTATIONAL BAYESIAN METHODS II Chair: Cathy Chen

C051: **Copulas selection in pairwise Markov chain**
Presenter: **Wojciech Pieczynski**, Telecom SudParis, France
Co-authors: Stephane Derrode

The Hidden Markov Chain (HMC) model considers that the process of unobservable states is a Markov chain. The Pairwise Markov Chain (PMC) model however considers the couple of processes of observations and states as a Markov chain. It has been shown that the PMC model is strictly more general than the HMC one, but retains the ease of processings that made the success of HMC in a number of applications. We are interested in the modeling of class-conditional densities appearing in PMC by bi-dimensional copulas and the mixtures estimation problem. We study the influence of copula shapes on PMC data and the automatic identification of the right copulas from a finite set of admissible copulas, by extending the general "Iterative Conditional Estimation" parameters estimation method to the context considered. A set of systematic experiments conducted with eight families of one-parameters copulas parameterized with Kendall's tau is proposed. In particular, experiments show that the use of false copulas can degrade significantly classification performances.

C268: **Bayesian model selection in multidimensional scaling by the product space method**
Presenter: **Kensuke Okada**, Senshu University, Japan

Comparing models that differ, e.g., in number of dimensions or representation of asymmetry is one of the key issues in the application of Bayesian multidimensional scaling (MDS). Generally there have been two approaches used in previous studies for this purpose: (i) those based on an information criterion called MDSIC and its variants; and (ii) those based on posterior distribution of a particular parameter set. However, no studies are available that directly estimate the Bayes factor, which is one of the most popular and useful statistic in Bayesian model selection. In this study, a method for directly estimating the Bayes factors to compare two MDS models is proposed. We adopt the product space method which is based on trans-dimensional Markov chain Monte Carlo algorithm. To apply this method, a mixture model in which the parameter sets of the two comparable MDS models are combined is constructed. Then, random samples are generated from the joint posterior distribution for the model index and all model parameters. The Bayes factor is estimated using Monte Carlo samples. The numerical comparison of the proposed method with existing methods are made to demonstrate the effectiveness of the proposed method.

C378: **Bayesian estimation of parameters and bandwidths for semiparametric GARCH models**

Presenter: **Xibin Zhang**, Monash University, Australia

Co-authors: Maxwell L. King

The aim is to investigate a Bayesian sampling approach to parameter estimation in the semiparametric GARCH model with an unknown conditional error density, which we approximate by a mixture of Gaussian densities centered at individual errors and scaled by a common standard deviation. This mixture density has the form of a kernel density estimator of the errors with its bandwidth being the standard deviation. The proposed investigation is motivated by the lack of robustness in GARCH models with any parametric assumption of the error density for the purpose of error-density based inference such as value-at-risk (VaR) estimation. The contribution is to construct the likelihood and posterior of model and bandwidth parameters under the proposed mixture error density, and to forecast the one-step out-of-sample density of asset returns. The resulting VaR measure therefore would be distribution-free. Applying the semiparametric GARCH(1,1) model to daily returns of stock indices, we find that this semiparametric GARCH model is often favored against the GARCH(1,1) model with Student t errors, and that the GARCH model underestimates VaR compared to its semiparametric counterpart. We also investigate the use and benefit of localized bandwidths in the proposed mixture density of the errors.

Tuesday 28.08.2012 09:00 - 10:40 Parallel Session E

IS07 Room R1: Demetra SIGNAL EXTRACTION AND FILTERING Chair: Stephen Pollock

C290: **Metropolising forward particle filtering backward simulation and Rao-Blackwellisation using multiple trajectories**
Presenter: **Tobias Ryden**, KTH Royal Institute of Technology, Sweden

Particle filters, or sequential Monte Carlo methods, are simulation-based methods for estimation of the latent states in state-space models. These methods have developed tremendously since invented almost 20 years ago, but common to all of them is a set of so-called particles that dynamically evolve in the state space as more observations become available; particles that fit observations well are duplicated to further explore the region where they are located while particles producing a poor fit are killed. The result is an estimate of the posterior distribution of the latent states given observations. Recently it has been demonstrated that by combining particle filter dynamics with a Metropolis-Hastings step deciding whether to accept a proposed set of particle locations, one can remove the estimation bias arising from the finite number of particles altogether. We will discuss how this approach can be extended to the case when particle trajectories are not drawn from the ancestral tree created by the particle evolution, but by backward sampling in the corresponding trellis. We also point to the potential of using multiple simulated trajectories as a means to profit on the simplicity and speed of this operation, as opposed to the more expensive simulation of a complete particle history. Statistically, this can be interpreted as partial Rao-Blackwellisation over the set of all backwards trajectories. Numerical examples on estimation of latent states, and parameter estimation through Monte Carlo EM, will be used to illustrate the above ideas.

C395: **The generalised autocovariance function**
Presenter: **Alessandra Luati**, University of Bologna, Italy
Co-authors: Tommaso Proeitti

The generalized autocovariance (GACV) function is defined as the Fourier transform of the power transformation of the spectral density. Depending on the value of the transformation parameter, the GACV nests the inverse autocovariances and the traditional autocovariance function. A frequency domain non-parametric estimator based on the power transformation of the pooled periodogram is considered and its asymptotic distribution is derived. The results are employed to construct classes of tests of the white noise hypothesis, clustering and discrimination of stochastic processes and to construct a feature matching estimator of time series parameters.

C291: **Cycles, syllogisms and semantics: Examining the idea of spurious cycles in macroeconomic data**
Presenter: **Stephen Pollock**, University of Leicester, United Kingdom

The claim that linear filters are liable to induce spurious fluctuations has been repeated many times of late. However, there are good reasons for asserting that this cannot be the case for the filters that, nowadays, are commonly employed by econometricians. If these filters cannot have the effects that have been attributed to them, then one must ask what effects the filters do have that could have led to the aspersions that have been made against them.

TS02 Room R8: Era TUTORIAL: KNOWLEDGE EXTRACTION THROUGH PREDICTIVE PATH MODELING Chair: Gil Gonzalez-Rodriguez

C138: **Knowledge extraction through predictive path modeling**
Presenter: **Vincenzo Esposito Vinzi**, ESSEC Business School of Paris, France

This tutorial will initially focus on the predictive modelling of relationships between latent variables in a multi-block framework by referring to the component-based approach of Partial Least Squares Path Modelling and its most recent variants and alternatives. We will consider several theoretical and methodological issues related to each modelling step, from measurement and structural model specification to model estimation, from the assessment of model quality to the interpretation of results. We will also provide some insights on the statistical criteria optimized by the presented approaches and their practical relevance enriched by a specific discussion on the outer weights and the dimensionality of latent variables. The difficulty of analysis is often due to the complexity of the network of hypothesized (but often hidden) and presumably causal (but mostly predictive) relationships between tangible phenomena or intangible concepts. Therefore, we will discuss the problem of extracting knowledge from uncertain models as compared to modelling the uncertainty in a specific model defined on a priori information. All material will be illustrated by examples and software to understand the results in practice.

OS15 Room R2: Ares NEW METHODS FOR ANALYZING MULTISET DATA Chair: Tom Wilderjans

C112: **Missing values in multi-level simultaneous component analysis**
Presenter: **Julie Josse**, Agrocampus, France
Co-authors: Marieke Timmerman, Henk A.L. Kiers

A common approach to deal with missing values in component analysis methods, such as principal component analysis, is to ignore the missing values by minimizing the loss function over all non missing elements. This can be achieved by EM-type algorithms where an iterative imputation of the missing values is performed during the estimation of the scores and loadings. However, such approaches are prone to overfitting problems. As an alternative, we propose a regularized iterative algorithm to deal with MCAR and MAR missing values in multi-level simultaneous component analysis (MLSCA), a method dedicated to explore data with groups of individuals. We will discuss the properties of this spectral thresholding iterative algorithm and explain the rationale of the regularization term, and draw attention to several points such as scaling issues and dysmonotony problems. We attach importance to separating the deviations due to sampling fluctuations and due to missing data. Finally, we show the performances of the method on the basis of a comparative extensive simulation study, and of a real dataset from psychology.

C124: **Clusterwise simultaneous component analysis for capturing between-block structural differences and similarities in multivariate multiblock data**
Presenter: **Eva Ceulemans**, University of Leuven, Belgium
Co-authors: Kim De Roover, Marieke Timmerman, Patrick Onghena

Many studies yield multivariate multiblock data, that is, multiple data blocks that all involve the same variables (e.g., the scores of different groups of subjects on the same variables). To explore the structure of such data, component analysis can be very useful. Specifically, two approaches are often applied: principal component analysis (PCA) on each block separately and simultaneous component analysis (SCA) on all blocks simultaneously. Recently, we introduced a generic modeling strategy, called Clusterwise SCA that comprises the separate PCA approach and SCA as special cases. Clusterwise SCA classifies the blocks into a number of clusters on the basis of their underlying structure. Blocks that belong to the same cluster are modeled using the same loadings, and the loadings may differ across the clusters. We discuss different members of the Clusterwise SCA family: the original Clusterwise SCA-ECP method which imposes Equal Cross-Product constraints on the component scores of the blocks within a cluster, the more general Clusterwise SCA-P method which allows for within-cluster differences in variances and correlations of component scores, as well as a model adaptation for letting the number of components vary across clusters.

C140: **Preprocessing in component analysis of high-dimensional data from experimental designs**
Presenter: **Marieke Timmerman**, University of Groningen, Netherlands
Co-authors: Eva Ceulemans, Age Smilde

In many studies experimental designs are used to generate multivariate data. To analyze those data, while exploiting the multivariate structure and utilizing the underlying design, one could use a component analysis (CA) approach. In CA, one usually 'preprocesses', which typically involves some form of centering and scaling of the observed data. Since experimental design data can be preprocessed in many different ways and rather different analysis results may occur, preprocessing is far from trivial in this case. We argue that modeling is preferable over preprocessing. We will outline the framework for high-dimensional fixed-effects ANOVA and show how it can be used to select a proper component model for data from an experimental design. We will show how different multi-set component models fit into this framework. Since all effects are part of the model, the 'centering' step is no longer needed. Scaling involves some prior weighting of observed variables. We will show that scaling in CA can be viewed as performing one or more weighted least squares analyses, and discuss the implications of this notion for empirical use. The use and usefulness of the framework is illustrated with an empirical example.

C170: **Clusterwise HICLAS: A generic modeling strategy to trace similarities and differences in multi-block binary data**
Presenter: **Tom Frans Wilderjans**, KU Leuven, Belgium
Co-authors: Eva Ceulemans

When studying multi-block binary data (e.g., successive multivariate binary observations of some subjects), a major challenge pertains to uncovering the differences and similarities between the structural mechanisms that underlie the different data blocks. To tackle this challenge for the case of a single data block (e.g., a single subject), one may rely on *Hierarchical Classes Analysis* (*HICLAS*). In case of multiple binary data blocks, one may perform *HICLAS* to each data block separately. However, such an analysis strategy obscures the similarities and, in case of many data blocks, also the differences between the blocks. To resolve this, the new *Clusterwise HICLAS* generic modeling strategy is proposed, in which the different data blocks are assumed to form a set of mutually exclusive clusters. For each cluster, different underlying mechanisms are derived. As such, blocks belonging to the same cluster have the same underlying mechanisms, whereas blocks of different clusters are modeled with different mechanisms. We discuss the results of a simulation study on the performance of the *Clusterwise HICLAS* algorithm and apply *Clusterwise HICLAS* to empirical data from the domain of developmental disabilities.

CS07 Room R4: Aph.+Pos. ROBUST STATISTICS I **Chair: Mia Hubert**

C306: **The MCS estimator of location and scatter**
Presenter: **Kaveh Vakili**, KU Leuven, Belgium
Co-authors: Mia Hubert, Peter Rousseeuw

A new robust estimator of multivariate location and scatter is introduced. Like MVE and MCD, it searches for an h-subset which minimizes a criterion. The difference is that the new criterion attempts to measure the cohesion of the h-subset. The optimal h-subset is called the Most Cohesive Subset (MCS). The MCS estimator uses projections and is affine equivariant. We construct a fast algorithm for the MCS estimator, and simulate its bias under various outlier configurations.

C144: **On the robustness of bootstrap tests for the homoscedasticity of random fuzzy sets**
Presenter: **Ana Belen Ramos-Guajardo**, University of Oviedo, Spain
Co-authors: Gil Gonzalez-Rodriguez, Maria Asuncion Lubiano

A comparison among some statistics for testing the equality of variances (or homoscedasticity) of k random fuzzy sets is carried out. The statistics proposed are defined on the basis of Levene's and Bartlett's classical procedures. The variance involved in the statistics is based on a generalized metric for fuzzy sets. Some asymptotic and bootstrap tests for equality of variances of random fuzzy sets have been recently developed by taking into account a Levene-based statistic. An introductory comparison between this statistic and a Bartlett-based one has been also carried out in this framework and the simulations indicated that the first approach was more conservative. The aim is to make an extensive comparison of more possible statistics (defined from the classical ones) by using bootstrap techniques. The robustness of those statistics is empirically analyzed by considering contamination in the samples. The analysis is carried out by means of simulation studies regarding both type I and type II errors.

C191: **Empirical comparison of the robustness of two L1 type medians for random fuzzy numbers**
Presenter: **Beatriz Sinova**, University of Oviedo, Spain
Co-authors: Maria Angeles Gil, Gil Gonzalez-Rodriguez, Stefan Van Aelst

To model certain experimental data, the scale of fuzzy numbers can be used. It integrates the mathematical manageability and variability of real numbers with the interpretability and expressivity to reflect the underlying imprecision (of valuations, ratings, judgements or perceptions) of the categorical scale. Examples and applications can be found in very different fields, from Engineering to Social Sciences. The notion of median, to avoid the sensibility of the Aumann expected value as central tendency measure of a random fuzzy number, has been defined by the moment through two L^1 metrics. The robustness of these two measures will be compared by means of some simulations.

C391: **Robust linear programming with coherent-risk constraints**
Presenter: **Pavel Bazovkin**, University of Cologne, Germany
Co-authors: Karl Mosler

Linear optimization problems with uncertain parameters are discussed. We apply coherent distortion risk measures to capture the possible violations of the restrictions. For each constraint there exists an uncertainty set (UcS) of coefficients that is determined by the choice of risk measure. Given an external sample of the coefficients, the UcS is a convex polytope that can be exactly calculated. The general uncertainty set is the Cartesian product of UcS corresponding to the individual constraints. However, we show that, under a natural assumption on the constraints (non-substitutability), the dimension of the general UcS reduces to the dimension of the UcS in the single-constraint case. We demonstrate an efficient algorithm that solves the SLP under spectral risk constraints. The solution is geometrically obtained by partially calculating the surface of the UcS.

CS10 Room R7: Ath. 4 CLUSTERING AND CLASSIFICATION II **Chair: Alfonso Gordaliza**

C194: **A new clustering approach based on cluster data feature maximization**
Presenter: **Jean-Charles Lamirel**, Synalp Team - LORIA, France

We present the IGNGF clustering method: a new incremental neural "winner-take-most" clustering method belonging to the family of the free topology neural clustering methods. Like other neural free topology methods such as Neural Gas (NG), Growing Neural Gas (GNG), or Incremental Growing Neural Gas (IGNG), the IGNGF method makes use of Hebbian learning for dynamically structuring the learning space. However, in contrast to these methods, the use of a standard distance measure for determining a winner is replaced in IGNGF by feature maximization. Feature maximization is a new cluster quality metric which associates each cluster with maximal features i.e., features whose Feature F-measure is maximal. Feature F-measure is the harmonic mean of Feature Recall and Feature Precision representing themselves new basic unsupervised cluster quality

measures. The paper details the operating mode of the IGNGF algorithm and illustrates that the method can outperform existing algorithms in the task of clustering of high dimensional heterogeneous data, whilst presenting, for the first time in the domain, a genuine incremental behavior.

C249: **Clusterwise non-negative matrix factorization (NMF) for capturing variability in time profiles**
Presenter: **Joke Heylen**, Katholieke Universiteit Leuven, Belgium
Co-authors: Eva Ceulemans, Iven Van Mechelen, Philippe Verduyn

In many domains, researchers are interested in capturing variability in time profiles. For example in emotion research, the time dynamics of emotions is a hot topic; hence, researchers recently have started gathering data on the intensity of different emotion components (e.g., appraisals, physiological features, subjective experience) at several time points during an emotion episode. To capture the inter- and/or intra–individual variability in such profiles, one can use functional component analysis or K-means clustering. Both strategies have some advantages but also some drawbacks. We propose a new method that combines the attractive features of these two strategies: Clusterwise Non-negative Matrix Factorization (NMF). This method assigns observations (e.g., emotional episodes, persons) into clusters according to the associated time profiles. The profiles within each cluster are decomposed into a general profile and an intensity score per observation that indicates the intensity of the general profile for specific observations. As Clusterwise NMF model is closely related to Mixtures of Factor Analyzers, we will discuss the similarities and differences of both methods. Finally, we will apply Clusterwise NMF to intensity profiles of emotional episodes.

C289: **Optimal cut finding for the hierarchical clustering using background information**
Presenter: **Askar Obulkasim**, VU University medical center, Netherlands
Co-authors: Mark van de Wiel, Gerrit Meijer

In hierarchical clustering, clusters are defined as branches of a tree. The constant height branch cut, a commonly used method to identify branches of a cluster tree, may not be ideal for cluster identification in complicated dendrograms. We introduce a novel method called piecewise-cut which uses main data type and available background information to find an optimal cutting point in the dendrogram. The crux of our main idea as follows: construct hierarchical clustering dendrogram with given molecular data. The resulting dendrogram is likely to express the tendency of samples are formed clusters guided by their molecular signatures. We presumed that besides molecular data, we may have the clinical information of the same patients (e.g. patient's survival time). Use this clinical data as background information, search through the cluster space which is composed of all possible partitions in the dendrogram obtained from molecular data. Find a partition in which every cluster is composed of patients which are homogenous in terms of their background information. Resulting clusters represents the cluster structure contained in the molecular data and easy to interpret. We applied our approach to many publicly available molecular data for which patients follow up information is also available. Experimental results show that our approach able to find a clustering structure which can not be obtained by commonly used constant height cutting approach.

C248: **Cluster analysis of age-grouped suicide data in Japan with spatio-temporal structure**
Presenter: **Takafumi Kubota**, The Institute of Statistical Mathematics, Japan
Co-authors: Makoto Tomita, Fumio Ishioka, Toshiharu Fujita

Cluster analysis was applied to the spatio-temporal suicide data in Japan. The objective of the analysis was to detect the characteristics of clusters and to compare the results among age groups. The data is divided in terms of area, year, age group and sex. The cluster analysis was applied by the following procedure. First, the *silhouette width* was used to detect the appropriate number of clusters. Second, a hierarchical cluster analysis was applied to the data. Third, a nonhierarchical cluster analysis, k-means clustering, was applied to the data. Also, the meanings of the clusters, which were determined by the results, are explained by the background in Japan. Finally, the contrasts are discussed by comparing the initial terms of area, year, age group and sex.

CS22 Room R5: Ath.1+2 HIGH-DIMENSIONAL DATA ANALYSIS II Chair: Maria-Pia Victoria-Feser

C135: **Model selection for high-dimensional data using the R package robustHD**
Presenter: **Andreas Alfons**, KU Leuven, Belgium

In applied data analysis, there is an increasing availability of data sets with a large number of variables, often larger than the number of observations. Variable selection for high-dimensional data allows to (i) overcome computational problems, (ii) improve prediction performance by variance reduction, and (iii) increase interpretability of the resulting models due to the smaller number of variables. However, robust procedures are necessary to prevent outlying data points from affecting the results. The R package robustHD contains functionality for robust linear model selection with complex high-dimensional data. The implemented robust methods are thereby based on least angle regression and sparse regression. Functionality is not limited to selecting individual variables, but also includes methods for robust groupwise variable selection, for instance groups of dummy variables representing categorical variables, or present and lagged values of time series data. The package follows a clear object-oriented design and parts of the code are written in C++ for fast computation. In addition, cross-validation functionality and various plots to select and evaluate the final model are available in robustHD.

C117: **Evaluation of microarray classification studies: Factors influencing the classification performance**
Presenter: **Putri Novianti**, University Medical Center Utrecht, Netherlands
Co-authors: Kit Roes, Marinus Eijkemans

Supervised methods used in microarray studies for gene expression are diverse in the way they deal with the underlying complexity of the data, as well as in the technique used to build the classification model and various classification performance measures are used to evaluate the model. The MAQC II study on cancer classification problems found that the variability of the performances was affected by factors such as the classification algorithm, cross validation method, number of genes, and gene selection method. We focus on evaluating these factors on the classification performance for non-cancer medical studies. Additionally, the effect of sample size, class imbalance, medical question (diagnostic, prognostic or treatment response), microarray platform, and disease type is evaluated. A systematic literature review was used to gather the information from 37 published studies. Pairwise associations between medical question, microarray color system, cross validation and gene selection method suggested that confounding at the study level plays a role. The impact of the various predictive factors on the reported classification accuracy was analyzed through random-intercept logistic regression, which always contained the degree of class imbalance as a predictive factor. The method of cross validation and medical question dominated the explained variation in the accuracy among studies, followed by gene selection method and microarray platform. In total, 63.32% of the between study variation was explained by these factors. In contrast to our expectation, disease category had a small impact on the classification performance. The accuracy of the classification models based on gene expression microarrays depends on study specific and problem specific factors.

C130: **Combining clustering of variables and random forests for high-dimensional supervised classification**
Presenter: **Robin Genuer**, University of Bordeaux, France
Co-authors: Marie Chavent, Jerome Saracco

The main goal is to tackle the problem of dimension reduction for high-dimensional supervised classification. The motivation is to handle gene expression data, for which the number p of variables is much larger than the number n of observations. The proposed method works in 3 steps. First, one eliminates redundancy using clustering of variables, based on the R-package ClustOfVar. Second, the most important synthetic variables

(summarizing the clusters obtained at the first step) are selected using an automatic procedure based on random forests. Third, a set of important variables is selected within each previously selected cluster. We stress that the two first steps reduce the dimension and give linear combinations of original variables (synthetic variables), while third step gives a set of relevant original variables. At each step, a supervised classification method (e.g. random forests, LDA, logistic regression) is used to build a predictor. Numerical performances of the predictors are compared on a simulated dataset and a real gene expression dataset for which n is about 100 and p about 7000.

C084: **Robust VIF Regression for high dimentional datasets**
Presenter: **Maria-Pia Victoria-Feser**, University of Geneva, Switzerland
Co-authors: Debbie Dupuis

The sophisticated and automated means of data collection used by an increasing number of institutions and companies leads to extremely large datasets. Subset selection in regression is essential when a huge number of covariates can potentially explain a response variable of interest. The recent statistical literature has seen an emergence of new selection methods that provide some type of compromise between implementation (computational speed) and statistical optimality (e.g. prediction error minimization). Global methods such as Mallows' C_p have been supplanted by sequential methods such as stepwise regression. More recently, streamwise regression, faster than the former, has emerged. A recently proposed streamwise regression approach based on the variance inflation factor (VIF) is promising but its least-squares based implementation makes it susceptible to the outliers inevitable in such large data sets. This lack of robustness can lead to poor and suboptimal feature selection. This article proposes a robust VIF regression, based on fast robust estimators, that inherits all the good properties of classical VIF in the absence of outliers, but also continues to perform well in their presence where the classical approach fails. The analysis of two real data sets shows the necessity of a robust approach for policy makers.

CS37 Room R6: Ath. 3 COMPUTATIONAL ECONOMETRICS III Chair: Simona Sanfelici

C241: **Improving model averaging in credit risk analysis**
Presenter: **Silvia Figini**, University of Pavia, Italy
Co-authors: Paolo Giudici

Classical and Bayesian Model Averaging (BMA) models for logistic regression in credit risk are compared. We also investigate regression models for rare event data, using Generalised Extreme Value regression. On the basis of a real data set, we show that Bayesian Model Averaging models outperform classical regression in terms of percentage of correct classifications and related performance indicators. We show that Bayesian Model Averaging is a technique designed to account for the uncertainty inherent in the model selection process, something which traditional statistical analysis often neglects. In credit risk, by averaging over many different competing models, BMA incorporates model uncertainty into conclusions about parameters and prediction. We also report the empirical evidence achieved on a real data sample containing rare events provided by a rating agency.

C239: **State dependence and the determination of sovereign credit ratings: Evidence from a panel of countries 2000-2010**
Presenter: **Stefanos Dimitrakopoulos**, University of Warwick, UK
Co-authors: Michalis Kolossiatis

It is analyzed whether time dependence in sovereign credit ratings could arise due to previous ratings or due to country-related unobserved components that are correlated over time. We set out to identify the sources and the degree of persistence in the determination of sovereign credit ratings using data from Moody's and Fitch for a set of countries covering the period 2000-2010. In particular, we disentangle two possible sources of inertia in rating agencies' decisions; the true state dependence and the spurious state dependence. We use an ordered probit model that controls for rating history via lagged dummies for each rating category in the previous period and unobserved heterogeneity via a sovereign-specific random effect. We name this model, which is new to the empirical literature in question, dynamic random effects panel ordered probit model. A nonparametric structure, based on the Dirichlet process, for the random effects is assumed while we also address the initial conditions problem. A Markov Chain Monte Carlo (MCMC) algorithm is developed to estimate the proposed model. Moreover, the proposed methodology itself advances the Bayesian literature on dynamic ordered probit models. Last, we compare our model against an alternative ordered probit model that has previously been used in the literature of sovereign credit ratings.

C058: **Understanding exchange rates dynamics**
Presenter: **Peter Martey Addo**, University of Paris 1, France
Co-authors: Dominique Guegan, Monica Billio

With the emergence of the chaos theory and the method of surrogates data, non-linear approaches employed in analysing time series typically suffer from high computational complexity and lack of straightforward explanation. Therefore, the need for methods capable of characterising time series in terms of their linear, non-linear, deterministic and stochastic nature are preferable. We provide a signal modality analysis on a variety of exchange rates. The analysis is achieved by using the recently proposed 'delay vector variance' (DVV) method, which examines local predictability of a signal in the phase space to detect the presence of determinism and non-linearity in a time series. Optimal embedding parameters used in the DVV analysis are obtain via a differential entropy based method using wavelet-based surrogates. A comprehensive analysis of the feasibility of this approach is provided. The empirical results show that the DVV method can be opted as an alternative way to understanding exchange rates dynamics.

C399: **Portfolio optimization algorithm based on data analysis and mean-risk models**
Presenter: **Florentin Serban**, Bucharest University of Economic Studies, Romania
Co-authors: Maria Viorica Stefanescu, Mihail Busu

An algorithm for solving portfolio optimization problems is proposed. First mean-variance model is compared with mean–Value-at-Risk (mean-VaR) model and the link between the mean-variance efficient set and the mean-VaR efficient set is investigated. Then two portfolio optimization approaches are analyzed. The first one is a two-stage portfolio optimization approach using, in order, both mean-variance and mean-VaR approaches. The second is a general mean-variance-VaR approach, using both variance and VaR as a double-risk measure simultaneously. Finally the case of an equity portfolio at the Italian Stock Market is considered. Data analysis techniques are used for portfolio selection, then risk estimation is performed for each stock and the portfolio optimization problem in the mean – VaR framework is solved.

PS01 Room Athenaeum Terrace POSTER SESSION I Chair: Patricia Roman-Roman

C061: **Stochastic gamma diffusion process with exogenous factors: Application to a real case**
Presenter: **Eva Maria Ramos-Abalos**, Universidad de Granada, Spain
Co-authors: Ramon Gutierrez-Jaimez, Ramon Gutierrez-Sanchez, Ahmed Nafidi

The aim is to study a new extension of the one-dimensional stochastic gamma diffusion process by introducing external time functions as exogenous factors. This original process can be used profitably in a variety of disciplines including Biology, Geology, Agriculture, Environmetrics and Population Dynamics. Firstly, we determine the probabilistic characteristics of the studied process as the analytical expression of the process,

its transition probability density function and their trend functions. Secondly, we study the statistical inference in this process: we estimate the parameters present in the model by using the maximum likelihood estimation method in basis of the discrete sampling of the process, thus obtaining the expression of the likelihood estimators. Finally, we applied this methodology to the real case, Gross Domestic Product (GDP) and CO^2 emissions from the consumption and flaring of natural gas in Spain. This implementation is carried out on the basis of annual observations of the variables over the period 1986-2009.

C076: **Asymptotic expansions for the ability estimator in item response theory**
Presenter: **Haruhiko Ogasawara**, Otaru University of Commerce, Japan

Asymptotic approximations to the distributions of the ability estimator and its transformations in item response theory are derived beyond the usual normal one when associated item parameters are given as in tailored testing. For the approximations, the asymptotic cumulants of the estimators up to the fourth order with the higher-order asymptotic variances are obtained under possible model misspecification. For testing and interval estimation of abilities, the asymptotic cumulants of the pivots studentized in four ways are derived. Numerical examples with simulations including those for confidence intervals for abilities are given using the three-parameter logistic model.

C054: **On testing multi-directional hypotheses in categorical data analysis**
Presenter: **Grzegorz Konczak**, University of Economics in Katowice, Poland

The problem of homogeneity data in r categorical population is analyzed. If the data are homogenous, the proportions of the observations in the j^{th} category will be equal in all samples. In the case 2 x 2 contingency tables it is possible to employ a one-tailed alternative hypothesis. In homogeneity testing, where in the contingency table the number of rows or columns is greater than two, it is possible to run a multi-tailed test. The proposal of the test multi-directional hypothesis is presented. This method is based on a permutation test. The procedure for homogeneity testing in two-way contingency tables is described. The properties of this test are analyzed in the Monte Carlo study.

C080: **Time series of macroeconomic indicators for the Czech Republic**
Presenter: **Jaroslav Sixta**, University of Economics, Czech Republic
Co-authors: Jakub Fischer

Long macroeconomic time series are very often required by economists and statisticians. Indicators like gross domestic product, household consumption or capital formation are widely used for different types of models. Even in the European Union, long time series are available for some countries only. A bit more complicated situation is for post-communist countries like the Czech Republic. Originally used Material Product System was different to currently used National Accounts. It means that official Czech figures starts in 1990 and before that there are no data available. During the communism, there were done estimates of gross domestic product for Czechoslovakia only. Moreover, these estimates were based on an old standard of national accounts. The transformation of original figures on Czech national income into gross domestic product is considered. The Czech statistics during communism had a good quality and all the units were fully covered by questionnaires. That time no sample surveys were used. The limits of this statistics were given by the concepts because it was focused on material production and supportive services. It means that the fast increase of significance of services like telecommunications in 1980s was omitted.

C086: **Prediction for negative binomial time series models**
Presenter: **Vasiliki Christou**, University of Cyprus, Cyprus
Co-authors: Konstantinos Fokianos

We consider tools for the evaluation of the predictive performance and the assessment of the assumed distributional assumptions based either on negative binomial or Poisson distribution. We strive to maximize the sharpness of the predictive distribution subject to calibration. We propose the use of a nonrandomized version of the probability integral transformation histogram to assess the probabilistic calibration, while for the assessment of the marginal calibration we consider the use of the marginal calibration plot. Addressing sharpness can be made via scoring rules, that is negatively oriented penalties that the forecaster wishes to minimize. The diagnostic approach is illustrated by an evaluation and ranking of the two competing forecasters for the transactions data of the stock ericsson B, measles data in Sheffield and breech births data in a hospital of South Africa.

C087: **Robust inference for count time series**
Presenter: **Stella Kitromilidou**, University of Cyprus, Cyprus

A log-linear Poisson model for count time series is considered. We study it under three forms of interventions: an Additive Outlier (AO), a Transient Shift (TS)and a Level Shift (LS). We estimate the parameters using the Maximum Likelihood Estimator (MLE), the Conditionally Unbiased Bounded-Influence estimator (CUBIF), and the Mallows Quasi-Likelihood estimator (MQLE), and compare all three estimators in terms of their MSE, bias and MAD. Our empirical results show that under a level shift, a transient shift or a combination of the two there are no significant differences between the three estimators. However, in the presence of large additive outliers, CUBIF dominates the other estimators.

C090: **On the analysis of high-dimensional data in immunology**
Presenter: **Ene Kaarik**, University of Tartu, Estonia
Co-authors: Tatjana von Rosen, Anne Selart

Peptide micro-arrays have become increasingly accessible in recent years. With the huge volumes of data being generated, appropriate statistical methods are crucial for extracting valid information concerning biological processes in the immune system. The immunological studies produce raw data that is not only voluminous but is typically highly skewed with artifacts due to technical issues, thus requiring appropriate cleaning, transformation and standardization prior to statistical analysis. Several results concerning methodology for analyzing immunological data will be presented. Among others, a new procedure for data pre-processing is proposed. A blocking/clustering strategy is worked out in order to detect informative sets of variables in a high-dimensional statistical setting. These results are essential when performing profile analysis when studying the behavior of the immune system over time. The obtained results will be illustrated on real data.

C081: **Bayesian analysis of a skewed exponential power distribution**
Presenter: **Lizbeth Naranjo**, University of Extremadura, Spain
Co-authors: Carlos J Perez, Jacinto Martin

A Bayesian analysis of a skewed exponential power distribution has been performed. The skewed exponential power family includes the symmetric exponential power distribution as a particular case and provides flexible distributions with lighter and heavier tails compared to the normal one. The distributions in this family can successfully handle both symmetry/asymmetry and light/heavy tails simultaneously. Even more, the distributions can fit each tail separately. The computational difficulties of tackling the posterior distribution have been avoided by proposing a scale mixture of uniforms representation. This representation has allowed the derivation of an efficient Gibbs sampling algorithm. The proposed approach represents a viable alternative to analyze data providing flexible fits.

C394: **Random walk approach to a prime factor of every odd perfect number which exceeds** 10^9
Presenter: **Ryuichi Sawae**, Okayama University of Science, Japan
Co-authors: Yoshiyuki Mori, Miho Aoki, Ishii Daisuke

A positive integer n is said to be perfect if $\sigma(n) = 2n$, where $\sigma(n)$ denotes the sum of positive divisors of n. At present, forty seven even perfect numbers are known, however, it is still open whether or not odd one does exist. Many necessary conditions for their existence have been found. One of these is to search the largest prime factor of an odd perfect number. Recently , it has proved that every odd perfect number must be divisible by a prime greater than 10^8. The method was based on Jenkins's and others results that needed about 26000 hours for computing time. In order to prove that a prime factor of every odd perfect number which exceeds 109, we must search the acceptable values of the cyclotomic numbers $\Phi_r(p)$. Our idea is based on finding these acceptable values by random walks on prime numbers.

PS02 Room Athenaeum Terrace POSTER SESSION II Chair: Patricia Roman-Roman

C216: **A new model for time series with trend and seasonal components**
Presenter: **Norio Watanabe**, Chuo University, Japan

A new model is proposed for time series including trend and seasonal components based on a fuzzy trend model. A fuzzy trend model is a statistical model which is represented by fuzzy if-then rules. The original fuzzy trend model can be applied to such time series but the number of parameters will increase extremely when a seasonal component exists, since a trend and a seasonal component cannot be separated. On the other hand the proposed model is an additive model with trend and seasonal components. Estimation and identification problems are discussed and applicability of the proposed model is shown by simulation studies. A practical example is also given.

C223: **Forecasting fruit caliber by means of diffusion processes: An application to Valencia late oranges**
Presenter: **Patricia Roman-Roman**, Universidad de Granada, Spain
Co-authors: Francisco Torres-Ruiz

The stochastic modeling by means of diffusion processes to forecast fruit caliber is considered. In a first phase, a diffusion process that adequately fits the available data on the time of fruit growth, is constructed and estimated. Then the use of the probability transition distributions of the fitted process for allocating caliber to each fruit is proposed. Tables are constructed with the values that discriminate between calibers at the time of harvest, which allows us to make a prediction for each previous time instant. Finally, the mean conditional functions to predict the percentages of each size at the time of harvest are considered. A practical application to the caliber of Valencia late oranges by considering a modified variety of the Bertalanffy process is presented. Such a process is modified by including in its trend a time dependent function used to model the observed deviations in the data about the evolution of the diameter of oranges with respect to the trajectories of the original process.

C233: **Nonparametric classification of functions based on depth**
Presenter: **Stanislav Nagy**, Charles University Prague, Czech Republic

The classification task for data coming from certain subspaces of continuous functions will be discussed. The functions will be of noisy nature and no further assumptions about the distributions will be stated. Special attention will be paid to depth-based classification and its possible generalisations. Several established depth functional classifiers will be compared. The outcoming drawbacks of these methods will be fixed by considering the derivatives of the smoothed versions of functions, although the observations do not have to be differentiable itself. Thus, a new version of Fraiman-Muniz depth capable of measuring the centrality of a differentiable function is introduced. Its classification performance is compared to known classifiers and we show that proper derivative using in combination with DD-plot (depth–depth plot) techniques is a powerful tool not only for the classification of functional observations.

C349: **Selection algorithm in regression models. `FWDselect` package**
Presenter: **Marta Sestelo**, University of Vigo, Spain
Co-authors: Nora M. Villanueva, Javier Roca-Pardinas

In multiple regression models, when there is a large number (p) of explanatory variables which may or may not be relevant for making predictions about the response, it is useful to be able to reduce the model. To this end, it is necessary to determine the best subset or subsets of q ($q \leq p$) predictors which will establish the model or models with the best prediction capacity. We present a new approach to this problem, where we will try to predict a new emission pollution incident, but focusing our attention on the importance to know the best combinations of time instants to obtain the best prediction. The proposed method is a new forward stepwise-based selection procedure that selects a model containing a subset of variables according to an optimal criteria (cross-validation determination coefficient) and taking into account the computational cost. Additionally, bootstrap resampling techniques are used to implement tests capable of detecting whether significant effects of the unselected variables are present in the model. The developed procedure have been implemented in a `R` package.

Tuesday 28.08.2012 11:10 - 12:50 Parallel Session F

IS02 Room R1: Demetra STATISTICAL SOFTWARE IN R WITH APPLICATIONS Chair: Peter Filzmoser

C177: **The `tclust` R package: A flexible tool to perform robust cluster analysis**
Presenter: **Alfonso Gordaliza**, Universidad de Valladolid, Spain
Co-authors: Luis A. Garcia-Escudero, Carlos Matran, Agustin Mayo-Iscar

The `tclust` R package implements the TCLUST methodology. TCLUST is a flexible statistical methodology to perform Robust Cluster Analysis in the very general setting of clusters having heterogeneous elliptical scatter structures and in the presence of any kind of contamination. This methodology is based on the so-called principle of "impartial trimming" in the Cluster Analysis framework. The programmed algorithm delivers the TCLUST estimator which has nice asymptotic and robust properties. The package incorporates graphical exploratory tools to assist the users in making a sensible choice of the number of clusters and the level of contamination present in the data set, as well as to measure the strength of assignments of the individuals to the clusters. The package will incorporate soon, in the version available at CRAN, the possibility of addressing the Common Principal Components and the Fuzzy Cluster Analysis problems. Other extensions of the methodology will be discussed.

C204: **Robust statistics and R**
Presenter: **Matias Salibian-Barrera**, The University of British Columbia, Canada

The aim is to describe and illustrate several R packages that have been developed in recent years implementing different robust statistical methods. In particular the "robustbase" package will be described, but when appropriate other publicly available packages, like "rrcov", will be considered. Developing high quality implementations of robust statistical methods is a particularly delicate task. This is typically due to the high computational complexity of the corresponding algorithms and the difficult optimization problems involved. Furthermore, relatively simple extensions of robust procedures designed for one model to other very similar models often result in disappointing results. In addition to illustrating the capabilities of the packages mentioned above using real data sets, it will be stressed the importance that the relatively recent emergence of R, an open-source cross-platform statistical programming environment, has had in the dissemination of novel statistical methods, and robust techniques in particular. How these observations relate to achieving a higher degree of reproducibility in statistical research will be discussed.

C206: **Robust sparse PCA in R**
Presenter: **Peter Filzmoser**, Vienna University of Technology, Austria
Co-authors: Christophe Croux, Heinrich Fritz

A method for principal component analysis is proposed that is sparse and robust at the same time. The sparsity delivers principal components that have loadings on a small number of variables, making them easier to interpret. The robustness makes the analysis resistant to outlying observations. The principal components correspond to directions that maximize a robust measure of the variance, with an additional penalty term to take sparseness into account. We propose an algorithm to compute the sparse and robust principal components. The algorithm computes the components sequentially, and thus it can handle data sets with more variables than observations. The method is implemented in the R package pcaPP as function sPCAgrid. Diagnostic plots for detecting outliers and for selecting the degree of sparsity are provided.

OS12 Room R5: Ath.1+2 VARIABLE SELECTION AND FEATURE EXTRACTION IN PREDICTIVE MODELING Chair: Laura Trinchera

C062: **Embedded variable selection in classification trees**
Presenter: **Servane Gey**, University Paris Descartes, France
Co-authors: Tristan Mary-Huard

The problems of model and variable selections for classification trees are jointly considered. A penalized criterion is proposed which explicitly takes into account the number of variables, and a risk bound inequality is provided for the tree classifier minimizing this criterion. This penalized criterion is compared to the one used during the pruning step of the CART algorithm. It is shown that the two criteria are similar under some specific margin assumption. In practice, the tuning parameter of the CART penalty has to be calibrated by hold-out. Simulation studies are performed which confirm that the hold-out procedure mimics the form of the proposed penalized criterion.

C108: **A new criterion for sparse PLS regression**
Presenter: **Laura Trinchera**, AgroParisTech and INRA-UMRMIA518, France
Co-authors: Tzu-Yu Liu, Arthur Tenenhaus, Dennis Wei, Alfred Hero

Partial least squares (PLS) regression combines dimensionality reduction and prediction using a latent variable model. It provides better predictive ability than principle component analysis by taking into account both the independent and response variables in the dimension reduction procedure. However, PLS suffers from over-fitting problems for few samples but many variables. We formulate a new criterion for sparse PLS by adding a structured sparsity constraint to the global SIMPLS optimization. The constraint is a sparsity-inducing norm, which is useful for selecting the important variables shared among all the components. The optimization is solved by an augmented Lagrangian method to obtain the PLS components and to perform variable selection simultaneously. We propose a novel greedy algorithm to overcome the computation difficulties. Experiments demonstrate that our approach to PLS regression attains better performance with fewer selected predictors.

C164: **Kernel discrimination and explicative features: An operative approach**
Presenter: **Caterina Liberati**, University of Milano-Bicocca, Italy
Co-authors: Furio Camillo, Gilbert Saporta

Kernel-based methods such as SVMs and LS-SVMs have been successfully used for solving various supervised classification and pattern recognition problems in machine learning. Unfortunately, they are heavily dependent on the choice of the optimal kernel function and from tuning parameters. Their solutions, in fact, suffer of complete lack of interpretation in terms of input variables. That is not a banal problem, especially when the learning task is related with a critical asset of a business, like credit scoring, where deriving a classification rule has to respect an international regulation (Basel II-III). The following strategy is proposed for solving problems using categorical predictors: replace the predictors by components issued from MCA, choice of the best kernel among several ones (linear, RBF, Laplace, Cauchy, etc.), approximation of the classifier through a linear model. The loss of performance due to such approximation is balanced by a better interpretability for the end user, employed in order to understand and to rank the influence of each category of the variables set in the prediction. This strategy has been applied to real risk-credit data of small enterprises. Cauchy kernel was found the best and leads to a score much more efficient than classical ones, even after approximations.

C224: **Categorial effect modifiers in generalized linear models**
Presenter: **Margret-Ruth Oelker**, LMU Munich, Germany
Co-authors: Jan Gertheiss, Gerhard Tutz

Varying-coefficient models are efficient tools to model interactions of covariates. We focus on the rarely treated case of categorical effect modifiers. Lasso-type regularization techniques are proposed that work for nominal and ordinal effect modifiers within the framework of generalized linear

models. They allow us to select predictors and to identify categories that have to be distinguished with respect to the response variable. Large sample properties are investigated; in simulation studies and a real data example, it is demonstrated that the method works well and is a strong tool to reduce the complexity of the predictor.

OS07 Room R2: Ares ADVANCES IN COMPUTATIONAL ECONOMETRICS Chair: Erricos Kontoghiorghes

C063: **Firm's volatility risk under microstructure noise**
Presenter: **Simona Sanfelici**, University of Parma, Italy
Co-authors: Flavia Barsotti

Equity returns and credit spreads are strictly interrelated financial variables capturing the credit risk profile of a firm. We consider CDS premium as direct measure of credit spreads and use high-frequency equity prices in order to estimate the volatility (and jump) risk component of a firm. We follow a structural credit risk modeling approach and we consider a general framework by introducing microstructure noise. The aim is to explain the CDS premium by identifying the volatility risk component of a firm through a nonparametric volatility estimator based on Fourier analysis, which is robust to microstructure noise.

C166: **Filtering with heavy tails**
Presenter: **Alessandra Luati**, University of Bologna, Italy
Co-authors: Andrew Harvey

The problem of modeling a time varying signal embedded in non-Gaussian noise is addressed. We consider a class of observation driven models where, conditional on past information, the observations are generated by a Student-*t* distribution. The dynamics of the time varying location are governed by the score of the conditional distribution and the time varying signal is updated by a suitably defined filter with ARMA-type structure. This filter may be interpreted as an approximation to a computer intensive solution for the corresponding unobserved component models. The attraction of regarding it as a model in its own right is that it allows us to derive the asymptotic distribution of the maximum likelihood estimator. We provide illustrations based on simulated and real data.

C212: **Bayesian evaluation of quantile forecasts**
Presenter: **Cathy WS Chen**, Feng Chia University, Taiwan
Co-authors: Edward Lin, Richard Gerlach

Many forecasting VaR (Value-at-Risk) measures have been proposed in a Bayesian framework. However, when dealing with back-testing, most jump outside the Bayesian framework, or use informal criteria, to compare VaR models. An important issue that arises in this context is how to evaluate the performance of VaR models/methods. It is desirable to have formal testing procedures for comparison, which do not necessarily require knowledge of the underlying model, or if the model is known, do not restrict attention to a specific estimation procedure. The motivation of the study is to propose a back-testing to evaluate VaR models via quantile regression, which does not rely solely on binary variables like violations. Using bivariate quadrature methods and the standard asymmetric Laplace quantile likelihood, we analytically estimate the Bayes factor in favour of the proposed model's forecasts being accurate and independent. For the empirical study, we compare the performance of the proposed method, and another three non-Bayesian back-testing methods, to figure out which back-tests are the most reliable, and most suitable for model validation processes. The proposed Bayesian tests provide sound methods to assess the finite sample performance of a quantile model.

C320: **Neglecting structural breaks in DCC models**
Presenter: **Christos Savva**, Cyprus University of Technology, Cyprus
Co-authors: Andreea Halunga

Parameter regime changes in dynamic conditional correlation (DCC) model and corrected DCC (cDCC) are considered. It is shown that if these parameter regime changes are not accounted for in estimations, it causes the sum of the estimated short and long run persistence parameters to converge to one. This is usually taken as evidence of high persistence in conditional correlations, but in the case of neglected parameter changes, it leads to spurious inferences. Results suggest that not accounting for shifts in the unconditional correlations of the stochastic processes may bias upward short and long run estimates of persistence in correlation. Thus, they contaminate the usefulness of the DCC models in situation in which the degree of permanence is important.

CS04 Room R4: Aph.+Pos. ADVANCES IN DATA ANALYSIS Chair: Patrick Groenen

C128: **Evaluating the performance of a Bayesian-multiplicative treatment of zeros in compositional count data sets**
Presenter: **Josep Antoni Martin-Fernandez**, University of Girona, Spain
Co-authors: Karel Hron, Matthias Templ, Peter Filzmoser, Javier Palarea-Albaladejo

Counts data are discrete vectors representing the numbers of outcomes falling into each of several mutually exclusive categories. Compositional techniques based on log-ratio methodology are appropriate in those cases where the total sum of the vector is not of interest. Such compositional count data sets usually contain zero values resulting from insufficiently large samples. That is, they refer to unobserved positive values that may have been observed with a larger number of trials. Because the log-ratio transformations require data with positive values, any statistical analysis of count compositions must be preceded by a proper replacement of the zeros. In recent years, a Bayesian-multiplicative approach, coherent with the vector space structure of the simplex, has been proposed for addressing the "zero problem". This treatment involves the Dirichlet prior distribution as the conjugate distribution of the multinomial and a multiplicative modification of the non-zero values in the vector of counts. Different parameterizations of the prior distribution provide different zero replacement results. Their performance will be evaluated from both theoretic and computational points of view.

C161: **Bivariate mixture with heavy tails along 1D projections**
Presenter: **Julie Carreau**, Institut de Recherche pour le Developpement, France
Co-authors: Philippe Naveau, Malaak Kallache

Multivariate heavy-tailed data occur in numerous fields such as finance and climate where extreme events can have dramatic consequences. The dependence structure of multivariate extremes is non-parametric and its estimation is challenging. Standard ways to proceed involve defining multivariate extremes by establishing a threshold or by taking componentwise maxima. This way, only a subset of the multivariate density is estimated. The authors propose a bivariate density estimator which addresses the need to simultaneously model central and extremal data. The estimator could be extended to a higher dimensional setting. Heavy-tailed bivariate data is modelled with a mixture of bivariate Gaussians which have a heavy tail along a one-dimensional (1D) projection. This estimator is non-parametric in its central part. The extremal dependence structure is assumed to have the following shape: extremes occur in a discrete number of directions given by the 1D projections of the mixture components. An algorithm is developed which first computes an estimate of the extremal dependence structure and then provides good initial values of the mixture parameters. This algorithm is evaluated on synthetic data with known discrete extremal directions. The proposed mixture is then compared with a mixture of bivariate Gaussians on heavy-tailed rainfall data.

C257: **Missing values imputation for mixed data based on principal component methods**
Presenter: **Francois Husson**, Agrocampus Rennes, France
Co-authors: Vincent Audigier, Julie Josse

Several methods are available to deal with missing values in exploratory multivariate data analysis, such as regularized iterative PCA for continuous variables or regularized iterative multiple correspondence analysis for categorical variables. We extend the methodology to deal with missing values in a principal component method dedicated to both types of variables (mixed data). To estimate the scores and the loadings from an incomplete dataset, we use an EM-type algorithm based on iterative singular value decompositions on a particular matrix. One property of this algorithm is that an imputation of the missing values is performed during the estimation step. Consequently, this method can be seen as a new single imputation method for mixed-data. This is appealing since many imputation methods are restricted to only one type of variables. The imputation takes into account the similarities between individuals and the relationships between all the variables (continuous and categorical). After studying the properties of this new method, we assess its performances with benchmark datasets and simulations. Regarding the quality of the prediction of the missing values, the results are promising compared to other methods such as an imputation method based on a random forest recently proposed in the literature.

C243: **Two-stage acceleration for non-linear PCA**
Presenter: **Yuichi Mori**, Okayama Universiy of Science, Japan
Co-authors: Masahiro Kuroda, Michio Sakakihara, Masaya Iizuka

Principal components analysis (PCA) is a descriptive multivariate method for analyzing quantitative data. For PCA of a mixture of quantitative and qualitative data, quantification of qualitative data requires obtaining optimal scaling data and using ordinary PCA. The extended PCA, including such quantification, is called non-linear PCA. Then, the alternating least squares (ALS) algorithm is used as the quantification method. However, the ALS algorithm for non-linear PCA of large data requires many iterations and much computation time due to its linear convergence. We provide a new acceleration method for the ALS algorithm using the vector ε (vε) and Graves-Morris (GM) algorithms. Both acceleration algorithms speed up the convergence of a linearly convergent sequence generated by the ALS algorithm. Acceleration of the ALS algorithm can be performed in two stages: 1) the vε algorithm generates an accelerated sequence of the ALS sequence and 2) the convergence of the vε accelerated sequence is accelerated using the GM algorithm. Thus, we expect that, by accelerating the convergence of the vε accelerated sequence, the GM algorithm improves the overall computational efficiency of the ALS algorithm. Numerical experiments examine the performance of the two-stage acceleration for non-linear PCA.

CS26 Room R7: Ath. 4 CLUSTERING AND CLASSIFICATION III Chair: Elise Dusseldorp

C325: **An overlapping clustering method for signed graphs**
Presenter: **Michiharu Kitano**, Graduate School of Doshisha University, Japan
Co-authors: Hiroshi Yadohisa

Signed graphs represent complex structures comprising both positive and negative relationships. In social network analysis using signed graphs, several methods for detecting communities, which are the groups within which positive relationships are dense and between which negative relationships are also dense, have been proposed to better understand complex structures. However, these methods do not allow for overlapping of communities even though it is dicult to model all communities in the real world as being disjoint. We propose a method for the detection of overlapping communities by splitting vertices which implies transforming a vertex into two vertices. For splitting vertices, we introduce a new index of centrality as the degree of mediation. With our approach, we can convert any method that detects disjoint communities into one that detects overlapping communities for signed graphs, and can clearly extract the mediation structure.

C285: **Stacking prediction for a binary outcome**
Presenter: **Charles Gomes**, L Oreal, France
Co-authors: Hicham Nocairi, Marie Thomas, Fabien Ibanez, Jean-Francois Collin, Gilbert Saporta

A large number of supervised classification models have been proposed in the literature. In order to avoid any bias induced by the use of one single statistical approach, they are combined through a specific stacking meta-model. To deal with the case of a binary outcome and of categorical predictors, we introduce several improvements to stacking: combining models is done through PLS-DA instead of OLS due to the strong correlation between predictions, and a specific methodology is developed for the case of a small number of observations, using repeated sub-sampling for variables selection. Five very different models (Boosting, Naïve Bayes, SVM, Sparse PLS-DA and Expert Scoring) are combined through this improved stacking, and applied in the context of the development of alternative strategies for safety evaluation where multiple *in vitro*, *in silico* and physico-chemical parameters are used to classify substances in two classes: Sensitizer and No Sensitizer. Results show that stacking meta-models have better performances than each of the five models taken separately, and furthermore, stacking provides a better balance between sensitivity and specificity.

C288: **Diagnostics and improvements on statistical sensitivity analysis in a subspace method**
Presenter: **Kuniyoshi Hayashi**, Okayama University, Japan
Co-authors: Fumio Ishioka, Hiroshi Suito, Koji Kurihara

We focus primarily on the CLAFIC method, which is a discriminant method for pattern recognition, and present the framework for diagnostics based on the influence function for this method. Furthermore, we suggest improvements and, through a simulation study, show the difference in performance in terms of prediction accuracy.

C228: **A new method for advanced subgroup analysis: QUINT**
Presenter: **Elise Dusseldorp**, TNO and Katholieke Universiteit Leuven, Netherlands
Co-authors: Iven Van Mechelen

When several treatments are available for some disease, for example A and B, it is often difficult to decide which treatment is best for an individual patient given his/her pretreatment characteristics. To support such a decision, empirical evidence is needed about which subgroup of patients displays a better outcome with treatment A than with B, and for which subgroup the reverse is true. This phenomenon is called a qualitative treatment-subgroup interaction. In case of data from randomized controlled trials with many patient characteristics that could interact with treatment in a complex way, a suitable statistical approach to identify qualitative treatment-subgroup interactions is not yet available. For this purpose, we propose a new method, called QUalitative INteraction Trees (QUINT). QUINT results in a binary tree that subdivides the patients into subgroups on the basis of pretreatment characteristics; these subgroups are assigned to one of three classes: a first one for which A is better than B, a second one for which B is better than A, and an optional third one for which type of treatment makes no difference. We will report the results of an extensive simulation study that was conducted to evaluate the optimization and recovery performance of QUINT.*quad*

CS27 Room R6: Ath. 3 MULTIVARIATE DATA ANALYSIS I Chair: Kohei Adachi

C315: **Exploring multivariate data via the DIVE system**
Presenter: **Mohit Dayal**, Indian School of Business, India

For the purpose of visualization, it is often useful to think of a linear projection of multivariate data, as viewing the high dimensional point cloud from a particular direction. Thus, if one utilizes several projections together, useful understanding about the features of the data may be gained. DIVE (Drawing Interactively for Visual Exploration) is a new system for multivariate data visualization based on this principle. Prime features of DIVE are its easy comprehensibility and extremely high level of interactivity, unsurpassed by any such system for multivariate data. DIVE embeds the multiple view paradigm by allowing the user to display an arbitrary number of (static) data projections that may be selected on any criteria. Next, a rich variety of point manipulation tools are presented, that essentially allow the analyst to specify the kind of projections that she wants to see next. This is akin to specifying a composite hypothesis about the features of the high dimensional point cloud, which the system then attempts to validate. Internally, these manipulation controls are implemented in the unified context exploratory projection pursuit framework. We contrast this static, high interaction approach, to that of dynamic graphics, like data tours, which automate the projection selection.

C341: **Three-way multidimensional scaling of percentile-velued dissimilarities with the non-concentric hyperbox model**
Presenter: **Yoshikazu Terada**, Osaka University, Japan
Co-authors: Hiroshi Yadohisa

The purpose of multidimensional scaling is to embedding the objects into a low dimensional space, in such a way that the given dissimilarities are approximated by the distances in the configuration. In symbolic data analysis, a dissimilarity between two objects can be described by not only a single value, but also an interval, a distribution, and so on. An interval-valued dissimilarity often consists of the maximum and minimum values. However, in some cases, maximum and minimum values are not stable. Therefore, percentile-valued data, which consists of nested percentile intervals, is sometime better than interval-valued one. In our previous study, MDS algorithm for one-mode two-way percentile-valued dissimilarity data, called Percen-Scal, have been proposed. Each object is represented by the non-concentric nested hyperboxes in this model. A new MDS model for two-mode three-way percentile-valued dissimilarity data is proposed. This model is based on the weighted Euclidean model and considered as the natural extension of 3WaySym-Scal. In this model, each object is also represented by non-concentric nested hyperboxes corresponding to percentile intervals in both common and individual spaces. Moreover, an efficient algorithm based on iterative majorization for this model, called 3WayPercen-Scal, is developed.

C149: **Variable selection in the context of multivariate process monitoring**
Presenter: **Luan Jaupi**, CNAM, France
Co-authors: Dariush Ghorbanzadeh, Philippe Durand, Dyah E. Herwindiati

Variable selection for multivariate process control is considered. A novel algorithm for determining a subset of variables that preserves, to some extent, the structure and information carried by all the original variables is proposed. This approach was motivated by the need to reduce the dimension of primary variables before carrying out a process monitoring. Such a requirement is more and more a feature of many application areas involving huge amounts of data, where it is useful for external utility information to influence the selection process. Utility information is attached to specific variables. For example, some measurements may be cheaper and easier to carry out than others. It is straightforward to incorporate such an information into a selection algorithm. The proposed method is a stepwise procedure. At each step, we select the most informative variable among the primary variables that have not yet been selected. The new variable is selected by its ability to supply complementary information for the whole set of variables. The method is tested and compared with previous variable selection methods in a simulation study. It is further illustrated with real data from the automotive industry.

C174: **The REBUS-PLS algorithm as a strategy for classifying regional labour markets**
Presenter: **Frank Pelzel**, Institute for Employment Research -IAB-, Germany

To overcome the problem of unobserved heterogeneity of regional labour markets, it is aspired to use the response based procedure for detecting unit segments in partial least squares path modeling (REBUS-PLS) to identify different region-types. The great advantage of REBUS-PLS is that the clustering of regions and the calculation of variables' weights occur simultaneously by means of an iterating mathematical algorithm. Furthermore, the estimations are based on a complex economic model developed on basis of theoretical macroeconomic considerations. By the use of this holistic, model-based approach, the bias resulting from separately estimated models is avoided and an objective selection and weighting procedure for the included variables is applied. Finally, using the structure of the goodness of fit index (GOF), the algorithm leads to local models for each segment with individual parameters for both structural and measurement model. This should increase the overall predictive performance and allow for a more precise cause-effect-analysis. As an application, German labor market regions, defined by their commuting linkages, are classified to unit segments. Afterwards the results from the REBUS-PLS algorithm are compared with alternative methods.

PS03 Room Athenaeum Terrace POSTER SESSION III Chair: Ana B. Ramos-Guajardo

C100: **Evaluation of annual number of deaths in Japanese hemophiliacs with HIV infection**
Presenter: **Takahiko Ueno**, St Marianna University School of Medicine, Japan
Co-authors: Shinobu Tatsunami, Junichi Mimaya, Akira Shirahata, Masashi Taki

Having passed more than 15 years since the introduction of protease inhibitor for the therapy of HIV infection, the life expectancy among people with HIV has been elongated. However, the risk for death among them is sometimes reported to be much higher than people without HIV. We tried comparison of annual number of death between Japanese hemophiliacs with Japanese population. The used the data by the research committee on the national surveillance on coagulation disorders in Japan. The Kaplan-Meyer plotting was used for the HIV-infected hemophiliacs' survival description after 1983. We applied the age dependent annual rate for death in Japanese population proposed by Statistics and Information Department, Minister's Secretariat, Ministry of Health, Labour and Welfare. Observed annual number of deaths was 12 in 2010, and was 14 in 2011. The expected numbers from Kaplan-Meyer method were 16 and 20, respectively. On the other hand, when we applied the Japanese population' rate dated in 2009, the estimated number of deaths in 2010 became 1.7. Therefore, it is suggested that Japanese HIV infected hemophiliacs are still exposed to higher risk for death compare to Japanese general population.

C101: **A model for understanding the importance of adherence to antiretroviral therapy**
Presenter: **Shinobu Tatsunami**, Saint Marianna University School of Medicine, Japan
Co-authors: Takahiko Ueno, Rie Kuwabara

Some of the specific aspects of disease development after infection with HIV can be described by viral dynamics due to the non-linearity of the equations. We tried to express the importance of adherence to antiretroviral therapy by using such kind of equation. Viral dynamics are fundamentally composed of three parts: viral replication, suppression by immune activity, and elimination by drugs. Therefore, the simplest dynamic equation is containing three main variables of viral concentration, magnitude of the immune activity, and concentration of antiviral drug. The most important assumption in the present formulation is that the probability of appearance of a drug-resistant virus depends on the time

derivative of viral concentration. Under this assumption, the dynamic equation can describe the divergence of viral concentration even after the viral concentration attains the lower detectable limit. For example, if the interruption of drug administration occurs on two successive days between two periods with perfect adherence, divergence of viral concentration within a month occurred 83 times in 1000 runs using different sequences of random numbers. The present model could demonstrate the fragility of antiretroviral therapy. This will be useful in explaining the importance of adherence to antiretroviral therapy.

C129: **Three-level analysis of raw single-case experimental data: A Monte Carlo simulation study**

Presenter: **Mariola Moeyaert**, Katholieke Universiteit Leuven, Belgium, Belgium

Co-authors: Maaike Ugille, John Ferron, Tasha Beretvas, Wim Van den Noortgate

The multilevel approach and its flexibility are appealing for combining single-case data within and across studies. To investigate under which realistic conditions the three-level model works to synthesize single-case data, a Monte Carlo simulation study is used to vary a subsequent set of parameters: the value of the immediate treatment effect and the treatment effect on a time trend, the number of studies, the number of cases, the number of measurements per case and the between-case and between-study variance. In order to evaluate the three-level approach, the bias and the mean squared error of the point estimates of the fixed effects and the corresponding standard errors, the coverage proportion of the confidence interval, the power for testing the fixed effects and the bias of the point estimates of the variance components is investigated. The simulation study shows that the three-level approach results in unbiased estimates of both kinds of treatment effects. In order to have a reasonable power for testing the treatment effects (.80 or higher), researcher are recommended to use strict inclusion criteria, resulting in a homogeneous set of studies, and to involve a minimum of 30 studies in their three-level analysis of single-case results.

C134: **The analysis of Latvia financial reference amount for state scientific institutions**

Presenter: **Irina Arhipova**, Latvia University of Agriculture, Latvia

The driving force of the future development of the Latvia economy will be increase of international economy competition. Its development temps will be determined by research-based production proportion in the national economics, as well as amount of high-tech in the export and innovation in all sectors of economics. For this purpose one of the main tasks is effective use of Latvia state budget resources for scientific activity: research grants schemes and the financial reference amount for state scientific institutions. The research funding instruments should always seek to measure scientific quality. At the same time it is important to analyze the factors that affect the results of scientific activities. It is necessary for the following improvement of existing research funding system as well to promote the research activities in research institutions. The analysis of Latvia financial reference amount for state scientific institutions was made using the last three years data. The casual relationships between the funding and development of research are evaluated using econometrical methods. The scientific institutions funding model was evaluated due the sensibility of research quality criteria. For the funding model development other quality criteria are considered to investigate the effects that research funding has on the research society.

C173: **An university efficiency evaluation by using its reputational component**

Presenter: **Vsevolod Petruschenko**, NRU HSE, Russia

Co-authors: Fuad Aleskerov, Irina Abankina, Veronica Belousova, Anastasia Bonch-Osmolovskaya, Vyacheslav Yakuba, Kirill Zinkovsky

Education makes contribution into economic growth. Therefore we estimate efficiency scores for Russian universities on the basis of a data set of input and output criteria by using Data Envelopment Analysis approach. One of the output parameters is a reputation index. To construct it 4000 contexts are analyzed and 11 reputation topics are found. The Copeland's method is used to aggregate them into a reputation indicator. In addition, which factors tend to make a university to be efficient are investigated.

C392: **Multistate analysis with data cleansing applied to marriage career**

Presenter: **Jana Langhamrova**, University of Economics Prague, Czech Republic

Co-authors: Martina Miskolczi, Jitka Langhamrova

Multistate analysis of demographic data concerning marriage carrier are made. Multistate mathematical demography arises from multiregional demographics method and combines transitions among states and shift in time in selected (sub)population. Here, statuses 'single', 'married', 'divorced', 'widowed' and 'dead' are used for the population of the Czech Republic women in 1990–2010. In the theoretical part, intensities of probability of transition and transition probabilities are derived. A special data cleaning procedure is presented both in a theoretical model and an application. Usually, crude nuptiality, divorce and mortality rate are used for estimation of intensities. Here, numerator and denominator in crude rates has been examined and adjusted in order to capture reality more accurately, for example some delay between marriage and divorce is applied. Final multistate life tables verify changes in the behaviour of women in the Czech Republic related to their marriage decision over last twenty years: women more often decide not to marry and stay unmarried; especially this tendency is stronger among young women under 30 years. Length of the time when a woman is married is stable (not necessarily in one marriage) as well as the time of being divorced or widowed.

C326: **A fast computational framework for genome-wide association studies with neuroimaging data**

Presenter: **Benoit Da Mota**, Microsoft Research - INRIA joint centre, France

Co-authors: Vincent Frouin, Edouard Duchesnay, Soizic Laguitton, Gael Varoquaux, Jean-Baptiste Poline, Bertrand Thirion

In the last few years, it has become possible to acquire high-dimensional neuroimaging and genetic data on relatively large cohorts of subjects, which provides novel means to understand the large between-subject variability observed in brain organization. Genetic association studies aim at unveiling correlations between the genetic variants and the numerous phenotypes extracted from brain images and thus face a dire multiple comparisons issue. While these statistics can be accumulated across the brain volume for the sake of sensitivity, the significance of the resulting summary statistics can only be assessed through permutations. Fortunately, the increase of computational power can be exploited, but this requires designing new parallel algorithms. The MapReduce framework coupled with efficient algorithms permits to deliver a scalable analysis tool that deals with high-dimensional data and thousands of permutations in a few hours. On a real functional MRI dataset, this tool shows promising results with a genetic variant that survives the very strict correction for multiple testing.

C405: **Multi-stage product development with exploration, value-enhancing, preemptive and innovation options**

Presenter: **Nicos Koussis**, Frederick University, Cyprus

Co-authors: Spiros Martzoukos

We provide a real options framework for the analysis of product development that incorporates research and exploration actions, product attribute value-enhancing actions with uncertain outcome, and preemption and innovation options. We derive two-stage analytic formulas and propose a general multi-period solution using a numerical lattice approach. Our analysis reveals that exploration actions are more important when the project is out or at-the-money (near zero NPV) and less important for high project values. In a multi-stage setting, exploration actions are important even for in-the-money projects when follow-on actions exist that can enhance the expected value of the project. With path-dependency, early actions are more valuable since they enhance the impact or reduce the cost of subsequent actions. By analyzing parameter estimates of jump diffusion models, we find that preemptive controls affecting rare event (jump) frequency and innovations that introduce positive jumps are more valuable for firms with higher frequency of competitive threats involving low volatility.

C342: **Fiscal policy sustainability, economic cycle and finanancial crises: The case of the GIPS**
Presenter: **Gabriella Legrenzi**, Keele University, United Kingdom
Co-authors: Costas Milas

The aim is to extend previous works on the sustainability of the government's intertemporal budget constraint by allowing for non-linear adjustment of the fiscal variables, conditional on (i) the sign of budgetary disequilibria and (ii) the phase of the economic cycle. Further, the endogenously estimated threshold for the non-linear adjustment is not fixed; instead it is allowed to vary over time and with the incidence of financial crises. This analysis presents particular interest within the current European economic scenario of financial crises, poor growth and debt crises, providing empirical evidence of a threshold behaviour for the GIPS, that only correct large unbalances (which, in the case of Greece and Portugal, are higher than the EGSP criteria). Financial crises further relax the threshold for adjustment: during financial crises, only very large budgetary unbalances are corrected.

PS04 Room Athenaeum Terrace POSTER SESSION IV Chair: Ana B. Ramos-Guajardo

C324: **The problem of yearly inflation rate**
Presenter: **Josef Arlt**, University of Economics, Prague, Czech Republic
Co-authors: Marketa Arltova

Inflation rate is important macroeconomic indicator, which measures the change in the general level of prices of goods and services. It plays a crucial role in monetary policy, it is used for the valorization of wages, pensions, social benefits etc. It is extremely important to find a good way to measure inflation, as well as a method for inflation forecasting. The yearly inflation rate is not the appropriate measure of inflation, mainly due to the fact that it does not provide up-to-date information on the level of inflation. The spectral time series analysis shows that the yearly inflation rate delays the information with respect to the monthly and annualized inflation rate approximately about six months. This conclusion is extremely important for the process of economic decision making. It leads also to the proposal of a new nontraditional method for the yearly inflation rate forecasting. Advantage of this method is that it is able to capture breaks and other instabilities in the future development of the inflation rate.

C363: **Calibration of three test forms using the factor analytic method under item response theory**
Presenter: **Sayaka Arai**, The national center for university entrance examinations, Japan
Co-authors: Shin-ichi Mayekawa

Latent trait theory is often referred to as item response theory (IRT) in the area of educational testing and psychological measurement. IRT methods are used in many testing applications, since these provide a major advantage: comparability of test scores from different test forms. To accomplish this, all parameters of the test items need to be put onto a common metric, namely, item parameter calibration. Although several calibration methods have been proposed, the factor analytic method has not been well-examined. A comparison study of the calibration of two forms showed that both the factor analytic method, and fixed common item parameter method performed well. However, there is no study on the calibration of more than two forms using the factor analytic method. In the calibration of three models, there are three linking patterns. A simulation study considering the linking patterns is done and its performance is compared to that of other methods.

C189: **Email categorization and spam filtering by random forest with new classes of quantitative profiles**
Presenter: **Marian Grendar**, Slovanet, Slovakia
Co-authors: Jana Skutova, Vladimir Spitalsky

Traditional content-based approaches to spam filtering and email categorization are based on heuristic rules, naive Bayes filtering and text-mining methods. In the quantitative profiles (QP) approach, an email is represented by a p-dimensional vector of numbers. Several classes of QPs are introduced: grouped character profile, histogram binary profile, moving window profile, which complement the already considered character and binary profiles. A QP is processed by Random Forest (RF). At low computational costs the QP-based RF classifiers attain in spam filtering comparable and in email categorization even better performance than the optimized SpamAssassin and Bogofilter. Since majority of the proposed QPs have similar performance, vulnerability of the resulting email filter can be lowered by applying the batch of QPs, or by predicting the category of a new email by a randomly selected QP-based RF classifier.

C375: **Comparing distributions using dependent normalized random measure mixtures**
Presenter: **Michalis Kolossiatis**, Cyprus University of Technology, Cyprus
Co-authors: Jim Griffin, Mark Steel

An area of Bayesian statistics that has attracted much attention recently is nonparametric modelling. One important reason for this is the advancement of various simulation methods, and especially Markov chain Monte Carlo methods. These models are particularly useful for flexible modelling of various data sets. This is usually done by modelling the underlying distributions of the data. For example, it is assumed that the distribution of the data in each group consist of a common component, shared by all distributions, and an idiosyncratic component. Dependence between the groups is induced by the common part. A methodology for the simultaneous Bayesian nonparametric modelling of several dependent distributions is described. The approach makes use of normalised random measures with independent increments and builds dependence through the superposition of shared processes. Theoretical properties of these priors are described, as well as the different modelling possibilities. Regarding posterior analysis, efficient slice sampling simulation methods are developed. A new graphical method for comparing pairs of distributions, which allows better understanding of the differences between the dependent distributions, is also described. The proposed models are finally applied to a stochastic frontier setting and used to analyse the efficiency of some hospital firms, with interesting results.

C376: **An extraction approach of latent knowledge in collective text data**
Presenter: **Ken Nittono**, Hosei University, Japan

In the field of mining for collective text data, a wide variety of analyzing methods and its applications have been studied. In such diversification of the field, the approach of this research aims to extract latent knowledge from a large set of texts or documents and accumulate some selected documents for efficient knowledge acquisition. The target bunch of documents is basically formulated as a term-document matrix and association analysis, which is a fundamental approach in data mining, is applied to find some essential combinations of terms. There are many kind of studied for analyzing relation between terms to extract features of documents structurally, however, the proposed association rules, that are support and confidence principle, are considered here as significant features related to some latent knowledge contained by the documents. Based on some essential terms extracted by the rules, a few documents are selected using latent semantic analysis for term-document matrix and information retrieval method. The documents are retrieved on the basis of similarity between the terms and each document in a concept space which is generated by singular value decomposition. And the few documents out of a large collective set are considered as significant or potential context for knowledge eventually.

Wednesday 29.08.2012 08:40 - 10:45 Parallel Session G

IS03 Room R1: Demetra TIME SERIES MODELLING AND COMPUTATION Chair: Roland Fried

C412: **Sparse estimation of vector autoregressive models**
Presenter: **Christophe Croux**, KU Leuven, Belgium

A sparse estimation technique for the typically heavy parametrized vector autoregressive market response model is introduced. The advantages of the method are more accurate parameter estimation and prediction for short time series in high dimensions. For regression models, applicable to cross-sectional data, the Lasso method became a popular technique for estimating regression models when the number of observations is small with respect to the number of parameters to estimate. The Lasso minimizes the least squares criterion penalized for the sum of the absolute values of the regression parameters. It simultaneously performs parameter estimation and variable selection. The VAR model differs from a regression model in two aspects, making it necessary to extend the lasso technique to make it appropriate for our setting. First, a VAR model contains several equations, and corresponds to a multivariate regression model. The presence of correlation between the error terms of the different equations needs to be taken into account when performing sparse estimation. Second, a VAR model is dynamic, containing lagged versions of the same time series as right hand side variables of the regression equation. We consider all lagged values of the same time series as a single group of explanatory variables, and we want to avoid that one variable in a group is selected, while some other variables of the same group are not. For example, we want to exclude the possibility to select the second lagged series of a certain predictor but not the first. We show on simulated and real data the good performance of the sparse estimation technique for multivariate time series when the number of sample points is small with respect to the dimension. This paper is joint work with Sarah Gelper and Stefan Stremersch.

C092: **Quasi–Likelihood inference and prediction for negative binomial time series models**
Presenter: **Konstantinos Fokianos**, University of Cyprus, Cyprus
Co-authors: Vasiliki Christou

We study inference and diagnostics for count time series regression models which include a feedback mechanism. In particular, we are interested on negative binomial processes for count time series. We study probabilistic properties and quasi likelihood estimation for this class of processes. We show that the resulting estimators are consistent and asymptotically normally distributed. The key observation in developing the theory is a mean parameterized form of the negative binomial distribution. In addition, we provide probabilistic forecasts based on the assumption of Negative Binomial or Poisson distribution and we propose the use of probability integral transformation histogram, marginal calibration plot and scoring rules to assess the predictive performance and rank the competing forecast models.

C093: **On the robust analysis of periodic nonstationary time series**
Presenter: **Roland Fried**, TU Dortmund University, Germany
Co-authors: Nils Raabe, Anita Thieler

Robust methods for the detection of periodic signals in time series are presented and discussed. The motivation are applications to the modeling of drilling processes and in astroparticle physics. The considered signals which are measured during drilling processes consist of equidistant observations with known periodicity and gradually changing periodic structure. The basic objective is to understand the granularity of different materials by analyzing the periodic structure in order to design suitable simulation models, which render subsequent optimization of the system possible. For this aim, robust nonparametric smoothers and edge detectors for the reconstruction of periodic jump surfaces are developed and combined. In astroparticle physics, the situation is worse because of very irregular observation times and heteroscedastic noise. The main interest is in the detection and identification of periods, if they exist at all. Suitably modified robust nonparametric smoothers allow construction of generalized periodograms. Significant periodicities are identified by application of rules for outlier detection to such periodograms. All methods are investigated by applications to simulated and real data with and without outliers.

TS03 Room R8: Era TUTORIAL: NUMERICAL METHODS AND OPTIMIZATION IN STATISTICAL FINANCE Chair: Simona Sanfelici

C300: **Numerical methods and optimization in statistical finance**
Presenter: **Manfred Gilli**, University of Geneva, Switzerland

Many optimization problems in theoretical and applied science are difficult to solve: they exhibit multiple local optima or are not well-behaved in other ways (e.g., have discontinuities in the objective function). The still-prevalent approach to handling such difficulties – other than ignoring them – is to adjust or reformulate the problem until it can be solved with standard numerical methods. Unfortunately, this often involves simplifications of the original problem; thus we obtain solutions to a model that may or may not reflect our initial problem. But there is yet another approach: the application of optimization heuristics like Simulated Annealing or Genetic Algorithms. These methods have been shown to be capable of handling non-convex optimization problems with all kinds of constraints, and should thus be ideal candidates for many optimization problems. In this talk we motivate the use of such methods by first presenting some examples from finance for which optimization is required, and where standard methods often fail. We briefly review some heuristics, and look into their application to finance problems. We will also discuss the stochastics of the solutions obtained from heuristics, in particular we compare the randomness generated by the optimization methods with the randomness inherent to the problem.

OS06 Room R7: Ath. 4 ISBIS SESSION ON INFORMATION MEASURES AND TECHNOLOGY Chair: Christos Kitsos

C068: **On statistical inference for grouped censored data**
Presenter: **Ilia Vonta**, National Technical University of Athens, Greece

The problem of determining the appropriate model for a given data set is important for reducing the possibility of erroneous inference. Additional issues are raised in biomedicine and biostatistics where the existence of censoring schemes in survival modelling makes the determination of the proper model an extremely challenging problem. We focus on hypothesis testing and in particular on a class of goodness of fit tests based on Csiszar's family of measures and propose a general family of test statistics for treating the case of censored data with applications extended from survival analysis to reliability theory. First we present the Csiszar's class of measures, we formulate the problem and propose a class of test statistics based on the Csiszar's class for testing a null hypothesis about the true distribution of lifetimes subject to right censoring. Then, we present theoretic results about the asymptotic distribution of the test statistic under the null hypothesis and under contiguous alternatives. Both the simple and the composite null hypothesis where the proposed distribution depends on a finite dimensional parameter are treated. Finally, we provide simulations and real data that illustrate the performance of the proposed test statistic.

C397: **Ineffectiveness of the FIM in selecting optimal BIB designs for testing block effects**
Presenter: **Teresa Oliveira**, CEAUL and UNIVERSIDADE ABERTA, Portugal
Co-authors: Amilcar Oliveira

Information Theory is a relatively new branch of applied mathematics, emerging in the 1940s. It deals with all theoretical problems connected with the transmission of information over communication channels and includes the quantification of information and the study of uncertainty. It has found deep applications in probability theory and statistics as well as in many social, physical and biological sciences. The most important information measures in Information Theory can be divided in three categories: entropy, non-parametric and parametric types. In parametric types we have the example of Fisher Information. A review on Fisher Information and on Fisher Information Matrices (FIM) will be presented. Experimental Designs, emerging in the 1920s, have been widely used in agricultural, industrial and computer experiments, in order to reduce experimental cost and to provide efficient parameter estimation. Some topics on Balanced Incomplete Block (BIB) designs with and without block repetition are shown. Besides some optimal design criteria are presented showing the important role of the FIM in the field of optimal designs. It is found that FIM is not useful to find a BIB Design (BIBD) being optimal to contrast block effects.

C357: **Clustering cross-validation and mutual information indices**
Presenter: **Maria Amorim**, ISEL and BRUE-IUL, Portugal
Co-authors: Margarida Cardoso

The use of cross-validation enables to evaluate clustering solutions internal stability. In this context mutual information indices (relying on the concept of entropy), enable to measure the agreement between diverse clustering solutions (partitions, in particular) based on diverse sub-samples drawn from the same source. In order to adequately do so, the proposed approach uses indices values corrected for agreement by chance. The corrected values are obtained using simulated indices values, corresponding to cross-classification tables generated under the hypothesis of restricted independence. The simulated tables have fixed marginal totals that match the ones derived from cross-validation. The proposed method is illustrated using real data referring to a retail application (with unknown structure) and using simulated data, with known structure with diverse clusters' weights and diverse degrees of cluster's overlap being considered. Furthermore, for simulated data, the agreement between clustering solutions obtained and the real partitions is also analyzed, enabling to discuss the observed relationship between stability and agreement with ground truth.

C331: **Supporting institutional research evaluation through cluster analysis and parallel coordinators techniques**
Presenter: **Effie Papageorgiou**, Technological Education Institute of Athens, Greece
Co-authors: Anastasios Tsolakidis, Cleo Sgouropoulou, Christos Koilias

Higher education institutions need to have a social responsibility and attainable objectives regarding their education and research activities. We present two different perspectives in order to support research evaluation and unveil the research communities. Using Hierarchical cluster analysis we find relatively homogeneous clusters of cases (faculty members) based on measured characteristics such as h-index, papers productivity, fp-index and research projects. Introducing an ontology-based software system architecture that supports research policy evaluation processes and decision-making strategies, using Parallel Coordinators, we provide an interactive way for the evaluation of faculty members' research activities. The used data represent the scientific publication activity of the 25 members of the academic staff of the Department of Informatics, at the Technological Educational Institute of Athens. A comparison of the resulting clusters is presented.

C083: **An entropy type measure of complexity**
Presenter: **Christos Kitsos**, Technological Educational Institute of Athens, Greece
Co-authors: Thomas Toulias

In a continuous system the Shannon entropy is defined as the expected content of a random variable X. We assume that the random variable X comes from the generalized γ–ordered Normal distribution $\mathcal{N}_\gamma$ which can be used to applications concerning heavy–tailed distributions. The generalized Shannon entropy also introduced through the generalized Fisher′s information. Therefore a useful measure to technological applications introduced and studied, extending the SDL measure of complexity used in the study of the EEG signals on epileptic seizures.

OS09 Room R2: Ares IFCS SESSION ON FINITE MIXTURE MODELS Chair: Shu-Kay Ng

C070: **New parameter estimates for random graph mixture models**
Presenter: **Christophe Ambroise**, Universite d'Evry val d'Esssonne, France

Random-graph mixture models are very popular for modelling real data networks. Parameter estimation procedures usually rely on variational approximations, either combined with the expectation–maximization (EM) algorithm or with Bayesian approaches. Despite good results on synthetic data, the validity of the variational approximation has not been established yet. Moreover, these variational approaches aim at approximating the maximum likelihood or the maximum a posteriori estimators. The behaviour of such estimators in an asymptotic framework (as the sample size increases to infinity) remains unknown for these models. We show that, in many different affiliation contexts (for binary or weighted graphs), parameter estimators based either on moment equations or on the maximization of some composite likelihood are strongly consistent and convergent, when the number n of nodes increases to infinity. As a consequence, our result establishes that the overall structure of an affiliation model can be (asymptotically) caught by the description of the network in terms of its number of triads (order 3 structures) and edges (order 2 structures). Moreover, these parameter estimates are either explicit (as for the moment estimators) or may be approximated by using a simple EM algorithm, for which the convergence properties are known. We illustrate the efficiency of our method on simulated data and compare its performances with other existing procedures. A data set of cross-citations among economics journals is also analysed.

C188: **Model based clustering of multivariate spatio-temporal data: A matrix-variate approach**
Presenter: **Cinzia Viroli**, University of Bologna, Italy

Multivariate spatio-temporal data arise from the observation of a set of measurements in different times on a sample of spatially correlated locations. They can be arranged in a three-way data structure characterized by rows, columns and layers. In this perspective each obseved statistical unit is a matrix of observations instead of the conventional p-dimensional vector. We propose a model based clustering for the wide class of continuous three-way data by a general mixture model with components modelled by matrix-variate Gaussian distributions. The effectiveness of the proposed method is illustrated on multivariate crime data collected on the Italian provinces in the years 2005-2009.

C205: **Constraints to avoid spurious solutions in finite mixture model estimation**
Presenter: **Agustin Mayo-Iscar**, Universidad de Valladolid, Spain
Co-authors: Luis A. Garcia-Escudero, Alfonso Gordaliza, Carlos Matran

The starting point is the problem caused by the high prevalence of spurious local maximizers in the likelihood function corresponding to finite mixture models. Our proposal is to use a restricted maximum likelihood estimator. It is computationally feasible by using the EM algorithm with a modification in the M step, motivated by the constraints. It is possible that expert statisticians can manage the suggested problem by analysing all the local maximizers. However, non-expert practicioners may alternatively use the restricted estimator that we propose. It is necessary to choose the level of the restrictions in advance. Users can choose it in a meaningful way. Alternatively, they can run the procedure for different restrictions levels and choose the best solution after analysing the very small set of obtained estimates.

C142: **A hierarchical modeling approach for clustering probability density functions**
Presenter: **Daniela Giovanna Calo**, University of Bologna, Italy
Co-authors: Angela Montanari, Cinzia Viroli

The problem of clustering probability density functions is emerging in different scientific domains. The methods proposed so far in the statistical literature are mainly focused on the univariate settings and are based on heuristic clustering solutions. The aim is to address new aspects of the problem: the multivariate setting and a model based perspective. The novel approach relies on a hierarchical mixture modeling of the data and on a factorial model performing dimension reduction. The proposed method is illustrated on two real data sets.

CS01 Room R4: Aph.+Pos. ROBUST STATISTICS II Chair: Stefan Van Aelst

C242: **Robust estimation of model with fixed and random effects**
Presenter: **Jan Amos Visek**, Charles University in Prague, Czech Republic

Robustified estimation of the linear regression model is recalled, with the inclusion of some pros and cons of individual methods. Then a straightforward generalization of the least weighted squares for model with the fixed or random effects is proposed and its properties, including the consistency, briefly mentioned. Patterns of the results of an extensive numerical study, employing an empirically optimized weight function, are presented. It can be of interest that the product of the optimized weight function and of the individual terms from the normal equations is close to Hampel's redescending ψ-function. Finally, conclusions implied by the results of numerical study conclude the paper.

C258: **Student depth in robust economic data stream analysis**
Presenter: **Daniel Kosiorowski**, Cracow University of Economics, Poland

In case of the economic data stream, since the data are arriving continuously and there is no known end to it, the usual approach of reading in all data and then processing them is not feasible. We cannot successfully analyze them by means of classical Fisherian statistics with well-defined experiment, data generated by regular model. Data streams carry signals that appear randomly, are irregularly spaced and the time duration between successive signals is not deterministic, but random. Additionally data streams generally are generated by non-stationary models of unknown form. Standard econometric time series analytical tools are generally inapplicable. The aim is to study properties of Mizera and Müller location-scale depth procedures in a context of robust data stream analysis. We look into a probabilistic information on the underlying data stream model carried by this depth function. We study the robustness and the utility in a decision making process. In particular we investigate properties of the moving Student median (two dimensional Tukey median in a location–scale problem). Results of our considerations are in favor for depth based procedures in comparison to Bayesian, local nonparametric regression or estimation-model selection approaches.

C251: **Selection of tuning parameters in robust sparse regression modeling**
Presenter: **Heewon Park**, Chuo university, Japan
Co-authors: Fumitake Sakaori, Sadanori Konishi

There is currently much research on the lasso-type regularized regression, which is a useful tool for simultaneous estimation and variable selection. Although the lasso-type regularization has several advantages in regression modeling, it suffers from outliers and thus yields unstable model estimates, because of using penalized least squares methods. A robust lasso-type approaches are constructed by using the robust loss function with L_1-type penalty. The desirable performance of the robust lasso-type regularization methods heavily depends on an appropriate choice of the regularization parameters and also a tuning constant in outlier detection. The choice of these tuning parameters can be viewed as model selection and evaluation problem. We present a model selection criterion for evaluating models constructed by the robust lasso-type approach from an information-theoretic view point. Monte Carlo simulations are conducted to investigate the effectiveness of our procedure for evaluating the robust sparse regression models. We observed that the proposed modeling strategy performs well in the presence of outliers.

C153: **Normalized estimating equation for robust parameter estimation**
Presenter: **Hironori Fujisawa**, Institute of Statistical Mathematics, Japan

Robust parameter estimation has been discussed as a method for reducing a bias caused by outliers. Recently, an estimating equation using a weighted score function has been proposed with a bias-correction ensuring Fisher consistency. Although a typical estimating equation is unnormalized, this paper considers a normalized estimating equation, which is corrected to ensure that the sum of the weight is one. In robust parameter estimation, it is important to control the difference between the target parameter and the limit of the robust estimator, which is referred to as the latent bias in this paper. The latent bias is usually discussed in terms of influence function and breakdown point. It is illustrated by some examples that the latent bias can be sufficiently small for the normalized estimating equation even if the proportion of outlier is significant, but not for the unnormalized estimating equation. Furthermore, this behavior of the normalized estimating equation can be proved under some mild conditions. The asymptotic normality of the robust estimator is also presented and then it is shown that the outliers are naturally ignored with an appropriate proportion of outlier from the viewpoint of asymptotic variance. The results can be extended to a regression case.

C245: **Robust statistics for classification of remote sensing data**
Presenter: **Dyah Erny Herwindiati**, Tarumanagara University, Indonesia
Co-authors: Maman Abdurahman Djauhari, Luan Jaupi

Remote sensing is the science and art of obtaining information about an object through the analysis of data acquired by a device that is not in contact with the object under investigation. In many aspects, remote sensing can be thought of as a reading process, using various sensors. The analysis of the information remote sensing data obtained is though visual and digital image processing. We discuss the robust classification of remote sensing data from Landsat 7 satellite. The area under investigation is Jakarta Province. The supervised land classification is done with two process; i.e. the training sites and classification process. The outcomes of training site are the spectral imaging references of water catchment area and vegetation area; which are useful for classification process. A robust computationally efficient approach is applied for training site to deal with the large remote sensing data set of Jakarta. A new depth function which is equivalent to Mahalanobis depth is used, the function is able to replace the inversion process of covariance matrix. The objective is to introduce the depth function for robust estimation of a multivariate location parameter minimizing vector variance for classification of data remote sensing.

CS08 Room R5: Ath.1+2 BIOSTATISTICS AND BIOCOMPUTING Chair: Joyce Niland

C067: **Experimental designs for drug combination studies**
Presenter: **Bader Almohaimeed**, Manchester University, United Kingdom
Co-authors: Alexander Donev

The significance of drug combination studies is increasing due to the opportunities they create to achieve a desirable therapeutic effect using smaller doses of two or more drugs rather than a large dose of a single drug, thus reducing the danger of undesirable side effects. The study of the joint action of two drugs is more complex and typically takes up longer time and substantial resources compared to studies of single drugs. This is why using an appropriate experimental design is very important. Recently there has been substantial progress with the development of relevant statistical methods and tools for statistical analysis of the data obtained in such studies, but much less attention has been given to the

choice of suitable experimental designs. We propose an approach using R software for design of combination studies that takes into account the statistical analysis that will be carried out as well as the distribution of the response which can be one of those belonging to the exponential family of distributions. We provide specific examples for the most common cases. We also propose simple and flexible experimental designs that permit a variety of combinations studies to be designed in such a way that low or high doses are possible to avoid and the statistical analysis can be simplified.

C281: **Efficient penalised likelihood estimation for the cox model**
Presenter: **Stephane Heritier**, George Institute and Sydney University, Australia
Co-authors: Jun Ma, Serigne Lo

Penalised likelihood to fit the Cox model is proposed to be used. We first adopt an approximation such as discretization to the baseline hazard function, and then estimate this approximated baseline hazard and the regression coefficients simultaneously. A new iterative optimization algorithm, which combines the Newton's algorithm and a multiplicative iterative algorithm, is developed. This algorithm has two interesting properties: 1) it always increases the penalised likelihood ensuring fast convergence; 2) the baseline hazard is always positive. We show that, under independent censoring, the maximum penalised likelihood estimator is consistent, asymptotically normal, and retains full efficiency provided that the smoothing parameter tends to zero sufficiently fast. A simulation study reveals that this method can be more efficient than the partial likelihood, particularly for small to moderate samples. In addition, the new estimator is substantially less biased under informative censoring. The approach provides new insight on a critical care dataset.

C222: **Comparison among spatial clustering methods on hierarchical structures for DNA markers**
Presenter: **Makoto Tomita**, Tokyo Medical and Dental University, Japan

Single nucleotide polymorphisms (SNPs) are the most abundant form of genetic variation. As an association study, linkage disequilibrium analysis for SNP data is particularly important. In DNA sequences, domain hotspots exist at which recombinations have occurred briskly. Conversely, large domains with infrequent recombinations in which linkage disequilibrium is maintained also exist. Such domain called a haplotype block or LD block. They have approached the new method to identify LD blocks using Echelon analysis which is the one of spatial clustering methods. We approached other spatial clustering methods, then these results were compared.

C244: **Application of k-word match statistics to the clustering of proteins with repeated domains**
Presenter: **Conrad Burden**, Australian National University, Australia
Co-authors: Susan Wilson, Junmei Jing, Sylvain Foret

An algorithm is developed for clustering similar protein sequences. The algorithm is based on a word-match statistic defined as a weighted inner product between count vectors of words of pre-specified length k. The weighting matrix used is designed to account for higher substitution rates between chemically similar amino acids. The method depends on similarity scores calculated from p-values under a null hypothesis which takes into account that proteins frequently contain multiple repeats of identifiable domains. The algorithm is applied to the Beta-catenin and Notch families of proteins, both of which are rich in domain repeats and are common throughout the animal kingdom.

C366: **The distribution of short word matches between Markovian sequences**
Presenter: **Conrad Burden**, Australian National University, Australia
Co-authors: Paul Leopardi, Sylvain Foret

The D_2 statistic is defined as the number of short word matches of pre-specified length k between two sequences of letters from a finite alphabet. It is potentially a useful statistic for measuring similarity of biological sequences in cases where long alignments may not be appropriate. Examples of this are cases where insertions, deletions and repetitions of extended stretches are present (as in regions of genome with copy number variations or proteins with a domain structure), or when detecting parts of a genome rich in regulatory motifs. The distribution of the D_2 statistic has been characterised extensively for the null hypothesis of sequences consisting of identically and independently distributed letters. However, it is well known that biological sequences are generally better modelled as higher-order Markovian sequences. A theory will be developed for the distribution of the D_2 statistic between Markovian sequences of arbitrary order. To facilitate the derivation of exact analytic formulae for the properties of the distribution, the concept of Markovian sequences with periodic boundary conditions will be introduced.

CS23 Room R6: Ath. 3 MULTIVARIATE DATA ANALYSIS II Chair: Vincenzo Esposito Vinzi

C219: **Constrained multilevel latent class models for the analysis of three-way three-mode binary data**
Presenter: **Michel Meulders**, HUBrussel, Belgium

Probabilistic feature models (PFMs) can be used to explain binary associations between two types of elements (e.g., products and attributes) on the basis of binary latent features. In particular, to explain observed associations between elements, PFMs assume that respondents classify each element with respect to a number of binary latent features, and that the observed association between the two elements is derived as a specific mapping of these classifications. PFMs assume that the association probability of elements are the same across respondents, and that all observations are statistically independent. As both assumptions may be unrealistic, a multilevel latent class extension of PFMs is proposed which allows feature-element parameters to be different across latent rater classes, and which allows us to model dependencies between associations with a common element by assuming that the link between features and elements is fixed across judgements. Formal relationships with existing multilevel latent class models for three-way data are described. As an illustration, the models are used to analyze data on product perception and anger-related behavior.

C335: **Batch self organizing maps for interval and histogram data**
Presenter: **Antonio Irpino**, Second University of Naples, Italy
Co-authors: Francisco de Assis Tenorio de Carvalho, Rosanna Verde

An extension of Batch Self Organizing Map (BSOM) is here proposed for non classic data as interval valued and histogram data. These kind of data have been defined in the context of symbolic data analysis. The BSOM cost function is then based on two distance functions: the Euclidean distance and the Wasserstein distance. This last distance has been widely proposed in several techniques of analysis (clustering, regression) when input data are expressed by distributions (empirical by histograms or theoretical by probability distributions). The peculiarity of such distance is to be an Euclidean distance between quantile functions so that all the properties proved for L2 distances are verified again. An adaptive versions of BSOM is also introduced considering an automatic system of weights in the cost function in order to take into account the different effect of the several variables in the Self-Organizing Map grid. Applications on real and synthetic data set are proposed to corroborate the procedures.

C218: **Some new classes of copulas and their application in finance**
Presenter: **Jozef Komornik**, Comenius University, Bratislava, Slovakia, Slovakia
Co-authors: Radko Mesiar, Magdalena Komornikova

We introduce 2 recently proposed constructions of derived copulas (UCS and DUCS) as well as new constructions of quadrant dependent copulas into modelling of the financial time series. All those classes show a potential to improve models for relations between financial time series, as we have demonstrated on the models of the relations between the returns on investments in stocks and gold.

C231: **ECO-POWER: A novel method to reveal common mechanisms underlying linked data**
Presenter: **Martijn Schouteden**, KULeuven, Belgium
Co-authors: Katrijn Van Deun, Iven Van Mechelen

Often data are collected consisting of different blocks that all contain information about the same entities. A main challenge in the analysis of such data is to reveal underlying mechanisms that are common to all data blocks. An interesting class of methods for this purpose is the family of simultaneous component analysis methods, which yield dimensions underlying the data at hand. Unfortunately, in results of such analyses information that is common to all data blocks and information that is specific for a certain data block or a few of them are mixed up. To solve this problem, we present a novel optimization criterion that relies on ideas underlying power regression, along with an associated iterative algorithm. We evaluate this algorithm in a comprehensive simulation study, and we present an application to empirical data.

C230: **BIC selection of the number of classes in latent class models with background variables**
Presenter: **Tomoki Tokuda**, University of Leuven, Belgium
Co-authors: Iven Van Mechelen, Gerda Claeskens, Francis Tuerlinckx

The Bayesian Information Criterion (BIC) is widely used for selecting the number of classes in mixture models, including latent class models. In normal mixture models, BIC is known to suffer from the problem of underestimating the true number of classes in high dimensional data where background variables, irrelevant to the clustering, are present. However, this problem has received less attention in the context of latent class models. We study the behavior of BIC in latent class models. First, we derive an analytical approximation of the expectation of BIC. Using this result, we show that also in latent class models with background variables BIC suffers from underestimating the true number of classes. Second, we propose a solution to this problem in terms of a corrected BIC. Finally, we report the results of a limited simulation study, which gives a first indication that the corrected BIC may have a good performance with regard to model selection.

PS05 Room Athenaeum Terrace POSTER SESSION V Chair: Cristian Gatu

C236: **Modification of CHF coefficient for evaluation of clustering with mixed type variables**
Presenter: **Tomas Loster**, University of Economics Prague, Czech Republic
Co-authors: Jitka Langhamrova

Cluster analysis involves a broad scale of techniques. Determining the optimal number of clusters is very important. Current literature draws attention particularly to the evaluation of clustering when individual objects are characterized only by quantitative variables. New coefficients are suggested for the evaluation of resulting clusters based on the principle of the variability analysis. Furthermore, only coefficients for mixed type variables based on a combination of sample variance and one of the variability measures for nominal variables will be presented. On the basis of real data files analyses there were compared newly proposed indices with the known BIC criterion. The number of object groups was known and there was interested in agreement of the found optimal number of clusters with the real number of groups. There was analyzed several (15) data sets. In analysed data sets was found that CHF-G index (based on the Gini coefficient) was successful. According to the experience when was analyzing more data files, the CHF-G index determines the correct number of clusters in most cases. The second successful criterion is the CHF-H (based on Entropy) index. The BIC-H and BIC-G indices are less successful.

C199: **Individual control treatments in designed agricultural experiments**
Presenter: **Stanislaw Mejza**, Poznan University of Life Sciences, Poland
Co-authors: Iwona Mejza

A common aim of genetical and agricultural experiments is to compare the test treatments with an individual control (standard) treatment. Two kinds of experiments are considered, namely; 1) - nonreplicated genetical experiments performed at early stage breeding program and 2) - the factorial experiments with crossed and nested structures of factors. Response surface methodology is proposed for the analysis of nonreplicated breeding experiments. First, estimates of the yield response surface based on check plots as supporting points are obtained. Then the treatment (genotype, hybrid) effect is estimated as the difference between the observation obtained for the treatment and the response surface forecast. The consequences of density and arrangements of controls (check plots) on statistical inference using both simulation and uniformity trials are investigated. Factorial experiments with nested and crossed factorial structure (split block designs, split plot designs) are considered in detail. In particular arrangements of individual controls in the incomplete split plot designs and incomplete split block designs are considered. Two aspects of these experiments, namely constructing methods leading to optimal designs and design efficiency, are examined. The Kronecker and the so-called semi-Kronecker product of designs are applied to generate new designs with desirable properties.

C250: **Distributed fusion filter for stochastic systems with markovian random delays and uncertain observations**
Presenter: **Maria Jesus Garcia-Ligero**, Universidad de Granada, Spain
Co-authors: Aurora Hermoso-Carazo, Josefa Linares-Perez

The least-squares linear filtering problem for discrete-time stochastic systems with uncertain observations which can be randomly delayed by one sampling time is addressed when these are acquired from multiple sensors. The uncertainties in the observations and the delays are modeled by sequences of Bernoulli random variables with different characteristics; specifically, the uncertainty is described by independent Bernoulli random variables whereas the delays are modeled by homogeneous Markov chains. Assuming that the state-space model for the signal is not fully known, the filter is obtained by using a distributed fusion method. This method is structured in two stages; in the first one, local filters for each sensor are derived by using the information provided by the covariance functions of the processes involved in the observation equations, as well as the probability distribution of the variables modeling the delays and the uncertainty. In the second one, the distributed fusion filter is obtained as the linear combination of the local linear filters verifying that the mean squared error is minimum. Filtering error covariance matrices are also given to measure the goodness of the proposed algorithm.

C255: **Stochastic forecast of fertility**
Presenter: **Mare Vahi**, University of Tartu, Estonia

A new method for stochastic forecast of the age-specific fertility rates is proposed. This approach can be summarized as follows. (1) Fit the suitable family of distributions for modelling the fertility rates. The age-specific fertility pattern has a typical shape through years. In order to describe this shape a number of distributions have been proposed. Most commonly used distributions are beta distribution, gamma distribution, Hadwiger distribution, mixture of beta distributions and mixture of Hadwiger distributions. (2) Estimate the parameters of distribution. (3) Forecast the parameters using the time series model. (4) Use the forecast parameters to forecast the model for age-specific fertility rates by one-year age groups. The fertility model is then used in simulation of future fertility. The methodology is applied to Estonian fertility data.

C279: **Stochastic discretization schemes for the simulation of a class of hysteretic systems: A comparison**
Presenter: **Paula Milheiro-Oliveira**, FEUP and CMUP, Portugal
Co-authors: Pedro Vieira, Alvaro Cunha

The analysis of the stochastic behaviour of hysteretic systems requires, in many cases, the ability to numerically simulate the solutions of non linear stochastic differential equations (SDE) of a certain type. Problems like those arising in structural engineering involving the analysis of the effect

of earthquakes on bridges or buildings are among those where an SDE with hysteretic components should be adopted to model the phenomena. Stochastic discretization schemes need to be used prior to the numerical simulation algorithm being designed. The single degree of freedom Noori–Baber–Wen model is considered for the structure perturbed by the effect of a Gaussian white noise as a simple model for the external forces. Two stochastic schemes of the Runge-Kutta family and two schemes of Newmark family, adapted from the literature, are applied in view of the discretization and further simulation of the model. The results are compared in terms of the second statistical moments of the displacement of the structure. These moments are computed on Monte Carlo simulations. Two hundred trajectories were simulated for values of the model parameters commonly used in practice. One concludes that all the schemes produce similar trajectories for earlier times but clearly differ for larger times. For both families of schemes the standard deviation of the response in terms of displacement differs more from that obtained by solving the moments equation as the non-stationarity gets larger.

C299: **Forecast of the population of the Czech regions by sex, age and the highest education level**
Presenter: **Tomas Fiala**, University of Economics Prague, Czech Republic
Co-authors: Jitka Langhamrova, Jana Langhamrova

Classic population projections provide the forecast only of the sex and age structure of the population in each year of the projected period. They give no information about the qualitative side of the population. The aim is to describe very briefly the methodology of the nor commonly computed projection taking into account also the education level of each person which is a new method of modeling and forecasting the development of human capital. The computation is based on the classical component projection method with simplified model of migration (only immigration at the level of net migration is assumed, emigration is supposed to be zero). Uniform distribution of the time of immigration and uniform distribution of the date of birth during the year is supposed. The multistate projection method has been applied, increase of education level is regarded as a transition (internal "migration") from one subpopulation to another. The main results of computation of such projection for the case of the regions of the Czech Republic are given.

C310: **Analysis of neurosurgery data: A statistical and data mining approach**
Presenter: **Petr Berka**, University of Economics/Institute of Finance and Administration, Prague, Czech Republic
Co-authors: Michal Vrabec

The data concerning the outcomes of surgical clipping and endovascular treatment in acute aneurysmal subarachnoid hemorrhage (SAH) patients have been analyzed to reveal relations between subjective neuropsychological assessments, measurable characteristics of the patient and the disease, and the type of treatment the patient had undergone one year before. We build upon results of previous analyses where have been found that the differences in neuropsychological assessment of the patients treated by either coiling or clipping was small and slightly in favor of surgical group. Using this data, we compare the "classical" statistical and data mining approach. While statistics offers techniques based on contingency tables, where the compared variables have to be manually selected, data mining methods like association rules, decision rules or decision trees offer the possibility to generate and evaluate a number of more complex hypotheses about the hidden relationships. We used SAS JMP to perform the statistical analysis. Our original LISp-Miner system based on the GUHA method was used for the data mining experiments.

C393: **On smoothing time series with low average counts**
Presenter: **Koen Simons**, Scientific Institute of Public Health, Belgium

Generalized Additive Models have been widely adopted for studies of acute effects of particulate matter on mortality and morbidity. Monitoring of pollutants and health outcomes increased worldwide and investigators thus increasingly relied on automatic selection methods that exist of summary statistics such as AIC and PACF. Methodological studies have used simulations to compare selection methods and their impact on large scale multi-city analyses and concluded that aggressive smoothing is to be preferred. For smaller groups, these effects can be visualised with simple residual plots. Data from Belgian cities is used to illustrate the effect of over-smoothing on time series with low average counts.

C354: **Strand asymmetries in mitogenomes: Toward detection of change points**
Presenter: **Nora M Villanueva**, University of Vigo, Spain
Co-authors: Miguel M Fonseca, Marta Sestelo, Javier Roca-Pardinas

Identifying the mutational processes that shape mitochondrial DNA (mtDNA) sequences is fundamental to better how mitogenomes evolve. The replication mechanism, during which the strands are exposed to an elevated mutational damage, has been described as one of the main sources of compositional bias in mtDNA. Different methods have been proposed to analyze such asymmetries in the past, but lack any measure of statistical support. We introduce a simple method to detect compositional changes or singularities in mtDNA based on regression models and their derivatives. The methodology was implemented in an `R` package from the change point research community within an easy to use framework.

PS06 Room Athenaeum Terrace POSTER SESSION VI Chair: Cristian Gatu

C401: **Some extensions of a test data analysis platform based on multi-dimensional item response theory**
Presenter: **Tomoya Okubo**, The National Center for University Entrance Examinations, Japan

A test data analysis platform is presented using item response theory extended with mixed multi-beta distribution. We focus on extending the system that enables users to analyze test data using many types of item response models including multi-dimensional item response theory. The system allows us to assume multi-dimensions to items and multi-groups to respondents. Further, we will show the merits of applying mixed multi-beta distributions to multi-dimensional Item Response Theory in this presentation using actual educational data. The system has been implemented 1, 2, 3-parameter logistic model for graded response model, Partial Credit Model, Generalized Partial Credit Model and Order-constrained Nominal Categories Model for graded response scale, and Nominal Categories Model for items in a nominal scale as well as the multi-dimensional models. Further, the system estimates item parameters of tests containing mixed type of items. The system is an open-platform and a client-server model web application, and run only on a Web browser.

C355: **Basis pursuit approach to estimate a mean**
Presenter: **Jiri Neubauer**, University of Defence, Czech Republic
Co-authors: Vitezslav Vesely

The contribution is focused on the estimation of the mean of a stochastic process (the sample path) by sparse parameter estimation from an overparametrized model. A covariance stationary stochastic process with change in the mean is estimated using dictionary consisting of Gaussian and Heaviside functions. The basis pursuit algorithm is used to get sparse parameter estimates.

C382: **Simultaneous equations solved by TSLS with interval predictions**
Presenter: **Jitka Langhamrova**, University of Economics Prague, Czech Republic
Co-authors: Martina Miskolczi, Jana Langhamrova

Labour markets and their modelling are considered. Two different simultaneous models were carefully developed. Number of employed and unemployed individuals and total number of economically active individuals are considered as endogenous variables in the first model, rate of unemployment and inflation are considered as endogenous variables in the second model. Macroeconomic variables, for example GDP growth rate, consumption, investments, export, wage and some delayed variables were used as predictors. Two-stage least square method (TSLS) was

used to estimate unknown coefficients of structural form of proposed simultaneous models, based on quarterly distributed data from the Czech Republic 2004–2010. Further, reduced form, elasticity coefficients and predictions were calculated. Moreover, predictions for 2011–2012 based on interval estimates of predictors were prepared with respect to rules of calculation with interval matrices. As results, intervals of endogenous variables were obtained and compared with both real observations (available till Q1/2012) and 95% prediction intervals. This approach combines two different procedures and is not presented in common literature. Interval technique in combination with other statistical methods is very well usable particularly in period of crisis when required variables tremble and are unstable. Here, interval predictions have higher usability compared to classical predictions.

C389: **On analysing finite mixture models**

Presenter: **Amilcar Oliveira**, CEAUL and Universidade Aberta, Portugal

Co-authors: Teresa Oliveira

The problem of finite mixtures in normal populations will be discussed, with reference to the traditional methods currently used in univariate and multivariate cases. A practical method on the parameters estimation in a mixture of normal populations for the univariate case will be presented, as well as the respective validation. Using R software and simulation techniques a comparison between the proposed method and others, namely the EM algorithm, will be presented. Some perspectives for future research in this area will be approached.

Wednesday 29.08.2012 11:15 - 12:55 Parallel Session H

IS05 Room R1: Demetra ROBUST MULTIVARIATE STATISTICAL METHODS Chair: Mia Hubert

C107: **Robust likelihood ratio type tests for regression τ estimators**
Presenter: **Stefan Van Aelst**, Ghent University, Belgium
Co-authors: Matias Salibian-Barrera, Victor Yohai

ANOVA tests are the standard tests to compare nested linear models fitted by least squares. These tests are equivalent to likelihood ratio tests and thus are very powerful. Since least squares estimators are very vulnerable to outliers, we consider regression τ estimators to estimate the parameters in the linear models. We introduce robust likelihood ratio type test statistics based on the τ estimates of the error scale. For the null distribution of the test statistics we either use the asymptotic approximation or the fast and robust bootstrap.

C139: **Robust classification**
Presenter: **Ruben Zamar**, University of British Columbia, Canada
Co-authors: Mohua Podder, Will Welch, Scott Tebbutt

A classification method will be presented which is robust not only to outliers in the training data, but also in the test data. We achieve that by using an ensemble of robust classifiers based on mixture models. Our methodology is applied for classification of single nucleotide polymorphism (SNP) genotypes. A mixture model for classification provides robustness against outlying values of the explanatory variables. Furthermore, different sets of explanatory variables are generated by deliberate redundancy in the genotyping chemistry, and the algorithm chooses among these sets in a dynamic way, prediction by prediction. This provides robustness against outliers in the test set.

C169: **A new depth-based approach for detecting outlying curves**
Presenter: **Mia Hubert**, KU Leuven, Belgium
Co-authors: Gerda Claeskens, Bart De Ketelaere, Kaveh Vakili

Depth functions are statistical tools, used to attribute a sensible ordering to observations in a sample from the center outwards. Recently several depth functions have been proposed for functional data. These depth functions can for example be used for robust classification and for the detection of outlying curves. A new depth function is presented, which can be applied to multivariate curves and which takes the local changes in the amount of variability in amplitude into account. It is illustrated on an industrial data set how this depth function can be useful to detect globally outlying curves as well as curves that are only outlying on parts of their domain. Several graphical representations of the curves and their degree of outlyingness are presented.

TS01 Room R8: Era TUTORIAL BY ARS OF IASC: BAYESIAN COMPUTING AND APPLICATIONS Chair: Rand WIlcox

C213: **Bayesian computing and applications**
Presenter: **Cathy WS Chen**, Feng Chia University, Taiwan

The objective is to introduce the Bayesian approach to statistical inference with applications and to describe effective approaches for Bayesian modeling and computation. Modern Bayesian inference relies heavily on computational algorithms, and hence a substantial part of the tutorial will be focused on posterior simulation. We describe Markov chain Monte Carlo (MCMC) methods in detail. Commonly used posterior simulators such as Gibbs sampling and random walk and independent kernel Metropolis-Hastings algorithms will be briefly reviewed. MCMC diagnostics will also be discussed including several approaches to monitoring convergence. We will illustrate Bayesian estimation with some popular models, such as regression models with change points, threshold autoregressive models, and GARCH models etc. Model selection and testing model adequacy in Bayesian framework are also discussed.

OS14 Room R2: Ares COMPONENT-BASED METHODS FOR SEM AND MULTI-BLOCK DATA ANALYSIS Chair: Vincenzo Esposito Vinzi

C186: **Advances on regularized generalized canonical correlation analysis**
Presenter: **Arthur Tenenhaus**, Supelec, France

Multi-block data analysis concerns the analysis of several sets of variables (blocks) observed on the same set of individuals. In this case, each block is a data matrix and represents a set of variables observed on a set of individuals. The number and the nature of the variables differ from one block to another but the individuals must be the same across blocks. In this framework, we are interested in evaluating relationships between blocks. For instance, in biology, with the availability of many 'omics' datasets measured on the same set of patients, the development of methods capable to analyze conjointly multiple datasets becomes crucial. Such development remains a major technical and computational challenge as most approaches suffer from high data dimensionality. We will talk about Regularized Generalized Canonical Correlation Analysis (RGCCA) which is very general and flexible method for multi-block data analysis. We will also present recent extensions of RGCCA such as Kernel Generalized Canonical Correlation Analysis and Sparse Generalized Canonical Correlation Analysis.

C203: **Multidimensional latent variables in PLS path modeling**
Presenter: **Laura Trinchera**, AgroParisTech - INRA, France
Co-authors: Giorgio Russolillo, Andrea Capellini, Vincenzo Esposito Vinzi

Partial Least Squares Path Modeling (PLS-PM) is the most widely used component-based approach to Path Models of Latent Variables. Two different estimation procedures based on OLS regression (known as Mode A and Mode B) can be used in PLS-PM to estimate the outer weights, i.e. the weights used for building the latent variable scores. However, both Mode A and Mode B assume a unique latent variable behind each block of manifest variables. Recently we have proposed two new outer estimation procedures within the PLS-PM algorithm for handling multidimensional latent variables: the PLScore Mode and the PLScow Mode. In both procedures PLS Regression replaces OLS regression, but the PLScore Mode keeps the classical PLS-PM constraint of unitary variance for the latent variable scores, while in PLScow Mode the outer weights are constrained to have a unitary norm, thus importing the normalization constraints of PLS-R. We investigate the performance of these modes using a popular data-set broadly used in classical SEM (LISREL approach) in which synthetic variables are added in order to simulate multidimensional latent variables. Moreover, we will compare PLS modes to the classical Mode B and Mode A and to Regularized Generalized Canonical Correlation Analysis that is another recent approach to multiple table analysis.

C234: **Principal covariates regression versus exploratory structural equation modeling**
Presenter: **Marlies Vervloet**, KU Leuven, Belgium
Co-authors: Eva Ceulemans, Marieke Timmerman, Wim Van den Noortgate

Structural Equation Modeling (SEM) is a very popular technique which consists of a structural part, showing how criterion variables depend on the predictor variables, and a measurement part, showing which latent factors underlie the observed variables involved. Recently, Exploratory SEM (ESEM) was introduced, which is a SEM model in which Exploratory, rather than Confirmatory Factor Analysis is used for the measurement

part of the model. ESEM, however, is very similar to a deterministic model that exists for quite some time: Principal Covariates Regression (PCovR). PCovR reduces predictor variables to components, and simultaneously uses those components to predict the criterion, which respectively corresponds with the measurement and the structural part of ESEM. The differences and similarities between PCovR and ESEM regarding model parameters, model estimation, and output are discussed. Furthermore, the performance of both methods under different conditions is examined by means of a simulation study.

C256: **Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm**
Presenter: **Xavier Bry**, University Montpellier 2, France
Co-authors: Catherine Trottier, Thomas Verron, Frederic Mortier

In the current estimation of a GLM model, the correlation structure of regressors is not used as the basis on which to lean strong predictive dimensions. Looking for linear combinations of regressors that merely maximize the likelihood of the GLM has two major consequences: 1) collinearity of regressors is a factor of estimation instability, and 2) as predictive dimensions may lean on noise, both predictive and explanatory powers of the model are jeopardized. For a single dependent variable, attempts have been made to adapt PLS Regression, which solves this problem in the classical Linear Model, to GLM estimation. We propose a technique, Supervised Component Generalized Linear Regression (SCGLR), that extends the Fisher Scoring Algorithm so as to combine PLS regression with GLM estimation in the multivariate context. SCGLR is tested on both simulated and real data.

CS06 Room R4: Aph.+Pos. TIME SERIES ANALYSIS III Chair: Rodney Wolff

C096: **Wavelet variance based estimation for composite stochastic processes**
Presenter: **Stephane Guerrier**, University of Geneva, Switzerland
Co-authors: Yannick Stebler, Jan Skaloud, Maria-Pia Victoria-Feser

A new estimation method for the parameters of a times series model is presented. We consider here composite Gaussian processes that are the sum of independent Gaussian processes which in turn explain an important aspect of the time series, as is the case in engineering and natural sciences. The proposed estimation method offers an alternative to classical estimation based on the likelihood that is straightforward to implement and often the only feasible estimation method with complex models. The estimator results as the optimization of a criterion based on a standardized distance between the sample wavelet variances (WV) estimates and the model based WV. Indeed, the WV provides a decomposition of the variance process through different scales, so that they contain the information about different features of the stochastic model. We derive the asymptotic properties of the proposed estimator for inference and perform a simulation study to compare our estimator to the MLE and the LS with different models. We also use it to estimate the stochastic error's parameters of gyroscope composing inertial navigation systems by means of a sample of over 800,000 measurements, for which no other estimation method can be used.

C160: **Advances in automatic time series forecasting**
Presenter: **Rob Hyndman**, Monash University, Australia

Many applications require a large number of time series to be forecast completely automatically. For example, manufacturing companies often require weekly forecasts of demand for thousands of products at dozens of locations in order to plan distribution and maintain suitable inventory stocks. In population forecasting, there are often a few hundred time series to be forecast, representing various components that make up the population dynamics. In these circumstances, it is not feasible for time series models to be developed for each series by an experienced statistician. Instead, an automatic forecasting algorithm is required. Some algorithms recently developed for automatically forecasting various types of time series will be considered, including approaches for handling functional time series, hierarchical time series, and time series with multiple seasonality. Some tools for monitoring automated forecasts and identifying problem series will also be described.

C371: **The pair Levy copula construction**
Presenter: **Stephan Nicklas**, University of Cologne, Germany
Co-authors: Oliver Grothe

Levy processes are common in continuous time modeling, but are mostly applied in low dimensional settings. To capture the dependence structure of Levy processes, Levy copulas have been developed. However, a remaining challenge is to find flexible but still applicable Levy copula models for higher dimensions. We develop a pair construction for Levy copulas from bivariate functions and demonstrate that this provides a flexible framework to model the dependence structure of high-dimensional Levy processes. The basic idea of this approach is inspired by the theory of pair copula constructions for distributional copulas. We provide simulation algorithms and derive a sequential estimator where the asymptotic results are based on the Godambe information matrix.

C409: **Forecasting crude oil price using soft-computing methods and Google insight for search**
Presenter: **Rodney Wolff**, The University of Queensland, Australia
Co-authors: Imad Haidar

We investigate if Google users' search activities can predict short-term crude oil price, detailing a critical analysis of our method. As a pilot study, an abstract list of words was constructed which we believed related to the crude oil market. These words were inserted into Google Insight for Search and each phrase's search query was retrieved. Non-linear AutoRegressive with eXogenous variable (NARX) networks were selected as a forecasting tool for price, because of their capability as a non-linear and universal function approximation model. We further used a hybrid NeuroEvolution-based method, namely the NeuroEvolution for Augmenting Topology (NEAT) combined with a supervised neural network for fine-tuning. We introduce a new fuzzy fitness function for NEAT that takes account of both risk-adjusted return and absolute error. Our empirical results show (i) hybrid-NEAT outperformed NARX networks; (ii) using NARX networks alone, none of the phrases selected from Google Insight for Search helped in improving the forecast for crude oil weekly price; but (iii) we find significant improvements of the sign prediction using error correction hybrid NEAT. Google Insight for Search data, combined with the proposed hybrid approach, appears to provide much needed extra information to achieve better forecasting results for complex time series.

CS12 Room R6: Ath. 3 NONPARAMETRIC STATISTICS I Chair: Ingrid Van Keilegom

C121: **A unifying approach to the shape and change-point hypotheses with unequal spacing of events**
Presenter: **Chihiro Hirotsu**, Meisei University, Japan
Co-authors: Shoichi Yamamoto

A change-point model usually assumes a step-type change at some point of time series. The max acc. t test for the isotonic hypothesis has been shown to be appropriate for detecting the change-point. Each corner vector of the polyhedral cone defined by the isotonic hypothesis corresponds to a component of the change-point model. The max acc. t is essentially the maximal component of the projections of the observation vector on to those corner vectors. It has an advantage in forming the simultaneous confidence intervals for the isotonic contrasts among all the maximal contrast type statistics and also the isotonic regression. The unifying approach to the isotonic and step-type change-point hypotheses has been extended to the convexity and slope change hypotheses. The proposed statistics are the maximal contrast type based on the cumulative and doubly cumulative

sums for the isotonic and convexity hypotheses, respectively. A very efficient and exact algorithm for the probability calculation for the exponential family has been obtained by the Markov property of the serial component statistics. The algorithm is extended further to the unequal spacing of events.

C308: **Comparison of kernel density estimators with assumption on number of modes**
Presenter: **Raphael Coudret**, INRIA and University of Bordeaux, France
Co-authors: Gilles Durrieu, Jerome Saracco

A method to choose the bandwidth of a kernel density estimator assuming it has k modes is described. This requires that the density f to estimate has at most k modes and the relevant bandwidth called critical bandwidth is denoted h_{crit}. When the kernel of the estimator is the Gaussian one, L_1 and L_∞ convergences are proven. These results cannot be extended to the uniform kernel because of a property on h_{crit}. Estimators with both kernels are compared with two others in a simulation study. The first one is the kernel density estimator with a plug-in bandwidth selection method. The second one is based on estimates of density contour clusters which are subspaces of $\mathbb{R}$, depending on a positive real λ, on which $f(t)$ is greater than λ. Estimated densities come from a beta mixture or from a Gaussian mixture called asymmetric claw density. The numerical behavior of these estimators is assessed by computing their integrated absolute error and by measuring the precision of the estimation of the position of the minimum of the local minima of f.

C381: **L-moments under nuisance regression**
Presenter: **Jan Picek**, Technical University of Liberec, Czech Republic

The aim is to extend L-moments in the presence of nuisance regression. L-moments are typically based on the linear combinations of order statistics. The considered extension of L-moments to linear regression models is based on the regression quantiles, which can be seen as a possible analogy to order statistics. The work is motivated by real problems from climatology, where the nuisance regression is considered due to existing trends in the data. The method of L-moments, as an alternative of maximum likelihood or conventional moments, is used for the estimation of the parameters of the underlying error distribution in the linear regression model. A simulation study illustrates the performance of L-moments method under various distributions of the model errors and compares it with the conventional methods.

C185: **An omnibus CUSUM chart for monitoring time to event data**
Presenter: **Ioannis Phinikettos**, Self employed, Cyprus
Co-authors: Axel Gandy

Control charts are an important tool in statistical process control and they were primarily used in manufacturing and process industries. Applications have spread to finance, medicine, biology and reliability. In some monitoring situations, it is of interest to detect smaller persistent changes and not just sudden step changes. In such cases cumulative sum (CUSUM) charts have been one of the main tools for process control. A non-parametric method for monitoring time-to-event data will be introduced. A CUSUM chart is constructed that is able to detect any unknown out-of-control state. This method exploits the absolute differences between the Kaplan-Meier estimator and the in-control distribution over specific time intervals. The efficiency of the algorithm is studied via simulations and tested over existing parametric methods.

CS16 Room R5: Ath.1+2 COMPUTATIONAL ECONOMETRICS IV Chair: Stephen Pollock

C064: **New insights into optimal control of nonlinear dynamic econometric models: Application of a heuristic approach**
Presenter: **Ivan Savin**, Friedrich Schiller University Jena and the Max Planck Institute of Economic, Germany
Co-authors: Dmitri Blueschke, Viktoria Blueschke-Nikolaeva

Optimal control of dynamic econometric models has a wide variety of applications including economic policy relevant issues. There are several algorithms extending the basic case of a linear-quadratic optimization and taking nonlinearity and stochastics into account, but being still limited in a variety of ways, e.g., symmetry of the objective function. To overcome these problems, an alternative approach based on heuristics is suggested. To this end, we apply a 'classical' algorithm (OPTCON) and a heuristic approach (Differential Evolution) to three different econometric models and compare their performance. Among scenarios considered are symmetric and asymmetric quadratic objective functions and different data frequencies between control variables. Results provide a strong support for the heuristic approach encouraging its further application to optimum control problems.

C106: **Empirical pricing kernel estimation using a functional gradient descent algorithm based on splines**
Presenter: **Pirmin Meier**, University of St Gallen, Switzerland
Co-authors: Francesco Audrino

A new methodology to estimate the empirical pricing kernel implied from option data is proposed. In contrast to most of the previous studies that use an indirect approach, i.e. first estimating the physical and risk-neutral densities and obtaining the pricing kernel in a second step, we follow a direct approach. Departing from an adequate parametric and economically motivated pricing kernel, we apply a functional gradient descent (FGD) algorithm based on B-splines. This approach allows us to locally modify the initial pricing kernel and hence to improve the final estimate. We empirically illustrate the estimation properties of the method and test its predictive power on S&P 500 option data, comparing it as well with other recent approaches introduced in the empirical pricing kernel literature.

C286: **A spline dynamic correlation model for high and low frequency correlations**
Presenter: **Vasiliki Skintzi**, University of Peloponnese, Greece

A new approach for modeling the dynamics of correlation between asset returns is proposed. The correlation matrix is decomposed into a high frequency and a low frequency component combining the univariate spline GARCH model and a simple DCC specification. The dynamics of the short-term component of both volatilities and correlations are driven exclusively by GARCH models while the long-term components of volatilities and correlation are slowly changing deterministic functions of time. Conditional correlations are allowed to mean-revert to the slowly time-varying unconditional correlations. The estimated unconditional correlations can be examined in relation with macroeconomic fundamentals. The proposed methodology is used for modeling return correlation between twelve major stock markets. For all pairs of countries the proposed spline dynamic correlation model outperforms the standard DCC model and attains an increase in the log likelihood. Using seemingly unrelated regressions and dynamic panel models the unconditional correlations are found to increase/decrease over time as a result of increases/decreases in macroeconomic factors such as bilateral imports and exports, inflation differential etc.

C295: **MARC-MARS: Modeling asset returns via conditional multivariate asymmetric regime-switching**
Presenter: **Pawel Polak**, University of Zurich and Swiss Finance Institute, Switzerland
Co-authors: Marc Paolella

A new multivariate mixture model for asset returns is motivated and studied. It allows for volatility clustering, excess kurtosis, asymmetry, and dynamics in the dependency between assets over time. It nests several models previously proposed, and is demonstrated to outperform all of them in out-of-sample exercises. An EM algorithm is developed for estimation and is demonstrated to be far faster than existing methods, and thus crucial for use with a large number of assets. The predictive distribution is a (possibly special case of a) multivariate generalized hyperbolic, so that sums of margins (as required for portfolios) is tractable. Each marginal is endowed with a common univariate shock, interpretable as a common market

factor, and this stochastic process has a predictable component. This leads to the new model as being a hybrid of GARCH and stochastic volatility, but without the estimation problems associated with the latter. An extension of the model which allows for different regimes in the dependency is introduced.

CS19 Room R7: Ath. 4 STATISTICS FOR INTERVAL DATA Chair: Gil Gonzalez-Rodriguez

C145: **Modeling interval variables with probabilistic background based on copulas**
Presenter: **Eufrasio Lima Neto**, Federal University of Paraiba, Brazil
Co-authors: Ulisses dos Anjos, Silva Alisson

Sometimes, the available data can be presented as intervals (for instance, the daily temperature or the daily stock price). Recently, different approaches have been proposed to fit a regression model for interval-valued variables. Let consider the interval variable Y = [Y1, Y2] as a bivariate random vector. We suggest the use of copula's theory to propose a general bivariate distribution (GBD) for Y. The copula interval regression method (CIRM) will consider in their random component an GBD distribution. This makes it possible the use of statistical inference techniques for the CIRM approach. The regression structure of the CIRM approach is composed by a systematic component with a set of explanatory interval-valued variables X1, ..., Xp and link functions that connect the systematic component to the mean of the variables Y1 and Y2, that represent the lower and the upper bounds of the intervals or any pair of features that represent an interval (for instance, the center and the spread). The coefficients of the CIRM will be estimated through BFGS algorithm. We will consider synthetic interval-valued data sets based on Monte Carlo framework to evaluate the prediction performance of the new approach in comparison with others methods.

C136: **Some computational properties of possibilistic regression models with interval data**
Presenter: **Michal Cerny**, University of Economics Prague, Czech Republic
Co-authors: Jaromir Antoch, Milan Hladik

Assume a linear regression model with input data X (observations of the independent variables) and output data y (observations of the dependent variable), where some or all data are intervals. We consider both the general case of interval input & interval output models as well as a special case of crisp input & interval output models. The possibilistic OLS set is defined as the set of all OLS estimates of the model as X and y range over the corresponding intervals. We study complexity-theoretic properties of the possibilistic OLS set. In particular, we show that even very basic questions about the set in the general model are coNP-hard. We also show examples of performance of some known methods for construction of enclosures of the OLS sets for particular interval models (referred to as "interval least squares methods"). The examples illustrate that the state-of-the-art methods for interval least squares can hardly be considered as satisfactory. We also show some computational properties of the OLS set in the crisp input & interval output model.

C398: **Modelling an interval-valued variable through some real/interval-valued variables**
Presenter: **Marta Garcia-Barzana**, University of Oviedo, Spain
Co-authors: Angela Blanco-Fernandez, Peter Winker, Henning Fischer

A multiple linear regression model to predict an interval-valued variable in terms of either interval-valued or real-valued variables is introduced. The model is based on the interval arithmetic and the canonical representation of intervals. It allows us to take into account separately the contribution of spreads and mid-points of the explanatory interval variables to estimate the spreads and mid-points of the dependent variable. It extends previously studied simple and multiple regression models for interval data. Some numerical optimization methods are applied to solve the associated least squares estimation problem. The properties of the model are demonstrated through Monte Carlo simulations and an application to real data.

C370: **Minimum Dk-distance estimation: A robust estimator for linear interval time series model**
Presenter: **Ai Han**, Chinese Academy of Sciences, China
Co-authors: Yongmiao Hong, Shouyang Wang

A minimum Dk-distance estimator is developed.The efficiency of the proposed estimator is further improved with a two-stage adaptive strategy. The consistency and the asymptotic normality of the estimator is established. It shows a robust performance in terms of efficiency for various data generating processes of the true interval errors, e.g., bivariate processes for left and right bounds of the random intervals or an interval linear process. Simulation experiments demonstrate the efficiency gain of our estimator compared with the quasi-maximum likelihood estimation (QMLE) in the direction of asymmetric densities such as the Bivariate Mixture densities, as well as the symmetric yet the departure of the innovation distribution from the Bivariate Normal is large, e.g., the Bivariate Student-t densities. Simulation results confirm that the resulting estimator is asymptotically efficient in the situation where the data generating process of true interval errors is Bivarate Normal, the same as MLE. Also the two-stage minimum Dk-distance estimator is more efficient than QMLE as the interval error is captured by bivariate Student-t densities or an interval linear process. A primary one-stage minimum Dk-distance estimation is sufficient for Bivariate Mixture densities.

Thursday 30.08.2012 08:40 - 10:20 Parallel Session I

IS06 Room R1: Demetra SMALL AREA ESTIMATION Chair: Domingo Morales

C151: **Small area estimation for skewed variables**
Presenter: **Alessandra Petrucci**, University of Firenze, Italy
Co-authors: Emanuela Dreassi, Emilia Rocco

Linear mixed models are widely used on small area estimation (SAE). These models include independent area effects to account for between area variation beyond that explained by auxiliary variables. However, these models appear inadequate for variables that have a portion of values equal to zero and a continuous, skewed distribution, with the variance that increase with the mean among the remaining values. This situation arises in many fields of applied research, such as socio-economic, environmental and epidemiological science. The aim is to suggest a small area model for this kind of data. We define a two-part SAE model consisting of a logistic regression model for the probability of a nonzero occurrence and a regression model based on a skewed distribution for the mean of the non-zero value. A real example motivates and illustrates the proposed method.

C154: **Predicting risks in space-time disease mapping**
Presenter: **Lola Ugarte**, Universidad Publica de Navarra, Spain
Co-authors: Tomas Goicoa, Jaione Etxeberria, Ana F. Militino

Cancer mortality risk estimates are essential for planning resource allocation and designing and evaluating cancer prevention and management strategies. However, mortality figures generally become available after a few years, making necessary to develop reliable procedures to provide current and near future mortality risks. A spatio-temporal P-spline model is proposed to be used to provide predictions of mortality/incidence counts. The model is appropriate to capture smooth temporal trends and to predict cancer mortality/ incidence counts in different regions for future years. The prediction mean squared error of the forecast values as well as an appropriate estimator are derived. Spanish prostate cancer mortality data in the period 1975-2008 will be used to illustrate results with a focus on cancer mortality forecasting in 2009-2011.

C098: **Testing for zero variance in the Fay-Herriot model**
Presenter: **Domingo Morales**, University Miguel Hernandez of Elche, Spain
Co-authors: Yolanda Marhuenda, M. del Carmen Pardo

The Fay-Herriot model is a linear mixed model that plays a relevant role in small area estimation (SAE). Under the SAE setup, tools for selecting an adequate model are required. Applied statisticians are often interested on deciding if it is worthwhile to use a mixed effect model instead of a simpler fixed-effect model. This problem is not standard because under the null hypothesis the random effect variance is on the boundary of the parameter space. The likelihood ratio test and the residual likelihood ratio test are proposed and their finite sample distribution are derived. Finally, we analyze their behavior under simulated scenarios and we also apply them to real data in SAE problems.

TS04 Room R8: Era TUTORIAL BY IFCS: MIXTURE MODELS FOR HIGH-DIMENSIONAL DATA Chair: Christophe Croux

C141: **Mixture models for high-dimensional data**
Presenter: **Geoffrey McLachlan**, University of Queensland, Australia

Finite mixture distributions are being increasingly used to model heterogeneous data and to provide a clustering of such data. For multivariate continuous data attention is focussed on mixtures of normal distributions and extensions of the latter, including mixtures of t-distributions for data with longer tails than the multivariate normal. For clustering purposes, the fitting of a g-component mixture model provides a probabilistic clustering of the data into g clusters in terms of the estimated posterior probabilities of component membership of the mixture for the individual data points. An outright clustering is obtained by assigning each data point to the component to which it has the highest (estimated) posterior probability of belonging. We shall focus on the case where the number of experimental units n is comparatively small but the underlying dimension p is extremely large, as, for example, in microarray-based genomics and other high-throughput experimental approaches. We shall consider ways including the use of factor models to reduce the number of parameters in the specification of the component-covariance matrices. The proposed methods are to be demonstrated in their application to some high-dimensional data sets from the bioinformatics literature.

OS08 Room R4: Aph.+Pos. FUZZY CLUSTERING Chair: Ann Maharaj

C301: **Fuzzy clustering model based on operators on a product space of linear spaces**
Presenter: **Mika Sato-Ilic**, University of Tsukuba, Japan

A family of fuzzy clustering models based on a new aggregation operator defined on a product space of linear spaces is presented. The purpose is to control the variability of similarity of objects in the fuzzy clustering models. In order to consider the variability, we propose to exploit an aggregation operator. Although this aggregation operator can represent the variety of the common degree of belongingness of a pair of objects to pairs of fuzzy clusters, the representation of the variability of similarity still has a constraint. That is, the similarity still has to be explained by a linear structure with respect to pairs of fuzzy clusters. This is caused by the metric constraint and the fact that the aggregation is a binary operator. In order to solve this problem, we require a new definition of a function family on a product space of linear spaces and have similar conditions of the aggregation operator to represent the variability of similarity. Therefore, we propose a new aggregation operator, called generalized aggregation operator, which presents a better performance for fuzzy clustering.

C089: **On fuzzy clustering with entropy regularization**
Presenter: **Maria Brigida Ferraro**, Sapienza University of Rome, Italy
Co-authors: Paolo Giordani

The Fuzzy k-Means (FkM) algorithm is a tool for clustering n units in k homogeneous clusters. FkM is able to detect only spherical shaped clusters, hence it may not work properly when clusters have ellipsoidal shapes. For this purpose a variant of FkM is the Gustafson-Kessel (GK) algorithm, which can recognizes the shapes of the cluster by the computation of a covariance matrix for each cluster. The fuzziness of the FkM and GK partitions is tuned by the so-called parameter of fuzziness which is an artificial device lacking a physical meaning. In order to avoid this inconvenience a fuzzy clustering algorithm based on entropy regularization can be used. The idea consists in tuning the amount of fuzziness of the obtained partition by the concept of entropy. Unfortunately, such a clustering algorithm can identify only spherical clusters. In this respect, we introduce a GK-like algorithm with entropy regularization capable to discover ellipsoidal clusters.

C319: **A new biplot procedure with joint classification of objects and variables by fuzzy c-means clustering**
Presenter: **Naoto Yamashita**, Tokyo Institute of Technology, Japan
Co-authors: Shin-ichi Mayekawa

Biplot provides a two-dimensional configuration of a data matrix in which both the rows (objects) and the columns (variables) of the matrix are plotted jointly. However, the biplots with a large number of objects and variables would be difficult to interpret. We developed a new biplot

procedure which facilitates the interpretation of large data matrix. In particular, the objects and variables are classified into a small number of clusters using fuzzy c-means clustering, and resulting clusters are simultaneously biplotted in lower dimensional space. This procedure allows us to capture configurations easily, and, in addition, to grasp the fuzzy memberships of the objects and the variables to the clusters. A simulation study and a real data example were given to show the effectiveness of the proposed procedure.

C066: **A review of feature-based fuzzy clustering techniques for time series**
Presenter: **Ann Maharaj**, Monash University, Australia
Co-authors: Pierpaolo D'Urso

Traditional methods for clustering time series assign them to mutually exclusive groups. Given the dynamic nature of time series over different intervals of time, some time series under consideration may belong to more than one group simultaneously. Clustering based on fuzzy logic will enable the determination of the degree to which a time series may belong to each of several clusters. Traditional and fuzzy cluster analyses are applicable to variables whose values are uncorrelated. Hence, in order to cluster time series data that are usually serially correlated, one needs to extract features from the time series, the values of which are uncorrelated. Several feature-based fuzzy clustering models are presented. In particular the models are based on the time series features of autocorrelations, spectral ordinates and wavelet coefficients and related functions. Comparisons are made between the various models and a determination is made as to which models exhibit superior performance under specific circumstances. Applications showing the benefits of fuzzy clustering methods over traditional clustering methods are presented.

OS17 Room R2: Ares ADVANCES IN COMPUTATIONAL STATISTICS AND DATA ANALYSIS Chair: Elvezio Ronchetti

C270: **Estimating the general linear and SUR models after deleting observations**
Presenter: **Stella Hadjiantoni**, Queen Mary University of London, United Kingdom
Co-authors: Erricos Kontoghiorghes

A new method to estimate the (downdating) problem of removing observations from the general linear model (GLM) after it has been estimated is proposed. It is verified that the solution of the downdated least squares method can be obtained from the estimation of an equivalent GLM, where the original model is updated with the imaginary deleted observations. This updated-GLM has a non-positive definite dispersion matrix which comprises complex covariance values. The model is formulated as a generalized linear least squares problem (GLLSP) and its solution is obtained by employing the generalized QR decomposition (GQRD) using hyperbolic Householder transformations. Previous computations are efficiently utilized. The method is based on hyperbolic reflections but no complex arithmetic is used in practice. The problem of deleting observations from the seemingly unrelated regressions (SUR) model is also considered.

C400: **New results on comparing distributions**
Presenter: **Rand Wilcox**, University of Southern California, United States

New results are described for comparing distributions. The first set of results deal with comparing groups via quantiles. Both independent and dependent groups are considered. Typically, extant techniques assume sampling is from a continuous distribution. There are exceptions, but generally, when sampling from a discrete distribution where tied values are likely, extant methods can perform poorly, even with a large sample size. The main result is that when using the Harrell-Davis estimator, good control over the Type I error probability can be achieved in simulations via a standard percentile bootstrap method, even when there are tied values, provided the sample sizes are not too small. In addition, the methods considered here can have substantially higher power than alternative procedures. Illustrations demonstrate that comparing quantiles can be used to gain a deeper understanding of how groups differ. Finally, two new methods for comparing two independent, discrete random variables are described for the situation where the cardinality of the sample space is small. The first is a global test of the hypothesis that the cell probabilities of two multinomial distributions are equal. The other is a multiple comparison procedure.

C407: **Effect of Influential observations on penalized linear regression estimators**
Presenter: **Karen Kafadar**, Indiana University Bloomington, United States
Co-authors: Guilherme V. Rocha

In current problems (e.g., microarrays, financial data) where the number of variables can greatly exceed the number of observations (*big p, small n*), penalized regression has been advocated as a way to identify informative variables by setting to zero a large subset of the regression coefficients. This approach to *model selection* aims for good fits to the data, but the resulting nonzero coefficients are often interpreted. While much attention has focused on the penalty term, robustness considerations would dictate focus on the loss function, and the usual squared error loss function is well known to be highly sensitive to outliers. Here we examine the effect of influential points (outliers and leverage points) on L1-penalized regression estimators with different loss functions (L2, L1, biweight), with a focus on accuracy and precision of the active coefficients. We show that the conventional L2 loss/L1 penalty (lasso) is sensitive to outliers and heavy-tailed error distributions.

C294: **Algorithms based on directed graphs to find the best grouping for each possible number of clusters**
Presenter: **Cristian Gatu**, Alexandru Ioan Cuza University of Iasi, Romania
Co-authors: Cornel Barna, Erricos John Kontoghiorghes

A graph structure which can be employed to enumerate and evaluate all possibilities to cluster a number of observation points is introduced. Any complete traversal of the graph generates all possible cluster solutions. The structure of the graph is exploited in order to design an efficient branch-and-bound algorithm that finds the optimal solution for each number of clusters without traversing the whole graph. The proposed algorithm is based on exhaustive search and therefore provides the optimum solution. A heuristic version of the branch-and-bound algorithm that improves the execution speed at the expense of the solution quality is also presented. In addition, a p-combination method that considers at each level of the graph not all, but only the best p groupings is also investigated. The Ward method is a special case for $p = 1$. This allows for trading between exploration and computational efficiency. Experimental results are presented and analyzed. The proposed strategies are found to be a viable approach to the clustering problem when the number of observation points is not too large.

CS15 Room R5: Ath.1+2 TIME SERIES ANALYSIS IV Chair: Roland Fried

C364: **Forecasting time series through reconstructed multiple seasonal patterns using empirical mode decomposition**
Presenter: **Farouk Mhamdi**, ENIT, Tunisia
Co-authors: Jean-Michel Poggi, Meriem Jaidane-Saidane

Empirical mode decomposition (EMD) is especially well suited for seasonal time series analysis. Indeed, it represents the initial series as a superposition of oscillatory components (modes) and a final residue. Then, it allows to conveniently select and aggregate some of these modes to recover seasonal components and to extract the trend. Based on these ideas, it is possible to investigate if forecasting through reconstructed multiple seasonal patterns is better than two natural alternative schemes. Namely, direct forecasting original series as well as completely desegregated based forecasting. Finally our method is applied to the daily Tunisian electricity load mid-term forecasting.

C387: **Testing of linearity against MSW type of nonlinearity using autocopulas**
Presenter: **Jana Lencuchova**, Slovak University of Technology in Bratislava, Slovakia

Autocopulas seem to be a useful tool for the description of the lagged interdependence structure of stationary time series. It has become natural to use them because they involve nonlinear dependencies too. We extend a previous work, where the autocopula approach is applied to test of heteroscedasticity in AR-ARCH model, to testing linearity against Markov-switching type of nonlinearity. Using several well-known families of copulas and simulations, we look for the copula which is able to distinguish between linear and Markov-switching model.

C360: **Copula approach to testing of the linearity and remaining nonlinearity in the SETAR models**
Presenter: **Anna Petrickova**, Slovak University of Technology Bratislava, Slovakia

A k-lag autocopula is coupled with a bivariate joint distribution function of the bivariate random vector (Y_t, Y_{t-k}), where k is the time lag of the values of the random variables that generate time series. We utilize the idea of the autocopulas, used for testing of the heteroscedasticity in AR-ARCH models. We apply the test as an alternative approach to testing linearity against SETAR type nonlinearity and the remaining nonlinearity in the SETAR models.

C383: **Sequential monitoring procedures in RCA models**
Presenter: **Zuzana Praskova**, Charles University in Prague, Czech Republic

Sequential monitoring procedures to detect stability of parameters in autoregressive models with random coefficients (RCA) are considered. The problem of detecting a change is solved as a sequential hypotheses testing. Detector test statistics are based either on the quasi-maximum likelihood principle and score statistics, or on cumulative sums of weighted residuals and conditionally weighted least-squares estimators that are computationally more simple. Asymptotic properties of test statistics are studied both under the null hypothesis (no change) and under alternatives, and compared numerically.

CS34 Room R6: Ath. 3 COMPUTATIONAL ECONOMETRICS V Chair: Dick van Dijk

C220: **Portfolio selection through an extremality order**
Presenter: **Henry Laniado**, Universidad Carlos III de Madrid, Spain
Co-authors: Rosa E. Lillo, Juan Romo

Markowitz defined the efficient frontier as the set of the feasible portfolios which cannot be improved in terms of risk and return simultaneously. The aim is to introduce new concepts of efficient frontier depending on some indexes that the investor can choose which can be different to the classical variance- return in Markowitz's model. Feasible portfolios are built with Montecarlo simulations and the new efficient frontiers are estimated by using an extremality order which will also be introduced. The performance of the selection method is illustrated with real data.

C275: **Balancing an ill-behaved national accounts table under power law property hypothesis: a cross-entropy perspective**
Presenter: **Second Bwanakare**, University of Information Technology and Management in Rzeszow, Poland

The Shannon-Kullback-Leibler cross-entropy (SKLCE) is particularly useful when ergodic system inverse problems require a solution. However, in spite of relative success recently met by the approach-particularly in balancing macroeconomic social accounting matrix (SAM), it suffers from its ergodicity character owing to bounding all micro-states of the system to identical odds of appearing. We postulate that economic activity is characterized by complex behavioural interactions between economic agents or/and economic sectors and that, additionally, statistical data gathering survey process itself may locally lead to systematic errors. Consequently to the above, long range correlation of aggregated accounts data, their observed time invariant scale economic structure or Levy- unstable distribution law of disturbances may constitute a rule rather than an exception. Then, we generalize the SKLCE approach and set up a power law -related non extensive cross-entropy econometric model(NECE). The latter is applied to successfully balance a Polish SAM expected to exhibiting warlasian general equilibrium features. The Rao-Cramer-Kullback inferential information indexes are computed. We note that increasing relative weight on disturbance component of the dual criterion function leads to higher values of the q-Tsallis complexity index while smaller disturbance weights produce q values closer to unity, the case of Gaussian distribution.

C361: **Uncovering intraday volatility and cross-correlations: The cases of Nikkei225 and S&P500**
Presenter: **Bahar Ghezelayagh**, University of East Anglia, United Kingdom

The intraday volatility and cross-correlations of two of the world's most important stock market indexes, Nikkei225 and S&P500, are analyzed based on a 9-year sample. The intraday volatility exhibits a doubly U-shaped (i.e. W-shaped) pattern associated with the opening and closing of the separate morning and afternoon trading sessions. Strong intraday periodic patterns are filtered out using a time-frequency technique which is free of selection parameters. The high-frequency returns reveal the existence of long-memory intraday volatility dependencies. The results indicate that stock market volatilities follow different scaling laws at different time horizons and that cross-correlation depends on the time scale examined. The correlation structure diverges when the investment horizon spans over months in contrast to over days. This result is very important for investors when taking into account their investment horizon and when they make risk management decisions based on the correlation structure between indexes.

C385: **Importance sampling based estimation methods for limited value-at-risk and limited conditional tail expectation measures**
Presenter: **Silvia Dedu**, Bucharest University of Economic Studies, Romania
Co-authors: Vasile Preda

We propose Importance Sampling based methods for estimating Limited Value-at-Risk and Limited Conditional Tail Expectation measures. The aggregate loss corresponding to the Stop-Loss reinsurance model with multiple retention levels is computed, using Limited Value-at-Risk and Limited Conditional Tail Expectation. Optimization problems associated to this reinsurance model are formulated and the existence of the optimal solution is investigated. Necessary and sufficient conditions for the existence of the optimal retention levels are obtained. Computational results are provided.

CS30 Room R7: Ath. 4 SURVIVAL ANALYSIS Chair: Martina Mittlboeck

C264: **A hidden competing risk model for survival data with a cured fraction**
Presenter: **Polychronis Economou**, University of Patras, Greece
Co-authors: Chrys Caroni

In survival and reliability studies there often appears to be a fraction of individuals who will not experience the event of interest even if followed for a very long period. These individuals are usually treated as they were never actually at risk (cured fraction models), despite the fact that this assumption is not always realistic. To describe such situations a new model for lifetime data is proposed in which all units start out susceptible to the event of interest, but can move into a non-susceptible group if another event intervenes. This model also results in the appearance of a cured fraction, but the membership of this group is not fixed from the outset as in cured fraction models. Likelihood ratio tests and diagnostic plots are proposed for the proposed model and examples of applications are provided.

C225: **Nonparametric modelling of clustered survival data**
Presenter: **Iris Ivy Gauran**, University of the Philippines Diliman, Philippines
Co-authors: Erniel Barrios

A nonparametric regression model to characterize clustered survival data is postulated. It incorporates the random clustering effect into the Cox Proportional Hazards model. The model is subsequently estimated via the backfitting algorithm. The simulation study yields evidence that clustered survival data can be better characterized in a nonparametric model. Predictive accuracy in the nonparametric model increases with the number of clusters. The distribution of the random component intended to account for clustering effect also contributes into the model. As the functional form of the covariate departs from linearity, the nonparametric model is becoming more advantageous over the parametric method. Furthermore, the nonparametric model is better than the parametric model when the data is highly heterogeneous and/or there is misspecification error.

C240: **Critical illness diagnosis rates by causes in stratified sampling setting**
Presenter: **Valeria D'Amato**, Salerno, Italy
Co-authors: Maria Russolillo

Mortality forecasts are used in a wide variety of academic fields, and for global and national health policy making, medical and pharmaceutical research, and social security and retirement planning. According to the World Health Organization a "right" detection of the causes of death is vital for forecasting more accurately mortality. An overview of mortality projection by cause of death has been previously done. Also a Bayesian model to use information on causes of death to estimate more accurately mortality has been proposed. Furthermore, the US Social Security Administration carries out projections by cause of death, as described by Social Security and Medicare Boards of Trustees. The aim is to propose a methodology for deriving the diagnosis rates by resorting to an analysis by cause. In fact, for the insurance company management perspective it is essential to avoid mismatch between the actual settled claims with expected diagnosed claims by accurately evaluating the diagnosis rates, in order to measure the actual future outflows. In this order of ideas, the research focuses on investigating computational methods for performing a survival analysis in the framework of the stratified sampling technique. In particular, we propose a methodology for deriving the diagnosis rates by resorting to an analysis by cause and we suggest to express the diagnosis rates for the specific illness by estimating the cause-specific survival probabilities. To this aim, we model the number of deaths within a generalised Lee Carter framework with a Poisson error structure and propose a stratification by death causes, to reduce standard error by providing some control over variance. The empirical results are presented using a range of numerical and graphical analyses.

C312: **Graphical representation of survival probabilities for a binary time-dependent covariate**
Presenter: **Martina Mittlboeck**, Medical Universtiy of Vienna, Austria
Co-authors: Ulrike Poetschger, Harald Heinzl

The time-dependent nature of covariates for a time-to-event outcome has to be properly addressed in the statistical analysis. In the case of a binary time-dependent covariate indicating a non-reversible change, effect estimation and testing can be easily performed within the Cox regression model using standard statistical software. Surprisingly, graphical representation of survival curves is still not satisfactorily solved although several suggestions have been made so far. Among others, a comparison of patients survival whose covariate has non-reversibly changed over time to the survival of all patients has been suggested. Another approach is the so-called landmark-method, where a predefined time interval is skipped and the starting point for survival curves is chosen so that in a sufficient number of patients a change in the time-dependent covariate has already been observed. A new approach is suggested, which combines Cox model estimates with the landmark-method. The statistical properties of the mentioned methods are investigated by a computer simulation study and by means of the Stanford heart transplant data set.

Thursday 30.08.2012 13:40 - 15:45 Parallel Session L

IS08 Room R1: Demetra NEW DEVELOPMENTS IN COMPUTATIONAL ECONOMETRICS Chair: Dick van Dijk

C297: **Large-scale accurate multivariate return density forecasting: A survey and new results**
Presenter: **Marc Paolella**, University of Zurich, Switzerland

Accurate prediction of the entire multivariate distribution of asset returns is useful for numerous purposes. One particularly important application is the generation of the distribution of a portfolio, i.e., a weighted sum of the marginals. This in turn can be used for risk management and portfolio optimization with respect to any chosen risk measure (value at risk, expected shortfall, etc.). We discuss several new methods for doing this, all of which can be used with a large set of assets, all of which deal with fat-tails, heteroskedasticity (in different ways) and the dynamics in the dependency between assets, and all of which lead to superior out-of-sample density forecasts compared to use of existing applicable GARCH models, such as CCC, DCC, and their various extensions. The methods include use of copulae, mixture distributions, regime switching structures, ICA-based algorithms, and multivariate densities with common market factors. The particular benefits and disadvantages of each model will be discussed.

C359: **Model uncertainty and forecast combination in high dimensional multivariate volatility prediction**
Presenter: **Alessandra Amendola**, University of Salerno, Italy
Co-authors: Giuseppe Storti

The identification of the optimal forecasting model for multivariate volatility prediction is not always feasible due to the curse of dimensionality typically affecting multivariate volatility models. In practice only a subset of the potentially available models can be effectively estimated after imposing severe constraints on the dynamic structure of the volatility process. This situation leaves scope for the application of forecast combination strategies as a tool for improving the predictive accuracy. The aim is to propose some alternative combination strategies and compare their performances in forecasting high dimensional multivariate conditional covariance matrices for a portfolio of US stock returns. In particular, we will consider the combination of volatility predictions generated by multivariate GARCH models, based on daily returns, and dynamic models for realized covariance matrices, built from intra-daily returns.

C333: **Macroeconomic and financial forecasting with a factor augmented smooth transition regression model**
Presenter: **Dick van Dijk**, Erasmus University Rotterdam, Netherlands

The smooth transition (auto)regressive [ST(A)R] model is a useful device for modelling and forecasting nonlinear time series variables. Empirical applications in macroeconomics and finance typically involve the univariate STAR model or the multivariate STR model but with a rather limited set of variables. We consider the use of a large number of predictors in the smooth transtion framework by means of a Factor Augmented STR [FASTR] model. The conventional STR modeling cycle is adapted to this context, and we consider empirical applications in macro and finance.

OS02 Room R8: Era ADVANCES IN THE ANALYSIS OF COMPLEX DATA Chair: Maria Paula Brito

C099: **Quantile visualisation for complex data analysis**
Presenter: **Monique Noirhomme**, Namur University, Belgium
Co-authors: Teh Amouh, Paula Brito

In symbolic data analysis, each data unit is set-valued. An interval, a histogram or a set of categories are examples of data units that can be found in a cell of a symbolic data table. The quantile representation of each of these data units provides a common framework for representing symbolic data described by features of different types. Basing on the quantile representation of symbolic data, some authors have proposed methods for PCA, clustering, and other data analysis tasks. Visual presentation allows for fast and effective information communication. In order to allow users to compare symbolic data at a glance, we suggest a technique for the simultaneous visualization of quantiles from different symbolic data. Two approaches were examined for plotting quantiles. The first approach is a boxplot-like representation. Regarding the task of comparing different distributions, the problem with the box plot visualization is that it does not make explicit any information about the proportions, as the original goal of the box plot is to highlight outliers rather than distributions. We therefore consider a second approach based on approximated distribution functions simultaneously plotted on the same graph. In order to improve the appearance of the graph and allow for visual comparison of symbolic data, our visualization manages to minimize the overlapping of the graphical elements used to plot the approximated distributions.

C226: **Batch self-organizing maps based on city-block distances for interval variables**
Presenter: **Francisco de Assis Tenorio de Carvalho**, Universidade Federal de Pernambuco - UFPE, Brazil
Co-authors: Filipe M. de Melo, Patrice Bertrand

The Kohonen Self Organizing Map (SOM) is an unsupervised neural network method with a competitive learning strategy which has both clustering and visualization properties. Interval-valued data arise in practical situations such as recording monthly interval temperatures at meteorological stations, daily interval stock prices, etc. Batch SOM algorithms based on adaptive and non-adaptive city-block distances, suitable for objects described by interval-valued variables, that, for a fixed epoch, optimize a cost function, are presented. The performance, robustness and usefulness of these SOM algorithms are illustrated with real interval-valued data sets.

C209: **The data accumulation PCA to analyze periodically summarized multiple data tables**
Presenter: **Manabu Ichino**, Tokyo Denki University, College of Science and Engineering, Portugal
Co-authors: Paula Brito

A vastly many official statistics have been opened to the public by various media including the Internet. Among them, we often encounter periodically summarized data tables. This paper presents the data accumulation method of PCA to analyze periodically summarized multiple data tables at once. When we have n periodically summarized (N objects)$\times$(d features) data tables, the data accumulation method transforms these data tables to a single ($N \times m$ sub-objects)$times$(d features) standard numerical data table, where m is a preselected integer number according to the type of the n given data tables. Then, we apply the standard PCA to the transformed data table. For each of N objects, the data accumulation guarantees the monotone property of m sub-objects in the $d-$dimensional feature space. In the factor planes by the PCA, each object is reproduced as a series of arrow lines that connect m sub-objects. We present several experimental results in order to show the effectiveness of the proposed data accumulation method.

C195: **Linear regression models in data frameworks with variability**
Presenter: **Paula Brito**, Universidade Porto, Portugal
Co-authors: Sonia Dias

In the classical data framework one numerical value or one category is associated to each individual (micro data). However, the interest of many studies is based in groups of records gathered according to a set of characteristics of the individuals, leading to macro-data. The classical solution for these situations is to associate to each 'higher-level' unit the mean or the mode of the corresponding records; however with this option the variability across the records is lost. For such situations, Symbolic Data Analysis proposes that to each 'higher-level' unit is associated the

distribution or the interval of the individual records' values. Accordingly, it is necessary to adapt concepts and methods of classical statistics to different kinds of variables. One such type of symbolic variable are histogram-valued variables, where to each group corresponds a distribution that can be represented by a histogram or a quantile function. We propose new linear regression models for histogram-valued data - Histogram and Symmetric Histogram Models - where distributions of values can be explained by a linear regression on other distributions. Examples on real data as well as simulated experiments illustrate the behavior of the proposed models, and a goodness–of–fit measure shows the quality of the forecast distributions.

OS16 Room R2: Ares ERCIM SESSION ON COMPUTATIONAL AND NUMERICAL METHODS IN STATISTICS Chair: Ana Colubi

C082: **PORT-PPWM extreme value index estimation**
Presenter: **Ivette Gomes**, University of Lisbon, Portugal
Co-authors: Frederico Caeiro, Ligia Henriques-Rodrigues

Making use of the peaks over random threshold (PORT) methodology and the Pareto probability weighted moments (PPWM) of the largest observations, and moreover dealing with the extreme value index (EVI), the primary parameter in statistics of extremes, new classes of location-invariant EVI-estimators are built. For finite samples, and through a Monte-Carlo simulation study, these estimators, the so-called PORT-PPWM EVI-estimators, are compared with the generalized Pareto probability weighted moments (GPPWM) and a recent class of minimum-variance reduced-bias (MVRB) EVI-estimators.

C131: **A goodness-of-fit test for a class of bivariate Laplace distributions**
Presenter: **Virtudes Alba-Fernandez**, University of Jaen, Spain
Co-authors: M.Dolores Estudillo-Martinez, M.Dolores Jimenez-Gamero

A goodness-of-fit test for a class of bivariate Laplace distributions is proposed and studied. The distributions in this class are obtained as differences of two Moran-Downton bivariate exponential distributions. The resultant distributions have correlated univariate Laplace marginals. The class of distributions considered can be successfully used to model financial, biological or engineering data. To construct a goodness-of-fit test for this family, we consider a test statistic that takes advantage of the convenient formula of the characteristic function of this distribution and compares it with the empirical characteristic function of the sample. Large sample properties of the proposed test are studied. Some numerical results are reported which study the finite sample performance of the proposed test, as well as some real data set applications which illustrate the practical application of the proposed technique.

C266: **Functional depth based normalization of microarray probe intensities for RMA analysis and outlier detection**
Presenter: **Alicia Nieto-Reyes**, Universidad de Cantabria, Spain
Co-authors: Javier Cabrera

Microarray normalization is a standard step of the RMA method to convert probe intensities into gene expressions. We propose a new method to apply the idea of functional depth to the problem of microarray normalization. The method consists in making the distributions of the gene expression as similar as possible across the microarrays of the sample by making it similar to the distribution of gene expressions of the deepest array. In addition this method is applicable to other normalization problems in genomics and elsewhere. Microarray data are known for containing outliers. Although seeking for the outlier genes in a given microarray has being broadly studied, it does not apply for detecting microarray outliers, lying the difficulty on the nature of the data: small samples and high dimensional spaces. The proposed procedure is analytical and it is based on the functional depth and the Tukey's concept of outlier. We compare it with some graphical procedures. We will illustrate this with examples.

C238: **3-dimensional Archimax copulas and their fitting to real data**
Presenter: **Radko Mesiar**, STU Bratislava, Slovakia
Co-authors: Tomas Bacigal

Based on a partition-based approach and generalized convex combination method, we introduce a special class of dependence functions known from EV-copulas theory. These dependence functions are a basis for Archimax copulas we use for fitting to capture the dependence structure of random variables. Examples arising from real 3-dimensional data illustrate our approach.

C408: **A GSVD strategy for estimating the simultaneous equations model**
Presenter: **Mircea Ioan Cosbuc**, Alexandru Ioan Cuza University of Iasi, Romania
Co-authors: Cristian Gatu, Erricos John Kontoghiorghes

A strategy for estimating the Simultaneous Equations Model with non full rank variance-covariance matrix is considered. The Generalized Singular Value Decomposition is then main tool used in the estimation. The block diagonal and banded structures of the matrices involved in the factorization are exploited in order to reduce the computational burden.

OS18 Room R5: Ath.1+2 SFDS SESSION ON CO-CLUSTERING METHODS AND THEIR APPLICATIONS Chair: Mohamed Nadif

C328: **Model selection in block clustering by the integrated classification likelihood**
Presenter: **Aurore Lomet**, Universite de Technologie de Compiegne, France
Co-authors: Gerard Govaert, Yves Grandvalet

Block clustering (or co-clustering) aims at simultaneously partitioning the rows and columns of a data table to reveal homogeneous block structures. This structure can stem from the latent block model which provides a probabilistic modeling of data tables whose block pattern is defined from the row and column classes. For continuous data, each table entry is typically assumed to follow a Gaussian distribution. For a given data table, several candidate models are usually examined: they may differ in the numbers of clusters or in the number of free parameters. Model selection then becomes a critical issue, for which the tools that have been derived for model-based one-way clustering need to be adapted. In one-way clustering, most selection criteria are based on asymptotical considerations that are difficult to render in block clustering due to dual nature of rows and columns. We circumvent this problem by developing a non-asymptotic criterion based on the Integrated Classification Likelihood. This criterion can be computed in closed form once a proper prior distribution has been defined on the parameters. The experimental results show steady performances for medium to large data tables with well-separated and moderately-separated clusters.

C254: **Model selection for the binary latent block model**
Presenter: **Christine Keribin**, Universite Paris Sud, France
Co-authors: Vincent Brault, Gilles Celeux, Gerard Govaert

The latent block model is a mixture model that can be used to deal with the simultaneous clustering of rows and columns of an observed numerical matrix, known as co-clustering. For this mixture model unfortunately, neither the likelihood, nor the EM algorithm are numerically tractable, due to the dependence of the rows and columns into the label joint distribution conditionally to the observations. Several approaches can be considered to compute approximated solutions, for the maximum likelihood estimator as well as for the likelihood itself. The comparison of a determinist approach using a variational principle with a stochastic approach using a MCMC algorithm is first discussed and applied in the context of binary

data. These results are then used to build and compute ICL and BIC criteria for model selection. Numerical experiments show the interest of this approach in model selection and data reduction.

C362: **Co-clustering by bi-stochastic matrix**
Presenter: **Lazhar Labiod**, Paris Descartes University, France
Co-authors: Mohamed Nadif

Motivated by several applications in text mining, market-basket analysis, and bioinformatics, the co-clustering problem has attracted much attention in the past few years. The matrix approximation approaches like SVD and NMF, for instance, have recently been shown to be useful and effective to tackle this problem. We embed the co-clustering in a Bistochastic Matrix Approximation (BMA) framework and we derive from the double kmeans objective function a new formulation of the criterion. We also show that the double kmeans is equivalent to algebraic problem of BMA under some suitable constraints. To optimize it, an iterative process seeking for the optimal simultaneous partitions of rows and columns data is proposed, the solution is given as the steady state of a Markov chain process. We develop two iterative algorithms: the first one leads to learn rows and columns similarities matrices and the second one to obtain the rows and columns partitions. Numerical experiments on simulated and real datasets demonstrate the interest of our approach.

C287: **Mutual information and chi-square in co-clustering**
Presenter: **Mohamed Nadif**, University of Paris Descartes, France
Co-authors: Gerard Govaert

Many classical clustering procedures aim to construct separately an optimal partition of objects or, sometimes, of variables; in contrast the co-clustering methods consider simultaneously the two sets and organize the data into homogeneous blocks. This kind of methods has practical importance in a wide of variety of applications such as text and market basket data analysis. Typically, the data that arise in these applications are arranged as two-way contingency tables. To take into account this type of data, we have recently proposed a latent block model using Poisson distributions and, setting it under the classification approach, we have developed a Block Classification EM algorithm maximizing a complete data log-likelihood. We show that under some constraints this criterion corresponds to mutual information and is approximatively equal to chi-square associated to a couple of partitions. Comparisons and evaluations of the different algorithms used, including the block EM algorithm, are performed on simulated and real data sets.

OS19 Room R4: Aph.+Pos. BRANCHING MODELS, DERIVED MODELS, AND THEIR APPLICATIONS Chair: Manuel Molina

C340: **Simulation, estimation and robustification against outliers in branching processes: A computational approach**
Presenter: **Vessela Stoimenova**, Sofia University, Bulgaria
Co-authors: Dimitar Atanasov

The estimation of the parameters of branching processes (BP) is an important issue used for studying and predicting their behavior. In this field of prime interest is to study the asymptotic behaviour of the different types of parameter estimators. An interesting technique for BP robust estimation is to combine these asymptotic results with a generic method for constructing robust estimators, based on the trimmed likelihood and called weighted least trimmed estimators (WLTE). The procedure is computationally intensive, but gives well interpretable results even in the case of minor a priori satisfied asymptotic requirements. For a better understanding the main outlines of this routine are explained and some classical estimators and their robust modifications in classes of discrete-time branching processes are shown. In this regard a software package for MATLAB for simulation, plotting and estimation of the process parameters is presented. The package is available on the Internet, under the GNU License.

C088: **ABC algorithms for bisexual multitype branching models in a genetic context**
Presenter: **Manuel Mota**, University of Extremadura, Spain
Co-authors: Miguel Gonzalez, Cristina Gutierrez, Rodrigo Martinez

Branching Processes have shown to be useful models in the field of Population Genetics. In particular, some Multitype Bisexual Branching Processes can explain accurately some aspects related to the extinction-fixation-coexistence of some alleles of sex-linked genes. Some of these models are complicated with many underlying parameters. ABC methods could be appropriated to address some inferential problems related to these models. Assuming the number of females and males of each phenotype in several generations is available, rejection inference algorithms for estimating the posterior densities of these parameters can be proposed. In these algorithms, most of the available information, such as the number of couples, the mating scheme, etc., should be used. We study how to implement all this information in the algorithms and to what extent the accuracy of the estimates is improved. A sensitivity analysis is also carried out, with discussion about the choice of the prior distributions, the metric or the tolerance.

C102: **Some methodological and computational results in the class of birth-death branching models**
Presenter: **Alfonso Ramos**, University of Extremadura, Spain
Co-authors: Casimiro Corbacho, Manuel Molina, Manuel Mota

Branching models have a long history of biological applications, particularly in population dynamics. Our interest is the development of mathematical models to describe the demographic dynamics of socially structured animal populations, focusing our attention on lineages, usually matrilines, as the basic structure in the population. We introduce new classes of birth-death branching models where we assume that both, the offspring probability distribution and the death probabilities may be different in each generation, changing either predictably or unpredictably in relation to habitat features. We consider the genealogical tree generated by observation of the process until a pre-set generation. We determine the probability distributions of the random variables representing the number of dead or living individuals having at least one ancestor alive, living individuals whose ancestors are all dead, and dead individuals whose ancestors are all dead, explicitly obtaining their principal moments. Also, we derive the probability distributions corresponding to the partial and total numbers of such biological variables, obtaining in particular the distribution of the total number of matriarchs in the genealogical tree. Using the computer software Mathematica and the R language and environment for statistical computing and graphics, we apply the proposed models to describe the demographic dynamics of African elephant populations living in different habitats.

C200: **Skeletons of near-critical Bienayme-Galton-Watson processes**
Presenter: **Maria Conceicao Serra**, Minho University, Portugal
Co-authors: Serik Sagitov

Skeletons of branching processes are defined as trees of lineages characterized by an appropriate feature that ensures future reproduction success. In the supercritical case a natural choice is to look for the lineages that survive forever. In the critical case it was earlier suggested to distinguish the particles with the total number of descendants exceeding a certain threshold. These two definitions lead to asymptotic representations of the skeletons as either pure birth process (in the slightly supercritical case) or critical birth-death processes (in the critical case conditioned on exceeding a high threshold value). The limit skeletons reveal typical survival scenarios for the underlying branching processes. We consider near-critical Bienaymé-Galton-Watson processes proposing a flexible way for building the skeletons. In the single type case, each vertex of the family tree is independently marked with a small probability. The branch connecting the root with a marked vertex is called a marked branch. The

marked branches form a subtree of the family tree of the branching process and this will be called a skeleton. Such a skeleton is approximated, when marking is a rare event, by a birth-death process. Some simulations results illustrating the limit behaviour will be presented also.

C190: **Estimation of the infection parameter in the different phases of an epidemic modeled by a branching process**
Presenter: **Sophie Penisson**, Universite Paris-Est Creteil, France
Co-authors: Christine Jacob

We build and compare estimators of the infection parameter in the different phases of an epidemic (growth and extinction phases). The epidemic is modeled by a Markovian process of order $d \geq 1$, and can be written as a multitype branching process. We propose three estimators suitable for the different classes of criticality of the process, in particular for the subcritical case corresponding to the extinction phase. We prove their consistency and asymptotic normality for two asymptotics, when the number of ancestors (resp. number of generations) tends to infinity. We finally illustrate our results with the study of the infection intensity in the extinction phase of the BSE epidemic in Great-Britain.

C075: **Some inferential and computational contributions for the class of two-sex branching models with progenitor couples in a random environment**
Presenter: **Manuel Molina**, University of Extremadura, Spain
Co-authors: Manuel Mota, Alfonso Ramos

In the overall context of stochastic modeling, branching processes are widely used as appropriate mathematical models to describe the probabilistic behavior of systems whose components after a certain life period reproduce and die. Branching process theory has become an active research area of interest and applicability. It has especially played a major role in modeling general population dynamics. We are particularly interested in the development of stochastic models to describe the demographic dynamics of biological populations with sexual reproduction. To this end several classes of two-sex branching models have been studied. We continue the research about the class of two-sex branching models with progenitor couples in a random environment. Under a nonparametric setting, we determine Bayesian estimators for the offspring probability distribution and for its main moments. Also, we present the software in R that we have developed for its study and provide a computational method to determine the highest posterior density credibility sets for the main parameters involved in the probability model. By way of illustration, we include an application in population dynamics.

CS18 Room R6: Ath. 3 STATISTICAL SOFTWARE Chair: Peter Filzmoser

C338: **Electronic statistics textbook by EPUB 3.0**
Presenter: **Liang Zhang**, Okayama University, Japan
Co-authors: Masaya Iizuka, Tomoyuki Tarumi

Recently, there has been a rapid increase in the number of e-book readers, tablet PC, touch screen devices, and electronic books. There are standard file formats for e-books, such as PNG (Portable Network Graphics) and JPEG (Joint Photographic Experts Group) formats for images, and general document formats such as PDF (Portable Document Format) or text files. To teach the concept of statistics, it is effective to illustrate the statistics material in an interactive manner. In this study, we used the EPUB format to develop a statistics electronic textbook that has a e-book standard. We have some reasons why we develop the e-book by applying EPUB format. One of the reason is the EPUB 3.0 format conforms to the latest Web standards, including HTML5, JavaScript, and CSS3. Second, EPUB is compatible with various multimedia formats. Third, EPUB has high scalability of expression for distribution in a single-file format. EPUB is designed for reflowable content, which means that the text, illustrations, and photographs, etc. can be optimized for whichever display device is used by the reader of the EPUB-formatted book, although EPUB also supports fixed-layout content. To this end, we used these advantages and digitized a book to the subject of statistics. We configured the electronic textbook to have integrated content, including presentation handouts, animations, exercises, etc. The aim of using these teaching materials is to increase the appeal and effectiveness of the subject area. We selected a required statistics topics from which the contents were obtained. The electronic textbook that was developed can be used in self-study, as well as during classroom lessons. In this study, we develop the e-textbook including animation, video, and teaching materials together with simulation, which was developed by Project CASE(Computer Assisted Statistical Education). We first present an overview of the EPUB and EPUB 3.0 specification, and describe the electronic textbook for which it is being used to teach statistics.

C347: **Muste - editorial computing environment within R**
Presenter: **Reijo Sund**, National Institute for Health and Welfare, Finland
Co-authors: Kimmo Vehkalahti, Seppo Mustonen

Statistical computing requires appropriate tools. We describe the integrated editorial environment for data analysis and related tasks developed since the 1960s under the name of Survo and its recent open source R package implementation known as Muste. As examples of the editorial approach in Muste, we consider a statistical application on influence curves for a correlation coefficient as well as on simulation and contour curves of bivariate Gaussian data.

C356: **Visual time series analysis**
Presenter: **Paul Fischer**, Technical University of Danmark, Denmark
Co-authors: Astrid Hilbert

We introduce a platform which supplies an easy-to-handle, interactive, and fast analysis tool for time series analysis. In contrast to other softwares suits like Maple, Mathlab, or R, which use a command-line-like interface and where the user has to memorize/look-up the appropriate commands, our application is select-and-click-driven. It allows us to derive many different sequences of deviations for a given time series and visualize them in different ways in order to judge their expressive power. For many transformations or model-fits, the user may choose between manual and automated parameter selection. The user can define new transformations and add them to the system. The application contains efficient implementations of advanced and recent techniques for time series analysis including techniques related to extreme value analysis and filtering theory. It has been successfully applied to time series in economics, e.g reinsurance, and to vibrational stress data for machinery. The software is web-deployed, but runs on the user's machine, allowing to process sensitive data locally without having to send it away. The software can be accessed under http://www.imm.dtu.dk/ paf/TSA/launch.html.

C132: **JMP for visual analytics**
Presenter: **Volker Kraft**, SAS Institute, Germany

JMP is a desktop product from SAS, with hundreds of man-years of development. Its particular strength lies in fostering a visual approach to analysis, important as a precursor to model building, when the focus of your analysis is in exploration and hypothesis generation, but also in model interpretation. As well as being very capable in its own right, JMP can work with both R and SAS, either in an ad-hoc fashion or by building custom applications that exploit the unique strengths of JMP. Some real-world examples are shown.

CS29 Room R7: Ath. 4 CONTRIBUTIONS IN COMPUTATIONAL STATISTICS Chair: Domingo Morales

C201: **Computational aspects of discrepancies for equidistribution on the hypercube**
Presenter: **Raffaello Seri**, University of Insubria, Italy
Co-authors: Christine Choirat

We study the asymptotic statistical properties of some discrepancies defined on the unit hypercube, originally introduced in Numerical Analysis to assess the equidistribution of low-discrepancy sequences. We show that they have highly desirable properties. Nevertheless, it turns out that the limiting distribution is an (infinite) weighted sum of chi-squared random variables. This raises some problems concerning the approximation of the asymptotic distribution. These issues are considered in detail: several solutions are proposed and compared, and bounds for the approximation error are discussed.

C267: **Multitask learning in mixture modelling framework via generalized linear mixed-effects models**
Presenter: **Shu-Kay Ng**, Griffith University, Australia
Co-authors: Alfred Lam

Many real-world problems can be considered as a series of related tasks. For example, related tasks are to predict survival of patients from different hospitals. In these multitask problems, the data collected could exhibit a clustered structure due to the relatedness between multiple tasks. Mixture model-based methods assuming independence may not be valid for regression and cluster analyses of data arisen from multiple related tasks. Multitask learning is an inductive transfer mechanism to improve generalization accuracy by sharing task-specific information from different tasks to improve the learning process. This multitask learning mechanism is extended for mixtures of generalized linear models via random-effects modelling to handle multitask problems. The use of random-effects models implies that a soft sharing mechanism is adopted to leverage task-specific information from multiple tasks. The proposed method is illustrated using simulated and real data sets from various scientific fields.

C313: **Drift mining in data: A framework for addressing drift in classification**
Presenter: **Vera Hofer**, University of Graz, Austria
Co-authors: Georg Krempl

A new statistical approach for analysing population drift in classification is introduced. Such drift denotes changes in the joint distribution of explanatory variables and class label over time. If the posterior distribution has changed, the decision boundary has altered and a classifier's performance is affected. The drift mining approach presented aims at detecting such changes over time. It helps either to understand evolution in the data from an ex-post perspective, or ex-ante, to anticipate changes in the joint distribution. The latter aspects play an important role in the presence of verification latency, i.e. when recent labelled data is not available to re-estimate a classification model after a drift. The proposed drift mining technique is based on the assumption that the class prior changes by a factor from one point of time to the next one, and that the conditional distributions do not change within this time period. Thus, the conditional distributions can be estimated from recent labelled data, and then be used to express the unconditional distribution as a mixture of the conditional distributions. This allows an easy and fast estimation of the prior change in the presence of verification latency in subsequent years by comparing the mixture of the unconditional distributions to the estimation of the unconditional distribution found from new unlabelled data. The usefulness of this drift mining approach is demonstrated using a real-world dataset from the area of credit scoring.

C227: **Test of mean difference for longitudinal data using stationary bootstrap**
Presenter: **Hirohito Sakurai**, National Center for University Entrance Examinations, Japan
Co-authors: Masaaki Taguri

We propose a testing method for detecting the difference of two mean curves in longitudinal data using the stationary bootstrap when the data of two groups are not paired. The stationary bootstrap is used to approximate the null distributions of test statistics. For the test, we here consider the following four types of test statistics: (1) sum of absolute values of difference between two mean sequences, (2) sum of squares of difference between two mean sequences, (3) estimator of area-difference between two mean curves, and (4) difference of kernel estimators based on two mean sequences. Monte Carlo simulations are carried out in order to examine the sizes and powers of the proposed tests. We also show an example of how to use the above method for analyzing a real data.

C368: **How to choose threshold in a POT model**
Presenter: **Martin Schindler**, Technical University of Liberec, Czech Republic
Co-authors: Jan Kysely, Jan Picek

The peaks-over-threshold (POT) method with a nonstationary threshold for estimating high quantiles (return levels) is investigated. It was shown that using (95%) regression quantile as the time-dependent threshold instead of a constant treshold can be beneficial. It is assumed that a linear trend is present in the data and so a linear regression quantile as the threshold is used. The aim is to find the threshold (regression quantile) which would be optimal with respect to the reliability of the estimates of high quantiles by means of Monte Carlo simulations. Based on this criterion stationary and regression quantile thresholds are compared. It is described how the choice of the optimal threshold depends on the sample size, estimated quantile or the estimate itself.

Thursday 30.08.2012 16:15 - 17:30 Parallel Session M

OS13 Room R1: Demetra GENERALIZED CANONICAL CORRELATION ANALYSIS Chair: Michel van de Velden

C057: **Between-group metrics and their use in canonical variate analysis**
Presenter: **Casper Albers**, University of Groningen, Netherlands
Co-authors: John Gower

Canonical analysis deals with measurements on p variables for n samples in k groups. In canonical variate analysis, we aim to represent the data matrix X in a lower-dimensional space. This can be done by optimising the ratio form, or by constrained optimisation. In classical situations, these methods coincide. When there are more variables than samples, they generally do not. A method to generalise canonical variate analysis to this case will be represented. Also, as well as the usual canonical means in the range-space of the within-groups dispersion matrix, canonical means may be defined in its null space. In the range space, we have the usual Mahalonobis metric; in the null space explicit expressions are given and interpreted for a new metric.

C103: **Hierarchical relationships among multiple correspondence, nonmetric component, and PCA in three types of least squares formulations**
Presenter: **Kohei Adachi**, Osaka University, Japan
Co-authors: Takashi Murakami, Henk A. L. Kiers, Jos M. F. ten Berge

Multiple correspondence analysis (MCA) can be formulated in three different manners, which are distinguished by least squares loss functions to be minimized. One of the formulations is the minimization of loss of homogeneity (HM) and another is to regard MCA as a correspondence (CS) analysis problem whose loss function is defined for contingency tables. The third is to formulate MCA as a score-loading approximation (SL), that is, as approximating a quantified indicator matrix by the product of component score and loading matrices. However, the equivalence of the solutions between SL and the other two (HM and CS) formulations has never been proved exactly. The authors give this proof, which allows us to complete a table (matrix) of 3 formulations (HM; CS; SL) by 3 methods (MCA; NCA; PCA) containing nine (3 by 3) loss functions, where PCA and NCA abbreviate principal component and nonmetric component analyses, respectively, which can also be described by the HM, CS, and SL formulations. With the 3 by 3 table we discuss the hierarchical relationship among PCA, NCA, and MCA. Further, we discuss the differences of the goodness-of-fit indices and rotational freedom for MCA between the three formulations.

C192: **On the use of generalized canonical correlation analysis in genetics**
Presenter: **Jan Graffelman**, Universitat Politecnica de Catalunya, Spain

Multivariate techniques are becoming increasingly relevant in genetics, due to the automated generation of large databases of gene expression information, metabolite composition and genetic markers. Regarding the latter, single nucleotide polymorphisms (SNPs) have been become popular markers in gene-disease association studies. Most of these markers are bi-allelic, and individuals can be characterized generically as AA, AB or BB. SNP data is essentially multivariate categorical data, and the genotype information can be coded in indicator matrices, where different coding schemes can be used to represent the data. The correlation between a pair of markers is referred to as linkage disequilibrium in genetics. If data is represented by indicator matrices, then linkage disequilibrium can be studied by a canonical correlation analysis of two indicator matrices. Often the genetic markers can be grouped according to their location on chromosomes, in genes, in regions, in introns or exons, giving rise to more than two groups. We explore the use of Carroll's generalized canonical correlation analysis for the analysis of genetic markers expressed in binary form.

CS11 Room R2: Ares FUNTCIONAL DATA ANALYSIS Chair: Domingo Morales

C074: **Functional PCA of measures for investigating the influence of bioturbation on sediment structure: PCA of grain-size curves**
Presenter: **Claude Mante**, CNRS, France
Co-authors: Georges Stora

After describing the main characteristics of grain-size curves, we recall previous results about Principal Components Analysis of absolutely continuous measures, in connection with grain-size curves analysis. This method simultaneously takes into account a chosen reference probability (r.p.) μ (associated with a Radon-Nikodym derivation operator), and the imposed sampling mesh T_p. The point is that it amounts to usual PCA in some metrics $M^{-}\left(T_p;\mu\right)$; consequently, analyzing a set of grain-size curves in reference to different r.p.s amounts to carry out PCA with different metrics. Three complementary r.p.s were chosen to analyze a set of 552 grain-size curves issued from an experiment designed for investigating the influence of a *Polychaetes, Nereis diversicolor* , on the sediment structure. The obtained results show that this worm is actually able to alter the sediment. Furthermore, it is shown that its influence depends on its density in the sedimentary column, but not on its position.

C346: **Robust classification for functional data via spatial depth-based methods**
Presenter: **Carlo Sguera**, Universidad Carlos III de Madrid, Spain
Co-authors: Pedro Galeano, Rosa Lillo

Functional data are becoming increasingly available and tractable because of the last technological advances. We consider supervised functional classification problems, and in particular, we focus on cases in which the samples may contain outlying curves. For these situations, some robust methods based on the use of functional depths are available, and the performances of these methods depend in part on the chosen functional depth. We widen this choice by defining two new functional depths that are based on a spatial approach: the functional spatial depth (FSD), that shows an interesting connection with the functional extension of the notion of spatial quantiles, and the kernelized functional spatial depth (KFSD), which is useful for studying functional samples that require an analysis at a local level, such as contaminated datasets. By means of a simulation study, we evaluate the performances of FSD and KFSD as depth functions for the depth-based methods. The results indicate that a spatial depth-based classification approach may result helpful when the datasets are contaminated, and that in general, it is stable and satisfactory if compared with a benchmark procedure such as the functional k-nearest neighbor classifier. Finally, we also illustrate our approach with a real dataset.

C327: **Clustering multivariate functional data**
Presenter: **Julien Jacques**, University Lille 1 and INRIA, France
Co-authors: Cristian Preda

Model-based clustering is considered for Gaussian multivariate functional data as an extension of the univariate functional setting. Principal components analysis is introduced and used to define an approximation of the notion of density for multivariate functional data. An EM like algorithm is proposed to estimate the parameters of the reduced model. Application on climatology data illustrates the method.

CS21 Room R5: Ath.1+2 MONTE CARLO METHODS Chair: Demetris Lamnisos

C303: **Importance sampling using Renyi divergence**
Presenter: **Emanuel Florentin Olariu**, Alexandru Ioan Cuza University, Romania

An alternative approach to the problem of estimating probabilities of rare events using the class of Rényi divergences of order $\alpha > 1$ is presented. This approach can be also used for solving global optimization problems. The general procedure we describe does not involve any specific family of distributions, the only restriction is that the search space consists of product-form probability density functions. We discuss an algorithm to estimate the probability of rare events and a version for continuous optimization. The results of numerical experimentation with these algorithms support their performances.

C171: **Maximum likelihood estimation via Monte Carlo methods**
Presenter: **Wojciech Rejchel**, Nicolaus Copernicus University, Poland
Co-authors: Wojciech Niemiro, Jan Palczewski

Likelihood maximization is a well-known method used in statistical inference in parametric models. However, for many stochastic processes exact calculation of maximum likelihood estimates is difficult. Such problems occur if considered densities are known only up to a normalizing constant (for instance in spatial statistics). Monte Carlo methods can be used to overcome such difficulties by estimating the intractable constant by an average based on a new sample (importance sampling). We prove that estimators obtained using this method are asymptotically normal if sample sizes of the initial sample and Monte Carlo one tend to infinity. Moreover, we investigate practical efficiency of estimators by simulation studies in the autologistic model. Besides, we develop the adaptive method of choosing the instrumental density in important sampling that is essential to reduce the variance of estimators.

C073: **Adaptive Monte Carlo for Bayesian variable selection in regression models**
Presenter: **Demetris Lamnisos**, Cyprus University of Technology, Cyprus
Co-authors: Jim E. Griffin, Mark F. J. Steel

The availability of datasets with large number of variables has lead to interest in variable selection methods for regression models with many regressors. Bayesian methods can deal with uncertainty in variable selection, however they require the development of Markov Chain Monte Carlo (MCMC) algorithms that explore efficiently the complicated and vast model space. A Metropolis-Hastings sampler is implemented with a model proposal that generates a candidate model by randomly changing components of the current model. This is similar to a Random Walk Metropolis (RWM) sampler, which proposes a new state as perturbed version of the current state. The model proposal can be generalized to include a tuning parameter that determines the degree of "localness" for the sampler and behaves similarly to the scale parameter of the RWM. The application of this model proposal to datasets with large number of variables suggests that the optimal sampler occurs for a parameter which leads to an average acceptance rate close to 0.3. Therefore, an adaptive sampler is proposed that automatically specifies the optimal tuning parameter and allows efficient computation in these problems. The method is applied to examples from normal linear and logistic regression.

CS25 Room R7: Ath. 4 PARAMETRIC MODELS Chair: Maria Brigida Ferraro

C215: **On generalization of Batschelet distributions**
Presenter: **Toshihiro Abe**, Tokyo University of Science, Japan

We consider the general application to symmetric circular densities of two forms of change of argument. The first one produces extended families of distributions containing symmetric densities which are more flat-topped, as well as others which are more sharply peaked than the originals. The second one produces families which are skew. General results for the modality and shape characteristics of the densities which ensue are presented. Maximum likelihood estimation of the parameters of two extensions of the Jones–Pewsey family is discussed.

C269: **Estimation methods in non-homogeneous Poisson process models for software reliability**
Presenter: **Alicja Jokiel-Rokita**, Wroclaw University of Technology, Poland
Co-authors: Ryszard Magiera

A subclass of non-homogeneous Poisson processes (NHPP) for which the expected number of events over $(0,\infty)$ is finite (the so called NHPP-I processes) is considered which besides of its theoretically interesting structure it can be used to model software reliability. It is demonstrated that in certain cases the maximum likelihood (ML) estimators do not exist despite the fact that a large number of faults is observed. As alternative to the ML method, some other methods of estimating parameters in the process models are proposed. All the methods proposed are based on an analogue to the minimum distance method. The methods described provide the estimates of unknown parameters with satisfactory accuracy measured by mean squared error, and can be also applied in some process models in which the ML estimators do not exist. Some numerical results are presented illustrating the accuracy of the proposed estimators with comparison to the ML estimators for a special case of the Erlangian software reliability model.

C271: **Prediction in trend-renewal processes for repairable systems**
Presenter: **Ryszard Magiera**, Wroclaw University of Technology, Poland
Co-authors: Juergen Franz, Alicja Jokiel-Rokita

Some problems of point and interval prediction in trend-renewal processes (TRP's) are considered. TRP's, whose realizations depend on a renewal distribution as well as on a trend function, comprise the non-homogeneous Poisson and renewal processes and serve as useful reliability models for repairable systems. For these processes, some possible ideas and methods for constructing the predicted next failure time and the prediction interval for the next breakdown time are presented. A method of constructing the predictors is also presented in the case when the renewal distribution of a TRP is unknown (and consequently, the likelihood function of this process is unknown). The forecasts in such TRP model consist in 1) finding the estimates of unknown trend parameters by minimizing the sample variance of the transformed working times and 2) exploiting the inverse transformation of the cumulative intensity function of the TRP. Some models of a TRP are considered for which the statistical inference concerning point and interval prediction is analytically intractable. Using the prediction methods proposed, the predicted times and intervals for a TRP with completely unknown renewal distribution are compared numerically with the corresponding results for the TRP with a Weibull renewal distribution and power law type trend function. The prediction methods are also applied to some real data.

CS28 Room R4: Aph.+Pos. METHODS FOR APPLIED STATISTICS II Chair: Alessandra Petrucci

C235: **Applying small area estimation to the structural business survey of Statistics Netherlands**
Presenter: **Sabine Krieg**, Statistics Netherlands, Netherlands
Co-authors: Marc Smeets

Traditionally, statistical offices like Statistics Netherlands prefer design-based techniques, for example the generalized regression (GREG) estimator

to produce estimates from survey samples. The advantage of these techniques is that the estimates are always approximately design unbiased. GREG estimators, however, have relatively large design variances in the case of small sample sizes. Then a small area estimator can be considered as an alternative. By applying such an estimator, information from other subpopulations can be borrowed to improve the accuracy of the estimates. The approach is applied to the Structural Business Survey, an annual survey about the Dutch business. In this survey, consistent estimates of different target variables are needed. Register information, which can be useful as auxiliary information, is only available for a part of the population. The first applied method is the EBLUP, which is probably the most known method. The M-quantile estimator uses M-quantiles to compute the subpopulation estimates. As a third estimator, a modification of the EBLUP is considered where the target variable is transformed to reduce skewness. The accuracy of these techniques is compared in a simulation study.

C369: **Classification of EEG signals by an evolutionary algorithm**
Presenter: **Laurent Vezard**, INRIA Bordeaux, France
Co-authors: Pierrick Legrand, Marie Chavent, Frederique Faita-Ainseba, Julien Clauzel

The goal is to predict the alertness of an individual by analyzing the brain activity through electroencephalographic data (EEG) captured with 58 electrodes. Alertness is characterized as a binary variable that can be in a normal or relaxed state. We collected data from 44 subjects before and after a relaxation practice, giving a total of 88 records. After a pre-processing step and data validation, we analyzed each record and discriminate the alertness states using our proposed slope criterion. Afterwards, several common methods for supervised classification (k nearest neighbors, decision trees -CART-, random forests, PLS and discriminant sparse PLS) were applied as predictors for the state of alertness of each subject. The proposed slope criterion was further refined using a genetic algorithm to select the most important EEG electrodes in terms of classification accuracy. Results shown that the proposed strategy derives accurate predictive models of alertness.

C253: **Clustering human body shapes using k-means algorithm**
Presenter: **Guillermo Vinue**, Universitat de Valencia, Spain
Co-authors: Amelia Simo, M. Victoria Ibanez, Sandra Alemany, Guillermo Ayala, Irene Epifanio, Esther Dura

Clustering of objects according to the shapes has a key importance in many scientific fields. Different approaches can be identified in shape analysis based on how the object is treated in mathematical terms. The shape is proposed to be analyzed by means of a configuration matrix of landmarks. An important application will be shown: to define prototypes in apparel design from a 3D large database of women's bodies. One of the major clustering objects approaches is based on the sum-of-squares criterion and on the well-known k-means algorithm which tries to find a partition minimizing the sum-of-squares error between the empirical mean of a cluster and the objects in the cluster, and it approximates this optimum k-partition by iterating. It has been proposed in different forms and under different names: Lloyd's algorithm, Hartigan-Wong algorithm, etc. We suggest to integrate Procrustes distance and mean into k-means algorithm. There has been several attempts in that sense. Each one of them adapts a different version of k-means algorithm. We also present an exhaustive study comparing the performance of these versions in the field of statistical shape analysis. Moreover, we analyze in detail the shape variability using PCA. Shapes package for R will be used.

CS33 Room R6: Ath. 3 NONPARAMETRIC STATISTICS II Chair: Agustin Mayo

C123: **Some aligned rank tests in measurement error models**
Presenter: **Radim Navratil**, Charles university in Prague, Czech Republic
Co-authors: A.K.Md.E. Saleh

In practice it may often happen that the real value of the characteristic of our interest is not observed, instead the value affected by some measurement error is obtained. In this situation, parametric methods may not be suitable due to the absence of the knowledge of exact distributions of the measurement errors except for the restrictive assumption of normality of distributions. In linear regression model with measurement errors test of hypothesis about an intercept and a slope parameter, as well as test of hypothesis that several regression lines are parallel (testing of parallelism) are considered. In this case, aligned rank tests may lead us to easy, simple and accessible solution to this problem. It is shown, when and under what conditions, these aligned rank tests can be used effectively, and indicated the effect of measurement errors on the power of the tests.

C260: **Performance of interval estimation methods for relative effects**
Presenter: **Takashi Nagakubo**, Asubio Pharmaceuticals Inc, United States
Co-authors: Masashi Goto

When a response variable is measured repeatedly but does not have normal distribution, relative effects can be analyzed nonparametrically. The advantage of using the relative effect is the ability to present the treatment effect visually by means of confidence interval. The translation method, which is based on the asymptotic distribution, is one approach for the interval estimation of relative effects. We consider the accelerated bias-corrected percentile (BCa) method for interval estimation. To compare the two methods we conducted a simulation study of the coverage probability and the width of the confidence interval. This indicates that, when the relative effects are close to the boundary and the sample size is small, the coverage probability of the BCa method is close to the nominal level. In terms of the width of the confidence interval there is hardly any difference between the two methods.

C214: **The max-type multivariate two-sample rank statistic based on the Baumgartner statistic**
Presenter: **Hidetoshi Murakami**, National Defense Academy, Japan

The multivariate two-sample testing problems are examined based on the Jurečková-Kalina's ranks of distances. The multivariate two-sample rank test based on the modified Baumgartner statistic for the two-sided alternative is suggested. Simulations are used to investigate the power of the proposed statistic for various population distributions.

Authors Index

Abankina, I., 23
Abe, T., 46
Adachi, K., 45
Addo, P., 6, 16
Ahlgren, N., 3
Ailliot, P., 9
Alba-Fernandez, V., 41
Albers, C., 45
Alemany, S., 47
Aleskerov, F., 23
Alfons, A., 15
Alisson, S., 35
Almohaimeed, B., 27
Ambroise, C., 26
Amendola, A., 40
Amorim, M., 26
Amouh, T., 40
Antoch, J., 35
Antoniadis, A., 5
Aoki, M., 18
Arai, S., 10, 24
Arhipova, I., 23
Arlt, J., 24
Arltova, M., 24
Atanasov, D., 42
Audigier, V., 21
Audrino, F., 7, 34
Augugliaro, L., 11
Ayala, G., 47
Azen, S., 1

Bach, F., 6
Bacigal, T., 41
Bar-Hen, A., 7
Barna, C., 37
Barrios, E., 39
Barsotti, F., 20
Basta, M., 3, 8
Baxevani, A., 9
Bazovkin, P., 14
Beaujean, F., 2
Belousova, V., 23
Beretvas, T., 23
Berka, P., 30
Bernard, A., 5
Bertrand, P., 40
Billio, M., 6, 16
Blanco-Fernandez, A., 35
Blueschke, D., 34
Blueschke-Nikolaeva, V., 34
Bonch-Osmolovskaya, A., 23
Bonneau, M., 11
Brault, V., 41
Brito, P., 40
Broda, S., 7
Bry, X., 33
Burden, C., 28
Busu, M., 16
Bwanakare, S., 38

Cabrera, J., 41
Caeiro, F., 41
Caldwell, A., 2
Calo, D., 27
Camillo, F., 19
Capellini, A., 32
Cardoso, M., 26
Caroni, C., 38
Carota, C., 2
Carreau, J., 20
Catani, P., 3
Celeux, G., 41
Cerny, M., 35
Ceulemans, E., 13–15, 32
Chavent, M., 15, 47
Chen, C., 20, 32
Choirat, C., 44
Christou, V., 17, 25
Chu, C., 7
Ciampi, A., 4
Claeskens, G., 29, 32
Clauzel, J., 47
Collet, J., 7
Collin, J., 21
Colubi, A., 5
Coppi, R., 5
Corbacho, C., 42
Cosbuc, M., 41
Costantini, M., 3
Coudret, R., 34
Croux, C., 19, 25
Cunha, A., 29
Cuzol, A., 9

D'Amato, V., 39
D'Urso, P., 37
Da Mota, B., 23
Daisuke, I., 18
Davison, A., 9
Dayal, M., 22
de Carvalho, F., 28, 40
De Ketelaere, B., 32
de Melo, F., 40
De Roover, K., 13
Dedu, S., 38
Derquenne, C., 3
Derrode, S., 11
Dias, S., 40
Dimitrakopoulos, S., 16
Djauhari, M., 27
Donev , A., 27
dos Anjos, U., 35
Dreassi, E., 36
Duchesnay, E., 23
Dupuis, D., 16
Dura, E., 47
Durand, P., 22
Durrieu, G., 34
Dusseldorp, E., 21

Economou, P., 38
Eijkemans, M., 15
Eilers, P., 2
Epifanio, I., 47
Esposito Vinzi, V., 13, 32
Estudillo-Martinez, M., 41
Etxeberria, J., 36

Faita-Ainseba, F., 47
Ferraro, M., 36
Ferron, J., 23
Fiala, T., 30
Figini, S., 16
Filippone, M., 2
Filzmoser, P., 19, 20
Fischer, H., 35
Fischer, J., 17
Fischer, P., 43
Fleury, G., 5
Fokianos, K., 9, 17, 25
Fonseca, M., 30
Foret, S., 28
Franz, J., 46
Fried, R., 25
Friel, N., 4
Fritz, H., 19
Frouin, V., 23
Fujisawa, H., 27
Fujita, T., 15

Galeano, P., 45
Gandy, A., 34
Garcia-Barzana, M., 35
Garcia-Escudero, L., 19, 26
Garcia-Ligero, M., 29
Gatu, C., 2, 37, 41
Gauran, I., 39
Genuer, R., 15
Gerlach, R., 20
Gertheiss, J., 19
Gey, S., 19
Ghezelayagh, B., 38
Ghorbanzadeh, D., 22
Gil, M., 14
Gilli, M., 25
Giordani, P., 36
Giudici, P., 16
Gogonel, A., 7
Goicoa, T., 36
Gomes, C., 21
Gomes, I., 41
Gonzalez, M., 42
Gonzalez-Rodriguez, G., 5, 14
Gordaliza, A., 19, 26
Goto, M., 47
Govaert, G., 41, 42
Gower, J., 9, 45
Graffelman, J., 45
Gramfort, A., 6
Grandvalet, Y., 41
Greenwald, D., 2
Grendar, M., 24
Griffin, J., 24, 46
Groenen, P., 2, 9, 10
Grothe, O., 33
Guegan, D., 6, 16
Guerrier, S., 33
Guinot, C., 5
Gunter, U., 3
Gutierrez, C., 42
Gutierrez-Jaimez, R., 16
Gutierrez-Sanchez, R., 16

Hadjiantoni, S., 37
Haidar, I., 33
Halunga, A., 20
Han, A., 35
Hansen, N., 8, 11
Harvey , A., 20
Hastie, T., 1
Hayashi, K., 21
Heinzl, H., 6, 39
Henriques-Rodrigues, L., 41
Heritier, S., 28
Hermoso-Carazo, A., 29
Hero, A., 19
Herwindiati, D., 22, 27
Heylen, J., 15
Hilbert, A., 43
Hirotsu, C., 33
Hladik, M., 35
Hofer, V., 44
Hong, Y., 35
Hron, K., 20
Hubert, M., 14, 32
Hurley, N., 4
Husson, F., 21
Hyndman, R., 33

Ibanez, F., 21
Ibanez, M., 47
Ichino, M., 40
Iizuka, M., 21, 43
Irpino, A., 28
Ishioka, F., 6, 15, 21

Jacob, C., 43
Jacques, J., 45
Jaidane-Saidane, M., 37
Jaupi, L., 22, 27
Jenatton, R., 6
Ji, Y., 4
Jimenez-Gamero, M., 41
Jing, J., 28
Jokiel-Rokita, A., 46
Jongbloed, G., 5
Josse, J., 13, 21

Kaarik, E., 17
Kafadar, K., 37
Kallache, M., 20
Kamakura, T., 4, 6
Kasprikova, N., 10
Keribin, C., 41
Kharroubi, S., 2
Kiers, H., 13, 45
King, M., 12
Kitano, M., 21
Kitromilidou, S., 17
Kitsos, C., 26

Klufa, J., 10
Knaus, S., 7
Koilias, C., 26
Kollar, D., 2
Kolossiatis, M., 16, 24
Komornik, J., 28
Komornikova, M., 28
Konczak, G., 17
Konishi, S., 27
Kontoghiorghes, E., 2, 37, 41
Kosiorowski, D., 27
Koussis, N., 23
Kraft, V., 43
Krempl, G., 44
Krieg, S., 46
Kroeninger, K., 2
Kubota, T., 15
Kuensch, H., 6
Kunst, R., 3
Kurihara, K., 6, 21
Kuroda, M., 21
Kuwabara, R., 22
Kysely, J., 44

La Grange, A., 9
Labiod, L., 42
Laguitton, S., 23
Lahalle, E., 5
Lam, A., 44
Lam, K., 7
Lamirel, J., 14
Lamnisos, D., 46
Langet, H., 5
Langhamrova, J., 23, 29, 30
Laniado, H., 38
Le Roux, N., 9
Legrand, P., 47
Legrenzi, G., 24
Lencuchova, J., 38
Leombruni, R., 2
Leopardi, P., 28
Lesaffre, E., 2
Liberati, C., 19
Lillo, R., 38, 45
Lima Neto, E., 35
Lin, E., 20
Linares-Perez, J., 29
Liu, T., 19
Lo, S., 28
Lomet, A., 41
Loster, T., 29
Luati, A., 13, 20
Lubbe, S., 9
Lubiano, M., 14
Luo, X., 8

Ma, J., 28
Magiera, R., 46
Maharaj, A., 37
Mahmood, M., 3
Mante, C., 45
Marek, L., 10
Marhuenda, Y., 36
Martin, J., 17
Martin-Fernandez, J., 20
Martinez, R., 42
Martzoukos, S., 23
Mary-Huard, T., 19
Matran, C., 19, 26
Mayekawa, S., 10, 24, 36
Mayo-Iscar, A., 19, 26
McDaid, A., 4
McLachlan, G., 36
Meier, P., 34
Meijer, G., 15
Mejza, I., 29
Mejza, S., 29
Mesiar, R., 28, 41
Meulders, M., 28
Mhamdi, F., 37
Mielniczuk, J., 11
Milas, C., 24
Milheiro-Oliveira, P., 29
Militino, A., 36
Mimaya, J., 22
Mineo, A., 11
Miskolczi, M., 23, 30
Mitsuhiro, M., 3
Mittlboeck, M., 39
Moeyaert, M., 23
Molina, M., 42, 43
Monbet, V., 9
Montanari, A., 27
Morales, D., 36
Mori, Y., 18, 21
Mortier, F., 33
Mosler, K., 14
Mota, M., 42, 43
Moysiadis, T., 8
Mundra, P., 2
Murakami, H., 47
Murakami, T., 45
Murphy, T., 4
Mustonen, S., 43

Nadif, M., 42
Nafidi, A., 16
Nagakubo, T., 47
Nagy, S., 18
Naranjo, L., 17
Naveau, P., 20
Navratil, R., 47
Neubauer, J., 30
Ng, S., 44
Nicklas, S., 33
Nielsen, H., 10
Niemiro, W., 46
Nieto-Reyes, A., 41
Nittono, K., 24
Nocairi, H., 21
Noirhomme, M., 40
Novianti, P., 15

Obozinski, G., 6
Obulkasim, A., 15
Oelker, M., 19
Ogasawara, H., 17
Okada, K., 11
Okubo, T., 30
Okusa, K., 4, 6
Olariu, E., 46
Oliveira, A., 26, 31
Oliveira, T., 26, 31
Onghena, P., 13
Ota, Y., 3

Palarea-Albaladejo, J., 20
Palczewski, J., 46
Paolella, M., 34, 40
Papageorgiou, E., 26
Papritz, A., 6
Paragios, N., 5
Pardo, M., 36
Park, H., 27
Pashapour, S., 2
Pedeli, X., 9
Pelzel, F., 22
Penisson, S., 43
Perez, C., 17
Petrickova, A., 38
Petrucci, A., 36
Petruschenko, V., 23
Peyrard, N., 11
Phinikettos, I., 34
Picek, J., 34, 44
Pieczynski, W., 11
Podder, M., 32
Poetschger, U., 39
Poggi, J., 37
Polak, P., 34
Poline, J., 23
Pollock, S., 13
Praskova, Z., 38
Preda, C., 45
Preda, V., 38
Proeitti, T., 13

Raabe, N., 25
Rahbek, A., 10
Raillard, N., 9
Rajapakse, J., 2
Ramos, A., 42, 43
Ramos-Abalos, E., 16
Ramos-Guajardo, A., 5, 14
Rejchel, W., 46
Rezac, M., 7
Riddell, C., 5
Roca-Pardinas, J., 18, 30
Rocco, E., 36
Rocha, G., 37
Roes, K., 15
Roman-Roman, P., 18
Romo, J., 38
Ronchetti, E., 1
Rousseeuw, P., 14
Russolillo, G., 10, 32
Russolillo, M., 39
Ryden, T., 13

Sabbadin, R., 11
Sagitov, S., 42
Sakakihara, M., 21
Sakaori, F., 27
Sakurai, H., 44
Saleh, A., 47
Salibian-Barrera, M., 19, 32
Sanfelici, S., 20
Saporta, G., 5, 19, 21
Saracco, J., 15, 34
Sato-Ilic, M., 36
Savin, I., 34
Savva, C., 20
Sawae, R., 18
Schindler, M., 44
Schoonees, P., 10
Schouteden, M., 29
Schwarz, M., 5
Schwierz, C., 6
Selart , A., 17
Serban, F., 16
Seri, R., 44
Serra, M., 42
Sestelo, M., 18, 30
Sgouropoulou, C., 26
Sguera, C., 45
Shevchenko, P., 8
Shirahata, A., 22
Sikorska, K., 2
Silvia, P., 2
Simo, A., 47
Simons, K., 30
Sinova, B., 14
Sixta, J., 17
Skaloud, J., 33
Skintzi, V., 34
Skutova, J., 24
Smeets, M., 46
Smilde, A., 14
Spitalsky, V., 24
Stahel, W., 6
Stebler, Y., 33
Steel, M., 24, 46
Stefanescu, M., 16
Stoimenova, V., 42
Stora, G., 45
Storti, G., 40
Suito, H., 21
Sund, R., 43
Sweeting, T., 2

Tagalakis, V., 4
Taguri, M., 44
Taki, M., 22
Tanioka, K., 4
Tarumi, T., 43
Tatsunami, S., 22
Tebbutt, S., 32
Teisseyre, P., 11
Templ, M., 20
ten Berge, J., 45
Tenenhaus, A., 5, 19, 32
Terada, Y., 22
Thieler, A., 25
Thirion, B., 6, 23
Thomas, M., 21
Timmerman, M., 13, 14, 32
Tokuda, T., 29
Tomita, M., 15, 28
Torres-Ruiz, F., 18
Toulias, T., 26
Trinchera, L., 19, 32
Trottier, C., 33
Trousset, Y., 5
Tsolakidis, A., 26
Tuerlinckx, F., 29
Tutz, G., 19

Ueno, T., 22
Ugarte, L., 36

Authors Index

Ugille, M., 23

Vahi, M., 29
Vakili, K., 14, 32
Van Aelst, S., 14, 32
Van de Velden, M., 10
van de Wiel, M., 15
van den Berg, R., 5
Van den Noortgate, W., 23, 32
Van Deun, K., 5, 29
van Dijk, D., 40
Van Keilegom, I., 5
Van Mechelen, I., 5, 15, 21, 29
Varoquaux, G., 6, 23
Vehkalahti, K., 43
Verde, R., 28
Verduyn, P., 15
Verron, T., 33
Vervloet, M., 32
Vesely, V., 30
Vezard, L., 47
Victoria-Feser, M., 16, 33
Vieira, P., 29
Villanueva, N., 18, 30
Vincent, M., 11
Vinue, G., 47
Viroli, C., 26, 27
Visek, J., 27
Vogl, P., 7
von Rosen, T., 17
Vonta, I., 25
Vosseler, A., 8
Vrabec, M., 10, 30

Waldhoer, T., 6
Waldl, H., 7
Wang, S., 35
Watanabe, N., 18
Wei, D., 19
Welch, W., 32
Welsch, R., 2
Wilcox, R., 37
Wilderjans, T., 5, 14
Wilson, S., 28
Winker, P., 35
Wit, E., 11
Wolff, R., 33

Yadohisa, H., 3, 4, 21, 22
Yakuba, V., 23
Yamamoto, S., 33
Yamashita, N., 36
Yano, K., 4
Yohai, V., 32

Zamar, R., 32
Zhang, L., 43
Zhang, X., 12
Zinkovsky, K., 23