DSSV 2017

DATA SCIENCE, STATISTICS & VISUALISATION

12-14 JULY 2017
IST, Lisbon, Portugal

DATA SCIENCE, STATISTICS & VISUALISATION
a Satellite Conference of the 61st World Statistics Congress,
promoted by IASC

BOOK OF ABSTRACTS

iasc isi
International Association for
Statistical Computing

TÉCNICO
LISBOA

Book of Abstracts

Data Science, Statistics & Visualisation 2017

12 – 14 July 2017

IST, Lisbon

DSSV2017.iasc-isi.org

Sponsors

Instituto Superior Técnico – University of Lisbon
International Association for Statistical Computing
International Federation for Information Processing
International Statistical Institute
Portuguese Statistical Society
University of Aveiro

Andritz Pulp & Paper
Springer

M. Rosário Oliveira, Alexandre Francisco, Conceição Amado, Francisco Santos,
Manuela Souto de Miranda, Patrícia Figueiredo, Ricardo Andrade, Sofia Naique

Book of Abstracts
Data Science, Statistics & Visualisation 2017

Publisher: IST Press

ISBN 978-972-99289-3-2

9 789729 928932

PRESS

Preface

The first conference *Data Science, Statistics & Visualisation* (DSSV 2017) is held at Instituto Superior Técnico in Lisbon, Portugal, from the 12th to the 14th of July 2017. It is a satellite meeting of the 61st World Statistics Congress, and is organized under the auspices of the *International Association for Statistical Computing*.

DSSV 2017 brings together researchers and practitioners in the interplay between computer science, statistics, and visualisation. The conference creates a forum for discussing emerging ideas and recent progress in these diverse disciplines. Our aim is to promote greater cooperation and to build lasting bridges between those fields.

DSSV 2017 features keynote addresses by Trevor Hastie (Stanford University, USA) and Daniel Keim (University of Konstanz, Germany), as well as invited lectures by Mário Figueiredo (Instituto Superior Técnico, Portugal), André F.T. Martins (Unbabel, Portugal), Peter Rousseeuw (University of Leuven, Belgium), Stephanie Sapp (Google Inc., USA), and Myra Spiliopoulou (University of Magdeburg, Germany). The conference also includes four invited topic sessions in collaboration with the Bernoulli Society, the European Association for Data Science, the International Society for Business and Industrial Statistics, and the Portuguese Statistical Society.

This book contains the scientific programme for DSSV 2017 and the abstracts for all the presentations made at the conference. For ease of reference, the abstracts are arranged according to their order in the programme.

We, the local organizers, wish to express our gratitude to all the people and institutions who made DSSV 2017 possible. We are grateful for the generous support provided by the conference's sponsors, in particular to Instituto Superior Técnico at the University of Lisbon. Special thanks go to the scientific committee and to the keynote and invited speakers. Finally, we thank each and every participant for making DSSV 2017 such a vibrant and wonderful conference.

Lisbon, July 2017

M. Rosário Oliveira
Alexandre Francisco
Conceição Amado
Francisco Santos
Manuela Souto de Miranda
Patrícia Figueiredo
Ricardo Andrade
Sofia Naique

Scientific Programme Committee

Peter Filzmoser, Vienna University of Technology, Austria (Chair)
David Banks, Duke University, USA
Fionn Murtagh, University of Derby, UK
M. Rosário Oliveira, Instituto Superior Técnico, ULisboa, Portugal
Patrick Groenen, Erasmus University Rotterdam, The Netherlands
Patrick Mair, Harvard University, USA
Silvia Miksch, Vienna University of Technology, Austria

Local Organizing Committee

M. Rosário Oliveira, Instituto Superior Técnico, ULisboa, Portugal (Chair)
Alexandre Francisco, Instituto Superior Técnico, ULisboa, Portugal
Conceição Amado, Instituto Superior Técnico, ULisboa, Portugal
Francisco Santos, Instituto Superior Técnico, ULisboa, Portugal
Manuela Souto de Miranda, University of Aveiro, Portugal
Patrícia Figueiredo, Instituto Superior Técnico, ULisboa, Portugal
Ricardo Andrade, Instituto Superior Técnico, ULisboa, Portugal
Sofia Naique, Instituto Superior Técnico, ULisboa, Portugal

Contents

Overview of the Programme

Wednesday, 12 July 2017

8:15–9:00	Registration
9:00–9:30	Opening Ceremony
9:30–10:30	**Keynote:** Trevor Hastie, *Statistical Learning with Sparsity*
10:30–11:00	Coffee Break
11:00–12:20	Contributed Paper Sessions I
12:20–14:00	Lunch Break
14:00–15:30	Invited Topic Sessions I
15:30–15:45	Coffee Break
15:45–16:45	Contributed Paper Sessions II
16:45–17:00	Coffee Break
17:00–18:00	Contributed Paper Sessions III
18:30–20:00	Welcome Reception

Thursday, 13 July 2017

9:00–9:45	André Martins, *From Softmax to Sparsemax*
9:45–10:30	Myra Spiliopoulou, *Learning on Timestamped Medical Data*
10:30–11:00	Coffee Break
11:00–12:20	Contributed Paper Sessions IV
12:20–14:00	Lunch Break
14:00–15:30	Invited Topic Sessions II
15:30–15:45	Coffee Break
15:45–17:05	Contributed Paper Sessions V
17:05–17:20	Coffee Break
17:20–18:40	Poster Session
20:30–	Conference Dinner

Friday, 14 July 2017

9:00–9:45	Mário Figueiredo, *Selection and Clustering of Correlated Variables*
9:45–10:30	Peter Rousseeuw, *Detecting Anomalous Data Cells*
10:30–11:00	Coffee Break
11:00–11:45	Stephanie Sapp, *Performance of Marketing Attribution Models*
11:45–12:45	**Keynote:** Daniel Keim, *The Role of Visualization in Data Science*
12:45–13:00	Closing Ceremony

Conference Programme

Wednesday, 12 July 2017

8:15–9:00	Registration		
9:00–9:30	Opening Ceremony		
9:30–10:30	**Keynote Lecture: Trevor Hastie** *Statistical Learning with Sparsity*		
10:30–11:00	Coffee Break		
11:00–12:20	Contributed Paper Sessions I		
	Robust Statistics I Room: EA2	**Biomedics** Room: EA3	**Clustering** Room: EA4
12:20–14:00	Lunch Break		
14:00–15:30	Invited Topic Sessions I		
	Classification and Network Modelling Room: EA2	**Statistical Learning in Data Science** Room: EA4	
15:30–15:45	Coffee Break		
15:45–16:45	Contributed Paper Sessions II		
	Robust Statistics II Room: EA2	**Regression and Beyond** Room: EA3	**Tools for Data Analytics** Room: EA4
16:45–17:00	Coffee Break		
17:00–18:00	Contributed Paper Sessions III		
	Applications I Room: EA2	**Big Data Platforms** Room: EA4	
18:30–20:00	Welcome Reception		

Thursday, 13 July 2017

9:00 – 9:45	**Invited Lecture: André Martins** *From Softmax to Sparsemax:* *A Sparse Model of Attention and Multi-Label Classification*		
9:45 – 10:30	**Invited Lecture: Myra Spiliopoulou** *Learning on Timestamped Medical Data*		
10:30 – 11:00	Coffee Break		
11:00 – 12:20	Contributed Paper Sessions IV		
	Methods for Data Science Room: EA2	**Time and Space** Room: EA3	**Visualization I** Room: EA4
12:20 – 14:00	Lunch Break		
14:00 – 15:30	Invited Topic Sessions II		
	Visualization and Analysis of Modern Data Room: EA2	**ISBIS Session** Room: EA4	
15:30 – 15:45	Coffee Break		
15:45 – 17:05	Contributed Paper Sessions V		
	Applications II Room: EA2	**Inference** Room: EA3	**Visualization II** Room: EA4
17:05 – 17:20	Coffee Break		
17:20 – 18:40	Poster Session		
20:30 –	Conference Dinner		

Friday, 14 July 2017

9:00 – 9:45	**Invited Lecture: Mário Figueiredo** *Selection and Clustering of Correlated Variables*
9:45 – 10:30	**Invited Lecture: Peter Rousseeuw** *Detecting Anomalous Data Cells*
10:30 – 11:00	Coffee Break
11:00 – 11:45	**Invited Lecture: Stephanie Sapp** *Performance of Marketing Attribution Models*
11:45 – 12:45	**Keynote Lecture: Daniel Keim** *The Role of Visualization in Data Science*
12:45 – 13:00	Closing Ceremony

Sessions and Presentations

The programme for DSSV 2017 comprises individual lectures, parallel sessions, and a poster session. The individual lectures, comprising the keynote and invited addresses, all take place in the Abreu Faro auditorium which is located in the building "Complexo Interdisciplinar". The parallel sessions take place in rooms EA2, EA3, EA4 located in the North Tower ("Torre Norte"). The poster session is also in the North Tower.

The keynote addresses by Trevor Hastie and Daniel Keim respectively begin and end the conference. There are also invited lectures by Mário Figueiredo, André Martins, Peter Rousseeuw, Stephanie Sapp, and Myra Spiliopoulou. These lectures all occur in the Abreu Faro auditorium at "Complexo Interdisciplinar".

There are four invited topic sessions, each of which features lectures by three invited speakers. These four sessions result from collaborations with the European Association for Data Science, the Portuguese Statistical Society, the Bernoulli Society for Mathematical Statistics and Probability, and the International Society for Business and Industrial Statistics.

Finally, the conference programme includes 14 contributed paper sessions as well as a poster session. Together with the invited topic sessions, they all take place in the North Tower at Instituto Superior Técnico.

We now list the sessions and presentations making up the scientific programme for DSSV 2017, sorted by time. Each session shows the name of the chairperson, the room, and the time, followed by the presentations in that session. Each presentation is identified with the title and with the authors of the corresponding abstract; the speaker is highlighted in bold.

Wednesday, 12 July 2017

Keynote Lecture Session **9:30–10:30**

Chair: Patrick Groenen, Room: Abreu Faro Wednesday 12th

Abstract on page 25

T. Hastie
Statistical learning with sparsity

Coffee Break: 10:30–11:00

Robust Statistics I **11:00–12:20**

Chair: Peter Filzmoser, Room: EA2 Wednesday 12th

Contributed paper session, Abstracts on pages 25–28

P.J. Rousseeuw, **J. Raymaekers**, M. Hubert
A measure of directional outlyingness with applications to image data and video
11:00–11:20

Y. Wang, S. Van Aelst
Sparse principal component analysis based on least trimmed squares
11:20–11:40

H. Cevallos Valdiviezo, S. Van Aelst, M. Salibian-Barrera
Least trimmed squares estimators for functional principal component analysis
11:40–12:00

M. Debruyne, **S. Höppner**, S. Serneels, T. Verdonck
On sparse directions of maximal outlyingness
12:00–12:20

Biomedics **11:00–12:20**

Chair: Wing Kam Fung, Room: EA3 Wednesday 12th

Contributed paper session, Abstracts on pages 28–31

M.B. Lopes, A. Veríssimo, E. Carrasquinha, S. Vinga
Consensus outlier detection in triple-negative breast cancer gene expression data
11:00–11:20

C. Arenas, **I. Irigoien**
Detection of differentially expressed genes by means of outlier detection
11:20 – 11:40

R. van den Berg, **L. De Mot**, G. Leroux-Roels, V. Bechtold, F. Clément, M. Coccia, E. Jongert, T.G. Evans, P. Gillard, R. van der Most
Transcriptional profiling of response to the candidate tuberculosis vaccine M72/AS01E for trial sampling time-point selection
11:40 – 12:00

A.H. Tavares, V. Afreixo, P. Brito
Clustering DNA words through distance distributions
12:00 – 12:20

Clustering

11:00 – 12:20

Chair: Ron Wehrens, Room: EA4

Wednesday 12th

Contributed paper session, Abstracts on pages 32 – 34

N. Bozkus
Lifting and clustering
11:00 – 11:20

S. Lubbe
Visualisations associated with bootstrapping cluster analysis
11:20 – 11:40

N.C. Chung
Evaluation of membership reliability in K-means clusters
11:40 – 12:00

R. Wehrens, J. Kruisselbrink
Data fusion with self-organising maps
12:00 – 12:20

Lunch Break: 12:20 – 14:00

Classification and Network Modelling **14:00–15:30**

Chair: Berthold Lausen, Room: EA2 Wednesday 12th

Invited topic session in collaboration with the European Association for Data Science
Abstracts on pages 35–36

B. Lausen
Ensemble classification
14:00–14:30

I. Gollini
Latent variable modelling of interdependent ego-networks
14:30–15:00

A. Caimo
Improving the efficiency of Bayesian computation for network models
15:00–15:30

Statistical Learning in Data Science **14:00–15:30**

Chair: Paula Brito, Room: EA4 Wednesday 12th

Invited topic session in collaboration with the Portuguese Statistical Society
Abstracts on pages 36–38

M.G.M.S. Cardoso
On clustering validation: the internal perspective
14:00–14:30

P.P. Rodrigues
Controversies in health data science
14:30–15:00

L. Torgo
Data pre-processing methods for forecasting with spatio-temporal data
15:00–15:30

Coffee Break: 15:30–15:45

Robust Statistics II **15:45–16:45**

Chair: Klaus Nordhausen, Room: EA2 Wednesday 12th

Contributed paper session, Abstracts on pages 38–40

P. Segaert, S. Van Aelst, T. Verdonck
Robust joint modelling of mean and dispersion for the GLM framework
15:45–16:05

A. Kharin, **E. Vecherko**
Performance and robustness in statistical testing of hypotheses for data with Markov dependencies
16:05 – 16:25

R. Crevits, C. Croux
Forecasting with robust exponential smoothing with trend and seasonality
16:25 – 16:45

Regression and Beyond **15:45 – 16:45**

Chair: Mário Figueiredo, Room: EA3 Wednesday 12th
Contributed paper session, Abstracts on pages 41 – 42

P. Macedo, J.P. Cruz
Entropy in high-dimensional variable selection
15:45 – 16:05

J. Wagner, R. Münnich
Regularized B-spline regression for high-dimensional scattered data
16:05 – 16:25

M.C. Costa, P. Macedo
Contributions to the analysis of inhomogeneous large-scale data using maximum entropy
16:25 – 16:45

Tools for Data Analytics **15:45 – 16:45**

Chair: Paulo C. Rodrigues, Room: EA4 Wednesday 12th
Contributed paper session, Abstracts on pages 43 – 45

T. Savage, C. Hansen, A. Seyb
Providing analysts the tools they need within a modern national statistics office
15:45 – 16:05

M. Rupp, R. Münnich
A flexible optimization tool for multivariate optimal allocation problems under high-dimensional data
16:05 – 16:25

V. Ardelean
Data analysis with contaminated data
16:25 – 16:45

Coffee Break: 16:45 – 17:00

Applications I

17:00 – 18:00

Chair: Pieter Segaert, Room: EA2

Wednesday 12th

Contributed paper session, Abstracts on pages 45 – 47

J.L. de Miranda, **M. Casquilho**
Computing over the Internet applied to data visualization: an illustrative example in geometry
17:00 – 17:20

J.M. Sanchez-Gomez, M.A. Vega-Rodriguez, C.J. Perez Sanchez, F. Calle-Alonso
Using word clouds as an e-learning analytics tool based on data visualization and statistics
17:20 – 17:40

D. Alptekin
Economic growth and tourism in Turkey: Hsiao's Granger causality analysis
17:40 – 18:00

Big Data Platforms

17:00 – 18:00

Chair: Klaus Nordhausen, Room: EA4

Wednesday 12th

Contributed paper session, Abstracts on pages 48 – 49

R.M. Silva, L. Sampaio, P. Calado, M.J. Silva, A. Delgado
Matching administrative data for census purposes
17:00 – 17:20

E.J. Harner, M. Lilback, W. Foreman
Rspark: running R and Spark in Docker containers
17:20 – 17:40

L. Borke, S. Bykovskaya
BitQuery – a GitHub API driven and D3 based search engine for open source repositories
17:40 – 18:00

Thursday, 13 July 2017

Invited Lecture Session

9:00 – 10:30

Chair: Berthold Lausen, Room: Abreu Faro

Thursday 13th

Abstracts on page 51

A.F.T. Martins
From softmax to sparsemax: a sparse model of attention and multi-label classification
9:00 – 9:45

M. Spiliopoulou
Learning on timestamped medical data
9:45 – 10:30

Coffee Break: 10:30 – 11:00

Methods for Data Science

11:00 – 12:00

Chair: Patrick Groenen, Room: EA2

Thursday 13th

Contributed paper session, Abstracts on pages 52 – 53

S. Zhang, H.-S. Chen
More powerful test procedures for multiple hypothesis testing
11:00 – 11:20

E. Stoimenova
Incomplete ranking
11:20 – 11:40

E. Macedo, **A. Freitas**, M. Vichi
Clustering and disjoint principal component analysis: an empirical comparison of two approaches
11:40 – 12:00

Time and Space

11:00–12:20

Chair: Myra Spiliopoulou, Room: EA3

Thursday 13th

Contributed paper session, Abstracts on pages 54–57

P. Otto
Estimation, simulation, and visualization of spatial and spatiotemporal autoregressive conditional heteroscedasticity
11:00–11:20

M.H. Gonçalves, M.S. Cabral
Performance of statistical approaches to model binary responses in longitudinal studies
11:20–11:40

B. Alptekin, C.H. Aladag
Air pollution forecasting with time series neural network models
11:40–12:00

C.S. Santos, I. Pereira, M.G. Scotto
Periodic multivariate INAR processes
12:00–12:20

Visualization I

11:00–12:20

Chair: Sugnet Lubbe, Room: EA4

Thursday 13th

Contributed paper session, Abstracts on pages 57–59

J. Rougier, **A. Zammit-Mangion**
Visualization for large-scale Gaussian updates
11:00–11:20

C. Bors, M. Bögl, T. Gschwandtner, S. Miksch
Visual support for rastering of unequally spaced time series
11:20–11:40

J.E. Lee, S. Ahn, D.-H. Jang
Visualization of three-dimensional data with virtual reality
11:40–12:00

M. Gallo, V. Todorov, M.A. Di Palma
R visual tools for three-way data analysis
12:00–12:20

Lunch Break: 12:20 – 14:00

Visualization and Analysis of Modern Data **14:00 – 15:30**

Chair: Po-Ling Loh, Room: EA2 Thursday 13th

Invited topic session in collaboration with the Bernoulli Society
Abstracts on pages 60 – 61

J.P. Long
Mapping the Milky Way halo: modeling and classification of sparsely sampled vector-valued functions
14:00 – 14:30

Y. Benjamini
Summarizing linearized prediction models along feature groups
14:30 – 15:00

T.H. McCormick
Using aggregated relational data to feasibly identify network structure without network data
15:00 – 15:30

ISBIS Session **14:00 – 15:30**

Chair: David Banks, Room: EA4 Thursday 13th

Invited topic session in collaboration with the International Society for Business and Industrial Statistics
Abstracts on pages 62 – 63

D. Banks
Statistical issues with agent-based models
14:00 – 14:30

T.A. Oliveira, A. Oliveira
Balanced incomplete block designs: some applications and visualization
14:30 – 15:00

P.C. Rodrigues, P. Tuy, R. Mahmoudvand
Randomized singular spectrum analysis for long time series
15:00 – 15:30

Coffee Break: 15:30 – 15:45

Applications II **15:45 – 16:45**

Chair: Paula Brito, Room: EA2 Thursday 13th

Contributed paper session, Abstracts on pages 63 – 65

F.G. Akgül, B. Şenoğlu
Alternative distributions to Weibull for modeling the wind speed data in wind energy analysis
15:45 – 16:05

Á.S.T. Sousa, M.G.C. Batista, O.L. Silva, M.C. Medeiros, H. Bacelar-Nicolau
Hierarchical cluster analysis in the context of performance evaluation: from classical to complex data
16:05 – 16:25

P. Brito, M.G.M.S. Cardoso, A.P. Duarte Silva
Building a map of Europe based on citizens' values: an interval data approach
16:25 – 16:45

Inference **15:45 – 16:45**

Chair: Peter Rousseeuw, Room: EA3 Thursday 13th

Contributed paper session, Abstracts on pages 66 – 68

T. Arslan, S. Acitas, **B. Şenoğlu**
Estimating the location and scale parameters of the Maxwell distribution
15:45 – 16:05

L. Qu
Copula density estimation by Lagrange interpolation at the Padua points
16:05 – 16:25

D. Kushary
Estimation of percentile of marginal distribution of order statistic with real life application
16:25 – 16:45

Visualization II **15:45 – 17:05**

Chair: Niel Le Roux, Room: EA4 Thursday 13th

Contributed paper session, Abstracts on pages 68 – 70

J. Nienkemper-Swanepoel, S. Gardner-Lubbe, N.J. Le Roux
The virtues and pitfalls of the visualisation of incomplete categorical data
15:45 – 16:05

A. de Falguerolles
Is Arthur Batut's geometric experiment a convincing argument in favor of Francis Galton's generic images?
16:05 – 16:25

V.O. Choulakian
Visualization of sparse two-way contingency tables
16:25 – 16:45

A. Alexandrino da Silva
Cyclical history theory in data visualization: using a four-quadrant display to see history repeating itself
16:45 – 17:05

Coffee Break: 17:05 – 17:20

Poster Session

17:20 – 18:40

Location: North Tower, first floor

Thursday 13th

Abstracts on pages 71 – 81

D. Montaña, Y. Campos-Roca, **C.J. Perez**
Early detection of Parkinson's disease by considering acoustic features of plosive consonants

N.S. Ribeiro, J.O. Folgado, H.C. Rodrigues
Surrogate-based visualization of the influence of geometric design on the performance of a coronary stent

V.M. Lourenço, P.C. Rodrigues, A.M. Pires, H.-P. Piepho
A robust DF-REML framework for variance components estimation in genetic studies

N. Krautenbacher, F.J. Theis, C. Fuchs
A correction approach for random forest under sample selection bias

R. Gorter, E. Geuze
Measuring trends in depression symptoms in Dutch veterans from before, until five years after deployment

J. Pinto, S. Nunes, M. Bianciardi, L.M. Silveira, L.L. Wald, **P. Figueiredo**
Cluster-based lag optimization of physiological noise models in high-field resting-state fMRI

M.C. Botelho, E. Vilar, E. Cardoso, **A. Alexandrino da Silva**, P.D. Almeida, L. Rodrigues, A.P. Martinho, S. Rodrigues
Information visualisation quadrant display: a synergistic approach to a postgraduate program

J.R. da Cruz, M.H. Herzog, P. Figueiredo
An automatic pre-processing pipeline for EEG analysis based on robust statistics

D.M. Garvis
Adaptive learning and learning science in a first course in university statistics

J. Korzeniewski
A novel technique of symbolic time series representation aimed at time series clustering

A.C. Finamore, **A. Moura Santos**, A. Pacheco
Introducing formative assessment in probability and statistics course – analysis of the first data

---

Friday, 14 July 2017

Invited Lecture Session

9:00–10:30

Chair: Christophe Croux, Room: Abreu Faro

Friday 14th

Abstracts on page 83

M.A.T. Figueiredo
Selection and clustering of correlated variables
9:00–9:45

P.J. Rousseeuw, W. Van den Bossche
Detecting anomalous data cells
9:45–10:30

Coffee Break: 10:30–11:00

Invited Lecture Session

11:00–11:45

Chair: Peter Filzmoser, Room: Abreu Faro

Friday 14th

Abstract on page 84

S. Sapp
Performance of marketing attribution models

Keynote Lecture Session

11:45–12:45

Chair: Peter Filzmoser, Room: Abreu Faro

Friday 14th

Abstract on page 85

D.A. Keim
The role of visualization in data science

Abstracts

We now come to the main content of this book. Following are the abstracts for all the presentations given at the conference DSSV 2017.

This section is divided into three parts, one for each day of the conference. For ease of reference, the abstracts are ordered by the time of their presentation. Moreover, each abstract displays the session in which it was presented as well as the room, the time, and the day of the presentation. The speaker is highlighted in bold among the list of authors.

Wednesday, 12 July 2017

Statistical learning with sparsity

T. Hastie

Stanford University

Keynote Lecture Session, Room: Abreu Faro *Wednesday 12^{th}, 9:30– 10:30*

In a statistical world faced with an explosion of data, regularization has become an important ingredient. In many problems, we have many more variables than observations, and the lasso penalty and its hybrids have become increasingly useful. This talk presents a general framework for fitting large scale regularization paths for a variety of problems. We describe the approach, and demonstrate it via examples using our R package GLMNET [1, 3]. We then outline a series of related problems using extensions of these ideas [2].

Joint work with Jerome Friedman, Rob Tibshirani and students, past and present.

Keywords: wide data, regularization, variable selection.

References

[1] J. Friedman, T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1), 1–22.

[2] T. Hastie, R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall, CRC Press.

[3] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. Tibshirani (2012). Strong rules for discarding predictors in lasso-type problems. *J. Royal Statistical Society B*, **74**.

A measure of directional outlyingness with applications to image data and video

P.J. Rousseeuw[a], **J. Raymaekers**[a], M. Hubert[a]

[a]*KU Leuven, Belgium*

Session: Robust Statistics I, Room: EA2 *Wednesday 12^{th}, 11:00– 11:20*

Images and video can be considered as functional data with a bivariate domain, where the data per grid point can be univariate (e.g. grayscale values) or multivariate (e.g. red, green, and blue intensities). This often yields large datasets in which outliers may occur that can distort the analysis.

At each grid point, we propose to compute a fast measure of outlyingness which accounts for skewness in the data. It can be used for univariate data and, by means of

projection pursuit, for multivariate data. The influence function of this outlyingness measure is computed as well as its implosion and explosion bias. We also construct a cutoff value for the outlyingness. Heat maps of the outlyingness indicate the regions in which an image deviates most from the majority of images.

The performance of the method is illustrated on real data. One example consists of MRI images augmented with their gradients. We also show an example of video surveillance data where we compare the exact method with faster approximations.

Keywords: outlyingness, functional data, robustness.

References

[1] P.J. Rousseeuw, J. Raymaekers, and M. Hubert (2016). A measure of directional outlyingness with applications to image data and video. arXiv:1608.05012

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

Sparse principal component analysis based on least trimmed squares

Y. Wang[a], S. Van Aelst[a]

[a]*Department of Mathematics, KU Leuven*

Session: Robust Statistics I, Room: EA2 *Wednesday 12th, 11:20– 11:40*

Principal Component Analysis (PCA) is an important tool for dimensional reduction. It searches for a set of linear combinations of the original variables, called Principal Components (PCs), that retain most of the covariance structure of the data set. There are two major disadvantages of the classical PCA method. First, it cannot produce sparse loading vectors. Therefore, the resulting PCs are often hard to interpret, especially in high-dimensional settings. Second, classical PCA is not resistant to outlying observations. To address these issues, two robust and sparse PCA methods, RSPCA [1] and ROSPCA [2], have been proposed. Despite their success in detecting the outliers and generating accurate sparse loadings, both methods are extensions of SCoTLASS [3] and so inherit its computational inefficiency. In this work, we proposed a fast sparse and robust PCA method. The new method is called Multivariate Least Trimmed Squares Sparse PCA (MLTS-SPCA), which uses the least trimmed squares loss function to evaluate the low rank approximation of the data matrix and incorporates a sparsity constraint as well. The PCs are searched sequentially using the deflation method [4]. The problem is solved by a modified Truncated Power method [5]. Simulation studies and real data analysis show that MLTS-SPCA can produce accurate estimation of the sparse loading vectors on both clean and contaminated datasets. Moreover, unlike RSPCA and ROSPCA, the LTS-SPCA solution for data sets with thousands of variables can be calculated within just a few minutes in R, including the selection of the sparsity control parameter.

Keywords: robust sparse PCA, least trimmed squares, truncated power method.

References

[1] C. Croux, P. Filzmoser, and H. Fritz (2013). Robust sparse principal component analysis. *Technometrics*, **55**, 202–214.

[2] M. Hubert, T. Reynkens, E. Schmitt, and T. Verdonck (2016). Sparse PCA for high-dimensional data with outliers. *Technometrics*, **58**, 424–434.

[3] I.T. Jolliffe, N.T. Trendafilov, and M. Uddin (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, **12**, 531–547.

[4] L. Mackey (2009). Deflation methods for sparse PCA. In *Advances in Neural Information Processing Systems 21*, 1017–1024.

[5] X.T. Yuan and T. Zhang (2013). Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, **14**, 899–925.

Least trimmed squares estimators for functional principal component analysis

H. Cevallos Valdiviezo[ab], S. Van Aelst[bc], M. Salibian-Barrera[d]

[a]*Escuela Superior Politécnica del Litoral (ESPOL), Facultad de Ciencias Naturales y Matemáticas (FCNM);* [b]*Ghent University, Department of Applied Mathematics, Computer Science and Statistics;* [c]*KU Leuven, Department of Mathematics, Section of Statistics;* [d]*The University of British Columbia, Department of Statistics*

Session: Robust Statistics I, Room: EA2 *Wednesday 12th, 11:40 – 12:00*

Classical functional principal component analysis can yield erroneous approximations in presence of outliers. To reduce the influence of atypical data we propose two methods based on trimming: a multivariate least trimmed squares (LTS) estimator and its coordinatewise variant. The multivariate LTS minimizes the multivariate scale corresponding to h-subsets of curves while the coordinatewise version uses univariate LTS scale estimators. Consider a general setup in which observations are realizations of a random element on a separable Hilbert space $\mathscr{H}$. For a fixed dimension q, we aim to robustly estimate the q-dimensional linear space in $\mathscr{H}$ that gives the best approximation to the functional data. Our estimators use smoothing to first represent irregularly spaced curves in a high-dimensional space and then calculate the LTS solution on these multivariate data. The solution is subsequently mapped back onto $\mathscr{H}$. Poorly fitted observations can therefore be flagged as outliers. Simulations and real data applications show that our estimators yield competitive results when compared to existing methods when a minority of observations is contaminated. When a majority of the curves is contaminated at some positions along its trajectory, coordinatewise methods like coordinatewise LTS are preferred over multivariate LTS and other multivariate methods which break down in this case.

Keywords: functional data analysis, outliers, principal components, robust methods.

On sparse directions of maximal outlyingness

M. Debruyne[a], **S. Höppner**[b], S. Serneels[c], T. Verdonck[b]

[a]*Dexia;* [b]*KU Leuven;* [c]*BASF Corp.*

Session: Robust Statistics I, Room: EA2 *Wednesday 12th, 12:00 – 12:20*

Nowadays, many robust statistical methods are available to detect outliers in multivariate data, both in high and low dimensions, e.g. robust covariance estimators such as MCD and robust PCA methods. Once an observation has been flagged as an outlier, it can be interesting to know which variables contribute most to its outlyingness. In practice, it is perfectly reasonable that an outlying observation deviates from the majority of the data points only in a few variables. Obviously, finding this subset of variables would be of high practical interest. Instead of downweighting the outlier in further analysis, we could then set the cells in the data set corresponding to these variables to missing, and hence the good part of the outlier can still be fully used. Therefore, consider the following problem: given a multivariate data set X and the fact that observation x_i is an outlier with large outlyingness, find the subset of variables contributing most to the outlyingness of x_i. This problem is akin to variable selection, with the objective of determining those variables contributing most to outlyingness instead of to predictive power.

A simple idea to find relevant variables is to check the univariate direction in which the observation is most outlying. The problem of estimating this direction of maximal outlyingness can be rewritten as the normed solution of a classical least squares regression problem. We propose to compute regularized directions of maximal outlyingness by sparse Partial Least Squares (PLS) regression, preferably by the fast SNIPLS algorithm.

Keywords: outlier, variable selection, partial least squares.

Consensus outlier detection in triple-negative breast cancer gene expression data

M.B. Lopes, A. Veríssimo, E. Carrasquinha, S. Vinga

IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Portugal

Session: Biomedics, Room: EA3 *Wednesday 12th, 11:00 – 11:20*

Triple-Negative Breast Cancer (TNBC) is the most heterogeneous group of breast cancers, with significantly shorter survival times. It is characterized by the lack of expression of estrogen (ER), progesterone (PR) and human epidermal growth factor 2 (HER2) receptors [1]. Endocrine and HER2-targeted therapies therefore fail, which fosters the need for new biomarkers and druggable targets for effective clinical management. Identifying outliers when predicting the cancer type based on genomic data is crucial in the context of precision medicine, as these observations might bias a correct understanding of the cancer and precipitate treatment failure.

The Rank Product (RP) test provides a consensus approach given different model-based observation rankings regarding a deviance measure, e.g. Cook's distance. The RP has been successfully used to detect differentially regulated genes in replicated microarray experiments [2] and for the meta-analysis of transcriptomic studies [3].

In this work we measure the outlierness of 1019 patients (TNBC and non-TNBC) by logistic regression based on RNA-Seq data (19688 covariates) from The Cancer Genome Atlas (TCGA). Since variable selection in high-dimensional data is a key step before outlier detection, three data reduction strategies were evaluated: i) Elastic net regularization; ii) Partial Least Squares - Discriminant Analysis (PLS-DA); and iii) sparse PLS-DA. The RP test was able to identify 21 observations that were systematically classified as influential (potential outliers), independently of the model chosen, from which 7 were previously marked as suspect cases regarding their labeling. These results represent a valuable insight towards precision medicine and for the development of clinical decision support systems in oncology.

Keywords: triple-negative breast cancer, outlier detection, dimensionality reduction.

References

[1] W.D. Foulkes, I.E. Smith, and J.S. Reis-Filho (2010). Triple-negative breast cancer. *The New England Journal of Medicine*, **363**, 1938–48.

[2] R. Breitling, P. Armengaud, and P. Herzykr (2004). A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, **573**, 83–92.

[3] J. Caldas and S. Vinga (2014). Global meta-analysis of transcriptomics studies. *PloS One*, **9**(2), e89318.

Detection of differentially expressed genes by means of outlier detection

C. Arenas[a], **I. Irigoien**[b]

[a] *University of Barcelona;* [b] *University of the Basque Country UPV/EHU*

Session: Biomedics, Room: EA3 *Wednesday 12th, 11:20 – 11:40*

Let X_g and Y_g be the random variables representing the expression level of gene g in tissues A and B, respectively ($g = 1, \dots, G$). A gene g is not important if $F^{-1}_{X_g}(p) = F^{-1}_{Y_g}(p)$ where F is the cumulative distribution function and $p \in [0, 1]$. Otherwise, the gene g is differently expressed or important. The proposed approach focuses on estimating quantiles and their differences among samples: $V_p = \hat{F}^{-1}_{X_g}(p) - \hat{F}^{-1}_{Y_g}(p)$, $p \in C_p$ where C_p is a set of probabilities. For instance, $C_p = \{0.25, 0.5, 0.75\}$ is adequate for small sample sizes. Broadly speaking, the matrix $\mathbf{V} = (V_{gp})$ with $g = 1, \dots, G$ and $p \in C_p$ must contain small values corresponding to the majority of non-important genes. Some of them should show a different behaviour, the most differently expressed genes being outliers in $\mathbf{V}$. To identify differentially expressed genes, the approach employs two main steps. In the first step, a robust index of outlyingness [1] is computed and, based on permuted samples, non-important genes are discarded at $(1-\alpha)100\%$ confidence-level. The remaining genes are suspected of being important. In the second step, the so-called suspicious genes and the corresponding permuted samples are mingled so that for each suspicious gene a 10-Nearest Neighbourhood is analysed. This way, a measure of "False Positiveness in Neighbourhood" (FPN) is computed. Finally, among the suspicious genes those with high value of outlyingness along with low value of FPN are reported as important. Simulation studies showed that the proposed procedure is able to discriminate differently expressed genes in several scenarios. We also analysed real data and we got competitive results compared to the results obtained with other well-known methods in this field, such as SAM [2] and eBayes [3].

Keywords: differentially expressed genes, outlier.

References

[1] C. Arenas, C. Toma, B. Cormand, and I. Irigoien (2017). Identifying extreme observations, outliers and noise in clinical and genetic data. *Curr. Bioinform.*, **12**(2), 101–117.

[2] V.G. Tusher, R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. *P. Natl. Acad. Sci. USA*, **98**(8), 5116–5121.

[3] G.K. Smyth (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**(1), 1544–6115.

Transcriptional profiling of response to the candidate tuberculosis vaccine M72/AS01E for trial sampling time-point selection

R. van den Berg[a], **L. De Mot**[a], G. Leroux-Roels[b], V. Bechtold[a], F. Clément[a], M. Coccia[a], E. Jongert[a], T.G. Evans[c], P. Gillard[a], R. van der Most[a]

[a]*GSK Vaccines;* [b]*Department of Clinical Chemistry, Microbiology and Immunology, Ghent University, Belgium;* [c]*Aeras and Vir Biotechnology, USA*

Session: Biomedics, Room: EA3 *Wednesday 12th, 11:40 – 12:00*

The M72/AS01E candidate tuberculosis vaccine is intended to address the unmet medical need in TB endemic countries and is currently being evaluated in a Phase II efficacy study in Africa (TB-018/NCT01755598). An ancillary study (C-041-972/NCT02097095), sponsored by Aeras, is planned for transcriptomics analysis.

The analysis presented here was used to guide time-point selection for the transcriptomics study. We chose to investigate, in transcriptomics data from another clinical study (TB019/NCT1669096), the behavior of a gene signature associated with protection in a controlled human malaria infection study (RTS,S/AS01 vaccine candidate). We indeed hypothesized that this signature would allow us to identify time-points displaying potentially biologically meaningful variation in a vaccine with the same adjuvant, even though protection mechanisms may differ between TB and malaria.

The statistical challenges were to (i) identify clusters of subjects that would present similar biological variation as protected and non-protected subjects from the malaria study (protection status is unknown in the TB study), (ii) validate the findings to control for the bias induced by searching for specific signatures. After signature-driven group assignment and statistical validation using a resampling approach and Monte Carlo simulations, we were able to show the presence of the signature at specific time-points after vaccination in TB019/NCT1669096. The optimal time-points, as defined by the capacity of the signature to reveal response variation, included day 7 post dose 2, which was selected for the ancillary transcriptomics study (C-041-972/NCT02097095).

References

[1] R.A. van den Berg, M. Coccia, W. Ripley Ballou, K.E. Kester, C.F. Ockenhouse, J. Vekemans, E. Jongert, A.M. Didierlaurent, and R. van der Most. Predicting RTS,S vaccine-mediated protection from transcriptomes in a malaria-challenge clinical trial. *Under review.*

Clustering DNA words through distance distributions

A.H. Tavares[a], V. Afreixo[a], P. Brito[b]

[a]*CIDMA and iBiMED, University of Aveiro;* [b]*FEP and LIAAD-INESC TEC, University of Porto*

Session: Biomedics, Room: EA3 *Wednesday 12th, 12:00 – 12:20*

Functional data appear in several domains of science, for example, in biomedical, meteorologic or engineering studies. A functional observation can exhibit an atypical behaviour during a short or a large part of the domain and this may be due to magnitude or to shape features. Over the last ten years many outlier detection methods have been proposed. In this work we use the functional data framework to investigate the existence of DNA words with outlying distance distribution, which may be related with biological motifs.

A DNA word is a sequence defined in the genome alphabet $\{ACGT\}$. Distances between successive occurrences of the same word allow defining the *inter-word distance* distribution, interpretable as a discrete function. Each word length k is associated with a functional dataset formed by 4^k distance distributions. As the word length increases, greater is the diversity of observed patterns in the functional dataset and larger is the number of distributions displaying strong peaks of frequency.

We propose a two-step procedure to detect words with an outlying pattern of distances: first, the functions are clustered according to their global trend; then, an outlier detection method is applied within each cluster. Each distribution trend is obtained by data smoothing, which avoids some distributions' peaks, and similarities between smoothed data are explored through hierarchical complete linkage clustering. The dissimilarity between functions is evaluated using the Euclidean distance or the Generalized Minimum distance [1], which considers the dependence between domain points. The resulting dendograms are then cut leading to a partition of the distance distributions. For the second step we use the Directional Outlyingness measure which assigns a robust measure of outlyingness to each domain point and is the building block of a graphical tool for visualization of the centrality of the curves [2].

We focus on the human genome and words of length $k \leq 7$. Results are compared with those obtained by applying only the second step of the procedure [3].

Keywords: distance distribution, DNA word, directional outlyingness.

References

[1] X. Zhao, E. Valen, B.J. Parker, and A. Sandelin (2011). Systematic clustering of transcription start site landscapes. *PloS one*, **6**(8), e23409.

[2] P.J. Rousseeuw, J. Raymaekers, and M. Hubert (2016). A measure of directional outlyingness with applications to image data and video. arXiv:1608.05012

[3] A.H. Tavares, V. Afreixo, P. Brito, and P. Filzmoser (2016). Directional outlyingness applied to distances between genomic words. In *RECPAD 2016*, 108–110.

Lifting and clustering

N. Bozkus

University of Leeds

Session: Clustering, Room: EA4 *Wednesday 12th, 11:00– 11:20*

A popular question in hierarchical clustering is how many clusters exist in a data set, or where to 'cut the tree'. Even though many methods have been proposed, this topic still attracts the interest of researchers. Previous indices capture the number of clusters quite well if clusters are well separated, but when the clusters overlap or have unusual shapes, their performance deteriorates. I propose a new method based on a multiscale technique called lifting which has recently been developed to extend the 'denoising' abilities of wavelets to data on irregular structures.

My method applies lifting to the structure of a dendrogram. I then assume that the distances between data points and cluster centroids are affected by noise. Denoising the mean distances of data points from each cluster centroid helps decide where to cut the tree. This method will be illustrated with both real and simulated examples.

Keywords: wavelets, lifting, cluster validity index.

Visualisations associated with bootstrapping cluster analysis

S. Lubbe

Stellenbosch University, South Africa

Session: Clustering, Room: EA4 *Wednesday 12^{th}, 11:20 – 11:40*

In the application presented here cluster analysis is performed with hierarchical clustering and Wards method. The clustering methodology is viewed as a given and the stability of the solution is evaluated by applying bootstrap methods. In the literature the bootstrap is used in a cluster analysis context to assist in deciding on the number of clusters, say k. Here, the focus is not on a single quantity k, but on replicating the complete clustering solution a large number of times. In order to summarise the bootstrap results, multivariate visualisation will play a pivotal role. Multidimensional scaling and Generalised Procrustes Analysis are used to obtain a representation of all the bootstrap replicates. In a further step, a dissimilarity matrix is constructed based on the bootstrap replicates and multidimensional scaling applied again to obtain a summary of the complete bootstrap process.

Keywords: multidimensional scaling, cluster analysis, bootstrap.

Evaluation of membership reliability in K-means clusters

N.C. Chung

Institute of Informatics, Faculty of Mathematics, Informatics and Mechanics, University of Warsaw

Session: Clustering, Room: EA4 *Wednesday 12^{th}, 11:40 – 12:00*

Unsupervised learning is crucial in computational statistics and data science, where the amount and the complexity of unlabeled data are rapidly increasing. There exist a number of algorithms to classify m unlabeled variables into $k = 1, \dots, K$ clusters, that are often designed for specific data types and sampling mechanisms. Particularly with extremely large numbers of variables (m) as encountered in genomics and biomedicine, there may exist variables that are mainly noise or technical artifacts. We investigated the next analytic step: Given m_k variables assigned to the k^{th} cluster, can we assign probabilistic measures evaluating the reliability of their *cluster membership* assignments? Topically, our research is distinct from a choice of k, cluster stability, and related questions (e.g. [1] and [2]) that assess the clusters themselves.

We developed a resampling-based strategy to evaluate robustness of membership in K-means clusters. Mirroring the *jackstraw* method introduced for principal component analysis [3], the proposed strategy learns over-fitting characteristics of the clustering algorithm and the input observed data. When resampling a small portion $s \ll m$ of observed variables and clustering this partially resampled data, some of s synthetic null variables may get artificially assigned to the k^{th} cluster. This relationship is used to evaluate whether the original m_k variables are reliably members of the k^{th} cluster. We demonstrate favorable operating characteristics in high-throughput genomics while noting limitations inherent in the nature of unsupervised classification and potential misspecification of parameters. The proposed strategy reduces noise in high-dimensional data, improves clustering, and aids the visualization of networks.

Keywords: clustering, resampling, membership.

References

[1] R. Tibshirani, G. Walther, and T. Hastie (2001). Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society, B*, **63**(2), 411–423.

[2] A. Rakhlin and A. Caponnetto (2007). Stability of *k*-means clustering. *Advances in Neural Information Processing Systems*, **19**, 1121–1128.

[3] N.C. Chung and J.D. Storey (2015). Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*, **31**(4), 545–554.

Data fusion with self-organising maps

R. Wehrens[a], J. Kruisselbrink[a]

[a]*Biometris, Wageningen University & Research, Netherlands*

Session: Clustering, Room: EA4 *Wednesday 12th, 12:00– 12:20*

Self-organising maps (SOMs) enjoy great popularity as a clustering method with very good visualization possibilities. Because they are based on distances, high-dimensional data are treated with ease, and because of their iterative training, it is easy to use them to map also large data sets with millions of records.

In this era of Big Data the trend is to combine data from different sources. Especially in the so-called Omics sciences, very often multiple detectors are used to characterize a sample. In our **kohonen** package [1] for the R language, this is enabled by allowing multiple information layers for all records, which are combined into one distance as a weighted sum of individual layer distances. Recently, we have extended the package [2], allowing different distance measures for individual layers, including user-defined distances. This greatly improves the utility of the package. In addition, the code has received a complete overhaul to improve efficiency, especially for large data sets. The **kohonen** package is available from the CRAN repository.

Here, we will present the basic ideas implemented in the package, and give examples of potential useful applications.

Keywords: self-organising maps, clustering, visualization.

References

[1] R. Wehrens and L. Buydens (2007). Self- and super-organizing maps in R: the kohonen package. *Journal of Statistical Software*, **21**(5), 1–19.

[2] R. Wehrens and J. Kruisselbrink (2017). Flexible self-organising maps in kohonen v3.0. *Submitted for publication.*

Ensemble classification

B. Lausen

University of Essex, Colchester, UK

Session: Classification and Network Modelling, Room: EA2 *Wednesday 12th, 14:00 – 14:30*

We review methods to use ensembles of selected classifiers to achieve classification rules with increased accuracy [1, 3]. Feature selection methods are often used as preprocessing method. We discuss a proposal to improve feature selection of microarray data based on a proportional overlapping score [2]. The methods are compared with other recent proposals using benchmarking.

References

[1] A. Gul, A. Perperoglou, Z. Khan, O. Mahmoud, M. Miftahuddin, W. Adler, and B. Lausen (2016). Ensemble of a subset of kNN classifiers. *Advances in Data Analysis and Classification*. doi:10.1007/s11634-015-0227-5

[2] O. Mahmoud, A.P. Harrison, A. Perperoglou, A. Gul, Z. Khan, M. Metodiev, and B. Lausen (2014). A feature selection method for classification within functional genomics experiments based on the proportional overlapping score. *BMC Bioinformatics*, **15**(1), 274–294.

[3] Z. Khan, A. Gul, O. Mahmoud, M. Miftahuddin, A. Perperoglou, W. Adler, and B. Lausen (2016). An ensemble of optimal trees for class membership probability estimation. In A. Wilhelm, H.A. Kestler (eds.), *Analysis of Large and Complex Data*, Springer-Verlag, Berlin, 395–409.

Latent variable modelling of interdependent ego-networks

I. Gollini

Birkbeck College, London, UK

Session: Classification and Network Modelling, Room: EA2 *Wednesday 12th, 14:30 – 15:00*

Ego-networks consist of a focal node (“ego”) and the nodes to whom ego is directly connected to (“alters”). We present a Bayesian latent variable network modelling framework for describing the connectivity structure of interdependent ego-networks (network of ego-networks) by using the latent space approach. This allows us to visualise both the ego-ego and ego-alter relational structures by estimating the positions of all the individuals in a latent “social space”. We apply this new methodology using an efficient variational algorithm in order to explore the structure and roles of human smuggling network out of Libya (operation Glauco II) consisting of 29 interconnected ego-networks and involving more than 15 thousand alters.

Improving the efficiency of Bayesian computation for network models

A. Caimo

Dublin Institute of Technology, Ireland

Session: Classification and Network Modelling, Room: EA2 *Wednesday 12th, 15:00 – 15:30*

Recent research in statistical network analysis has demonstrated the advantages and effectiveness of Bayesian approaches to relational data. In this talk we present a Markov Chain Monte Carlo algorithm based on a pre-computation strategy which samples from the posterior parameter space approximating the intractable likelihood of exponential random graph models and therefore circumventing the need of computing their doubly intractable posterior distribution. The approaches turn out to significantly improve the efficiency and scalability of Bayesian methods for exponential random graph models.

On clustering validation: the internal perspective

M.G.M.S. Cardoso

Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit, Portugal

Session: Statistical Learning in Data Science, Room: EA4 *Wednesday 12th, 14:00 – 14:30*

The internal evaluation of a clustering solution relies on the clustering base data, no class structure being *a priori* known. It essentially regards the properties of cohesion-separation and stability of clusters found in data. Sometimes, the determination of the number of clusters is also considered in this setting, which can also be viewed within a model selection perspective.

To discuss clustering evaluation (or clustering validation as it is usually designated), a tour to some of the numerous clustering validation indices is conducted. For the indices of agreement a typology is proposed that relies on their behaviour under the hypothesis of agreement by chance, e.g. [1]. A weighted clustering cross-validation approach is also presented [4]. Finally, the importance of, in a specific application, obtaining an interpretable clustering solution is underlined and some instruments to provide such a solution (e.g. [2, 3]) are referred.

An application to data from the more recent round of European Social Survey illustrates the process of clustering validation. The data regards citizens' political engagement: sets of questions referring to whether the citizens were involved "in different ways of trying to improve things in their country or help prevent things from going wrong" (e.g. "taken part in lawful public demonstration in last 12 months").

Keywords: clustering validation, cohesion-separation, stability, indices of agreement, European Social Survey.

References

[1] M.J. Amorim and M.G.M.S. Cardoso (2015). Comparing clustering solutions: the use of adjusted paired indices. *Intelligent Data Analysis*, **19**(6), 1275–1296.

[2] J.-P. Baudry, M.G.M.S. Cardoso, G. Celeux, M.J. Amorim, and A.S. Ferreira (2015). Enhancing the selection of a model-based clustering with external categorical variables. *Advances in Data Analysis and Classification*, **9**(2), 177–196.

[3] M.G.M.S. Cardoso (2012). Logical discriminant models. In L. Moutinho and K.-H. Huarng (eds.), *Quantitative Modeling in Marketing and Management*, World Scientific.

[4] M.G. Cardoso, K. Faceli, and A.C. de Carvalho (2010). Evaluation of clustering results: the trade-off bias-variability. In *Classification as a Tool for Research*, Springer.

Controversies in health data science

P.P. Rodrigues

Fac. Medicina and CINTESIS, Univ. Porto, Portugal

Session: Statistical Learning in Data Science, Room: EA4 *Wednesday 12th, 14:30 – 15:00*

During this talk we will discuss three controversies which are still creating friction for the full application of data science in healthcare and medical research. First, we approach van der Lei's 1st law of medical informatics, which states that data shall be used only for the purpose for which they were collected, still subject to fierce debate, but supported by numerous studies on data quality and association confounding. Then, we discuss to which extent innovations in analytical methods have alleviated the need to conduct expensive randomised clinical trials, as data science enthusiasts claim that big data and machine learning can be used to answer all research questions, traditionalists claim that there is no replacement for randomised experiments when we are interested in causation. Finally, we address access to data, as many researchers have the opinion that all medical and healthcare data should be made freely available to them without any restrictions, in order to accelerate research and improve medical knowledge, while data custodians fear privacy breaches and loss of public trust, especially with new restrictive European directives turned into law. The goal is to foster discussion, probably providing more questions than answers, but highlighting the current pitfalls of health data science.

Data pre-processing methods for forecasting with spatio-temporal data

L. Torgo

INESC TEC / FCUP, Univ. Porto, Portugal

Session: Statistical Learning in Data Science, Room: EA4 *Wednesday 12th, 15:00 – 15:30*

In this work we present, discuss and compare different data pre-processing techniques designed to help in building effective forecasting models for spatio-temporal data sets. This type of data has dependencies among observations which can be temporal, spatial or spatio-temporal. Ignoring these dependencies at model building leads to unreliable prediction models. We present a series of data pre-processing techniques designed with the goal of deriving new predictor variables that encode these observation dependencies to improve the predictive accuracy of the resulting models. We describe these techniques and evaluate them on real world spatio-temporal data.

Robust joint modelling of mean and dispersion for the GLM framework

P. Segaert[a], S. Van Aelst[a], T. Verdonck[a]

[a]*KU Leuven, Belgium*

Session: Robust Statistics II, Room: EA2 *Wednesday 12th, 15:45 – 16:05*

Generalized linear models form a unified way of modelling the mean response under different distributions belonging to the exponential family. Because of their flexibility, they have been widely studied and have become a powerful tool in statistics. Yet in practice, real data often show a larger or smaller variability than expected from the model. In these cases the data is said to be overdispersed and underdispersed, respectively. The dispersion itself may also change for different observations in the data. For example, two groups — indicated by a factor variable — may have a different dispersion factor. The dispersion may thus also depend on a set of predictors.

Properly accounting for dispersion is crucial for several reasons. Firstly, confidence intervals for coefficients of the mean depend on the estimated dispersion. Secondly, neglecting dispersion may result in a loss of efficiency and bias in the estimation of the coefficients. Thirdly, the dispersion model itself may be the focus of interest.

A typical problem in analysing real data is the possible presence of outliers in the data. As classical methods try to fit an optimal model for all observations they are highly susceptible to these atypical observations. It is important to note that these outliers are not necessarily errors in the data. Their presence may reveal that the data is more heterogeneous than assumed. They may also come in clusters, indicating there are subgroups in the population that behave differently. A robust analysis can thus provide better insights in the data and reveal underlying structures that would remain hidden in a classical analysis.

Therefore, we propose a robust procedure for jointly modelling the mean and dispersion under the GLM framework. Using the ideas of the double exponential distribution proposed by Efron [2], we extend the work of Cantoni and Ronchetti [1] who proposed a robust GLM estimator for the mean response. Our methodology does not suppose constant dispersion but models both mean and dispersion behaviour based on a possibly different set of predictors. As such, our methodology is very flexible and models both over- and underdispersion. We will briefly discuss properties of the proposed methodology and discuss the problem of robust inference. The performance of the estimator and the proposed test will be validated by a simulation study.

Keywords: generalized linear models, dispersion, outliers.

References

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

[2] B. Efron (1986). Double exponential families and their use in generalized linear regression, *Journal of the American Statistical Association*, **81**, 709–721.

Performance and robustness in statistical testing of hypotheses for data with Markov dependencies

A. Kharin[a], **E. Vecherko**[a]

[a]*Belarusian State University*

Session: Robust Statistics II, Room: EA2 *Wednesday 12th, 16:05 – 16:25*

The problem of statistical analysis of dependencies in data appears in many applications. Markov dependencies serve as convenient probability models for an effective solution of this problem, especially for discrete data. Although the optimal statistical decision rules are constructed for certain hypothetical models, data often does not follow the models exactly and, even under minor deviations from the hypothetical models, the procedures lose their optimality [1]. Here, we consider the following important problems: performance analysis of statistical tests under deviations from the hypothetical models of Markov dependence, deviations identification, and robust tests construction.

We develop our previous results [2]. For a discrete Markov chain $\{x_t\}$ under deviations of the factual transition probabilities matrices $\tilde{P}^{(0)}$, $\tilde{P}^{(1)}$ from the hypothetical values $P^{(0)}$, $P^{(1)}$ (“dependency outliers”), we construct asymptotic expansions for the sequential test performance characteristics — error probabilities and expected numbers of observations. With the minimax criterion [3], the robust sequential test is constructed.

The problem of statistical detection of “white noise embeddings” (replacement outliers) in a stationary binary Markov chain $\{x_t\}$ is also considered. The “white

noise" is a Bernoulli stationary process with parameter $p = 1/2$. Under the influence of "white noise" the Markov chain $\{x_t\}$ becomes a non-Markovian binary process $\{y_t\}$, and the order of stochastic dependence increases: $s \to \infty$. Statistical tests for detecting "white noise embeddings" in a Markov chain are constructed; they are based on run statistic, short run statistic and likelihood ratio statistic. For a family of contiguous alternatives the asymptotic powers of the tests based on run statistic and short run statistic are found. The performance of statistical inferences is evaluated. The computer experiments and illustrations are presented for different cases and show when it is possible to distinguish $\{y_t\}$ from a Markov chain.

Keywords: Markov chain, statistical test, outlier.

References

[1] Yu. Kharin (2013). *Robustness in Statistical Forecasting*. Springer, New York.

[2] A. Kharin (2004). Robustness in sequential discrimination of Markov chains under "contamination". In *Theory and Applications of Recent Robust Methods*, Birkhauser, Basel, 165–171.

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

Forecasting with robust exponential smoothing with trend and seasonality

R. Crevits, C. Croux

KU Leuven

Session: Robust Statistics II, Room: EA2 *Wednesday 12th, 16:25 – 16:45*

Simple forecasting methods such as exponential smoothing are very popular in business analytics. This is not only due to their simplicity, but also because they perform very well. Incorporating trend and seasonality into an exponential smoothing method is standard. In a highly cited paper, Hyndman and Khandakar [1] developed an automatic forecasting method using exponential smoothing, available as the R package `forecast`. We propose the package `robets`, an outlier robust alternative of the function `ets` in the `forecast` package. For each method of a class of exponential smoothing variants we made a robust alternative. The class includes methods with a damped trend and/or seasonal components. The robust method is developed by robustifying the original exponential smoothing variant: we provide robust forecasting equations, robust initial values, robust smoothing parameter estimation and a robust information criterion. The method is an extension of [2]. The code for our R package is based on the function `ets` of the `forecast` package. The usual functions for visualizing the models and forecasts also work for `robets` objects. Additionally, there is a function `plotOutliers` which highlights outlying values in a time series.

Keywords: time series, forecasting, robust estimation, R package.

References

[1] R.J. Hyndman and Y. Khandakar (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, **27**(4).

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

Entropy in high-dimensional variable selection

P. Macedo[a], J.P. Cruz[a]

[a]*CIDMA – Center for Research and Development in Mathematics and Applications, Department of Mathematics, University of Aveiro, Portugal*

Session: Regression and Beyond, Room: EA3 *Wednesday 12th, 15:45 – 16:05*

In 1996, in their book *Maximum Entropy Econometrics*, Golan, Judge and Miller proposed a variable selection procedure using normalized entropy measures [1]. To our knowledge, the idea has not received special attention in the literature since then, probably due to the fact that supports for the model parameters are needed. To overcome this difficulty, we propose a novel approach to define the supports based on the ridGME procedure, which combines the analysis of a ridge trace and generalized maximum entropy estimation [2]. The well-known Hald and Prostate Cancer data sets are used to illustrate the good performance of this variable selection procedure.

Keywords: maximum entropy, ridge trace, variable selection.

References

[1] A. Golan, G.G. Judge, and D. Miller (1996). *Maximum Entropy Econometrics: Robust Estimation with Limited Data.* Wiley, Chichester, UK.

[2] P. Macedo (2017). Ridge regression and generalized maximum entropy: an improved version of the ridge–GME parameter estimator. *Communications in Statistics – Simulation and Computation*, doi:10.1080/03610918.2015.1096378

Regularized B-spline regression for high-dimensional scattered data

J. Wagner[a], R. Münnich[b]

[a]*Trier University, DFG-RTG Algorithmic Optimization, Germany;* [b]*Trier University, Economic and Social Statistics, Germany*

Session: Regression and Beyond, Room: EA3 *Wednesday 12th, 16:05 – 16:25*

Estimating timber reserves in several forest districts is a fundamental task in the German national forest inventory (GNFI). Recently, remote sensing data were used in a regularized B-spline approach [3]. To improve the estimates, it is desirable to incorporate multiple input data, but all such methods suffer from the curse of

dimensionality [1] which describes an exponential growth of the underlying linear system in the dimension of the input data. Even for moderate dimensions, storing the system matrix exceeds the memory capacity of common computational systems and efficient numerical methods are required to solve the multiple regression problem.

To address this challenge, a regularized tensor-product B-spline [2] is proposed, yielding a system matrix with a special structure. Exploiting this structure, the system is solved with low storage costs using suitable numerical methods. This allows incorporating multiple input data into regression models and especially into the GNFI estimation process. To verify the proposed approach, the timber reserves are estimated and compared to the original estimates obtained by the one-dimensional model.

Keywords: high-dimensional regression, regional estimates, regularization, scattered data approximation.

References

[1] R. Bellman (1957). *Dynamic Programming.* Princeton University Press.

[2] C. de Boor (1978). *A Practical Guide to Splines*. Springer.

[3] J. Wagner, R. Münnich, J. Hill, J. Stoffels, and T. Udelhoven (Forthcoming, 2017). Nonparametric small area models using shape-constrained penalized B-splines. *Journal of the Royal Statistical Society A*. doi:10.1111/rssa.12295

Contributions to the analysis of inhomogeneous large-scale data using maximum entropy

M.C. Costa[a], P. Macedo[a]

[a]*University of Aveiro and CIDMA*

Session: Regression and Beyond, Room: EA3 *Wednesday 12th, 16:25 – 16:45*

It was already in the fifties of the last century that the relationship between information theory, statistics, and maximum entropy was established, following the works of Kullback, Leibler, Lindley and Jaynes. However, the applications were restricted to very specific domains and it was not until recently that the convergence between information processing, data analysis and inference demanded the foundation of a new scientific area, commonly referred to as Info-Metrics [1].

As huge amount of information and large-scale data have become available, the term “big data” has been used to refer to the many kinds of challenges presented in its analysis: many observations, many variables (or both), limited computational resources, different time regimes or multiple sources. In this work, we consider one particular aspect of big data analysis which is the presence of inhomogeneities, compromising the use of the classical framework in regression modelling. Maximin effects [2] and magging [3] can be presented as recent approaches to this problem.

A new approach is proposed in this work, with the introduction of the concepts of info-metrics to the analysis of inhomogeneous large-scale data. The framework of information-theoretic estimation methods is presented, along with some information measures. In particular, the normalized entropy is tested in aggregation procedures and some preliminary simulation results are presented.

Keywords: big data, info-metrics, maximum entropy.

References

[1] A. Golan (2013). On the state of the art of info-metrics. In V.N. Huynh et al, *Uncertainty Analysis in Econometrics with Applications*. Springer-Verlag, 3–15.

[2] N. Meinshausen and P. Bühlmann (2015). Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, **43**(4), 1801–1830.

[3] P. Bühlmann and N. Meinshausen (2014). Magging: maximin aggregation for inhomogeneous large-scale data. arXiv:1409.2638v1

Providing analysts the tools they need within a modern national statistics office

T. Savage[a], C. Hansen[a], A. Seyb[a]

[a] *Stats NZ*

Session: Tools for Data Analytics, Room: EA4 *Wednesday 12th, 15:45 – 16:05*

Stats NZ are looking to move away from the collection and publication of standalone sample surveys to making use of a wide range of data sources and estimation strategies. A key component to enabling this change is to develop the infrastructure which allows analysts to explore, test and use a range of tools which are not traditionally heavily used within National Statistics Offices. This talk will focus on two ways we are looking to do this. The first is the development of enterprise level support for the statistical package R which includes the creation of internal RStudio and Shiny servers. This complements the already heavily supported use of SAS within the organisation and has enabled users to make use of a wide range of tools. The second is a series of initiatives developed through a process where external companies and internal employees were asked to pitch projects that would leverage technology to transform the way we delivered information and data to users. Our goal was to break away from our current way of thinking and doing things, and leverage external expertise rather than trying to do everything in house. The initiatives range from using Big Data tools as a service, to formalising the process for producing reproducible research. We'll outline what the initiatives are and how they have progressed.

Keywords: official statistics, RStudio, big data.

A flexible optimization tool for multivariate optimal allocation problems under high-dimensional data

M. Rupp[a], R. Münnich[a]

[a]*Economic and Social Statistics Department, University of Trier*

Session: Tools for Data Analytics, Room: EA4 *Wednesday* 12^{th}*, 16:05 – 16:25*

The aim of modern surveys is to provide accurate information on a large variety of variables on different regional levels and subclasses of high-dimensional populations. Hence, optimal allocation of a fixed total sample size has to consider a vast number of strata along with optimization conflicts due to the complementary information of the variables of interest and uncertainty of auxiliaries. Furthermore, particular quality and cost restrictions might be taken into account.

In this paper, we present an efficient and flexible optimization tool for solving multivariate optimal allocation problems. Thereby, while possibly including a large variety of constraints, we achieve a significant reduction of the computational burden compared to other classical allocation methods, and even for large problem instances.

The allocation problem is stated as a multi-objective optimization problem as shown in [1]. Taking advantage of its special structure and applying Pareto optimization, the problem can be equivalently reformulated as a significantly lower-dimensional non-smooth problem, as described in [2] and [3] for the univariate case. This problem is solved via a semi-smooth Newton method in analogy to [1].

The performance of the developed optimization tool is tested on a household data set of Germany.

Keywords: data analysis, design optimization, non-smooth optimization.

Acknowledgements: This research is supported within the project *Research Innovation for Official and Survey Statistics* (RIFOSS), funded by the German Federal Statistical Office, and by the research training group 2126 *Algorithmic Optimization* (ALOP), funded by the German Research Foundation DFG.

References

[1] U. Friedrich, R. Münnich, and M. Rupp (2017). Multivariate optimal allocation under box-constraints. *Submitted for publication.*

[2] S. Gabler, M. Ganninger, and R. Münnich (2012). Optimal allocation of the sample size to strata under box constraints. *Metrika*, **75**(2), 151–161.

[3] R. Münnich, E. Sachs, and M. Wagner (2012). Numerical solution of optimal allocation problems in stratified sampling under box constraints. *AStA Advances in Statistical Analysis*, **96**(3), 435–450.

Data analysis with contaminated data

V. Ardelean

GfK SE

Session: Tools for Data Analytics, Room: EA4 *Wednesday 12th, 16:25 – 16:45*

When analysing data there is an implicit assumption that the data is uncontaminated, i.e. all observed data are from the same data-generating process. We consider two types of contamination, anomalies and noisy data.

Anomalies or outliers are observations that seem to be inconsistent with the assumed model for the data. Measurement error or noise in the data is the difference between the true but unknown value and the observed value. This variability is an inherent part of the measurement process, often found in data used for market research purposes.

In order to avoid a negative impact of such data constellations we investigate different options to clean contaminated data.

Keywords: noise filtering, anomaly detection, predictive analytics.

Computing over the Internet applied to data visualization: an illustrative example in geometry

J.L. de Miranda[ac], **M. Casquilho**[bc]

[a]Escola Superior de Tecnologia e Gestão (School of Technology and Management), Instituto Politécnico de Portalegre; [b]Instituto Superior Técnico, University of Lisbon; [c]CERENA (Centre for Natural Resources and the Environment), Instituto Superior Técnico

Session: Applications I, Room: EA2 *Wednesday 12th, 17:00 – 17:20*

We address a Statistics-related illustrative example that is solved by combining the following constituents: computing, the Internet, Statistics, and visualization. These parts are put together in the example [1] described below and solved by applying Monte Carlo simulation, a technique that typically generates large amounts of data. The particular statistical problem aims to assess the behaviour of the distance from a given source point (in or out of a circle) to random points in a circle of unit radius (without loss of generality). We thus wish to view an empirical PDF (probability density function) of the distance, and have some related information, such as its mean and standard deviation.

This example is one of our many problems (in Engineering, etc.) solved by "computing over the Internet", that is: making the problem available on a Web page and solving it with the user's data, just with a browser and no software installation. This permits access from any terminal (PC, smartphone, etc.) with any common operating system. We are aware of only one author with the same approach, V.M. Ponce [3], whose impressive Vlab solves numerous problems in Hydraulics, e.g. [2], although without dynamic images.

In our problem, the user supplies: x_0 of the source point (abscissa alone, w.l.o.g. from circular symmetry); and N, the number of random destination points in the circle. The result for, say, $x_0 = 0.6$ and $N = 10^6$ is surprising and made clear only by viewing the PDF plot at the cited Web site [1]. Observe that the analytically derivable result for $x_0 = 0$ is simply $y = 2x$ (the CDF x^2). Visualization of the plots is very beneficial for understanding the results, as is often the case for Monte Carlo methods.

Keywords: computing over the Internet, engineering, visualization.

References

[1] M. Casquilho (2017). Distance to points in circle. `http://web.tecnico.ulisboa.pt/~mcasquilho/compute/or/Fx-distInCircle.php`, accessed January 2017.

[2] V.M. Ponce (2017). Normal depth in a prismatic channel. `http://ponce.sdsu.edu/onlinechannel01.php`, accessed January 2017.

[3] V.M. Ponce (2017). Visualab (Vlab). `http://ponce.sdsu.edu/992epilogue.html`, accessed January 2017.

Using word clouds as an e-learning analytics tool based on data visualization and statistics

J.M. Sanchez-Gomez[a], M.A. Vega-Rodriguez[b], C.J. Perez Sanchez[c], F. Calle-Alonso[d]

[a]Cátedra ASPgems, University of Extremadura, Spain; [b]Department of Technologies of Computers & Communications, University of Extremadura, Spain; [c]Department of Mathematics, University of Extremadura, Spain; [d]Research & Development Department, ASPgems SL, Spain

Session: Applications I, Room: EA2 *Wednesday 12th, 17:20 – 17:40*

Nowadays, most online educational platforms do not have tools for e-learning analytics. In addition, they also do not have visual representations of the information contained in the platform. In this work, word clouds are implemented as a tool to provide an overview of the main words (concepts) used by the students in their communications [1]. In this way, teachers can get a thorough understanding of the student learning process through the texts that they write [2]. Furthermore, the proposed word-cloud tool can help teachers to make decisions about strategies and pedagogical guidance, and they can promote and support students' participation and activity [3]. In particular, this tool consists of three different word clouds: the ideal word cloud, with the learning words according to the theoretical weight indicated by the teacher; the real word cloud, with the most used words by the students; and the mixture of both, with the learning words according to their use by the students. These three word clouds provide a quick visual summary at the learning unit level or at the course level. Besides, the word clouds allow the teacher to compare word clouds of different students with the ideal word cloud to rate them. This rate is a metric calculated automatically by the set of statistics provided by the tool. To conclude, the word-cloud tool has many advantages from a didactic point of view. A future research line is to test this tool with a high number of students during one or several courses.

Keywords: word clouds, e-learning analytics, data visualization and statistics.

Acknowledgements: This research has been supported by the Spanish Ministry of Economy, Industry and Competitiveness (Centro para el Desarrollo Tecnológico Industrial, contract IDI-20161039), the Government of Extremadura (contract AA-16-0017-1, and projects GR15106 and GR15011), Cátedra ASPgems, and the European Union (European Regional Development Fund).

References

[1] M. Resendes, M. Scardamalia, C. Bereiter, B. Chen, and C. Halewood (2015). Group-level formative feedback and metadiscourse. *International Journal of Computer-Supported Collaborative Learning*, **10**(3), 309–336.

[2] I. Tuns (2016). The potential of Semantic Web technologies for student engagement analytics. *LSBM Working Paper Series*, **1**(1), 18–37.

[3] A.M. Tervakari, K. Silius, J. Koro, J. Paukkeri, and O. Pirttila (2014). Usefulness of information visualizations based on educational data. In *2014 IEEE Global Engineering Education Conference (EDUCON)*, Istanbul, 142–151.

Economic growth and tourism in Turkey: Hsiao's Granger causality analysis

D. Alptekin

Department of Statistics, Hacettepe University, Turkey

Session: Applications I, Room: EA2 *Wednesday 12th, 17:40 – 18:00*

A country's development and economic growth depends on finding solutions to economic problems such as the balance of payments deficit, unemployment. The tourism sector has huge impacts on the economy, so it has become a leading sector in the world over the past decades. In the literature, the relationship between economic growth and tourism for both developed and developing countries has been extensively researched since 1990. Knowledge of the causal relationship between economic growth and tourism is of importance to policy makers, as tourism policies are becoming major concerns for these countries. The aim of this study is to estimate the causal relationship between economic growth and tourism in Turkey over the period 1980–2016 by applying the Phillips–Perron unit root test, the Johansen co-integration test and Hsiao's Granger causality. Tourism revenues, international tourist arrivals and economic growth series are stationary in first difference and neither series is co-integrated. According to Hsiao's Granger causality, it is found

that there is no causal relationship between economic growth and tourism in Turkey over the period 1980–2016.

Keywords: economic growth, Hsiao's Granger causality, tourism.

References

[1] J.P. Cerdeira Bento (2016). Tourism and economic growth in Portugal: an empirical investigation of causal links. *Tourism & Management Studies*, **12**(1), 164–171.

[2] Q. Hou (2009). The relationship between energy consumption growth and economic growth in China. *International Journal of Economics and Finance*, **1**(2), 232–237.

[3] S. Creaco and G. Querini (2003). The role of tourism in sustainable economic development. In *43rd Congress of the European Regional Science Association*, Finland, 1–17.

Matching administrative data for census purposes

R.M. Silva[a], L. Sampaio[a], P. Calado[a], M.J. Silva[a], A. Delgado[b]

[a]*INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal;* [b]*Statistics Portugal*

Session: Big Data Platforms, Room: EA4 *Wednesday 12th, 17:00– 17:20*

A feasibility study is under way at Statistics Portugal to transform the Portuguese Census, based on the classical door-to-door distribution of questionnaires, to a combined or a register-based census, in which the data is to be progressively obtained from Public Administration databases. Our objective is to identify and register each individual resident in Portugal in a specific year in a Resident Population Database, to be derived from those data sources. There are multiple hurdles to the creation of this database: records have inconsistencies and errors due to manually inserted data, and the Data Protection Authority (CNPD) imposes anonymisation criteria on the datasets. Attempting the record linkage between sources using exact comparison methods would leave out many potential matches (roughly 10% in our case). Our presentation details the followed approach and the results of the initial experiments on matching some of the available information sources. So far, we were able to train a probabilistic model for matching the Civil Population Register with the Tax Authority Register, using manual pairings from a third data source as gold standard. The precision of the method is about 99% on the gold standard. Also, applying the training model to 6,933,367 records from Civil Population Register and 4,414,595 records from the Tax Authority Register that were not in the gold standard, we were able to find 77,649 new matches and confirm 3,258,614 records were already in the BPR with our method.

Keywords: record linkage, census, machine learning applications.

Rspark: running R and Spark in Docker containers

E.J. Harner[a], M. Lilback[b], W. Foreman[b]

[a]*West Virginia University;* [b]*Rc²ai*

Session: Big Data Platforms, Room: EA4 *Wednesday 12th, 17:20 – 17:40*

Docker containers allow Hadoop, Spark, and other big-data platforms to be run virtually on multiple host platforms with identical configurations and functional capabilities. This allows collections of containers to be launched together with specific requirements, e.g. a specified version of Spark and compatible versions of R and other resources. Using Docker, DevOps can automate the deployment of big-data platforms that meet defined specifications.

Rspark is a collection of Docker containers for running R, Hadoop, and Spark with various persistent data stores including PostgreSQL, HDFS, HBase, Hive, etc. At this time, the server side of Rspark runs on Docker's Community Edition, which can be: on the same machine as the client, on a server, or in the cloud. Currently, Rspark supports an RStudio client, but any R-based web client could be adapted to work.

The direction of computing is in virtualization. The Docker containers in Rspark can be orchestrated by Kubernetes to build arbitrarily large virtual clusters for R and/or for Hadoop/Spark. A virtual cluster of Spark containers using Kubernetes can be built with a persistent distributed data store, e.g. HDFS. The ultimate goal is to build data science workflows, e.g. ingesting streaming data into Kafka, modulating it into a data store, and passing it to Spark Streaming.

Keywords: Spark, Docker containers, Kubernetes.

BitQuery – a GitHub API driven and D3 based search engine for open source repositories

L. Borke[a], S. Bykovskaya[b]

[a]*Humboldt-Universität zu Berlin;* [b]*Lomonosov Moscow State University*

Session: Big Data Platforms, Room: EA4 *Wednesday 12th, 17:40 – 18:00*

With the growing popularity of GitHub, the largest host of source code and collaboration platform in the world, it has evolved to a Big Data resource offering a variety of open source repositories (OSR). Multiple libraries and package managers, among them CRAN, CPAN, WordPress and many others, mirror their data in GitHub organizations, while a growing number of developers are releasing their packages directly on GitHub. We present BitQuery [1], a new GitHub API driven search engine which (I) provides an automatic OSR categorization system for data science teams and software developers, and (II) establishes visual data exploration (VDE) and topic driven navigation of GitHub organizations for collaborative reproducible research and web deployment.

The BitQuery architecture consists of three abstraction layers, following the ETL paradigm (Extract, Transform, Load), or, equivalently, the visual analytics approach (data management, analysis, and visualization). First, the information is extracted via the GitHub API based parser layer. Next, the Smart Data layer transforms Big Data into value, processing the data semantics and metadata via dynamic calibration of metadata configurations, text mining (TM) models and clustering methods [2, 3]. One of the examined TM models is the latent semantic analysis (LSA) technique which measures semantic relations and allows dimension reduction. Both layers were implemented via several novel R packages, forming the basis of a self-contained "GitHub Mining infrastructure in R" [4]. Thus derived Smart Data is loaded into the web-based visual analytics application (VA-App) realized via the D3-3D Visu layer, which is powered by two JavaScript libraries for producing dynamic data visualizations in web browsers: D3.js and Three.js. The D3-3D Visu layer was designed in full compliance with the so-called visual information seeking mantra: "Overview first, zoom and filter, and then details-on-demand". Various techniques and interactive interfaces perform VDE from multiple perspectives.

The application spectrum of BitQuery is illustrated by the exploration of the "R universe", a massive collection of all R packages on GitHub including CRAN and Bioconductor. This example shows a great potential of BitQuery as a VA-App which increases the visibility and discoverability of any organization or digital library hosted on GitHub.

Keywords: software mining, clustering analysis, visual analytics.

References

[1] `http://bitquery.borke.net/`

[2] L. Borke and W.K. Härdle (2017). Q3-D3-LSA. In W.K. Härdle, H.H. Lu, and X. Shen (eds.), *Handbook of Big Data Analytics*, Springer.

[3] L. Borke and W.K. Härdle (2017b). GitHub API based QuantNet mining infrastructure in R. *SFB 649 Discussion Paper*, Humboldt Universität zu Berlin.

[4] L. Borke and S. Bykovskaya (2017). GitHub mining infrastructure in R. *Forthcoming*.

Thursday, 13 July 2017

From softmax to sparsemax: a sparse model of attention and multi-label classification

A.F.T. Martins

Unbabel, Portugal; Instituto de Telecomunicações, Instituto Superior Técnico, Portugal

Invited Lecture Session, Room: Abreu Faro *Thursday 13th, 9:00 – 9:45*

The softmax transformation is a key component of several statistical learning models, encompassing multinomial logistic regression, action selection in reinforcement learning, and neural networks for multi-class classification. Recently, it has also been used to design attention mechanisms in neural networks, with important achievements in machine translation, image caption generation, speech recognition, and various tasks in natural language understanding and computation learning.

In this talk, I will describe sparsemax, a new activation function similar to the traditional softmax, but able to output sparse probabilities. After deriving its properties, I will show how its Jacobian can be efficiently computed, enabling its use in a neural network trained with backpropagation. Then, I will propose a new smooth and convex loss function which is the sparsemax analogue of the logistic loss. An unexpected connection between this new loss and the Huber classification loss will be revealed. We obtained promising empirical results in multi-label classification problems and in attention-based neural networks for natural language inference. For the latter, we achieved a similar performance as the traditional softmax, but with a selective, more compact, attention focus.

Keywords: sparse modeling, Huber loss, neural networks.

Learning on timestamped medical data

M. Spiliopoulou

Faculty of Computer Science, Otto-von-Guericke-University Magdeburg

Invited Lecture Session, Room: Abreu Faro *Thursday 13th, 9:45 – 10:30*

There is a proliferation of mining methods for the collection, curation and analysis of timestamped clinical data for identifying comorbidities, predicting response to treatments and supporting clinical decisions. Moreover, data miners are being called to support epidemiologists in understanding how participants of epidemiological studies evolve over time. In this talk, I present results on mining the participant data of a population-based study and the patient data of a study using an mHealth application. I elaborate on the challenge of learning on systematically incomplete timestamped data, and on the potential of semi-supervised learning approaches.

Keywords: medical mining data, mining systematically incomplete data.

References

[1] T. Hielscher, M. Spiliopoulou, H. Völzke, and J.-P. Kühn (2014). Mining longitudinal epidemiological data to understand a reversible disorder. In *Proc. of the 13th Int. Symposium on Intelligent Data Analysis (IDA 2014)*, Springer.

[2] T. Hielscher, M. Spiliopoulou, H. Völzke, and J.-P. Kühn (2014). Using participant similarity for the classification of epidemiological data on hepatic steatosis. In *Symposium on Computer-Based Medical Systems*, IEEE.

[3] T. Hielscher, M. Spiliopoulou, H. Völzke, and J.-P. Kühn (2016). Identifying relevant features for a multi-factorial disorder with constraint-based subspace clustering. In *Symposium on Computer-Based Medical Systems*, IEEE.

[4] U. Niemann, T. Hielscher, M. Spiliopoulou, H. Völzke, and J.-P. Kühn (2015). Can we classify the participants of a longitudinal epidemiological study from their previous evolution? In *Symposium on Computer-Based Medical Systems*, IEEE.

[5] U. Niemann, H. Völzke, J.-P. Kühn, and M. Spiliopoulou (2014). Learning and inspecting classification rules from longitudinal epidemiological data to identify predictive features on hepatic steatosis. *Expert Systems with Applications*, **41**(11), 5405–5415.

[6] V. Unnikrishnan (2017). Analysis of patient evolution on time series of different lengths. Master's thesis at the Faculty of Computer Science, Otto-von-Guericke-University Magdeburg. Under supervision of M. Spiliopoulou.

More powerful test procedures for multiple hypothesis testing

S. Zhang[a], H.-S. Chen[b]

[a]Department of Statistics, University of Central Florida, USA; [b]Statistical Methodology and Application Branch, Division of Cancer Control and Population Sciences, National Cancer Institute, MD, USA

Session: Methods for Data Science, Room: EA2 *Thursday 13th, 11:00 – 11:20*

We propose a new multiple test called *minPOP* and one modified version (the left truncated) for testing a large number of multiple hypotheses simultaneously. We show that multiple testing procedures based on these tests have strong control of the family-wise error rate. A method for finding the p-values of the multiple testing procedures after adjusting for multiplicity is also developed. Simulation results show that the *minPOP* tests in general have higher global power than the existing well-known multiple tests, especially when the number of hypotheses being compared is relatively large. Among the multiple testing procedures we developed, we find that the left truncated *minPOP* test has lower number of false rejections than the existing multiple testing procedures.

Keywords: multiple test, stepwise procedure, adjusted p-value.

Incomplete ranking

E. Stoimenova

Institute of Information and Communication Technologies and Institute of Mathematics and Informatics, Bulgarian Academy of Sciences

Session: Methods for Data Science, Room: EA2 *Thursday 13th, 11:20– 11:40*

There are many situations, in which people are presented a large number of items to rank and they do not need to completely specify the ranking of all n items. The goal of the experiment might be to rank only their favourite k out of n items or just to choose their k favourite items. In other cases it is important to classify items into groups or categories according to some reasonable criterion of "goodness".

The general partitioning problem can be described as follows. Let $\{1, \dots, n\}$ be n given items. We wish to partition them into a fixed number of disjoint categories, such that each category contains a certain preassigned number of items. The first category contains n_1 favourite items, the second category contains the n_2 next preferred items, and so on; the final category contains the n_r least favourite items, where $\sum n_i = n$, $n_i \geq 1$. We do not state any preferences among members of the same category. If we assign values to r and n_i we obtain several special cases of interest.

(1) To choose the best single item ($r = 2, n_1 = 1, n_2 = n - 1$);

(2) To choose the best k items without regard to order ($r = 2, n_1 = k, n_2 = n - k$);

(3) To choose the best k items with regard to order ($r = k + 1, n_1 = \cdots = n_k = 1, n_{k+1} = n - k$);

(4) To order all items ($r = n, n_1 = \cdots = n_r = 1$);

(5) To partition the items into a fixed number of categories.

Many of the decision procedures one might use within the scope of these ranking problems have a structure which is invariant under a group of transformations. We consider suitable models for the analysis of such partially ranked data.

Keywords: top k ranking, partial ranking, distances on partial rankings.

Clustering and disjoint principal component analysis: an empirical comparison of two approaches

E. Macedo[a], **A. Freitas**[b], M. Vichi[c]

[a]*TEMA, University of Aveiro, Portugal;* [b]*DMat & CIDMA, University of Aveiro, Portugal;* [c]*University "La Sapienza", Italy*

Session: Methods for Data Science, Room: EA2 *Thursday 13th, 11:40– 12:00*

A new constrained principal component analysis for multivariate data, called Clustering and Disjoint Principal Component Analysis (CDPCA), was proposed to identify

clusters of objects and, simultaneously, describe the data matrix in terms of sparse and disjoint components, which become useful for interpretation and visualization purposes. Recently, two different heuristic iterative procedures, one described as an alternating least squares method (ALS) and another based on semidefinite programming models (Two-Step-SDP), were suggested to perform CDPCA.

We empirically compare and evaluate the performance of these algorithms using three real gene expression data sets, with different number of classes of objects and where the true classification of objects is known. The model error, the between cluster deviance, the proportion of explained variance by the new components, the accuracy of object classification and the running time were computed to assess the quality of the overall fit of the CDPCA model and the efficiency of each algorithm.

Our numerical tests show that both procedures perform well and suggest that the Two-Step-SDP approach provides faster results, while the ALS algorithm is better in terms of solution precision.

Keyword: principal component analysis.

Acknowledgements: This work was supported by Portuguese funds through CIDMA – Center for Research & Development in Mathematics and Applications, and the Portuguese Foundation for Science and Technology (FCT – Fundação para a Ciência e a Tecnologia), within project UIDMAT041062013.

References

[1] E. Macedo (2015). Two-step-SDP approach to clustering and dimensionality reduction. *Stat., Optim. Inf. Comput.*, **3**(3), 294–311.

[2] E. Macedo and A. Freitas (2015). The alternating least-squares algorithm for CDPCA. In A. Plakhov et al (eds.), *Optimization in the Natural Sciences*, Springer-Verlag, 173–191.

[3] M. Vichi and G. Saporta (2009). Clustering and disjoint principal component analysis. *Computational Statistics & Data Analysis*, **53**, 3194–3208.

Estimation, simulation, and visualization of spatial and spatiotemporal autoregressive conditional heteroscedasticity

P. Otto

Department of Statistics, European University Viadrina, Germany

Session: Time and Space, Room: EA3 *Thursday 13th, 11:00 – 11:20*

Otto, Schmid, and Garthoff (2016) introduce a new spatial model that incorporates heteroscedastic variance depending on neighboring locations [1]. The proposed process is regarded as the spatial equivalent to the temporal autoregressive conditional heteroscedasticity (ARCH) model. In contrast to the temporal ARCH model, in which the distribution is known given the full information set of the prior periods, the distribution is not straightforward in the spatial and spatiotemporal setting. However, it is possible to estimate the parameters of the model using the maximum-likelihood approach. Moreover, we combine the well-known spatial autoregressive model with the spatial ARCH model assuming heteroscedastic errors. In this talk, I focus on the estimation from a computational and practical point of view. From this perspective, the log-likelihood function is usually sufficient to get accurate parameter estimates by using any non-linear, numerical optimization function. To compute the likelihood for a certain set of parameters, the determinant of the Jacobian matrix must be computed, which often requires large computational capacities, especially for large data sets. The estimation procedure is implemented in the R-package `spGARCH`.

Keywords: spatial ARCH process, maximum likelihood estimation, visualization.

References

[1] P. Otto, W. Schmid, and R. Garthoff (2016). Generalized spatial and spatiotemporal autoregressive conditional heteroscedasticity. *European University Viadrina Frankfurt (Oder), Discussion Paper Series*, **387**, 1–40.

Performance of statistical approaches to model binary responses in longitudinal studies

M.H. Gonçalves[a], M.S. Cabral[b]

[a]*CEAUL and Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade do Algarve, Portugal;* [b]*CEAUL and DEIO, Faculdade de Ciências da Universidade de Lisboa, Portugal*

Session: Time and Space, Room: EA3 *Thursday 13th, 11:20 – 11:40*

Longitudinal binary studies are a powerful design and they are commonly encountered in experimental and observational research. In these studies, repeated observations of a response variable are taken over time on each individual. In such cases, the repeated measures are likely to be correlated and the autocorrelation structure plays a significant role in the estimation of regression parameters. A frequent problem in longitudinal studies is the presence of missing data. The generalized linear mixed effects model (GLMM) is recommended to analyse this kind of data when the goal of the study is a subject-specific interpretation because it allows missing values on the response, provided they are missing at random (MAR), and it accounts for the correlation among the repeated observations by the inclusion of random effects in the linear predictor. However, GLMM assumes that observations of the same subject are independent conditional to the random effects and covariates which may not be true. To overcome this problem, the methodology implemented in the `R` package `bild` uses GLMM with binary Markov chain (GLM3C) as the basic stochastic mechanism to accommodate serial dependence and the odds-ratio to measure dependence between successive observations [1]. Taking into account the correlation structure in the GLM3C approach, missing values on the response are allowed provided they

are MAR with some further restrictions. The aim of this paper is to give a statistical assessment of the approaches considered in terms of efficiency and coverage probability. To achieve that goal, a simulation study was carried out in which the `R` packages `bild` and `lme4` were used.

Keywords: generalized linear mixed effects model, correlation structure, missing data.

References

[1] M.H. Gonçalves, M.S. Cabral, and A. Azzalini (2012). The R package bild for the analysis of binary longitudinal data. *Journal of Statistical Software*, **46**(9), 1–17.

Air pollution forecasting with time series neural network models

B. Alptekin[a], C.H. Aladag[b]

[a]*Dept. of Statistics, Middle East Technical University, Turkey;* [b]*Dept. of Statistics, Hacettepe University, Turkey*

Session: Time and Space, Room: EA3 *Thursday 13th, 11:40– 12:00*

Air pollution is a vital issue for human beings and a very common problem in large cities around the world. Forecasting air pollution levels is therefore an important task, one which has been addressed in various studies. One way to forecast air pollution levels is to use time series methods. Recently, artificial neural networks have been widely used as an effective time series forecasting tool in the related literature. This study employs multiple neural network models to forecast air pollution in the city of Istanbul. This approach requires identifying which neural network models produce good forecasts. In this study, we apply the weighted information criterion (WIC) in order to reach a high level of accuracy. It is shown that using WIC to find a good forecasting model produces very accurate forecasts for air pollution in Istanbul.

Keywords: air pollution, artificial neural networks, forecasting.

References

[1] C.H. Aladag (2011). A new architecture selection method based on tabu search for artificial neural networks. *Expert Systems with Applications*, **38**, 3287–3293.

[2] E. Egrioglu, C.H. Aladag, and S. Gunay (2008). A new model selection strategy in artificial neural network. *Applied Mathematics and Computation*, **195**, 591–597.

[3] M. Boznar, M. Lesjak, and P. Mlakar (1993). A neural network-based method for short-term predictions of ambient SO_2 concentrations in complex terrain. *Atmospheric Environment*, **27B**(2), 221–230.

[4] M.W. Gardner and S.R. Dorling (1999). Neural network modeling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. *Atmospheric Environment*, **33**, 709–719.

Periodic multivariate INAR processes

C.S. Santos[a], I. Pereira[b], M.G. Scotto[c]

[a]*Coimbra College of Agriculture, Polytechnic Institute of Coimbra and CIDMA;* [b]*DMat and CIDMA, University of Aveiro;* [c]*DMat and CEMAT, IST, University of Lisbon*

Session: Time and Space, Room: EA3 *Thursday 13th, 12:00 – 12:20*

In this work a multivariate integer-valued autoregressive model of order one with periodic time-varying parameters, and driven by a periodic innovations sequence of independent random vectors is introduced. We extend the results of [1] to the multivariate case. Keeping an eye on the practical applications we will restrict our attention to the diagonal matrix case [2]. Emphasis is placed on models with periodic multivariate negative binomial innovations. Basic probabilistic and statistical properties of the novel model are discussed. Aiming to reduce the computational burden arising from the use of the conditional maximum likelihood method, a composite likelihood-based approach is adopted. The performance of such method is compared with that of some traditional competitors, namely moment estimators and conditional maximum likelihood estimators. Furthermore, forecasting is also addressed. An application to a multivariate data set of time series concerning the monthly number of fires in three districts in mainland Portugal is also presented.

Keywords: multivariate models, binomial thinning operator, composite likelihood.

References

[1] M. Monteiro, M.G. Scotto, and I. Pereira (2015). A periodic bivariate integer-valued autoregressive model. In J.P. Bourguignon , R. Jelstch, A. Pinto, and M. Viana (eds.), *Dynamics, Games and Science - International Conference and Advanced School Planet Earth, DGS II*, Springer International Publishing, 455–477.

[2] X. Pedeli and D. Karlis (2011). A bivariate INAR(1) process with application. *Statistical Modelling*, **11**(4), 325–349.

Visualization for large-scale Gaussian updates

J. Rougier[a], **A. Zammit-Mangion**[b]

[a]*University of Bristol, UK;* [b]*University of Wollongong, Australia*

Session: Visualization I, Room: EA4 *Thursday 13th, 11:00 – 11:20*

In geostatistics and also in other applications in science and engineering, it is now common to perform updates on Gaussian process models with many thousands or even millions of components. These large-scale inferences involve modelling, representational and computational challenges. We describe a visualization tool for large-scale Gaussian updates, the 'medal plot' [1]. The medal plot shows the updated uncertainty at each observation location and also summarizes the sharing of information across observations, as a proxy for the sharing of information across

the state vector (or latent process). As such, it reflects characteristics of both the observations and the statistical model. We illustrate with an application to assess mass trends in the Antarctic Ice Sheet, for which there are strong constraints from the observations and the physics [2].

Keywords: uncertainty quantification, visualizing uncertainty, medal plot.

References

[1] J. Rougier and A. Zammit-Mangion (2016). Visualization for large-scale Gaussian updates. *Scandinavian Journal of Statistics*, **43**(4), 1153–1161.

[2] A. Zammit-Mangion, J. Rougier, N. Schön, F. Lindgren, and J. Bamber (2015). Multivariate spatio-temporal modelling for assessing Antarctica's present-day contribution to sea-level rise. *Environmetrics*, **26**(3), 159–177.

Visual support for rastering of unequally spaced time series

C. Bors[a], M. Bögl[a], T. Gschwandtner[a], S. Miksch[a]

[a] *Vienna University of Technology, Austria*

Session: Visualization I, Room: EA4 *Thursday 13th, 11:20 – 11:40*

Cleansing and wrangling [1, 2] — preprocessing data and transforming it into a usable form — constitutes an important step for subsequent analysis. In many application domains, e.g., environmental sensor measurements, datasets are created with varying interval lengths. Specifically with time series data, established analysis methods require the data to be structured, e.g., being equally spaced. By rastering time series, unevenly distributed time points and their corresponding values are aggregated and binned into evenly spaced time intervals, while still retaining the original data's structure. Rastering the original data alters it to (1) trade consistent value distribution for accuracy of the original values, (2) achieve more accurate value representation by smoothing measurement inaccuracies, and (3) reduce data size by lowering the time series resolution. Users require knowledge about the data domain and temporal aspects of the data to generate an adequately transformed time series usable for subsequent analysis. Rastering introduces uncertainty, which users are predominantly not made aware of in further analysis.

We propose a Visual Analytics (VA) approach to effectively support users during the analysis and validation of time series raster parametrizations. VA intertwines interactive visualization, analytical methods, perception and cognition to ease the information discovery process. Our conceptualized VA framework allows users to transform unequally spaced time series data into equally spaced rasters. It facilitates finding appropriate parametrizations and analyzing the rastering outcome. By providing quality measures and uncertainty information, users receive contextual knowledge on quality issues occurring during rastering to assess the outcome of the time series rastering. For different time series characteristics it is necessary to adapt the rastering algorithm and feedback information accordingly. We provide considerations for handling special use cases and domain specific properties and suggest well-fitting measures to deal with intricacies in the data.

Keywords: time series preprocessing, visualization, visual analytics.

References

[1] S. Kandel, J. Heer, C. Plaisant, et al (2011). Research directions in data wrangling: visualizations and transformations for usable and credible data. *Information Visualization*, **4**, 271–288.

[2] T. Gschwandtner, W. Aigner, S. Miksch, et al (2014). TimeCleanser: a visual analytics approach for data cleansing of time-oriented data. *Proc. of the 14th International Conf. on Knowledge Technologies and Data-driven Business*, 18:1–18:8, ACM.

Visualization of three-dimensional data with virtual reality

J.E. Lee[a], S. Ahn[a], D.-H. Jang[a]

[a]*Department of Statistics, Pukyong National University*

Session: Visualization I, Room: EA4 *Thursday 13th, 11:40 – 12:00*

A variety of data visualization methods are utilized to analyze huge amounts of data. Among various methods, a three-dimensional image requires rotations to show a stereo image on a two-dimensional screen. This study discusses two methods (batch method and real-time method) which make it possible to analyze the construction of stereo images to improve the restriction of the three-dimensional image display with virtual reality. This investigation can be useful to explore three-dimensional data structures more clearly.

Keywords: data visualization, virtual reality, stereo image, R package.

R visual tools for three-way data analysis

M. Gallo[a], V. Todorov[b], M.A. Di Palma[a]

[a]*University of Naples "L'Orientale", Italy;* [b]*United Nations Industrial Development Organization (UNIDO), Austria*

Session: Visualization I, Room: EA4 *Thursday 13th, 12:00 – 12:20*

Standard multivariate analysis addresses data sets represented as two-dimensional matrices. In recent years, an increasing number of application areas like chemometrics, computer vision, econometrics and social network analysis involve analysing data sets that are represented as multidimensional arrays and multiway data analysis has become popular as an exploratory tool (see [1]). The most popular multiway models are CANDECOMP/PARAFAC and TUCKER3 [3]. The results from a three-way analysis can be presented in several different ways (see [3]), the first one being tables of the coefficients or loadings for each mode, either rotated or not. While it is important to inspect the numerical output of the methods for the analysis of three-way data (the component matrices and the core array) in order to properly interpret the results, of great help can be different visual representations of these outcomes. The most typical plots are: (i) Pair-wise graphs of the components for each mode separately, (ii) All-components plots which will show all components of a single mode using the levels of the mode as X-axis, (iii) Per-component plot, showing a single component on all modes simultaneously in the same plot, (iv) Joint biplots for Tucker 3 models and (v) Trajectory plots.

We present an R package, **rrcov3way**, implementing a set of functions for the analysis of multiway data sets, including PARAFAC and TUCKER3 as well as their robust alternatives. Apart from basic tools for data handling and preprocessing of multidimensional arrays, tools for displaying raw data as well as the models in two- and three-dimensional plots are provided. Several examples based on the data sets available with the package are used to demonstrate the basic usage of the functions and illustrate some of the graphical results obtainable with the software. These graphical procedures, mainly based on [2] and [3], are flexible enough to give the user the possibility to design the graphs according to the needs and the data at hand but at the same time provide suitable default parameters which ease their use.

Keywords: three-way analysis, visualization, R.

References

[1] E. Acar and B. Yener (2009). Unsupervised multiway data analysis: a literature survey. *IEEE Trans. Knowl. Data Eng.*, **21**(1), 6–20.

[2] H.A. Kiers (2000). Some procedures for displaying results from three-way methods. *Journal of Chemometrics*, **14**(3), 151–170.

[3] P.M. Kroonenberg (2008). *Applied Multiway Data Analysis*. Wiley series in probability and statistics, John Wiley and Sons, Hoboken, NJ.

Mapping the Milky Way halo: modeling and classification of sparsely sampled vector-valued functions

J.P. Long

Department of Statistics, Texas A&M University, United States

Session: Visualization and Analysis of Modern Data, Room: EA2 *Thursday 13th, 14:00 – 14:30*

Modern time domain astronomical surveys measure temporal changes in brightness of tens of millions of stars. Each star is recorded as a vector-valued function sampled at irregular intervals. We discuss current statistical challenges and methodology for modeling and classifying these vector-valued functions. Using Dark Energy Survey

data, we illustrate progress towards using variable stars to map structure in the Milky Way Halo, the region of space that surrounds our galaxy.

Keywords: functional data, classification, astronomy.

Summarizing linearized prediction models along feature groups

Y. Benjamini

Hebrew University of Jerusalem, Israel

Session: Visualization and Analysis of Modern Data, Room: EA2 *Thursday 13th, 14:30– 15:00*

Linearized regression models have been getting big, especially with the increased use of basis expansions and convolutional networks for representing the data before the regression. Often, when the representation is sparse, a dense regression obtained by ridge or nuclear norm regularization can achieve better prediction accuracy than sparse alternatives. Alas, such dense rules are much harder to interpret, and often remain black-boxes.

One way to visualize and interpret such large models is to summarize the weight vector for different feature groups. In this talk, I discuss several desired properties for these summaries, and suggest an *impact statistic* that can summarize the signed effect of the group. As a motivating example, I will discuss how the proposed method helps describe the responses of neurons in the V4 cortical area to an observed sequence of natural images. Using this method, we can generate insights regarding the fitted models and develop new hypotheses for function of these neurons.

This is based on joint work with Julien Marial and Bin Yu, and in collaboration with Jack Gallant's vision lab at UC Berkeley.

Keywords: regression, interpretation, visual cortex.

Using aggregated relational data to feasibly identify network structure without network data

T.H. McCormick

University of Washington

Session: Visualization and Analysis of Modern Data, Room: EA2 *Thursday 13th, 15:00– 15:30*

An individual's social environment influences many economic and health behaviors. Social network data, consisting of interactions or relationships between individuals, provide a glimpse of this environment but are extremely arduous to obtain. Collecting network data via surveys is financially and logistically prohibitive in many circumstances, whereas online network data are often proprietary and only informative about a subset of possible relationships. Designing efficient sampling strategies,

and corresponding inference paradigms, for social network data is, therefore, fundamental for scalable, generalizable network research in the social and behavioral sciences. This talk proposes methods that estimate network features (such as centrality or the fraction of a network made up of individuals with a given trait) using data that can be collected using standard surveys. These data, known as aggregated relational data (ARD), poll individuals about the number of connections they have with certain groups in the population, but do not measure any links in the graph directly. We demonstrate the utility of the proposed models using data from a savings monitoring experiment in India. This is joint work with Emily Breza (Harvard), Arun Chandrasekhar (Stanford), and Mengjie Pan (University of Washington).

Statistical issues with agent-based models

D. Banks

Dept. of Statistical Science, Duke University, United States

Session: ISBIS Session, Room: EA4 *Thursday 13th, 14:00 – 14:30*

Many, many fields use agent-based models (ABMs) for prediction and insight. Google holds virtual auctions, economists model markets, urban planners study traffic flow, ecologists examine species interactions, and epidemiologists forecast disease spread. But the statistical properties of such models are almost unstudied. ABMs are different from standard statistical models, such as the linear model, in that one can rarely write out the likelihood function, and thus our usual strategies for making quantified inferences and fitting parameters are unworkable. This talk reviews the history and scope of ABMs, followed by a description of how emulators can be used to make approximate inference without an explicit likelihood.

Balanced incomplete block designs: some applications and visualization

T.A. Oliveira[ab], A. Oliveira[ab]

[a]*Universidade Aberta, Portugal;* [b]*Center of Statistics and Applications, University of Lisbon, Portugal*

Session: ISBIS Session, Room: EA4 *Thursday 13th, 14:30 – 15:00*

In the area of Experimental Design, a new method of arranging variety trials involving a large number of varieties was introduced by Yates (1936): the Balanced Incomplete Block Designs (BIBD). When the block size is not enough to accommodate all the varieties of an experiment, BIBD allows their allocation in such a way that every variety appears the same number of times in the design and all pairs of varieties concur together the same number of times, along the blocks. Initially proposed to face problems in the area of Agriculture, very quickly these designs were explored in applications on many other areas and emerged as one of the most interesting top-quality research topics of the new century. The huge number of application areas, mathematical and optimal properties of these designs have been highlighted and developed not only by statisticians, but by mathematicians and computer scientists. In this work we will follow the research evolution emphasizing the main reasons why BIBDs are known as a rigorous and powerful tool providing features to continue at the future forefront, accomplishing the technological and computational evolution. Some visualization examples will be provided.

Randomized singular spectrum analysis for long time series

P.C. Rodrigues[ab], P. Tuy[b], R. Mahmoudvand[c]

[a]*University of Tampere, Finland;* [b]*Federal University of Bahia, Brazil;* [c]*Bu-Ali Sina University, Iran*

Session: ISBIS Session, Room: EA4 *Thursday 13th, 15:00 – 15:30*

Singular spectrum analysis (SSA) is a relatively new and powerful nonparametric method for analyzing time series that is an alternative to the classic methods. This methodology has proved effective for analyzing time series in various disciplines as the assumptions of stationarity and Gaussian residuals can be relaxed. The Era of Big Data has brought very long and complex time series. Although SSA provide advantages over traditional methods, the computational time needed for the analysis of long time series might make it inappropriate. In this work, we propose the randomized SSA which intends to be an alternative to SSA for long time series without losing the quality of the analysis. The SSA and the randomized SSA are compared in terms of quality of the analysis and computational time, using Monte Carlo simulations and real data about the daily prices of five of the major world commodities.

Alternative distributions to Weibull for modeling the wind speed data in wind energy analysis

F.G. Akgül[a], B. Şenoğlu[b]

[a]*Artvin Çoruh University;* [b]*Ankara University*

Session: Applications II, Room: EA2 *Thursday 13th, 15:45 – 16:05*

Identification of the wind speed characteristics is very important for the researchers and the practitioners working in that area. There exist a considerable number of studies about modeling the wind speed characteristics in the literature, see [1, 2, 3]. Although, one of the most widely used statistical distributions for modeling the wind speed data is Weibull, it may not provide the best fit for all wind regimes. For this reason, alternative distributions are used for modeling wind speed data.

This study uses the Inverse Weibull, Burr Type III and Extreme Value distributions to model wind speeds as an alternative to the Weibull distribution. We wish to identify the distributions which provide the best fit for different wind regimes encountered in nature.

We compare fitting performances of these distributions to real data sets obtained from different stations in Turkey. In estimating the distribution parameters, the maximum likelihood (ML) methodology is used. We also determine the distribution which has better modeling performance for each data set by using the root mean squares error and coefficient of determination criteria. It is shown that the Inverse Weibull and Burr Type III distributions fit the wind speed data better than Weibull. In conclusion, alternative distributions can be preferred over Weibull in identifying the wind speed characteristics for various wind regimes encountered in nature.

Keywords: maximum likelihood, wind speed distributions, wind energy.

References

[1] J.A. Carta, P. Ramirez, and S. Velazquez (2009). A review of wind speed probability distributions used in wind energy analysis: case studies in the Canary Islands. *Renew. Sustain. Energy Rev.*, **13**, 933–955.

[2] E.C. Morgan, M. Lackner, M.V. Richard, and L.G. Baise (2011). Probability distributions for offshore wind speeds. *Energy Convers. Manage.*, **52**, 15–26.

[3] F.G. Akgül, B. Şenoğlu, and T. Arslan (2016). An alternative distribution to Weibull for modeling the wind speed data: inverse Weibull distribution. *Energy Convers. Manage.*, **114**, 234–240.

Hierarchical cluster analysis in the context of performance evaluation: from classical to complex data

Á.S.T. Sousa[a], M.G.C. Batista[a], O.L. Silva[b], M.C. Medeiros[c], H. Bacelar-Nicolau[d]

[a]*University of Azores, CEEAplA;* [b]*University of Azores, CICS.NOVA.UAc;* [c]*University of Azores;* [d]*University of Lisbon*

Session: Applications II, Room: EA2 *Thursday 13th, 16:05 – 16:25*

Given the computational and methodological advances of recent decades, it is possible to synthesize data in terms of their most relevant concepts, which may be described by different types of complex data, also known as symbolic or complex data. In this work, we assess the performance of a set of employees of a company in the dentistry sector, based on a classical data matrix and on a complex data matrix.

Twenty-three individuals participated in the study, and only those who perform functions at the office and reception assistance level (twelve) were evaluated. The remainder are included in this process as evaluators. Data were collected through two questionnaires of “Performance Evaluation” (PE): questionnaire 1 (“Self-evaluation” / “Evaluation by the Superior”) and questionnaire 2 (“360° Evaluation”). The first one corresponds to the application of two traditional Performance Evaluation (PE) methods, and the second one to a modern method, called “360° Evaluation” (an important tool of Human Resources Management). In the present work, each collaborator is evaluated directly by all individuals who interact with him (various organizational actors). Therefore, we obtained a complex data table where the data units (each one of the collaborators) are described by variables whose values are frequency distributions with different number of modalities. The data were analysed based on several statistical methods, among which we highlight some graphical visualization methods (e.g. Zoom-Star, 2D) and some algorithms of Ascendant Hierarchical Cluster Analysis (AHCA). The AHCA was based on the weighted generalized affinity coefficient [1] combined with classical and probabilistic aggregation criteria.

In general, we conclude that the individuals present a satisfactory performance, evidencing some differences between self-perception and external perception. The clustering structures obtained, in the case of the AHCA of the individuals, allowed to detect groups of collaborators with results referring to these types of evaluation relatively similar. The results obtained allowed to identify factors susceptible of influencing the performance, in order to adopt measures that promote the continuous improvement of the performance of each employee.

Keywords: performance evaluation, hierarchical cluster analysis, visualization.

References

[1] H. Bacelar-Nicolau, F.C. Nicolau, Á. Sousa, and L. Bacelar-Nicolau (2009). Measuring similarity of complex and heterogeneous data in clustering of large data sets. *Biocybernetics and Biomedical Engineering*, **29**(2), 9–18.

Building a map of Europe based on citizens' values: an interval data approach

P. Brito[a], M.G.M.S. Cardoso[b], A.P. Duarte Silva[c]

[a]Faculdade de Economia & LIAAD INESC TEC, Universidade do Porto, Portugal; [b]Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit, Portugal; [c]Católica Porto Business School & CEGE, Universidade Católica Portuguesa, Portugal

Session: Applications II, Room: EA2 *Thursday 13th, 16:25 – 16:45*

How similar are the values of Europeans across regions? Do neighbour regions share similar values? Which profiles may be identified? To investigate such issues, data from the European Social Survey are analyzed at region level. This is a transnational survey of European citizens taking place every two years throughout Europe since 2001. The survey applies Schwartz' human values scale [5]. Values such as "Important to think new ideas and being creative" or "Important to live in secure and safe surroundings" are measured on a 6-point ordinal scale, from 1="Very much like me" to 6="Not like me at all". The data on 40,185 citizens are aggregated into regions and weighted by population size combined with post-stratification weight.

In a first step, Categorical Principal Component Analysis is applied to the individual ordinal data, leading to eleven factors, which jointly account for around 72% of the total inertia. Factorial scores are then aggregated by region, in the form of intervals,

resulting in a 250×11 interval data array [1]. Outlier detection in the multivariate interval data [4] is then addressed, using the Gaussian parametric modelling for the intervals' MidPoints and Log-Ranges proposed in [2]. With the goal of identifying groups of regions with similar values, a model-based approach to the clustering of interval data is applied [3]. According to the BIC criterion, a homoscedastic solution with five clusters is selected. The corresponding profiles bring new insights regarding the homogeneity-heterogeneity of human values across the regions.

Keywords: European Social Survey, interval data, interval clustering.

References

[1] P. Brito (2014). Symbolic data analysis: another look at the interaction of data mining and statistics. *WIREs Data Mining and Knowledge Discovery*, **4**(4), 281–295.

[2] P. Brito and A.P. Duarte Silva (2012). Modelling interval data with normal and skew-normal distributions. *Journal of Applied Statistics*, **39**(1), 3–20.

[3] P. Brito, A.P. Duarte Silva, and J.G. Dias (2015). Probabilistic clustering of interval data. *Intelligent Data Analysis*, **19**(2), 293–313.

[4] A.P. Duarte Silva, P. Filzmoser, and P. Brito (2015). Outlier detection in interval data. SDA 2015, 5th Workshop on Symbolic Data Analysis, Orléans, France.

[5] S.H. Schwartz (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, **2**(1).

Estimating the location and scale parameters of the Maxwell distribution

T. Arslan[a], S. Acitas[b], **B. Şenoğlu**[c]

[a]*Eskisehir Osmangazi University;* [b]*Anadolu University;* [c]*Ankara University*

Session: Inference, Room: EA3 *Thursday 13th, 15:45 – 16:05*

In this study, we obtain the modified maximum likelihood (MML) estimators of the location and scale parameters of the Maxwell distribution. See Maxwell [1] and Dey et al [2] for further information about the Maxwell distribution. See also Tiku [3] for more detailed information about the MML methodology. Then we compare the efficiencies of the MML estimators with the corresponding maximum likelihood (ML), moments and least squares (LS) estimators by using the Monte Carlo simulation study. In the comparisons, we use the bias and the mean square error (MSE) criteria. Simulation results showed that the MML estimators have more or less the same efficiency as the corresponding ML estimators. Moreover, the ML method produces more accurate estimators. On the other hand, MML estimators are preferred if the focus is on the computational cost. A real data set is analyzed at the end to illustrate the implementation of the proposed methodologies.

Keywords: Maxwell distribution, modified maximum likelihood, efficiency.

References

[1] J.C. Maxwell (1860). Illustrations of the dynamical theory of gases. Part I. On the motions and collisions of perfectly elastic spheres. *Philosophical Magazine*, 4th series, 19, 19–32.

[2] S. Dey, T. Dey, S. Ali, and M.S. Mulekar (2016). Two-parameter Maxwell distribution: properties and different methods of estimation. *Journal of Statistical Theory and Practice*, **10**(2), 291–300.

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

Copula density estimation by Lagrange interpolation at the Padua points

L. Qu

Department of Mathematics, Boise State University, Idaho, USA

Session: Inference, Room: EA3 *Thursday 13th, 16:05 – 16:25*

A multivariate distribution with continuous margins can be uniquely decomposed via a copula and its marginal distributions. The problem considered here is to estimate the copula density function non-parametrically by Lagrange interpolation at the Padua points. The so-called “Padua points” [2] give a simple, geometric and explicit construction of bivariate polynomial interpolation in a square. Moreover, the associated Lebesgue constant has minimal order of growth (log square of the degree). Fast algorithms have been implemented for bivariate Lagrange interpolation at the Padua points in a square [1]. When a copula density is approximated by Lagrange interpolation at the Padua points, the likelihood function can be expressed in terms of the coefficients of the Chebyshev polynomials. The uniform margins and symmetry constraints for a copula density are enforced by linear equality constraints on the interpolation coefficients. The positivity constraints for a density are enforced by linear inequality constraints. The likelihood subject to the linear equality and inequality constraints is maximized by an augmented Lagrangian method [3] which is particularly suitable for the large-scale optimization problem. A data-driven selection of the regularization parameter — total degree of the polynomial — is made through information criteria for model selection. Simulation and real data application show the effectiveness of the proposed approach.

Keywords: density estimation, low-rank approximation, Chebyshev polynomials.

References

[1] M. Caliari, S. De Marchi, A. Sommariva, and M. Vianello (2011). Padua2DM: fast interpolation and cubature at the Padua points in Matlab/Octave. *Numer. Algorithms*, **56**, 45–60.

[2] M. Caliari, S. De Marchi, and M. Vianello (2005). Bivariate polynomial interpolation on the square at new nodal sets. *Appl. Math. Comput.*, **165**, 261–274.

[3] J. Nocedal and S. Wright (2006). *Numerical Optimization.* Springer-Verlag.

Estimation of percentile of marginal distribution of order statistic with real life application

D. Kushary

Rutgers University – Camden, New Jersey, USA

Session: Inference, Room: EA3 *Thursday 13th, 16:25 – 16:45*

The distribution of the kth order statistic from a continuous distribution is well known when the sample size n is known. But many times in practice the sample size n is random while k is fixed. Hence it is necessary to compute and use the marginal distribution of the kth order statistic. In a real life application in auto industry it is required to estimate the percentile of the marginal distribution. In this presentation, we propose some parametric and nonparametric solutions. A computational algorithm is suggested and its precision is measured using simulation.

The virtues and pitfalls of the visualisation of incomplete categorical data

J. Nienkemper-Swanepoel[a], S. Gardner-Lubbe[a], N.J. Le Roux[a]

[a] *Stellenbosch University, South Africa*

Session: Visualization II, Room: EA4 *Thursday 13th, 15:45 – 16:05*

An ongoing investigation into the optimal visualisation of categorical data sets with missing values will be presented. Visualising incomplete data enables the recognition of response patterns and evaluation of the effect of the unobserved information on the interpretable information. Configurations obtained from subset multiple correspondence analysis (sMCA) are compared to configurations resulting from multiple imputation procedures for incomplete simulated categorical data sets. sMCA preserves the original scaffolding of the data while enabling a focused view of a subset of variables. The first step in applying sMCA to missing data is to create a missing value category level for each variable. Once the missing value category level is allocated, the observed and unobserved information are easily distinguishable. Non-responses can be evaluated by visual exploration of the missing category levels. It is anticipated that imputation methods could be more suitable for data sets with a high fraction of missing values, especially data sets suffering from a large amount of lost information. To determine the validity of sMCA, generalised Procrustes analysis and Rubin's rules (GPAbin) will be used to compare the sMCA biplots with the MCA biplots of combined multiple imputed data sets. Results from a simulation study will be presented to evaluate the configurations from the sMCA and GPAbin methods in the context of the MCA biplot obtained from the complete data.

Keywords: biplots, incomplete categorical data, subset multiple correspondence analysis.

Is Arthur Batut's geometric experiment a convincing argument in favor of Francis Galton's generic images?

A. de Falguerolles

Université de Toulouse III, France

Session: Visualization II, Room: EA4 *Thursday 13th, 16:05 – 16:25*

In a recent book, Sephen M. Stigler [5] recalls Francis Galton's construction (1879) of generic images obtained by superimposing and merging photographic images [4] and their links to the question of averaging data (a series of photographs, in this context). It turns out that Galton's idea was rapidly entertained in France by some photographers. In particular, Arthur Batut (1822–1911) published a booklet [1] which shows that he completely mastered the technique in 1887.

I will recall Galton's proposal and investigate how his idea was received near Toulouse by such an accomplished amateur photographer (see [3]). Note in passing that some of Arthur Batut's photographs have been exhibited at the Metropolitan Museum of Art of New York, at Houston, and at Washington.

I will then present how Arthur Batut [2] devised an experiment to illustrate the efficiency of composite portraiture for finding a hidden feature. This experiment can be seen as a simple case of pattern recognition in a geometric background. But its pretense of objectivity makes it pleasant to revisit.

Keywords: statistics, visualization, composite portraiture.

References

[1] A. Batut (1887). *La Photographie Appliquée à la Production du Type d'une Famille, d'une Tribu ou d'une Race*. Paris: Gauthier-Villars.

[2] A. Batut (1890). Étude sur la formation des images composites. *La Nature*, **18**, 188–190.

[3] D. Blanc, C. Cusani, and A. de Falguerolles (2017). Arthur Batut (1846–1918) et les portraits composs de Francis Galton (1822–1911). Submitted to the *Journées de Statistique 2017.*

[4] F. Galton (1879). Generic images. *Proceedings of the Royal Institution of Great Britain.* **9**, 161–170.

[5] S.M. Stigler (2016). *The Seven Pillars of Statistical Wisdom*. Harvard University Press, Cambridge (Mass.) and London.

Visualization of sparse two-way contingency tables

V.O. Choulakian

Département de Math/Statistique, Université de Moncton, NB, Canada

Session: Visualization II, Room: EA4 *Thursday 13th, 16:25 – 16:45*

The aim of this talk is two-fold. First, we attempt to quantify sparsity in contingency tables by a 7-number summary based on the minimal size of an equivalent contingency table. We use the invariance property of correspondence analysis (CA) and taxicab correspondence analysis (TCA) to construct the equivalence class of contingency tables. Second, we argue that a comparison of CA and TCA maps is enriching.

Keywords: sparse contingency table, 7-number summary, visualization.

References

[1] M. Greenacre (2013). The contributions of rare objects in correspondence analysis. *Ecology*, **94**(1), 241–249.

[2] J. Tukey (1977). *Exploratory Data Analysis*. Addison-Wesley, Massachusetts.

[3] V. Choulakian (2017). Taxicab correspondence analysis of sparse two-way contingency tables. arXiv:1508.06885v2

Cyclical history theory in data visualization: using a four-quadrant display to see history repeating itself

A. Alexandrino da Silva

Instituto Universitário de Lisboa (ISCTE-IUL), CIES-IUL, Portugal

Session: Visualization II, Room: EA4 *Thursday 13th, 16:45 – 17:05*

Statistical graphics history is recent, about 250 years old. Much older are cartography, the musical notation or the Cartesian axes on which graphics are based [1]. The evolution that has taken place since the first line, bar or pie charts were invented by William Playfair is remarkable, particularly after the computer advent. The rise of statistical graphics can't be dissociated from the historical period they're in: the Industrial Revolution, the development of sciences, the 1920's, the growing role of literacy or socialization of information that was followed by the obscurantist phase called "Modern Dark Ages" [2]. "The re-birth of data visualization" with the invention of new exploratory information graphics [3] to handle huge amounts of data was only possible with the increase of computer processing. The availability of big data coming from the Internet and the Internet of Things (IoT) associated with computation power has created the need for new techniques for visualizing/understanding data. One could think that the end of the story was near, but we're far from it: data visualization is constantly reinventing itself.

This paper aims to contribute to the study of theoretical and historical approaches to data visualization. It proposes a quadrant model of analysis based on two perpendicular axes that translate the lesser or greater complexity of the graphical design options and the lesser or greater complexity of the concepts represented. These axes allow us to model four profiles that range from complex graphic design that describe elaborated concepts (1st quadrant) to simple graphic design that show plain ideas (3rd quadrant). The position in the quadrant is related to the technology development process, meaning the farther from the center the more technology is used. Data visualization uses a similar framework as 100 years ago: simplicity in graphic design to describe different ideas. This cyclical process can be described through an upward spiral shape along the four quadrants pushed towards the periphery due to technology and computer development.

Keywords: data visualization, information graphics, history milestones.

References

[1] W.S. Cleveland (1987). Research in statistical graphics. *J. American Statistical Association*, **398**(82), 419–423.

[2] M. Friendly and D.J. Denis (2001). *Milestones in the history of thematic cartography, statistical graphics, and data visualization*. Web document, `http://www.datavis.ca/milestones/`, accessed April 2017.

[3] J.W. Tukey (1977). *Exploratory Data Analysis*. Addison-Wesley.

Early detection of Parkinson’s disease by considering acoustic features of plosive consonants

D. Montaña[a], Y. Campos-Roca[a], **C.J. Perez**[b]

[a]*Department of Technologies of Computers & Communications, University of Extremadura, Spain;* [b]*Department of Mathematics, University of Extremadura, Spain*

Poster Session, Location: North Tower, first floor *Thursday 13th, 17:20 – 18:40*

An approach is proposed for automatic detection of early-stage Parkinson’s Disease (PD) by using Diadochokinesis (DDK) tests. The method is based on temporal and spectral features (Mel Frequency Cepstrum Coefficients (MFCCs) and spectral moments) extracted from the Voice Onset Time (VOT) segments of /ka/ syllables.

An important drawback to test the effectiveness of a detection methodology in an early-stage scenario is the scarcity of data. Therefore, a voice recording database has been collected. It is composed of 27 individuals diagnosed with PD and 27 healthy controls (HCs). The average disease stage is 1.85 ± 0.55 according to the Hoehn and Yahr (H&Y) scale.

Three different classification experiments on the three plosive segments (/p/, /t/ and /k/) have been performed combining feature selection and Support Vector

Machine learning. The approach based on /k/ plosive shows the best performance in comparison to the corresponding versions based on the other two plosives (/p/ and /t/), achieving an accuracy rate of 92.2% (using 10-fold cross-validation). It is notable that these results were obtained on a database with a lower average disease stage than previous articulatory databases presented in the literature [1, 2]. The sensitivity is estimated as 97.3%. This high sensitivity is especially relevant in primary health care, because it would allow the physician to identify PD in early stages and refer the patient to a neurological unit, reducing the cases of undiagnosed PD.

Keywords: acoustic features, classification, Parkinson's disease.

Acknowledgements: This research has been supported by the Spanish Ministry of Economy, Industry and Competitiveness (Project MTM2014-56949-C3-3-R), the Government of Extremadura (Projects GR15106 and GR15011), and European Regional Development Funds.

References

[1] M. Novotný, J. Rusz, R. Čmejla, and E. Růžička (2014). Automatic evaluation of articulatory disorders in Parkinson's disease. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22**, 1366–1378.

[2] J.R. Orozco-Arroyave, F. Hönig, J.D. Arias-Londoño, J. Vargas-Bonilla, E. Nöth, K. Daqrouq, S. Skodda, and J. Rusz (2016). Automatic detection of Parkinson's disease in running speech spoken in three different languages. *Journal of the Acoustic Society of America*, **139**, 481–500.

Surrogate-based visualization of the influence of geometric design on the performance of a coronary stent

N.S. Ribeiro[a], J.O. Folgado[a], H.C. Rodrigues[a]

[a]*IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Portugal*

Poster Session, Location: North Tower, first floor *Thursday 13th, 17:20 – 18:40*

Coronary artery disease is one of the most common causes of death in developed countries and is characterized by the narrowing of coronary arteries caused by atherosclerosis which may lead to a heart attack. A standard treatment is the deployment of a coronary stent, a medical device that keeps the artery open so that blood flow can be restored to appropriate levels. Despite its widespread use, the long-term performance of these devices can be compromised by complications such as in-stent restenosis and stent thrombosis which are strongly influenced by the geometric design.

The performance of a stent is determined by several functional attributes such as the elastic recoil, flexibility, radial strength and fatigue resistance. Other attributes like vascular trauma, dogboning, longitudinal strength and tissue prolapse have an impact on the aforementioned post-stenting related adverse events. All these characteristics can be computationally investigated through properly set up virtual tests which capture the actual physical behaviour of stents. Moreover, all mentioned attributes are dependent on the geometric features of the stent configuration. Therefore, through computational models, one is able to estimate how changes in stent geometry influence stent performance and to identify designs with improved clinical performance.

The computational time and cost required to evaluate the physics-based numerical models is high, compromising the application of methods that demand a high number of evaluations such as visualization. One approach to reduce the computational burden is to use surrogate models that emulate the expensive computer simulation, such as Gaussian process regression models [1]. Once a surrogate model is built, applications like design space visualization become feasible in a reasonable time span.

In this work, the Efficient Global Optimization algorithm [2] was employed to obtain accurate surrogate models of the functional attributes that characterize stent performance. The inputs for these are variables controlling the geometry of the stent configuration. The surrogate models were used to obtain a multidimensional visualization of the response surfaces allowing us to understand the complex interplay between the design variables and the stent functional attributes.

Keywords: visualization, Gaussian process modelling, coronary stents.

References

[1] J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Henry (1989). Design and analysis of computer experiments. *Statistical Science*, 409–423.

[2] D.R. Jones, M. Schonlau, and W.J. Welch (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, **13**(4), 455–492.

A robust DF-REML framework for variance components estimation in genetic studies

V.M. Lourenço[a], P.C. Rodrigues[bc], A.M. Pires[d], H.-P. Piepho[e]

[a]CMA and Department of Mathematics, FCT – NOVA University of Lisbon, Portugal; [b]Department of Statistics, Federal University of Bahia, Brazil; [c]CAST, University of Tampere, Finland; [d]CEMAT and Department of Mathematics, IST – University of Lisbon, Portugal; [e]Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Germany

Poster Session, Location: North Tower, first floor *Thursday 13th, 17:20 – 18:40*

In genetic association studies, linear mixed models (LMMs) are used to test for associations between phenotypes and candidate single nucleotide polymorphisms (SNPs). These same models are also used to estimate heritability, which is central not only to evolutionary biology but also to the prediction of the response to selection in

plant and animal breeding, as well as the prediction of disease risk in humans. However, when one or more of the underlying assumptions are violated, the estimation of variance components may be compromised and therefore so may the estimates of heritability and any other functions of these. Considering that datasets obtained from real life experiments are prone to several sources of contamination, which usually induce the violation of the assumption of the normality of the errors, a robust derivative-free restricted-maximum-likelihood framework (DF-REML) together with a robust coefficient of determination are proposed for the LMM in the context of genetic association studies of continuous traits. The proposed approach, in addition to the robust estimation of variance components and robust computation of the coefficient of determination, allows in particular for the robust estimation of SNP-based heritability by reducing the bias and increasing the precision of its estimates. The performance of both classical and robust DF-REML approaches is compared via a Monte Carlo simulation study and examples of application of the methodologies to real datasets are given in order to validate the usefulness of the proposed robust approach.

Keywords: linear mixed model, SNP markers, heritability.

References

[1] E. Demidenko (2013). *Mixed Models: Theory and Applications with R*, 2nd Edition. John Wiley & Sons, Inc., Hoboken, New Jersey.

[2] K. Meyer (1989). Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative free algorithm. *Genet. Select. Evol.*, **21**, 317–340.

[3] D. Speed, G. Hemani, M.R. Johnson, and D.J. Balding (2012). Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.*, **91**, 1011–1021.

[4] J. Yu, G. Pressoir, W.H. Briggs, et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, **38**, 203–208.

A correction approach for random forest under sample selection bias

N. Krautenbacher[ab], F.J. Theis[ab], C. Fuchs[ab]

[a]Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Germany; [b]Department of Mathematics, Technische Universität München, Germany

Poster Session, Location: North Tower, first floor *Thursday 13th, 17:20 – 18:40*

Samples taken via complex survey design can lead to sample selection bias and can distort predictions when applying classifiers on unbiased data. Several methods correct for the sample selection bias, but their performance remains unclear especially for machine learning classifiers. We aim to assess which corrections to perform in which setting and to obtain methods suitable for machine learning techniques, especially the random forest. We propose two new resampling-based methods to resemble the original data and covariance structure: stochastic inverse-probability oversampling and parametric inverse-probability bagging. We compare all techniques for the random forest and other classifiers, both theoretically and on simulated and real data. Empirical results show that the random forest profits from only the parametric inverse-probability bagging proposed by us. For other classifiers, correction is mostly advantageous, and methods perform uniformly. We provide guidance for choosing correction methods when training classifiers on biased samples. For random forests, our method outperforms state-of-the-art procedures if distribution assumptions are roughly fulfilled.

Keywords: selection probabilities, sample selection bias, machine learning.

Measuring trends in depression symptoms in Dutch veterans from before, until five years after deployment

R. Gorter[ab], E. Geuze[ab]

[a]*University Medical Center Utrecht, The Netherlands;* [b]*Research Center – Military Mental Healthcare, Dutch Ministry of Defence*

Poster Session, Location: North Tower, first floor *Thursday 13th, 17:20 – 18:40*

The development of depression symptoms 5 years after deployment to Afghanistan was investigated using data from a longitudinal prospective cohort study with Dutch veterans. The questionnaire that was used to administer depression symptoms at the 5 years measurement contained a subset of the items used in the first five time points that took place 1 month before, 1 month, 6 months, 1 year, and 2 years after deployment. In order to make meaningful inferences on the development of the latent variable depression over time, we need to make sure that the construct measurements represent the same trait at all six measurement occasions (i.e. the assumption of measurement invariance (MI) must be met). The depression scale shows MI across time points when veterans with a changed level of depression at two time points have the same expected raw-score on the items of the questionnaire. Equivalently, if veterans with an unchanged level of depression have different expected raw-scores for one or more items of the questionnaire, the MI assumption is not met. With the lack of MI, the observed differences in mean scores can be either due to the change in depression or due to a different relation between the underlying constructs and the observed item scores. After measurement invariance is established, the development of depression symptoms over time can be investigated.

In the current study, MI was investigated using Bayesian SEM with small informative priors on the item cross loadings. The detected non-invariant factor loadings and thresholds were freely estimated in the next step while generating plausible values (PVs) for the latent variables. In the final step, a polynomial curve was estimated to model the development of depression symptoms.

Non-invariance was detected in several item factor loadings as well as thresholds. There were two items showing MI over all measurement occasions. Scale scores where estimated while freeing the non-invariant parameters using PV technology. The result from the growth curve model showed a significant positive quadratic trend in depression scores over time. Five years after deployment, Dutch veterans showed an increase in self-reported depression symptoms.

Keywords: measurement invariance, plausible values, questionnaire data.

Cluster-based lag optimization of physiological noise models in high-field resting-state fMRI

J. Pinto[a], S. Nunes[a], M. Bianciardi[b], L.M. Silveira[c], L.L. Wald[b], **P. Figueiredo**[a]

[a]ISR and Department of Bioengineering, Instituto Superior Técnico, Universidade de Lisboa, Portugal; [b]Department of Radiology, A.A. Martinos Center for Biomedical Imaging, MGH and Harvard Medical School, MA, United States; [c]INESC-ID and Department of Electrical and Computer Engineering, Instituto Superior Técnico, Universidade de Lisboa, Portugal

Poster Session, Location: North Tower, first floor *Thursday 13th, 17:20 – 18:40*

The last years have seen an increasing interest in the study of the brain's intrinsic functional connectivity, based on measurements performed during resting-state by functional magnetic resonance imaging (rs-fMRI). These connections are inferred from time synchronous fluctuations in signal across brain that are caused not only by neuronal activity but also by non-neuronal mechanisms, usually referred to as physiological noise. To date several strategies have been proposed to model and remove physiological noise from rs-fMRI data, particularly at high-field (7 Tesla), including contributions from respiratory volume rate (RV) and heart rate (HR) signal fluctuations. Recent studies suggest that these contributions are highly variable across subjects/brain and that physiological noise correction may thus benefit from more specific levels of optimization.

In this work, we propose a spatiotemporal clustering approach for the optimization of RV/HR models. This method is based on the k-means algorithm. Different values of k were tested $(2, 3, 4, 5, 6)$, the squared Euclidean distance was used as the distance metric and local minima were minimized by performing 10 replicates using different initial cluster centroid positions chosen at random. We systematically investigate the impact of the degree of spatial specificity, including the newly proposed cluster method, in the optimization of RV and HR models.

Voxelwise models explained more signal variance than less spatially specific models, as expected. However, the accuracy of functional connectivity strength measurements improved with the model spatial specificity up to a maximum at the cluster level, and subsequently decreased at the voxel level, suggesting that the latter incurs in over-fitting to local fluctuations with no physiological meaning. In conclusion, our results indicate that 7 Tesla rs-fMRI connectivity measurements improve if a cluster-based physiological noise correction approach is employed in order to take into account the individual spatial variability in the time-lags of HR and RV contributions.

Keywords: clustering, fMRI, modeling.

References

[1] R. Birn (2008). Influence of heart rate on the BOLD signal: the cardiac response function. *Neuroimage*, **44**, 857–69.

[2] C. Chang (2009). The respiration response function: the temporal dynamics of fMRI signal fluctuations related to changes in respiration. *Neuroimage*, **40**, 644–654.

Information visualisation quadrant display: a synergistic approach to a postgraduate program

M.C. Botelho[a], E. Vilar[b], E. Cardoso[c], **A. Alexandrino da Silva**[a], P.D. Almeida[b], L. Rodrigues[d], A.P. Martinho[a], S. Rodrigues[b]

[a]*Instituto Universitário de Lisboa (ISCTE-IUL), CIES-IUL;* [b]*Faculdade de Belas Artes, Universidade de Lisboa;* [c]*Instituto Universitário de Lisboa (ISCTE-IUL), INESC-ID;* [d]*Universidade dos Açores, IMAR – Instituto do Mar*

Poster Session, Location: North Tower, first floor *Thursday 13th, 17:20– 18:40*

Information visualization (Infovis) has become a privileged medium of mass visual communication [1]. The multidisciplinary nature of information visualization is today fairly consensual in both professional and academic communities. It is organized around four articulated areas — information design, data visualization, visual analytics, and data journalism — that operate within the information visualization domain. Infovis is increasingly becoming an independent research field with a specific research agenda [2], seeking to provide people with better and more effective ways to understand and analyse datasets.

This multidisciplinary approach and the profuse discussions about the teaching methods for Information Visualization are the foundation of a postgraduate program, which brings together two universities and three schools (in the areas of Social Studies, Design and Technologies).

This poster aims to contribute to a much needed and current debate on Information Visualization, considering four main areas. It also contributes to an innovative quadrant display, since it summons and discusses the multiple viewpoints of the four areas as well as the bridges between them. The connections between the four main areas are visually represented through a metaphor referring to the children’s game “the cootie catcher”. Each time a triangle is opened, an area’s perspective is revealed, as is its interaction with the others. These interactions or bilateral synergies enable deeper reflections on which contents should be present for the different curricular units.

It is concluded that, even if there are specificities in each area, a common language may be adopted and synergies may be generated. The debate on the matter will continue and new challenges will certainly arise upon the implementation of the postgraduate program.

Keywords: information visualization, data journalism, visual analytics.

References

[1] P.A. Hall (2011). Bubbles, lines and string: How visualization shapes society. In A. Blauvelt, E. Lupton, et al (eds.), *Graphic Design: Now in Production*. Walker Art Center, 170–185.

[2] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmannn (2010). *Mastering the Information Age: Solving Problems with Visual Analytics*. Eurographics Association.

An automatic pre-processing pipeline for EEG analysis based on robust statistics

J.R. da Cruz[ab], M.H. Herzog[b], P. Figueiredo[a]

[a]*Institute for Systems and Robotics (LARSyS) and Department of Bioengineering, Instituto Superior Técnico, Universidade de Lisboa, Portugal;* [b]*Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne, Switzerland*

Poster Session, Location: North Tower, first floor *Thursday 13th, 17:20 – 18:40*

The electroencephalogram (EEG) is a non-invasive tool commonly used for the investigation of the human brain function. With the advent of high-density EEG arrays and studies of large populations, conventional supervised methods for artifact rejection have become excessively time consuming. Here, we propose a novel automatic pipeline for the pre-processing of EEG data (App) based on state-of-the-art guidelines and robust statistics. App consists of: 1) high-pass filtering; 2) power-line noise removal; 3) re-referencing to a robust estimate of the mean of all channels; 4) removal and interpolation of bad channels; 5) removal of bad epochs; 6) independent component analysis (ICA) to remove eye-movement, muscular and bad-channel related artifacts; 7) removal of epoch artifacts. App was tested on event-related potential (ERP) data from both healthy and schizophrenia patients performing a visual task and eyes-closed resting-state (RS) data from healthy participants. The results were compared with the ones obtained using previously reported automatic methods (FASTER, TAPEEG and PREP) as well as supervised artifact detection. In general, App rejected a similar number of bad channels relative to the supervised method and fewer than the alternative automatic methods. In the ERP study, the proposed pipeline produced significantly higher amplitudes than FASTER, while no difference was found relative to the supervised scheme. In the RS study, the power

across different frequency bands obtained using App were found to correlate with TAPEEG, PREP and the supervised scheme. In conclusion, App effectively removed EEG artifacts, performing similarly to the supervised scheme and outperforming existing automatic alternatives.

Adaptive learning and learning science in a first course in university statistics

D.M. Garvis

Washington and Lee University, Lexington, Virginia, USA

Poster Session, Location: North Tower, first floor *Thursday 13th, 17:20 – 18:40*

Adaptive learning courseware originally developed in the Online Learning Initiative (OLI) at Carnegie Mellon University has been used in several disciplines in many U.S. colleges and universities. Original development of OLI courseware for instruction in a first course in Statistics included experienced course instructors as well as web designers and managers, software engineers, and learning science researchers. Accordingly, findings from learning science research were directly incorporated into the adaptive learning materials used for OLI-Statistics instruction.

Subsequently, OLI-Statistics has been used in first courses in Statistics using pure online, hybrid, blended, and supplemental models. Empirical research is consistent in showing that student outcomes in OLI-Statistics courses are comparable or better than in courses when the courseware is not used. In the most rigorous empirical research in this stream of work, double-blind, randomized experiments found that there was no significant difference between student assessment outcomes for US undergraduate public university students using the OLI courseware materials and traditional lecture pedagogies.

In this paper, data from a sample of undergraduate students in a small, selective private university in rural Virginia, USA is used to compare assessed outcomes in a first course in Statistics. Specifically, since 2014, versions of OLI-Statistics have been used for undergraduates taking an Applied Statistics course required for accounting, business, economics and politics majors. Assessment scores using the Comprehensive Assessment of Outcomes in a First Statistics course (CAOS Assessment) in courses taught using OLI-Statistics are compared to assessment outcomes for students nationwide. In addition, pedagogical advantages and teaching trade-offs from using the OLI-Statistics courseware, now also hosted by Stanford EdX, will be discussed.

Keywords: statistics, adaptive learning.

References

[1] D. Raths (2017). Scaling up with adaptive learning. In *Campus Technology*. `https://campustechnology.com/articles/2017/01/04/scaling-up-with-adaptive-learning.aspx`

[2] T. Walsh and W.G. Bowen (2010). *Unlocking the Gates.* Princeton University Press.

[3] M. Lovett, O. Meyer, and C. Thille (2008). The Open Learning Initiative: measuring the effectiveness of the OLI statistics course in accelerating student learning. *Journal of Interactive Media in Education*, 2008.1.

[4] W.G. Bowen, M.M. Chingos, K.A. Lack, and T.I. Nygren (2014). Interactive learning online at public universities: evidence from a six-campus randomized trial. *Journal of Policy Analysis and Management*, **33**(1), 94–111.

[5] R. delMas, J. Garfield, A. Ooms, and B. Chance (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, **6**(2), 28–58.

[6] C. Chung (2016). 5 higher-ed innovators share challenges, ideas for the future of digital learning. In *EdSurge.* `https://www.edsurge.com/news/2016-08-23-5-higher-ed-innovators-share-challenges-ideas-for-the-future-of-digital-learning`

A novel technique of symbolic time series representation aimed at time series clustering

J. Korzeniewski

University of Lodz

Poster Session, Location: North Tower, first floor *Thursday 13th, 17:20– 18:40*

The article proposes a novel technique which may be useful in time series data analysis. The technique is basically a part of symbolic time series analysis which has become very popular in recent years and which can be very successful in spite of the weakening of the measurement scale. When the time series is represented by a much smaller number of symbols in comparison with the original number of observations, it becomes reasonable to investigate the correlation among the symbols across all series. We try to identify more important symbols or segments of the time series which can significantly improve the starting point for different challenges connected with time series clustering. The main idea of the technique is to analyse the correlation coefficient between the distances (which can be found for any measurement scale) between objects measured on different sets of attributes. In such a setup, the attributes comprise the time series segments described in terms of symbols. We investigate the efficiency of the new technique in connection with other methods aimed at efficient time series representation such as PAA based methods. The technique is evaluated on synthetic data sets as well as stock exchange time series data. The results are promising.

Keywords: time series, clustering, correlation.

Introducing formative assessment in probability and statistics course – analysis of the first data

A.C. Finamore, **A. Moura Santos**, A. Pacheco

CEMAT, CEAFEL, and Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Portugal

Poster Session, Location: North Tower, first floor *Thursday 13th, 17:20 – 18:40*

Probability and Statistics is a course in Mathematics offered every semester to about 1,500 STEM students of the Instituto Superior Técnico of Lisbon from the Departments of Science and Engineering. This course has had one of the smallest IST approval rates and, at the same time, one of the largest dropout rates for several consecutive years. Facing the challenge to transform the reality of this course, the Department of Mathematics jointly with the Pedagogical Council proposed a new approach: combining two different types of assessment, summative and formative.

Until the first semester of 2016/2017, only the summative assessment had been used, in the form of two tests (midterm and final) and one exam. In this type of assessment, teachers can only verify the students' difficulties after the teaching and learning process had occurred [1]. In the first semester of 2016/2017, the formative assessment was also introduced using Online Electronic Quizzes with Random Parameters [2]. Throughout the semester, six non compulsory quizzes were introduced aiming to identify specific topics that the students were struggling with. This type of assessment promotes the alignment of the learning and teaching strategies used both by the students and teachers during the course.

By implementing the quizzes through the IST Learning Management System, we captured in only one semester 1,078 footprints of the students from 18 different degrees through the raw collected data. Based on the first results of this pedagogical strategy and conducting different types of statistical analysis on this data, our aim is to present our findings together with interesting insights and conclusions, providing academic coordinators the necessary support for making future decisions [3].

Keywords: online electronic quizzes, STEM education, data mining.

References

[1] N. Glazer (2014). Formative plus summative assessment in large undergraduate courses: why both? *International Journal of Teaching and Learning in Higher Education*, **26**(2), 276–286.

[2] A.C. Finamore, A. Moura Santos, and P. Ribeiro (2016). Fostering STEM formative assessment for lifelong learners. *ICERI 2016 Proceedings*. ISBN: 978-84-617-5895-1, doi:10.21125/iceri.2016.1031

[3] W. Greller and H. Drachsler (2012). Translating learning into numbers: a generic framework for learning analytics. *Educational Technology and Society*, **15**(3), 42–57.

Friday, 14 July 2017

Selection and clustering of correlated variables

M.A.T. Figueiredo

Instituto de Telecomunicações and Instituto Superior Técnico, Universidade de Lisboa, Portugal

Invited Lecture Session, Room: Abreu Faro *Friday 14th, 9:00 – 9:45*

In high-dimensional linear regression (and other supervised learning) problems, highly correlated variables/covariates create a challenge to variable selection procedures. In those scenarios, standard sparsity-inducing regularization (namely ℓ_1 regularization, also known as LASSO–*least absolute shrinkage and selection operator*) may be inadequate, as it leads to the selection of arbitrary convex combinations of those variables, maybe even of an arbitrary subset thereof. However, particularly in scientific contexts, it is often important to explicitly identify all the relevant covariates, as well as explicitly identify groups/clusters thereof. This talk addresses the recently introduced *ordered weighted* ℓ_1 (OWL) regularizer, which has been proposed for this purpose. We review several optimization aspects concerning this regularizer, namely computational methods to efficiently solve the corresponding regularized regression problems. In the analysis front, we give sufficient conditions for exact feature clustering (under squared error, absolute error, and logistic losses) and characterize its statistical performance.

Keywords: linear regression, logistic regression, variable selection, variable clustering.

Detecting anomalous data cells

P.J. Rousseeuw[a], W. Van den Bossche[a]

[a]*KU Leuven, Belgium*

Invited Lecture Session, Room: Abreu Faro *Friday 14th, 9:45 – 10:30*

A multivariate dataset consists of n cases in d dimensions, and is often stored in an n by d data matrix. It is well known that real data may contain outliers. Depending on the situation, outliers may be (a) undesirable errors which can adversely affect the data analysis, or (b) valuable nuggets of unexpected information. In statistics and data science, the word outlier usually refers to a row of the data matrix, and the methods to detect such outliers only work when at least half the rows are clean, see e.g. [3]. But often many rows have a few contaminated cell values, especially in high-dimensional data, which implies few rows are entirely clean [2].

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

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

We illustrate the method on several real data sets, where it uncovers more structure than found by purely columnwise methods or purely rowwise methods. The proposed method can help to diagnose why a certain row is outlying, e.g. in process control. It may also serve as an initial step for estimating multivariate location and scatter matrices, as in [1].

Keywords: algorithms, data science, outliers.

References

[1] C. Agostinelli, A. Leung, V.J. Yohai, and R.H. Zamar (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test*, **24**(3), 441–461.

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

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

Performance of marketing attribution models

S. Sapp

Google Inc., Mountain View, CA, USA

Invited Lecture Session, Room: Abreu Faro *Friday 14th, 11:00 – 11:45*

Attribution models allocate credit to marketing channels for cross-channel marketing campaigns. We present a process for evaluating the efficacy of attribution models using simulation. The proposed simulations generate user-level activity streams using a non-stationary Markov model. The transition matrix in this model is modified by the appearance of advertising impressions, which are injected into the activity stream with a specified probability. By increasing or decreasing this probability of an ad impression, it is possible to run virtual experiments to measure ad effectiveness. The results of these experiments are used to evaluate a set of commonly used attribution models.

The role of visualization in data science

D.A. Keim

Computer and Information Science, University of Konstanz

Keynote Lecture Session, Room: Abreu Faro *Friday 14th, 11:45 – 12:45*

Never before in history has data been generated and collected at such high volumes as it is today. Data Science tries to generate insights into the data but for this to be effective, it needs to include the human in the data exploration process and combine the flexibility, creativity, and general knowledge of the human with the enormous storage capacity and the computational power of today's computers. Visualizations may help in all phases of the data science process: Presenting data in an interactive, graphical form helps to get an initial understanding of the data and to form hypotheses about the data, but it may also significantly contribute to the actual knowledge discovery by guiding the analysis using visual feedback. Visual data exploration often complements the statistical analysis since it helps to identify properties that are hard to detect by statistical methods and it encourages the formation and validation of new hypotheses for better problem-solving and gaining deeper domain knowledge.

In putting visualization to work in data science, it is not obvious what can be done by automated statistical analyses and what should be done by interactive visual methods. In dealing with massive data, the use of automated methods is mandatory — and for some problems it may be sufficient to only use fully automated analysis methods — but there is also a wide range of problems where the use of interactive visual methods is crucial. Examples from a number of application areas illustrate the benefits of visualization methods in data science.

Keywords: data visualization, visual analytics.

Index