ZQS

13th

Workshop on Stochastic Models, Statistics and Their Applications

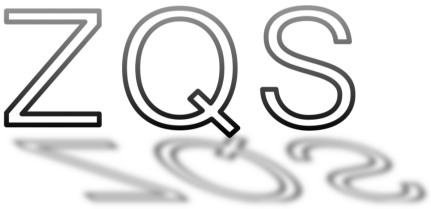
2017

February 20th to 24th, 2017
Humboldt-Universität zu Berlin
Erwin-Schrödinger Zentrum

SMSA 2017 is organized by:
Department of Mathematics, HU Berlin
Department of Computer Science, HU Berlin
Department of Mathematics, RWTH Aachen
AG Stochastische Modelle für Zuverlässigkeit, Qualität und Sicherheit e.V. (ZQS)

SMSA 2017 is technically supported by:
Computer and Media Service (CMS), HU Berlin

Our Sponsors:



Lehmanns Media GmbH
Rudower Chaussee 26
D-12489 Berlin

Printed by the in-house printing service of Humboldt-Universität zu Berlin

13th Workshop on Stochastic Models, Statistics and Their Applications

Humboldt-Universität zu Berlin
Erwin-Schrödinger-Zentrum
February 20th to February 24th, 2017

Editors: Ansgar Steland
Wolfgang Kössler

Printed by the in-house printing service of Humboldt-Universität zu Berlin

Conference logo designed by Ansgar Steland
Main cover, back cover and general layout: Frank Fuhlbrück
The back cover image can be generated by the following R program:

```
require(MASS)
many=13
set.seed(132017)
x=rchisq(many,13)
x=rep(x,times=many)
y=rchisq(many,2017)
y=rep(y,rep.int(many,many))
persp(kde2d(x,y,n=50),col=sample(rainbow(many^2)),
  theta=45,phi=45,box=F,lwd=0.5,
  border=rgb(33,65,202,maxColorValue=255))
```

Berlin, 2017

Preface

Following the successful previous workshops, the conference will put together recent advances and trends in areas related to stochastic modeling, statistical inference and their applications. It aims to bring together researchers working on methodologies and applications.
As usual, contributions motivated by or addressing issues in engineering, industry and natural sciences are particularly welcomed. Several invited sessions are devoted to these topics and we are very happy that two sessions are organized in cooperation with two other scientific organizations: With ENBIS (European Network for Business and Industrial Statistics) a session about industrial and business statistics has been organized and with DStatG (German Statistical Society) a session about statistics in technology, reliability and natural sciences.
The invited plenary presentations are held by Holger Dette (Ruhr-Universität Bochum), Allan Gut (Uppsala University) and Jana Jurečková (Charles University of Prague). Many renowned researchers have agreed to give invited presentations making this workshop an international one, which attracts more than 150 participants.
The 2017 workshop is organized by the Department of Computer Sciences and the Department of Mathematics of the Humboldt University of Berlin, in cooperation with the Institute of Statistics of the Department of Mathematics, RWTH Aachen University. It is supported by the President and the Faculty of Natural Sciences of Humboldt University, the Institute of Statistics of RWTH Aachen University, The German Statistical Society, The European Network for Business and Industrial Statistics, SAS Institute GmbH and Lehmanns Media.
It is our intention and hope that the workshop deepens existing cooperations and partnerships as well as stimulates new collaborations.

We hope you enjoy your stay and the talks.

Aachen and Berlin — Ansgar Steland
February 2017 — Wolfgang Kössler

Acknowledgements

We would like to thank the following colleagues for organizing invited and contributed sessions:
Edgar Brunner (Göttingen),
Steve Coad (London),
Thorsten Dickhaus (Bremen),
Maik Döring (Hohenheim),
Bernd Droge (Berlin),
Matthias Eckardt (Berlin),
Rainer Göb (Würzburg),
Piotr Jaworski (Warsaw),
Adam Krzyżak (Montreal),
Eckhard Liebscher (Merseburg),
Alexander Meister (Rostock),
Mirek Pawlak (Winnipeg),
Markus Reiß (Berlin),
Dana Simian (Sibiu),
Ewa Skubalska-Rafajłowicz (Wrocław),
Wolfgang Schmid (Frankfurt/Oder),
Ansgar Steland (Aachen),
Wolfgang Stummer (Erlangen-Nürnberg),
Krzysztof Szajowski (Wrocław),
Christian Weiß (Hamburg),
Florian Ziel (Frankfurt/Oder), and
Silvelyn Zwanzig (Uppsala).

We are also grateful to our institutions, RWTH Aachen, and Humboldt Universität zu Berlin, especially to the Faculty of Sciences and the Computer and Media Service for giving support.
Special thanks go our sponsors, especially SAS Institute GmbH who give considerable support to our GetTogether and to the Conference Dinner.
Furthermore, we are very grateful to Wolf F. Lesener who organized many financial issues and the social events, and to Frank Fuhlbrück who managed our website and the preparation of the abstract volume.

Aachen and Berlin — Ansgar Steland
February 2017 — Wolfgang Kössler

Program Comittee

Ansgar Steland (Chair, RWTH Aachen University)
Wolfgang Kössler (Humboldt-Universität zu Berlin)
Marco Burkschat (RWTH Aachen University)
Uwe Jensen (University of Hohenheim)
Sven Knoth (Helmut-Schmidt-Universität Hamburg)
Waltraud Kahle (Otto-von-Guericke-Universität Magdeburg)

Local Organizers

Markus Reiß (Chair)
Wolfgang Kössler
Wolf F. Lesener
Frank Fuhlbrück

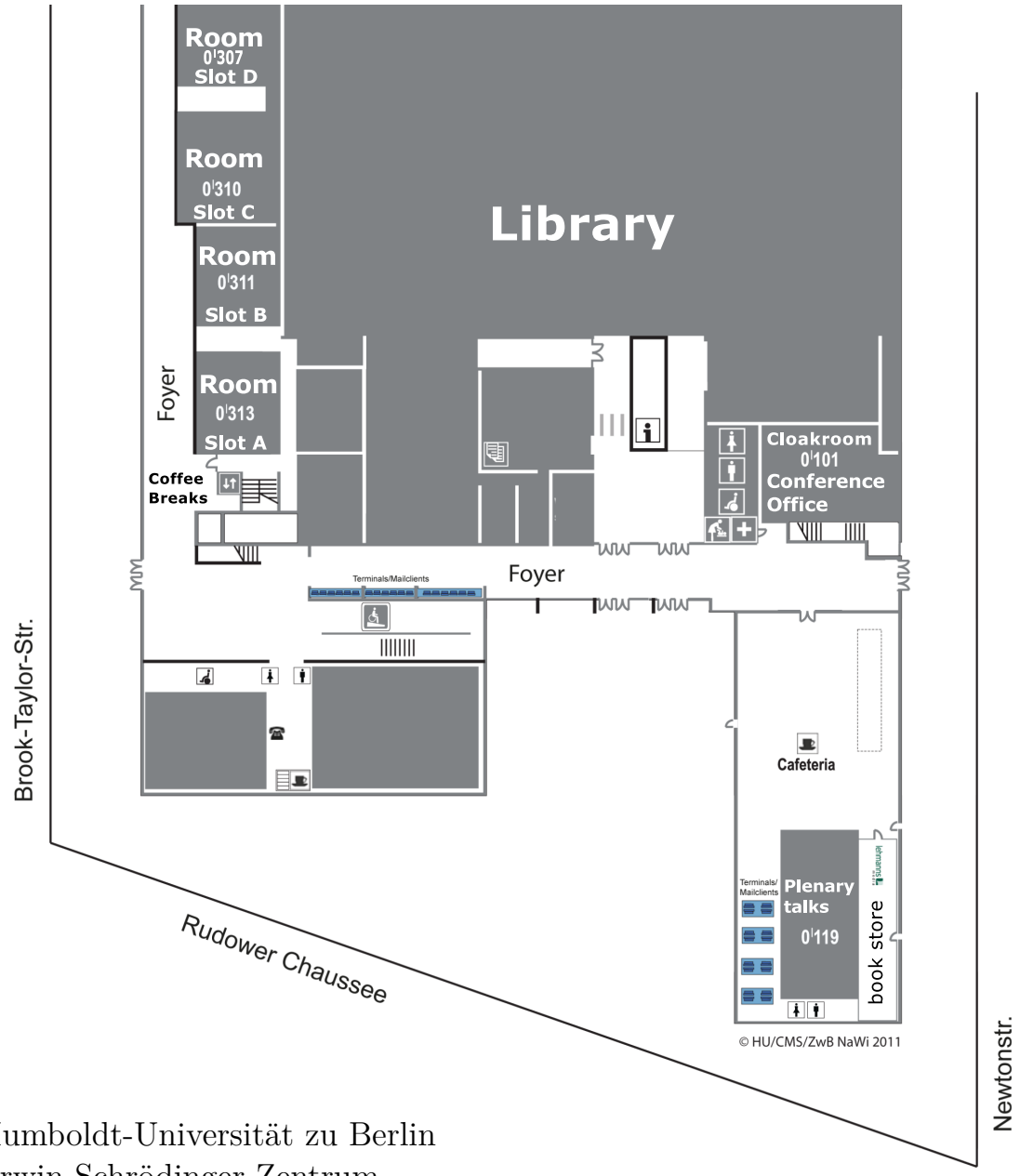
Contact and Important Addresses

Conference office is located in room 0'101 of Erwin-Schrödiger-Zentrum and includes the **cloakroom**.

Opening hours:

Conference Office	
Monday 20.2.	12:00 - 17:00
Tuesday 21.2.	08:30 - 17:00
Wednesday 22.2.	08:30 - 17:00
Thursday 23.2.	08:30 - 15:00
Friday 24.2.	08:30 - 13:00

Cloakroom	
Monday 20.2.	12:00 - 19:00
Tuesday 21.2.	08:30 - 18:30
Wednesday 22.2.	08:30 - 18:30
Thursday 23.2.	08:30 - 15:30
Friday 24.2.	08:30 - 15:00



Humboldt-Universität zu Berlin
Erwin-Schrödinger-Zentrum
Rudower Chaussee 26
12489 Berlin

Wifi and computer access

Participants coming from institutions that are part of the **eduroam** network should be able to use the mobile devices in the same way as usually. There is an additional SSID **eduroam_5GHz** which works with the same authentication mechanism as **eduroam** but uses 5 Ghz band instead of the heavily used 2.4 GHz band.
If your home institution is not part of eduroam, you may ask for an account for the wifi network with SSID **HU-Meeting** in the conference office. This network is unencrypted and asks for an account name and password when you request an an URL with your browser for the first time. You should take into account the following remarks:

- You cannot send e-mail via SMTP (i.e. with your dedicated e-mail client, like Thunderbird, Outlook etc.) unless you either use a SMTP server of HU Berlin or a VPN connection to your home institution.
- Everyone can read your data send/received via **HU-Meeting** unless you only use encrypted connections (e.g. https) or better a VPN.
- Transfer rate for all HU-Meeting users is limited, so please do not use it for private activities like video streaming.

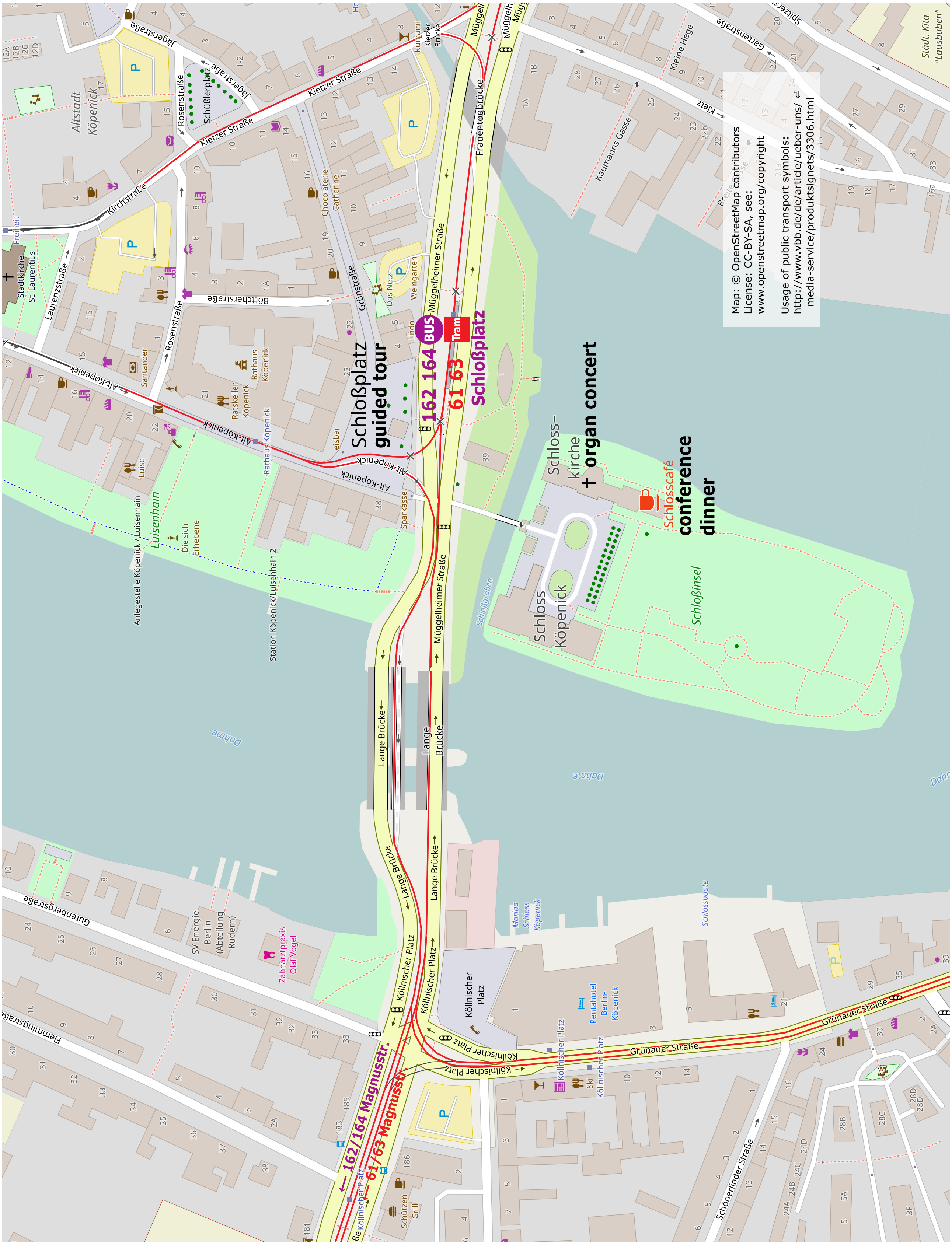
Access to PCs is also possible using the accounts for HU-Meeting. PCs (only Windows) can be found in the library or in the foyer.

Social activities

In your conference portfolio, you will find vouchers for the social activities you have selected. Please bring the respective voucher to each activity (even for those included in the general fee).
On Tuesday/Wednesday each aeronautic tour is about one and a half hour long and starts 18:00 in front of the conference office. Be sure to join the language group you have selected during registration (also noted on the voucher).
Thursday afternoon offers most of the social events:

- The meeting point for the **city tour** in Köpenick is front of the conference office at **15:00**. We will go by bus/tram to Schloßplatz Köpenick (place near the castle) which can be reached via tram 61 and 63 and bus 162 and 164.
- **The organ concert** starts immediately after the city tour (**17:00**). You will listen to an one-hour concert celebrating Händel's 332nd birthday in the church belonging to Schloss Köpenick.
- Finally, **the conference dinner** will take place in the castle's café (Schloss-café Köpenick). It lasts from **18:00 to 22:00**.

Please see the next page for a map of the area around Schloss Köpenick.


Schloßplatz
guided tour
BUS
162 164
Tram
61 63
Schloßplatz
Schloss-
kirche
+ organ concert
Schlosscafe
conference
dinner
Schloss
Köpenick
Schloßinsel
← 162/164 Magnusstr.
← 61/63 Magnusstr.
Altstadt
Köpenick
Luisenhain
Lange Brücke
Müggelheimer Straße
Alt-Köpenick
Kietzer Straße
Grünstraße
Böttcherstraße
Rosenstraße
Frauentogbrücke
Köllnischer Platz
Grünauer Straße
Gutenbergstraße
Dahme
Map: © OpenStreetMap contributors
License: CC-BY-SA, see:
www.openstreetmap.org/copyright
Usage of public transport symbols:
http://www.vbb.de/de/article/ueber-uns/
media-service/produktsignets/3306.html

Time Table

Monday 20.02.	Tuesday 21.02.	Wednesday 22.02.	Thursday 23.02.	Friday 24.02.
	09:00 - 11.00 Slot 3a-3d	09:00 - 11.00 Slot 6a-6c	09:00 - 11.00 Slot 9a-9c	09:00 - 11.00 Slot 11a-11c
	11:30 - 12:30 Plenary Talk J. Jurečková	11:30 - 12:30 Plenary Talk A. Gut	11:30 - 13:30 Slot 10a-10c	11:30 - 13:30 Slot 12a-12b
	12:30 Lunch	12:30 Lunch		
13:00 Opening H.Dette	13:30 - 15:30 Slot 4a-4d	13:30 - 15:30 Slot 7a-7d	13:30 Lunch	13:30 Lunch
14:30 - 16:30 Slot 1a-1c	16.00 - 18.00 Slot 5a-5c	16.00 - 18.00 Slot 8a-8c	15:00 City Tour Köpenick	
17:00 - 18:30 Slot 2a-2b	18.00 - 19.00 Aereonautic tour 1-2	18.00 - 19.00 Aereonautic tour 3-4	17:00 Organ Concert	
18.30 - 21.30 Get Together	18.00 - 19.00 ZQS Meeting		18.00 - 22.00 Conference Dinner	

Please bring a USB thumb drive and transfer your presentation (PPT or PDF) to the PC just before the session of your talk begins. You can also use your own laptop, tablet etc. if it has either a USB or VGA connector, but please test switching between the local PC and your device in advance to avoid unnecessary delays between the talks.

Slot	Session
1a	Stochastic Models in Technology, Reliability, and Quality 1
1b	Nonparametric Regression and Density Estimation 1
1c	Panel Data
2a	Analysis, Testing and Change Detection in High Dimensions 1
2b	Multivariate Distributions and Copula 1
3a	Statistics of Stochastic Processes 1
3b	Discrete-Valued Time Series 1
3c	Analysis, Testing and Change Detection in High Dimensions 2
3d	Survival Analysis
4a	Statistics of Stochastic Processes 2
4b	Discrete-Valued Time Series 2
4c	Nonparametric Methods 1
4d	High-Dimensional Problems in Engineering
5a	Statistics of Stochastic Processes 3
5b	Discrete-Valued Time Series 3
5c	Analysis, Testing and Change Detection in High Dimensions 3
6a	Statistics of Stochastic Processes 4
6b	Machine Learning 1
6c	Nonparametric and Semiparametric Testing
7a	Stochastic Models in Technology, Reliability, and Quality 2
7b	Machine Learning 2
7c	ENBIS
7d	Nonparametric Methods 2
8a	Stochastic Models in Technology, Reliability, and Quality 3
8b	Nonparametric Regression and Density Estimation 2
8c	Sequential Experimental Design
9a	Analysis, Testing and Change Detection in High Dimensions 4
9b	Simultaneous Statistical Inference
9c	Multivariate Distributions and Copula 2
10a	Nonparametric Methods 3
10b	Optimal Decision in Changepoint Models
10c	Functional Data Analysis
11a	Energy Statistics
11b	Errors-in-Variables Models
11c	Miscellaneous
12a	Graphical Models and Network Analysis
12b	Distance-based Statistical Methods

Detailed Program

Program

Plenary talk

Mon 13:00 - 14:00 in room 0'119 organized by *Ansgar Steland*

Holger Dette :
Statistical methodology for comparing curves **28**

Plenary talk

Tue 11:30 - 12:30 in room 0'119 organized by *Ansgar Steland*

Jana Jurečková :
Empirical Regression Quantile Process in Analysis of Risk **29**

Plenary talk

Wed 11:30 - 12:30 in room 0'119 organized by *Ansgar Steland*

Allan Gut :
Aspects on the Law of Large Numbers **32**

Stochastic Models in Technology, Reliability, and Quality 1

Slot 1a Mon 14:30 - 16:30 in room 0'313 organized by *Ansgar Steland*

Sonja Kuhnt **(invited talk)** :
Sensitivity analysis and optimization of computer experiments with an application to a centrifugal compressor impeller **33**

Marco Burkschat **(invited talk)** :
Stochastic comparisons of systems based on sequential order statistics **34**

Waltraud Kahle **(invited talk)** :
Incomplete Repair in Degradation Processes: A Kijima-Type Approach **36**

Hans Manner **(invited talk)** :
Modeling and forecasting multivariate electricity price spikes **38**

Nonparametric Regression and Density Estimation 1

Slot 1b Mon 14:30 - 16:30 in room 0'311 organized by *Adam Krzyżak*

Harro Walk **(invited talk)** :
Asymptotic normality of a nearest neighbor estimate of the second moment regression functional with an application to dimension reduction **39**

Adam Krzyżak **(invited talk)** :
Nonparametric regression estimation using hierarchical interaction models **41**

Benedikt Bauer **(invited talk)** :
On estimation of surrogate models for high-dimensional computer experiments **42**

Panel Data

Slot 1c Mon 14:30 - 16:30 in room 0'310 organized by *Bernd Droge*

Analysis, Testing and Change Detection in High Dimensions 1

Slot 2a Mon 17:00 - 18:30 in room 0'313 organized by *Ansgar Steland*

Multivariate Distributions and Copula 1

Slot 2b Mon 17:00 - 18:30 in room 0'311 organized by *Eckhard Liebscher*

Statistics of Stochastic Processes 1

Slot 3a Tue 9:00 - 11:00 in room 0'313 organized by *Markus Reiß*

Mark Podolskij **(invited talk) :**
Statistical inference for the fractional stable motion **60**

Markus Bibinger **(invited talk) :**
The discontinuous leverage effect in contemporaneous price and volatility jumps **61**

Mathias Trabs **(invited talk) :**
Volatility estimation for stochastic PDEs using high-frequency observations **63**

Jakob Söhl **(invited talk) :**
Nonparametric Bernstein–von Mises theorems for discretely observed compound Poisson processes **65**

Discrete-Valued Time Series 1

Slot 3b Tue 9:00 - 11:00 in room 0'311 organized by *Christian Weiß*

Robert Jung **(invited talk, 60 minutes) :**
Model Validation and Diagnostics for Discrete-Valued Time Series **66**

Christian Weiß **(invited talk) :**
A test procedure on the compounding structure of CP-INARCH models **67**

Analysis, Testing and Change Detection in High Dimensions 2

Slot 3c Tue 9:00 - 11:00 in room 0'310 organized by *Ansgar Steland*

Boris Darkhovsky **(invited talk) :**
Model-free off-line change-points detection in vector time series via ϵ-complexity: theoretical results. **69**

Alexandra Piryatinska **(invited talk) :**
Model-free off-line change-points detection in vector time series via ϵ-complexity: simulation and application **71**

Ewaryst Rafajłowicz **(invited talk) :**
Rule-based method of spike detection and suppression **73**

Ewa Skubalska-Rafajłowicz **(invited talk) :**
Dimensionality reduction by random projection: to orthogonalize or not to orthogonalize ? **75**

Survival Analysis

Slot 3d Tue 9:00 - 11:00 in room 0'307 organized by *Maik Döring*

Statistics of Stochastic Processes 2

Slot 4a Tue 13:30 - 15:30 in room 0'313 organized by *Markus Reiß*

Discrete-Valued Time Series 2

Slot 4b Tue 13:30 - 15:30 in room 0'311 organized by *Christian Weiß*

Nonparametric Methods 1

Slot 4c Tue 13:30 - 15:30 in room 0'310 organized by *Edgar Brunner, Ansgar Steland, Wolfgang Kössler*
Chair: *Edgar Brunner*

High-Dimensional Problems in Engineering

Slot 4d Tue 13:30 - 15:30 in room 0'307 organized by *Ewa Skubalska-Rafajlowicz*

Statistics of Stochastic Processes 3

Slot 5a Tue 16:00 - 18:00 in room 0'313 organized by *Markus Reiß*

Discrete-Valued Time Series 3

Slot 5b Tue 16:00 - 18:00 in room 0'311 organized by *Christian Weiß*

Analysis, Testing and Change Detection in High Dimensions 3

Slot 5c Tue 16:00 - 18:00 in room organized by *Ansgar Steland*

Statistics of Stochastic Processes 4

Slot 6a Wed 9:00 - 11:00 in room 0'313 organized by *Markus Reiß*

Machine Learning 1

Slot 6b Wed 9:00 - 11:00 in room 0'311 organized by *Dana Simian*

Nonparametric and Semiparametric Testing

Slot 6c Wed 9:00 - 11:00 in room 0'310 organized by *Mirek Pawlak*

Stochastic Models in Technology, Reliability, and Quality 2

Slot 7a Wed 13:30 - 15:30 in room 0'313 organized by *Ansgar Steland, Wolfgang Schmid*
Chair: *Wolfgang Schmid*

Machine Learning 2

Slot 7b Wed 13:30 - 15:30 in room 0'311 organized by *Dana Simian*

ENBIS

Slot 7c Wed 13:30 - 15:30 in room 0'310 organized by *Rainer Göb*

Nonparametric Methods 2

Slot 7d Wed 13:30 - 15:30 in room 0'307 organized by *Edgar Brunner, Ansgar Steland, Wolfgang Kössler*
Chair: *Markus Pauly*

Stochastic Models in Technology, Reliability, and Quality 3

Slot 8a Wed 16:00 - 18:00 in room 0'313 organized by *Ansgar Steland*
Chair: *Guido Knapp*

Nonparametric Regression and Density Estimation 2

Slot 8b Wed 16:00 - 18:00 in room 0'311 organized by *Adam Krzyżak*

Sequential Experimental Design

Slot 8c Wed 16:00 - 18:00 in room 0'310 organized by *Steve Coad*

Analysis, Testing and Change Detection in High Dimensions 4

Slot 9a Thu 9:00 - 11:00 in room 0'313 organized by *Ansgar Steland*

Marie Hušková **(invited talk, 45 minutes) :**
Change point detection with multivariate observations based on characteristic functions **182**

Zuzana Prášková **(invited talk, 45 minutes) :**
Change point problem in dynamic panel data **183**

Martin Wendler **(contributed talk) :**
Bootstrap and Change-Point Detection in Functional Time Series and Random Fields **184**

Simultaneous Statistical Inference

Slot 9b Thu 9:00 - 11:00 in room 0'311 organized by *Thorsten Dickhaus*

Werner Brannath **(invited talk) :**
Simultaneous Confindence Intervals for Graphical Multiple Tests **185**

Sebastian Döhler **(invited talk) :**
A modified Benjamini-Hochberg procedure for discrete data **187**

Helmut Finner **(invited talk) :**
From higher criticism tests and local levels of GOF tests to confidence bounds for the proportion of true nulls **188**

Florian Frommlet **(invited talk) :**
Asymptotic Bayes Optimality under Sparsity Revisited **189**

Multivariate Distributions and Copula 2

Slot 9c Thu 9:00 - 11:00 in room 0'310 organized by *Eckhard Liebscher*
Chair: *Piotr Jaworski*

Anouar El Ghouch **(invited talk) :**
Semiparametric copula quantile regression for complete or censored data **191**

Wolf-Dieter Richter **(invited talk) :**
Star-shaped distributions. A survey of recent results. **192**

Klaus Müller **(invited talk) :**
Exact distributions of order statistics from continuous $l_{n,p}$-symmetric sample distributions **194**

Eckhard Liebscher **(invited talk) :**
Modeling and statistical inference of copulas based on Frank's family **196**

Nonparametric Methods 3

Slot 10a Thu 11:30 - 13:30 in room 0'313 organized by *Edgar Brunner, Ansgar Steland, Wolfgang Kössler*
Chair: *Edgar Brunner*

Optimal Decision in Changepoint Models

Slot 10b Thu 11:30 - 13:30 in room 0'311 organized by *Krzysztof Szajowski*

Functional Data Analysis

Slot 10c Thu 11:30 - 13:30 in room 0'310 organized by *Alexander Meister*

Energy Statistics

Slot 11a Fri 9:00 - 11:00 in room 0'313 organized by *Florian Ziel*

Errors-in-Variables Models

Slot 11b Fri 9:00 - 11:00 in room 0'311 organized by *Silvelyn Zwanzig*

Miscellaneous

Slot 11c Fri 9:00 - 11:00 in room 0'310
Chair: *Marco Burkschat*

Graphical Models and Network Analysis

Slot 12a Fri 11:30 - 13:30 in room 0'313 organized by *Matthias Eckardt*

Distance-based Statistical Methods

Slot 12b Fri 11:30 - 13:30 in room 0'311 organized by *Wolfgang Stummer*

Michel Broniatowski **(invited talk, 40 minutes) :**
Estimating divergences through weighted bootstrap **230**

Wolfgang Stummer **(invited talk, 40 minutes) :**
A New Method of Robust Statistics **231**

Anna-Lena Kißlinger **(invited talk, 40 minutes) :**
A New Information-Geometric Method of Change Detection **232**

Abstracts

Statistical methodology for comparing curves

Session: **Plenary talk**

Holger Dette *Department of Mathematics, Ruhr-Universität Bochum, Germany*

Abstract: An important problem in drug development is to establish the similarity between two dose response curves (bridging studies). We propose new statistical methodology improving the current state of the art in at least two directions. On the one hand efficient designs are constructed minimizing the width of the confidence band for the difference between the regression functions, which is currently used for a test of similarity. The use of the new designs yields a reduction of the width of the confidence band by more than 50 percent and consequently to a substantially more powerful test. On the other hand – and more importantly – we propose new and substantially more powerful tests for the hypothesis of "similarity". In particular, we develop some non-standard parametric bootstrap procedure and prove its consistency. We also explain some not so well known results about classical goodness of fit tests (such as Kolmogorov-Smirnov-tests) under fixed alternatives.

Empirical Regression Quantile Process in Analysis of Risk

Session: **Plenary talk**

Jana Jurečková *Department of Probability and Statistics, Charles University, Czech Republic*

Abstract: The averaged regression α-quantile and its two-step modifications, involving R-estimators of the slope components of the linear model, turn out to be a useful tool in the situation with unknown nuisance regression. They are asymptotically equivalent to the α-quantile of the model errors, after a standardization. This allows us to make an inference without estimating the unknown parameters. The averaged regression quantile is interesting mathematically as a process with monotone step-function trajectories, whose inversion approximates the parent distribution function. On the other hand, its applications are of great interest. We shall concentrate on applications in the analysis and measuring the risk, but possible applications are also in testing hypotheses under nuisance regression, including goodness-of-fit testing, estimating various functionals of the risk, among others.

Averaged regression quantile

Let us describe the relation of the loss to covariates by the regression model

$$Y_{ni} = \beta_0 + \mathbf{x}_{ni}^\top \boldsymbol{\beta} + e_{ni}, \quad i = 1, \ldots, n \tag{1}$$

where $Y_{n1}, \ldots, Y_{nn}$ are observed responses, $e_{n1}, \ldots, e_{nn}$ are independent model errors, possibly non-identically distributed with unknown distribution functions $F_i,\ i = 1, \ldots, n$. The covariates $\mathbf{x}_{ni} = (x_{i1}, \ldots, x_{ip})^\top,\ i = 1, \ldots, n$ are random or nonrandom, and $\boldsymbol{\beta}^* = (\beta_0, \boldsymbol{\beta}^\top)^\top = (\beta_0, \beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^{p+1}$ is an unknown parameter. We also use the notation $\mathbf{x}_{ni}^* = (1, x_{i1}, \ldots, x_{ip})^\top,\ i = 1, \ldots, n$.
An important tool in the risk analysis is the regression α-quantile

$$\widehat{\boldsymbol{\beta}}_n^*(\alpha) = \Big(\hat\beta_{n0}(\alpha), (\widehat{\boldsymbol{\beta}}_n(\alpha))^\top\Big)^\top = \Big(\hat\beta_{n0}(\alpha), \hat\beta_{n1}(\alpha), \ldots, \hat\beta_{np}(\alpha)\Big)^\top.$$

It is a $(p+1)$-dimensional vector defined as a minimizer

$$\widehat{\boldsymbol{\beta}}_n^*(\alpha) = \arg\min_{\mathbf{b}\in\mathbb{R}^{p+1}} \Big\{ \sum_{i=1}^n \Big[\alpha(Y_i - \mathbf{x}_i^{*\top}\mathbf{b})^+ + (1-\alpha)(Y_i - \mathbf{x}_i^\top \mathbf{b})^-\Big]\Big\}$$
$$\text{where } \ z^+ = \max(z, 0) \ \text{ and } z^- = \max(-z, 0),\ z \in \mathbb{R}_1. \tag{2}$$

The solution $\widehat{\boldsymbol{\beta}}_n^*(\alpha) = (\hat\beta_0(\alpha), \widehat{\boldsymbol{\beta}}(\alpha))^\top$ minimizes the $(\alpha, 1-\alpha)$ convex combination of residuals $(Y_i - \mathbf{x}_i^{*\top}\mathbf{b})$ over $\mathbf{b} \in \mathbb{R}^{p+1}$, where the choice of α depends on the balance between underestimating and overestimating the respective losses Y_i. The increasing $\alpha \nearrow 1$ reflects a greater concern about underestimating losses Y, comparing to overestimating.
The quantile regression is an important method for investigation of the risk of an assett in the situation that it depends on some exogenous variables. An averaged regression quantile, introduced in [3], or some of its modifications, serve as a convenient tool for the global risk measurement in such a situation. The *averaged*

regression α-quantile is the weighted mean of components of $\widehat{\boldsymbol{\beta}}_n^*(\alpha),\ 0 \le \alpha \le 1$:

$$\bar{B}_n(\alpha) = \overline{\mathbf{x}}_n^{*\top} \widehat{\boldsymbol{\beta}}_n^*(\alpha) = \widehat{\beta}_{n0}(\alpha) + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij} \widehat{\beta}_j(\alpha), \quad \overline{\mathbf{x}}_n^* = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^* \tag{3}$$

A generally accepted measure of the financial or other risk is the expected shortfall, based on quantiles of a portfolio return or of an asset. Its properties were recently intensively studied. Acerbi and Tasche [1] speak on "expected loss in the $100\alpha\%$ worst cases", or shortly on "expected α-shortfall", $0 < \alpha < 1$. It is defined as

$$-I\!E\{Y | Y \le F^{-1}(\alpha)\} = -\frac{1}{\alpha} \int_0^{\alpha} F^{-1}(u) du, \tag{4}$$

where F is the distribution function of the asset Y. The quantity can be estimated by means of approximations of the quantile function $F^{-1}(u)$ by the sample quantiles. It is shown in [3] that $\bar{B}_n(\alpha) - \beta_0 - \bar{\mathbf{x}}_n^\top \boldsymbol{\beta}$ is asymptotically equivalent to the $[n\alpha]$-quantile $e_{n:[n\alpha]}$ of the model errors, if they are identically distributed. Hence, $\bar{B}_n(\cdot)$ can help to estimate the expected α-shortfall (4) even under the nuisance regression. It can also measure other types of risk, as in environment analysis and elsewhere. An alternative to the regression quantile is the *two-step regression α-quantile*, introduced in [2]. Here the slope components $\boldsymbol{\beta}$ are estimated by a specific rank-estimate $\tilde{\boldsymbol{\beta}}_{nR}$, which is invariant to the shift in location, and the intercept component is estimated by the α-quantile of residuals of Y_i's from $\tilde{\boldsymbol{\beta}}_{nR}$. While the averaged two-step regression quantile $\tilde{B}_n(\alpha)$ is asymptotically equivalent to $\bar{B}_n(\alpha)$ under a wide choice of the R-estimators of the slopes, the finite-sample behavior of $\tilde{B}_n(\alpha)$ is affected by the choice of R-estimator.

The averaged two-step regression quantile $\tilde{B}_n(\alpha)$ can be made motonone in α by a suitable choice of R-estimate $\tilde{\boldsymbol{\beta}}_{nR}$. Hence, then we can consider its inversion, which in turn will estimate the parent distribution F of the model errors. In such a way it will be convenient for an inference, and at the same time it is simpler than the inversion of $\bar{B}_n(\alpha)$. We shall describe this methodology in more detail.

Acknowledgment

The research was supported by the Grant GAČR 15-00243S.

References

[1] Acerbi, C. and Tasche, D. (2002). Expected shortfall: A natural coherent alternative to value at risk. *Economic Notes* 31, 379-388.

[2] Jurečková, J. and Picek, J. (2005). Two-step regression quantiles, *Sankhyā* 67, 227–252.

[3] Jurečková, J. and Picek, J. (2014). Averaged regression quantiles. In: Contemporary Developments in Statistical Theory (S. Lahiri et al., eds.), *Springer Proc. in Math. & Statistics* 68, 203–216.

[4] Koenker, R. (2005). *Quantile Regression*. Cambridge University Press

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

Aspects on the Law of Large Numbers

Session: **Plenary talk**

Allan Gut *Department of Mathematics, Uppsala University, Uppsala, Sweden*

Abstract: The law of large numbers is one of fundamental results, actually the first one, in probability theory; the two others being the central limit theorem and the law of the iterated logarithm. In this talk we discuss the law of large numbers in the i.i.d. case with some extensions and generalizations from the following list:

- Marcinkiewicz–Zygmund laws
- Convergence rates
- Precise asymptotics
- Uniform integrability and moment convergence
- Randomly indexed sums
- Weighted sums
- Summation methods; e.g. Cesàro summation
- Other normalizations
- Subsequences
- Arrays
- What happens when $p \nearrow 2$?
- Non-i.i.d. summands
- Multidimensional index sets; random fields
- Vector (Banach space) valued r.v's
- Dependent sequences
- Renewal theory (for random walks)
- Records

We close with some recent joint work with Ulrich Stadtmüller on some boundary problems.

Sensitivity analysis and optimization of computer experiments with an application to a centrifugal compressor impeller

Session: **Stochastic Models in Technology, Reliability, and Quality 1**

Sonja Kuhnt *Department of Computer Science, Dortmund University of Applied Sciences and Arts, Dortmund, Germany*

Abstract: Computer simulations (black-box experiments) of complex processes in engineering and natural sciences nowadays commonly replace real-life experiments. As these simulations can be very time-consuming and complex often a surrogate or meta-model is built first and sensitivity analysis and optimization are then based on this model. Gaussian process models, better known as Kriging models, are widely used for this purpose. Sensitivity analysis investigates how the input variables contribute to the variation of the outcome. A popular method is the use of Sobol indices which quantify the importance of individual input variables or groups of them [1]. We review the Sobol sensitivity indices and the more recently developed total interaction indices (TII) and show different ways to display the result in so called FANOVA graphs [2, 3]. The popular efficient global optimization (EGO) procedure provides a sequential Kriging-based optimization procedure with the expected improvement as criterion [4]. Motivated by an application to a centrifugal compressor impeller this contribution treats a situation where an extension to multiple responses as well as constraints is needed, where the constraint functions are also an output of the computer simulation.

References

[1] Sobol, I.M. (1993). Sensitivity estimates for non linear mathematical models. *Mathematical Modelling and Computational Experiments* **1**, 407-414.

[2] Fruth, J., Roustant, O. and Kuhnt, S. (2014). Total Interaction Index: A Variance-based Sensitivity Index for Second-order Interaction Screening. *Journal of Statistical Planning and Inference* **147**, 212–223.

[3] Roustant, O., Fruth, J., Iooss, B. and Kuhnt, S. (2014). Crossed-derivative based sensitivity measures for interaction screening. *Mathematics and Computers in Simulation* **105**, 105–118.

[4] Jones, D.R. and Schonlau, M. and Welch, W.J. (1998). Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization* **13**, 455–492.

Stochastic comparisons of systems based on sequential order statistics

Session: **Stochastic Models in Technology, Reliability, and Quality 1**

Marco Burkschat *Institute of Statistics, RWTH Aachen University, D-52056 Aachen, Germany*
Jorge Navarro *Facultad de Matemáticas, Universidad de Murcia, 30100 Murcia, Spain*

Abstract: The model of sequential order statistics has been proposed for describing increasingly ordered failure times of components in technical systems where failures may have an impact on the lifetimes of remaining components. In the considered systems the lifetime distributions of surviving components are allowed to change after the occurrence of a failure. In the talk systems based on sequential order statistics with underlying distributions possessing proportional hazard rates are studied. In that case the lifetime distribution of the system can be expressed as a distorted distribution. Motivated by the distribution structure in the case of pairwise different model parameters, a particular class of distorted distributions, the generalized proportional hazard rate model, is defined and characterizations of stochastic comparisons for several stochastic orders are obtained. Moreover, related asymptotic results on aging characteristics of general distorted distributions with applications to coherent systems based on sequential order statistics are also considered.

1 Introduction

Using the notion of coherent systems of [1], the lifetime T of a system consisting of n components can be expressed as $T = \phi(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ denote the lifetimes of the components and ϕ the coherent life function of the system. In the talk, the component lifetimes $X_1, \ldots, X_n$ are modeled via random variables with a dependence structure induced by sequential order statistics. In doing so, the lifetime distributions of surviving components may change after every failure of a unit. The model of sequential order statistics has been proposed in [2] for describing increasingly ordered failure times $X^*_{1:n} \leq \cdots \leq X^*_{n:n}$ of system components in such situations when the lifetimes of remaining units may be affected by component losses. The corresponding reliability of the system is then given by a mixture (see, e.g., [3])

$$P(T > t) = \sum_{i=1}^{n} s_i P(X^*_{i:n} > t), \quad t \in \mathbb{R}, \tag{1}$$

where the weights are determined by the signature $\mathbf{s} = (s_1, \ldots, s_n)$ of the system (see, e.g., [4]).

2 Representation via distorted distribution

It is assumed that the survival functions $\overline{F}_i, i = 1, \ldots, n$, belonging to different stages in the model of sequential order statistics are given by

$$\overline{F}_i = \overline{F}^{\alpha_i}, \quad i = 1, \ldots, n, \tag{2}$$

where $\overline{F}$ is a continuous survival function and $\alpha_1, \ldots, \alpha_n$ are positive parameters. Then $F_1, \ldots, F_n$ belong to a Proportional Hazard Rate model with baseline survival function $\overline{F}$. In this case the corresponding survival function of the system lifetime can be expressed as

$$P(T > t) = \overline{q}(\overline{F}(t)), \quad t \in \mathbb{R}, \tag{3}$$

with a distortion function $\overline{q}$ (for the definition, see, e.g., [5]), i.e. the distribution of the system lifetime can be interpreted as a distorted distribution. For pairwise different model parameters $\gamma_i = \alpha_i(n - i + 1), i = 1, \ldots, n$, the resulting distribution is contained in the generalized proportional hazard rate model (see [6]). In the talk, characterizations of different stochastic comparisons are treated in this setting. Moreover, results on the asymptotic behavior of the hazard rate and the mean residual life function of general distorted distributions are given. The results are supplemented with limiting properties of the systems in the case of possibly equal model parameters. Some examples are presented in order to illustrate the findings.

References

[1] Barlow, R. E. and Proschan, F. (1981). *Statistical Theory of Reliability and Life Testing*. To Begin With, Silver Spring, Maryland.

[2] Kamps, U. (1995). A concept of generalized order statistics. *Journal of Statistical Planning and Inference* 48, 1–23.

[3] Navarro, J. and Burkschat, M. (2011). Coherent systems based on sequential order statistics. *Naval Research Logistics* 58, 123–135.

[4] Samaniego, F. J. (2007). *System Signatures and their Applications in Engineering Reliability*. Springer, New York.

[5] Hürlimann, W. (2004). Distortion risk measures and economic capital. *North American Actuarial Journal* 8, 86–95.

[6] Burkschat, M. and Navarro, J. (2016). Stochastic comparisons of systems based on sequential order statistics via properties of distorted distributions. *To appear in Probability in the Engineering and Informational Sciences.*

Incomplete Repair in Degradation Processes: A Kijima-Type Approach

Session: **Stochastic Models in Technology, Reliability, and Quality 1**

Waltraud Kahle *Institute of Mathematical Stochastics, Otto-von-Guericke University Magdeburg, Germany*

Abstract: We consider a Wiener process with (possible nonlinear) drift for degradation modeling. Regularly, inspections are carried out, and the level of degradation is measured. At each inspection point, a decision about preventive maintenance actions (replace the system by a new one, let it as it is, ore make an incomplete repair with some degree) is possible. In the talk, we consider the influence of such maintenance actions to the further development of the degradation process and the resulting lifetime distribution. A connection between virtual age in Kijima-type models and degradation level in the underlying degradation process is developed. Our aim is to find optimal time points of inspection and optimal decisions about maintenance actions.

1 The degradation process and the resulting lifetime distribution

We consider a Wiener process with drift $Z(t)$:

$$Z(t) = z_0 + \sigma W(t - t_0) + \mu \cdot (t - t_0) \ , \quad t \geq t_0 \tag{1}$$

with
- z_0 - constant initial degradation ($z_0 \in \mathbb{R}$),
- t_0 - beginning of the degradation ($t_0 \in \mathbb{R}$),
- μ - drift parameter ($\mu \in \mathbb{R}$),
- σ - variance parameter ($\sigma > 0$),
- $W(t)$ - standard Wiener process on $[0, \infty)$.

The lifetime T_h is the time of first passing a given level h

$$T_h = \inf\{t \geq t_0 : Z(t) \geq h\} \ . \tag{2}$$

It is well known that the resulting lifetime distribution is the inverse Gaussian distribution:

$$f_{T_h}(t) = \frac{h - z_0}{\sqrt{2\pi\sigma^2(t - t_0)^3}} \exp\left(-\frac{(h - z_0 - \mu(t - t_0))^2}{2\sigma^2(t - t_0)}\right) I_{\{t > t_0\}}, \tag{3}$$

2 Kijima incomplete preventive maintenance

In Kijima-type models an item with initial failure rate $\lambda_1(t) = \lambda(t)$ is considered. At some time t_1 (this can be a failure time or a stopping time for preventive maintenance) the item will be repaired with the degree ξ_1. By this repair action the age of the item is decreased to $v_1 = \xi_1 t_1$. The distribution of the time until the next failure has the failure rate $\lambda_2(t) := \lambda(t - t_1 + v_1)$. Let $t_1, t_2, \ldots$ be the sequence of maintenance points. The process defined by

$$v(t) := t - t_n + v_n, \quad t_n \leq t < t_{n+1} \ , \ n \geq 1$$

is called the *virtual age process.* Kijima has considered two possible types for the reduction of the virtual age:

$$\begin{aligned} \text{Kijima Type I:} \quad v_k &= v_{k-1} + \xi_k(t_k - t_{k-1}) \\ \text{Kijima Type II:} \quad v_k &= \xi_k(v_{k-1} + (t_k - t_{k-1})) \end{aligned}$$

The distribution of the time until the next failure then has failure intensity $\lambda_{k+1}(t) = \lambda(t - t_k + v_k)$.

3 Application to degradation processes

We consider the degradation process $Z(t)$ with maintenance points τ_1, τ_2, ... Let the state of the process at first inspection be $Z(\tau_1-) = z_1$ and let ξ be the degree of repair.
There are 2 possibilities of an incomplete repair at the first inspection point: the reduction of degradation level : $z(\tau_1) = \xi \cdot z_1$ and the new virtual age is $v(\tau_1) = \xi \cdot z_1/\mu$, or the reduction of virtual age $v(\tau_1) = \xi \cdot \tau_1$ and the new state is $z(\tau_1) = \xi \cdot \mu \cdot \tau_1$.
At the second inspection point we have to make a difference between Kijima Type I and Kijima Type II models.

In the talk, we consider the influence of such maintenance actions to the further development of the degradation process and the resulting lifetime distribution. A connection between virtual age in Kijima-type models and degradation level in the underlying degradation process is developed.
Further, some questions about optimal maintenance are considered.

References

[1] Kijima M. (1989) Some results for repairable systems with general repair, *Journal of Applied Probability,* vol 26, pp 89-102.

[2] Kijima M., Morimura H., Suzuki Y. (1988) Periodical replacement problem without assuming minimal repair, *European Journal of Operational Research*, vol 37, pp 194-203.

Modeling and forecasting multivariate electricity price spikes

Session: **Stochastic Models in Technology, Reliability, and Quality 1**

Hans Manner *Institute of Econometrics and Statistics, University of Cologne, Germany*
Dennis Türk *RCMA Group Pte Ltd, Singapore*
Michael Eichler *Department of Quantitative Economics, Maastricht University, The Netherlands*

Abstract: We consider the problem of forecasting the occurrence of extreme prices in the Australian electricity markets from high frequency spot prices. In particular, we are interested in the simultaneous occurrence of such so-called spikes in two or more markets. Our approach is based on a novel dynamic model for multivariate binary outcomes, which allows the latent variables driving these observed outcomes to follow a vector autoregressive process. Furthermore the model is constructed using a copula representation for the joint distribution of the resulting innovations. This has several advantages over the standard multivariate probit model. First, it allows for nonlinear dependence between the error terms. Second, the distribution of the latent errors can be chosen freely. Third, the computational burden can be greatly reduced making estimation feasible in higher dimensions and for large samples. The model is applied to spikes in half-hourly electricity prices in four interconnected Australian markets. The multivariate model provides a superior fit compared to single-equation models and generates better forecasts for spike probabilities. Furthermore, evidence of spillover dynamics between the markets is revealed.

Asymptotic normality of a nearest neighbor estimate of the second moment regression functional with an application to dimension reduction

Session: **Nonparametric Regression and Density Estimation 1**

Luc Devroye *School of Computer Science, McGill University, Montreal, Canada*
László Györfi *Department of Computer Science and Information Theory, Budapest University of Technology and Economics, Budapest, Hungary*
Gábor Lugosi *ICREA and Department of Economics and Business, Pompeu Fabra University, Barcelona, Spain*
Harro Walk *Department of Mathematics, University of Stuttgart, Stuttgart, Germany*

Abstract: The problem of estimating the minimum achievable mean squared error in regression estimation is equivalent to estimating the second moment of the regression function of Y on $X \in \mathbb{R}^d$. We present a nearest-neighbor-based estimate. A normal limit law for the estimate is obtained under the only assumptions that X has a density and Y is bounded. The asymptotic variance is calculated. It depends only on d and on conditional moments of Y. We apply the estimate for testing whether a component of the vector X carries information.

1 Introduction

This paper considers a nearest-neighbor-based estimate of second moment of the regression function in multivariate problems. The interest in the second moment is motivated by the fact that by estimating it one obtains an estimation of the best possible achievable mean squared error, a quantity of crucial interest in statistics. The estimate is asymptotically normally distributed. It is remarkable that the asymptotic variance only depends on conditional regression moments. Moreover, the asymptotic variance is bounded by a constant that is independent of the dimension and of the underlying distributions. We apply these results to construction of a test for deciding whether a component of the observational vector has predictive power. Let (X, Y) be a pair of random variables such that $X = (X^{(1)}, \ldots, X^{(d)})$ takes values in $\mathbb{R}^d$ and Y is a real-valued random variable with $\mathbb{E}[Y^2] < \infty$. We denote by $m(x) = \mathbb{E}\{Y \mid X = x\}$ the *regression function* of Y on X. The problem of estimating the minimum mean squared error $\mathbb{E}\left\{(m(X) - Y)^2\right\}$ is essentially equivalent to estimation of the second moment $S^* = \mathbb{E}\left\{m(X)^2\right\}$ of the regression function.

In this paper we present a nearest-neighbor-based estimator of S^*. The advantage of this estimator, apart from sharing the fast rates of convergence of previously defined estimators, is that its random fluctuations may be bounded by dimension and distribution-independent quantities. We show a central limit theorem for the estimator and design a test for deciding whether exclusion of a certain component of X increases the minimum mean squared error or not.

2 The result

Assume that we have $2n$ samples split into two halves as

$$D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \qquad \text{and} \qquad D'_n = \{(X'_1, Y'_1), \ldots, (X'_n, Y'_n)\}$$

such that $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n), (X'_1, Y'_1), \ldots, (X'_n, Y'_n)$ are i.i.d.
The nearest-neighbor-based estimator of S^* is defined as follows. Based on the data D_n construct the nearest-neighbor (1-NN) regression function estimator as follows. Let $X_{1,n}(x)$ be the first nearest neighbor of x from $X_1, \ldots, X_n$ and $Y_{1,n}(x)$ its label. The 1-NN regression estimator of the regression function m is defined as

$$m_n(x) = Y_{1,n}(x).$$

The splitting estimate of S^* is

$$S_n = \frac{1}{n} \sum_{i=1}^{n} Y'_i m_n(X'_i) \ .$$

Let $S_{x,r}$ denote the closed ball of radius $r > 0$ centered at x in $\mathbb{R}^d$ and let λ denote the Lebesgue measure on $\mathbb{R}^d$. Let V be a random vector uniformly distributed in $S_{0,1}$. Define $\overline{1} = (1, 0, 0, \ldots, 0) \in \mathbb{R}^d$ and let $\overline{S} = S_{\overline{1},1} \bigcup S_{V,\|V\|}$. Introduce the random variable $W = \lambda(\overline{S})/\lambda(S_{0,1})$. The constant

$$\alpha(d) = \mathbb{E}\left\{\frac{2}{W^2}\right\} \in [1, 2]$$

describes the asymptotic behavior of the second moments of Voronoi cell measures and plays an important role in Theorem 1 below.

Theorem 1 *(Devroye et al. [1].) Assume that the distribution μ of X has a density and that $|Y| < L$. Then*

$$\sqrt{n}\,(S_n - \mathbb{E}\{S_n\})\,/\sigma \xrightarrow{\mathcal{D}} N(0, 1),$$

where $\sigma^2 = \sigma_1^2 + \sigma_2^2 \leq 3L^4$ with

$$\sigma_1^2 = \int M_2(x)^2 \mu(dx) - \left(\int m(x)^2 \mu(dx)\right)^2 > 0$$

and

$$\sigma_2^2 = \alpha(d)\left(\int M_2(x) m(x)^2 \mu(dx) - \int m(x)^4 \mu(dx)\right)$$

and $M_2(X) = \mathbb{E}\{Y^2 \mid X\}$.

References

[1] Devroye, L., Györfi, L., Lugosi, G. and Walk, H.: A nearest neighbor estimate of a regression functional, (in preparation)

Nonparametric regression estimation using hierarchical interaction models

Session: **Nonparametric Regression and Density Estimation 1**

Michael Kohler *Fachbereich Mathematik, Technische Universität Darmstadt, Germany*
Adam Krzyżak *Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada*

Abstract: In this paper we introduce so-called hierarchical interaction models where we assume that the computation of the value of a function $m : \mathbb{R}^d \to \mathbb{R}$ is done in several layers, where in each layer a function of at most d^* inputs computed by the previous layer is evaluated. We investigate two different regression estimates based on polynomial splines and on neural networks and show that if the regression function satisfies a hierarchical interaction model and all occurring functions in the model are smooth, the rate of convergence of these estimates depends on d^* (and not on d). Hence in this case the estimates can achieve good rate of convergence even for large d and are in this sense able to circumvent the so-called curse of dimensionality.

On estimation of surrogate models for high-dimensional computer experiments

Session: **Nonparametric Regression and Density Estimation 1**

Benedikt Bauer *Fachbereich Mathematik, Technische Universität Darmstadt, Germany*
Michael Kohler *Fachbereich Mathematik, Technische Universität Darmstadt, Germany*
Adam Krzyżak *Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada*
Felix Heimrich *Fachbereich Maschinenbau, Technische Universität Darmstadt, Germany*

Abstract: Estimation of surrogate models for computer experiments leads to nonparametric regression estimation problems without noise in the dependent variable. We propose an empirical maximal deviation minimization principle to construct estimates in this context, and analyze the rate of convergence of corresponding quantile estimates. As an application we consider estimation of high-dimensional computer experiments by neural networks, and show that here we can circumvent the so-called curse of dimensionality by imposing rather general assumptions on the structure of the regression function.

1 Introduction

Physical phenomena are nowadays often described by mathematical models, which enables the use of so-called computer experiments instead of real experiments in order to analyze them. In the simplest case the mathematical model is described by a function $m : \mathbb{R}^d \to \mathbb{R}$, which models the relation between d–dimensional input and a real-valued output. Due to uncertainty in nature, we consider

$$Y = m(X), \tag{1}$$

where X is an $\mathbb{R}^d$-valued random variable and $m : \mathbb{R}^d \to \mathbb{R}$ is a real–valued function. Here this function can be, e.g., the solution of a partial differential equation system, where the value of X determines the values of parameters and initial conditions of this system. The aim of studying the physical phenomenon is to derive characteristics of the outcome Y. In the mathematical model Y is a real-valued random variable, and we are interested in the distribution of this random variable. It is often possible to write a complex computer program which generates the values of function m. Since computing of values of $m(X)$ is often extremely time-consuming in practice, it is not possible to generate a large sample size to analyze Y. One idea to circumvent this problem is to use so-called surrogate models for m. Here we begin by generating data

$$(X_1, m(X_1)), \ldots, (X_n, m(X_n)) \tag{2}$$

by evaluating the computer model for realizations of n independent copies of X and using this data to construct an estimate

$$m_n(\cdot) = m_n(\cdot, (X_1, m(X_1)), \ldots, (X_n, m(X_n))) : \mathbb{R}^d \to \mathbb{R} \tag{3}$$

of m. Given the estimate one can generate a modified sample

$$Y_{n+1} = m_n(X_{n+1}), \ldots, Y_{n+N_n} = m_n(X_{n+N_n}) \tag{4}$$

of Y for a sample size N_n, which is much larger than n, and this sample can be used together with nonparametric statistics in order to estimate some aspects of the distribution of Y, e.g., its quantiles (cf., e.g., [1]).

2 Methods

In order to estimate m, we choose a set $\mathcal{F}_n$ of functions $f : \mathbb{R}^d \to \mathbb{R}$ and select from this set a function, which fits our data within a given set $B_n \subseteq \mathbb{R}^d$ well. For this purpose, we minimize the maximal absolute error on the observed data contained in B_n. Formally, the resulting least empirical deviation estimate is defined by

$$m_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \max_{\substack{i=1,\dots,n, \\ X_i \in B_n}} |f(X_i) - m(X_i)| . \tag{5}$$

Here we assume for simplicity, that the above minimum exists (otherwise minor modification of the criterion).

3 Results

After deriving some general results regarding the approximation quality of function estimates of the above type, we focused on a special case. We assume that m satisfies a so-called (p, C)-smooth generalized hierarchical interaction model of order $d^* \leq d$ and with $p \leq 1$ (cf., [2]), a structure that includes many other types of functions assumed in the literature. Choosing $\mathcal{F}_n$ as a certain set of not completely connected multilayer neural networks, we can circumvent the curse of dimensionality and derive rates of convergence, that are independent of d, for the error of the estimate m_n as well as the corresponding quantile estimate. By using simulated data we demonstrate in several high-dimensional settings that our newly proposed quantile estimate can outperform other quantile estimates even for finite sample size.

References

[1] Enss, G., Kohler, M., Krzyżak, A., and Platz, R. (2016). Nonparametric quantile estimation based on surrogate models. *IEEE Transactions on Information Theory*, **62**, pp. 5727–5739.

[2] Kohler, M., and Krzyżak, A. (2016). Nonparametric regression based on hierarchical interaction models. To appear in *IEEE Transaction on Information Theory* (2017).

A panel cointegration rank test with structural breaks and cross-sectional dependence

Session: **Panel Data**

Antonia Arsova *Center of Methods, Leuphana Univerität Lüneburg, Germany*

Deniz Dilan Karaman Örsal *Center of Methods and Institute of Economics, Leuphana Univerität Lüneburg, Germany*

Abstract: This paper proposes a new panel cointegration rank test which allows for a linear time trend with breaks and cross-sectional dependence. The new correlation-augmented inverse normal (CAIN) test is based on a novel modification of the inverse normal method and combines the p-values of individual likelihood-ratio trace statistics. A Monte Carlo study demonstrates its robustness to cross-sectional dependence and its superior size and power properties compared to other meta-analytic tests used in practice.

1 Introduction

In this paper we propose a new meta-analytic approach (CAIN) to test for the cointegrating rank in panels where structural breaks and cross-sectional dependence are allowed for. We extend the TSL (see [1]) test to the panel setting resorting to a new p-value combination method which allows the p-values to be correlated. The CAIN-TSL test is based on a modification of the popular inverse normal method (see [2]) for p-values combination, employing a novel estimator for the unknown correlation between the probits $\tilde{\rho}_t$. p-value combination approaches offer much more flexibility than traditional pooling of individual test statistics, as they allow the specification of the deterministic terms, the lag order, the number and the location of the breaks, and even the time span of the data to vary over cross-sections.

2 The method

Let the observed data $Y_{it} = (Y_{1,it}, \ldots, Y_{m,it})'$ for cross-sectional unit i $(i = 1, \ldots, N)$ be generated by a stochastic VAR(s_i) process X_{it} added to a deterministic process. The latter consists of a constant, linear time trend and structural breaks in both the level and the trend slope at known individual-specific time(s) τ_i:

$$Y_{it} = \mu_{0i} + \mu_{1i}t + \delta_{0i}d_{it} + \delta_{1i}b_{it} + X_{it}, \quad t = 1, \ldots, T. \tag{1}$$

Here μ_{ji} and δ_{ji} $(j = 0, 1)$ are unknown $(m \times 1)$ parameter vectors, while d_{it} and b_{it} are dummy variables defined by $d_{it} = b_{it} = 0$ for $t < \tau_i$, and $d_{it} = 1$ and $b_{it} = t - \tau_i + 1$ for $t \geq \tau_i$. The break dates are assumed to occur at individual-specific fixed fractions of the sample size: $\tau_i = [T\lambda_i]$ with $0 < \underline{\lambda_i} < \lambda_i < \overline{\lambda_i}$, where $\underline{\lambda_i}$ and $\overline{\lambda_i}$ are specified real numbers and $[\cdot]$ denotes the integer part of the argument. In other words, the breaks are assumed not to occur in the very beginning or in the very end of the sample, while $\underline{\lambda_i}$ and $\overline{\lambda_i}$ are allowed to be arbitrarily close to 0 and 1, respectively. The stochastic processes X_{it} are assumed to be at most $I(1)$ and cointegrated with cointegrating rank at most r:

$$X_{it} = A_{1i}X_{i,t-1} + \ldots A_{s_i i}X_{i,t-s_i} + \varepsilon_{it}, \quad t = 1, \ldots, T_i. \tag{2}$$

It is assumed that the $(m \times 1)$ vector $\varepsilon_{it} \sim iid(0, \Omega_i)$, where Ω_i is a positive definite matrix for each i. Further it is assumed that ε_{it} have finite moments of order $(4+\nu)$ for some $\nu > 0$, $\forall i$. Denoting the pairwise cross-sectional correlations of the elements of ε_{it} by $\rho_{il,jk} := corr(\varepsilon_{it,l}, \varepsilon_{jt,k})$ for $i, j = 1, \ldots, N$ and $l, k = 1, \ldots, m$, we make the following assumptions.

Assumption 1

$$\lim_{N\to\infty} \frac{1}{mN(N-1)} \sum_{i\neq j}^{N} \sum_{l=1}^{m} |\rho_{il,jl}| = \rho_{\varepsilon}. \tag{3}$$

Assumption 2

$$\lim_{N\to\infty} \frac{1}{mN(N-1)} \sum_{i\neq j}^{N} \sum_{l\neq k}^{m} |\rho_{il,jk}| = 0. \tag{4}$$

The panel test statistic for the composite null hypothesis

$$H_0 : r_i = r, \forall i = 1, \ldots, N \quad vs. \quad H_1 : r_i > r, \exists i \tag{5}$$

is given by

$$t(\tilde{\rho}) = \frac{\sum_{i=1}^{N} t_i}{\sqrt{N + (N^2 - N) \cdot \tilde{\rho}_t}} \tag{6}$$

where r is the number of cointegrating relations. The CAIN-TSL test can be applied for testing $H_0 : r_i = r, \forall i$ at each step $r = 0, \ldots, m-1$ of the sequential rank testing procedure. We estimate $\tilde{\rho}_t$ as a function of the dimension of the system m, the hypothesized cointegrating rank r and the estimated mean absolute correlation ρ_{ε} between the innovations of the individual DGPs. This function results from a response surface regression of the values of $\tilde{\rho}_t$, estimated by simulation, on m, r and ρ_{ε}. The latter is easily estimable in practice and provides an easy-to-interpret measure of the degree of cross-sectional dependence. In a Monte Carlo study we demonstrate the superior properties of the CAIN-TSL test in comparison with other p-value combination tests.

References

[1] Trenkler C., Saikkonen P., Lütkepohl H. (2007). Testing for the Cointegrating Rank of a VAR Process with Level Shift and Trend Break. *Journal of Time Series Analysis* 29: 331-358. ISSN 0143- 9782.

[2] Choi I. 2001. Unit root tests for panel data. *Journal of International Money and Finance* 20: 249-272.

Forecasting Regional Unemployment With Spatial Panel Data Models

Session: **Panel Data**

Elena Semerikova *Department of Econometrics, Humboldt University of Berlin, Germany*
Olga Demidova *Department of Applied Economics, Higher School of Economics, Russia*

Abstract: We consider forecasting unemployment in Russian and German regions with the help of econometric panel data models. Using regional data from 2005 till 2012 we show that spatial panel data models perform better in terms of forecasting accuracy than other models (on average and at least for some distinct regions) such as non-spatial panel data models, pooled OLS, models without exploratory variables and naive forecasts (average value for one or several previous periods).

In the current paper we make an attempt to predict unemployment rates at the regional level for Russia and Germany. With the help of panel data structure we account for the regional heterogeneity and take into account spillover effects between regions with the help of spatial econometric approach. We employ data from 2005 till 2012 (for Russia) and from 2005 till 2011 (for Germany) for estimation and leave two years (2011–2012) for prediction assuming that we know the values of the exploratory variables. We also make prediction for one additional year assuming that the values are not known, and therefore employ their dynamic lags for the prediction.

The main objective of the paper is to examine to what extent the introduction of the spatial effects into model increases the quality of the forecast. We compare predictions made with the help of spatial econometric models with predictions obtained with the help of non-spatial panel data models. Besides we compare the quality of the forecast with naive predictions (previous value of unemployment rates and average value of previous periods). In order to measure the forecasting accuracy of the models we calculate root mean squared error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE) and symmetric absolute percentage error (sMAPE).

We find that in case of known future values of the exploratory variables spatial models allow to obtain the best forecasting accuracy. In case of using the lags of the exploratory variables spatial models on average perform a little bit worse than naive models, however they still allow to obtain better predictions for some certain regions. Thus, we conclude that accounting for spatial correlation allows to obtain better prediction quality.

Local political budget cycles in a federation: Evidence from West German cities

Session: **Panel Data**

Marina Furdas *Department of Economics, Humboldt University Berlin, Germany*
Katerina Homolkova *University of Kiel, Germany*
Krisztina Kis-Katos *Department of Economics, University of Göttingen, Germany*

Abstract: This paper analyzes the occurrence of political budget cycles in 604 West German cities between 1975 and 2007. Due to the idiosyncratic timing of state and local elections, the budgetary changes before electionsat two tiers of the federalist government can be separately estimated and can also be distinguished from common time effects. Both local and state elections result in pre-election manipulation of the local finances of moderate size. Before both types of elections, we observe an increase in building investments, accompanied by increasing intergovernmental grants for investment purposes but also a halt in the increase of local tax rates. By contrast, elections at the two tiers of the government affect the size of the current budget differently: current revenues and expenditures decrease before local but increase before state elections, suggesting a difference in the tightness of the local budget constraint. The extent of these political budget cycles is more pronounced in municipalities that are politically aligned with the state governments and are politically more contested.

1 Introduction

In multi-tiered governmental systems, local budgets can be affected by electoral incentives at various levels of the government. Due to its federalist structure and electoral system, Germany offers a unique case study on the interaction between local and state elections and their effects on local budget cycles. The idiosyncratic, state-wise varying timing of elections at the state and local level of government enables us to identify the effects of both types of elections on the size and composition of local budgets and to separate them from common country-wide shocks to the economy and the political system. Moreover, we can investigate whether the amplitude of the cycles varies with differences in the political environment.

2 Methodology

As a baseline model, we estimate the following equation:

$$Y_{it} = \beta\, LEL(\tau)_{it} + \delta\, SEL(\tau)_{it} + \gamma X_{it} + \lambda_i + \kappa_t + \epsilon_{it} \quad (1)$$

where Y_{it} denotes the budgetary variable of municipality i in year t. $LEL(\tau)_{it}$ and $SEL(\tau)_{it}$ consist of two vectors of indicators for the timing of local and state elections, where the elements of the vector $LEL(\tau) = [LEL(T-1), LEL(T), LEL(T+1)]$ take 1 if $t = \tau$ (where t stands for the current year, $\tau \in \{T-1, T, T+1\}$ and T is the election year) and 0 otherwise. SEL is determined in the same way. X_{it} is the vector of further controls (in the baseline, population size and an indicator for directly elected mayors), λ_i and κ_t are municipality and time fixed effects, and ϵ_{it} is the error term. In the baseline we estimate panel data models that are purged of

municipality fixed effects through a demeaning transformation. Municipality fixed effects control for time-invariant sources of heterogeneity across municipalities like their institutional status and geographic location. The time fixed effects control for economic and political shocks common to all West-German cities as well as common changes in budgetary procedures over time. We cluster standard errors at municipality level to correct for the autocorrelation in budgetary variables.
In our empirical analysis, we acknowledge that all budgetary variables are jointly determined within the same budgeting process and hence cannot be considered to be independent from each other. In statistical terms, we allow for a correlation between our six selected budgetary variables by estimating them jointly in a seemingly unrelated regression (SUR) model and base our estimates of significance of the specific coefficients on Bonferroni-adjusted p-values. The Bonferroni method adjusts the individual p-values by k, the number of concurrent tests being executed (in our baseline case, $k = 6$) and results in $p_i^B = \min(kp_i, 1)$. This method is relatively strict, yielding conservative p-values, and is very unlikely to overestimate the significance of an electoral cycle.

3 Empirical results

The empirical results showed the most pronounced cycles in local tax setting (tax rates are less likely to be increased before elections), but also in building investments and the related investment grants, all of which changed by about 8-10% of a standard deviation around election years. This contradicts the notion that political budget cycles should vanish in countries with more mature democracies and more experienced voters. Further results suggest that state elections were related to an overall increase in current expenditures and potentially even revenues. By contrast, local elections involved not only decreasing taxes but also lower current expenditures and revenues as well as a shift from administrative towards investment grants. The two sets of results are supportive of the idea of a harder local budget constraint before local than before state elections.

References

[1] Rogoff, K. and A. Silbert (1988). *Elections and Macroeconomic Policy Cycles*, Review of Economic Studies 55(1), 1-16.

[2] Shi, M. and Svensson, J. (2006). *Political budget cycles: Do they differ across countries and why?* Journal of Public Economics 90(8-9), 1367–1389.

[3] Veiga, L. G. and Veiga, F. J. (2013). *Intergovernmental fiscal transfers as pork barrel.*, Public Choice 155(3/4), 335–353.

The temporal stability of initial nonresponse in panel surveys.

Session: **Panel Data**

Ulrich Rendtel *Economic Department, FU Berlin, Germany*
Juha Alho *Department of Social Statistics, University Helsinki, Finland*
Gerrit Müller *IAB, Nürnberg*

Abstract: Nonresponse in a panel survey is analysed here under a Markov chain setting. Under certain regularity conditions it can be proven that the distribution on the state space for the respondents and non-respondents converge in later panel waves. The speed of convergence depends on the temporal stability of the variable of interest. Empirical results are displayed for the German Panel on Labour Market and Social Security (PASS).

1 Extended Abstract

The term "Fade Away Effect" was coined to name a phenomenon that can be observed in panel surveys which were sampled from registers. If one has access to the register information of all units, including the non-respondents of the initial wave of the panel, one will observe that a non-response bias in the first wave declines in the sub- sequent waves by itself. [1] has used a Markov chain approach to model this fade-away of the initial nonresponse. The key idea is the existence of a steady state distribution of the Markov chain. Initital nonresponse is then interpreted as a deviation from the steady state distribution. The distribution on the state space swings into the steady state distribution with a velocity that depends on the transition matrix of the Markov chain. Empirical results for the EU-SILC were given in [1]. In this work we investigate a sample that is far away from a steady state distribution. Also the transition matrix is not stable over time. So in a strict sense the steady state distribution is not present in this case. Yet it can be demonstrated that a Fade Away Effect exists in this case. The empirical data base is the German Panel on Labour Market and Social Security (PASS). Here a sample of social benefit receivers was selected from the German Social Security Register. It was possible to get access to the employment information from also for the non-responders of the first wave. The analysis covers the first 5 waves of the PASS. The non-response rate at the start was about 80 percent. We analyse transitions into and out-of social aid payments and unemployment.

Keywords: Panel surveys, nonresponse, Markov chains, steady state distribution, social aid payments.

References

[1] Rendtel U. (2013). The fade away effect of initial nonresponse in panel urveys: Empirical results for EU-SILC, Eurostat Methodologies and Working Papers. http://ec.europa.eu/eurostat/documents/3888793/5857657/KS-RA-13-012-EN.PDF

Distance Correlation Screening for Detecting Nonlinear Associations in Ultrahigh Dimensional DNA Methylation Data

Session: **Analysis, Testing and Change Detection in High Dimensions 1**

Dominic Edelmann *Division of Biostatistics, German Cancer Research Center (DKFZ)*
Axel Benner *Division of Biostatistics, German Cancer Research Center (DKFZ)*

Abstract: We introduce a novel method for variable screening via a modification of the distance correlation coefficient. The consistency of this method in lowdimensional Gaussian linear models is derived. A large simulation study outlines the efficacy of our method compared to existing methods. Finally, we apply our method on epigenome wide DNA methylation data.

1 Introduction

With the rise of high-throughput data, computationally and statistically efficient variable selection techniques have emerged, e.g. the LASSO, the elastic net or LARS, to name only a few. However, these valuable procedures become more and more inefficient as the dimension p increases. Consequently, they often perform poorly on ultra-high-dimensional data such as DNA methylation measurements from microarray experiments ($p > 450.000$). In the setting of linear models, Fan, et al. [3] have proposed the sure independence screening (SIS) method, which allows for screening of random variables in ultra-high-dimensional data. Recently, Li, et al. [1] have modified this approach by replacing Pearson correlation with distance correlation, a novel measure of independence introduced by Székely, et al. [6]. However, as pointed out by many authors, Distance Correlation-SIS (DC-SIS) fails to detect important predictors which are marginally independent of the response due to correlation among the predictors. To overcome this issue, Zhong, et al. [4] and Yenigün, et al. [5] have proposed stepwise DC-SIS methods. While they demonstrate the efficacy of their techniques in broad simulation studies, no theoretical guarantees are given.

In the first part of this talk, we point out pitfalls of distance correlation methods when comparing vectors with different dimensions. In particular, we investigate the asymptotic behavior of the distance correlation coefficient for the case where the dimension of one vector is fixed while the dimension of the other vector goes to infinity. We present suggestions for modified distance measures that show a more meaningful behavior in high dimensions.

In the second part of the talk, we present a new approach to feature screening via the affinely invariant distance correlation [2]. We prove that this technique consistently estimates the set of predictors in Gaussian linear models. Unfortunately, this method is computationally very expensive since it involves an exhaustive search over all possible subsets of predictors. Alternative stepwise procedures suited for ultra-high dimensions are proposed. To outline their performance, we apply our estimators both on simulated data and a large DNA methylation data set, which has been measured using the Illumina 450k array.

2 The method

Let $\mathcal{V}$ and $\mathcal{R}$ denote the standard distance covariance and distance correlation, respectively [6].
The *affinely invariant distance covariance* and *affinely invariant distance correlation*, respectively are defined as [6, 2]

$$\widetilde{\mathcal{V}}(X,Y) = \mathcal{V}(\Sigma_X^{-1/2}\, X, \Sigma_Y^{-1/2}\, X), \quad \widetilde{\mathcal{R}}(X,Y) = \mathcal{R}(\Sigma_X^{-1/2}\, X, \Sigma_Y^{-1/2}\, X),$$

where Σ_X and Σ_Y denote the covariance matrices of X and Y respectively.

For a Gaussian linear model, both $\widetilde{\mathcal{V}}(X,Y)$ and $\widetilde{\mathcal{R}}(X,Y)$ can be expressed as continuous, monotone functions of the *multiple correlation coefficient* $R(X,Y)$. This motivates the use of these measures for variable screening.

References

[1] Li, R., Zhong, W., and Zhu, L. (2012). *Feature screening via distance correlation learning.* Journal of the American Statistical Association, 107(499), 1129-1139.

[2] Dueck, J., Edelmann, D., Gneiting, T., and Richards, D. (2014). *The affinely invariant distance correlation.* Bernoulli, 20(4), 2305-2330.

[3] Fan, J., and Lv, J. (2008). *Sure independence screening for ultrahigh dimensional feature space.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(5), 849-911.

[4] Zhong, W., and Zhu, L. (2015). *An iterative approach to distance correlation-based sure independence screening.* Journal of Statistical Computation and Simulation, 85(11), 2331-2345.

[5] Yenigün, C. D., and Rizzo, M. L. (2015). *Variable selection in regression using maximal correlation and distance correlation.* Journal of Statistical Computation and Simulation, 85(8), 1692-1705.

[6] Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). *Measuring and testing dependence by correlation of distances.* The Annals of Statistics, 35(6), 2769-2794.

Spatial change point models for automatical detection of tumors in CT scans

Session: **Analysis, Testing and Change Detection in High Dimensions 1**

Philipp Otto *Department of Statistics, European University Viadrina, Frankfurt (Oder), Germany*
Wolfgang Schmid *Department of Statistics, European University Viadrina, Frankfurt (Oder), Germany*

Abstract: We discuss the detection of tumors in computed tomography scans using a new spatial change-detection procedure. In particular, a test procedure is proposed to detect change points of multidimensional autoregressive processes. The considered process differs from typical applied spatial autoregressive processes in that it is assumed to evolve from a predefined centre into every dimension. Additionally, structural breaks in the process can occur at a certain distance from the predefined center. That distance can be used to measure the size of the tumor. Moreover, the cancerous disease is automatically classified into a certain stage according to the TNM staging system (T-component). The proposed test procedure is based on the likelihood-ratio approach. We show that the generalized Gumbel distribution seems to be a suitable limiting distribution of the proposed test statistic.

1 Introduction

CT scans are of high clinical relevance for medical diagnoses. The voxels of the CT can be seen as a random process in the three-dimensional space of integers. One important issue is that the random process will have a specific three-dimensional mean due to the anatomical structure, i.e., voxels in the area of bones have a different mean than voxels in the area of soft tissue.
The main aim of our paper is to detect structural breaks of such processes, which occur at an unknown point in space. Therefore, we consider the pulmonary carcinoma as a random process in a three-dimensional space having some centre of origin. The cancer tissue spreads out from the centre in every direction. Our procedure allows us to detect structural breaks in this process, i.e., to detect the distance from the centre, where the cancer ends. Moreover, we propose a testing procedure to check whether the process has changed within an arbitrarily chosen distance from the centre. In addition, the distance, where the change point occurs, is estimated and can be used for the classification of the tumour size according to the TNM staging system.

2 Spatial Change-Point Model

Let $\{Y(s) \in \mathbb{R}^p : s \in D_s\}$ with $D_s \subset \mathbb{R}^q$ be a p-dimensional stochastic process. D_s denotes the region of interest, and s represents a specific location. The resulting process is a spatial process in the plane if q is chosen to be equal to 2. For $q = 3$, the process lies in a three-dimensional space. It is worth noting that this definition also covers spatiotemporal settings, as one could consider the temporal dimension as one dimension of $\mathbb{R}^q$.
Let $\{X(s) : s \in D_s\}$ be the observed process. If an arbitrary change

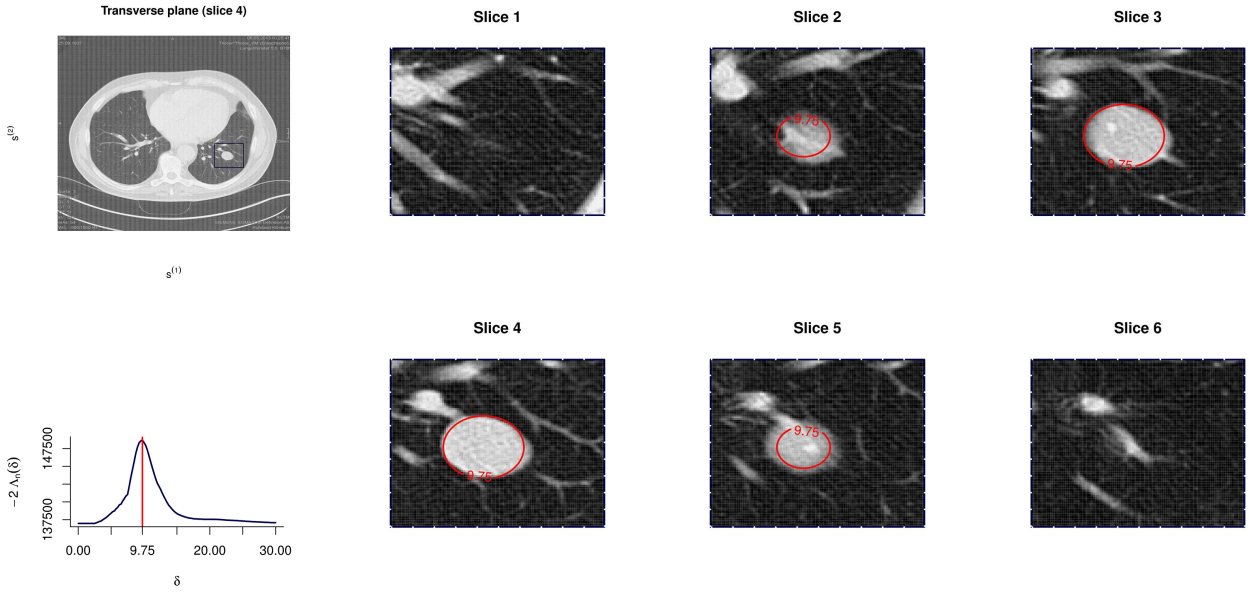


Figure 1: Transverse CT scans of the lungs (left, above) and analysed details of six parallel sections (right, slice 1 - 6), including the detected change point $\hat{\delta} = 9.75$ coloured in red. Moreover, the value the likelihood ratio $-2\Lambda_n$ is plotted against the possible values of δ (left, below).

$a = (a_1, \ldots, a_p)' \in \mathbb{R}^p \setminus \{0\}$ in the mean parameter occurs at distance δ, then the observed process can be specified as

$$X(s) = \begin{cases} Y(s) & \text{if } D(s) < \delta \\ a + Y(s) & \text{if } D(s) \geq \delta \end{cases} \tag{1}$$

for each location s. Furthermore, there may be changes in the autoregressive parameter of the process as well. The results of this setting can be found in [1]. It is worth noting that changes in the autoregressive parameters lead to changes in the covariance matrix of the process.
In Figure 1, the results of the change detection test are illustrated. In particular, the full CT scan in the transverse plane as well as the analysed details are shown. The estimated spatial change point $\hat{\delta} = 9.75$ is printed as bold line.

References

[1] Otto, P., and Schmid, W. (2016). Detection of spatial change points in the mean and covariances of multivariate simultaneous autoregressive models, *Biometrical Journal*, 58(5), pp. 1113–1137

Dimension reduction with partially integrated penalized likelihood

Session: **Analysis, Testing and Change Detection in High Dimensions 1**

Piotr Sobczyk *Wrocław University of Science and Technology, Poland*
Małgorzata Bogdan *Wrocław University, Poland*
Josse Julie *Ecole Polytechnique, France*

Abstract: We consider approximate Bayesian approach for automatic choice of the number of principal components in PCA. Because the number of parameters in PCA grows with the number of observations n and the number of variables p, it is impossible to directly apply Bayesian Information Criterion to a full fixed effect model. However, in case when only n or only p is large, one can specify and integrate out the prior on the respective subset of parameters and use the Laplace approximation for the resulting marginal likelihood. We consider both cases when n or p $\to \infty$ and define the respective model selection criteria called PEnalized SEmi-integrated Likelihood (PESEL).

1 Introduction

Reducing dimension of the data is crucial for performing statistical analysis of small populations. PCA is a very popular method used for this task, it requires however a way for choosing a number of truly important dimensions, the rate of dimension reduction. This is a non trivial problem and there are a number of different techniques to deal with it. In their paper [1] Tipping and Bishop proposed parametric model for PCA called Probabilistic Principal Component Analysis. This parametric framework still lacked a clear way of estimating dimensionality. In [2] Bishop extended it by proposing a full Bayesian approach to resolve this issue. As exact computation of MAP estimators was not possible, he calculated an approximated solution. Later in [3] Minka chose different priors than [2] and, instead of integrating out all parameters in likelihood, he approximated posterior probability getting a criterion called Laplace evidence.

2 The method

We present a new method for dimensionality reduction via PCA. Consider fixed effect model

$$\mathbf{t}_n = \mathbf{W}\mathbf{x}_n + \epsilon_n, \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2 I), \quad n = 1, \dots N,$$

where $\mathbf{t}_n$ in n^{th} observation, $\mathbf{x}_n$ in n^{th} factor and $\mathbf{W}$ is a matrix of coefficients. We aim for estimating factor dimension. To do that we use Bayesian approach. Specific prior is imposed for coefficients in matrix $\mathbf{W}$ rather than factors $\mathbf{x}_n$, as is the case in PPCA. $w_{\cdot,i} \sim \mathcal{N}(0, I)$. Different and insightful way of seeing this model is that each variable is a linear combination of random factors plus some noise. Maximum likelihood estimator we get for matrix $\mathbf{X}$ are principal components for covariance matrix of transposed data.

We require some regularity for remaining priors and, using Laplace approximation, we get

$$\int p(D|\mathbf{X},\mu,\sigma^2)d\mathbf{X}d\mu dv \approx \log p(D|\hat{\mathbf{X}},\hat{\mu},\hat{\sigma^2}) - \log(\#\text{ of variables})\frac{\#\text{ of parameters}}{2}$$

Our method proved to be effective way of reducing dimensionality. The asymptotics we use for Laplace approximation is with number of variables $p \to \infty$ while number of observations is constant. This is the setting one observes for small populations studies.
Method is implemented in the `R`package *varclust*, paper is available on arxiv [4].

References

[1] Michael E. Tipping and Chris M. Bishop, Probabilistic Principal Component Analysis, Journal of the Royal Statistical Society, Series B, 1999

[2] Bishop, Christopher M., Bayesian PCA, Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II, 1999

[3] Thomas P. Minka, Automatic choice of dimensionality for PCA, NIPS, 2000

[4] Bayesian dimensionality reduction with PCA using penalized semi-integrated likelihood, Piotr Sobczyk, Malgorzata Bogdan, Julie Josse, 2016, on arxiv (`http://arxiv.org/abs/1606.05333`)

A conditional copula-based technique to impute complex multivariate dependent data

Session: **Multivariate Distributions and Copula 1**

F. Marta L. Di Lascio *Faculty of Economics and Management, Free University of Bozen-Bolzano, Italy*
Simone Giannerini *Department of Statistical Sciences, University of Bologna, Italy*

Abstract: We present a copula-based imputation method that allows the imputation of multivariate missing data with generic patterns and complex dependence structure. We describe the basic idea of the method and provide a summary of the empirical analysis performed.

1 Introduction

Missing data occur in almost all the surveys and data collections no matter the field of application. Restricting the analysis to complete cases, i.e. observations without any missing values, is the simplest solution but leads to severe loss of information that might invalidate any inferential process. Hence, missing data are commonly treated by imputation methods that fill in missing data with plausible values. The literature of imputation methods is very large and the choice of the most appropriate method depends on many elements. We propose a method that can be used when it is important to preserve the multivariate dependence structure of the data generating process.

2 The method

The imputation method, called CoImp [1, 2], is based on the copula function [4] and makes it possible to impute multivariate missing data with generic patterns and complex dependence structure. The CoImp is a stochastic single imputation method and employs conditional density functions of the missing variables given the observed ones to fill in each multivariate missing value.
Suppose we have p continuous variables $X_1, \ldots, X_j, \ldots, X_p$ with distribution functions $F_1, \ldots, F_j, \ldots, F_p$ and densities $f_1, \ldots, f_j, \ldots, f_p$, such that their probability integral transforms are $U_1 \sim F_1(X_1), \ldots, U_j \sim F_j(X_j), \ldots, U_p \sim F_p(X_p)$, respectively. Further, assume that for some records X_j is missing whereas the remaining $p-1$ variables are observed. We derive the conditional density function $f_j(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)$ through (i) the canonical representation of the joint density via the density copula $c(\cdot)$ (see [5]):

$$f(x_1, \ldots, x_j, \ldots, x_p) = c(F_1(x_1), \ldots, F_j(x_j), \ldots, F_p(x_p)) \prod_{j=1}^{p} f_j(x_j). \qquad (1)$$

and (ii) the conditional copula density $c(u_j|u_1, \ldots, u_{j-1}, u_{j+1}, \ldots, u_p)$ defined by using Bayes' rule (see [5], p.89) so that:

$$f(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p) = c(u_j|u_1, \ldots, u_{j-1}, u_{j+1}, \ldots, u_p) f_j(x_j). \qquad (2)$$

Equation (2), which can be easily extended to the case of multivariate missing vari-

ables, is derived analytically once parametric models for the margins and the copula are specified. When analytical derivations are not feasible, due to the complexity of the distributions and/or the high-dimensionality of the data generating process, it is possible to follow a semi-parametric approach. In the latter case, the margins can be modelled non-parametrically by means of a localized version of the log-likelihood function [3] and the dependence parameter θ of the copula model c is estimated on the available data by the pseudo maximum log-likelihood method (for technical details see [1]). Missing data are imputed by drawing observations from the conditional densities in (2) by means of a Monte Carlo method like the Hit or Miss described in [2].

3 Results

The method has been developed both when it is possible to derive analytically the conditional densities used to perform the imputation and when this is not possible by using a semi-parametric approach. The performance of the two solutions has been investigated by means of a large simulation study. The results show that (i) the analytical version compares favourably with classical methods (see [2] for details) and (ii) the performance of the semi-parametric approach is very similar to that of the analytic approach (see [1] for details). The CoImp method has a wide range of applicability and has been implemented in an R software package called `CoImp` available on the CRAN at `http://cran.r-project.org/web/packages/CoImp/index.html`.

References

[1] Di Lascio, F.M.L. and Giannerini, S. (2014). Imputation of complex dependent data by conditional copulas: analytic versus semiparametric approach, *Book of proceedings of the 21st International Conference on Computational Statistics (COMPSTAT 2014)*, p. 491-497.

[2] Di Lascio, F.M.L., Giannerini, S. and Reale, A. (2015). Exploring copulas for the imputation of complex dependent data, *Statistical Methods & Applications*, 24(1), p. 159-175.

[3] Loader, C.R. (1996). Local likelihood density estimation, *The Annals of Statistics*, 24(4), 1602–1618.

[4] Sklar, A. (1959). Fonctions de répartition à n dimensions et leures marges, *Publications de l'Institut de Statistique de L'Université de Paris*, 8, p. 229–231.

[5] Trivedi, P.K., Zimmer, D.M. (2005). *Copula Modeling: An Introduction for Practitioners*, Found. Trends Econom., 1(1), p. 1-111.

Identifying misspecified tail-dependence / Bernoulli matrices in high-dimensions via linear programming

Session: **Multivariate Distributions and Copula 1**

Daniel Krause *Lehrstuhl für Finanzmathematik, Technische Universität München, Germany*
Matthias Scherer *Lehrstuhl für Finanzmathematik, Technische Universität München, Germany*
Jonas Schwinn *Institut für Mathematik, Universität Augsburg, Germany*
Ralf Werner *Institut für Mathematik, Universität Augsburg, Germany*

Abstract: In the context of market-, credit-, and operational risk, stochastic models allowing for tail dependence are considered indispensable in modern risk-management. Being difficult to estimate, it is often a matter of expert judgment to define a matrix of pairwise tail-dependence coefficients. Given such a matrix, however, it is rather difficult to decide if (i) the given matrix is indeed a possible tail-dependence matrix, and (ii) how a stochastic model can be constructed representing it. These problems, and the one-to-one connection to Bernoulli matrices, has been studied on a theoretical level, but efficient numerical tests are rare. We add to the existing literature by exploiting the polyhedral geometry of the set of Bernoulli matrices. This allows to translate the above questions into a linear optimization problem with exponentially many variables. Moreover, we partially overcome the curse of dimensionality by a specific column generation ansatz.

Limiting properties of the modified Conditional Value-at-Risk (CoVaR)

Session: **Multivariate Distributions and Copula 1**

Piotr Jaworski *Institute of Mathematics, University of Warsaw, Poland*

Abstract:
CoVaR (conditional Value at Risk) is a newly introduced risk measure which is oriented on systemic risk. Assume that we are measuring the risk basing on some extra information. For example, we want to determine what size bailout would be required to keep a financial institution **Y** solvent with probability at least $1-\beta$ when a financial institution **X** would incur significant losses. Conditional Value at Risk (CoVaR) introduced by Adrian and Brunnermeier in 2008 ([1]) and its later modifications by Mainik and Schaanning in [4] proved to be very useful tools to deal with such phenomena.
If random variables X and Y are modelling our phenomena, say welfares of banks or gains from the investments, modified CoVaR of Y with respect to X is VaR of conditional Y under the condition that X is below the threshold. In more details

$$CoVaR_{\alpha,\beta}(Y|X) = VaR_{\beta}(Y|X \leq -VaR_{\alpha}(X)).$$

The above can be expressed in terms of quantiles

$$CoVaR_{\alpha,\beta}(Y|X) = -Q^{+}_{\beta}(Y \mid X \leq Q^{+}_{\alpha}(X))$$

and furthermore in terms of copulas

$$CoVaR_{\alpha,\beta}(Y|X) = -Q^{+}_{w_*}(Y),$$

where the threshold $w_* = w_*(\alpha, \beta, C)$ depends only on the linking copula C and significance levels α and β ([2, 3]). In my talk I will discuss limiting properties of the implied threshold w_* when α and β are close to the origin. Presented results will be illustrated on the example of several families of copulas, like survival extreme value, left truncation invariant, Archimedean, t-Student, Gaussian, FGM and Marshall-Olkin copulas.

References

[1] Adrian, T., Brunnermeier, M.K. (2016). CoVaR, *The American Economic Review* **106.7**, 1705-1741.

[2] Bernardi,M., Durante, F., Jaworski, P. (2017) CoVaR of families of copulas. *Statistics and Probability Letters* **120**, 8-17.

[3] Jaworski,P. (2017) On the Conditional Value at Risk (CoVaR) for tail-dependent copulas, *Dependence Modeling to appear.*

[4] Mainik G., Schaanning E (2014). On dependence consistency of CoVaR and some other systemic risk measures. *Stat. Risk Model.*, **31**, 49–77.

Statistical inference for the fractional stable motion

Session: **Statistics of Stochastic Processes 1**

Mark Podolskij *Department of Mathematics, Aarhus University, Denmark*
Stepan Mazur *Department of Mathematics, Aarhus University, Denmark*

Abstract: In this talk we investigate the parametric inference for the linear fractional stable motion in high and low frequency setting. The symmetric linear fractional stable motion is a three-parameter family, which constitutes a natural non-Gaussian analogue of the scaled fractional Brownian motion. It is fully characterised by the scaling parameter $\sigma > 0$, the self-similarity parameter $H \in (0, 1)$ and the stability index $\alpha \in (0, 2)$ of the driving stable motion. The parametric estimation of the model is based upon the limit theory for stationary increments Lévy moving average processes. More specifically, we combine power variation statistics and empirical characteristic functions to obtain consistent estimates of (σ, α, H). We present the law of large numbers and fully feasible central limit theorems.

The discontinuous leverage effect in contemporaneous price and volatility jumps

Session: **Statistics of Stochastic Processes 1**

Markus Bibinger *Faculty of Mathematics and Computer Science, Philipps-Universität Marburg*
Christopher Neely *Research Division, Federal Reserve Bank of St. Louis*
Lars Winkelmann *Department of Economics, Freie Universität Berlin*

Abstract: We discuss nonparametric inference for price and volatility jump sizes in high-frequency data with market microstructure noise. Local tests for price jumps and volatility jumps are presented as well as their estimation under the alternative that jumps occur. The methods facilitate a new way to assess and test for a (tail) discontinuous leverage effect understood as the discontinuous part of the quadratic covariation between logarithmic price and volatility by contemporaneous price and volatility jumps.

We conduct an empirical study based on five years of NASDAQ-transaction data from 360 stocks.

1 Statistical model

We work with noisy discrete observations $Y_{t_i}, i = 0, \dots, n$, $Y_t = X_t + \epsilon_t$ on $[0,1]$ of an Itô semimartingale $(X_t)_{t\in[0,1]}$ with latent volatility $(\sigma_t)_{t\in[0,1]}$. The process ϵ_t captures market microstructure frictions and is centered, $\mathbb{E}[\epsilon_t] = 0$. With the usual notation for jumps $\Delta X_t = X_t - X_{t-}$, where $X_{t-} = \lim_{s<t, s\to t} X_s$, one quantity of interest in this talk is

$$[X, \sigma^2]_T^d(a) = \sum_{s\le T} \Delta X_s \big(\sigma_s^2 - \sigma_{s-}^2\big)\, \mathbf{1}_{\{|\Delta X_s| > a\}} \tag{1}$$

which is introduced as the *tail discontinuous leverage effect* in [1].

2 Statistical methods

Consider some specified time $\tau \in (0,1)$. From noisy observed returns we can not directly see if $\Delta X_\tau > 0$ and of course neither if $\Delta\sigma_\tau^2 > 0$. We discuss statistical devices

- to test $\Delta X_\tau = 0$ vs. $\Delta X_\tau \neq 0$;
- to estimate ΔX_τ;
- to test $\Delta\sigma_\tau^2 = 0$ vs. $\Delta\sigma_\tau^2 \neq 0$;
- to estimate $\Delta\sigma_\tau^2$.

Considering jumps of finite activity or $a > 0$ in (1), related to [4] and [2], these local methods facilitate a statistical test for the hypothesis $[X, \sigma^2]_T^d(a) = 0$ against $[X, \sigma^2]_T^d(a) \neq 0$ and estimation of the (tail) discontinuous leverage effect.

3 Empirical evidence from the NASDAQ order book

We study contemporaneous price and volatility jumps of 360 individual NASDAQ stocks from 2010 to 2015. In line with the previous literature, we find mixed and

mostly insignificant discontinuous leverage effects across stocks when using all detected price and volatility jumps. We show that the event specific nature and distinct sources of jumps obscures the correlation between price and volatility jumps. We find, however, strong and significant discontinuous leverage effects across firms studying positive and negative price jumps separately and conditioning on market-wide vs. idiosyncratic jumps.

Our findings have some implications for parametric asset prices models. In particular, some prominent models including relations between price and volatility jump are rejected by the empirical findings.

Our results suggest that the relation between price and volatility jumps is quite different to the generally negative leverage effect of the continuous components. News impact models by [5] and systematic volatility studied by [3] seem more appropriate to describe the driving force behind simultaneous prices and volatility jumps. We find a positive correlation between price and volatility jumps for market-wide upward price jumps and a negative correlation for market-wide downward price jumps. Our analysis reveals that some stocks exhibit a stronger relation between price and volatility jumps than other stocks.

References

[1] Aït-Sahalia,Y., Fan, J., Laeven, R.J.A., Wang, C.D. and Yang, X. (2016) *Estimation of the continuous and discontinuous leverage effect*, forthcoming in Journal of the American Statistical Association.

[2] Bibinger, Winkelmann (2016) *Common price and volatility jumps in noisy high-frequency data*, preprint, arxive: 1407.4376.

[3] Cremers, M., Halling, M. and Weinbaum, D. (2015) *Aggregate jump and volatility risk in the cross-section of stock returns*, Journal of Finance 70, 577–614.

[4] Jacod, J., Klüppelberg, C. and Müller, G. (2016) *Testing for non-correlation between price and volatility jumps*, forthcoming in Journal of Econometrics

[5] Pástor, L. and Veronesi, P. (2012) *Uncertainty about government policy and stock prices*, Journal of Finance 67, 1219–1264.

Volatility estimation for stochastic PDEs using high-frequency observations

Session: **Statistics of Stochastic Processes 1**

Markus Bibinger *Fachbereich Mathematik und Informatik, Philipps-Universität Marburg, Germany*
Mathias Trabs *Fachbereich Mathematik, Universität Hamburg, Germany*

Abstract: We study the parameter estimation for parabolic, linear, second order, stochastic partial differential equations observing a mild solution on a discrete grid in time and space. A high-frequency regime is considered where the mesh of the grid in the time variable goes to zero. Focusing on volatility estimation, we provide a simple and easy to implement method of moments estimator based on the squared increments of the process. The estimator is consistent and admits a central limit theorem. It should be emphasized that the theory considerably differs from the statistics for semi-martingale literature. The performance of the method is illustrated in a simulation study.

1 Introduction and model

Nowadays models based on stochastic partial differential equations (SPDEs) become increasingly popular, like the stochastic Navier–Stokes equation to describe the motion of fluid substances, SPDE-models for neuronal systems [1] or in financial applications where, for instance, SPDEs are used to model interest rate fluctuations like the yield curve model in [2].
In probability theory there are very recently enormous efforts to advance research for SPDEs. In contrast, statistical inference for SPDEs is a relatively young research field, cf. [3] for a survey paper on the existing results.
We will study the following linear parabolic SPDE (with one space dimension)

$$\begin{aligned} \mathrm{d}X_t(y) &= \Big(\theta_2 \frac{\partial^2 X_t(y)}{\partial y^2} + \theta_1 \frac{\partial X_t(y)}{\partial y} + \theta_0 X_t(y)\Big)\,\mathrm{d}t + \sigma\,\mathrm{d}B_t(y), \qquad (1)\\ &\forall t \geq 0, y \in [0,1], \quad X_t(0) = X_t(1) = 0, \end{aligned}$$

where B_t is defined as a cylindrical Brownian motion in the Sobolev space on $[0,1]$ and with parameters $\theta_0, \theta_1 \in \mathbb{R}$ and $\theta_2 > 0$ and some volatility parameter σ. Although (1) is a relatively simple SPDE model, it already covers many interesting applications, including the *stochastic heat equation*, the *cable equation* as a basic PDE-model in neurobiology and the *term structure model* in [2].
The scenario where the solution of the SPDE is observed only at discrete points in time and space, say on the gird $(t_i, y_j), i = 0, \ldots, n, j = 1, \ldots, j$, is, to the best of the authors' knowledge, only considered in [4] who has derived asymptotic normality and efficiency for the maximum likelihood estimator in a parametric problem. In contrast to [4] we consider a high-frequency regime setting $t_i = i/n$ for $i = 0, \ldots, n$ and let $n \to \infty$. The number of spatial observations m may be fixed or go to infinity.

2 Method and results

We focus on estimating σ^2, assuming for simplicity that $\theta = (\theta_0, \theta_1, \theta_2)$ is known. Volatility estimation relies typically on squared increments of the observed process. We develop a method of moment estimator utilizing squared increments $(X_{t_i} - X_{t_{i-1}})^2(y_j), 1 \leq i \leq n$, also. However, the behavior of the latter is quite different compared to the standard setup of high-frequency observations of semi-martingales: Since X can be understood as infinite dimensional stochastic differential equation, the increment $\Delta_i X := X_{t_i} - X_{t_{i-1}}$ is given as a weighted mean of infinitely many independent Ornstein-Uhlenbeck processes with growing mean reversion rate.
In view of the complex model, our asymptotic analysis brings forth a surprisingly simple estimator for the volatility

$$\hat{\sigma}^2_{n,m} := \frac{\sqrt{\pi\vartheta_2}}{\sqrt{n}\, m} \sum_{i=1}^{n} \sum_{j=1}^{m} (\Delta_i X)^2(y_j) e^{y_j\, \vartheta_1/\vartheta_2}.$$

Up to a bias correction we show that the mean squared error of this estimator is of order $(nm)^{-1}$. In particular, $\hat{\sigma}^2_{n,m}$ is already consistent for $m = 1$. We moreover proof a central limit theorem that allows for the construction of confidence intervals. We briefly touch time dependent volatility and the estimation of θ. The numerical performance of the method is demonstrated in simulations.

References

[1] Tuckwell, H. C. (2013). Stochastic partial differential equations in neurobiology: Linear and nonlinear models for spiking neurons. In *Stochastic Biomathematical Models*, pages 149–173. Springer.

[2] Cont, R. (2005). Modeling term structure dynamics: an infinite dimensional approach. *Int. J. Theor. Appl. Finance*, 8(3):357–380.

[3] Lototsky, S. V. (2009). Statistical inference for stochastic parabolic equations: a spectral approach. *Publ. Mat.*, 53(1):3–45.

[4] Markussen, B. (2003). Likelihood inference for a discretely observed stochastic partial differential equation. *Bernoulli*, 9(5):745–762.

Nonparametric Bernstein–von Mises theorems for discretely observed compound Poisson processes

Session: **Statistics of Stochastic Processes 1**

Richard Nickl *Statistical Laboratory, University of Cambridge, UK*
Jakob Söhl *Delft Institute of Applied Mathematics, TU Delft, The Netherlands*

Abstract: We show a Bernstein–von Mises theorem for discretely observed compound Poisson processes. The compound Poisson processes are observed at low frequency, i.e., the distance between the observations is fixed while more and more observations are accrued as the observation time increases. The inference is on the Lévy density which is model by a wavelet series prior. The Bernstein–von Mises theorem holds in the bounded Lipschitz metric for weak convergence in a multi-scale space. Along the way a posterior contraction rate for the Lévy density is derived. This is done by showing concentration of frequentist estimators and verifying the small ball probability condition.

Model Validation and Diagnostics for Discrete-Valued Time Series

Session: **Discrete-Valued Time Series 1**

Robert Jung *Department of Economics, University of Hohenheim, Germany*

Abstract: Checking the adequacy of a specified model is an important part of any iterative modelling exercise in applied time series analysis. For linear Gaussian time series models, or those based on the framework of Generalized Linear Models, there exist well-developed tools for this purpose that are readily available and routinely employed in applied work. However, for time series models for discrete-valued data, this is not the case. Nevertheless, the need to compare two or more competing model specifications, or evaluate the adequacy of fit of a chosen model is obvious. To help address this gap, a range of diagnostic and model validation methods designed to lead to data coherent models that achieve good probabilistic forecasting outcomes are surveyed.

A test procedure on the compounding structure of CP-INARCH models

Session: **Discrete-Valued Time Series 1**

Christian H. Weiß *Department of Mathematics and Statistics, Helmut Schmidt University, 22008 Hamburg, Germany.*
Esmeralda Gonçalves *CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal*
Nazaré Mendes-Lopes *CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal*

Abstract: To distinguish between a simple Poisson INGARCH and a true compound Poisson (CP) INGARCH model, we develop a test based on the probability generating function of the compounding. For the particular case of an INARCH(1) process, the test statistics' asymptotic normal distribution is derived, either in the case, where the model parameters are specified, or in that one, important in practice, where such parameters are consistently estimated. A simulation study illustrating the performance of this test methodology concludes the paper.

1 Introduction

The INGARCH models have known, since their introduction by [3, 1], great extension and development namely through the assumption of new conditional distributions in alternative to the Poisson one, initially considered by those authors. Recently, [2] introduced a wide class of this type of models, the CP-INGARCH with compound Poisson conditional distribution, which includes the main INGARCH models present in literature and, particularly, the simple Poisson INGARCH ones. In order to contribute to the distinction between a simple Poisson INGARCH and a true CP-INGARCH model, we develop a test based on the form of the probability generating function (pgf) of the compounding distribution related to the model conditional law.

2 Our approach

We consider the CP-INARCH model defined by

$$\begin{aligned} \mathrm{pgf}_{X_t|X_{t-1},\ldots}(z) \;&=\; \exp\Big(\tfrac{M_t}{H'(1)}\big(H(z)-1\big)\Big) \\ \text{with } M_t \;&:=\; E[X_t \mid X_{t-1},\ldots] \;=\; \alpha_0 + \textstyle\sum_{i=1}^{p}\alpha_i X_{t-i}, \end{aligned}$$

where $H(z)$ denotes the pgf of the compounding distribution. So the testing problem becomes

$$H_0 : \text{Poisson INARCH with } H(z)=z; \quad H_1 : \text{CP-INARCH with } H(z)\neq z.$$

Given the past $X_{t-1},\ldots$, the conditional CP model implies that first a stopping count N_t is generated according to $\mathrm{Poi}\big(M_t/H'(1)\big)$, and then (independently) the N_t i.i.d. counts $Y_{t,1},\ldots,Y_{t,N_t}$ according to the compounding model with pgf $H(z)$. The next observation is obtained as $X_t = Y_1 + \ldots + Y_{N_t}$.
To distinguish between the null hypothesis H_0 and the alternative hypothesis H_1, information about $H(z)$ is required, the unique pgf of the $Y_{t,i}$. In fact, it suffices

to check if the mean $H'(1)$ of the compounding distribution is equal to 1 (H_0) or larger than 1 (H_1). Hence, the mean statistic

$$\frac{1}{T}\sum_{t=1}^{T}\frac{Y_{t,1}+\ldots+Y_{t,N_t}}{N_t} \;=\; \frac{1}{T}\sum_{t=1}^{T}\frac{X_t}{N_t}$$

would be a reasonable candidate to infer $H'(1)$. But we do not observe N_t in practice, we only know that it has mean $M_t/H'(1)$. Therefore, we consider

$$\widehat{C}_{p;\,1} \;:=\; \frac{1}{T-p}\sum_{t=p+1}^{T}\frac{X_t}{M_t} \;=\; \frac{1}{T-p}\sum_{t=p+1}^{T}\frac{X_t}{\alpha_0+\sum_{i=1}^{p}\alpha_i X_{t-i}}.$$

Under H_1, the variance of $\widehat{C}_p$ is inflated by $1+H''(1)/H'(1)$ (while $H''(1)=0$ under H_0), but the mean of $\widehat{C}_p$ is always 1. Therefore, we consider higher-order extensions of $\widehat{C}_p$ such that also the mean is affected if violating H_0:

$$\widehat{C}_{p;\,r} \;:=\; \frac{1}{T-p}\sum_{t=p+1}^{T}\frac{(X_t)_{(r)}}{M_t^r} \;=\; \frac{1}{T-p}\sum_{t=p+1}^{T}\frac{X_t(X_t-1)\cdots(X_t-r+1)}{\left(\alpha_0+\sum_{i=1}^{p}\alpha_i X_{t-i}\right)^r}.$$

3 Outline

For the particular case of an INARCH(1) process, the normality of the test statistics' asymptotic distribution is established either in the case, where the model parameters are specified, or in that one, important in practice, where such parameters are consistently estimated. Involving the test statistics' inverse conditional moments of CP-INGARCH process, the analysis of their existence, calculation or estimation constitutes a relevant aspect of this study. A simulation study illustrating the performance of this test methodology concludes the paper.

References

[1] Ferland, R., Latour, A., Oraichi, D. (2006), Integer-valued GARCH processes. *J. Time Ser. Anal.*, 27(6), 923–942.

[2] Gonçalves, E., Mendes-Lopes, N., Silva, F. (2015), Infinitely divisible distributions in integer-valued GARCH models. *J. Time Ser. Anal.*, 36(4), 503–527.

[3] Heinen, A. (2003), Modelling time series count data: an autoregressive conditional Poisson model. CORE Discussion Paper, 2003-63, University of Louvain, Belgium.

Model-free off-line change-points detection in vector time series via ϵ-complexity: theoretical results.

Session: **Analysis, Testing and Change Detection in High Dimensions 2**

Boris Darkhovsky *Institute for Systems Analysis FRC CSC RAS, Higher School of Economics, Moscow, Russia*

Alexandra Piryatinska *Department of Mathematics, San Francisco State University, U.S.A.*

Abstract: We propose fundamentally new methodology of change-points detection (in off-line statement) for *vector time series of arbitrary nature*. The main advantage of our methodology is the fact that it *does not use any model* of the vector time series, and so it is *model-free*. In this first part of our talk we will present the theoretical results concerning our new theory of *ϵ-complexity of continuous vector-functions*. This theory is the foundation of our methodology.

1 Introduction

We consider vector time series with possible changes in generating mechanism. In case when collected data are generated by some probabilistic mechanism, this problem is a well-known *change-point detection problem* for stochastic processes.
In econometric literature "change points" are called *"structural breaks"*. Recently in this area a significant attention is given to detect change points in so called *panel data*. In panel data the problem of detection of changes in mathematical expectation of *vector time series* is considered. The authors of these papers study the asymptotic properties of the corresponding estimates when cumulative value of changes in mathematical expectation (by all components) has certain relation with the sample size.
We consider time series with fixed number of components, and our methodology does not use assumptions about connection between the number of components and the sample size. But main distinctive feature of our methodology: *we do not use any model of time series.*

2 Main Definitions

Let $x(\cdot) : \mathbb{I} = [0,1] \to \mathbb{R}^d$ be a continuous vector-function. Let $R_i = \max_{t\in\mathbb{I}} |x_i(t)|, i \in I = \{1,\dots,d\}$. We will assume that $\min_{i\in I} R_i > 0$.
Let $\hat{x}_i(\cdot)$ be an approximation of the i-th component $x_i(\cdot)$, $i \in I$ of vector-function $x(\cdot)$ constructed by its values at the nodes of a uniform grid with spacing h by one of the allowable methods of function reconstruction from some given collection $\mathcal{F}$.
Put $\delta_i^{\mathcal{F}}(h) = \inf_{\hat{x}_i(\cdot)\in\mathcal{F}} \sup_{t\in\mathbb{I}} |x_i(t) - \hat{x}_i(t)|$, $i \in I$. The function $\delta_i^{\mathcal{F}}(h)$ is called *absolute recovery error of component $x_i(\cdot)$ by methods $\mathcal{F}$.*
The function $x_i(\cdot)$ is called *$\mathcal{F}$-nontrivial* if it can not be recovered with an arbitrary small error by methods $\mathcal{F}$ for any $h > 0$.
Analogously, denote $\delta_i(h) = \inf_{\tilde{x}_i(\cdot)} \sup_{t\in\mathbb{I}} |x_i(t) - \tilde{x}_i(t)|$, where $\tilde{x}_i(\cdot)$ is an *arbitrary computable estimate* of function $x_i(\cdot)$ by its values on uniform grid with spacing h.
The function $x_i(\cdot)$ is called *totally-nontrivial* if it can not be recovered with an arbitrary small error by *any collection of computable methods* for any $h > 0$.
Vector-function $x(\cdot)$ is called *$\mathcal{F}$-nontrivial (correspondingly, totally nontrivial)* if it

has at least one $\mathcal{F}$-*nontrivial (correspondingly, totally nontrivial)* component. The set of totally nontrivial functions is *dense* in $C(\mathbb{I},\mathbb{R})$. Therefore, in any neighborhood of zero in $C(\mathbb{I},\mathbb{R}^d)$ there exists a vector-function with *all totally nontrivial components*. So, the set of vector-functions with all totally nontrivial components is a *generic set* (i.e., "small perturbation" of arbitrary continuous map makes all its components totally non-trivial). Therefore, we can say that *"almost any"* continuous vector-function has all its components *totally nontrivial*.
Denote

$$\tilde{I} = \{i \in I : x_i(\cdot) \,\text{is}\, \mathcal{F} - \text{nontrivial function}\}$$

Put $(\forall \epsilon \geq 0)$

$$h_x^*(\epsilon,\mathcal{F}) = \begin{cases} \inf\{h \leq 1 : \sum_{i\in\tilde{I}} \frac{\delta_i^{\mathcal{F}}(h)}{R_i} > \epsilon\}, & \text{if}\, \tilde{I} \neq \emptyset \\ 1, & \text{in opposite case} \end{cases}$$

We call $\frac{\delta_i^{\mathcal{F}}(h)}{R_i}$ *related recovery error of component* $x_i(\cdot)$ *by methods* $\mathcal{F}$.
The number

$$\mathbb{S}_x(\epsilon,\mathcal{F}) = -\log h_x^*(\epsilon,\mathcal{F})$$

is called the $(\epsilon,\mathcal{F})$-complexity of an individual continuous vector-function $x(\cdot)$.
This definition is in line with Kolmogorov's idea about "complexity" of objects and is a generalization of our main definition for scalar functions (see [1])

3 Main result

Taking into account that *"almost any"* continuous vector-function has all its components *totally nontrivial*, we will consider further only such vector-functions.
Let $\mathcal{T}$ be a set of vector-functions satisfying Hölder condition.

Theorem

For any vector-function $x(\cdot)$ from certain dense subset of $\mathcal{T}$ and for any (sufficiently small) $r > 0$, $\gamma > 0$, there exist real numbers $\alpha > 0$, $\Delta > 0$, $\mathbb{A}$, $\mathbb{B}$, $|\mathbb{B}| \geq b\,(x(\cdot)) > 0$, a family of approximation methods $\mathcal{F}^*$, functions $\theta(\epsilon), \zeta(\epsilon)$, and set $M \subset [\alpha, \alpha + \Delta]$ with Lebesgue measure $\mu(M) > \Delta - r$ such that the following relations hold on the set M for all families of approximation methods $\mathcal{F} \supset \mathcal{F}^*$:

$$\mathbb{S}_x(\epsilon,\mathcal{F}) = \mathbb{A} + \mathbb{B}\log\epsilon + \theta(\epsilon)\log\epsilon + \zeta(\epsilon),\ \sup_{\epsilon\in M} \max\left(|\theta(\epsilon)|, |\zeta(\epsilon)|\right) \leq \gamma.$$

References

[1] B.S.Darkhovsky and A. Piryatinska (2014). New Approach to the Segmentation Problem for Time Series of Arbitrary Nature, *Proceedings of the Steklov Institute of Mathematics*, pp.54-67, Vol.,287.

Model-free off-line change-points detection in vector time series via ϵ-complexity: simulation and application

Session: **Analysis, Testing and Change Detection in High Dimensions 2**

Boris Darkhovsky *Institute for Systems Analysis FRC CSC RAS, Higher School of Economics, Moscow, Russia*

Alexandra Piryatinska *Department of Mathematics, San Francisco State University, U.S.A.*

Abstract: This is the second part of our joint talk. In this talk we define the ϵ-complexity for vector-functions given by their values at some uniform grid, and present our model-free change-points detection methodology. Our approach contains two steps. Firstly, we utilize our concept of the ϵ-complexity to create so called "diagnostic sequence". Secondly, we apply the non-parametric change-point detection procedure for detecting of changes in mathematical expectation of the "diagnostic sequence". Some results of simulations and application to real data will be given.

Let us explain on semantic level the concept of the ϵ-complexity of a continuous vector-function given by its n values on some uniform grid.

We choose a number $0 < S < 1$ and discard $[(1-S)n]$ vector-function values. Using remaining values we approximate values of the vector-function at the discarded points utilizing the set of given approximation methods $\mathcal{F}$. For each component we find the best approximation in the sense of minimal relative recovery error. Let the minimal relative recovery error of the i-th component be equal to ϵ_i, $i \in I = \{1, \dots, d\}$. We define the relative recovery error of the vector-function as $\epsilon = \sum_{i \in I} \epsilon_i$.

The ϵ-complexity of a continuous vector-function given by its values at a uniform grid is defined as $(-\log S(\epsilon, \mathcal{F}))$, where $S(\cdot)$ is a fraction of the vector-function values which should be retained to reconstruct this vector-function in the discarded points with a relative error not large then ϵ.

The foundation of the methodology is following theorem:

Theorem. *For any vector-function $x(\cdot)$ from a dense subset of set of vector-functions satisfying Hölder condition, any (sufficiently small) $\kappa > 0, \delta > 0$, and $n \geq n_0(x(\cdot))$ there exist a set of approximation methods $\mathcal{F}^*$, numbers $0 < \alpha(n, x(\cdot)) < \beta(n, x(\cdot)) < 1$, functions $\rho(S), \xi(S)$ and a set $N \subset Q = [\alpha(\cdot), \beta(\cdot)], \mu(N) > \mu(Q) - \delta$ ($\mu(\cdot)$ is Lebesgue measure) such that for all $\mathcal{F} \supseteq \mathcal{F}^*$ and $S \in N$ the following relations hold:*

$$\log \epsilon = A + (B + \rho(S)) \log S + \xi(S), \ \sup_{S \in N} \max(|\rho(S)|, |\xi(S)|) \leq \kappa.$$

It follows from the above theorem that (in the case of sufficiently rich family of approximation methods $\mathcal{F}$ and sufficiently large n) for vector-functions satisfying Höllder condition and defined by their n values at a uniform grid the ϵ-complexity is characterized by a pair of real numbers (A, B) via the formula

$$\log \epsilon \approx A + B \log S.$$

These two parameters A, B are called the ϵ-complexity coefficients. They will be

utilized for creating diagnostic sequences to detect change points in vector time series.

Let $X = \{x(t)\}_{t=1}^{N}$ be a vector-time series with unknown change points $t_i, i = 2, \ldots, k$ (it is unknown if there are changes or not). The type of generating mechanisms are also unknown and could be stochastic, deterministic or mixed. Any segment $[t_i, t_{i+1}]$, $t_1 = 1, \ldots, t_{k+1} = N$, which is generated by the same mechanism is called *homogeneous*. We assume that homogeneous segments are sufficiently long.

Due to our Theorem the ϵ-complexity is uniquely characterized by a pair of coefficients $\mathcal{R} = (A, B)$. For given time series we separate it into disjoint intervals of length n or consider a sliding window of size n. For each interval we estimate the ϵ-complexity coefficients. As a result we obtain the *diagnostic vector sequence* $\{\mathcal{R}(j)\}_{j=1}^{[N/n]}$. Similar vector sequence of the ϵ-complexity coefficients can be also calculated for sliding windows.

The keystone of the proposed methodology for segmentation of time series into homogeneous increments is following

Conjecture. At i-st segment of homogeneity $[t_i, t_{i+1}]$ of the time series X for $t_i \leq t$, $(t + n) < t_{i+1}$ the corresponding ϵ-complexity coefficients $\mathcal{R}(j)$ satisfy the relation

$$\mathcal{R}(j) = \mathcal{R}_i + \xi_i(j),$$

where $\xi_i(j)$ is a random process with zero expectation.

Conjecture implies that the expected values of the ϵ-complexity coefficients of a vector time series are constant within segments of homogeneity.

Thus, assuming the Conjecture is true, the change-point detection problem is reduced to the detection of changes in mathematical expectation of the diagnostic vector sequence $\mathcal{R}$.

To solve this problem, we propose to use the family of statistics introduced by Brodsky and Darkhovsky in the beginning of 80th (see [2]).

The results of simulations and applications to real data will be presented.

References

[1] B.S.Darkhovsky and A. Piryatinska (2014). New Approach to the Segmentation Problem for Time Series of Arbitrary Nature, *Proceedings of the Steklov Institute of Mathematics*, pp.54-67, Vol., 287.

[2] B.E. Brodsky and B.S. Darkhovsky (2000). *Non-parametric Statistical Diagnosis: Problems and Methods*, Kluwer, Dordrecht.

Rule-based method of spike detection and suppression

Session: **Analysis, Testing and Change Detection in High Dimensions 2**

Ewaryst Rafajłowicz *Department of Control Systems and Mechatronics, Wrocław University of Science and Technology, Poland*

Wojciech Rafajłowicz *Department of Engineering Informatics, Wrocław University of Science and Technology, Poland*

Abstract: We propose a rule based method of spike detection and suppression. Its elementary properties are established and the example of application for a laser power control is discussed.

1 Introduction

The problem of spike (peak, cusp) detection attracted an increasing interest of researchers in many fields (see [3], [1], [2] for selected contributions). The problem of spike detection can formally be reduced to the problem of change detection by replacing the original sequence of observations x_i, $i = 1, 2, \ldots$ by $\kappa_n = \sum_{i=1}^{n} x_i$. However, it is not easy to suppress the heights of the detected spikes in a controllable way. Spikes can be outliers and then they should be completely removed. On the other hand, they can bear information, but their heights are too large and should be suppressed (reduced) – this is exactly the case considered in our example of laser power control. Motivated by these applications we propose the method of simultaneous spike detection and suppression (SDS). This method is an extension of the jump detector that was proposed in [4] (see also [5] and the bibliography therein).

2 Simultaneous spike detection and suppression method

The idea of SDS method is sketched in Fig. 2 – left panel. It is based on the contents of buffer (box) $B(n,\, M,\, H) \stackrel{def}{=} \{x_{n-k} : |x_n - x_{n-k}| \leq H,\, k = 0,\, 1, \ldots,\, M\}$ of the length of $M + 1$ current and past observations, having the height $2\,H > 0$ with x_n placed at the middle of the right hand side of the box. If the box is in flat areas (positions a, f), then it contains most of the M past observations that are avareged and provided as output y_n. When the number of observations is less than $M_{ke} < M$, say, then jump is signaled (positions b, c), while if the same observations would be cateched by the box (position e), then the presence of a spike is declared. In fact, the rules of SDS method are somewhat more complicated. Depending on selected combinations of M_{kr}, M and H and the growth rate of a sampled signal, SDS method preserves jumps, linear functions (see the middle plot of the right panel of Fig. 2). For more rapidly changing signals, SDS Method detects short ramps, narrow triangles (see the upper plot of the right panel of Fig. 2) and other spikes and then, it provides again y_{n-1} as the output, leading to suppression of spikes (see slightly higher horizontal line at the same plot). In a laser cladding process the laser power is controlled basing on temperature measurements provided by a pirometer, which – in certain cicumstances – provides false observations (too high or too low spikes) that has to be reduced. In the lower plot of the right panel of Fig. 2 the

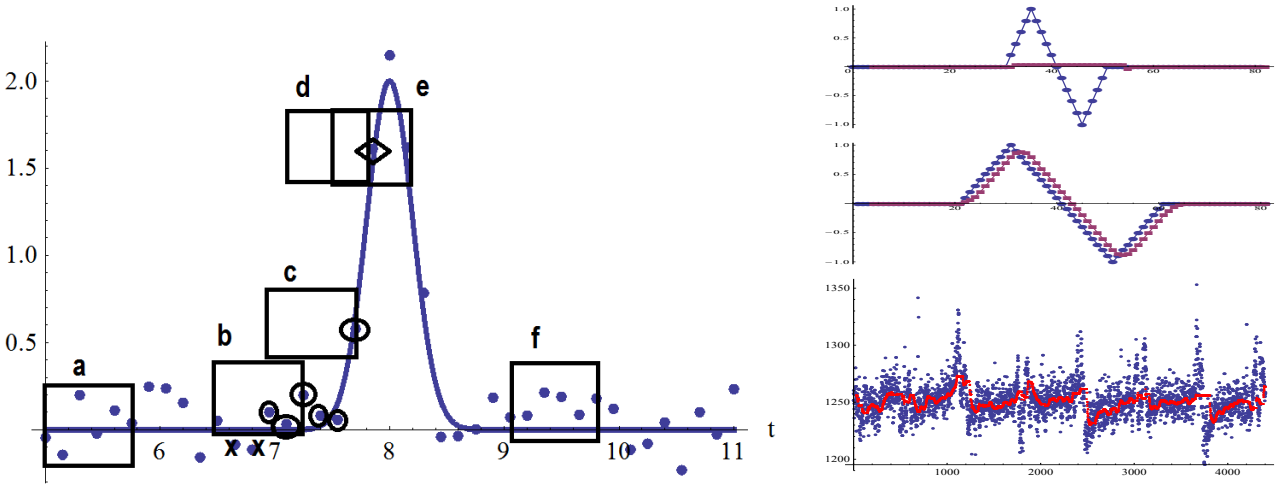


Figure 2: Left panel – the idea of SDS method. Right panel : upper plot – detection and suppression of narrow triangle spikes, middle plot – SDS method preserves leass rapidly changing triangles, lower plot – the time series of temperatures and the result of applying SDS method.

result of applying SDS method is shown (fat red curve).

References

[1] A. C. Atkinson, S.-J. Koopman, and N. Shephard. Detecting shocks: outliers and breaks in time series. *Journal of Econometrics*, 80(2):387–422, 1997.

[2] H. Azami and S. Sanei. Spike detection approaches for noisy neuronal data: Assessment and comparison. *Neurocomputing*, 133:491–506, 2014.

[3] S. Mukhopadhyay and GC Ray. A new interpretation of nonlinear energy operator and its efficacy in spike detection. *IEEE Transactions on Biomedical Engineering*, 45(2):180–187, 1998.

[4] E. Rafajlowicz, M. Pawlak, and A. Steland. Nonparametric sequential change-point detection by a vertically trimmed box method. *IEEE Transactions on Information Theory*, 56(7):3621–3634, 2010.

[5] A. Steland. Nonparametric monitoring of financial time series by jump-preserving estimators. *Statistical Papers*, 43:361–377, 2002.

Acknowledgements: The research of the 1-st author has been supported by the National Science Center under grant: 2012/07/B/ST7/01216.

Dimensionality reduction by random projection: to orthogonalize or not to orthogonalize ?

Session: **Analysis, Testing and Change Detection in High Dimensions 2**

Ewa Skubalska-Rafajłowicz *Department of Computer Engineering, Faculty of Electronics, Wrocław University of Science and Technology, Wrocław, Poland*

Abstract: We have shown that orthogonal Gaussian random projections, however more time consuming, are more accurate and should be preferred in many applications in which dimensionality reduction is based on the Johnson-Lindenstrauss lemma.

1 Introduction

Random projections are closely related to the Johnson-Lindenstrauss lemma [3], [4], which states that any set S, say, of N points in an Euclidean space can be embedded in an Euclidean space of a lower dimension ($\sim O(logN/\varepsilon^2)$) with a relatively small distortion ε of the distances between any pair of points from S. The Johnson-Lindenstrauss lemma has been shown to be useful in applications in computer science and engineering.

In the random projection methods, the original high-dimensional observations are projected onto a lower-dimensional space using a suitably scaled random matrix with independent, typically, normally distributed entries. It is well known that if $A \in R^{k\times d}$ is randomly generated matrix such that each its entry is independently generated from standard normal distribution $\mathcal{N}(0.1)$, then $AA^T \in R^{k\times k}$ is with probability 1 a full rank, symmetric (and positive definite) matrix. Thus, randomly obtained matrix A generates random k-dimensional subspace of R^d. $(AA^T)^{-1/2}A$ consists of k orthonormal row vectors (with probability 1) and it is known that these vectors are chosen according to the Haar measure on the compact Stiefel manifold $S_d(k) = \{B \in R^{k\times d} : B^TB = I_k\}$.

$(AA^T)^{-1/2}A$ (with factor $\sqrt{\frac{d}{k}}$) is an orthogonal Gaussian random projection (GRP) from R^d onto R^k and $P_A = A^T(AA^T)^{-1}A$ is a corresponding orthogonal GRP on a random k dimensional subspace in R^d generated by matrix A.

Other possibility [4] is to use non-orthogonal projection, i.e., multiplying by randomly generated matrix A (with factor $\sqrt{\frac{1}{k}}$) .

We will concentrate on the accuracy of the GRP in terms of probability of obtaining an $\varepsilon-$ isometry for the finite set of points transformed from R^d into R^k, where $k << d$.

It will be shown that for the same projection dimension k and the Euclidean metric distortion ε the probability that the orthogonal GRP fails is usually smaller than the adequate probability obtained for non-orthogonalized GRP.

It is known (see, for example, [2] Theorem 1.1) that if θ is the angle between vector $x \in R^d$ and random k-dimensional subspace in R^d, then the random variable $\cos(\theta)^2$ has the beta distribution $\beta(k/2, (d-k)/2)$. Consequently, the random variable $\sin(\theta)^2 = 1 - \cos(\theta)^2$ has the beta distribution $\beta(d-k)/2, k/2)$.

Thus, for any vector $x \in R^d$, randomly generated Gaussian matrix A and the corresponding orthogonal projection P_A,

$$||P_A \frac{x}{||x||}||^2 = ||(AA^T)^{-1/2} A \frac{x}{||x||}||^2$$

has the beta distribution $\beta(k/2, (d-k)/2)$ (random with respect to A).

2 Probability bound for orthogonal GRP

We will use Theorem 5.1 formulated by Frankl and Maehara [2] providing the bounds on how $X \sim \beta(p, q)$ deviates from its expected value $\mu = \frac{p}{p+q}$.
Probability that for given $k < d$ and $\varepsilon \in (0, 1)$ orthogonal GRP fails is bounded by

$$UO = \frac{N(N+1)}{\varepsilon/2 - \varepsilon^2/3} \sqrt{\frac{d}{\pi k(d-k)}} \exp\left(-\frac{k}{2}(\varepsilon^2/2 - \varepsilon^3/3)\right).$$

The corresponding known probability bound for non-orthogonal GRP (based on chi-square distribution with k degrees of freedom) is

$$UNO = N(N+1) \exp\left(-\frac{k}{2}(\varepsilon^2/2 - \varepsilon^3/3)\right).$$

It will be shown that for reasonable values of parameters (k/d,k, ε) the probability upper bound UO is visibly smaller than UNO what is important when set of points S is not fully known in advance.

References

[1] Achlioptas, D.(2003) Database friendly random projections:Johnson-Lindenstrauss with binary coins *Journal of Computer and System Sciences,* Vol. 66, pp. 671–687

[2] Frankl, P.and Maehara, H.(1990). Some geometric applications of the beta distribution, *Annals of the Institute of Statistical Mathematics,* Vol.42 (3) pp. 463–474.

[3] Johnson, W. B. and Lindenstrauss, J.(1984). Extensions of lipshitz mapping into Hilbert space, *Contemporary Mathematics* Vol. 26 pp. 189–206

[4] Matoušek, J.(2008) On Variants of the Johnson-Lindenstrauss Lemma, *Random Structures and Algorithms* Vol. 33(2) pp. 142–156.

Tests for proportional hazards under censored data

Session: **Survival Analysis**

Marc Ditzhaus *Heinrich-Heine-University of Düsseldorf.*

Abstract: In this talk we consider a two-sample survival setting under right censoring. Due to the famous work of Cox [1] it is of great interest to test whether the hazard rates of the two samples are proportional. By extending an idea of Neuhaus [6] concerning the testing problem of equal distributions in the two-sample setting we developed a studentized permuation test for the proportional hazard testing problem. Beside the theory we present a simulation study, where we compare our test with the test of Grambsch and Therneau [3] and the test of Gill and Schumacher [2].

1 Introduction

Let $T_1, \dots, T_{n_1}$ and $C_1, \dots, C_{n_1}$ be survival times with CDF F_1 and censoring times with CDF G_1, respectively. Analogously, let $T_{n_1+1}, \dots, T_n$ and $C_{n_1+1}, \dots, C_n$ be given for the second group with CDF F_2 and G_2, respectively. Suppose that Lebesgue-densities f_1 and f_2 of T_1 and T_n exist. Then the hazard rate λ_1 of the first group, i.e. of T_1, is given by

$$\lambda_1(t) := \frac{f_1(t)}{1-F_1(t)} \;=\; \lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon} P\Big(T_1 \in [t, t+\varepsilon) \Big| T_1 \geq t\Big)$$

for all $t > 0$. Let λ_2 be the hazard rate of the second group, i.e. of T_n. Due to the famous work of Cox [1] it is of great interest to test wheter the hazard rates λ_1 and λ_2 of the first and the second group, respecitvely, are proportional, i.e.

$$\lambda_2(t) \;=\; \exp(\beta)\, \lambda_1(t) \;=\; \vartheta\, \lambda_1(t)$$

for all $t > 0$ and some $\beta \in \mathbb{R}$, $\vartheta > 0$. In practice it is often not possible to observe the survival times $T_1, \dots, T_n$. Instead of them the (possibly censored) times $X_i = \min(T_i, C_i)$ and the censoring status $\Delta_i = \mathbf{1}\{T_i \leq C_i\}$ can be observed. Neuhaus [6] suggested a permutation test based on studentized survival rank tests for the null $\mathcal{H}_0^= : F_1 = F_2$, where the censoring distributions can differ, i.e. $G_1 \neq G_2$. To explain his idea let $X_{i:n}$ be the ith order statistic of $X_1, \dots, X_n$ and let $\Delta_{i:n}$ be the corresponding censoring status. The idea is to permutate the obervation times $X_1, \dots, X_n$ while keeping the vector $(\Delta_{i:n})_{1 \leq i \leq n}$ fixed. This leads to an asymptotically exact permutation test for $\mathcal{H}_0^=$. Monte-Carlo simulations of Neuhaus [6] and Heller and Venkatraman [4] show a good finite performance of the test.

2 Our null

Let us now consider triangular schemas of random variables, e.g. $T_{n,1}, \dots, T_{n,n}$. Let $\lambda_{n,1}$ and $\lambda_{n,2}$ be the hazard rates of the survival times within the two samples, which now depend on the sample size n. Following the ideas in [5, 6] we developed a studentized permutation test for the following null $\mathcal{H}_0$, which reminds on local

alternatives:

$$\mathcal{H}_0 : \lambda_{1,n} = \vartheta_n \lambda_{2,n}\,, \text{ where } \vartheta_n = 1 + \frac{\gamma}{\sqrt{n}} \text{ for some } \gamma > 0 \text{ and all } n \in \mathbb{N}.$$

Our test is asymptotically exact for $\mathcal{H}_0$. One advantage of the permuation idea is that the corresponding test is even finite exact under the more restrictive null

$$\widetilde{\mathcal{H}}_0 : F_{n,1} = F_{n,2}, G_{n,1} = G_{n,2} \text{ for all } n \in \mathbb{N},$$

where $F_{n,j}$ and $G_{n,j}$ are the CDF of the surival times and the censoring times, respecitvely, of group $j \in \{1, 2\}$.

References

[1] Cox, D.R. (1972). *Regression models and life-tables.* J. R. Stat. Soc. Ser. B, Vol. 34, p. 187–220.

[2] Gill, R.D. and Schumacher, M. (1987), *A simple test of the proportional hazards assumption.* Biometrika 74, No. 2, 289-300.

[3] Grambsch, P.; Therneau, T (1994). *Proportional hazards tests and diagnostics based on weighted residuals.* Biometrika 81, No. 3, 515–526.

[4] Heller, G. and Venkatraman, E.S. (1996), *Resampling procedures to compare two survival distributions in the presence of right-censored data.* Biometrics 52, No. 4, 1204–1213.

[5] Janssen, A. and Mayer, C.-D. (2001). *Conditional Studentized survival tests for randomly censored models.* Scand. J. Stat. 28, No. 2, 283–293.

[6] Neuhaus, G. (1993). *Conditional rank tests for the two-sample problem under random censorship.* Ann. Stat. 21, No. 4 , 1760–1779.

Multivariate Survival Analysis for random right censorship models

Session: **Survival Analysis**

Arnold Janssen *Heinrich-Heine-University of Düsseldorf.*

The talk considers multivariate survival models under random right censorship. Already at dimension two the estimation of the two dimensional survival function is a difficult affair, see Dabrowska, Gill, Stute or van der Laan. This talk is hazard based within a multivariate setting. The survival functions are expressed by hazard exponent measures which describe higher order dependence. They are multivariate versions of the univariate cumulative hazard functions. Our approach offers an efficient method for estimation of the hazard exponent measures of dependence. Exponent dependence measures were introduced in Bendel, Dobler and Janssen, where the following theorem can be found.

Theorem (see [1] and [4])

For all $t = (t_1, \ldots, t_d) \in (0, \infty)^d$ we have

$$S(t) = \prod_{i=1}^{d} S_i(t_i) \exp\left(\sum_{I \subset \{1,\ldots,d\},\, |I| \geq 2} (-1)^{|I|} \Lambda_I(t) \right),$$

where S_i marginal survival function

Λ_I exponent measures of dependence (signed measures) for $I \subset \{1, \ldots, d\}$.

For short:

$$S(t) = \prod_{\emptyset \neq I \subset \{1,\ldots,d\}} S^I\left((t_i)_{i \in I} \right), \quad \text{where } S^I = \exp\left[(-1)^{|I|}\ \Lambda_I \right].$$

References

[1] Bendel, J. ; Dobler, D. ; Janssen, A. (2014). *Exponent dependence measures of survival functions and correlated frailty models.* arXiv:1409.6854.

[2] Dabrowska, D. M. (1988). *Kaplan Meier Estimate on the Plane.* Ann. Stat. 16, No. 4, 1475–1489.

[3] Dabrowska, D.M. (1989). *Kaplan Meier Estimate on the Plane: Weak Convergence, LIL, and the Bootstrap.* J. Multivariate Anal. 29, 308–325.

[4] Gill, R. (1992). *Multivariate survival analysis.* Theory Probab. Appl. 37, No.1, 18–31, No.2, 284–301.

[5] Gill, R. ; van der Laan, M. J. ; Wellner, J.A. (1995). *Inefficient estimators*

of the bivariate survival funtion for three models. Ann. Inst. Henri Poincare, Probab. Stat. 31, No.3, 545–597.

[6] Janssen, A.; Rahnenführer, J. (2002). *A Hazard-based approach to dependence tests for bivariate censored models.* Math. Methods Statist. 11, 297–322.

[7] Sen, A. ; Stute, W. (2007). *A bi-variate Kaplan-Meier estimator via an integral equation.* Technical Report # 03/07, Department of Mathematics and Statistics, Concordia University.

[8] van der Laan, M.J. (1996). *Efficient estimation in the bivariate censoring model and repairing NPMLE.* Ann. Statist. 24, 596–627.

Empirical Martingale Spaces with Applications

Session: **Survival Analysis**

Winfried Stute *Mathematical Institute, University of Giessen, Germany*
David Hess *Mathematical Institute, University of Giessen, Germany*

Abstract: In this talk we present a full description of empirical martingales, i.e., functions of the empirical distribution function which are martingales. Many of these martingales are highly nonlinear. Interestingly enough, they may be applied for goodness-of-fit tests in data situations, which are typical in an engineering context.

Reliability Prediction using Regression Models

Session: **Survival Analysis**

Maik Döring *University of Hohenheim, Germany*

Abstract: In order to know the failure behavior of some mechatronic systems and components under specified operating conditions, plenty of endurance tests have to be run. However, such tests are often time-consuming and expensive. In practice manufacturers may have lots of failure data of similar products using the same technology basis under different operating conditions. Thus, one can try to derive predictions for newly developed components or new application environments through the existing data using regression models.

Three regression models are considered: a parametric, a semi-parametric and a nonparametric approach. First, the parameters of the Weibull-distribution are modelled as linear functions of the covariates. Second, the Cox proportional hazards model, well-known in survival analysis, is applied. Finally, a kernel estimator is used to interpolate between empirical distribution functions. Further we discuss a model selection procedure.

A statistical test is proposed, to decide which of two competing model classes approximate the underlying distribution better. The Cramér-von Mises distance is used to define the closeness between two distribution functions. To illustrate this method of reliability prediction, the three classes of regression models are applied to real test data of motor experiments. Further the advantages and the disadvantages of the approaches are shown by Monte Carlo simulations.

Limit order book modeling and quasi likelihood analysis

Session: **Statistics of Stochastic Processes 2**

Nakahiro Yoshida *Graduate School of Mathematical Sciences, University of Tokyo, Japan*

Abstract: Modeling of limit order book is discussed. We propose use of a point process regression model that incorporates effects of covariates through the intensity process. The proposed models give a good fit to the real data. We discuss applications of the quasi likelihood analysis (QLA) to point processes. The QLA provides a new framework of statistical inference for stochastic processes.

1 Modeling of the limit order book

Today data analysis of the limit order book (LOB) is requiring new developments in statistical inference for stochastic processes. Since the sampling frequency is ultra-high, the microstructure becomes a subject of modeling, rather than eliminated as noise. There is no Brownian motion as a process driving the system since the central limit theorem is not effective at this level of description, differently from the standard framework. However, point processes give a promising approach to a description of LOB.

We discuss an attempt of LOB model building from the data, along Muni Toke and Yoshida [2]. The intensity of the counting process N^M of the market orders on bid/ask side is modeled as a function

$$\begin{aligned}\lambda^M(t,\beta) \;=\; \exp\Big[&\beta_0+\beta_1\log S(t)+\beta_{11}\big(\log S(t)\big)^2\\ &+\beta_2\log(1+q_1(t))+\beta_{22}\big(\log(1+q_1(t))\big)^2\\ &+\beta_{12}\log S(t)\;\log(1+q_1(t))\Big],\end{aligned}$$

where $\beta=(\beta_0,\beta_1,\beta_2,\beta_{11},\beta_{22},\beta_{12})$, $S(t)$ is the spread and $q_1(t)$ is the volume at the best quote on the side of submission. Variables $S(t)$ and $q_1(t)$ are treated as covariates and the quasi likelihood analysis is applied to

$$\ell_T^M(\beta) \;=\; \int_0^T \log\big(\lambda^M(t,\beta)\big)dN_t^M-\int_0^T\lambda^M(t,\beta)dt.$$

The proposed model gives overall a good fit to the real data. The market order intensity model is a so-called point process regression model. The advantage of this formulation is that we can very freely incorporate dependency of the intensity on various covariates.

It is possible to take a similar approach to the limit order by an intensity model with covariates $S(t)$ and the total volume $Q_{10}(t)$ available in the LOB up to the tenth limit of the side of submission. Moreover, the placements of limit orders are modeled by a normal mixture model $\pi^L(p)=\sum_{i=1}^G\pi_i\phi(p;\mu_i,\sigma_i^2)$.

For the cancellation process, we introduce a new *priority index* as a main modelling variable, which turns out to be very efficient.

All proposed models were fitted to a database of 10 consecutive trading days (17-28 January 2011) for 6 different liquid stocks traded on the Paris stock exchange in the construction of the model, then to much longer sample (2011-2013) to test robustness of the results.

2 Quasi likelihood analysis for point processes

The quasi likelihood analysis (QLA) is a systematic analysis of the quasi likelihood random field and the associated estimators ([5]). The QLA features a large deviation method that provides more precise tail probability estimates for the random field and estimators than those limit distributions give. The QLA serves to form the basis of developments of prediction, model selection, information criteria, asymptotic expansion, higher-order inferential theory, sparse estimation, etc. Thanks to flexibility, the QLA was successfully applied to asymptotic statistics of diffusion processes, jump-diffusion processes, non-synchronous smapling, model selection and point processes ([5], [4] and Uchida AISM2010, Uchida-Y SPA2013, Ogihara-Y SISP2011, Ogihara-Y SPA2014, among many others).
QLA can be constructed for point processes in both ergodic and non-ergodic cases ([1], [3]) as reviewed in this talk. Recently the QLA is used for sparse estimation.

References

[1] Clinet, S., Yoshida, N. (2016). Statistical inference for ergodic point processes and application to Limit Order Book. *Stochastic Processes and Their Applications*, on line.

[2] Muni Toke, I., Yoshida, N. (2016). Modelling intensities of order flows in a limit order book. *Quantitative Finance*, on line.

[3] Ogihara, T, Yoshida, N. (2015). Quasi likelihood analysis of point processes for ultra high frequency data. arXiv:1512.01619

[4] Uchida, M., Yoshida, N. (2016). Model selection for volatility prediction. *The Fascination of Probability, Statistics and their Applications. In Honour of Ole E. Barndorff-Nielsen* Mark Podolskij, Robert Stelzer, Steen Thorbjørnsen, Almut E. D. Veraart, eds, 343-360, Springer.

[5] Yoshida, N. (2011). Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. *Annals of the Institute of Statistical Mathematics*, 63, 431–479.

Absolute continuity of the invariant measure in systems of interacting neurons.

Session: **Statistics of Stochastic Processes 2**

Eva Löcherbach *Laboratoire AGM CNRS UMR 8088, Université de Cergy-Pontoise, France*

Abstract: We consider the following model of a finite system of interacting neurons. Each neuron is represented by its membrane potential. After a random time (which has an intensity depending on the potential of the neuron), a neuron "spikes", i.e. emits an action potential. At this spiking time, the potential of the spiking neuron is reset to zero, while all other neurons receive an additional amount of potential. In between successive spikes, the value of the potential of a neuron follows a deterministic evolution (basically, some leak effects imply that the potential values are all attracted to some equilibrium potential). As a consequence, the system follows an evolution which is described by a Piecewise Deterministic Markov Process (PDMP) with degenerate transitions.
We are interested in the estimation of the unknown spiking rate function of each single neuron, by using a Nadaray-Watson type kernel estimator.
It is well known that such a kernel estimator works well if the invariant measure of the underlying process has a sufficiently regular Lebesgue density. This talk will discuss this question and the associated difficulties which are mainly due to the fact that the transition kernel of the above system is very degenerate.

References

[1] Hodara, P., Krell, N., Löcherbach, E. (2016). Nonparametric estimation of the spiking rate in systems of interacting neurons, , DOI 10.1007/s11203-016-9150-4.

[2] Löcherbach, E. (2016). Absolute continuity of the invariant measure in Piecewise Deterministic Markov Processes having degenerate jumps, accepted for publication in *Stoch. Proc. Appl.*

Wasserstein distances between discretely observed Lévy processes

Session: **Statistics of Stochastic Processes 2**

Ester Mariucci *Humboldt Universität zu Berlin, Germany*
Markus Reiß *Humboldt Universität zu Berlin, Germany*

Abstract: We present some upper bounds for the Wasserstein distance of order p between the product measures associated with the increments of two independent Lévy processes with possibly infinite Lévy measures. As an application, we derive an upper bound for the total variation distance between the marginals of two independent Lévy processes with possibly infinite Lévy measures and non-zero Gaussian components. Also, a lower bound for the Wasserstein distance of order p between the marginals of two independent Lévy processes is discussed.

1 Introduction

We are interested in understanding the geometry of the space of Lévy processes. In particular, we want to define (or find) a good metric to measure the distance between the increments of two independent and discretely observerd Lévy processes in terms of their Lévy triplets.
One feature we look for is a distance for which the contribution given by the increments of the small jumps can be confused with those of a Gaussian random variable having the same mean and variance.
The starting point of this investigation is to consider Wasserstein distances.

2 Techniques

The Wasserstein distance of order p between the measures μ and ν is

$$\mathcal{W}_p(\mu,\nu) := \inf\left\{\left[\mathbb{E}[|X'-Y'|^p]\right]^{\frac{1}{p}},\ \text{law}(X')=\mu,\ \text{law}(Y')=\nu\right\},$$

where the infimum is taken over all X' and Y' having laws μ and ν, respectively.
The properties of such distances that are important for our scopes are:

- Homogeneity and sub-additivity.
- A good behavior with respect to product measures.
- An accurate central limit theorem, see [1].

These three properties can be exploited to obtain an upper bound of the Wasserstein distance of order p between the increments of two given Lévy processes.
To obtain a lower bound, the Wasserstein distances are compared with the Toscani distance [2].

3 Conclusions

The main points that will be presented in this talk are:

- Upper and lower bounds for the Wasserstein distances between the increments of two independent Lévy processes in terms of their Lévy triplets.
- A fine control for the Wasserstein distances between the marginal of a pure

jump Lévy process and a Gaussian distribution.
- Upper bound for the total variation distance between the marginals of any pair of independent Lévy processes.

References

[1] Rio, E. (2009). Upper bounds for minimal distances in the central limit theorem. *Annales IHP* **3**, 802–817.

[2] Toscani, G. and Villani, C. (1999). Probability metrics and uniqueness of the solution to the Boltzmann equation for a Maxwell gas. *J. Statist. Phys.* 94, 3/4, 619–637.

Estimating occupation time functionals

Session: **Statistics of Stochastic Processes 2**

Randolf Altmeyer *Humboldt-Universität zu Berlin, Germany*
Jakub Chorowski *Humboldt-Universität zu Berlin, Germany*

Abstract: An occupation time functional is a time integral $\int_0^T f(X_t)dt$ for a function f and a continuous-time stochastic process $(X_t)_{0\leq t\leq T}$. Given discrete-time observations of the process we approximate the occupation time functional by a Riemann-sum estimator and study the rate of covergence. For Sobolev-smooth functions f we establish surprising upper and lower bounds on the approximation error for many important processes such as Markov processes, semimartingales, and Gaussian processes, e.g. fractional Brownian motion. We also provide a generalized Itô formula for continuous Itô semimartingales, which is of independent interest, and apply it to prove stable central limit theorems.

References

[1] Altmeyer, R., Chorowski, J. (2017). Estimating occupation time functionals, *in preparation.*

[2] Altmeyer, R., Chorowski, J. (2016). Estimation error for occupation time functionals of stationary Markov processes, *arxiv preprint.*

Seeing the unseen: INAR(1) models with under-reported data

Session: **Discrete-Valued Time Series 2**

Pedro Puig *Department of Mathematics, Universitat Autònoma de Barcelona, Spain*
Amanda Fernández-Fontelo *Department of Mathematics, Universitat Autònoma de Barcelona, Spain*
Alejandra Cabaña *Department of Mathematics, Universitat Autònoma de Barcelona, Spain*
David Moriña *Institut Català d'Oncologia (ICO), Spain*

Abstract: Dealing with under-reported data is a quite frequent problem in public health practice. We present a hidden Markov chain model to deal with time series of counts coming from an INAR(1) underlying process.

1 Introduction

Human papillomavirus (HPV) is one of the most prevalent sexually transmitted infection such that nearly all sexual people have it at some point in their lives [1]. Normally, the infection disappears on its own without inducing health problems, but in few cases it can be related to several cancers which usually are diagnosed many years after the infection. Figure 3 shows the number of cases per week of human papillomavirus (HPV) in Girona (a province of Catalonia, Spain) from 2010 to 2014.

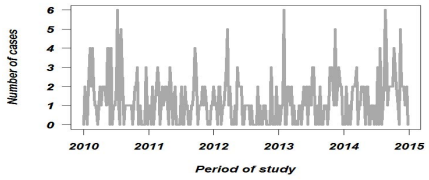


Figure 3: Number of weekly cases of HPV recorded in Girona from 2010 to 2014.

It seems reasonable to consider that the number of cases of HPV could be under-reported since most of the sexual active people are infected without manifesting symptoms or health problems. Then, what can we say about the unobserved cases? How can we see the unseen?

2 An INAR(1)-hidden Markov chain model

Because the correlation structure of the observed time series Y_n looks approximately like an AR(1), we assume that the total number of cases of HPV (not observed), denoted as X_n, follows a classical Integer-Autoregressive process of order 1 (INAR(1)), that is,

$$X_n = \alpha \circ X_{n-1} + W_n,$$

where $\alpha \in (0, 1)$ is a parameter and W_n are iid Poisson(λ) distributed. Moreover, X_{n-1} and W_n are assumed to be independent at any time n. The $\circ$ operator, called

binomial thinning, is defined as follows,

$$\alpha \circ X_{n-1} = \sum_{i=1}^{X_{n-1}} \xi_i,$$

being ξ_i iid Bernoulli random variables with probability of success equal to α.
Let Y_n be the observed time series. The way we allow Y_n to be under-reported is by assuming the observed counts as,

$$Y_n = \begin{cases} X_n & : \text{ with probability } 1-\omega \\ q \circ X_n & : \text{ with probability } \omega \end{cases} \tag{1}$$

Note that the observed time series Y_n coincides with the underlying series X_n (and therefore the observed count at time n is not under-reported) with probability $1-\omega$. Otherwise, Y_n is a *binomial thinning* of the hidden process X_n, representing an under-reported phenomenon coming from an underlying INAR(1) process. Parameter ω can be interpreted as the proportion of times that X_n is not completely observed (under-reported) while q quantifies the intensity of the under-reportation. The closer to zero is q, the more intense is the problem of under-reporting in the series.
Because the hidden process X_n is markovian, the observed process Y_n is a hidden Markov chain model with emission probabilities,

$$P(Y_i = j \mid X_i = k) = \begin{cases} 0 & if \quad k < j \\ (1-\omega) + \omega q^k & if \quad k = j \\ \omega \binom{k}{j} q^j (1-q)^{k-j} & if \quad k \geq j. \end{cases} \tag{2}$$

Details of these models, properties and several examples of application can be found in [2]. Using the methodology of hidden-Markov chains (forward and Viterbi algorithms, etc) we can estimate the parameters of the model and reconstruct the hidden process. In particular, we find that public health policy makers only record in average 38% of the total HPV cases in Girona, being a highly under-reported time series with an estimated frequency $\hat{\omega} = 0.922 (s.e. = 0.073)$.

References

[1] Dunne EF, Markowitz LE, Saraiya M, Stokley S, Middleman A, Unger ER, Williams A and Iskander J (2014). CDC grand rounds: Reducing the burden of HPV-associated cancer and disease. *Morbidity and mortality weekly report*, **63**(4), p. 69–72.

[2] Fernández-Fontelo A, Cabaña A, Puig P and Moriña D (2016). Under-reported data analysis with INAR-hidden Markov chains. *Statistics in Medicine*, **35**(26), p. 4875–4890.

Modelling Time Series of Counts under Censoring or Truncation

Session: **Discrete-Valued Time Series 2**

Maria Eduarda Silva *Faculdade de Economia, Universidade do Porto & CIDMA, Portugal*
Isabel Pereira *Departamento de Matemática & CIDMA, Universidade de Aveiro, Portugal*
Brendan McCabe *The University of Liverpool, U.K.*

Abstract: Data resulting from censoring and truncation are frequently encountered in diverse fields including environmental monitoring, medicine, economics and social sciences. Censoring occurs when observations are available only for a restricted range, e.g., due to a detection limit. Truncation, on the other hand, occurs if observations in some range are lost. This work considers the analysis of time series of counts under censoring based on first order integer autoregressive (INAR) models. The focus is on estimation and inference problems.

1 Introduction

There is an extensive literature on regression models in which dependent variables are censored or underlying distributions are truncated, see [3] for a comprehensive review including discrete dependent variables. The analysis of time series under censoring or truncation has received little attention in the literature. [4] and [1] consider censored continuous valued time series but censored count time series also arise in many areas.
Consider a time series of counts $X_1, \ldots, X_T$, generated according to an INteger AutoRegressive, INAR, model

$$X_t = \alpha_1 \diamond X_{t-1} + \alpha_2 \diamond X_{t-2} + \ldots + \alpha_p \diamond X_{t-p} + \epsilon_t \tag{1}$$

where the arrival process ϵ_t is a sequence of independent and identically distributed non-negative integer valued random variables and, conditional on X_{t-i}, $\alpha_i \diamond X_{t-i}$ is an integer valued random variable whose probability distribution with respect to the counting measure ν is denoted by $f(\cdot|X_{t-i}; \alpha_i)$ and depends on the parameter α_i which may be a vector of parameters; the variables X_{t-i}, $\alpha_i \diamond X_{t-i}$, $i \in \{1, \ldots, p\}$ conditional on X_{t-i} are mutually independent. Thus, $'\diamond'$ denotes a random operator, usually called thinning operator, which always produces integer values and introduces serial dependence via the conditioning on X_{t-i}. There are several possibilities for the conditional distribution of $\alpha_i \diamond X_{t-i}$, for a review see [2]. For the remainder of this work, we consider the operator $\alpha_i \diamond X_{t-i}$ as binomial thinning, Then if ϵ_t is a Poisson variable, the model gives rise to the standard INAR(p)-P model; when $p = 1$ X_t has a marginal Poisson distribution and the model is called PoINAR(1). For the remainder of this work we consider $p = 1$, for easiness of notation.

2 A censored count model

Censoring occurs if observations Y_t are available for a restricted range due to aggregation or detection limits. We can say that censored data are "piled up" at a censoring point. We define a **censored at N model** as

$$
\begin{aligned}
X_t &= \alpha \diamond (X_{t-1}) + \epsilon_t \\
Y_t &= \min\{X_t, N\} = \begin{cases} X_t, & \text{if } X_t \leq N \\ N, & \text{if } X_t > N \end{cases}
\end{aligned} \tag{2}
$$

Neglecting censoring and truncation in the time series hinders meaningful statistical inference, leading to model misspecification, biased parameter estimation, and poor forecasts. We study the regression properties of the censored INAR(1) process and show that least squares estimation of the parameters is no longer appropriate. Likelihood analysis is developed, including maximum likelihood estimation and computation of score and information quantities. Small sample properties of ML estimators are studied in the particular case of Poisson marginals. Forecasting is also considered.

Acknowledgments

This work was partially supported by the Portuguese Foundation for Science and Technology (FCT-Fundação para a Ciência e a Tecnologia), through CIDMA - Center for Research and Development in Mathematics and Applications, within project UID/MAT/04106/2013.

References

[1] Choi, Seokwoo Jake and Portnoy, Stephen (2016). Quantile Autoregression for Censored Data. *J. Time. Ser. Anal.* 37, p. 603–623.

[2] Scotto, M.G. Weiss, C.H. and Gouveia, S. (2015). Thinning-based models in the analysis of integer-valued time series: a review. *Statistical Modelling* 15, 590-618.

[3] Greene, W. (2005). Censored Data and Truncated Distributions, *Working Papers 05-08*, New York University, Leonard N. Stern School of Business, Department of Economics.

[4] Park, Jung Wook and Genton, Marc G. and Ghosh, Sujit K. (2007). Censored time series analysis with autoregressive moving average models. *The Canadian Journal of Statistics*, 35, pp. 151-168.

Periodic INAR(1) models based on the signed thinning operator

Session: **Discrete-Valued Time Series 2**

Isabel Pereira *Departamento de Matemática & CIDMA, Universidade de Aveiro, Portugal*
Cláudia Santos *CIDMA & Departamento de Ciências Exatas, Escola Superior Agrária de Coimbra, Portugal*
Manuel G. Scotto *Departamento de Matemática & CEMAT, Instituto Superior Técnico, Portugal*

Abstract: INteger-valued AutoRegressive (INAR) processes play a central role in the statistical analysis of integer-valued time series. In this work we propose two INAR (univariate and bivariate) models with periodic structure, S-PINAR(1) and BS-PINAR(1) respectively. Both models are based on the signed thinning operator allowing for positive and negative counts. We examine the basic probabilistic and also the statistical properties of the periodic models. Innovations are modeled by univariate and bivariate Skellam distributions, respectively. To study the performance of the conditional least squares and conditional maximum likelihood estimators, a simulation study is conducted for the S-PINAR(1) model.

1 Introduction

The binomial thinning operator can only be applied to count variables therefore it has been generalized in a number of different ways. [4] introduced the signed binomial thinning operator, allowing time series with negative values, the so-called $\mathbb{Z}$-valued time series. [3] established a slightly different signed thinning operator also allowing for negative values both for the series and its autocorrelation function. Recently, [1] proposed an extension of the preceding signed thinning operator to the bivariate case. Many phenomena have in their essence a periodic structure and there are several potential applications for this class of models. In this work, we introduce two new first-order integer-valued autoregressive models with time-varying parameters and sequences of innovations with periodic structure. Both models are based on the signed thinning operator defined in the univariate case by [3] and in the bivariate case by [1], adapted to the periodic case, accordingly.

2 Development

Basic notations and definitions concerning the periodic signed thinning operator are established as well as some of its properties. Emphasis will be placed on models with innovations following Skellam distribution and bivariate Skellam distribution, respectively. Therefore, a brief description of the periodic Skellam distribution for both univariate and bivariate distributions defined on the whole set of integers is also provided.
In extending the model proposed by [2] to the periodic case, we introduce a univariate signed periodic INAR(1) process (S-PINAR(1) for short) with period s, by considering a parametric assumption on the common distribution of the periodic counting sequence of the model. We focus on a specific parametric case which arises under the assumption of periodic Skellam distributed innovation. The properties

of the S-PINAR(1) model with period s are discussed. Due to some limitations of the periodic signed thinning, only the conditional moments of first and second-order of the process were established. Regarding parameter estimation, two methods are considered: conditional least squares and conditional maximum likelihood. The performance of the proposed estimation methods for the S-PINAR(1) model is accomplished through a simulation study. Numerical results from the simulation study suggested that the proposed model is suitable for practical use.
Within the bivariate setting, the work of [1] has motivated a new periodic bivariate model. The generalization of the previous signed model with period s to the bivariate case is denoted by BS-PINAR(1). Several statistical properties of this periodic model are derived. The assumption of a diagonal autoregressive matrix is made, thus the correlation is achieved through their innovation processes, where the distribution of the innovation processes is set a priori which consequently determines the distribution of the underlying time series. Hence, the discrete bivariate distribution on $\mathbb{Z}^2$ assigned to the distribution of the innovations is the periodic bivariate Skellam distribution. Parameter estimation of the unknown parameters of the BS-PINAR(1) model with period s is provided through conditional maximum likelihood method.

Acknowledgements

This work was supported in part by the Portuguese Foundation for Science and Technology (FCT-Fundação para a Ciência e a Tecnologia), through CIDMA - Center for Research and Development in Mathematics and Applications, within project UID/MAT/04106/2013.

References

[1] Bulla, J., Chesneau, C. and Kachour, M. (2016). A bivariate first-order signed integer-valued autoregressive process. *Communications in Statistics-Theory and Methods* (accepted).

[2] Chesneau, C. and Kachour, M. (2012). A parametric study for the first-order signed integer-valued autoregressive process. *Journal of Statistical Theory and Practice*, 6, pp. 760–782.

[3] Kachour, M. and Truquet, L. (2011). A p-order signed integer-valued autoregressive (SINAR(p)) model. *Journal of Time Series Analysis*, 32, pp. 223–236.

[4] Kim, H.-Y. and Park, Y. (2008). A non-stationary integer-valued autoregressive model. *Statistical Papers*, 49, pp. 485–502.

Time reversal and perfect simulation for the INAR(1) autoregressive process

Session: **Discrete-Valued Time Series 2**

Andreas Löpker *HTW Dresden, Germany*

Abstract: The INAR(1) process is the integer valued counterpart of the classic AR(1) autoregressive process and is used to model count data time series. Under mild conditions the process has a unique stationary distribution and the question arises how one can sample from that distribution. We show how an exact sample can be obtain by using a method called coupling from the past, employing the time reversal of the process.

Uniqueness of characterization of continuous distributions by single regression of generalized order statistics

Session: **Nonparametric Methods 1**

Mariusz Bieniek *Institute of Mathematics, Maria Curie Skłodowska University, Lublin, Poland*

Abstract: We consider the problem of the unique characterization of continuous distributions by regression function of generalized order statistics. Using Markov property of generalized order statistics from continuous distributions we show that the characterization is unique if and only if corresponding system of differential equations has the unique solution. This approach provides new proof of characterization of power, exponential and Pareto distributions by linearity of corresponding regression.

1 Motivation and results

Let $X_*^{(r)}, X_*^{(r+\ell)}$, $r, \ell \geq 1$, denote generalized order statistics (GOSs, for short) with fixed parameters $\gamma_1, \ldots, \gamma_{r+\ell}$, based on a continuous distribution function F supported on the interval (α, β) as defined by Kamps [6]. The most important special cases of GOSs include ordinary order statistics $X_{1:n} \leq \ldots \leq X_{n:n}$ of a random sample of size $n \in \mathbb{N}$ or record values of a sequence of i.i.d. observations. Other special cases are Pfeifer's record values, progressively censored type II order statistics and sequential order statistics.

Let $h : (\alpha, \beta) \to \mathbb{R}$ be a known continuous and strictly increasing function such that $E\big|h\big(X_*^{(r+\ell)}\big)\big| < \infty$. Define the regression function of $h\big(X_*^{(r+\ell)}\big)$ given $X_*^{(r)}$ as

$$\xi(x) = E\big(h(X_*^{(r+\ell)}) \mid X_*^{(r)} = x\big), \quad x \in (\alpha, \beta).$$

Then ξ is also continuous and increasing, and $\xi > h$ on (α, β). Moreover each continuous distribution F uniquely determines such regression function ξ. We consider the inverse problem of unique identification of F by the knowledge of ξ.

This problem is solved completely in the adjacent case, i.e. for $\ell = 1$. Then simple reasoning shows that F can be recovered from ξ and h as

$$F(x) = 1 - \exp\left(-\frac{1}{\gamma_{r+1}} \int_\alpha^x \frac{d\xi(y)}{\xi(y) - h(y)}\right), \quad x \in (\alpha, \beta),$$

see e.g. [2], [4] and [5]. However, due to complicated expressions for the conditional densities of GOSs in terms of Meijer's G–function, the arguments for the adjacent case cannot be extended to nonadjacent case.

For $\ell \geq 2$, if $h(x) = x$ and ξ is a linear function of the form $\xi(x) = ax + b$, then either $a \in (0, 1)$ and F is unique power distribution, or $a = 1$ and F is exponential, or $a > 1$ and F is Pareto distribution, see [1], [2]. Unfortunately, the method of the proof introduced in [3], utilizing so called integrated Cauchy functional equation, cannot be applied to nonlinear regression ξ.

We propose new approach to the problem, utilizing Markov property of GOSs based on continuous distributions. In the case of absolutely continuous distributions with continuous density we show that for $\ell \geq 2$ the uniqueness of the characterization of

the underlying F by the knowledge of the regression ξ is equivalent to the uniqueness of solution to the corresponding system of $\ell - 1$ ordinary differential equations. For instance, for $\ell = 2$ the regression ξ determines F uniquely if and only if the ordinary differential equation

$$y' = \frac{\gamma_{r+2}}{\gamma_{r+1}} \frac{y - h(x)}{\xi(x) - y} \xi'(x)$$

has the unique solution φ such that $h(x) < \varphi(x) < \xi(x)$ for all $x \in (\alpha, \beta)$. Moreover, then

$$F(x) = 1 - \exp\left(-\frac{1}{\gamma_{r+1}} \int_{\alpha}^{x} \frac{\xi'(y)}{\xi(y) - \varphi(y)} dy \right).$$

For continuous F the equivalent condition for uniqueness of the characterization is expressed in terms of an appropriate system of integral equations.
This approach provides new almost elementary proof of the characterization of exponential, power and Pareto distributions by linearity of the regression $\xi(x) = ax + b$ in the case when $\ell = 2$.

References

[1] Bieniek, M., Szynal, D. (2003). Characterizations of distributions via linearity of regression of generalized order statistics. *Metrika* 58, 259–271.

[2] Cramer, E., Kamps, U., Keseling, C. (2004). Characterizations via linear regression of ordered random variables: a unifying approach. *Comm. Statist. Theory Methods* 33 , 2885–2911.

[3] Dembińska, A., Wesołowski. J. (1998). Linearity of regression for non-adjacent order statistics, *Metrika* 48, 215–222.

[4] Franco, M., Ruiz, J. M. (1995). On characterization of continuous distributions with adjacent order statistics. *Statistics* 26, 375–385.

[5] Franco, M., Ruiz, J. M. (1996). On characterization of continuous distributions by conditional expectation of record values. *Sankhyā Ser. A* 58, 135–141.

[6] Kamps, U. (1995). A concept of generalized order statistics. *J. Statist. Plann. Inference* 48 , 1–23.

Efficient multivariate entropy estimation via k-nearest neighbour distances

Session: **Nonparametric Methods 1**

Thomas Berrett *Statistical Laboratory, University of Cambridge, UK*
Richard Samworth *Statistical Laboratory, University of Cambridge, UK*
Ming Yuan *Department of Statistics, University of Wisconsion–Madison*

The concept of entropy plays a central role in information theory, and has found a wide array of uses in other disciplines, including statistics, probability and combinatorics. The *(differential) entropy* of a random vector X with density function f is defined as

$$H = H(X) = H(f) := -\mathbb{E}\{\log f(X)\} = -\int_{\mathcal{X}} f(x) \log f(x)\, dx$$

where $\mathcal{X} := \{x : f(x) > 0\}$. It represents the average information content of an observation, and is usually thought of as a measure of unpredictability.
In statistical contexts, it is often the estimation of entropy that is of primary interest, for instance in goodness-of-fit tests of normality or uniformity, tests of independence, independent component analysis and feature selection in classification. Many nonparametric techniques exist including those based on sample spacings (in the univariate case), histograms and kernel density estimates. The estimator of [1] is particularly attractive as a starting point, both because it generalises easily to multivariate cases, and because, since it only relies on the evaluation of kth-nearest neighbour distances, it is straightforward to compute.
To introduce this estimator, let $X_1, \ldots, X_n$ be independent random vectors with density f on $\mathbb{R}^d$. Write $\|\cdot\|$ for the Euclidean norm on $\mathbb{R}^d$, and for $i = 1, \ldots, n$, let $X_{(1),i}, \ldots, X_{(n-1),i}$ denote a permutation of $\{X_1, \ldots, X_n\} \setminus \{X_i\}$ such that $\|X_{(1),i} - X_i\| \leq \ldots \leq \|X_{(n-1),i} - X_i\|$. The Kozachenko–Leonenko estimator of the entropy H is given by

$$\hat{H}_n = \hat{H}_n(X_1, \ldots, X_n) := \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{\|X_{(k),i} - X_i\|^d V_d (n-1)}{e^{\Psi(k)}} \right),$$

where $V_d := \pi^{d/2}/\Gamma(1+d/2)$ denotes the volume of the unit d-dimensional Euclidean ball and where Ψ denotes the digamma function. In fact, this is a generalisation of the estimator originally proposed in [1], which was defined for $k = 1$. Noting that $e^{\Psi(k)}/k \to 1$ as $k \to \infty$, this estimator can be regarded as an attempt to mimic the 'oracle' estimator $H_n^* := -n^{-1} \sum_{i=1}^{n} \log f(X_i)$, relying on the approximation

$$\frac{k}{n-1} \approx V_d \|X_{(k),1} - X_1\|^d f(X_1).$$

The initial purpose of this talk is to present new results that describe the theoretical properties of the Kozachenko–Leonenko estimator. In particular, when $d \leq 3$, we have that under a wide range of choices of k (which must diverge to infinity with

the sample size) and under regularity conditions, the estimator satisfies

$$n\mathbb{E}\{(\hat{H}_n - H_n^*)^2\} \to 0. \tag{1}$$

This immediately implies that in these settings, $\hat{H}_n$ is efficient in the sense that

$$n^{1/2}(\hat{H}_n - H) \xrightarrow{d} N\big(0, \operatorname{Var} \log f(X_1)\big).$$

The fact that the asymptotic variance is the best attainable follows from standard results on semiparametric information bounds. When $d = 4$, we have that (1) no longer holds but the Kozachenko–Leonenko estimator is still root-n consistent provided k is bounded. Moreover, when $d \geq 5$, a non-trivial bias means that the rate of convergence is slower than $n^{-1/2}$ in general, regardless of the choice of k.

I will next talk about our second main contribution, the proposal of a new entropy estimator, formed as a weighted average of Kozachenko–Leonenko estimators for different values of k. I will present results which show that it is possible to choose the weights in such a way as to cancel the dominant bias terms, thereby yielding an efficient estimator in arbitrary dimensions, given sufficient smoothness.

There have been several studies of the Kozachenko–Leonenko estimator, but results on the rate of convergence have until now confined either to the case $k = 1$ or (very recently) the case where k is fixed as n diverges. Our results assume substantially weaker regularity conditions compared with previous works, and show that with an appropriate choice of k, efficient entropy estimators can be obtained in arbitrary dimensions, even in cases where the support of the true density is the whole of $\mathbb{R}^d$. Such settings present significant new challenges and lead to different behaviour compared with more commonly-studied situations where the underlying density is compactly supported and bounded away from zero on its support. When f is bounded below the entropy functional is a smooth functional and can be estimated by standard techniques. However, when f is allowed to have unbounded support the entropy functional is no longer smooth, due to the non-differentiability of the funtion $x \mapsto x \log x$ at the origin, and estimation becomes more difficult. To the best of our knowledge, therefore, this is the first time that a nonparametric entropy estimator has been shown to be efficient in multivariate settings for densities having unbounded support. Although the focus of this work is on efficient estimation, our results have other methodological implications: first, they suggest that prewhitening the data, i.e. replacing X_i with $Y_i := \hat{\Sigma}^{-1/2} X_i$ before computing the estimator, where $\hat{\Sigma}$ denotes the sample covariance matrix, can yield substantial bias reduction. Second, they yield asymptotically valid confidence intervals and hypothesis tests in the standard way.

References

[1] Kozachenko, L. F. and Leonenko, N. N. (1987) Sample estimate of the entropy of a random vector, *Probl. Inform. Transm.*, **23**, 95–101.

Nonparametric estimation in a multiplicative censoring model

Session: **Nonparametric Methods 1**

Fabienne Comte *MAP5 Université Paris Descartes, France*
Charlotte Dion *SAMM Paris 1 Sorbonne, France*

Abstract:
We consider the multiplicative noise model, with a uniform noise with mean 1, independent of the signal. This model represents an amplification/attenuation phenomenon of a signal.
We study nonparametric estimation of the density and the corresponding survival function $\overline{F}$ of the hidden variable of interest. Our strategy is based on a projection estimator. Risk bounds in term of integrated squared error are provided, showing that the dimension of the projection space has to be chosen in order to realize a compromise. Thus, a model selection strategy is proposed. The resulting estimators are proven to reach the best possible risk bounds and to be numerically efficient. This work is published in Journal of Nonparametric Statistics ([3]).

1 Introduction

We consider the following model

$$Y_i = X_i U_i,\ i = 1, \ldots, n, \quad U_i \sim \mathcal{U}_{[1-a,1+a]},\ 0 < a < 1 \tag{1}$$

where $(X_i)_{\{i=1,\ldots,n\}}$ and $(U_i)_{\{i=1,\ldots,n\}}$ are two independent samples. The U_i's are independent and identically distributed (*i.i.d.*) random variables from uniform density on an interval $[1-a, 1+a]$ of $\mathbb{R}^+$ with $0 < 1-a < 1+a$ and a is assumed to be known. The X_i's are *i.i.d.* from an unknown density f on $\mathbb{R}^+$. Only the Y_i's are observed. The model implies that they are *i.i.d.* and we denote by f_Y their density on $\mathbb{R}^+$. Our goal is to estimate nonparametrically the density f of the X_i's from the observations Y_i's.
Classical models involving measurement errors are often additive but the multiplicative noise model centred in one appears naturally considering that this additive noise is proportional to the signal.
Equation (1) expresses an approximate transmission of the information: the recorded values Y_i correspond to the value of interest X_i, up to an error of order of $\pm 100a\%$.
Very few studies of this model have been conducted in the literature. We mainly found it in [4], who study a form of data masking. Nevertheless, multiplicative noise models can be found with other distributions for the noise U. The case of U following a uniform distribution on $[0, 1]$ has been introduced by [5] who called it a "multiplicative censoring" model. This model was studied for example in [1], [2].
The strategies on $[0, 1]$ cannot be applied on an other interval. In this work, we build estimators of the density f and of the survival function $\overline{F}$. The operator linking the density of the observations and the density of interest is given by

$$f_Y(y) = \frac{1}{2a} \int_{\frac{y}{1+a}}^{\frac{y}{1-a}} \frac{f(x)}{x} dx, \quad y \in]0, +\infty[, \tag{2}$$

and the inversion of formula (2) is not obvious. Our strategy relies on two steps.

2 Method

First, we approach an auxiliary function g expressed as a function of f and a. We prove that for an explicit transformation $t \in \mathbb{L}^2(\mathbb{R}^+) \mapsto \psi_t$ and this function g in $\mathbb{L}^2(\mathbb{R}^+)$, we have

$$\mathbb{E}[\psi_t(Y_1)] = \langle t, g \rangle, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product of $\mathbb{L}^2(\mathbb{R}^+)$. Relation (3) is used to build projection estimators of g. Indeed, considering the collection of spaces $\mathcal{S}_m = \text{Vect}\{\varphi_0, \varphi_1, \ldots, \varphi_{m-1}\}$ where $(\varphi_j)_{j\geq 0}$ is an orthonormal basis of $\mathbb{L}^2(\mathbb{R}^+)$, the orthogonal projection g_m of g on S_m is given by $g_m = \sum_{j=0}^{m-1} a_j \varphi_j$, with $a_j = \langle g, \varphi_j \rangle$. From relation (3), we notice that $a_j = \mathbb{E}[\psi_{\varphi_j}(Y_1)]$ and replacing the expectation by its empirical counterpart $\widehat{a}_j$, we obtain the estimator $\widehat{g}_m = \sum_{j=0}^{m-1} \widehat{a}_j \varphi_j$. We deduce, by inverting the relation between f and g, collections of estimators of f, for which risk bounds are provided in term of mean integrated squared error on $\mathbb{R}^+$. Model selection criterion are proposed to automatically select m, and to make the adequate tradeoff between bias and variance. The survival function is estimated analogously.

References

[1] Asgharian, M., Carone, M., and Fakoor, V. (2012), *Large-sample study of the kernel density estimators under multiplicative censoring*, Ann.Statist., 40, 159–187.

[2] Belomestny, D., Comte, F., and Genon-Catalot, V. (2016), *Nonparametric Laguerre estimation in the multiplicative censoring model*, Electronic Journal of Statistic, 10, 3114–3152.

[3] Comte, F. & Dion, C. (2016). *Nonparametric estimation in a multiplicative censoring model with symmetric noise*. Journal of Nonparametric Statistics, 28, 768–801.

[4] Sinha, B., Nayak, T., and Zayatz, L. (2011), *Privacy protection and quantile estimation from noise multiplied data*, Sankhya B, 73, 297–315.

[5] Vardi, Y. (1989), *Multiplicative Censoring, Renewal Processes, Deconvolution and Decreasing Density: Nonparametric Estimation*, Biometrika, 76, 751–761.

Outlier Detection in High-Dimensional Data – Applied for Open-Set Text Classification

Session: **High-Dimensional Problems in Engineering**

Szymon Datko *Department of Computer Engineering, Wrocław University of Science and Technology, Poland*
Henryk Maciejewski *Department of Computer Engineering, Wrocław University of Science and Technology, Poland*

Abstract: We have performed an overview on the outlier detection techniques, especially in terms of high-dimensional data. Then we took an effort to apply such techniques in the task of text documents classification. The goal is to implement methods of so called the open-set classification. Although our research is still ongoing, we have already achieved some promising results. During the presentation we would like to introduce the topic, shortly describe our observations and plans for future work.

1 Introduction

One of important application areas of the Text Mining is an automatic subject classification of text documents. Most promising approaches to solve this problem are based on methods of the Natural Language Processing (NLP), coupled with methods of the Machine Learning, [2] [5] [4]. NLP methods are used for constructing feature vectors representing text documents, which typically results in the high-dimensional training data; based on this, supervised-learning algorithms are used then to train classification models. As a result, generated classification models for text documents are based on training collections of data with a known subject category.

2 Main concept

We would like to prepare a procedure of classification that would be appropriate also for the new categories of data, not known in the original training set. An important specific requirement for this procedure is that the text categorization should be performed as the open-set classification. This means that new samples (i.e. documents), which are not similar enough to any of the classes available in the training data, should be labelled as *unrecognized*, not as one of the known classes. Methods of building open-set classifiers based on the high-dimensional training data are in the focus of our work. There are a lot of potential applications for such methods, including improvements to an automatic contextual classification and a detection of new topics in threads.

3 General schema

We propose a generic approach to open-set classification which combines standard closed-set classifiers with methods of outlier detection in high-dimensional data. The idea is that from the second test (outlier detection) we expect to receive some indicator which will describe how the result of closed-set classification is trust-worth.

In case of many doubts, the detailed look at the data and possibly re-training of the model shall be taken into account.

4 Summary

In our work we focus how to solve the challenging task of outlier detection in high-dimensional data. We discuss recently proposed methods of the outlier detection, see e.g. [1], and point the limitations of many of the methods when dealing with high-dimensional data. We present an approach to outlier detection based on angles between (high-dimensional) feature vectors, inspired by [7]. We provide a comprehensive analysis of performance for this approach using simulated data. Finally, we apply this method for analysis of similarity between text documents represented by 'bag-of-words' feature vectors. In the latter analysis we used a collection of English Wikipedia articles grouped into several subject categories. We show that articles not belonging to any of the training classes can be recognized as such using the proposed outlier detection technique.

References

[1] Chandola V., Banerjee A., Kumar V. (2009). *Anomaly detection: A survey*, ACM computing surveys (CSUR), 41(3), 15.

[2] Jurafsky, D., Manning, C. (2012). *Natural Language Processing*, Lecture Slides from the Stanford Coursera course.

[3] Kriegel H.P., Zimek A. (2008). *Angle-based outlier detection in high-dimensional data*, In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 444–452. ACM.

[4] Semberecki P., Maciejewski H. (2016). *Distributed Classification of Text Documents on Apache Spark Platform*, In International Conference on Artificial Intelligence and Soft Computing, pp. 621–630. Springer International Publishing.

[5] Torkkola, K. (2004). *Discriminative features for text document classification*, Formal Pattern Analysis and Applications, 6.4, pp. 301–308

[6] Vidhya K.A., Aghila G. (2010). *A Survey of Naïve Bayes Machine Learning approach in Text Document Classification*, International Journal of Computer Science and Information Security, vol. 7, no. 2, pp. 206–211.

[7] Zimek A., Schubert E., Kriegel H.P. (2012). *A survey on unsupervised outlier detection in high-dimensional numerical data*, Statistical Analysis and Data Mining, 5(5), pp. 363–387.

Testing for Image Symmetries: An Information Theoretic Approach

Session: **High-Dimensional Problems in Engineering**

Agata Migalska *Department of Control Systems and Mechatronics, Wrocław University of Technology, Poland*

Abstract: In this paper a statistical test for symmetry detection in images is proposed. The development of the test is based on the principle that certain statistics of the probability distribution of pixel intensities are preserved when the image is averaged with its copy obtained by applying symmetry transformation to the original image. These statistics include negentropy, an information theoretical measure of normality within a probability distribution. On this basis, testing for symmetry can be simplified to testing for a zero difference in the means of pixel intensities. The improvement over several previous methods for symmetry detection along with the applicability to such tasks as shape inspection and space symmetry detection in crystals, are in support of the claim that our contribution is important in the field of symmetry detection.

1 Introduction

Reliable symmetry detection is crucial to numerous tasks originating from such domains as computer vision and computer graphics, image analysis and understanding, and quality control of manufactured goods, to name a few. Motivated by unsatisfied demands for a robust symmetry detector capable of detecting both rotational and reflectional symmetries in images, allowing to manipulate the precision of detection up to a high level, and having a well-established interpretation, we propose a novel statistical test for symmetry detection in images.

2 The method

The entry point to the proposed test lies in image averaging. Suppose image intensities recorded on a square, equally spaced Cartesian grid. An image is considered symmetric if there exists at least one transformation such that when applied to this image leaves the image identical to the original.
Let us also consider a copy of the original image, obtained by applying certain transformation to it, and an average of the two – of the original and of the copy. An example of these three images is given in Figure 4.
The proposed test is based on a principle that if an image is symmetric and an averaged image is obtained through applying the true symmetry transformation, then the amount of normality within the distribution of the averaged image does not differ from the amount of normality within the distribution of the original one. Conversely, the absence of symmetry or an incorrectly chosen transform will result in the amount of normality in the distribution of the averaged image being altered. The amount of normality is assessed by means of negentropy, a measure originating from information theory.
Utilizing the approximation scheme, in which negentropy is given by the sum of

(a) (b) (c)

Figure 4: (a) Original image. (b) Copy obtained by a transformation T. (c) An averaged image obtained by averaging pixel intensities of the original image (a) and pixel intensities of the transformed copy (c).

expected values of nonlinear functions [1], testing for symmetry can be transformed into the problem of testing for the differences between the two sample means for independent samples of pixel intensities, when the standard deviations are unknown and the samples are large.

3 Results

In order to verify that the accuracy of the proposed information theoretical statistical test for symmetry an extensive experimental verification was conducted on a set of 794 symmetric images. We found that our method outperforms several other methods for symmetry detection. Moreover, our test can be successfully applied to shape inspection, one of the objectives in quality control of manufactured goods, as well as to the space symmetry detection of crystals and molecules in their X-ray diffraction patterns.

Acknowledgements

This work has been supported by the National Science Center under grant: 2012/07/B/ST7/01216, internal code 350914, of the Wrocław University of Technology.

References

[1] Hyvärinen, A., Karhunen, J., & Oja, E. (2004). *Independent component analysis*. John Wiley & Sons.

Hammerstein system nonlinearity recovery by standard and aggregative estimates

Session: **High-Dimensional Problems in Engineering**

Przemysław Śliwiński *Department of Control Systems and Mechatronics, Wrocław University of Science and Technology, Wrocław, Poland*

Paweł Wachel *Department of Control Systems and Mechatronics, Wrocław University of Science and Technology, Wrocław, Poland*

Abstract: We apply two regression estimates, the standard orthogonal series and the aggregative one, to recover a nonlinearity in a discrete-time Hammerstein system. Comparison of their non-asymptotic properties is supplemented by the results of a numerical experiment in which the system nonlinearity is followed by a dynamics containing an accumulation block.

1 Introduction

In the note we compare two algorithms recovering a nonlinearity in a discrete-time Hammerstein system. In both cases the a priori knowledge about the system characteristics is nonparametric. The first algorithm is based on a standard orthogonal series estimate while the other exploits the convex programing approach. Their formal properties are compared for small sample records. Then, the estimates are used in a numerical experiment in which the Hammerstein system of interest has a dynamic subsystem consisting of a non-stable, accumulating block.

1.1 Motivations

There is a vast literature pertaining to identification of block-oriented systems and the Hammerstein systems in particular. The nonparametric orthogonal series algorithms are extensively presented and examined in the textbook [1]. These algorithms are based on a truncated orthogonal series and thus are sometimes referred to as *the linear algorithms* (there were some attempts to use the non-linear approximation schemes instead, but the obtained results were rather mediocre). The orthogonal series algorithms are very simple – they merely consists in computing the estimates of the truncated expansion coefficients. In this note (for the sake of direct comparison with the aggregative algorithms) we assume that the input is a uniformly distributed random signal so that these estimates are just arithmetic means. The aggregative algorithms – in turn – are based on constrained convex optimization and their coefficient estimates are obtained by a convex programming routine. They seem to have promising small sample size properties and work particularly well when the number of coefficients to be evaluated is comparable with (or even grater than) the number of available measurements. The aggregation algorithms were proposed by [2], and then adapted to nonlinear system modelling in [3]. They reveal a resemblance to the popular in signal and image processing areas algorithms based on regularization (with the LASSO method being the most prominent example), however, they do not require the nonlinearity to have a sparse representation.

1.2 Contribution

In the note we examine two regression estimation algorithms applied to recover a static nonlinearity in a block-oriented discrete-time Hammerstein system. We compare their non-asymptotic properties formally and then we check their performance in two numerical experiments in which the measurements:

- come from the system working in a steady-state, or
- are collected as soon as the system starts (so that the system is in a transient state that makes the output process non-stationary).

Similar approach was already used in previously to improve the small sample size properties of the nonparametric kernel regression estimates, however, the transient effect was not taken into account in the formal analysis of the algorithm.

1.3 Applications

The Hammerstein system is an example of a block-oriented nonlinear system and – as a very simple extension of linear systems – can be used to model the input nonlinearity of various devices, including – with the aforementioned modification of the estimation algorithms that allows accumulating dynamics – odometers, charge-amplifiers used in CCD/CMOS imaging or piezoelectric acoustic sensors, or accelerometers (with or without the analog—to-digital converter put prior the accumulation element).
Both algorithms can be adopted to the reverse case when the nonlinearity is preceded by accumulating dynamics – like in Wiener systems. It seems to be particularly simple when a user can control the input signal and make it Gaussian and i.i.d.; see [1].

References

[1] W. Greblicki and M. Pawlak, *Nonparametric System Identification.* New York: Cambridge University Press, 2008.

[2] A. Juditsky and A. Nemirovski, "Functional aggregation for nonparametric regression," *The Annals of Statistics*, vol. 28, no. 3, pp. 681–712, 2000.

[3] P. Wachel and P. Śliwiński, "Aggregative modelling of nonlinear systems," *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1482–1486, 2015.

Statistical investigation of the required space for inland vessels

Session: **High-Dimensional Problems in Engineering**

Nicolas Fischer *Faculty of Transportation and Traffic Sciences "Friedrich List", Dresden University of Technology, Germany*

Ostap Okhrin *Faculty of Transportation and Traffic Sciences "Friedrich List", Dresden University of Technology, Germany*

Abstract: An important factor in the design of fairways for inland waterway traffic is the width a vessel requires for safe navigation. This is not only given by the constant dimensions of the vessel, but highly influenced by its surroundings. Furthermore it depends on the vessel speed which can be regulated by the steersman. We investigate a complex physical model describing these dependencies and looking for a simpler approach in terms of a multiple regression and a kernel supervised principal component regression. Thereby we reduce the dimensionality and identify the most important influencing variables.

1 Introduction

Inland vessels navigating rivers are highly influenced by their surroundings. The vessel demands a part of the fairway width. This part is not only given by its own beam, but by an additional width which highly depends on the physical conditions to which the vessel is exposed. In order to design a fairway, particularly define its width, the widths required for certain combinations of passing vessels have to be considered. Complicated models to determine the additional widths in curves, cross flows and regarding the vessel instability have been developed by the Federal Waterways Research and Engineering Institute, an introduction is given in [2]. These specific additional widths compose the total additional width (Δb) of a single vessel, given by the function

$$\Delta b \;=\; f\left(l_{\mathrm{s}}, b_{\mathrm{s}}, t_{\mathrm{s}}, \gamma, \gamma_{\mathrm{lat}}, h, v_{\mathrm{str}}, \kappa, v_{\mathrm{cross}}, v_{\mathrm{stw}}\right),$$

where the influencing variables can be divided into three main groups:

1. static vessel parameters, i.e. the vessel length l_{s}, beam b_{s}, draught t_{s}, block coefficient γ and the block coefficient of the lateral plane γ_{lat};

2. dynamic physical parameters, i.e. the water depth h, river flow speed v_{str}, the cross flow speed v_{cross} and the track's curvature κ. These variables depend on the vessels position on the river;

3. the vessel speed through water v_{stw} which can be regulated by the steersman.

The signs of κ and v_{cross} describe the direction of the curvature and the cross flow, respectively. Vessels traveling upstream have a negative river flow speed v_{str}.
An example of the high dependence of the total additional width on the physical parameters described in terms of a position on the river is given in figure 5.
We aim to identify the most important influencing variables by a dimension reduction and to find a simple model which is capable of describing the dependency.

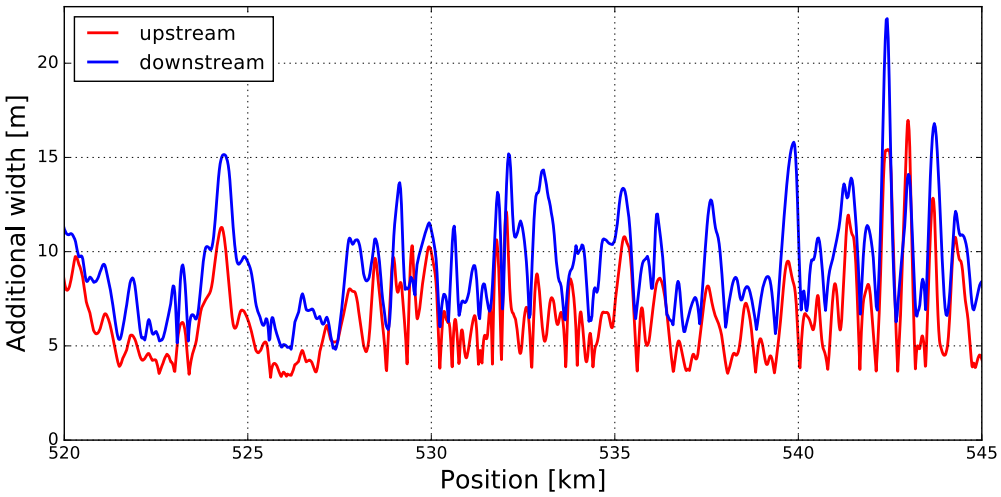


Figure 5: Simulated total additional widths of a typical vessel traveling the river Rhine in up- and downstream direction. Peaks correspond to high curvature or cross flow.

2 The method

To achieve this goal we use two different methods to investigate the dependence of Δb on the input variables. First, we model the additional widths using a simple penalized multiple regression model. Descriptive statistics of the pairwise relationships between the additional width and some variables lead us to use linear as well as nonlinear components, e.g. the dependency on the vessel speed might be quadratic or reciprocal. Obtained results are used as a benchmark to a second model which is a kernel supervised principal component regression (KSPCR) as described in [1]. The KSPCR is used to determine principal components on which a linear regression is applied. The advantage of this approach is that nonlinearities are automatically taken into account by a kernel function. This leads to a better description of the data without providing any information on the probable shape in advance.
Since real data from measurements are not provided, we simulated different vessels on a river for the estimation of both models. We take training and test data sets from the generated data and use the RMSE to compare the accuracy of the models.

References

[1] Barshan, E., Ghodsi, A., Azimifar, Z., & Jahromi, M. Z. (2011). Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44(7), 1357-1371.

[2] Bundesanstalt für Wasserbau (2016). *Driving Dynamics of Inland Vessels*, Karlsruhe.

An Inverse Problem for Infinitely Divisible Moving Average Random Fields

Session: **Statistics of Stochastic Processes 3**

Stefan Roth *Institute of Stochastics, Ulm University, Germany.*
Jochen Glück *Insitute of Applied Analysis, Ulm University, Germany*

Abstract: Given a sample from a discretely observed infinitely divisible moving average random field $\{X(t) = \int_{\mathbb{R}^d} f(t-x)\Lambda(dx);\ t \in \mathbb{R}^d\}$, we study the problem of nonparametric estimation of the Lévy characteristics of the independently scattered random measure Λ.

1 Introduction

Let Λ be a stationary infinitely divisible independently scattered random measure with Lévy characteristics (a_0, b_0, v_0), where $a_0 \geq 0$, $b_0 \in \mathbb{R}$ and v_0 is a Lévy density. Let furthermore $X = \{X(t);\ t \in \mathbb{R}^d\}$ be a moving average infinitely divisible random field on $\mathbb{R}^d$ defined by

$$X(t) = \int_{\mathbb{R}^d} f(t-x)\Lambda(dx), \quad t \in \mathbb{R}^d, \tag{1}$$

with Lévy characteristics (a_1, b_1, v_1), where f is a Λ-integrable function. Suppose a sample $(X(t_1), \ldots, X(t_n))$ from X is available. We studied the problem of nonparametric estimation of (a_0, b_0, v_0).

2 The method

We focused our studies on the estimation of the Lévy density v_0. Therefore we studied the bounded linear operator $\mathcal{G} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$,

$$\mathcal{G}w(x) = \int_{\{u:\ f(u)\neq 0\}} sign(f(u)) w\left(\frac{x}{f(u)}\right) du, \quad w \in L^2(\mathbb{R}), \tag{2}$$

that is motivated by the relation $v_{1,1} = \mathcal{G}v_{0,1}$, for $v_{i,1}(x) = xv_i(x)$, $i = 0, 1$ (see e.g. [5]). It turned out that $\mathcal{G}$ is in fact a multiplication operator whose image can be described by properties of the function f. Under certain conditions the equation $v_{1,1} = \mathcal{G}v_{0,1}$ can explicitely be solved and thus be used to construct a plug-in estimator for v_0. This approach requires an estimator for $v_{1,1}$ that can be obtained in certain cases by using the empirical characteristic function (and its derivatives) of the sample $(X(t_1), \ldots, X(t_n))$, together with Fourier techniques. Those ideas are commonly used in the theory of Lévy processes (see e.g. [1], [2], [3] and many others).

3 Example

Consider a stationary moving average random field

$$X(t) = \int_{\mathbb{R}^d} f(t-x)\Lambda(dx) = \sum_{j=1}^{k} f_j \Lambda(t - \Delta_j), \quad t \in \mathbb{R}^d,$$

for congruent bounded pairwise disjoint Borel sets $\Delta_1, \ldots, \Delta_k$ in $\mathbb{R}^d$, $f(x) = \sum_{j=1}^{k} f_j \mathbb{1}_{\Delta_j}(x)$, $x \in \mathbb{R}^d$ a simple function and $\Lambda(A)$ a homogeneous compound Poisson measure, where the underlying Poisson counting measure is assumed to have intensity $\lambda = 1$. Figure 6 shows the numerical result for $n = 150^2$, $f_1 = 0.95$, $f_2 = f_3 = f_4 = 0.1$ and $|\Delta_j| = 1$, $j = 1, \ldots, 4$.

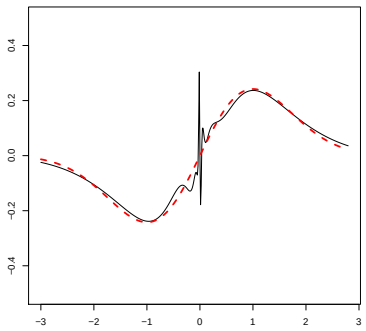


Figure 6: Estimated $v_{0,1}$ (black) vs. true $v_{0,1}$ (red).

References

[1] Comte F. and Genon–Catalot V. (2009), *Nonparametric estimation for pure jump Lévy processes*, pp. 4088–4123, Stoch. Proc. Appl., Vol. 119.

[2] Comte F. and Genon–Catalot V. (2010), *Nonparametric adaptive estimation for pure jump irregularly sampled or noisy Lévy processes*, pp. 290-313, Statistica Neerlandica, Vol. 64.

[3] Neumann M. and Reiss M. (2009), *Nonparametric estimation for Lévy processes from low-frequency observations*, pp. 223–248, Bernoulli, Vol. 15.

[4] Gugushvili S. (2009), *Nonparametric estimation of the characteristic triplet of a discretely observed Lévy process*, pp. 321–343, J. Nonparametr. Stat., Vol. 21.

[5] Rajput B. and Rosinski J. (1989). *Spectral representations of infinitely divisible processes*, pp. 451–487, Probab. Th. Rel. Fields, Vol. 82.

Model-Free Approaches to Discern Non-Stationary Noise in High-Frequency Data

Session: **Statistics of Stochastic Processes 3**

Richard Chen *Department of Statistics, University of Chicago, U.S.A.*
Per Mykland *Department of Statistics, University of Chicago, U.S.A.*

Abstract: In this paper, we provide non-parametric statistical tools to test stationarity of microstructure noise in general hidden Itô semimartingales, and discuss how to measure liquidity risk using high frequency financial data. In particular, we investigate the impact of non-stationary microstructure noise on some volatility estimators, and design three complementary tests by exploiting edge effects, information aggregation of local estimates and high-frequency asymptotic approximation. The asymptotic distributions of these tests are available under both stationary and non-stationary assumptions, thereby enable us to conservatively control type-I errors and meanwhile ensure the proposed tests enjoy the asymptotically optimal statistical power. Besides it also enables us to empirically measure aggregate liquidity risks by these test statistics. As byproducts, functional dependence and endogenous microstructure noise are briefly discussed. Simulation with a realistic configuration corroborates our theoretical results, and our empirical study indicates the prevalence of non-stationary microstructure noise in New York Stock Exchange.

1 Introduction

High-frequency observations of time series on a finite horizon $[0, T]$ are often encountered in electricity networks, climatology, neuroscience, and recently finance. The underlying time series Y , in many cases, can be modeled by an Itô semimartingale X plus a noise process ϵ, i.e. $Y = X + \epsilon$ where

$$X_t = X_0 + \int_0^t b_s \mathrm{d}s + \int_0^t \sigma_s \mathrm{d}W_s + J_t \tag{1}$$

and the noise ϵ_t might be attributable to rounding, measurement errors or intrinsic machineries of underlying stochastic systems.

The main task of our paper is to design some nonparametric tests to investigate the stationarity of the process $g = (\mathbb{E}\epsilon_t^2)$ which are asymptotically optimal in terms of statistical power for any specified type-I error.

Stationarity of the error is important for our choice of estimators of parameters in the process (1) [1, 2]. Besides, statistical properties of ϵ is extremely informative regarding the stochastic system under study, for example, in financial econometrics ϵ is termed as "microstructure noise" and reflects market liquidity [3].

2 The method

Our theory lives in high-frequency asymptotic regime on compact interval allowing irregular sampling. 3 nonparametric tests N, V, $\overline{V}$ are designed as functionals of the sample path of Y, and their central limit theorems under both null and alternative hypotheses are accessible. Under the null hypothesis, all the tests converge to the

standard normal in law; under the alternative, all the statistical powers converge to 1.
The 3 tests are complementary: 1) N has the fastest convergence rate under the null, but V and $\overline{V}$ enjoy large statistical powers; 2) $\overline{V}$ is more accurate in controlling type-I error than V in finite sample, however, the finite-sample power of V is larger than that of $\overline{V}$. How to choose the tests are discussed in the context of high-frequency financial data (an example is shown in Figure 7 below).
Meanwhile, properly scaled $\overline{V}$ converges to the integrate volatility $[g, g]$ and the associated CLT is also available when g is an Itô diffusion. Thus the scaled $\overline{V}$ is a consistent estimate of the level of non-stationarity, hence a estimator of "liquidity risk" in financial application.

3 Example

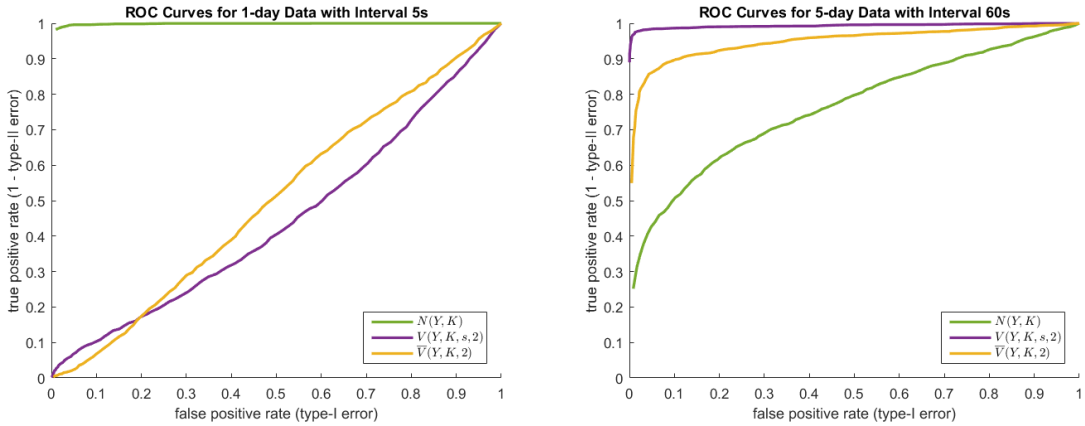


Figure 7: ROC Curves

References

[1] Li, Y. Y., Mykland, P. (2007). Are volatility estimator robust with respect to modeling assumptions, *Bernoulli, Vol. 13(3), 601-622.*

[2] Kalnina, I., Linton, O. (2008). Estimating quadratic variation consistently in the presence of endogenous and diurnal measurement error, *Journal of Econometrics, Vol. 147, 47-59.*

[3] Aït-Sahalia, Y., Yu, J., (2009). High frequency market microstructure noise estimates and liquidity measures *The Annals of Applied Statistics, Vol. 3(1), 422-457.*

Volatility estimation for stochastic PDEs and related processes

Session: **Statistics of Stochastic Processes 3**

Carsten Chong *Department of Mathematics, Technical University of Munich, Germany*

Abstract: The limit theory of realized power variations is well studied for processes like semimartingales or stationary moving averages. By contrast, apart from very particular examples, there has been no systematic analysis of realized power variations for tempo-spatial processes such as stochastic PDEs. In this talk, we discuss work in progress about laws of large numbers and central limit theorems for the realized power variations of stochastic PDEs and related processes. In particular, we show that the form of the integrated power variations heavily depends on whether the underlying equation is, for example, of parabolic or hyperbolic type. From a statistical point of view, this amounts to the estimation of the integrated volatility or intermittency process based on high-frequency observations.

Nonparametric estimation of the coefficients of a jump diffusion

Session: **Statistics of Stochastic Processes 3**

Émeline Schmisser *Laboratoire Painlevé, Université Lille 1*

Abstract: We construct non parametric adaptive L^2 estimators of the coefficients of a jump diffusion and bound their risks.

1 Introduction

Diffusions with jumps become powerful tools to model processes in biology, physics, social sciences, medical sciences, econemics, and finance. They are used in the sudy of dynamical systems when the noise is discontinuous or too intensive to be modeled by a Brownian motion, like polymerarization phenomenons, telephone noise or infinite capacity dam. (see [1] and [4] for instance). They already exists some nonparametric estimators of the diffusion coefficients, but they are pointwise and converge only in the finite intensity case (see [3] and [2]).

2 The method

We consider a diffusion with jumps satisfying the following stochastical differential equation:

$$dX_t = b(X_{t^-})dt + \sigma(X_{t^-})dW_t + \xi(X_{t^-})dL_t, \quad X_0 = \eta$$

with η a random variable, $(W_t)_{t\geq 0}$ a Brownian motion independent of η and $(L_t)_{t\geq 0}$ a pure jump centered Lévy process independent of $(\eta, (W_t)_{t\geq 0})$. This process is observed at discret instant $0, \Delta, \ldots, n\Delta$ where Δ tends to 0 and $n\Delta$ tends to infinity.It is assumed to be ergodic, stationary and exponentially β-mixing. Our aim is to provide nonparametric adaptive estimators of the functions $g = \sigma^2 + \xi^2$, σ^2 and ξ^2 on a compat set A and to bound their risks.
For this purpose, we consider the random variables

$$T_{k\Delta} = \frac{(X_{(k+1)\Delta} - X_{k\Delta})^2}{\Delta} = \sigma^(X_{k\Delta}) + \xi^2(X_{k\Delta}) + noise + remainder,$$

$$Y_{k\Delta} = T_{k\Delta}\mathbf{1}_{|X_{(k+1)\Delta} - X_{k\Delta}| \leq c\Delta^{1/2}} = \sigma^2(X_{k\Delta}) + noise + remainder$$

and

$$Z_{k\Delta} = \frac{(X_{(k+1)\Delta} - X_{k\Delta})^4}{\Delta} = \xi^4(X_{k\Delta^-}) + noise + remainder$$

We then introduce a sequence of increasing vectorial subspaces S_m of $L^2(A)$ and we construct a sequence of nonparametric estimators $(\hat{g}_m, \hat{\sigma}^2_m, \hat{\xi}^4_m)$ by minimizing over each S_m some constrast functions. Indeed,

$$\hat{g}_m = \arg\min_{t\in S_m} \gamma_n(t) \quad \gamma_n(t) = \frac{1}{n}\sum_{k=1}^{n}(T_{k\Delta} - t(X_{k\Delta})^2$$

The risk of this estimator is bounded by a bias term, $\|g - g_m\|^2_{L^2}$, and a variance term, $D_m/(n\Delta)$ (where g_m is the orthogonal projection of g over S_m and D_m is the dimension of S_m). The variance increase when the dimension increases, whereas the bias decreases. To construct the adaptive estimators, we introduce a penalty function proportional to the variance to select the dimension:

$$\hat{m} = \arg \min_{m, D_m \leq \sqrt{n\Delta}} \gamma_n(\hat{g}_m) + pen(m)$$

The estimators $(\hat{g}_{\hat{m}}, \hat{\sigma}^2_{\hat{m}}, \hat{\xi}^4_{\hat{m}})$ automatically realise a bias-variance compromise and satisfy an oracle inequality.

References

[1] Aït-Sahaalia, Y. and Jacod, J. (2009) Estimation the degree of activity of jumps in high frequency data. *Ann. Statist.*,37(5A):2202-2244.

[2] Hanif, M., Wang, H., Lin, Z. (2012). Reweighted Nadaray-Watson estimation of jump-diffusion models. *Sci. China Math.*, 55(5):1005-1016.

[3] Mancini, C. and Renò, R. (2011). Theshold estimation of Markov models zith jumps and interest rate modeling. *J. Econometrics*, 160(1):77-92

[4] Protter, P. and Talay, D. (1997). The Euler scheme for Lévy driven stochastic differential equations. *Ann. Probab.*, 25(1):393-423.

[5] Schmisser, e, (2014). Nonparametric estimation of the coefficients of a diffusion with jumps, preprint.

Modeling Zero Inflation in Count Data Time Series with Bounded Support

Session: **Discrete-Valued Time Series 3**

Tobias Möller *Helmut Schmidt University, Hamburg, Germany*
Christian H. Weiß *Helmut Schmidt University, Hamburg, Germany*
Hee-Young Kim *Helmut Schmidt University, Hamburg, Germany*
Andrei Sirchenko *Higher School of Economics, Moscow, Russia*

Abstract: Real count data time series often show an excessive number of zeros, which can form quite different patterns. We develop four extensions of the binomial autoregressive model for autocorrelated counts with a bounded support, which can accommodate a broad variety of zero patterns. Stochastic properties of these models are derived, ways of parameter estimation and model identification are discussed. The usefulness of the models is illustrated, among others, by an application to the monetary policy decisions of the National Bank of Poland.

1 Introduction

We start with the *binomial autoregressive model of order 1 (BAR(1) model)* by [1], which uses the *binomial thinning* operation "$\circ$" of [2].

Definition 1.1 *Let* $\rho \in \left(\max\left\{-\frac{\pi}{1-\pi}, -\frac{1-\pi}{\pi}\right\}\ ;\ 1\right)$ *and* $\pi \in (0;1)$. *Define* $\beta := \pi\,(1-\rho)$ *and* $\alpha := \beta + \rho$. *Fix* $n \in \mathbb{N}$. *A BAR(1) process* $(X_t)_{\mathbb{N}}$ *with finite support* $\{0, 1, \ldots, n\}$ *is defined by the recursion*

$$X_t \;=\; \alpha \circ X_{t-1} \;+\; \beta \circ (n - X_{t-1}) \qquad \text{with } X_0 \sim \mathrm{Bin}(n, \pi),$$

where thinnings are performed independently and independent of $(X_s)_{s<t}$.

2 BAR(1)-Extensions to Model Zero Inflation

We denote the BAR(1) process by $(X_t)_{\mathbb{N}}$, the zero-inflated processes by $(Z_t)_{\mathbb{N}}$.

- BAR(1) Process with Zeros at Random:

$$Z_t = \kappa_t \circ X_t \quad \text{with } \kappa_t \sim \mathrm{Bin}(1, 1-\omega), \tag{1}$$
$$X_t = \alpha \circ X_{t-1} \;+\; \beta \circ (n - X_{t-1}). \tag{2}$$

- BAR(1) Process with Innovational Zeros:

$$Z_t = \kappa_t \circ X_t \quad \text{with } \kappa_t \sim \mathrm{Bin}(1, 1-\omega), \tag{3}$$
$$X_t = \alpha \circ Z_{t-1} \;+\; \beta \circ (n - Z_{t-1}). \tag{4}$$

- ZIB-AR(1) Process using ZIB Thinning:
(using *zero-inflated binomial thinning*, see [3])

$$Z_t \;=\; (\alpha, \omega_\alpha) \odot Z_{t-1} \;+\; (\beta, \omega_\beta) \odot (n - Z_{t-1}). \tag{5}$$

- BAR(1) Process with Zero Threshold:

	Model			Zero pattern		
	Markov	CLAR(1)	No. param.	scattered	long runs	jumps
RZ-BAR(1)	✘	✘	3	✔	✘	up & down
IZ-BAR(1)	✔	✔	3	(✔)	✔	down
ZIB-AR(1)	✔	✔	3–4	✔	✔	up & down
ZT-BAR(1)	✔	✘	2–3	✘	✔	✘

Table 1: Characteristic features of zero-inflated BAR(1) models.

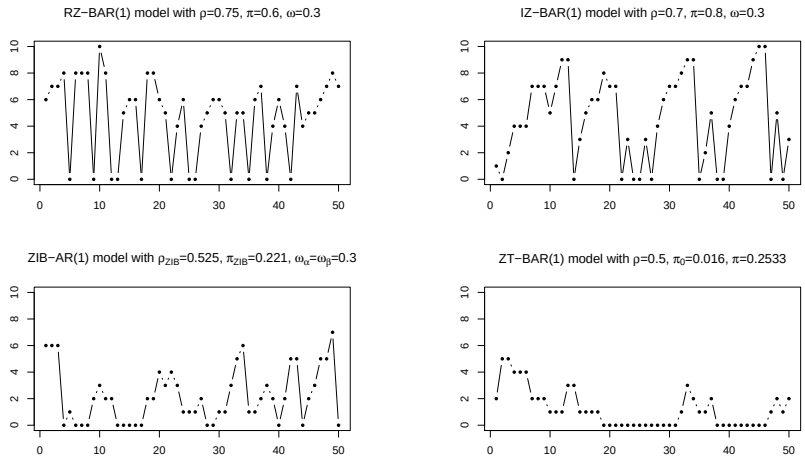


Figure 8: Simulated sample paths for zero-inflated BAR(1) models.

(adapts "LSET-BAR(1) model" by [4], *self-exciting threshold* mechanism)

$$Z_t \;=\; \begin{cases} \beta_0 \circ n & \text{if } Z_{t-1} = 0, \\ \alpha \circ Z_{t-1} \;+\; \beta \circ (n - Z_{t-1}) & \text{if } Z_{t-1} > 0. \end{cases} \tag{6}$$

Model properties and features of sample paths are surveyed in Table 1. The example paths shown in Figure 8 illustrate, e. g., the "shark fin pattern" of the IZ-BAR(1) model, or the tendency to long runs of the ZT-BAR(1) model.

References

[1] McKenzie, E. (1985) Some simple models for discrete variate time series. *Water Resources Bulletin* **21**(4), 645–650.

[2] Steutel, F.W., van Harn, K. (1979) Discrete analogues of self-decomposability and stability. *Annals of Probability* **7**(5), 893–899.

[3] Weiß, C.H., Homburg, A., Puig, P. (2016) Testing for zero inflation and overdispersion in INAR(1) models. *Statistical Papers*, in press.

[4] Möller, T.A., Silva, M.E., Weiß, C.H., Scotto, M.G., Pereira, I. (2016) Self-exciting threshold binomial autoregressive processes. *AStA Advances in Statistical Analysis* **100**(4), 369–400.

Evaluating Approximations of Count Data Distributions

Session: **Discrete-Valued Time Series 3**

Annika Homburg *Department of Mathematics and Statistics, Helmut Schmidt University, Germany*

Abstract: Key indicators are specified to evaluate approximations of count data distributions considering different application scenarios and their specific requirements. This is illustrated by investigating the goodness of normal approximations to the Poisson distribution. By means of an INAR(1) model, the application to count data process approximations is outlined.

1 Introduction

We find many examples of count data process approximations in the literature. However, the goodness of such an approximation depends on the requirements of its application scenario. The evaluation of the approximate probability functions is therefore as essential as the evaluation of certain approximated characteristics of the target process. In this paper a program will be defined to evaluate approximations of count data distributions. An outlook exemplifies its application on the evaluation of count data process approximations. Here, the difficulties and first results of the approximation of a Poisson INAR(1) forecast distribution will be outlined.

2 The evaluation program

We approximate the distribution function of a random count data variable X and start the evaluation program with a visual comparison of true and approximated probability mass and distribution function. The diversity of the target probability mass function p_1 and its approximation p_2 is further quantified by the *Kullback Leibler Divergence* measure

$$d_{KL}(p_1, p_2) = \sum_x p_1(x) \ln\left(\frac{p_1(x)}{p_2(x)}\right)$$

with $d_{KL} \approx 0 \;\Leftrightarrow\; p_1 \approx p_2$. Having evaluated the "overall goodness", we take a look at more specific characteristics of the target distribution. Here, we consider the approximation of mean, variance, zero probability and certain quantiles. Furthermore, the approximation of the risk measures *Value at Risk* (VaR), *Tail Conditional Expectation* (TCE) and *Expected Shortfall* (ES) with a risk level ρ is discussed:

$$\begin{aligned}
\mathrm{VaR}_\rho &= \min\{x \in \mathbb{N}_0 | P(X \leq x) \geq \rho\} = x_\rho \\
\mathrm{TCE}_\rho &= \frac{\mu - \sum_{x=0}^{\mathrm{VaR}_\rho} x \cdot P(X = x)}{1 - F(\mathrm{VaR}_\rho)} \\
\mathrm{ES}_\rho &= \mathrm{TCE}_\rho \times \frac{1 - F(\mathrm{VaR}_\rho)}{1 - \rho} + \mathrm{VaR}_\rho \times \frac{F(\mathrm{VaR}_\rho) - \rho}{1 - \rho}
\end{aligned}$$

3 Example: Normal approximation of Poisson distribution

We approximate the distribution function of $X \sim \mathrm{Poi}(\lambda)$ by the simple normal approximation $\Phi\left(\frac{x-\lambda}{\sqrt{\lambda}}\right)$ and the continuity corrected normal approximation $\Phi\left(\frac{x-\lambda+0.5}{\sqrt{\lambda}}\right)$

for $x \in \mathbb{N}_0$. We determine true and approximated probability mass function, mean and variance, zero probability, quantiles and the risk measures VaR_ρ, TCE_ρ and ES_ρ. While the zero probability is poorly approximated by both methods, the continuity corrected normal approximation provides clearly better results for the approximation of probability mass and distribution function, mean and central quantiles. The upper extreme quantiles and the associated risk measures are better approximated by the simple normal approximation. The variance of the target distribution is never exactly approximated due to an additional variance caused by the discretization. We see that the goodness of an approximation has to be evaluated concerning the varying requirements of its application scenario, even in the commonly used case of a normal approximation. An insufficient evaluation may lead to poor approximations remaining undetected.

References

[1] Alwan, L.C., Weiß C.H. (2016) INAR implementation of newsvendor model for serially dependent demand counts. *International Journal of Production Research*, to appear.

[2] Maiti, R., Biswas, A., Das, S. (2016) Coherent forecasting for count time series using Box-Jenkinss AR(p) model. *Statistica Neerlandica* **70**(2), 123–145

[3] Göb, R. (2011) Estimating Value at Risk and Conditional Value at Risk for Count Variables. *Quality and Reliability Engineering International* **27**, 659–672

[4] Johnson, N.L., Kemp, A.W., Kotz, S. (2005) Univariate discrete distributions. – 3rd ed. *Wiley Series in Probability and Statistics*

[5] Parzen, E. (1996) Concrete statistics. *Statistics of Quality* (S. Ghosh, W. Schucany and W. Smith, eds.) 309–332.

[6] Schneeweiss, H., Komlos, J. & Ahmad, A.S. (2010) Symmetric and asymmetric rounding: a review and some new results. *AStA Advances in Statistical Analysis* **94**(3), 247–271.

A Full ARMA Model for Counts with Bounded Support

Session: **Discrete-Valued Time Series 3**

Tobias Möller *Department of Mathematics and Statistics, Helmut Schmidt University, Germany*
Christian H. Weiß *Department of Mathematics and Statistics, Helmut Schmidt University, Germany*
Sónia Gouveia *Institute of Electronics and Informatics Engineering and Centre for R&D in Mathematics and Applications, University of Aveiro, Portugal*
Manuel G. Scotto *Department of Mathematics, University of Lisbon, Portugal*

Abstract: We introduce a generalization of the NDARMA model by [1] for count data time series. One main characteristic of the NDARMA model is the tendency to show runs of values, which is untypical for real count data. The newly proposed model does not show this characteristic, but preserves properties of ARMA-like models.

1 Introduction

The NDARMA(p, q) model proposed by [1] offers a model for count data processes with an ARMA-like structure, also for processes with a bounded support. However, it has the tendency to show long runs of values. This is not a feature of most real data examples. We will introduce a generalization of the NDARMA(p, q) process that generates higher volatility.

2 Binomial variation

To introduce more volatility into the process we define the operation of binomial variation for a random variable X with range $\{0, \ldots, n\}$ as

$$\mathrm{bv}_n(X)\,\Big|_X \;\sim\; \mathrm{Bin}\big(n,\; X/n\big), \tag{1}$$

where Bin is the binomial distribution.
By conditioning it follows that

$$\begin{aligned} E\big[\mathrm{bv}_n(X)\big] &= E[X], \\ V\big[\mathrm{bv}_n(X)\big] &= (1-\tfrac{1}{n})\,V[X] \;+\; E[X]\,(1-\tfrac{1}{n}\,E[X]). \end{aligned} \tag{2}$$

The first equation shows that the binomial variation maintains the mean of X, while the second equation shows the enlarged variance. The binomial index of dispersion

$$I_X := \frac{nV[X]}{E[X](n-E[X])} \tag{3}$$

supports the exhibition of extra-binomial variation:

$$I_{\mathrm{bv}_n(X)} \;=\; 1 + \left(1-\frac{1}{n}\right) I_X \;> 1. \tag{4}$$

3 The bvARMA(p, q) model

Let $(X_t)_{\mathbb{Z}}$ denote the observations process with range $\{0, \ldots, n\}$. Let $(\epsilon_t)_{\mathbb{Z}}$ be the process of i.i.d. innovations with range $\{0, \ldots, n\}$, where ϵ_t is independent of

$(X_s)_{s<t}$. Let the random vectors $(D_{t,-q}, \ldots, D_{t,0}, \ldots, D_{t,p})$ be independent and multinomially distributed according to $MULT(1;\ \phi_{-q}, \ldots, \phi_0, \ldots, \phi_p)$ with $\phi_{-q} + \ldots + \phi_0 + \ldots + \phi_p = 1$ and $0 < \phi_k < 1$ for $k = -q, \ldots, 0, \ldots, p$, being also independent of $(\epsilon_t)_{\mathbb{Z}}$ and of $(X_s)_{s<t}$. $(X_t)_{\mathbb{Z}}$ is said to follow a bvARMA(p, q) model if

$$X_t \ = \ \sum_{i=1}^{p} D_{t,i} \, \mathrm{bv}_n(X_{t-i}) \ + \ D_{t,0} \, \epsilon_t \ + \ \sum_{j=1}^{q} D_{t,-j} \, \mathrm{bv}_n(\epsilon_{t-j}). \tag{5}$$

The NDARMA(p, q) model would be obtained by replacing the binomial variation with the deterministically identity function.
The talk will consider model properties, special instances of the model and discuss the topic of parameter estimation. Furthermore, some examples will be shown.

References

[1] Jacobs, P.A., Lewis, P.A.W. (1983). Stationary discrete autoregressive-moving average time series generated by mixtures. *Journal of Time Series Analysis* ***4****(1), 19–36.*

A Goodness-of-Fit Test for Integer-Valued Autoregressive Processes

Session: **Discrete-Valued Time Series 3**

Sebastian Schweer *P3 automotive GmbH, Stuttgart, Germany*

Abstract: In this talk, a goodness-of-fit test for autoregressive count data time series based on the empirical joint probability generating function is considered. The underlying process is contained in a general class of Markovian models satisfying a drift condition.

Asymptotic theory for the test statistic is provided, including a functional central limit theorem for the non-parametric estimation of the stationary distribution and a parametric bootstrap method. Connections between the new approach and existing tests for count data time series based on moment estimators appear in limiting scenarios. Finally, the test is applied to a real data set.

Inference for High-Dimensional Split-Plot-Designs

Session: **Analysis, Testing and Change Detection in High Dimensions 3**

Markus Pauly *University of Ulm, Institute of Statistics*
Paavo Sattler *University of Ulm, Institute of Statistics*

Abstract: Statisticians increasingly face the problem to reconsider the adaptability of classical inference techniques. In particular, diverse types of high-dimensional data structures are observed in various research areas; disclosing the boundaries of conventional multivariate data analysis. Such *large d and small n problems* are, e.g., met in life sciences whenever it is easier or cheaper to repeatedly generate a large number d of observations per subject than recruiting many, say n, subjects. In this talk we discuss inference procedures for such situations in repeated measures and heteroscedastic split plot designs. These will, e.g., be able to answer questions about the occurrence of certain time, group and interactions effects or about particular profiles. The test procedures are based on standardized quadratic forms involving unbiased and dimension-stable estimators of different traces of the underlying unrestricted covariance structures. In a general asymptotic framework ($\min(n, d) \to \infty$) its limit distributions are analyzed and additional small sample approximations are proposed. Beneath discussing the theoretical properties, its small sample performance is investigated in simulations. Moreover, the procedure is illustrated by a real data application

References

[1] Bai, Z. and Saranadasa, H. (1996). Effect of High Dimension:by an Example of a Two Sample Problem. *Statistica Sinica* **6**, 311–329.

[2] Chen, S.X. and Qin, Y.-L. (2010). A Two-Sample Test for High-Dimensional Data with Applications to Gene-Set Testing. *The Annals of Statistics* **38**, 808–835.

[3] Gregory, K. B., Carroll, R. J., Baladandayuthapani, V., and Lahiri, S. N. (2015). A two-sample test for equality of means in high dimension. *Journal of the American Statistical Association*, **110**, 837–849.

[4] Jordan, W., Tumani, H., Cohrs, S., Eggert, S., Rodenbeck, A., Brunner, E., Rüther, E., and Hajak, G. (2004). Prostaglandin D Synthase (β-trace) in Healthy Human Sleep. *Sleep* **27**, 867–874.

[5] Pauly, M., Ellenberger, D., and Brunner, E. (2015). Analysis of High-Dimensional One Group Repeated Measures Designs. *Statistics* **9**, 1243–1261.

[6] Srivastava, M.S. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *Journal of Multivariate Analysis* **100**, 518–532.

Exact and Asymptotic Tests on a Factor Model in Low and Large Dimensions with Applications

Session: **Analysis, Testing and Change Detection in High Dimensions 3**

Taras Bodnar *Department of Mathematics, Stockholm University, Sweden*
Markus Reiß *Department of Mathematics, Humboldt-University zu Berlin, Germany*

Abstract: We suggest three tests on the validity of a factor model which can be applied for both, small-dimensional and large-dimensional data. The exact and asymptotic distributions of the resulting test statistics are derived under classical and high-dimensional asymptotic regimes. It is shown that the critical values of the proposed tests can be calibrated empirically by generating a sample from the inverse Wishart distribution with identity parameter matrix. The powers of the suggested tests are investigated by means of simulations. The results of the simulation study are consistent with the theoretical findings and provide general recommendations about the application of each of the three tests. Finally, the theoretical results are applied to two real data sets, which consist of returns on stocks from the DAX index and on stocks from the S&P 500 index. Our empirical results do not support the hypothesis that all linear dependencies between the returns can be entirely captured by the considered factors.

1 Introduction

Let X_{it} be the observation data for the i-th cross-section unit at time t. For instance, in the case of portfolio theory, X_{it} represents the return of the i-th asset at time t. Let $\mathbf{X}_t = (X_{1t}, \ldots, X_{pt})^\top$ be the observation vector at time t and let $\mathbf{f}_t$ be a K-dimensional vector of common observable factors at time t. Then the factor model in vector form is expressed as

$$\mathbf{X}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t \tag{1}$$

where $\mathbf{B}$ is the matrix of factor loadings and $\mathbf{u}_t$, $t = 1, \ldots, T$, are independent errors with covariance matrix $_u$. It is also assumed that $\mathbf{f}_t$ are independent in time as well as independent of $\mathbf{u}_t$.
Under the generic assumption that $_u$ is a diagonal matrix, the dependence between the elements of $\mathbf{X}_t$ is fully determined by the factors $\mathbf{f}_t$. This means that the precision matrix of $\mathbf{Y}_t = (\mathbf{X}_t^\top, \mathbf{f}_t^\top)^\top$ has the following structure

$$= \{\mathrm{cov}(\mathbf{Y}_t)\}^{-1} = \begin{pmatrix} _{11} & _{12} \\ _{21} & _{22} \end{pmatrix}, \tag{2}$$

where $_{21} = {}^\top_{12}$ is a $p \times K$ matrix and $_{11}$ is a diagonal $p \times p$ matrix, if the factor model (1) is true, i.e., if all linear dependencies among the components of $\mathbf{X}_t$ are fully captured by the factor vector $\mathbf{f}_t$. As a result, the test on the validity of the factor model (1) is equivalent to testing

$$H_0 : _{11} = \mathrm{diag}(\omega_{11}, \ldots, \omega_{pp}) \qquad \text{versus} \qquad H_1 : _{11} \neq \mathrm{diag}(\omega_{11}, \ldots, \omega_{pp}) \tag{3}$$

for some positive constants $\omega_{11}, \ldots, \omega_{pp}$.

2 Three tests

A test on the hypothesis (3) can be performed in three different ways.

- Approach 1: All non-diagonal elements of $_{11}$ are equal to zero

  $$H_0 : \omega_{ij} = 0, 1 \leq j < i \leq p \quad \text{against} \quad H_1 : \omega_{ij} \neq 0 \text{ for at least one } (i,j)$$

  where $= (\omega_{ij})_{i,j\in\{1,\ldots,p+T\}}$.
- Approach 2: $^{-1}_{11} = (\omega^{(-)}_{ij})_{i,j\in\{1,\ldots,p\}}$

  $$H_0 : \omega_{jj}\omega^{(-)}_{jj} = 1, 1 \leq j \leq p \quad \text{against} \quad H_1 : \omega_{jj}\omega^{(-)}_{jj} > 1 \text{ for at least one } j$$
- Approach 3: Hadamard's inequality (for any positive definite symmetric matrix $\mathbf{A}$ we get $det(\mathbf{A}) \leq \prod_{i=1}^{p} a_{ii}$ with equality iff $\mathbf{A}$ is diagonal)

  $$H_0 : \frac{\prod_{i=1}^{p} \omega_{ii}}{det(_{11})} = 1 \qquad \text{against} \qquad H_1 : \frac{\prod_{i=1}^{p} \omega_{ii}}{det(_{11})} > 1\,.$$

3 Analysis of Stocks Included into the DAX Index

We perform the T_1, T_2, and T_3 tests on the validity of factor models fitted to the returns of 20 stocks included into the DAX index. These 20 stocks are chosen randomly out of all 30 stocks which determine the value of the DAX index. Repeating this procedure 10^4 times, 10^4 models are fitted and tests on the validity of each model are performed. As factors, we use the returns of the DAX index in the first approach. In the second approach, we included three further factors, namely, the STOXX50E index, the TecDAX index, and the MDAX index. In all cases, weekly returns are considered from the 11th of June 2012 to the 10th of June 2014 ($T = 104$ observations) obtained from the Yahoo! finance web-page.

$K = 1$					
Test\ Quantile	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
T_1	0.0000	0.0000	0.0000	0.0002	0.7296
T_2	0.0000	0.0000	0.0000	0.0000	0.0270
T_3	0.0000	0.0000	0.0000	0.0000	0.0000
$K = 4$					
Test\ Quantile	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
T_1	0.0000	0.0002	0.0002	0.0011	0.9241
T_2	0.0000	0.0000	0.0000	0.0000	0.1024
T_3	0.0000	0.0000	0.0000	0.0000	0.0000

Table 2: Quantiles of the p-values calculated from the empirical distribution functions T_1, T_2, and T_3 with $p = 20$, $T = 104$, and $K \in \{1, 4\}$.

References

[1] Bodnar, T., Reiß (2016). Exact and Asymptotic Tests on a Factor Model in Low and Large Dimensions with Applications, *Journal of Multivariate Analysis* **150**, 125-151.

Power Analysis of Tests on Independence of Large Dimensional Variables

Session: **Analysis, Testing and Change Detection in High Dimensions 3**

Taras Bodnar *Department of Mathematics, Stockholm University, SE-10691 Stockhom, Sweden*
Holger Dette *Department of Mathematics, Ruhr University Bochum, D-44870 Bochum, Germany*
Nestor Parolya *Institute of Statistics, Leibniz University Hannover, D-30167 Hannover, Germany*

Abstract: In the paper, we derive three tests motivated from MANOVA on the independence of two sets of many variates, i.e., block-diagonality of high-dimensional covariance matricx. In our study the dimension p of the covariance matrix is increasing together with the sample size n so that $p/n \to c > 0$. Moreover, the dimension of each ith block may grow to the infinity as well, i.e., $p_i/n \to c_i < 1$, $i \in \{1, 2\}$. Under alternative hypothesis we derive the powers of all test based on the central Fisher matrices. The main theoretical contribution is the proof of the central limit theorem for the linear spectral statistics of the large non-central Fisher matrices in the case when the non-centrality matrix is of full rank.

1 Introduction

The present paper is devoted to the problem of testing the independence of the sets of variables, i.e., block-diagonal structure of the covariance matrix from the point of view of MANOVA. The related topic has been discussed under the assumption of normality via likelihood ratio criterion where the aim was to test the independence between several groups of random variables (see, e.g., Yao et al. (2015)). We contribute to the existent literature by deriving further methods which are based on large dimensional random matrix theory. The results are obtained in the case of the finite number of blocks, where the dimension of each block has the same order as the sample size. In the derivation, we make use of the distributional properties of Wishart random matrices and the recent results from random matrix theory dealing with central and non-central Fisher random matrices. The asymptotic distributions of the test statistics are derived under both the null and the alternative hypotheses. Moreover, in order to find the powers of the proposed tests we prove a new central limit theorem for non-central Fisher random matrix with non-centrality parameter matrix of full rank.

2 The method

Let $\mathbf{x}_1, ..., \mathbf{x}_n$ be a sample of i.i.d. observations with $\mathbf{x}_1 \sim \mathcal{N}_p(\mathbf{0},)$ (p-dimensional normal distribution with zero mean vector and covariance matrix). We define the $p \times n$ dimensional observation matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ and denote by

$$\mathbf{S} = \frac{1}{n}\mathbf{X}\mathbf{X}' \tag{1}$$

the sample covariance matrix, which is used as an estimate of . Using the relationship between the matrix variate normal distribution and the Wishart distribution we get that $n\mathbf{S}$ has a p-dimensional Wishart distribution with n degrees of freedom and covariance matrix , i.e., $n\mathbf{S} \sim W_p(n,)$.

Recalling the definition of the matrices and $\mathbf{S}$ we introduce the decomposition

$$= \begin{pmatrix} \widetilde{11} & \widetilde{12} \\ \widetilde{21} & \widetilde{22} \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \widetilde{\mathbf{S}}_{12} \\ \widetilde{\mathbf{S}}_{21} & \widetilde{\mathbf{S}}_{22} \end{pmatrix},$$

where the matrices $\widetilde{12}$, $\widetilde{\mathbf{S}}_{12}$ are $p_1 \times (p-p_1)$ matrices and $\widetilde{11}$, $\widetilde{\mathbf{S}}_{11}$ are $p_1 \times p_1$ matrices. We further denote by $\widetilde{22\cdot 1} = \widetilde{22} - \widetilde{21}\widetilde{11}^{-1}\widetilde{12}$ and $\widetilde{\mathbf{S}}_{22\cdot 1} = \widetilde{\mathbf{S}}_{22} - \widetilde{\mathbf{S}}_{21}\mathbf{S}_{11}^{-1}\widetilde{\mathbf{S}}_{12}$ the corresponding Schur complements. It can be shown that

$$\begin{aligned} p_1\mathbf{W} &= \widetilde{\mathbf{S}}_{21}\mathbf{S}_{11}^{-1}\widetilde{\mathbf{S}}_{12} \sim W_{p-p_1}(p_1, \widetilde{22\cdot 1}), \\ (n-p_1)\mathbf{T} &= \widetilde{\mathbf{S}}_{22\cdot 1} \sim W_{p-p_1}(n-p_1, \widetilde{22\cdot 1}), \end{aligned}$$

and $\mathbf{W}$ and $\mathbf{T}$ are independent. Under the alternative hypothesis H_1, $\mathbf{W}$ possess non-central Wishart distribution conditionally on $\mathbf{S}_{11}$
The distributional properties of $\mathbf{W}$ and $\mathbf{T}$ are very similar to the ones observed for the within and between covariance matrices in the multivariate ANOVA model. This similarity motivates the application of three tests which are usually used in the MANOVA. They are given by

(i) Wilks' Λ statistics: $T_W = -\log(|\mathbf{T}|/|\mathbf{T}+\mathbf{W}|) = \log(|\mathbf{I}+\mathbf{W}\mathbf{T}^{-1}|) = \sum_{i=1}^{p-p_1}\log(1+v_i)$,

(ii) Lawley-Hotelling's trace criterion: $T_{LH} = tr(\mathbf{W}\mathbf{T}^{-1}) = \sum_{i=1}^{p-p_1} v_i$,

(iii) Bartlett-Nanda-Pillai's trace criterion: $T_{BNP} = tr(\mathbf{W}\mathbf{T}^{-1}(\mathbf{I}+\mathbf{W}\mathbf{T}^{-1})^{-1}) = \sum_{i=1}^{p-p_1}\frac{v_i}{1+v_i}$,

where $v_1 \geq v_2 \geq ... \geq v_{p-p_1}$ denote the ordered eigenvalues of the matrix $\mathbf{W}\mathbf{T}^{-1}$.
It is remarkable that all of the three test statistics are functions of eigenvalues of $\mathbf{W}\mathbf{T}^{-1}$ only and, consequently, they can be presented as linear spectral statistics calculated for the random matrix $\mathbf{W}\mathbf{T}^{-1}$, which is the so-called Fisher matrix under the null hypothesis H_0 (cf. Zheng (2012)) and, moreover, under the alternative H_1 conditionally on $\mathbf{S}_{11}$ it is a noncentral Fisher matrix. We make use of the above observation and derive the central limit theorems for the proposed test statistics in case of both H_0 and H_1 hypotheses.

References

[1] Yao, J., Bai, Z. and Zheng, S. (2015). Large Sample Covariance Matrices and High-Dimensional Data Analysis, number 39, Cambridge University Press.

[2] Zheng, S. (2012). Central limit theorems for linear spectral statistics of large dimensional F-matrices. In Annales de l'Institut Henri Poincaré, Probabilités et Statistiques (Vol. 48, No. 2, pp. 444-476). Institut Henri Poincaré.

Heavy-tailed - based approach for signals modeling in application to technical diagnostics

Session: **Analysis, Testing and Change Detection in High Dimensions 3**

Grzegorz Żak *Diagnostics and Vibro-Acoustics Science Laboratory, Wroclaw University of Science and Technology, Na Grobli 15, 50-421 Wroclaw, Poland*
Agnieszka Wyłomańska *Hugo Steinhaus Center, Faculty of Pure and Applied Mathematics, Wroclaw University of Science and Technology, Janiszewskiego 14a, 50-370 Wroclaw, Poland*
Radosław Zimroz *Diagnostics and Vibro-Acoustics Science Laboratory, Wroclaw University of Science and Technology, Na Grobli 15, 50-421 Wroclaw, Poland*

Abstract: Recent years have shown that there is increasing trend in application of heavy-tailed based processes in multiple fields. One of these fields is condition monitoring. It plays major role in predictive maintenance of machines. The use of the condition monitoring allows one for the maintenance to be scheduled accordingly to the state of the machine components. It has main advantage as one can address the developing fault before it leads to the major failure of the whole machine. Usage of heavy-tailed distributions in this field shows its importance when dealing with data possessing impulsive components. Such impulsive component is related to the aforementioned developing fault. One of the members of heavy-tailed distributions family is the α-stable distribution. The class of this distribution is especially important in the context of modeling of data with impulsive nature but it should be mentioned that for stability parameter (called α parameter) equal to 2 the stable distribution reduces to Gaussian one. However application of stable distribution requires novel methods for analysis of the signal. Especially it is related to the fact that for stable distribution the classical measures of dependence as autocovariance or autocorrelation cannot be used because for most of the parameters of stable distribution they are not properly defined. Therefore we consider here alternative measures. One can recall few of them, such as codifference, covariation and fractional lower-order covariance. On the basis of the time-frequency maps constructed using those measures we can enhance the classical spectrogram in order to detect impulsive nature of signals with complex behavior. Proper application of the heavy-tailed counterparts of standard dependency measures results in better and more precise local damage detection. In the analysis we will use real world data acquired from the mining industry.

Efficient asymptotic variance reduction when estimating volatility in high frequency data

Session: **Statistics of Stochastic Processes 4**

Yoann Potiron *Faculty of Business and Commerce, Keio University, Tokyo*
Simon Clinet *Graduate School of Mathematical Sciences, The University of Tokyo*

Abstract: This paper shows how to carry out efficient asymptotic variance reduction when estimating volatility in the presence of stochastic volatility and microstructure noise with the realized kernels (RK) from [1] and the quasi-maximum likelihood estimator (QMLE) studied in [2]. To obtain such a reduction, we chop the data into B blocks, compute the RK (or QMLE) on each block, and aggregate the block estimates. The ratio of asymptotic variance over the bound of asymptotic efficiency converges as B increases to the ratio in the parametric version of the problem, i.e. 1.0025 in the case of the fastest RK Tukey-Hanning 16 and 1 for the QMLE. The finite sample performance of both estimators is investigated in simulations, while empirical work illustrates the gain in practice.

Over the past decades, the availability of high frequency data has led to a better understanding of asset prices. The main object of interest, the quadratic variation, can be used for example as a proxy for the spot volatility or the volatility parameter of a time-varying model. Moreover, forecasts of future volatility can be improved with it. Without microstructure noise, the realized variance (RV) estimator is both consistent and efficient. The convergence rate $n^{1/2}$ and the asymptotic variance (AVAR) were established.

Under market frictions, the RV is no longer consistent. Zhang et al. (2005) bring forward the Two-Scale Realized Volatility nonparametric estimator, the first consistent estimator in the presence of noise and with a relatively slow convergence rate of $n^{1/6}$. Zhang (2006) modifies it to provide the Multi-Scale Realized Volatility (MSRV) which features the optimal rate of convergence $n^{1/4}$ as documented in Gloter and Jocod (2001). Other approaches consist in and are not limited to: pre-averaging (PAE) the observations (Jacod et al. (2009)), [1] advocates for the realized kernels (RK) and [2] studies the quasi-maximum likelihood estimator (QMLE). Those three approaches share the optimal rate property and only differ through edge-effects which impact their respective AVAR.

The nonparametric AVAR bound of efficiency is equal to $8\omega T^{\frac{1}{2}} \int_0^T \sigma_u^3 du$, where T stands for the horizon time and ω^2 corresponds to the noise variance. This was shown in [3] under the deterministic volatility and Gaussian noise setting, but it is commonly assumed that it stays true under stochastic volatility. Subsequently, in a recent breakthrough paper, Altmeyer and Bibinger (2015) found an estimator based on the spectral approach introduced in [3] which reaches the bound in a very general situation. More recently, [4] proposed an adapted version of the pre-averaging estimator using local estimates as in [3] which gave rise to estimators that are within 7% of the bound.

When comparing fairly how perform several estimators, we need the candidates to be equipped with the same technology. Following closely the local technique

used in [3] and [4], we aim to adapt accordingly the RK and the QMLE. Indeed, although both estimators behave remarkably well when volatility is constant, i.e. in the parametric case the ratio of AVAR over the bound of asymptotic efficiency is 1.0025 when considering the most efficient Tukey-Hanning 16 RK and 1 for the QMLE, they can actually be highly inefficient in the non parametric setting. Under time-varying volatility, we aim to reduce significantly their AVAR and rend them as efficient as in the parametric problem. Although it would reduce the AVAR the same way, we didn't implement the local version of the MSRV. One reason is that the MSRV doesn't perform as well with a parametric ratio around 1.15.

To reduce the variance, we divide the interval $\big[0, T\big]$ into B non-overlapping regular blocks $\big[0, T/B\big]$, $\big[T/B, 2T/B\big]$, $\cdots$, $\big[(B-1)T/B, T\big]$. We then compute the RK (QMLE) on each block, and take the sum of the B estimates. We show that the estimator ratio converges to the parametric ratio as B increases. More importantly for practical applications, the convergence is very fast, and the gain is already important in the case $B = 2$ blocks. There are two essential reasons from a practical perspective why in our setting the number of blocks B is fixed whereas Altmeyer et al. (2015) and [4] let $B \to \infty$. The first reason is that we find that the AVAR is in most cases already within 4% of the bound when estimating daily volatility with $B = 8$ blocks, which is reasonably close enough. The second reason is that we want to quantify the expected gain when choosing a fixed number of blocks. which might not be optimal, we can fix B with the PAE and compute the associated AVAR.

References

[1] Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76(6), 1481-1536.

[2] Xiu, D. (2010). Quasi-maximum likelihood estimation of volatility with high frequency data. *Journal of Econometrics*, 159(1), 235-250.

[3] Reiß, M. (2011). Asymptotic equivalence for inference on the volatility from noisy observations. *The Annals of Statistics*, 39(2), 772-802.

[4] Jacod, J. and Mykland, P. A. (2015). Microstructure noise in the continuous case: Approximate efficiency of the adaptive pre-averaging method. *Stochastic Processes and their Applications*, 125(8), 2910-2936.

Nonparametric gaussian inference for stable processes

Session: **Statistics of Stochastic Processes 4**

Fabian Mies *Institute of Statistics, RWTH Aachen, Germany* †
Ansgar Steland *Institute of Statistics, RWTH Aachen, Germany*

Abstract: Jump processes driven by α-stable Levy processes impose inferential difficulties as their increments are heavy-tailed and the intensity of jumps is infinite. This paper considers the estimation of the functional drift and diffusion coefficients from high-frequency observations of a stochastic differential equation. By transforming the increments suitably prior to a regression, the variance of the emerging quantities may be bounded while allowing for identification of drift and diffusion in the limit. The findings are applied to obtain asymptotic normality of a nonparametric kernel estimator and of a parametric volatility estimator for the Ornstein-Uhlenbeck process. The limit theory suggests a semiparametric estimator for the index of stability α.

† Corresponding author. Mail: mies@stochastik.rwth-aachen.de

1 SDE Model

We consider the scalar stochastic differential equation (SDE)

$$dX_t = \mu(X_{t-})dt + \sigma(X_{t-})dZ_t \tag{1}$$

where Z_t is a symmetric α-stable Levy motion, i.e. $Ee^{i\lambda Z_t} = e^{-t|\lambda|^\alpha}$. Solutions to (1) exist for example if μ and σ are globally Lipschitz. Inference for the functional parameters μ, σ may be motivated by a local Euler expansion

$$X_{t+h} \approx X_t + h\mu(X_t) + \sigma(X_t)(Z_{t+h} - Z_t), \tag{2}$$

for small $h > 0$. Expression (2) can be read as an auto-regressive formula in order to estimate the drift function μ by nonparametric smoothing techniques [1, 2]. Since α-stable random variables have infinite variance, the limit results for standard Kernel smoothers admit a slow rate and heavy tailed limit distributions. Similar results hold in a parametric setting [3].

2 Tempering the increments

We suggest to handle the heavy-tailed innovations by a suitable tempering scheme. For smooth bounded functions $g \in C^\infty$ such that $g^{(k)} \in L_\infty$ for all $k \in \mathbb{Z}_{\geq 0}$, we show that

$$\frac{1}{h}E(g(X_{t+h} - X_t)|X_t = x) = \mu(x)g'(0) + \sigma(x)^\alpha g^{(\alpha)}(0) + \mathcal{O}(h^{\frac{2}{\alpha}-1}) \tag{3}$$

where $g^{(\alpha)}$ denotes the fractional derivative. As relation (3) depends on g only via two scalars, it again motivates an auto-regressive scheme. For example, odd functions g identify the drift only, whereas even functions can be used to infer the volatility function σ. Furthermore, all powers of g satisfy the conditions of relation (3).

3 Functional estimation

Evoking a regression scheme $X_t \to g(X_{t+h} - X_t)$ for the tempered increments, the noise of estimation is roughly quantified by $g(X_{t+h} - X_t)^2$. Thus, applying (3) to the function g^2 recovers a setting of finite variances. In combination with the explicit approximation rate of (3), functional estimation of μ and σ is amenable to a wide range of parametric and nonparametric regression techniques. In particular, we show that the Nadaraya-Watson estimator admits a Gaussian limit under mixing and stationarity assumptions.
An interesting special case arises by choosing an even function g, i.e. $g'(0) = 0$. Then, for any $\lambda > 0$,

$$E(g(\lambda(X_{t+h} - X_t))|X_t) \approx \lambda^\alpha E(g(X_{t+h} - X_t)|X_t). \tag{4}$$

This relation can be used to infer the parameter α by performing inference based on (4) for various scaling factors λ. Since all estimators have finite variance, α may then be estimated by a least squares fit. This approach resembles the estimators of [4] and [5] for the jump activity index in a more general model, who consider the asymptotic regime $\lambda \to \infty$.

References

[1] Long, H. and Qian, L. (2013). Nadaraya-Watson estimator for stochastic processes driven by stable Lévy motions *Electronic Journal of Statistics*, 7, 1387–1418.

[2] Wang, Y. and Zhang, L. (2013). Local linear estimation for stochastic processes driven by α-stable Levy motion *Statistical Inference for Stochastic Processes*, 16:2, 161–171.

[3] Hu, Y. and Long, H. (2009). Least squares estimator for Ornstein–Uhlenbeck processes driven by α-stable motions *Stochastic Processes and their applications*, 119:8, 2465–2480.

[4] Aït-Sahalia, Y. and Jacod, J. (2009). Estimating the degree of activity of jumps in high frequency data, *The Annals of Statistics*, 37:5, 2202–2244.

[5] Todorov, V. (2015). Jump activity estimation for pure-jump semimartingales via self-normalized statistics, *The Annals of Statistics*, 43:4, 1831–1864.

An application of non-homogeneous Poisson process in presence of change-points

Session: **Statistics of Stochastic Processes 4**

Jorge Achcar *Department of Social Medicine, FMRP, University of São Paulo, Ribeirão Preto, SP, Brazil*

Emílio Coelho-Barros *Department of Mathematics, Technical Federal University of Paraná, Cornélio Procópio, PR, Brazil*

Roberto Souza *Department of Mathematics, Technical Federal University of Paraná, Cornélio Procópio, PR, Brazil*

Abstract: Rain precipitation in the last years has been very atypical in different regions of the world, possibly, due to climate changes. We analyze SPI (Standard Precipitation Index) measures (1, 3, 6 and 12 month timescales) for a large city in Brazil: Campinas located in the southeast region of Brazil, São Paulo State, ranging from January 01, 1947 to May 01, 2011. A Bayesian analysis of non-homogeneous Poisson processes in presence or not of change-points is developed using Markov Chain Monte Carlo methods in the data analysis. We consider a special class of models: the power law process. We also discuss some discrimination methods for the choice of the better model to be used for the rain precipitation data.

1 Introduction

A popular index introduced in the literature and used by many countries around the world, more simple, easy to calculate, statistically relevant and effective in analysing wet periods/cycles or dry periods/cycles is the Standardized Precipitation Index (SPI) introduced by [2]. A classification system introduced by [2] is used to define drought intensities obtained from the SPI. They also defined the criteria for a drought event for different timescales. A drought event occurs any time the SPI is continuously negative and reaches an intensity of -1.0 or less. The event ends when the SPI becomes positive. Each drought event, therefore, has a duration defined by its beginning and end, and an intensity for each month that the event continues. The positive sum of the SPI for all the months within a drought event can be termed the drought's "magnitude".

2 The method

For the modelling the number of violation occurrences of drought in Campinas city, we consider a point process to count violations. Let $N(t)$ be the cumulative number of violations that are observed during the interval $(0, T)$ and assume that $N(t)$ is modelled by a non-homogeneous Poisson process (NHPP). To have the intensity function $\lambda(t) = \frac{dm(t)}{dt} = \frac{dE[N(t)]}{dt}$ where $m(t)$ is the mean value function, a monotonic function of t, we could consider different parametrical forms introduced in the literature [1, 3, 4]. A Bayesian analysis of non-homogeneous Poisson processes in presence or not of change-points is developed using Markov Chain Monte Carlo methods in the data analysis. We consider a special class of models: the power law process (PLP). We also discuss some discrimination methods for the choice of the better model to be used for the rain precipitation data.

3 Example

We analyze SPI (standard precipitation Index) measures (1, 3, 6 and 12 - month timescales) for a large city in Brazil: Campinas located in the in the southeast region of Brazil, São Paulo State, ranging from January 01, 1947 to May 01, 2011. In place to model the SPI time series, since a drought event occurs any time the SPI is continuously negative and reaches an intensity of -1.0 or less, we consider the calendar time or epochs of droughts less or equal to the treshold $L = -1.0$ for Campinas, during the period ranging from January 01, 1947 to May 01, 2011, which corresponds to a period of $T = 770$ months.
The use of non-homogeneous Poisson processes assuming a specified parametrical form for the intensity function could be very useful to analyze precipitation data of cities throughout the world and explain some atypical drought event periods. This data would provide information about the epochs of occurrence of drought event violations of environmental standards during a specific period of time, as is the case of the SPI measures (1, 3, 6 and 12-month timescales) for the Campinas city, Brazil. A common problem with precipitation or other environmental data is the presence of one or more change-points, possible due to climate changes in the last years. In this case, we also observe that better NHPP models could be of great use.

References

[1] Cox, D.R; Lewis, P.A. (1966). *Statistical Analysis of Series of Events*, Methuem: UK.

[2] Mckee, T. B.; Doesken, N. J.; Kleist, J. (1993). The relationship of drought frequency and duration to times scale, *Conference on Applied Climatology*, Boston: American Meteorological Society, pp. 179-184.

[3] Musa, J. D.; Okumoto, K (1984). A logarithmic Poisson execution time model for software reliability measurement, *Proceedings of 7th International Conference on Software Engineering*, Orlando, pp. 230-238.

[4] Musa, J. D.; Iannino, A.; Okumoto, K. (1987). *Software reliability: measurement, prediction, application*, McGraw Hill: USA.

Autonomous watermark identification based on machine learning techniques

Session: **Machine Learning 1**

Dana Simian *Faculty of Science, Department of Mathematics and Informatics, "Lucian Blaga" University of Sibiu, Romania*

Ralf Fabian *Faculty of Science, Department of Mathematics and Informatics, "Lucian Blaga" University of Sibiu, Romania*

Abstract: The fast growth of multimedia content generation and the easy access to it, led to digital embedding techniques that add supplementary information for later access and with multiple purposes, e.g., author identification, content classification and profile identification. Intended or unintended alterations of digital documents may occur in handling processes or in transmission, obscuring any later information extraction. The paper proposes a model of a system that allows autonomous identification of embedded information in media documents, even if they have undergone subsequent transformations that changed their initial state. The system is built on watermarking and machine learning techniques. A study conducted upon the influence of various parameters on the system's performance is presented in detail.

1 Introduction

Nowadays including additional information in multimedia content has become increasingly used, allowing ownership identification and multimedia documents' classification. Our goal is to provide an autonomous way to optimize the recognition and interpretation of the results when host multimedia data suffers common alterations that destructively affect embedded information.

2 Proposed model

We aim to model a system working on two directions: a) embedding additional information in digital content for later access; b) extracting the additional embedded information and use it for multiple purposes: author identification, profile identification, automate content classification, building a history of all transformations applied to the original content before and/or within transmission of the multimedia content, detecting any kind of operation meant to alter or eliminate the additional information. We choose watermarking techniques for embedding additional information; they are suitable for any multimedia content. Watermark recovering and identification is highly dependent on the content requiring an adaptive recovering method. Our proposed model includes subsystems for: pre-processing; watermark embedding; watermark extracting; classification; building the history of the host content transformations. The classification subsystem supposes three processes: training/testing set construction, training process and performance evaluation. These processes are discussed in detail.

3 Theoretical results

Several watermarking techniques [1] have been analyzed and compared, taking into account the quantization step size, redundancy level, computational complexity, watermark size and field of application. For the recognition process, the focus was on machine learning techniques suitable for the design of an adaptive subsystem: neural networks, support vector machines (SVM) and hybrid techniques [2].

4 Experimental results

We implemented the proposed model for digital images. We extended our system presented in [3] and conducted experiments in two directions: a) determining a suitable machine learning technique for the pattern recognition process and studying the influence of different parameters on the system's performances; b) tuning over benchmarking with several test images and patterns. For the considered test scenario we used watermarking techniques working in frequency domain and neural networks based classifier.

Acknowledgement: the first author, Dana Simian, was supported by the research grant LBUS-IRG-2015-01, project financed by the "Lucian Blaga" University of Sibiu.

References

[1] Khan A., Siddiqa A., Munib S., and Malik S. A. (2014): A recent survey of reversible watermarking techniques, *Information Sciences*, Volume 279, pp. 251-272.

[2] Kuncheva L.I.(2014): *Combining pattern classifiers: methods and algorithms*. Second edition, John Wiley & Sons, Inc., Hoboken, New Jersey.

[3] Simian, D. and Fabian, R. (2016): Ownership tracking with dynamic identification of watermark patterns, *Proceedings of the MDIS 2015*, Sibiu, Romania, pp. 113-123, "Lucian Blaga" Univ. Press.

Support Vector Machine Optimized by Fireworks Algorithm for Handwritten Digits Recognition

Session: **Machine Learning 1**

Milan Tuba *Faculty of Mathematics, University of Belgrade, Serbia*
Eva Tuba *Graduate School of Computer Science, John Naisbitt University, Serbia*
Raka Jovanovic *Qatar Envir. and Energy Research Institute, Hamad Bin Khalifa University, Qatar*

Abstract: Handwritten digits recognition is an important subarea in the object recognition research area. Support vector machines represent a very successful recent binary classifier. Basic support vector machines have to be improved in order to deal with real world problems. Introduction of soft margin for outliers and misclassified samples as well as kernel function for non linearly separably data leads to the hard optimization problem of selecting parameters for these two modifications. Grid search which is often used is rather inefficient. In this paper we propose the use of one of the latest swarm intelligence algorithms, the fireworks algorithm, for the support vector machine parameters tuning. With selected set of simple features we obtained better results compared to other approaches from literature.

1 Introduction

In computer vision object recognition represents a significant research area which deals with recognition of specific objects in digital images. Digit recognition is a subarea which can be divided into printed digit recognition (for license plate recognition, street numbers recognition, etc.) and handwritten digit recognition (for reading bank checks, in post offices for mail sorting, etc.). Recognition of handwritten digits is much harder since numerous handwriting styles exist.
Template matching is an old and simple technique for object recognition but it is not suitable for handwritten digit recognition. Techniques based on more advanced features like projections histograms, invariant moments and DCT coefficient are more appropriate.

2 The proposed method

Support vector machine (SVM) represents one of the latest supervised learning classifiers. SVM determines a hyperplane that separates data from different classes. Basic support vector machines have to be improved in order to deal with real world problems. However, the introduction of soft-margin for outliers and misclassified samples as well as kernel function for nonlinearly separably data leads to the hard optimization problem of selecting parameters for these two modifications. Grid search on the log-scale of the parameters, combined with cross validation procedure may result in huge computational time and far from optimal selection of parameters. Selecting good values for parameters is a hard optimization problem and for such

[1]This research was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, Grant No. III-44006

problems, stochastic population search algorithms, particularly swarm intelligence, were studied and successfully used.
One of the latest swarm intelligence algorithms is the fireworks algorithm (FWA). Authors of the original fireworks algorithm continued to work on its improvement. The results are four upgraded versions of the FWA, enhanced FWA, corporative framework for FWA, FWA with enhanced interaction and the latest update is guided FWA (GFWA).
In this paper we propose using SVM optimized by the latest GFWA algorithm for handwritten digit recognition using intentionally weak features with which other approaches would not give good results. Our proposed algorithm was tested on standard MNIST dataset for handwritten digit recognition and performance was better than other approaches from literature [1], [2], [3].

3 Experimental results

Digit	**MLNN**	**SVM-BAT**	**SVM-FWA**
0	86.45	99.00	99.00
1	94.39	99.00	99.00
2	88.73	97.00	98.00
3	77.02	89.00	92.00
4	76.12	98.00	99.00
5	84.10	91.00	93.00
6	78.81	100.00	100.00
7	77.12	93.00	95.00
8	79.03	95.00	95.00
9	49.64	95.00	96.00
Global	79.14	95.60	**96.60**

Table 3: Accuracy of classification reported in [3], [1] and our proposed method

References

[1] Tuba, E., Tuba, M., Simian, D. Handwritten Digit Recognition by Support Vector Machine Optimized by Bat Algorithm, *Computer Science Research Notes CSRN 2602 (Vaclav Scala, Editor): Papers from the 24th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, Plzen, Czech Republic, 2016, pp. 369-376

[2] Tuba, E., Bacanin N. An Algorithm for Handwritten Digit Recognition Using Projection Histograms and SVM Classifier, *23rd Telecommunications Forum TELFOR*, Belgrade, Serbia, 2015, pp. 464-467

[3] Kessab, B. E., Daoui, C., Bouikhalene, B. Fakir, M., Moro, K. Extraction method of handwritten digit recognition tested on the MNIST database. *International Journal of Advanced Science and Technology*, vol. 50, no. 6, 2013, pp. 99–110

Ontology-based Decisions in Software Engineering

Session: **Machine Learning 1**

Franz-Felix Füßl *Department of Computer Science and Automation, Technische Universität Ilmenau, Germanyy*
Detlef Streitferdt *Department of Computer Science and Automation, Technische Universität Ilmenau, Germany*
Elena Rozova *Department of Computer Science and Automation, Technische Universität Ilmenau, Germany*
Livia Sangeorzan *Department of Mathematics and Computer Science, Transilvania University of Brasov*
Nicoleta Enache-David *Department of Mathematics and Computer Science, Transilvania University of Brasov*

Submission for the 13^{th} Workshop on Stochastic Models, Statistics and Their Applications.
detlef.streitferdt@tu-ilmenau.de

Abstract: Current software development efforts are facing very short development cycles for complex systems with high demands on the quality of the resulting product. At the same time the environments for software developers are manifold and are getting more and more complex as well, what requires relevant additional efforts to setup and maintain such environments. Development methods and processes as well as the required, corresponding tools are part of the developers environments. A large number of decisions need to be taken to finally get to an optimized development process and tool set, the developers environment. Such decisions are influenced by technical and evenly important personal constraints. Thus, a knowledge model is needed to capture the knowledge and enable an automated deduction of an optimized developers environment.

1 New Knowledge Model

In this contribution a knowledge model [1] is presented, see Fig. 9, based on the PhD work of Franz Füßl with five abstraction levels to capture and maintain constraints, interconnect them and use the model for automated decisions using deduction and ontology learning [3]. This models extends [2] by the inclusion of arbitrary metrics (e.g., based on measurements) and social factors.

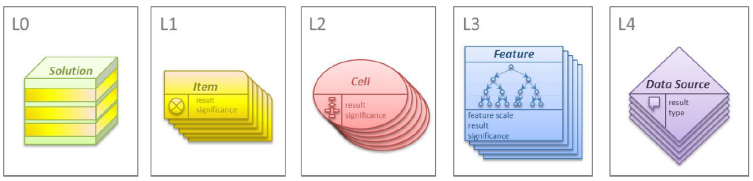


Figure 9: The New Five Level Knowledge Model

The 4^{th} level in Fig. 9 hosts *data sources* representing very simple issues which are captured with corresponding multiple-choice or single-choice questions. Data sources may also use sensor values or measurements. The 3^{rd}level includes the *features* of a project (e. g. budget, operating system or personal motivation). The features are connected to at least one data source element. Each feature is measured on a nominal, ordinal or metric scale which also corresponds to the connected data source element type. Features are connected to the *cells* on the 2^{nd} level. Cells generate knowledge based on the connected features. The 1^{st}level hosts *items* (e. g., requirements engineering, software architecture pattern) which use the information stored in the cells to model abstract components for the solution. Finally, *solutions* are at level zero

and represent developer packets to be used in a given development effort. The selection of the solution packets is based on their feasibility to fulfill the items of the first level. The knowledge model is a directed graph. Arbitrary associations can be realized in this graph. Currently five associations have been defined (is path, has path, can path, part-of path, and used-for path) and fully realized in a software tool.

2 Example

Five algorithms for the deduction process have been developed so far. The "isn't-it"-algorithm to ask whether an element is of a specified type, the "kind-of"-algorithm to ask an element for it's type, the "parts"-algorithm to ask and return the sub-elements of a given element, the "characteristics"-algorithm to identify the features of an element, and the "find"-algorithm to find elements of a given class fulfilling a given purpose. With these algorithms we can *ask* the model e. g., "Find a software to describe requirements". The capabilities of such a model are shown in the requirements example in Fig. 10.

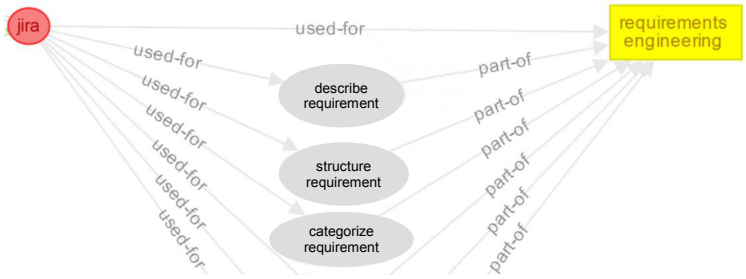


Figure 10: Requirements Engineering Example

Finally, the learning capabilities of the model are based on weighted edges. In the above example the top most *used-for*-edge between "jira-node" and "requirements engineering-node" is the result of a repeated process of selecting the tool Jira for describing, structuring and categorizing requirements. With a given threshold for the weighted *used-for*-edges between the "jira-node" and the describe-, structure-, and categorize-nodes the system learns by inductive reasoning. This results in the additional, new *used-for*-edge between the "jira-node" and "requirements engineering-node".
In this contribution we present the usage and benefit (higher acceptance of the individually "selected" developers environment) of the model. While considering all the constraints stored in the knowledge model, the future goal is the fully automated generation of a complete developers environment for software development projects.

References

[1] CW Chan, Y Peng, and LL Chen. Knowledge acquisition and ontology modelling for construction of a control and monitoring expert system. *International Journal of Systems Science*, 33(6):485–503, May 2002.

[2] Giancarlo Guizzardi and Veruska Zamborlini. Using a trope-based foundational ontology for bridging different areas of concern in ontology-driven conceptual modeling. *Science of Computer Programming*, 96(4):417–443, Dec 15 2014.

[3] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. *Ontology learning from text: methods, evaluation and applications*, volume 123. IOS press, 2005.

On the relationship between the statistical properties of the time series and goodness of fit of GRNN models, with application to financial data

Session: **Machine Learning 1**

Alina Barbulescu *ETS (Mathematics, Physics and Natural Sciences), Higher Colleges of Technology, Sharjah, U.A.E.*

Abstract: In this paper we address the performances of the Generalized Regression Neural Networks (GRNN) on time series modeling, function of their statistical properties, with application to financial time series.

1 Introduction

Last years, researchers have exposed an increasing concern in using artificial intelligence methodology for financial markets analysis. The most difficult problem of economy and finance is the predictability of future events. High non-linearity and volatility of financial instruments make the forecasting a challenging process. The basic idea of forecasting is to identify a similarity of mapping between the input and output data in order to extract the implicit rules governing the detected evolutions [4].
New studies on forecasting with Generalized Regression Neural Network (GRNN) show that this technique can be a promising option to nonlinear time series modeling, in general, and for financial series, in particular [3].
Since systematic studies concerning the GRNN's performances function of the statistical properties of the series have not been performed, we aim to provide such an analysis. The study is done on the BET series registered in the period October 2000 - September 2014.

2 The method

Firstly, we test the series randomness, the existence of an increasing or nonlinear trend, its stationarity around a deterministic trend and the breakpoints existence [1] [2] .
Secondly, we built GRNN models for the entire series, the subseries detected after the segmentation and the deseasonalized ones.
Finally we compare the modeling results for the brute and deseazonalized series, to see the influence of the seasonality removal on the models' quality.
We conclude that some regularity properties (as normality and homoskedasticity) of series have no influence on the models quality, but the seasonality influences it.

References

[1] Gilbert R (1987) *Statistical Methods for Environmental Pollution Monitoring*, Wiley, New York.

[2] Killick R, Fearnheadt P, Eckley IA(2012) Optimal detection of changepoints with a linear computational cost. J Am Stat Assoc 107(500): 1590-1598.

[3] Li W, Liu J, Le J (2005) Using GARCH-GRNN model to forecast financial time series, Computer and Information Sciences - ISCIS 2005. 20th International Symposium, Istanbul, Turkey, October 26-28, 2005. Proceedings: 565-574.

[4] Marzi H, Turnbull M, Marzi E (2008) Use of neural networks in forecasting financial market, Soft Computing in Industrial Applications, 2008. SMCia '08. IEEE Conference on: 240-245.

Validation of positive expectation dependence

Session: **Nonparametric and Semiparametric Testing**

Bogdan Ćmiel *Faculty of Applied Mathematics, AGH University of Science and Technology, Al. Mickiewicza 30, 30-059 Cracow, Poland.*

Teresa Ledwina *Institute of Mathematics, Polish Academy of Sciences, ul. Kopernika 18, 51-617 Wrocław, Poland.*

Abstract: We shall present new tests for positive expectation dependence. The solutions are weighted Kolmogorov-Smirnov type statistics. They originate from the function valued monotonic dependence function, describing local changes of the strength of the dependence. Therefore, the inference can be supported by a simple and insightful graphical device.

We shall show that an approach relying on multiplier central limit theorem and p-values allows to overcome inherent difficulties of this testing problem. Monte Carlo methods are used to assess the sizes and powers, and to compare new solutions to existing one. The simulations show that the new tests perform well in finite samples. Danish fire insurance data set shall be briefly examined to demonstrate the practical application of the proposed inference methods.

References

[1] Davydov, Yu., Egorov, V. (2000). Functional limit theorems for induced order statistics, *Mathematical Methods of Statistics* 9, 297-313.

[2] Kowalczyk, T. (1977). General definition and sample counterparts of monotonic dependence functions of bivariate distributions. *Mathematische Operationsforschung und Statististik, Ser. Statistics* 8, 351-365.

[3] Kowalczyk, T., Pleszczyńska, E. (1977). Monotonic dependence functions of bivariate distributions. *Annals of Statistics* 5, 1221-1227.

[4] Muliere, P., Petrone, S. (1992). Generalized Lorenz curve and monotone dependence orderings, *Metron* L, 19-38.

[5] Wright, R. (1987). Expectation dependence of random variables, with an application in portfolio theory. *Theory and Decision* 22, 111-124.

[6] Zhu, X., Guo, X., Lin, L., Zhu, L. (2016). Testing for positive expectation dependence. *Annals of the Institute of Statistical Mathematics* 68, 135-153.

Adaptive estimation in sup-norm for semiparametric conditional location-scale mixtures

Session: **Nonparametric and Semiparametric Testing**

Hajo Holzmann *Fachbereich Mathematik und Informatik, Philipps-Universität Marburg, Germany*
Heiko Werner *Fachbereich Mathematik und Informatik, Philipps-Universität Marburg, Germany*
Pierre Vandekerkhove *University Paris Val de Marne, Paris, France*

Abstract: Suppose that the conditional density of Y given X is a semiparametric location-scale mixture of two symmetric densities, one of which is known up to unknown scale, the other being unkown with additional unkown location. We discuss identification in such a model, and provide an estimator for the location, scale and mixing proportion functions. We analyze their rate of convergence in sup-norm, and propose an adaptive version of the estimator.

Semiparametric mixtures have recently been studied in various papers, see e.g. [2, 3]. [2] studied a conditional two-component location mixture model in a symmetric density. They proposed an estimator and derived its pointwise, non-adaptive rate of convergence. In this paper we consider a related problem. Suppose that the conditional density of Y given $X = x$ is given by

$$g(y|x) = p(x)\, f_x\big(y - \mu(x)\big) + \frac{1 - p(x)}{\sigma(x)}\, f_0\big(y/\sigma(x)\big),$$

where $p(x) \in (0, 1)$ is a smooth weight function, $\mu(x)$ a smooth location and $\sigma(x)$ a smooth scale function, f_0 a fixed, known symmetric density and f_x an unkown symmetric density depending smoothly on x.
We discuss identification of $\theta(x) = \big(p(x), \mu(x), \sigma(x), f_x\big)$, and propose a smoothed minimum-contrast estimators for these functions. The rate of convergence of the resulting estimators is analyzed in the sup-norm, it turns out to be the usual nonparametric rate in sup-norm for $p(x), \mu(x)$ and $\sigma(x)$. We also propose an adaptive version of the estimator based on the Lepski-scheme. The main technical tool is a version of the Bernstein-inequality for U-statistics from [1].

References

[1] Giné, E., Latala, R., Zinn, J. (2000). Exponential and moment inequalities for U-statistics. arXiv:math/0003228

[2] Butucea, C., Nguyepe Zumpe, R. , Vandekerkhove, P. (2015). Semiparametric topographical mixture models with symmetric errors. *Bernoulli*, to appear

[3] Hohmann, D., Holzmann, H. (2013). Semiparametric location mixtures with distinct components. *Statistics* 47 348-362

Statistical tests for signal models

Session: **Nonparametric and Semiparametric Testing**

Ulrich Stadtmüller *Department of Number & Probability Theory, Ulm University, Germany*
Mirek Pawlak *Department of Electical Engineering, University of Manitob, Winnipeg, Canada*

Abstract: Given noisy samples of a signal, the problem of testing whether the signal belongs to a restricted class of parametrically specified class of signals is considered. We examine the parametric situation for a well-defined null hypothesis signal model and compare it with broad alternative signal classes that cannot be parametrized. For such a setup, we introduce testing procedures relying on nonparametric kernel-type sampling reconstruction algorithms properly adjusted for noisy data. The proposed testing procedures utilizes the L_2 - distance between the kernel estimate and signals from the target class. The central limit theorems of the test statistics are derived yielding consistent testing methods. Hence, we obtain the testing algorithms with the desirable level of the probability of false alarm and the power tending to one. The asymptotic as well as the finite sample size comparisons of the introduced tests are given.

Bootstrap method applied to estimators of tempered stable distribution parameters

Session: **Nonparametric and Semiparametric Testing**

Piotr Kruczek *Faculty of Pure and Applied Mathematics, Hugo Steinhaus Center, Wroclaw University of Science and Technology Wybrzeze Wyspianskiego, 27, 50-370 Wroclaw, Poland, piotr.kruczek@pwr.edu.pl*

Agnieszka Wyłomańska *Faculty of Pure and Applied Mathematics, Hugo Steinhaus Center, Wroclaw University of Science and Technology Wybrzeze Wyspianskiego, 27, 50-370 Wroclaw, Poland*

Jacek Leśkow *Cracow University of Technology, Poland*

Radosław Zimroz *Diagnostics and Vibro-Acoustic Science Laboratory, Wroclaw University of Science and Technology, Na Grobli 15, 50-421 Wroclaw, Poland*

Abstract: Tempered stable distribution is an extension of well-known stable distribution. In our paper case with exponential tempering function is considered. The basic properties are recalled and described. In literature different estimators of tempered stable distribution can be found. In our paper three estimators are recalled, namely method of moments (MM), maximum likelihood method (MLE) and based on the characteristic function (CF). In order to verify estimators performance the bootstrap methods can be applied. In particular, the parametric and non-parametric bootstrap methods are used. They will provide the information, which of the estimators is the most effective. Furthermore, the tempered stable distribution is applied to model the vibration data acquired on the machine located in the underground mine. It is shown that such distribution can be useful in case of the condition monitoring. Applying the tempered stable distribution we are able to diagnose the local damage in the machine drive unit.

GARCH processes and the phenomenon of misleading and unambiguous signals

Session: **Stochastic Models in Technology, Reliability, and Quality 2**

Manuel Cabral Morais *Dept. Mathematics, Instituto Superior Técnico, ULisboa, Portugal*
Beatriz Sousa *Instituto Superior Técnico, ULisboa, Portugal*
Yarema Okhrin *Faculty of Business and Economics, University of Augsburg, Germany*
Wolfgang Schmid *Department of Statistics, European University Viadrina, Germany*

Abstract: In Finance it is quite usual to assume that a process behaves according to a previously specified target GARCH process. The impact of rumours or other events on this process can be frequently described by a change in the variance or an outlier responsible for a shift in the process mean, thus calling for the use of joint schemes for the process mean and variance, such as the ones proposed by [1] and [2]. Since changes in the mean and in the variance require different actions from the traders/brokers, this talk provides an account on the probabilities of misleading and unambiguous signals (PMS and PUNS) of those joint schemes, thus adding insights on their out-of-control performance.

References

[1] Schipper, S. (2001). *Sequential Methods for Detecting Changes in the Volatility of Economic Time Series*. Ph.D. thesis, European University, Department of Statistics, Frankfurt (Oder), Germany.

[2] Schipper, S. and Schmid, W. (2001). Control charts for GARCH processes. *Nonlinear Analysis: Theory, Methods & Applications* **47**, 2049–2060.

CUSUM-Shewhart charts for monitoring normal variance

Session: **Stochastic Models in Technology, Reliability, and Quality 2**

Sven Knoth *Institute of Mathematics and Statistics, Department of Economics and Social Sciences, Helmut Schmidt University Hamburg, Germany*

Abstract: Monitoring the normal variance experienced lows and highs in the SPC literature. Besides very common vehicles such as the R, S or S^2 Shewhart control charts, some more sophisticated tools such as EWMA (exponentially weighted moving average) and CUSUM (cumulative sum) charts derived from the mentioned statistics were introduced already decades ago. It is quite surprising that the analysis of combining the simple Shewhart with one of the more advanced charts gained not much interest in case of monitoring (normal) variance. Except for the classic [2], no further studies of the subject seem to be available. One potential reason is that despite the simple operation of the combo, the numerical ARL (average run length) analysis is a demanding task. Here, we want to provide some new insights following the more recent [1]. It is demonstrated that the CUSUM-Shewhart combo deploying the running sample variance S_i^2 provides a simple and powerful procedure to detect a wide range of potential changes in the variance level.

References

[1] Knoth S. (2016). New results for two-sided CUSUM-Shewhart control charts. In *Proceedings of the XIIth International Workshop on Intelligent Statistical Quality Control*, 269–287.

[2] Yashchin E. (1985). On the analysis and design of CUSUM-Shewhart control schemes. *IBM Journal of Research and Development*, 29, 377–391.

Statistical Process Control Chart by Copula Condition Distributions

Session: **Stochastic Models in Technology, Reliability, and Quality 2**

Jong-Min Kim *Statistics Discipline, Division of Science and Mathematics, University of Minnesota at Morris, U.S.A.*

Jaewook Baik *Department of Information Statistics, Korea National Open University, Seoul, Republic of Korea*

Mitch Reller *Statistics Discipline, Division of Science and Mathematics, University of Minnesota at Morris, U.S.A.*

Abstract: We propose new control charts by using conditional distribution with diverse copula functions to detect the change point of conditional variance. We generate the conditional asymmetric transformed data by using asymmetric functions including asymmetric copula functions and apply the conditional transformed data to the cumulative sum control (CUSUM) and exponentially weighted moving average (EWMA) statistics in order to detect the change point of conditional variance by using Average Run Length (ARL) of CUSUM control charts and EWMA control charts by using Monte Carlo Simulation Method. We show that the ARLs of change point of conditional variance by new conditional control charts are affected by the directional dependence by using the bivariate gaussian copula beta regression ([2]). In real example, we use our proposed quality control charts to identify outlying seasons in Major League Baseball(MLB). Seasons from 1998 to 2016 are analyzed to determine the presence of outlying seasons using Earned Run Average (ERA) and Batting Average. It can be seen that there is a large abnormal variation of MLB statistics across these years.

1 Introduction

Since datasets in the real situation usually do not follow the usual assumption of constant variance, CUSUM and $\ln(S^2)$-EWMA statistics can be used to detect these change points of variance. A change in the mean affects only the mean chart but a change in the variance affects both the mean and variance charts. Using an asymmetric copula function and applying the conditional transformed data to the CUSUM and $\ln(S^2)$-EWMA statistics, we detect the change point of conditional variance by using the Average Run Length (ARL) of CUSUM control charts and $\ln(S^2)$-EWMA control charts by using Monte Carlo Simulation Method.

2 The method

In this research, we use diverse copula functions to generate the conditional transformed data by using condition distribution by Archimedean copula functions. One of examples is
Corollary 1. Using one of Archimedean copulae, Clayton Copula, [1] is

$$C_\alpha(u, v, \theta_2) = (u^{\theta_2} + v^{\theta_2} - 1)^{-1/\theta_2},$$

for $\alpha > 0$, for two random variables X_1 and X_2, we can derive the conditional distributions $F_{1|2}(X_1|X_2)$ as follows:

$$F_{1|2}(X_1|X_2) = \frac{\partial C_\alpha(u, v, \theta_2)}{\partial u}.$$

3 Example

We use an asymmetric copula (Frank(5) $\times$ Gumbel(30)) simulated data with $\rho = 0.8$ (see [2]). Table 1 shows that the directional dependence of Y given on X is higher than the directional dependence of X given on Y.

Directional Dependence	Y → X	X → Y
$Var(E(X\|Y))$	0.0495	0.0535
$Var(X)$	0.0833	0.0833
$\rho^2_{Y\to X}$	0.5937	**0.6421**
Thousand Replications with Sample Size 1000		
$\mathrm{E}(\hat{\rho}^2_{Y\to X})$	0.5935	**0.6407**
Std. Error$(\hat{\rho}^2_{Y\to X})$	0.0003	0.0002
Median$(\hat{\rho}^2_{Y\to X})$	0.5937	**0.6410**
Bias$(\hat{\rho}^2_{Y\to X})$	**-0.0002**	**-0.0014**

Table 4: Directional dependence of gaussian copula beta regression model.

The following figure shows that the ARLs of copula conditional control charts (CUSUM and $\ln(S^2)$-EWMA (r=0.3)) in case of the Y given on X and $n = 1000$ is more efficient than the ARLs of unconditional control charts (CUSUM and $\ln(S^2)$-EWMA (r=0.3)) and the ARLs of copula conditional control charts (CUSUM and $\ln(S^2)$-EWMA (r=0.3)) in case of the X given on Y and $n = 1000$ is not efficient than the ARLs of unconditional control charts.

Y|X	S=1000			Estimate	Kendall Tau
	n=300	n=700	n=1000		
CUSUM	150.83	356.77	497.93		
Clayton-CUSUM	148.64	353.17	494.25	2.174	0.521
Frank-CUSUM	149.90	347.01	495.39	8.740	0.628
Gumbel-CUSUM	150.01	361.00	501.25	2.582	0.613
t-CUSUM	145.56	352.18	491.42	(0.831, 3.176)	0.625
EWMA (r=0.3)	175.54	404.01	591.73		
Clayton-EWMA (r=0.3)	170.31	407.62	569.73	2.174	0.521
Frank-EWMA (r=0.3)	176.35	402.80	569.90	8.740	0.628
Gumbel-EWMA (r=0.3)	172.83	404.23	574.24	2.582	0.613
t-EWMA (r=0.3)	172.78	412.87	571.78	(0.831, 3.176)	0.625
EWMA (r=0.7)	170.06	404.72	576.10		
Clayton-EWMA (r=0.7)	174.82	399.96	580.64	2.174	0.521
Frank-EWMA (r=0.7)	170.34	414.27	582.16	8.740	0.628
Gumbel-EWMA (r=0.7)	177.76	400.73	582.70	2.582	0.613
t-EWMA (r=0.7)	177.90	398.38	577.22	(0.831, 3.176)	0.625

(a) Y given on X

X|Y	S=1000			Estimate	Kendall Tau
	n=300	n=700	n=1000		
CUSUM	151.87	351.16	491.96		
Clayton-CUSUM	149.41	349.02	501.65	2.174	0.521
Frank-CUSUM	149.90	347.01	495.39	8.740	0.628
Gumbel-CUSUM	146.15	344.35	502.23	2.582	0.613
t-CUSUM	151.60	348.61	499.55	(0.831, 3.176)	0.625
EWMA (r=0.3)	180.27	415.31	577.89		
Clayton-EWMA (r=0.3)	173.56	403.14	584.48	2.174	0.521
Frank-EWMA (r=0.3)	177.60	399.82	579.10	8.740	0.628
Gumbel-EWMA (r=0.3)	172.73	407.11	574.71	2.582	0.613
t-EWMA (r=0.3)	177.03	403.23	588.98	(0.831, 3.176)	0.625
EWMA (r=0.7)	172.42	407.45	590.46		
Clayton-EWMA (r=0.7)	175.37	414.89	590.19	2.174	0.521
Frank-EWMA (r=0.7)	174.50	402.58	583.63	8.740	0.628
Gumbel-EWMA (r=0.7)	173.92	410.72	576.08	2.582	0.613
t-EWMA (r=0.7)	175.05	411.61	578.69	(0.831, 3.176)	0.625

(b) X given on Y

Figure 11: ARLs with asymmetric copula simulated data with $\rho = 0.8$ where S is the number of simulations and n is the size of data out of 8,888 simulated data.

References

[1] Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika*, **65(1)** 141-151.

[2] Kim, J.-M. and Hwang, S. (2016). Directional Dependence via Gaussian Copula Beta Regression Model with Asymmetric GARCH Marginals, *Communications in Statistics: Simulation and Computation*; **In Press**.

Text Classification Systems: JavaScript versus WEKA implementation approach

Session: **Machine Learning 2**

Livia Sangeorzan *Department of Mathematics and Informatics, Transilvania University of Brasov, Romania*

Nicoleta Enache-David *Department of Mathematics and Informatics, Transilvania University of Brasov, Romania*

George-Alex Stelea *Departament of Electronics, Telecommunications and Information Technology, Transilvania University of Brasov, Romania*

Detlef Streitferdt *Department of Computer Science and Automation, Technische Universität Ilmenau, Germany*

Elena Rozova *Department of Computer Science and Automation, Technische Universität Ilmenau, Germany*

Abstract: In this paper we present an overview of the text documents classification and an application for the classification of e-mail messages using the Naive Bayes algorithm. The application was implemented in two different development environments, JavaScript and WEKA. We also carry out a comparative study between the two approaches in terms of runtime and predictive probability.

1 Introduction

Nowadays, machine learning techniques have developed very quickly and they are being used to solve many real-life problems. Document classification is a very actual issue and is a continuous challenge; it is based on different techniques of machine learning including Bayesian classification [1], SVM classifiers (Support Vector Machine) [4], k-NN (k-Nearest-Neighbor) classifier [2], classification based on association rules [3], etc.

2 The method

Our application for the classification of e-mail messages was implemented in two different development environments: JavaScript and WEKA.
JavaScript is a programming language used in web application development. It is supported by most current browsers and the running environment is the browser itself [5].
The Naive Bayes classifier implemented in JavaScript is trained using three training datasets - SPAM, SPORT, SOCIAL MEDIA. The application returns the classification of the e-mail type and its probabilities. In order to calculate the runtime of the algorithm we used a counter that returns a ”timestamp” used to measure the time difference in milliseconds between two discrete points with an accuracy of 5 mS (microseconds), as long as the browser does not have hardware or software constraints and is able to offer such precision.
WEKA is a collection of machine learning algorithms used for data mining, which contains algorithms implemented in Java for data preprocessing, classification, regression, clustering and association rules [6]. For our application we used Naïve Bayes Multinomial Text and J48 classifiers in WEKA. Considering the three aforementioned training datasets, both algorithms had the same prediction rate of the correctly classified instances.

3 Conclusions

In terms of runtime, in both approaches we obtained almost identical results when the e-mail classification system with JavaScript runs on Google Chrome or Opera browsers. When the e-mail classification system with JavaScript runs on Mozilla Firefox, it is slower than the system modeled in WEKA.

Regarding the probability obtained for the test dataset classification, it was different in the two approaches and it depends heavily on the size of the training dataset.

References

[1] Mukherjee S., Sharma N. (2012). Intrusion Detection using Naive Bayes Classifier with Feature Reduction, *2nd International Conference on Computer, Communication, Control and Information Technology (C3IT-2012)*, Hooghly, West Bengal, India.

[2] Lu W., Shen Y, Chen S., Ooi B.C. (2012). Efficient Processing of K Nearest Neighbor Joins Using MapReduce, *Proceedings of the VLDB*, pp 1016-1027, Vol. 5 Issue 10.

[3] Nguyena L.T.T., Vob B., Hongc T.P., Thanhe H.C. (2012). Classification Based on Association Rules: A lattice-based approach, *Expert Systems with Applications*, Vol. 39 Issue 13.

[4] Qi Z., Tian Y., Shi Y. (2013). Robust Twin Support Vector Machine for Pattern Classification, *Pattern Recognition*, pp 305–316, Vol. 46 Issue 1.

[5] Sangeorzan L., Stelea G.A., Enache-David N. (2016). *Web Development Techniques For Applications and Websites*, Transilvania University of Brasov Press: Romania.

[6] Frank E., Hall M.A., Witten I.H. (2016). *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed., Morgan Kaufmann: Massachusetts.

Stochastic Dynamics in Ultimatum Game simulation

Session: **Machine Learning 2**

Florentin Bota *Centre for the Study of Complexity, Babes-Bolyai University, Cluj-Napoca, Romania*
Dana Simian *Department of Mathematics and CS, Lucian Blaga University of Sibiu, Romania*

Abstract: The purpose of this project is to design and develop an autonomous computational model that can be used to play the so-called "Ultimatum Game". This is an economic experiment and it represents the next step in our Unified Model research. We used a 3D approach to better illustrate the Emotional component in our model. In this paper we will present the state-of-the-art and the ultimatum game experiments followed by our proposed model and implementation. We used stochastic evolutionary game theory to observe the evolution of our agents in a complex dynamic environment. This project also intends to create a better simulation tool that will be used in further experiments and for educational purposes.

1 Introduction

The standard theory assumes that people behave like perfect rational self-interested agents and always make the best decisions in order to maximize their utility. In the real world however, the actors can and will make impulse, emotional or social driven decisions (see Gilovich, Griffin and Kahneman 2002; Goldstein 2009). This is where we try to implement our model, which is inspired from the human nature, and not from a singular utility-maximization function.

2 The method

We used the Ultimatum Game (UG) to demonstrate that people are also influenced by the payoffs of others [1, 2]. Two players have to divide a fixed sum of money: Player A, makes an offer and Player B can either accept the offer or refuse it. If the offer is accepted, the money is split as proposed by player A and if the offer is refused, neither player receives anything. In evolutionary game theory, the approach is usually deterministic, but in these models evolution favors self-interest [3]. Using stochastic evolutionary game theory [2], where agents make mistakes and have bounded rationality (Simon (1955), Arrow(1981), Samuelson(1993) and Sen(1995)) we also observed that natural selection favors fairness.
An agent's strategy can be easily specified with $\alpha, \beta \in [0, 1]$, where α is the amount offered as player A and β is the minimum amount demanded as player B. One of the tested selection dynamics is the pairwise comparison process, where an individual and a role model are sampled at random from the population and the focal individual accepts the strategy of the role model [4] with probability p, depending of their payoffs, as seen in 1.

$$p = \frac{1}{2} + w\frac{\pi_f - \pi_r}{\Delta\pi}. \tag{1}$$

Where $w(0 \leq w \leq 1)$ is the intensity of selection, π_f and π_r are the payoffs of the selected individuals and $\Delta\pi$ is the maximum payoff difference.

3 Example

We gathered 176 game results, enough to gain a basic idea of how our model compares to the real world data (Fig. 12). We can see that our model behaves in an almost human manner, better than the simple "always offer 1$" standard model. The acceptance and rejection of values are similar, except the distribution of low value offers.

4 Acknowledgements

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS - UEFISCDI, project number PN-II-RU-TE-2014-4-2560.

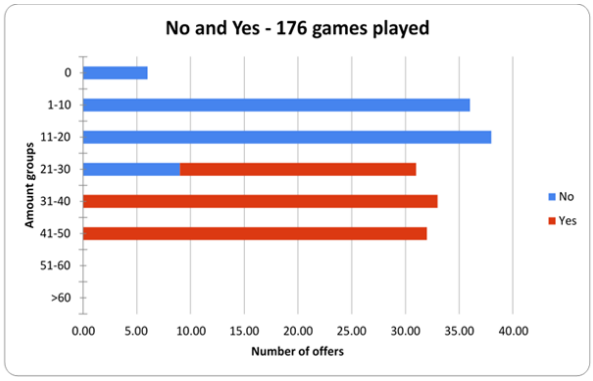


Figure 12: Intermediate statistical data from our simulations

In memory of prof. Dan Dumitrescu, who made this research possible.

References

[1] Camerer CF. (2003), *Behavioral Game Theory: Experiments in Strategic Interaction*, NJ: Princeton Univ Press, Princeton.

[2] Rand, D. G., Tarnita, C. E., Ohtsuki, H., Nowak, M. A. (2013). *Evolution of Fairness in the one-shot anonymous Ultimatum Game*, Proceedings of the National Academy of Sciences of the United States of America, 110(7), 2581–2586. http://doi.org/10.1073/pnas.1214167110.

[3] Nowak MA, Page KM, Sigmund K (2000), *Fairness versus reason in the ultimatum game.* Science. 2000 Sep 8; 289(5485):1773-5.

[4] Traulsen, A., Hauert, C. (2009). *Stochastic evolutionary game dynamics. Reviews of nonlinear dynamics and complexity*, 2, 25-61.

Gaussian Process Dynamic Mixture Models

Session: **Machine Learning 2**

Patrick Jähnichen *Machine Learning group, Humboldt-Universität zu Berlin, Germany*
Florian Wenzel *Machine Learning group, Humboldt-Universität zu Berlin, Germany*
Marius Kloft *Machine Learning group, Humboldt-Universität zu Berlin, Germany*

Abstract: We introduce a novel class of dynamic mixture models, one whose dynamics are driven by latent Gaussian processes. In doing so, we improve upon previous approaches that exclusively focus on the enclosed simpler case of Brownian motion dynamics of mixture components. We give an inference procedure for approximate posterior computation in the more general model class and proceed by also developing a scalable inference algorithm based on stochastic gradient descent.

1 Introduction

Despite their extraordinary capabilities to describe complex behavior in data, dynamic mixture models are not as heavily used as their static counterparts. Introducing dynamics to mixture models allows us to keep track of mixture components that are subject to a drift. Examples include the analysis of stock market data or time-stamped document collections and weather forecasting, among others. In our approach, the underlying dynamics are modelled via Gaussian processes (GPs), opening up for a wide range of dynamic priors in mixture models and models of mixed membership.

2 The model

We study a novel modelling class introducing new kinds of dynamic priors for mixture models on time series. The model under study is a mixture model of L D-dimensional Gaussian distributions whose time-dependent dynamics are governed by a GP as described by the following generative process:

1. for all $l = 1, \ldots, L$ draw $\beta_l \sim \mathcal{GP}(0, K)$
2. for all $t = 1, \ldots, T$ draw $\theta_t \sim \mathrm{Dir}_L(\alpha)$
3. for all $n = 1, \ldots, N$
   (a) draw a component: $z_n \sim \mathrm{Cat}(\theta_{t_n})$
   (b) draw data $x_n \sim \mathcal{N}(\beta_{z_n,t_n}, \sigma_X^2 \mathbf{I})$,

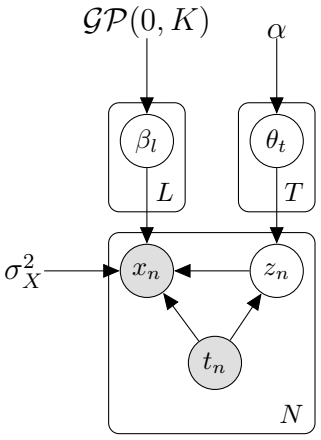


Figure 13: The GP dynamic mixture model.

where β_l are mixture components (each of which is a time-series over T steps), as given by a zero-mean GP prior with kernel function $k(\cdot, \cdot)$ and associated covariance matrix K. θ_t denotes the prior over mixing proportions for each data point at time t, σ_X^2 is a variance parameter and t_n is the observed time-stamp associated with observation x_n. A graphical representation of the model is given in Fig. 13.
We develop a variational inference algorithm for posterior approximation in this model, that, for a particular choice of covariance function[1], is equivalent to the variational Kalman filter (VKF) (cf. [5, 1]). Building on this, we introduce a scalable inference algorithm based on stochastic variational inference [3] and inducing point methods for lowering the complexity of posterior GP computation [4, 2].

[1] Wiener kernel function: $k(t_i, t_j) = \min(t_i, t_j)$

3 Example

We evaluate our approach on synthetic data generated according to our generative model for two settings ($T = 10, D = 5, L = 5$ and $T = 100, D = 50, L = 25$) and measure performance statistics for a growing number of observations N. We compare VKF with GP (batch) and GP (scalable), the two inference schemes described above. For these, we use the Wiener covariance function to be comparable to the VKF approach. As would be expected, the VKF and batch GP algorithm perform with similar performance, although the latter is clearly faster in terms of computation time. This can be explained by the need of numerical optimization in the VKF, while the batch GP uses a direct coordinate ascent update. For the simpler problem, our scalable GP algorithm performs slightly less accurate in terms of predictive quality. As it uses a lower rank approximation to the resulting covariance matrix of the full batch GP approach this is again expected behavior. With increasing model complexity, the batch GP approach is still much faster than the VKF, however, the stochastic variational inference approach benefits from utilizing a lower-rank approximation and its property to reach an optimum after having processed much less data points than needed by a batch algorithm.

References

[1] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML*, 2006.

[2] J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian Processes for Big Data. In *UAI*, 2013.

[3] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *JMLR*, 14(1):1303–1347, 2013.

[4] M. K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *AISTATS*, pages 567–574, 2009.

[5] C. Wang, D. M. Blei, and D. Heckerman. Continuous Time Dynamic Topic Models. In *UAI*, 2008.

Statistical Properties of Localized Support Vector Machines

Session: **Machine Learning 2**

Florian Dumpert *University of Bayreuth, Germany*

Abstract: Support Vector Machines (SVMs) are a nonparametric method of statistical machine learning with desirable properties like guaranteed existence, uniqueness, universal consistency and some kinds of robustness if only weak assumptions are satisfied. They are solutions of a regularized and kernel based risk minimization problem with respect to a loss function that can be chosen by the statistician. Nowadays, they therefore play an important role in statistics. Usually the SVM is learnt by a sample of observed data points which contain some information of their generating distribution. The SVM potentially uses all the information given by the data set. Two problems can arise. First: If there are too many data points and if it is not obvious which of them contain redundant information, learning the SVM may take too much time or may need too much storage. Second: If the generating distribution has different characteristics in different regions of the input space, learning only one "global" SVM usually leads to imprecise estimation. In this talk, we will show that some interesting statistical properties like consistency and some kinds of robustness are preserved in a class of local approaches which try to gather the different characteristics of the unknown generating distribution.

How to produce predictive models on an assembly line

Session: **ENBIS**

Andrea Ahlemeyer-Stubbe *Data Mining and More, Gengenbach*

Abstract: To detect fast changes in customer behavior or to react in an as focused manner as possible, predictive modeling must be done in good quality to get effective predictions of customer behavior and it has to be done fast to be relevant under business aspects. Modeling speed is of great importance in industry as time is a crucial factor. This necessity requires a different technical set up for model development to fulfill both needs: quality and development speed. Today most companies like to develop their models individually with the help of specialists. But for a lot of companies, this way takes too long. Even though the models are excellent, the time to develop them sometimes kills the advantages of a better prediction. This article describes the general structure and ideas how to implement industry-focused model production that will help to react quickly to changing behavior. We will discuss the key success factors and the pitfalls of this assembly line model product.

Monitoring a Wind Turbine by Combining Sensor Data

Session: **ENBIS**

Alessandro Di Bucchianico *Eindhoven University of Technology, Netherlands*
Stella Kapodistria *Eindhoven University of Technology, Netherlands*
Thomas Kenbeek *Eindhoven University of Technology, Netherlands*

Abstract: Undetected damage to parts of a wind turbine such as blade cracks due to lightning or broken gear wheels may have disastrous consequences possibly leading to loss of the entire wind turbine. It is therefore important to continuously monitor the condition of wind turbines, in particular when they are placed at remote locations (e.g., off-shore wind farms). Technological advances make it economically feasible to equip wind turbines with sensors for various physical variables (including vibration).
We describe our experiences when applying Statistical Process Control to monitor the condition of wind turbines in the Netherlands that are equipped with various sensors. Our approach is based on jointly monitoring variables using regression analysis to correct for external influences. This approach was an eye opener for the wind turbine engineers who used to think in threshold values for individual sensor variables. Analysis of historical data showed that malfunctioning of one the generators of a specific wind turbine could have been detected several months before the actual breakdown of the complete gearbox. Apart from describing our practical experiences, we also indicate the relatively unexplored methodological issues in applying SPC with regression models. This research was performed within the DAISY4OFFSHORE (Dynamic Asset Information System for Offshore Wind Farm Optimisation) project funded by the Dutch government through its "Wind at Sea" Top Consortium Knowledge and Innovation. The work of Kapodistria is also supported by the Dutch Science Foundation Gravitation Project "Networks" (`www.thenetworkcenter.nl`).

Data, methods and tools in support of smart and sustainable cities planning: an insight

Session: **ENBIS**

Alberto Pasanisi *Europäisches Institut für Energieforschung, Karlsruhe, Germany*
Andreas Koch *Europäisches Institut für Energieforschung, Karlsruhe, Germany*
Plessis Gilles *Europäisches Institut für Energieforschung, Karlsruhe, Germany*

Abstract: In a worldwide context of growing urbanization, there is an increasing need for a sustainable planning that optimises the capital expenditure and is compliant with environmental and climate change related issues. In addition, modern IT infrastructures allow to collect, exchange and analyse a huge quantity of data and open the way for new services and businesses. By means of feedback, coming from real use-cases or demonstrators, an insight is given of the main stakes to be tackled by planners today as well as how data, coming from different sources, coupled with specific domain-expertise (in particular in the field of energy), can provide efficient tools for diagnosis and forecasting at both the neighbourhood and the city scale.

Ranks and Pseudoranks
- Paradoxical Results of Rank Procedures in Case of Unequal Sample Sizes -

Session: **Nonparametric Methods 2**

Edgar Brunner *Institue of Medical Statistics, University of Göttingen, Germany*

Abstract: If rank methods are used for $d > 2$ samples $X_{ik} \sim F_i$, $i = 1, \ldots, d$; $k = 1, \ldots, n_i$, then paradoxical results may be obtained in case of unequal sample sizes since the quantities $p_i = \int H dF_i$, on which these procedures are based depend on sample sizes through the definition of the weighted mean distribution function $H = \frac{1}{N} \sum_{i=1}^{d} n_i F_i$ in the experiment. These undesirable property applies to the well-known Kruskal-Wallis (KW) test, the Hettmansperger-Norton (HN) test for trend, the Jonckheere-Terpstra (JT) test, as well as for the Akritas-Arnold-Brunner (AAB) procedures [1] in factorial designs. This is explained by analytical considerations as well as by some simulated data sets demonstrating that for the same set of alternatives $F_1, \ldots, F_d$ either a highly significant (p-value < 0.01) or a completely non-significant (p-value > 0.70) may be obtained just be altering the ratios $n_1 : n_2 : \cdots : n_d$ of the sample sizes while keeping fixed the total sample size N. For the JT-test and the HN-test one may even obtain a significantly increasing trend or a significantly decreasing trend for the same set of alternatives $F_1, \ldots, F_d$ just changing the ratios of the samples sizes. While in the one-way layout such paradoxical results can be obtained by a so-called non-transitive set of distributions (see, e.g, [5]) it can be demonstrated in a two-way layout that similar paradoxical results are also obtained by applying the AAB-procedure to shifted normal distributions. For example, in such designs interactions may appear or disappear just by changing the ratios of the samples sizes while keeping fixed the underlying normal distributions.
Such designs appear in the so-called sub-group analysis in clinical trials or in clinical epidemiology. Here the question is to examine whether an effect (or non-effect) detected in a large trial is also present (or non-present) in a small sub-group, such as diabetics, or people having some particular disease in a certain organ.
Another problem using the quantities $p_i = \int H dF_i$ as nonparametric effects is the fact that it seems to be not reasonable to formulate hypotheses on quantities depending on sample sizes - except in the case of equal sample sizes. Therefore, The AAB-procedures are based on hypotheses expressed in terms of the distribution functions $F_1, \ldots, F_d$ as $H_0^F : \mathbf{CF} = \mathbf{0}$, where $\mathbf{C}$ denotes an appropriate contrast matrix, $\mathbf{F} = (\mathbf{F_1}, \ldots, \mathbf{F_d})'$ the vector of distribution functions, and $\mathbf{0}$ the vector of 0-functions. Moreover it is not reasonable to use the plug-in estimators $\widehat{p}_i = \int \widehat{H} d\widehat{F}_i$ as an intuitive estimator of a fixed model quantity to describe a treatment effect or to compute confidence intervals for p_i - except in the case of equal sample sizes.
This problem can be solved by using the unweighted mean distribution $G = \frac{1}{d} \sum_{i=1}^{d} F_i$ of the distributions $F_1, \ldots, F_d$ in the experiment. Thus, the quantities $\psi_i = \int G dF_i$ are fixed model quantities which can be estimated by the simple plug-in estimator $\widehat{\psi}_i = \int \widehat{G} d\widehat{F}_i$. Note that $\widehat{p}_i$ can be represented by the ranks R_{ik} of the observations X_{ik} as

$$\widehat{p}_i \;=\; \frac{1}{N}\left(\overline{R}_{i\cdot} - \frac{1}{2}\right),$$

where $\overline{R}_{i\cdot} = \frac{1}{n_i} \sum_{k=1}^{n_i} R_{ik}$ denotes the mean of the ranks R_{ik}. In the same way, the quantities $\psi_i = \int G dF_i$ can be estimated by the simple plug-in estimators $\widehat{\psi}_i = \int \widehat{G} d\widehat{F}_i$ which can be represented by the so-called pseudoranks R_{ik}^{ψ} of the observations X_{ik} as

$$\widehat{\psi}_i \;=\; \frac{1}{N}\left(\overline{R}_{i\cdot}^{\psi} - \frac{1}{2}\right),$$

where $\overline{R}_{i\cdot}^{\psi} = \frac{1}{n_i}\sum_{k=1}^{n_i} R_{ik}^{\psi}$ denotes the mean of the pseudoranks of R_{ik}^{ψ} ([4], [3]). The pseudorank estimators $\widehat{\psi}_i$ have similar properties (unbiased, consistent, asymptotic normal distribution) as the usual rank estimators $\widehat{p}_i$.
The above listed paradoxical results do not appear using the unweighted nonparametric effects $\psi_i = \int GdF_i$ and their estimators based on the pseudoranks also not in extreme cases of unequal sample sizes. This is demonstrated reconsidering the above presented examples where paradoxical results were obtained using the usual rank estimators. Moreover, hypotheses based on the quantities ψ_i can be reasonably formulated and confidence intervals for these fixed model quantities can be computed.
It remains the problem of finding the asymptotic covariance matrix $\mathbf{V}$ of $\sqrt{N}\widehat{\psi} = \sqrt{N}(\widehat{\psi}_1, \ldots, \widehat{\psi}_d)'$. This matrix has been shown to be quite involved ([6]) and thus, Gao and Alvo [3] consider the more restrictive hypothesis $H_0^F : \mathbf{CF} = \mathbf{0}$ instead of $H_0^{\psi} : \mathbf{C}\psi = \mathbf{0}$. It is shown in [1] that under this restrictive hypothesis, the covariance matrix $\mathbf{V}$ of $\sqrt{N}\widehat{\mathbf{p}}$ has a diagonal structure and the diagonal elements can be easily estimated from the usual ranks.
A simple technique using only some matrix algebra, however, enables the estimation of $\mathbf{V}$ in the general case using pairwise rankings. The details are to be found in [2]. The R-package *rankFD* performing the computations in general factorial designs can be downloaded from CRAN.

References

[1] Akritas, M. G., Arnold, S. F. and Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association* **92**, 258-265.

[2] Brunner, E., Konietschke, F., Pauly, M., and Puri, M.L. (2017). Rank-Based Procedures in Factorial Designs: Hypotheses about Nonparametric Treatment Effects. *Journal of the Royal Statistical Society Series B*, (to appear).

[3] Gao, X. and Alvo, M. (2005*b*). A Unified Nonparametric Approach for Unbalanced Factorial Designs. *Journal of the American Statistical Association*, **100**, 926–941.

[4] Kulle, B. (1999). Nichtparametrisches Behrens-Fisher-Problem im Mehrstichprobenfall. Diploma Thesis, Inst. of Math. Stochastics, University of Göttingen.

[5] Peterson, I. (2002). Tricky Dice Revisited. *Science News* **161**, https://www.sciencenews.org/article/tricky-dice-revisited

[6] Puri, M. L. (1964). Asymptotic efficiency of a class of *c*-sample tests. *Annals of Mathematical Statistics* **35**, 102–121.

Asymptotic Permutation Tests in General Factorial Designs

Session: **Nonparametric Methods 2**

Frank Konietschke *Department of Mathematical Sciences, University of Texas at Dallas, U.S.A.*
Edgar Brunner *Department of Medical Statistics, University of Göttingen, Germany*
Markus Pauly *Institute of Statistics, Ulm University, Germany*

Abstract: We investigate permutation methods as small sample size approximations in general linear models without assuming normally distributed error terms or equal variances.

1 Introduction

Factorial designs

$$X_{ik} = \mu_i + \epsilon_{ik},\ i = 1, \ldots, a;\ k = 1, \ldots, n_i,$$

are common tools in experimental sciences, e.g., economics, ecology or medicine. Here, the index i denotes the level of a factor A, and k the unit within level i. Higher-way layouts can be modeled by sub-indexing the index i. For convenience, the expected values are collected in $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_a)'$. Here, $E(\epsilon_{i1}) = 0, Var(\epsilon_{i1}) = \sigma_i^2$, and $E(\epsilon_{i1}^4) < \infty$, thus, homogeneous variances across the different groups is not assumed. Traditionally, parametric inference methods such as the classical F-test for normal models are used for analyzing data. The classical F-test is derived under the assumptions of equal variances ($\sigma_i^2 \equiv \sigma^2$) and normally distributed errors ϵ_{ik}. However, these assumptions are often unsuccessful or are difficult to verify. Moreover, if they are not satisfied, the classical F-test may behave conservatively or even liberally under the null hypotheses. To overcome this problem in the case of the normal distribution, several approximations for designs with heteroscedastic variances have been proposed, e.g. the generalized Welch-James test or the ANOVA-type statistic by Brunner et al. (1997). However, if data is not normally distributed, all of these methods tend to result in liberal or conservative conclusions, depending on the shape (skewness) of the distributions. Thus, there is a need for statistical methods which do not rely on the assumptions of normality and homoscedastic variances.

2 The method

Let $\overline{X}_{i\cdot} = n_i^{-1} \sum_{k=1}^{n_i} X_{ik}$ and $\widehat{\sigma}_i^2 = (n_i - 1)^{-1} \sum_{k=1}^{n_i} (X_{ik} - \overline{X}_{i\cdot})^2$ denote the sample means and variances, respectively. In order to test the general linear hypothesis $H_0 : \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$, the Wald-type statistic (WTS)

$$\begin{aligned} Q_N(\mathbf{X}) &= N \overline{\mathbf{X}}_{\cdot}' \mathbf{H}' (\mathbf{H} \widehat{\mathbf{V}} \mathbf{H}')^{+} \mathbf{H} \overline{\mathbf{X}}_{\cdot}, \text{ where } \overline{\mathbf{X}}_{\cdot} = (\overline{X}_{1\cdot}, \ldots, \overline{X}_{a\cdot})' \text{ and} \\ \widehat{\mathbf{V}} &= diag(\frac{N}{n_i} \widehat{\sigma}_i^2, i = 1, \ldots, a), \end{aligned}$$

is an asymptotically valid procedure, however, very large sample sizes are necessary to achieve an accurate type-1 error rate control. This behavior of the Wald-type test motivated us to improve its behavior for small sample sizes.
We investigate and develop *resampling and permutation methods* for robust analysis of such models. In particular, permutation methods for the analysis of factorial designs have been widely discussed in the literature. Most of the contributions do not provide any theory satisfying the validity of the procedures and/or assume exchangeability across the permuted observations. A natural question that arises is "How to permute data in factorial designs?". In order to test the main effects A and B, say, data can be permuted across the levels of factor A or B individually,

but how to permute data to achieve a permutation test for the interaction effect? In this talk we solve this problem in a feasible way: The factorial structure can be neglected if an appropriate studentized statistic is permuted, i.e. data is permuted within the vector $\mathbf{X} = (X_{11}, \ldots, X_{an_a})'$. We show that the permutation distribution of the Wald-type statistic mimics the distribution of the Wald test. This means, instead of using the asymptotic distribution of the Wald-test (a χ^2-distribution), p-values and critical values can be estimated from its permutation distribution. Simulation studies show that the methods tend to produce accurate conclusions even with very small sample sizes ($n_i \leq 10$) and non-normal data with heteroscedastic variances. The key advantage of these procedures is that they keep the exactness property of permutation tests if data is exchangeable.

References

[1] Akritas, M. G., Arnold, S. F., and Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *J. Amer. Statist. Assoc.* **92**, 258–265.

[2] Pauly, M., Brunner, E., and Konietschke, F. (2015a). Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society: Series B* **77**, 461–473.

Resampling-based inference for the Wilcoxon-Mann-Whitney effect in survival analysis for possibly tied data

Session: **Nonparametric Methods 2**

Dennis Dobler *Institute of Statistics, Ulm University, Germany*
Markus Pauly *Institute of Statistics, Ulm University, Germany*

Abstract: In a two-sample survival setting with independent survival variables T_1 and T_2 and independent right-censoring, the Wilcoxon-Mann-Whitney effect $p = P(T_1 > T_2) + \frac{1}{2}P(T_1 = T_2)$ is an intuitive measure for discriminating two survival distributions. When comparing two treatments, the case $p > 1/2$ suggests the superiority of the first over the second. Nonparametric maximum likelihood estimators based on normalized Kaplan-Meier estimators naturally handle tied data, which are omnipresent in practical applications. Studentizations allow for asymptotically accurate inference for p. For small samples, however, coverage probabilities of confidence intervals are considerably enhanced by means of bootstrap and permutation techniques. The latter even yields finitely exact procedures in the situation of exchangeable data. Simulation results support all theoretic properties under various censoring and distribution set-ups.

1 Introduction

For the comparison of two survival functions S_1 and S_2, the *Wilcoxon-Mann-Whitney effect* is an intuitive measure. We consider a classical survival settings with continuous, independent life times $T_1 \sim S_1$ and $T_2 \sim S_2$ and independent right-censoring. In this talk, we face the practically relevant situation with ties in the data, i.e. hazard rates need not exist. Therefore, a general definition of the Wilcoxon-Mann-Whitney effect is $p = P(T_1 > T_2) + \frac{1}{2}P(T_1 = T_2)$. Hence, $p > 1/2$ implies a protective survival effect for group 1. We estimate p by Kaplan-Meier curves, which are the nonparametric maximum likelihood estimates of the survival functions S_1 and S_2. Futhermore, we focus on the null hypothesis $H_0^p : p = \frac{1}{2}$ instead of the typical $H_0^S : S_1 = S_2$ as considered in e.g. [2]. We adjust for tied data by using normalizations of the Kaplan-Meier estimators. In the uncensored case, such normalizations lead to mid-ranks. This more realistic assumption of ties in the observations accounts for natural circumstances in many study designs. Therefore, methodology for continuous data should not be applied. The present approaches yield one- and two-sided test procedures for the null hypothesis H_0^p of no group effect. In the uncensored case this is also called the *nonparametric Behrens-Fisher problem*, see e.g. [1] and [4].

2 The method

In the situation of two samples with i.i.d. right-censored survival observations with perhaps different survival and censoring distributions among both groups, p is estimated via $\hat{p} = \phi(\hat{S}_1, \hat{S}_2)$, a functional of both sample-specific Kaplan-Meier curves. A variance estimator $\hat{\sigma}^2$ yields the asymptotic pivot $\sqrt{\frac{n_1 n_2}{n}}\frac{\hat{p}-p}{\hat{\sigma}} \xrightarrow{d} N(0,1)$ as $\min(n_1, n_2) \to \infty$. In order to correct for inaccuracies in small samples, a pooled bootstrap and a permutation technique are used to obtain asymptotically exact critical values, but a different studentization is required. In case of exchangeable data in both sample groups, the permutation approach even yields finitely exact inference procedures, such as 95% confidence intervals for p.

3 Example

We consider a data-set with right-censored survival times of tongue cancer patients, cf. [3]. It contains 80 patients of which (group 1) $n_1 = 52$ are suffering from an aneuploid and (group 2) $n_2 = 28$ from a diploid tumor. 21 patients in group 1 and six patients in group 2 are

right-censored, otherwise the times of death have been recorded. The data set contains some tied event times. Figure 14 shows that the aneuploid Kaplan-Meier curve is always above the Kaplan-Meier curve of the diploid group. In this talk, we examine whether this gap already yields significant survival superiority of the first group, i.e. $p > \frac{1}{2}$ significantly.

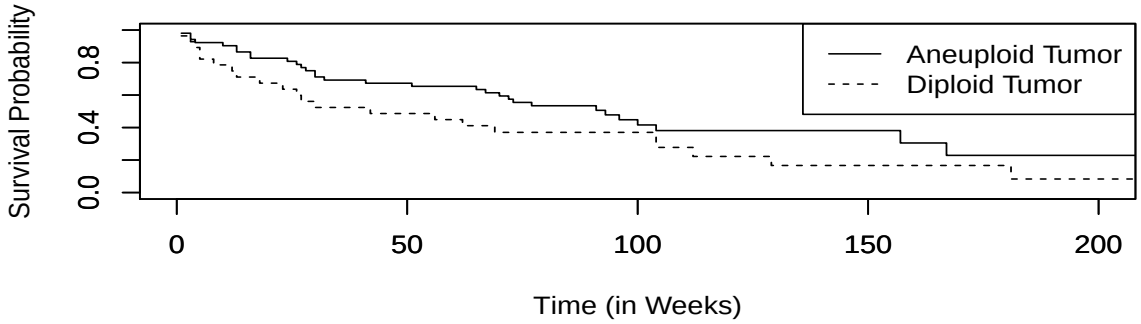


Figure 14: Kaplan-Meier curves for diploid (- -) and aneuploid tumor (—) groups.

References

[1] Brunner, E. and Munzel, U. (2000) *The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation*, Biometrical Journal, 42(1):17–25.

[2] Efron, B. (1967) *The two sample problem with censored data.* In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 4:831–853.

[3] Klein, J. P. and Moeschberger, M. L. (2003) *Survival Analysis: Techniques for Censored and Truncated Data*, Springer Science & Business Media.

[4] Neubert, K. and Brunner, E. (2007) *A studentized permutation test for the non-parametric Behrens–Fisher problem.* Computational Statistics & Data Analysis, 51(10):5192–5204.

Electroluminescence Image Analysis and Suspicious Areas Detection

Session: **Stochastic Models in Technology, Reliability, and Quality 3**

Evgenii Sovetkin *RWTH Aachen University, Wüllnerstr. 3, 52062, Aachen, Germany*
Ansgar Steland *RWTH Aachen University, Wüllnerstr. 3, 52062, Aachen, Germany*

Abstract: In this work we consider several problems arising in quality control analysis of electroluminescence (EL) images of photovoltaics (PV) modules. The EL image technique is a useful tool for investigating the state of a PV module and allows us to look inside a module and to analyse the crystalline structure at high resolution. However, there is a lack of methods to employ the information provided by EL images in the analysis of large PV systems.
We first consider several practical issues that arise in field studies, i.e. when images are taken under outdoor conditions and not in a lab. We discuss a new problem-specific procedure for automatic correction of rotation and perspective distortions, which to some extent employs statistical approaches such as robust regression; and a procedure for automatic detection of the module and its cell areas (by means of a modified version of the Hough Transform). Those techniques provide us with images of the PV module cells, intensity light of which are of the main interest in quality study.
Secondly, we discuss a simple test statistics to screen large databases of EL image data aiming at the detection of malfunctioning cells. The asymptotics is established for a general class of random fields, covering the asymptotic distribution and the estimation of the spatial covariances.
Lastly, we present a spatial version of the test which aims at detecting the position of a defect and discuss asymptotic properties. The finite sample properties of the procedure are studied by simulations.
Keywords: Image Processing, Nonparametrics, Change-Point, Spatial Statistics.

Multistage acceptance sampling plans for nonparametric quality control under dependent sampling designs

Session: **Stochastic Models in Technology, Reliability, and Quality 3**

Andreas Sommer *Institut für Statistik und Wirschaftsmathematik, RWTH Aachen University, Germany* †

Ansgar Steland *Institut für Statistik und Wirschaftsmathematik, RWTH Aachen University, Germany*

Abstract: An important problem arising in statistical quality control is to check whether or not a shipment (or lot) of produced items is in agreement with the specifications. The quality of the produced items can, however, change over time due to many reasons, so that the lot should be reinspected at later points in time, too. We extend the procedure of acceptance sampling by variables for an arbitrary number of inspection points under nonparametric models in the presence of additional data.

We derive approximations of the test statistics for independent as well as dependent sampling schemes leading to asymptotically optimal sampling plans, which are time-individual. Every sampling plan consists of the sample size and the corresponding critical value, which are itself random variables. Miscellaneous properties of the sampling plans are shown, in particular consistency and asymptotic normality. The stopping time of the procedure and the average outgoing quality are elaborated.

In a simulation study, optimal choices of parameters as well as finite sample properties of the sampling plans are examined. Moreover, we propose a bootstrap procedure in order to improve the accuracy of the procedure in terms of the consumer and producer risks.

† Corresponding author, e-mail address: `sommer@stochastik.rwth-aachen.de`

Vagueness, Imprecision, Belief and Tax Fraud Investigation

Session: **Stochastic Models in Technology, Reliability, and Quality 3**

Hans-J. Lenz *Freie Universität Berlin*

1 Introduction

Tax Fraud is a criminal activity done by a manager of a firm or by a tax payer who intentionally manipulates tax data to deprive the tax authorities of money for his own benefit. Tax fraud is a kind of *Data Fraud*, and happens everywhere and at any time.
While the hype about "*Big Data*" is dominating research in computer science tax fraud detection is intrinsically related to the "*Small Data*" area. Starting with a initial suspicion the human creativity is needed for a step-by-step investigation of the private or firm's environment hunting for the tax liability materialized in various forms like cash, foreign currency, gold, equities, real estate etc. The point is that knowledge cannot be created by data mining – due to an intrinsic lack of information in the beginning of any investigation. However, experience and imagination of the investigator lead to assumptions, hypotheses and beliefs about the tax betrayer's tricks and behavior, and substitute data to some extent. A tax fraud case is characterized by the vagueness of the belief, and the imprecision of the tax liability especially in the beginning.
The investigation can be embedded into the Bayesian Learning Theory. This approach is based on hints, investigation (unscrambling information) and the integration of partial information in a stepwise procedure. The kick-off is an initial suspicion issued by a tax officer, an insider like a fired employee, disappointed companion or wife, envious neighbor or inquisitive custom collector at the border of a tax harbor etc. This first step can be conceived as the fixing of the prior distribution $p(\theta)$ of the tax liability size, θ, of a tax betrayer. The next step is concerned with opening a new case at the tax authority's site, and getting access to the full tax file of the suspect. Thereby new evidence (x) and further hints are created. Formally, the likelihood of the tax fraud, $l(x|\theta)$, is established. This allows updating of the initial suspicion for gaining the posterior distribution $p(\theta|x) \propto l(x|\theta)p(\theta)$.
Learning is performed if further sequential investigations deliver more information on the non-conforming suspect's life style related to his annual taxable income and assets. The necessary investigations are tricky for getting insight into the betrayer's life style, and make use of criminal investigator's good practice like, for instance, "*Simple issues first!*" or "*Find the safe*".
More formally, we take the former posterior $p(\theta|x)$ as a new prior $p^*(\theta)$ and combine it with the new facts (x') about the tax crime using the likelihood $l^*(x'|\theta)$. This leads to the updated suspicion $p^*(\theta|x')$ as the new posterior. The investigation stops when the tax liability is fixed, and p* as a measure of certainty is near 100%, or expressed by the linguistic term "*is most probably, if not certainly*". Alternatively, the tax authorities may stop it when p^* drops down extremely, i.e. doubts increased. In the first case the charge is left to the judicial system to prosecute, judge and eventually arrest the betrayer.
Alternative approaches are for instance *Case Based Reasoning, Rule-based Systems, Fuzzy Logic* and *Social Network Analysis*. Finally, innocent and lawful people will be satisfied because betrayers never will be able to construct a really perfect manipulated ("artificial") world of facts and figures, and in the long run all of them will be captured as claimed by the former German investigator F. Wehrheim (2011).

Confidence Intervals for Standardized Mortality Ratios

Session: **Stochastic Models in Technology, Reliability, and Quality 3**

Johannes Rauh *Institute for Quality Assurance and Transparency in Healthcare (IQTIG), Germany*
Michael Höhle *Institute for Quality Assurance and Transparency in Healthcare (IQTIG), Germany*

Abstract: When measuring the quality of a hospital by event rates of certain indications or after specific treatment, it is important to perform a risk adjustment in order to account for the fact that different hospitals see patients that systematically differ in their characteristics, such as age or preexisting health conditions. One way to summarize the hospital results over such strata is the computation of standardized mortality (or morbidity) ratios (SMRs). In the present work we follow an approach by Clay and compute exact confidence intervals for SMRs based on the Poisson binomial distribution. Furthermore, we compare the coverage of such intervals with existing procedures.

1 Introduction

We consider a situation where the outcome for each patient can be regarded as dichotomous, where, e.g., one level can be characterized as a good outcome and the other level as a malign outcome. An example could be in-hospital death after a given treatment. Let n_i be the number of patients under a given treatment in hospital i, and let o_i be the number of the treatments among these that result in a malign outcome. Instead of comparing the raw rate $p_i = o_i/n_i$ among hospitals we would like to control for the effect of confounding due to, e.g., age. One classical way to do this is by indirect standardization [6]. Here, an expected risk $e_{i,j} \in (0,1)$ for a malign outcome is estimated for each patient j in hospital i by taking into account the patient's characteristics. The sum $e_i = \sum_{j=1}^{n_i} e_{i,j}$ of these risks for all patients in hospital i then equals the expected number of malign outcomes. The SMR of hospital i is defined as $\hat{\mathrm{SMR}}_i = o_i/e_i$. It indicates the performance of the hospital: A ratio smaller than one says that the number of observed complications is smaller than expected. For the purpose of our work, we assume that the expected risks $e_{i,j}$ are given, and we are interested in how to take statistical uncertainty into account when reporting the SMRs.

2 The Poisson binomial distribution

If there is no effect of the hospital on the treatment quality (and if the model for computing the $e_{i,j}$s contains all relevant adjustment factors appropriately), then o_i is a sum of independent Bernoulli random variables $o_{i,j}$ with parameters $e_{i,j}$, $j = 1, \dots, n_i$. Such a distribution is called *Poisson binomial distribution* (or *generalized binomial distribution*). We denote this distribution by $\mathrm{PBin}(\{e_{i,j}\})$.
To compute a confidence intervall for the SMR, we define the SMR of an arbitrary joint distribution for the $o_{i,j}$ as

$$\mathrm{SMR}_i = \mathrm{E}\left(\frac{\sum_{j=1}^{n_i} o_{i,j}}{e_i}\right).$$

For example, if the underlying distribution is indeed $\mathrm{PBin}(\{e_{i,j}\})$, then $\mathrm{SMR}_i = 1$. In general, the support of SMR_i (i.e. the set of possible values) is the interval $[0, n_i/e_i]$. Next, we discuss different ways of computing approximate and exact confidence intervals by approximating $\mathrm{PBin}(\{e_{i,j}\})$ by a well-known distribution or by embedding $\mathrm{PBin}(\{e_{i,j}\})$ in a family of Poisson binomial distributions.

2.1 Approximation by binomial and Poisson distribution

Since the Poisson binomial distribution has many parameters and since its probability mass function can be computationally demanding to compute, it is customary to approximate the Poisson binomial distribution by other distributions. One possibility is to replace all parameters $e_{i,j}$ by their average $\overline{e_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} e_{i,j}$. The Poisson binomial distribution then becomes a simple binomial distribution:

$$\mathrm{PBin}(\{e_{i,j}\}) \approx \mathrm{Bin}(\overline{e_i}, n_i),$$

For a binomial distribution, exact confidence intervals can be computed for the parameter $\overline{e_i}$. Alternatively, when n_i is large, one can compute asymptotic Wilson intervals for $\overline{e_i}$ (see, e.g., [3]). These confidence intervals can then be re-scaled by the factor n_i/e_i to obtain a confidence interval for SMR_i. If the parameters $e_{i,j}$ are sufficiently small, the binomial distribution can be further approximated by a Poisson distribution, for which exact confidence intervals can be computed and rescaled.
These approximations are only good if the parameters $e_{i,j}$ lie close to each other: The variances of $\mathrm{PBin}(\{e_{i,j}\})$ and $\mathrm{Bin}(\overline{e_i}, n_i)$ are related via

$$\mathrm{Var}(\mathrm{PBin}(\{e_{i,j}\})) = \mathrm{Var}(\mathrm{Bin}(\overline{e_i}, n_i)) - \sum_{j=1}^{n_i}(e_{i,j} - \overline{e_i})^2.$$

Thus, averaging the parameters $e_{i,j}$ overestimates the variance and leads to conservative confidence intervals that are too large. In our applications, we observe that the expected risk $e_{i,j}$ varies a lot between different patients, and so this effect becomes important.

2.2 Asymptotic Wald confidence intervals

Another possibility is to compute an asymptotic Wald confidence interval. Let

$$\sigma_i^2 = \mathrm{Var}(\mathrm{PBin}(\{e_{i,j}\})) = \sum_{j=1}^{n_i} e_{i,j}(1 - e_{i,j}).$$

Then the Wald confidence interval for the SMR is defined by

$$\hat{\mathrm{SMR}}_i \pm \frac{1}{e_i} z_{1-\alpha/2}\sigma_i.$$

The coverage of such intervals as well as bootstrap based intervals was already investigated in [4]. In our applications, we observe that the normal distribution often is not a good approximation, even for large hospitals with a few thousand treatments. The generalized binomial distribution has heavier tails, and so the Wald confidence intervals tend to be too short. On the other hand, if the data lies on the boundary (i.e. $\hat{\mathrm{SMR}}_i$ is close to either 0 or n_i/e_i), the Wald interval may reach out to values that lie outside of the support of SMR_i and contain either negative values or values larger than n_i/e_i.

2.3 Exact confidence intervals

We follow the approach of [2] and use the Poisson binomial to obtain exact confidence intervals for the hospital specific SMRs. Using a logistic regression setup we introduce a hospital specific parameter κ_i. To be precise, we assume that $o_{i,j}$ is Bernoulli-distributed with parameter $e'_{i,j}$, where $e'_{i,j}$ satisfies

$$\mathrm{logit}(e'_{i,j}) = \mathrm{logit}(e_{i,j}) + \kappa_i,\ \kappa_i \in \mathbb{R}.$$

This means that $\exp(\kappa_i)$ reflects the multiplicative increase in the odds of a malign outcome due to treatment in hospital i.
The SMR of $\mathrm{PBin}(\{e'_{i,j}\})$ is now a function of the parameter κ_i:

$$\mathrm{SMR}_i = \frac{\sum_{j=1}^{n_i} e'_{i,j}}{e_i}.$$

Thus, we can (numerically) compute an exact confidence interval for κ_i and transform this interval to obtain an exact confidence interval for SMR_i. Along the same lines, we can also compute mid-p confidence intervals for SMR_i [1].

3 Comparing the confidence intervals

To compare the different methods, we use data about pneumonia patients from the German hospital profiling in the year 2015 [5]. The outcome $o_i = 1$ in the specific example corresponds to in-hospital death. The patient risks $e_{i,j}$ were computed according to the 2015 logistic risk-adjustment model from the German hospital profiling. The confidence intervals were computed using R [7].
The different confidence intervals vary quantitatively and qualitatively. Most notably, the lower boundary of the Wald interval can be outside of the support of the SMR. Apart from this, the Wald intervals tend to be too small. On the other hand, the Poisson intervals tend to be too large.
To compare the coverage of the confidence intervals, we use the following setting: We fix a single hospital and the corresponding parameter $e_{i,j}$. For different values of κ_i (and hence SMR_i), we consider the Poisson binomial distribution $\mathrm{PBin}(\{e'_{i,j}\})$ and compute the coverage probability of the various confidence intervals, as a function of κ_i or, equivalently, as a function of SMR_i.
The coverage probability of the Wald interval can be small when the risks $e_{i,j}$ are heterogeneous, even when considering a few thousand patients. The Wilson intervals perform better, because two effects cancel each other: Replacing the $e_{i,j}$ by their average increases the variance and makes the intervals bigger, but the normal approximation neglects the heavy tails, making the intervals smaller. The Poisson approximation is not valid in our case, since the average risk is not small (larger than 10%), and it leads to very large intervals. Altogether, the exact intervals based on the Poisson binomial perform well in terms of coverage.

References

[1] Berry, G. and Armitage, P. (1995). Mid-P Confidence Intervals: A Brief Review, *The Statistician* 44.4, pp. 417–423.

[2] Clay, T. (2011). Exact confidence intervals for risk-adjusted rates versus trouble in River City, *Proceedings of the SAS® Global Forum 2011 Conference*, SAS Institute Inc.

[3] Fleiss, J. L., Bruce L., and Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*, 3rd ed., Wiley: Hoboken.

[4] Hosmer, D. W. and Lemeshow, S. (1995). Confidence interval estimates of an index of quality performance based on logistic regression models, *Stat Med* 14.19, pp. 2161–2172.

[5] IQTIG (2016). Qualitätsindikatoren und Bundesauswertungen. URL: `https://www.iqtig.org/ergebnisse/qs-verfahren/` (12/08/2016).

[6] Keiding, N. and Clayton, D. (2015). Standardization and Control for Confounding in Observational Studies: A Historical Perspective, *Statistical Science* 29.4, pp. 529–558.

[7] R Core Team (2016). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org`

Estimation of conditional distribution functions from data with additional measurement errors

Session: **Nonparametric Regression and Density Estimation 2**

Matthias Hansmann *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany*

Michael Kohler *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany*

Abstract: We study the problem of estimating conditional distribution functions from data that contain additional measurement errors. The only assumption on these errors is that a weighted sum of the absolute errors tends to zero with probability one for sample size tending to infinity. In particular, we do not assume that the measurement errors are independent with expectation zero.

We prove sufficient conditions on the weights of a local averaging estimate of the conditional distribution function, based on data with measurement errors, which ensure the strong pointwise consistency. We show that these conditions are fullfilled by the weights of the kernel estimate. Furthermore, we investigate the rate of convergence and give sufficient conditions on the weights, which ensure a certain rate of convergence.

The results are applied in the context of experimental fatigue tests to estimate the conditional distribution function of the number of cycles until failure and to optimize the geometry of a component w.r.t. this estimate.

Weighted Nearest Neighbor Estimates for Nonlinear Time Series

Session: **Nonparametric Regression and Density Estimation 2**

Włodzimierz Greblicki *Wrocław School of Information Technology "Horizon", Wejherowska 28, 54-239 Wrocław, Poland, email: wlodzimierz.greblicki@gmail.com*

Mirosław Pawlak *Department of Electrical and Computer Engineering, University of Manitoba, Canada, email: Miroslaw.Pawlak@umanitoba.ca*

Abstract: This paper considers the nonparametric estimation problem for a class of nonlinear time series systems that is characterized by its cascade structure. This is a series connection of a nonlinear memoryless subsystem followed by a linear dynamic system. The input-output training data generated from the system are dependent and they do not reveal the strong mixing property. The nonlinear part of the system is recovered with the weighted k-nearest neighbor regression estimate. The *a priori* information is nonparametric, both the nonlinear characteristic and the impulse response of the linear part are completely unknown and can be of any form. Local and global properties of the estimate are examined. Whatever the probability density of the input signal, the estimate converges at every continuity point of the characteristic as well as in the global sense. We derive the formulas for asymptotic bias and the variance and evaluate the corresponding rate of convergence. The convergence rate is independent of the shape of the input density and is proved to be optimal. These results allow us to find a set of optimal nonnegative weights that further improve the accuracy of our estimation algorithm. This reveals the advantage of the weighted k-nearest neighbor estimate over the traditional uniformly weighted k-nearest neighbor method. The obtained results are also extended to other types of nonparametric and semi-parametric time series models.

References

[1] G. Biau and L. Devroye (2015). *Lectures on the Nearest Neighbour Method*, New York: Springer.

[2] J. Fan and Q. Yao (2003). *A Nonlinear Time Series: Nonparametric and Parametric Methods*, New York: Springer.

[3] F. Giri and E.W. Bai (Eds.) (2010). *Block-Oriented Nonlinear System Identification*, New York: Springer-Verlag.

[4] W. Greblicki and M. Pawlak (2008). *Nonparametric System Identification*. Cambridge: Cambridge University Press.

[5] W. Greblicki and M. Pawlak (2017), "Hammerstein system identification with the nearest neighbor algorithm," *IEEE Trans. Information Theory*, to appear.

[6] R.J. Samworth (2012). "Optimal weighted nearest neighbour classifiers," *Annals of Statistics*, vol. 40, pp. 2733–2763.

[7] S. Yakowitz (1987). "Nearest-neighbour methods for time series," *Journal of Time Series*, vol. 8, pp. 235–247.

A Combined Criterion for Dose Optimisation in Early Phase Clinical Trials

Session: **Sequential Experimental Design**

Iftakhar Alam *Institute for Statistical Research and Training, University of Dhaka, Bangladesh*
Steve Coad *School of Mathematical Sciences, Queen Mary, University of London, U.K.*
Barbara Bogacka *School of Mathematical Sciences, Queen Mary, University of London, U.K.*

Abstract: A combined criterion is introduced for dose optimisation in seamless phase I/II clinical trials. The design considers efficacy and toxicity as endpoints. The criterion is based on the probability of a successful outcome and on the determinant of the Fisher information matrix for estimation of the dose-response parameters. Starting with the lowest dose, the design selects that dose for each subsequent cohort that maximises the defined criterion. The methodology is illustrated with a dose-response model that assumes trinomial responses. The aim is to investigate whether any bridge is possible between best intention and D-optimum designs. Simulation studies show that the method is capable of identifying the optimal dose accurately without exposing many patients to toxic doses.

1 Introduction

Different methods are being developed to increase the popularity of seamless phase I/II clinical trials. There are best intention designs [1] which aim to allocate the best dose to the cohort of patients based on the current knowledge. These designs may lead to poor learning of the dose-response relationship. In contrast, there are methods which rely on optimal design techniques [2]. All of these designs focus on the efficient estimation of the dose-response parameters, so that the optimum dose can be obtained more accurately. However, they may often expose patients to subtherapeutic or toxic doses on their way to efficient parameter estimation. The aim here is to develop a design that exposes not many cohorts in a trial to either subtherapeutic or toxic doses and that can also find the optimum dose accurately.

2 Methods

At each stage of a trial, the optimum dose is selected based on a newly-defined optimisation criterion. A linear combination of the probability of success and the determinant of the Fisher information matrix for the dose-response model is used. The procedure initially assigns the lowest dose to a cohort of patients. The successive cohorts then receive doses based on the criterion until the trial stops according to the stopping rules. To assess the performance of the proposed adaptive design, two efficiency measures [3] are introduced.

The proposed methodology is illustrated with an example which is based on the continuation ratio dose-response model [4]. For each patient, it is assumed that the outcomes can be categorised as neutral, success or toxic. Success is referred to as an outcome which is efficacious but non-toxic. The occurrence of these outcomes depends on dose. An experimental drug is assumed for which the probability of a neutral response decreases monotonically with dose and the probability of toxicity increases monotonically with dose. However, the probability of success may be non-monotonic, increasing or decreasing.

3 Results

Four dose-response scenarios are considered for the simulation study. Twenty doses ranging from 0.5 to 10.0 mg/kg body weight of an experimental drug are available. Each trial assigns the lowest available dose of 0.5 mg/kg body weight to the first cohort of patients. Escalation or de-escalation of doses to the first four cohorts is based on the up-and-down design. Since the

Fisher information matrix is singular, to have a non-zero value for the determinant immediately after the up-and-down stage, at least two different doses need to be assigned to the first four cohorts. After the up-and-down stage, the dose escalation is based on the combined criterion, using the updated estimates of the dose-response parameters at each stage. Each trial is stopped when the same dose is repeated for six cohorts or when the trial reaches the maximum number of 20 cohorts, whichever comes first. Results from 1,000 simulations of each of the scenarios show that, on the whole, combining the two approaches can improve the performance of the adaptive design.

References

[1] Zhang, W., Sargent, D. J., and Mandrekar, S. (2006). An adaptive dose-finding design incorporating both toxicity and efficacy, *Statistics in Medicine* 25: 2365-2383.

[2] Dragalin, V. and Fedorov, V. V. (2006). Adaptive designs for dose-finding based on efficacy-toxicity response, *Journal of Statistical Planning and Inference* 136: 1800-1823.

[3] Hardwick, J., Meyer, M. C., and Stout, Q. F. (2003). Directed walk designs for dose-response problems with competing failure modes, *Biometrics* 59: 229-236.

[4] Agresti, A. (1990). *Categorical Data Analysis*, Wiley: New York.

Inadequacy of the classical statistical test under response-adaptive randomization procedures

Session: **Sequential Experimental Design**

Maroussa Zagoraiou *Department of Business Administration and Law, University of Calabria, Italy*

Abstract

The goals of individual care and experimental information often come into conflict and the ensuing ethical problem is how to balance the welfare of the patients in the trial against a possible knowledge gain that will improve the care of future patients. In the context of clinical trials it is widely accepted that Response Adaptive (RA) randomization is a possible answer and this is the reason for which over the past two decades there has been a growing stream of statistical papers on this topic (for a recent review see [1]).
Often the conflicting goals related to the ethical demand of maximizing the subjects care and to the statistical aim of drawing correct inferential conclusions with high precision can be formalized through the adoption of target allocations of the treatments that could be regarded as a valid trade-off among ethics and inference. Generally, these targets depend on the unknown model parameters and they can be approached asymptotically by using suitable RA randomization procedures in order to converge to the chosen target such as the efficient randomized adaptive design [4].
In this context, the large majority of the design literature is focused on the problem of optimizing the estimation of the treatment effects, while little attention is devoted to hypotheses testing - almost exclusively in the case of binary response trials (see e.g. [3, 5]).
This presentation is based on the recently published paper [2], where the impact of RA randomization rules is analyzed in the case of normally response trials for testing the superiority of one of two available treatments. In particular, we show that several target allocations may induce an anomalous behaviour of the power of Wald test, that could be locally decreasing, or could vanish as the sample size of the difference between the treatment effects grows, leading to a consistent loss of inferential precision. Moreover, we suggest a modified version of Wald test which, by using the current allocation proportion to the treatments as a consistent estimator of the target, avoids some degenerate scenarios and so it should be preferable to the classical test. We also explore how a correct choice of the initial sample size allows one to overcome the above-mentioned drawbacks regardless of the adopted target.

References

[1] Atkinson, A. C. and Biswas, A. (2013). *Randomised Response-Adaptive Designs in Clinical Trials*, 1st ed., Chapman & Hall/CRC Press: Boca Raton.

[2] Baldi Antognini, A., Vagheggini, A. and Zagoraiou, M. (2016). Is the classical Wald test always suitable under response-adaptive randomization? *Statistical Methods in Medical Research*, available online, DOI: 10.1177/0962280216680241.

[3] Hu, F. and Rosenberger,W. F. (2003). Optimality, variability, power: evaluating response adaptive randomization procedures for treatment comparisons. *Journal of the American Statistical Association*, 98, 671-678.

[4] Hu, F., Zhang, L. X. and He, X. (2009). Efficient randomized adaptive designs. *The Annals of Statistics*, 37, 2543-2560.

[5] Yi, Y. and Wang, X. (2011). Comparison of Wald, score, and likelihood ratio tests for response adaptive designs. *Journal of Statistical Theory and Applications*, 10, 553-569.

A functional urn model for CARA designs

Session: **Sequential Experimental Design**

Andrea Ghiglietti *Department of Mathematics, Università degli Studi di Milano, Italy*

Abstract: We present a general class of covariate-adjusted response-adaptive (CARA) designs introduced in [1], which is based on a new functional urn model. We show strong consistency concerning the allocation probability and the proportion of subjects assigned to the treatment groups, in the whole study and for each covariate profile, allowing the distribution of the responses conditioned on covariates to be estimated nonparametrically.

1 Introduction

In CARA designs the patients in the trial are randomly assigned to $d \geq 2$ treatment groups with an allocation probability that depends on the current patient covariate profile and on the previous patients' covariates, allocations and responses (e.g. see [2]). In this framework, it is desirable that the proportion of subjects of each covariate profile assigned to the treatments converges to a desired target, defined as a function of the response distribution conditionally on the covariates. These kinds of results have been proved in [3] for the case of a known target function of a finite number of parameters, and patients with i.i.d. covariate profiles. Here we present a class of CARA designs introduced in [1], in which the allocation probability may depend by nonparametric estimates of the response distribution, and the patients' covariate profiles are not identically distributed.

2 The model

For any $n \geq 0$, let $\mathbf{Y}_n = (Y_n^1, .., Y_n^d)^\top$ be a vector of functions, with $Y_n^j : \tau \mapsto (0, 1)$, where τ is the covariate space. For any $t \in \tau$, $\mathbf{Y}_n(t)$ represents an urn containing $Y_n^j(t)$ balls of color $j \in \{1, .., d\}$ and $\mathbf{Z}_n(t) = \mathbf{Y}_n / \sum_{j=1}^d Y_n^j$ indicates the proportion of the colors.
When subject n enters the trial, his covariate profile T_n is observed. Then, a ball is sampled at random from the urn identified by T_n (i.e. with proportions $\mathbf{Z}_{n-1}(T_n)$), its color is observed and represented by $\mathbf{X}_n$: $\bar{X}_n^j = 1$ when the color is $j \in \{1, .., d\}$, $\bar{X}_n^j = 0$ otherwise. Then, subject n receives the treatment associated to the sampled color and a response $\bar{\xi}_n$ is collected. The functional urn is then updated as: $\mathbf{Y}_n = \mathbf{Y}_{n-1} + D_n \mathbf{X}_n$, where $\mathbf{X}_n$ and D_n are appropriately defined. Specifically, the weighting function $\mathbf{X}_n : \tau \mapsto [0, 1]^d$ should be such that, for any $t \in \tau$ and $j \in \{1, .., d\}$, $\sum_{j=1}^d X_n^j(t) = 1$ and $\mathbf{E}[X_n^j(t) | \mathcal{F}_{n-1}, T_n] = Z_n^j(t)$, where $\mathcal{F}_{n-1}$ is the σ-algebra of the information related with the first $(n-1)$ patients. This is straightforward for $t = T_n$ by setting $X_n^j(T_n) = \bar{X}_n^j$, since $\bar{X}_n^j$ is conditionally on $\mathcal{F}_{n-1}$ and T_n Bernoulli distributed with parameter $Z_n^j(T_n)$. Then, we define a family of Bernoulli random variables $\{\breve{X}_n^j(t); t \in \tau\}$ with parameters $\{Z_n^j(t); t \in \tau\}$, representing the color that would be sampled in the trial if the covariate profile of subject n was equal to any $t \in \tau$. Finally, we use the quantile function that links this family to compute $\mathbf{X}_n(t)$ for all $t \in \tau$ as $\mathbf{X}_n := \mathbf{E}[\mathbf{X}_n | \mathcal{F}_{n-1}, T_n, \mathbf{X}_n]$. Analogously, we can define the replacement functional matrix $D_n : \tau \mapsto [0, 1]^{d \times d}$ as $D_n := \mathbf{E}[\breve{D}_n | T_n, \mathbf{X}_n, \bar{\xi}_n]$, where $\breve{D}_n(t)$ is a function of a random variable having the same distribution of the response observed from a subject with covariate profile t, i.e. the response that would be observed in the trial if the covariate profile of subject n was equal to any $t \in \tau$. Naturally, $D_n(T_n) = \breve{D}_n(T_n)$. Since the quantile functions of the response distributions are typically unknown, D_n is computed by using the corresponding (parametric or nonparametric) estimators obtained with the information in $\mathcal{F}_{n-1}$.

3 Results

We now present some consistency results for (i) the probability of allocation of the subjects for each covariate profile ($\mathbf{Z}_n(t)$), (ii) the proportion of subjects associated to each covariate profile assigned to the treatments ($\mathbf{N}_{t,n}/\sum_{j=1}^d N_{t,n}^j$, where $\mathbf{N}_{t,n} := \sum_{i=1}^n \bar{\mathbf{X}}_i \mathbf{1}_{\{T_i=t\}}$), (iii) the proportion of subjects assigned to the treatments ($\mathbf{N}_n/n$, where $\mathbf{N}_n := \sum_{i=1}^n \bar{\mathbf{X}}_i$). Consider the following assumptions:

(A1) for any $t \in \tau$ and $n \geq 1$, $D_n^\top \mathbf{1} = \mathbf{1}$ (*constant balance*);

(A2) denoting by $H(t) := \mathbf{E}[\breve{D}_1(t)]$ the average replacement when the covariate profile is t, we assume that $H(t)$ is irreducible, diagonalizable and there exists $\alpha > 0$ such that $\mathbf{E}[|\mathbf{E}[D_n(t)|\mathcal{F}_{n-1}, T_n, \mathbf{X}_n] - H(t)|\mathcal{F}_{n-1}] = O(n^{-\alpha})$.

Denote by $\mathbf{v}(t)$ the right eigenvector of $H(t)$ associated to $\lambda = 1$, with $\sum_{j=1}^d v^j(t) = 1$, and let μ_{n-1} be the probability distribution of T_n conditioned on $\mathcal{F}_{n-1}$. Then,

(a) for any probability measure ν on τ, we have $\int_\tau \|\mathbf{Z}_n(t) - \mathbf{v}(t)\|\nu(dt) \overset{a.s.}{\to} 0$;

(b) if $\sum_{i=1}^n \mu_{i-1}(\{t\}) \overset{a.s.}{\to} \infty$, we have $\|\mathbf{N}_{t,n}/\sum_{j=1}^d N_{t,n}^j - \mathbf{v}(t)\| \overset{a.s.}{\to} 0$;

(c) if $\int_\tau |\mu_n(dt) - \mu(dt)| \overset{a.s.}{\to} 0$, we have $\|\mathbf{N}_n/n - \int_\tau \mathbf{v}(t)\mu(dt)\| \overset{a.s.}{\to} 0$.

The second-order asymptotic results on these CARA designs can be found in [1].

References

[1] Aletti G., Ghiglietti A., Rosenberger W.F. (2016). Nonparametric covariate-adjusted response-adaptive design based on a functional urn model, *Technical Report* arXiv:1611.09421.

[2] Hu F., Rosenberger W.F. (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*, John Wiley & Sons, New York.

[3] Zhang L.-X., Hu F., Cheung S. H., Chan W. S. (2007). Asymptotic properties of covariate-adjusted response-adaptive designs, *Ann. Stat.* 35, 1166-1182.

A Bayesian Adaptive Design for Clinical Trials in Rare Diseases

Session: **Sequential Experimental Design**

S. Faye Williamson *Department of Mathematics and Statistics, Lancaster University, UK*
Peter Jacko *Department of Management Science, Lancaster University, UK*
Sofía S. Villar *MRC Biostatistics Unit, Cambridge, UK*
Thomas Jaki *Department of Mathematics and Statistics, Lancaster University, UK*

1 Introduction

The main goal of the current gold standard design for clinical trials, the *fixed randomised*, is to learn about treatment effectiveness with a view to treat future patients outside of the trial. Its drawbacks for trials involving rare diseases motivate the use of *response-adaptive designs* in which the accruing data on patient responses are used to skew the allocation towards the superior treatments, with an alternative goal of treating the patients within the trial as effectively as possible.
The problem of designing a trial which aims to identify the superior treatment (exploration/learning) whilst treating the trial participants as effectively as possible (exploitation/earning) is a natural application area for *bandit models*, which seek to balance this trade-off in order to obtain an optimal allocation policy which maximises the expected number of patient successes during the given time horizon. We use a bandit model set in the framework of finite-horizon Markov decision processes, where dynamic programming (DP) can be used to develop a Bayesian response-adaptive design. Although the use of bandit models to optimally design a trial is often referred to as the primary motivation for their study, they have never been implemented in real clinical practice for reasons including lack of randomisation, low power, and biased treatment effect estimates [1].
We propose a novel bandit-based design which addresses these key issues in a very appealing way. We incorporate randomisation and add a constraint which penalises if a minimum number of patients are not recruited to each treatment arm. Simulation results for the proposed design show that: (i) the percentage of patients allocated to the superior arm is much higher than in the traditional fixed randomised design; (ii) relative to the optimal (non-rndomised and non-constrained) DP design, the power is largely improved upon and (iii) it exhibits only a very small bias and mean squared error of the treatment effect estimator.

2 The method

We consider a two-armed trial with binary endpoints, immediate responses and a finite number of patients. Patients enter sequentially over time, one-by-one, and each patient is allocated to either treatment A or B. Let X and Y denote the patient's response (either a success 1 or failure 0) from treatments A and B respectively, which we model as independent Bernoulli random variables,

$$X \sim \text{Bernoulli}(1,\ \theta_A) \ \text{ and } \ Y \sim \text{Bernoulli}(1,\ \theta_B), \ \text{ for } \ 0 \leq \theta_A,\ \theta_B \leq 1,$$

where θ_A (θ_B) is the unknown success probability of treatment A (B).
In Section 2.3 of [2] we develop the optimal design using *Constrained Randomised Dynamic Programming* (CRDP). We force actions to be randomised by assigning a probability so that each treatment has a probability of at least $1-p$ of being allocated, where $0.5 \leq p \leq 1$, and will be referred to as the *degree of randomisation*. Note that $p = 0.5$ corresponds to fixed equal randomisation design. We further add a constraint to ensure that we always obtain at least ℓ observations from each treatment arm, where ℓ is a fixed predefined value and will be referred

to as the *degree of constraining*. For details of how this design was implemented in R, refer to the online supplementary material of [2].

3 Overall Performance

Through extensive simulation studies we compare our proposed CRDP design with other designs, including Fixed (equal randomisation), RPW (randomised play-the-winner), DP (non-randomised and non-constrained), WI (the Whittle index approximation of DP) and RDP (randomised but non-constrained). Figure 15 summarises the key features of each design showing that our proposed CRDP design performs well with respect to all of the performance measures.

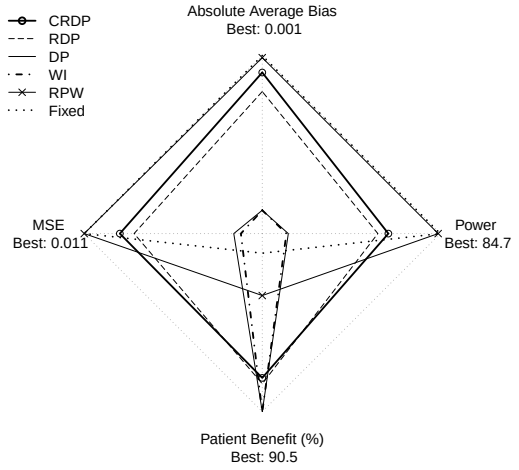


Figure 15: Star plot showing the performance of each design with respect to power, patient benefit, absolute average bias of the treatment effect estimator and MSE in a trial with 75 patients when $\theta_A = 0.5$ and $\theta_B = 0.2$. The best achieved values for each performance measure are depicted at the outer edge.

References

[1] Villar, S. S., Bowden, J., Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges, *Statistical Science* 30 (2), pp. 199–215.

[2] Williamson, S. F., Jacko, P., Villar, S. S., Jaki, T. (2016). A Bayesian adaptive design for clinical trials in rare diseases, *Computational Statistics and Data Analysis*, DOI: 10.1016/j.csda.2016.09.006.

Change point detection with multivariate observations based on characteristic functions

Session: **Analysis, Testing and Change Detection in High Dimensions 4**

Zdeněk Hlávka *Department of Statistics, Charles University,Prague, Czech Republic*
Marie Hušková *Department of Statistics, Charles University,Prague, Czech Republic*
Simos Meintanis *Department of Economics, National and Kapodistrian University, Athens, Greece*

Abstract: The talk concerns break-detection procedures for vector observations, both under independence as well as under an underlying structural time series scenario. The new methods are based on empirical characteristic functions. Asymptotics as well as Monte-Carlo results are presented. The new methods are also applied to time-series data from the financial sector.

1 Introduction

Recently there have been proposed and studied a number of statistical procedures employing empirical characteristics functions for various setups. Among others these are goodness-of-fit tests, model specification tests and tests for detection of changes. The overview paper on such procedures was published by S. Meintanis (2016). The aim of the talk is to present an extension of procedures based on empirical characteristic functions to detection of a change in distribution function of multivariate independent or dependent observations. Among various favorable features for using CFs is that with CFs vector observations are linearly projected onto the real line and the resulting statistics may be written in convenient closed-form expressions. This feature of simplicity is important when dealing with multivariate data.

2 Formulation of the problem

Let $\{X_t,\ t = 1, 2, \ldots, T\}$ be a sequence of random vectors of dimension d ($d \geq 1$), with X_t having the respective distribution function (DF) denoted by $F_t, 1 \leq t \leq T$. Then the classical change-point detection problem is formulated as:

$$\mathcal{H}_0 : F_t \equiv F_0 \text{ for all } t = 1, \ldots, T, \quad \text{vs.} \quad \mathcal{H}_1 : F_t \equiv F_0,\ t \leq t_0;\ F_t \equiv F^0,\ t > t_0, \tag{1}$$

where F_0, F^0 and t_0 are assumed to be unknown.
The null hypothesis can be equivalently formulated via characteristic functions. Here $\varphi_t(u) := \mathbb{E}(e^{iu'X_t})$ denotes the characteristic function (CF) of X_t. The proposed test is based on

$$\sup_t \int_{R^d} \left|\phi_t(u) - \phi^t(u)\right|^2 w(u)du, \tag{2}$$

where $w(\cdot)$ is a suitable weight function,

$$\phi_t(u) = \frac{1}{t}\sum_{\tau=1}^{t} e^{iu'X_\tau}, \qquad \phi^t(u) = \frac{1}{T-t}\sum_{\tau=t+1}^{T} e^{iu'X_\tau},$$

are the empirical CFs computed from $X_1, \ldots, X_t$ and $X_{t+1}, \ldots, X_T,\ t = 1, \ldots, T$, respectively. Large values of (2) indicate that the null hypothesis is violated.

The talk will include results on large sample behavior, discussion on computational aspects and the implementation of the procedures on the basis of suitable resampling techniques. Also results of a Monte Carlo study for the finite-sample properties of the methods along with some empirical applications will be presented.

References

[1] Meintanis S.G. (2016) *A review of testing procedures based on the empirical characteristic function* South African Statist.. J. 50, 1-14.

Change point problem in dynamic panel data

Session: **Analysis, Testing and Change Detection in High Dimensions 4**

Zuzana Prášková *Faculty of Mathematics and Physics, Charles University, Czech Republic*

We consider a dynamic panel data model

$$y_{it} = \beta_{i0} + \sum_{j=1}^{p} \beta_{ij} y_{i,t-j} + \epsilon_{it}, i = 1, \ldots, N,\ t = p+1, \ldots, T$$

where β_{i0} is a fixed effect of the $i-th$ panel, β_{ij} are coefficients with lagged variables $y_{i,t-j}$ and ϵ_{it} are errors.
Our goal is to detect change either in β_{i0} or β_{ij} for $i \in \mathcal{I} \subseteq \{1, \ldots, N\}$ at time $1 < t^* \leq T$ that is unknown and same for all panels. The problem is solved as the hypothesis testing in the model

$$y_{it} = (\beta_i + \delta_i \mathbb{I}[t > t^*])' x_{it} + \epsilon_{it},\ i = 1, \ldots, N,\ t = p+1, \ldots, T$$

where

$$\begin{aligned} \beta_i &= (\beta_{i0}, \beta_{i1}, \ldots, \beta_{ip})' \\ \delta_i &= (\delta_{i0}, \delta_{i1}, \ldots, \delta_{ip})',\ \delta_{ij} \geq 0 \\ x_{it} &= (1, y_{i,t-1}, \ldots, y_{i,t-p})' \end{aligned}$$

when we formulate the null hypothesis

$$H_0 : \delta_1 = \delta_2 = \cdots = \delta_N = 0$$

against the alternative

$$H_1 : \delta_i \neq 0 \ \text{ for some } \ i \in \{1, \ldots, N\}.$$

We generalize procedures developed in [1] and [2] to detect change in the mean of panel data. A test statistic is based on the least-squares estimators (LSE) of parameters and cumulative processes of weighted LSE residuals computed separately in each panel that are assumed to be mutually independent (though the errors in separate panels can be dependent), and on a high-dimensional aggregation of such cumulative processes. The asymptotic distribution of the test statistic under the null hypothesis is established as $N \to \infty$ and $T \to \infty$ but $N/T^2 \to 0$. It is also shown that the test is consistent under local alternatives. More general model that considers cross-panel dependence is shortly discussed.
The approach is completely different from [3] where number of observations T is fixed and/or small.

Acknowledgement. Research is supported by the Czech Science Foundation project GA15-09663S.

References

[1] Bai J. (2010). Common breaks in means and variances for panel data. *Journal of Econometrics*, 157, 78–92

[2] Horváth L., Hušková M. (2012). Change-point detection in panel data. *Journal of Time Series Analysis*, 33, 631–648

[3] De Wachter S., Tzavalis E. (2012). Detection of structural breaks in linear dynamic panel data models. *Computational Statistics and Data Analysis*, 56, 3020–3024.

Bootstrap and Change-Point Detection in Functional Time Series and Random Fields

Session: **Analysis, Testing and Change Detection in High Dimensions 4**

Martin Wendler *Institut für Mathematik und Informatik, Ernst-Moritz-Arndt-Universität Greifswald, Germany*

Abstract: We propose new change point tests for series or fields of weakly dependent random variables taking their value in a Hilbert space. The alternative is at most one change-point in the case of a time series, or an rectangular change-set in the case of random fields. To obtain critical values, we will generalize bootstrap methods to Hilbert space valued random variables.

1 Change Point Tests in Hilbert spaces

We consider sequences or fields of random variables taking their value in a Hilbert space. The random variables are assumed to be weakly dependent, meaning that they fulfill some mixing conditions. New test for the hypothesis of stationary are proposed. The alternative is at most one change-point in the case of a time series, or an rectangular change-set in the case of random fields.
The asymptoic distribution of the test statistics is obtained with the continuous mapping theorem from new functional central limit theorems for the partial sum process in a Hilbert space.

2 Bootstrap Methods

Because the limit distribution is difficult to evaluate and depends on a high-dimensional, difficult to estimate variance parameter, we propose to use bootstrap methods. In the case of time series, we study the nonoverlapping block bootstrap, in the case of random fields the dependent wild bootstrap and show the validity of these methods.
In a simulation study, we will show that our test outperforms previous proposals based on dimension reduction. Our new test can also be used to detect arbitrary change in the distribution function of real valued observations.

References

[1] O.Sh. Sharipov, J. Tewes, M. Wendler (2016): Sequential block bootstrap in a Hilbert space with application to change point analysis. *Canadian Journal of Statistics* 44 (3), 300-322.

[2] B. Bucchia, M. Wendler (2015): Change-Point Detection and Bootstrap for Hilbert Space Valued Random Fields. *preprint*, arXiv:1511.02609.

Simultaneous Confindence Intervals for Graphical Multiple Tests

Session: **Simultaneous Statistical Inference**

Werner Brannath *Institute for Statistics, Faculty 03, University Bremen, Germany*

Abstract: We will provide an introduction to graphical multiple tests and corresponding simultaneous confidence intervals (SCIs). We will illustrate that these confidence intervals do often not provide more information than the cheer hypothesis tests. We therefore introduce a new method to construct SCIs for graphical tests that are always more informative than the corresponding multiple hypothesis tests. We will illustrate the method by examples and simulation results.

1 Introduction

Graphical multiple tests, as introduced in [2], are a powerful tool to design multiple test procedures for clinical trials with multiple endpoints and multiple treatment doses or regimen. With a graphical multiple tests it is easy to account for preferences or hierarchies among the different null hypotheses. These tests control the family wise error rate in the strong sense. They are closed testing procedures and hence come along with major difficulties in extending them to simultaneous confidence intervals (SCIs).
The extension to SCIs is useful because the intervals (usually) provide more information than the multiple tests itself. We may, for instance, wish to exclude parameter values within the alternative hypothesis if the corresponding null hypothesis has been rejected. For instance, this permits to quantify sizes of treatment effects as suggested by regulatory guidelines for clinical trials. Simultaneous confidence intervals are also of particular value in non-inferiority trials in order to improve the pre-specified non-inferiority margin or to claim superiority when the data show compelling evidence for superiority.
Unfortunately, the currently known extensions of closed testing procedures lead to SCIs that often do not provide any additional information to the sheer hypothesis tests (see e.g. [5]). As it seems impossible to overcome this severe shortcoming, we suggest modifying and extending the given graphical test to receive SCIs that always provide additional information. The modification can be fine-tuned to balance potential power losses due to the modification with the information content gained by the SCI. Such SCIs were derived and implemented for the Bonferroni-Holm, Gate-Keeping and Fallback procedures ([1], [3], [4]). In our recent research we have extended the approach to the general class of graphical test procedures of [2].

2 The method

Consider a statistical experiment with a m-dimensional parameter $\mu \in \mathbb{R}^m$. Lower simultaneous confidence bounds for μ can be defined by the following general approach. First define for all intersection hypotheses

$$H_0^\mu = \cap_{j=1}^m H_j^{\mu_j} : \theta_1 \leq \mu_1, \ldots, \theta_m \leq \mu_m, \qquad \mu = (\mu_1, \ldots, \mu_m) \in \mathbb{R}^m$$

a level α test and build the $(1-\alpha)100\%$-confidence set

$$\mathcal{C} = \{\mu \in \mathbb{R}^m : H_0^\mu \text{ is not rejected}\}.$$

Finally determine the largest $L = (L_1, \ldots, L_m)$ such that $\mathcal{C} \subseteq \times_{j=1}^m [L_j, \infty)$. For a given graphical test, the new informative confidence interval is obtained by testing each H_0^μ with a specific modification of the given initial graphical test. We derived a simple, iterative algorithm that allows to numerically calculate L without the need to determine $\mathcal{C}$. The algorithm provides an increasing sequence of lower bounds L_n, $n \in \mathbb{N}$, and we can prove mathematically that L_n

converges to L. With a suitable data dependent stopping criteria N, the final L_N provides a conservative approximation of L that "exploits" (in a specific sense) the target level α up to a pre-specified small ϵ.

References

[1] Brannath, W. and Schmidt, S. (2014). A new class of powerful and informative simultaneous confidence intervals. *Statistics in Medicine* 33, 3365–86.

[2] Bretz, F., Maurer, W., Brannath, W., and Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 28, 586–604.

[3] Schmidt, S. and Brannath, W. (2014). Informative simultaneous confidence intervals in hierarchical testing. *Methods of Information in Medicine* 53, 278–283.

[4] Schmidt, S. and Brannath, W. (2015). Informative simultaneous confidence intervals for the fallback Procedure. *Biometrical Journal* 57(4):712–719.

[5] Strassburger, K. and Bretz, F. (2008). Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests. *Statistics in Medicine* 27, 4914–4927.

A modified Benjamini-Hochberg procedure for discrete data

Session: **Simultaneous Statistical Inference**

Sebastian Döhler *Department of Mathematics and Science, Darmstadt University of Applied Sciences, Germany*

Guillermo Durand *UPMC Univ Paris 06, UMR 7599, Laboratoire de Probabilités et Modèles Aléatoires (LPMA), France*

Etienne Roquain *UPMC Univ Paris 06, UMR 7599, Laboratoire de Probabilités et Modèles Aléatoires (LPMA), France*

Abstract: The Benjamini-Hochberg procedure is a classical method for controlling the false discovery rate for multiple testing problems. This procedure was originally designed with a view towards continuous test statistics. However, in many applications, such as the analysis of next-generation sequencing data, the test statistics are discretely distributed. While it is well known that the Benjamini-Hochberg procedure still controls the false discovery rate in the discrete paradigm, it may be unnecessarily conservative. In this talk we aim to improve the Benjamini-Hochberg procedure in such settings by incorporating the discreteness of the p-value distributions. We investigate the performance of these approaches for empirical and simulated data.

From higher criticism tests and local levels of GOF tests to confidence bounds for the proportion of true nulls

Session: **Simultaneous Statistical Inference**

Helmut Finner *Institute for Biometrics and Epidemiology, German Diabetes Center (DDZ), Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Germany*

Veronika Gontscharuk *Institute of Medical Statistics, Medical Faculty at the Heinrich Heine University, Düsseldorf, Germany*

Klaus Strassburger *Institute for Biometrics and Epidemiology, German Diabetes Center (DDZ), Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Germany*

Abstract: Local levels can be viewed as an interesting characteristic of union-intersection based overall tests. In some recent work (cf. [1] – [4]) we studied local levels of union-intersection based goodness of fit (GOF) tests including higher criticism tests, Kolmogorov-Smirnov type tests and Berk-Jones type tests. Local levels indicate regions of high and low sensitivity of such tests. An interesting issue is the asymptotic behavior of local levels for extreme, intermediate and central order statistics. Typically, local levels tend to zero or converge to some positive limit. Thereby, it is impossible that all local levels have a positive limit. Furthermore, by means of suitable *local level shape functions* we can design new GOF tests with pre-determined local level behavior. We illustrate the local level behavior of various GOF tests by animated plots. In some cases, the finite local level behavior is far away from the asymptotics even for huge sample sizes. Finally, we show how the concept of local levels can be adopted in order to design improved confidence bounds for the proportion of true and false null hypotheses in multiple testing problems with independent p-values.

References

[1] Finner, H., Gontscharuk, V. (2016). Two-sample Kolmogorov-Smirnov type tests revisited: old and new tests in terms of local levels. *Submitted.*

[2] Gontscharuk, V., Finner, H. (2016). Asymptotics of goodness-of-fit tests based on minimum p-value statistics. *Commun. Stat. – Theo. Meth.* 46, 2332–2342.

[3] Gontscharuk, V., Landwehr, S., Finner, H. (2016). Goodness of fit tests in terms of local levels with special emphasis on higher criticism tests. *Bernoulli* 22, 1331–1363.

[4] Gontscharuk, V., Landwehr, S., Finner, H. (2015). The intermediates take it all: asymptotics of higher criticism statistics and a powerful alternative based on equal local levels. *Biometrical J.* 57, 159–180.

Asymptotic Bayes Optimality under Sparsity Revisited

Session: **Simultaneous Statistical Inference**

Florian Frommlet *Department of Medical Statistics, Medical University of Vienna, Austria*
Małgorzata Bogdan *Department of Mathematics, University of Wroclaw, Poland*

Abstract: A while ago we introduced the concept of Asymptotic Bayes Optimality under Sparsity (ABOS) to evaluate certain frequentist multiple testing rules. The main result can be summarized by the fact that procedures controlling the family wise error rate (FWER) like Bonferroni are ABOS only in case of extreme sparsity, whereas procedures controlling the false discovery rate (FDR) like the Benjamini Hochberg procedure are ABOS under a much wider range of sparsity levels. The original results were presented only in case of independent test statistics. We are currently working towards extending these results for correlated test statistics and also for certain model selection criteria in a regression setting. This talk will present preliminary results.

1 Introduction

The notion of Asymptotic Bayes Optimality under Sparsity (ABOS) is based on a decision theoretic framework to analyze a two groups mixture model [1, 2]. The absolutely simplest setting might consider p random variables $X_1, \ldots, X_p$ with normal mixture distribution $X_i \sim \mathcal{N}(\mu_i, \sigma^2)$ where

$$\mu_i \sim (1-\eta)d_0 + \eta\mathcal{N}(0, \tau^2) \; .$$

Here d_0 denotes a point mass at zero under the null and the sparsity parameter η is assumed to be small. Our decision theoretic framework considers an additive overall loss based on individual losses of δ_0 for false positive and δ_A for false negative decisions. In case of independent test statistics it is then relatively easy to derive the rule which minimizes the expected loss which is none other than the Bayes classifier for each individual test.
To define ABOS we further consider an asymptotic framework where we are interested in $\eta \to 0$ (sparsity) as well as $p \to \infty$ (high dimensionality). To obtain meaningful asymptotic results one might either consider large effect sizes $\tau \to \infty$ as in [1] or large sample sizes $n \to \infty$ and $\sigma^2 \propto n^{-1}$ like in [2]. Specific relationships between sparsity η and effect size τ (or sample size n) are assumed to guarantee that the Bayes rule has asymptotically non-vanishing power.
Given this setting a multiple testing rule is called ABOS if its risk R behaves asymptotically like the risk R_{opt} of the optimal Bayes rule, that is $R/R_{opt} \to 1$. Our main results are that the Bonferroni rule is ABOS only in case of extreme sparsity which says that $p \cdot \eta$ essentially has to be constant. In contrast the Benjamini Hochberg procedure adapts to the unknown level of sparsity (see also [3]) in the sense that it is ABOS for a much wider range of sparsity levels.

2 Dependent test statistics

Trying to extend the above results to the case of dependent test statistics one is immediately confronted with the non-trivial problem to determine the rule which minimizes the additive loss function. Apparently one can no longer work with individual Bayes classifiers but one needs to define optimal multivariate decision boundaries. We made some progress in this direction where Figure 1 illustrates the amount of perturbation to be expected in comparison with the individual Bayes classifiers in case of $p = 2$. We will sketch how to use these results to extend the ABOS results from independence to a more general setting.

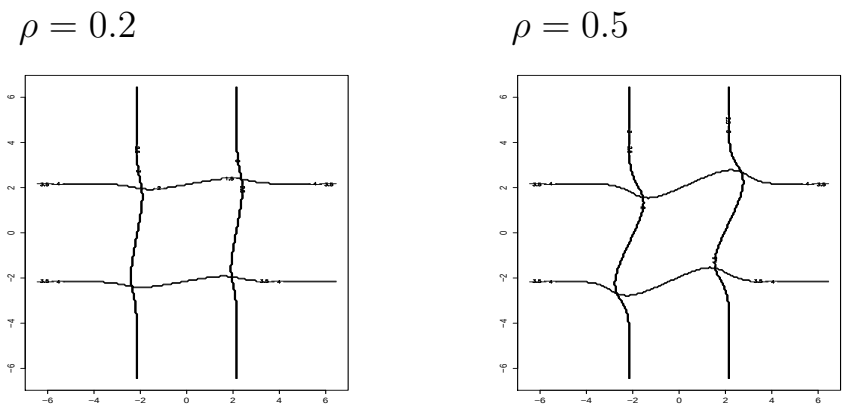


Figure 16: Decision boundaries for Bayes rule under correlation

3 Regression setting

Under orthogonality the ABOS results of multiple testing can be used to establish ABOS of certain model selection criteria. We will introduce several modifications of AIC and BIC which are designed to control FWER or FDR. We will discuss under which assumptions they are ABOS. Finally simulation results in the context of genome-wide association data will illustrate their performance in practice.

References

[1] Bogdan M., Chakrabati A., Frommlet F., Ghosh J.K. (2011) Asymptotic Bayes Optimality under sparsity of some multiple testing procedures. *Ann. Statist.*, **39**, 1551–1579.

[2] Frommlet F., Bogdan M. (2013) Asymptotic Some optimality properties of FDR controlling rules under sparsity. *Electronic Journal of Statistics,*, **7**, 1328–1368.

[3] Abramovich F., Benjamini Y., Donoho D. L., Johnstone I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34**, 584–653.

Semiparametric copula quantile regression for complete or censored data

Session: **Multivariate Distributions and Copula 2**

Anouar El Ghouch *Université catholique de Louvain, Belgium*
Ingrid Van Keilegom *KU Leuven, Belgium*
Mickael De Backer *Université catholique de Louvain, Belgium*

Abstract: When facing multivariate covariates, general semiparametric regression techniques come at hand to propose flexible models that are unexposed to the curse of dimensionality. A semiparametric copula-based estimator for conditional quantiles is investigated for complete or right-censored data. Extending recent work, the main idea consists in appropriately defining the quantile regression in terms of a multivariate copula and marginal distributions. Prior estimation of the latter and simple plug-in lead to an easily implementable estimator expressed, for both contexts with or without censoring, as a weighted quantile of the observed response variable. In addition, and contrary to the initial suggestion in the literature, a semiparametric estimation scheme for the multivariate copula density is studied, motivated by the possible shortcomings of a purely parametric approach and driven by the regression context. The resulting quantile regression estimator has the valuable property of being automatically monotonic across quantile levels, and asymptotic normality for both complete and censored data is obtained under classical regularity conditions. Finally, numerical examples as well as a real data application are used to illustrate the validity and finite sample performance of the proposed procedure.

Star-shaped distributions.

A survey of recent results.

Session: **Multivariate Distributions and Copula 2**

Wolf-Dieter Richter *Institute of Mathematics Mathematics, University of Rostock, Germany*

Abstract: This talk gives a survey of author's and co-authors' recent results on star-shaped distributions and several important particular cases. Geometric representations of such distributions and stochastic representations of correspondingly distributed random vectors are discussed as well as several of their probabilistic and statistical applications.

1 Classes of star-shaped distributions

Flexible contour specifications of density level sets beyond normality are dealt with in Osiewalski and Steel (1993), Fernandez, Osiewalski and Steel (1995), Gupta and Song (1997), Arnold, Castillo and Sarabia (2007), Sarabia and Gómez-Déniz (2008), Balkema, Embrechts and Nolde (2010). The p-generalized elliptically contoured distributions are one of many possible generalizations of the elliptically contoured distributions. Star-shaped distributions can be considered as a further generalization. If such distribution has a density this allows the representation $\varphi_{g,K}(x) = C(g,K)g(h_K(x-\mu)), x \in R^n,$ where K is a contour defining star body with boundary S and having the origin as an interior point, $h_K(x) = \inf\{\lambda > 0 : x \in \lambda K\}, x \in R^n$ means its Minkowski functional, $\mu \in R^n$, $g \geq 0$ is an arbitrary density generating function that satisfies $0 < I(g) = \int_0^\infty r^{n-1}g(r)dr < \infty$, $C(g,K) = 1/[\mathcal{O}_S(S)I(g)]$, and $\mathcal{O}_S(S)$ denotes the S-generalized surface content measure, see R (2014a). Particular representations of $\mathcal{O}_S$ on norm and antinorm spheres as well as on polyhedra are given in R (2015a,b, 2016b,c) and R+Schicker (2014, 16a,b). If, for example, $K = \{x \in R^n : |x_1|^p + ... + |x_n|^p \leq 1\}$ is an (a,p)-generalized ellipsoid, $p > 0$, then $\mathcal{O}_S(A) = \int\limits_{\{(x_1,...,x_{n-1})^T : \exists \eta s.t.(x_1,...,x_{n-1},\eta)^T \in A\}} \frac{d(x_1,...,x_{n-1})}{(1-\sum_1^{n-1}|x_i/a_i|^p)^{1-1/p}}$, $A \in \mathcal{B}^n \cap S$.
Norm and antinorm contoured distributions R(2015a,b) build important subfamilies of the star-shaped distribution family. If densities exist, they are convex or radially concave contoured. For the latter notion, see Moszynska+R (2012). The family of polyhedral star-shaped distributions introduced in R+Schicker (2014,2016a,b) allows flexible (approximative) modeling of many types of sample clouds. Families of non-centrally elliptically contoured distributions are considered in R(2014a), Dietrich+R (2016) and Liebscher+R (2016) for dimension equal and greater or equal to two, respectively, and allow modeling a certain type of skewness.

2 Stochastic vector and geometric measure representations

A star-shaped distributed random vector X satisfies the stochastic representation

$$X \diamond RU, \quad P(U \in A) = \omega_S(A) = \frac{\mathcal{O}_S(A)}{\mathcal{O}_S(S)},\ A \subset S \cap \mathcal{B}^n$$

where $\diamond$ means 'distributed as', and the S-generalized uniform basis U is independent of the generalized radius R, $R = h_K(X)$.
The measure $\Phi_{g,K}$ corresponding to $\varphi_{g,K}$ allows the geometric representation

$$\Phi_{g,K}(B) = \frac{1}{I(g)} \int\limits_0^\infty r^{n-1}g(r)\omega_S([\frac{1}{r}A] \cap S)dr,\ B \in \mathcal{B}^n.$$

Disintegration formulas of this type were proved and applied for purposes of large deviation and general distribution theory in R (1985, 1986, 1990, 1991, 1995, 2007, 2009), R+Steinebach (1994), Breitung+R (1996) and R+Schumacher (2000).

3 Probabilistic and statistical applications

Assume that $X \sim \Phi_{g,K}$ and let $T : R^n \to R$ be any statistic. If $A(t) = \{x \in R^n : T(x) < t\}$ then $P(T(X) < t) = \Phi_{g,K}(A(t))$, $t \in R$, and the geometric measure representation applies for deriving exact distributions under non-standard model assumptions, see R (1995, 2012). Particular applications are •the derivation of exact noncentral distributions and a g-robustness study in Ittrich et al. (2000) •the construction of an exact test in nonlinear regression in Ittrich+R (2005) •the unified measure-of-cone representation of skewed elliptically contoured distributions in Günzel et al. (2012), Batun-Cútz et al. (2013), R+Venz (2014), Müller+R (submitted) •the geometric generalization of the von Mises distribution in R (2014a) and Dietrich+R (2016) •the geometric dependence modeling in Dietrich, Kalke, R (2013), where correlation is shown to combine a variance ratio and a rotation •the analytical dependence modeling in Müller +R (2016a,b) where uncorrelated variables are dependent due to the choice of a non-standard density generating function •the derivation and application of ▷ exact distributions of generalized $\chi^2-, t-$, and F-statistics in R (1995, 2007, 2009, 2012, 2015b, 2016a) ▷exact probabilities of correct selection in R (1994) and correct classification in Krause+R (1994, 2004) ▷exact distributions of sums, products and ratios as well as of order statistics under non-standard model assumptions in Kalke et al. (2013), Müller+R (2015) and in Müller+R (2016a,b, and submitted), respectively ▷representations of standard Gaussian random variables in R (2014b). Successful applications of the stochastic representation are • the construction of skewed $l_{n,p}$-symmetric distributions in Arellano-Valle+R (2012) • the simulation of star-shaped distributed random vectors (or simply their marginal variables) in Kalke+R (2013), R (2015b, 2016a) and R+Schicker (2016b, submitted).

Exact distributions of order statistics from continuous $l_{n,p}$-symmetric sample distributions

Session: **Multivariate Distributions and Copula 2**

Klaus Müller *Institute of Mathematics, University of Rostock, Germany*

Abstract: First, definitions, properties and instances of continuous $l_{n,p}$-symmetric and skewed $l_{n,p}$-symmetric distributions are resumed. Second, the exact distributions of order statistics from continuous $l_{n,p}$-symmetric sample distribution are derived yielding numerous representations in terms of mixtures of skewed distributions.

1 Two classes of multivariate probability distributions

In this first section, two comprehensive classes of multivariate probability distributions are presented. On the one hand, this is the class of continuous $l_{n,p}$-symmetric distributions, see i.a. [5] and [8], which contains the class of continuous spherical distributions in the special case $p = 2$. In general, these are multivariate dependent distributions whose densities may be characterized by their level set which are $l_{n,p}$-spheres with generalized radii and by a density generator essentially specifying the tail behavior and the behavior at the distribution center. Showing the great flexibility of this class of distributions, the density generators of $l_{n,p}$-symmetric Kotz type, $l_{n,p}$-symmetric Pearson Type VII and $l_{n,p}$-symmetric Pearson Type II distributions are visualized. Furthermore, some distributional properties such as invariance properties and marginal and conditional distributions are discussed. On the other hand, according to [3], the class of skewed $l_{n,p}$-symmetric distributions is presented and the cumulative distribution function of skewed $l_{k,p}$-symmetric distribution with dimensionality parameter m is represented as the continuous $l_{k+m,p}$-symmetric measure evaluated at a cone from a specific class of cones, see [7]. Such measure-of-cone representation, based on the geometric measure representations of $l_{n,p}$-symmetric distributions, see [8], generalize analogue results for skewed elliptically contoured distributions in [9].

2 Exact distributions of order statistics from continuous $l_{n,p}$-symmetric sample distribution

The main concern of this second section is the determination of exact distributions of order statistics from continuous $l_{n,p}$-symmetrically distributed populations. To this end, first, the special case of p-generalized Gaussian distribution is considered with the help of classical results on order statistics from independent and identically distributed sample, see e.g. [4]. Afterwards, two different approaches are used to determine the exact distributions of order statistics from continuous $l_{n,p}$-symmetric sample distribution generalizing the results on the distributions of order statistics from spherical sample distributions in [1, 2] and [6]. The first approach is also used in [6] for the case of elliptically contoured sample distribution and the second utilizes the measure-of-cone representations of skewed $l_{n,p}$-symmetric distributions mentioned in Section 1. Each of these approaches yield numerous representations of the distribution considered here as mixtures of skewed $l_{k,p}$-symmetric distributions and the results following from the first approach are structurally different to that following from the second one. Subsequently, these structural differences are analyzed and the similarities with representations of distribution of order statistics from independent and identically distributed sample distributions are outlined. Finally, the densities of order statistics from continuous $l_{n,p}$-symmetric distributions are illustrated in some special cases also considered in Section 1.

References

[1] Arellano-Valle, R. B. and Genton, M. G. (2007). On the exact distribution of linear

combinations of order statistics from dependent random variables. *J. Multivariate Anal.*, 98(10):1876–1894.

[2] Arellano-Valle, R. B. and Genton, M. G. (2008). Corrigendum to "On the exact distribution of linear combinations of order statistics from dependent random variables": [J. Multivariate Anal. 98 (2007) 1876-1894]. *J. Multivariate Anal.*, 99(5):1013.

[3] Arellano-Valle, R. B. and Richter, W.-D. (2012). On skewed continuous $l_{n,p}$-symmetric distributions. *Chil. J. Stat.*, 3(2):193–212.

[4] David, H. A. and Nagaraja, H. N. (2003). *Order statistics.* Wiley, New York, 3rd edition.

[5] Gupta, A. K. and Song, D. (1997). l_p-norm spherical distributions. *J. Stat. Plann. Inference*, 60(2):241–260.

[6] Jamalizadeh, A. and Balakrishnan, N. (2010). Distributions of order statistics and linear combinations of order statistics from an elliptical distribution as mixtures of unified skew-elliptical distributions. *J. Multivar. Anal.*, 101(6):1412–1427.

[7] Müller, K. and Richter, W.-D. (2016). Extreme value distributions for dependent jointly $l_{n,p}$-symmetrically distributed random variables. *Depend. Model.*, 4:30–62.

[8] Richter, W.-D. (2009). Continuous $l_{n,p}$-symmetric distributions. *Lith. Math. J.*, 49(1):93–108.

[9] Richter, W.-D. and Venz, J. (2014). Geometric representations of multivariate skewed elliptically contoured distributions. *Chil. J. Stat.*, 5(2):71–90.

Modeling and statistical inference of copulas based on Frank's family

Session: **Multivariate Distributions and Copula 2**

Eckhard Liebscher *Department of Engineering and Natural Sciences, University of Applied Sciences Merseburg, Germany*

In the last decade a lot of interesting copula models were developed. In the first part of the talk we discuss copula models based on Frank's copula family. We will see that Frank's family fulfills an appealing property with respect to extensions of the model class. Let $\lambda : [0,1] \to [0,1]$ be a continuous 1-1 mapping with $\lambda(0) = 0, \lambda(1) = 1$. We consider the copula

$$C(u) = \psi\left(\varphi(u_1) + \ldots + \varphi(u_d)\right),$$

where $\psi = \varphi^{-1}$ and

$$\varphi(t) = -\ln\left(\frac{e^{-\theta\lambda(t)} - 1}{e^{-\theta} - 1}\right).$$

In the case $\lambda(t) = t$, one obtains the classical Frank copula. We provide a theorem providing sufficient conditions on λ to obtain an admissible copula model.
The proposed family has the following advantageous properties:
- C is an Archimedean copula,
- an algorithm for random number generation is available,
- asymmetric versions of copulas can be derived, by reflection method for example.

To expand the flexibility of the model, we deal with finite mixtures of copulas of the model class and transformed versions. The copula densities of the class show various shapes, see figures in the talk. In applications, it remains to find a suitable model for a given dataset.
Next some preliminaries concerning the parameter estimation. Let us denote the copula of the sample by C. This copula is modelled by a parametric family $\mathcal{M} = \{C_\theta\}_{\theta\in\Theta}$ of copulas. $\Theta \subset \mathbb{R}^q$ is the parameter space. We consider the *Cramér-von-Mises divergence* as a measure of discrepancy between the copula C and the model class $\mathcal{M}$:

$$\mathcal{D}(C, \mathcal{M}) = \inf_{\theta\in\Theta} \int_{\mathbb{R}^d} \left(H(x) - C_\theta(F(x))\right)^2 \, \mathrm{d}H(x).$$

The CvM divergence is estimated by

$$\widehat{\mathcal{D}}_n(C_\theta) = \frac{1}{n}\sum_{i=1}^{n} \left(\hat{H}_n(X_i) - C_\theta(\bar{F}_n(X_i))\right)^2.$$

The approximate minimum distance estimator $\hat{\theta}_n$ is an estimator satisfying

$$\widehat{\mathcal{D}}_n(C_{\hat{\theta}_n}) \le \min_{\theta\in\Theta} \widehat{\mathcal{D}}_n(C_\theta) + \varepsilon_n$$

with a sequence $\{\varepsilon_n\}$ of random numbers with $\varepsilon_n \to 0$ *a.s.* Under certain regularity conditions, one can prove that $\hat{\theta}_n$ is a consistent estimator for

$$\theta_0 = \arg\min_{\theta\in\Theta} \mathcal{D}(C, C_\theta)$$

and it is asymptotically normally distributed, see [1], [2]. Since it is appropriate for most situations in applications, we assume that $C \notin \mathcal{M}$. It should be pointed out that the best parameter θ_0 depends heavily on the approximation measure. There is no "true parameter" at

all. As a key result, we obtain asymptotic normality of $\widehat{\mathcal{D}}_n(C_{\hat{\theta}_n})$ with a certain variance which can be estimated in a suitable way.
The second part of the talk is devoted to the model selection in the above described framework. One idea for comparisons of various models is to look at the estimated measure for goodness-of-approximation:

$$\hat{\rho} = 1 - \frac{\widehat{\mathcal{D}}_n(C_{\hat{\theta}_n})}{\widehat{\mathcal{D}}_n(\Pi)}$$

where Π is the independence copula. The quantity $\hat{\rho}$ describes the quality of approximation. The case $\hat{\rho} = 1$ corresponds to perfect approximation.
Further we discuss comparisons of non-nested models. For this purpose, we develop a test which should provide a decision to the problem: what model is the best of two models $\mathcal{M}_1$ and $\mathcal{M}_2$ or are the two models equivalent in approximation quality. The test hypothesis is

$$H_0 : \mathcal{D}(C, \mathcal{M}_1) - \mathcal{D}(C, \mathcal{M}_2) \leq M, \quad H_1 : \mathcal{D}(C, \mathcal{M}_1) - \mathcal{D}(C, \mathcal{M}_2) > M,$$

where $M \geq 0$ is a given number. Finally, the situation of nested models is considered and the use of the methods is illustrated by examples.

References

[1] Liebscher, E. (2009). Semiparametric Estimation of the Parameters of Multivariate Copulas. *Kybernetika* 45, 972-991.

[2] Liebscher, E. (2015). Goodness-of-Approximation of Copulas by a Parametric Family. *in A. Steland, E. Rafajłowicz, K. Szajowski (Editors): Stochastic Models, Statistics and Their Applications.* New York, NY [u.a.] : Springer, 101-109.

Estimation in the Probabilistic Index Model

Session: **Nonparametric Methods 3**

Olivier Thas *Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Belgium, and National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Australia*

Karel Vermeulen *Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Belgium*

Gustavo Amorim *Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Belgium*

Jan De Neve *Department of Data Analysis, Ghent University, Belgium*

Stijn Vansteelandt *Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium*

Abstract: After a brief introduction to the Probabilistic Index Model (PIM) we focus on several improvements of the estimation procedures. In particular, we discuss semiparametric efficient estimation and a few methods that give improved inference in small samples.

1 Introduction

The Probabilistic Index Model (PIM) was introduced by [5]. It is a flexible class of semiparametric models that can be used to generate many classical rank tests, such as the Wilcoxon-Mann-Whitney, Kruskal-Wallis and Friedman tests, among others; see [2].
The PIM models the conditional probability $P(Y \leq Y'|\mathbf{X}, \mathbf{X}^*)$, where Y and Y' are independent random outcomes associated with covariates $\mathbf{X}$ and $\mathbf{X}^*$, respectively. With g a link function, a PIM is often of the form $g(P(Y \leq Y'|\mathbf{X}, \mathbf{X}^*)) = \beta_0 + \beta^t(\mathbf{X}^* - \mathbf{X})$. Thas *et al.* [5] proposed an estimator for β which is consistent and asymptotically normal under mild conditions. However, no semiparametric efficiency was proven and their simulation results indicate that the convergence to the asymptotic normal distributions is too slow for the method to be recommended for use in small samples.
In this presentation we will present some recent results on improved estimation methods. First, we propose semiparametric efficient estimators, and next we show how empirical likelihood methods can be employed to result in well-behaved inference in small samples.

2 Semiparametric Efficient Estimation

We derive the class of all consistent and asymptotically normal estimators for β in the semiparametric PIM by appealing to the theory of semiparametrics [6] and identify the efficient influence function for β. Next, we propose estimating equations to solve the efficient influence function relying on the theory of semiparametric two- step estimators. The efficient estimator is computationally more demanding than the original estimator. In addition, we work out estimators that statistically improve the latter while retaining their computationally attractive properties. These improved estimators are chosen in such a way that they decrease the second-order finite-sample bias as compared to the original estimator of Thas *et al.* [5].
We conclude that the semiparametric efficient estimator, which is computationally more intensive than the original estimator, does not result in much larger efficiencies in small to moderately large datasets. The biased reduced estimator seems to be a good compromise.

3 Improved Inference in Small Samples

We explore methods that are designed to give better small sample results. Resampling techniques, such as the bootstrap and jackknife, are often used as alternative approaches to increase accuracy in many statistical applications. However, they sometimes require strong computational power. We solve this issue by applying the bootstrapping U-statistics method of Jiang

and Kalbfleisch [3]. In addition to bootstrap, we also use methods based on empirical likelihood to improve small sample inference for probabilistic index models. In particular, we adapt the empirical likelihood methods of Jing *et al.* [4] and Chen *et al.* [1].
Our simulation results demonstrate that our empirical likelihood methods work well for samples with sizes as small as 20.

References

[1] Chen, J., Variyath, A. M., and Abraham, B. (2008). Adjusted empirical likelihood and its properties. *Journal of Computational and Graphical Statistics*, 17(2), 426-443.

[2] De Neve, J. and Thas, O., 2015. A regression framework for rank tests based on the probabilistic index model. *Journal of the American Statistical Association*, 110(511), 1276-128.

[3] Jiang, W., and Kalbfleisch, J. D. (2012). Bootstrapping U-statistics: applications in least squares and robust regression. *Sankhya B*, 74(1), 56-76.

[4] Jing, B.-Y., Yuan, J., and Zhou, W., 2009. Jackknife empirical likelihood. *Journal of the American Statistical Association*, 104(487), 1224-1232.

[5] Thas, O., Neve, J. D., Clement, L., and Ottoy, J.-P., 2012. Probabilistic index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4), 623–671.

[6] Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

Rank Repeated Measures Analysis of Covariance

Session: **Nonparametric Methods 3**

Chunpeng Fan *Sanofi US Inc. Bridgewater NJ, U.S.A.*
Donghui Zhang *Sanofi US Inc. Bridgewater NJ, U.S.A.*

Abstract: When analyzing a response variable at the presence of both factors and covariates, with potentially correlated responses and violated assumptions of the normal residual or the linear relationship between the response and the covariates, rank-based tests can be an option for inferential procedures instead of the parametric repeated measures analysis of covariance (ANCOVA) models. This article derives a rank-based method for multi-way ANCOVA models with correlated responses. The generalized estimating equations (GEE) technique is employed to construct the proposed rank tests. Asymptotic properties of the proposed tests are derived. Simulation studies confirmed the performance of the proposed tests.

1 Introduction

In many scientific fields, researchers frequently face the problem of analyzing a response variable at the presence of both factors and covariates. When the linearity and normality assumptions required by the parametric ANCOVA model is violated, the nonparametric rank-based method may be an appropriate alternative to their parametric counterpart. When the central goal is to assess the effects of the factors after adjustment for the covariates, [1] proposed the rank-based method for ANCOVA models in the framework of one-way completely randomized designs with independent observations. However, inferences for the rank-based ANCOVA models which allow multi-way factors and correlated observations seem not been done yet.

In this current work, we incorporate the rank transform statistic in the rank repeated measures ANCOVA model into the GEE framework and derive the asymptotic properties of the Wald-type rank test and the ANOVA-type rank test for testing contrast effects of the factors. The proposed GEE-based tests can be used in any ANCOVA model where GEE applies, which includes the multi-way ANCOVA model with multiple covariates and correlated observations. As the proposed tests retain the ease of being understood, it is also easy to be implemented using common statistical softwares, such as SAS® and R®.

2 The method

Suppose we have observations $\{(\boldsymbol{Y}_n, \boldsymbol{Z}_n) : n = 1, \cdots, N\}$ from N independent subjects with $J_n \times 1$ response vector $\boldsymbol{Y}_n = (Y_{n1}, \ldots, Y_{nJ_n})'$ and $1 \times S$ covariate vector $\boldsymbol{Z}_n = (Z_{1n}, \ldots, Z_{Sn})$ being a random vector correlated with $\boldsymbol{Y}_n$.

We assume that Y_{nj} comes from one of the C cumulative distribution functions (CDFs) $F_1(y), \ldots, F_C(y)$. Define the index $\{\boldsymbol{c}_n = (c_{n1}, \ldots, c_{nJ_n})' : n = 1, \ldots, N, Y_{nj} \in F_{c_{nj}}(y)\}$. In the vector form, we denote $\boldsymbol{F}(y) = \{F_1(y), \ldots, F_C(y)\}'$ and $\boldsymbol{F}_{\boldsymbol{c}_n}(y) = \{F_{c_{n1}}(y), \ldots, F_{c_{nJ_n}}(y)\}'$. The relationship between $\boldsymbol{F}(y)$ and $\boldsymbol{F}_{\boldsymbol{c}_n}(y)$ can be characterized by a $J_n \times C$ design matrix $\boldsymbol{X}_n$. For each subject n, $\boldsymbol{X}_n$ can be defined by letting the (j, c)th element of $\boldsymbol{X}_n$ being 1 if Y_{nj} comes from F_c and 0 otherwise. It can be seen $\boldsymbol{F}_{\boldsymbol{c}_n}(y) = \boldsymbol{X}_n \boldsymbol{F}(y)$.

The present work deals with the problem of simultaneously testing ℓ linear combinations of the elements of $\boldsymbol{F}$, $H_0 : \boldsymbol{L}'\boldsymbol{F} = 0$, for $\boldsymbol{L}$ a $C \times \ell$ fixed matrix. Also define quantity $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_C)'$ to characterize the covariate-adjusted nonparametric effects of the factors $\boldsymbol{F}$. For $c = 1, \ldots, C$, define $\beta_c = \int \bar{H}(y) dF_c(y)$, for $\bar{H}(y) = \sum_{n=1}^{N} \sum_{j=1}^{J_n} H_{c_{nj}}(y)/N$ and $H_c(y) = \{F_c(y) + F_c(y-)\}/2$. Denote the rank of Y_{nj} in the pooled sample to be R_{nj} and the rank of the covariate Z_{sn} in $Z_{s1}, \ldots, Z_{sN}$ to be T_{sn}. Denote the $J_n \times S$ design matrix for the covariates $\boldsymbol{X}_n^* = [\{T_{1n} - (N + 1)/2\}1_{J_n}, \ldots, \{T_{Sn} - (N + 1)/2\}1_{J_n}]$. With the covariate effect $\boldsymbol{\lambda}$, denoting $\boldsymbol{S}_n = [\boldsymbol{X}_n, \boldsymbol{X}_n^*/N]$,

by minimizing $\sum_{n=1}^{N} \boldsymbol{e}_n' \boldsymbol{e}_n$ with $\boldsymbol{e}_n = (\boldsymbol{R}_n - \boldsymbol{X}_n^* \boldsymbol{\lambda})/N - \boldsymbol{X}_n \boldsymbol{\beta}$, the GEE can be written as

$$\boldsymbol{U}_N(\boldsymbol{\theta}) = \sum_{n=1}^{N} \boldsymbol{S}_n' \Big(\frac{\boldsymbol{R}_n - \boldsymbol{X}_n^* \boldsymbol{\lambda}}{N} - \boldsymbol{X}_n \boldsymbol{\beta} \Big).$$

It is rigorously proved that under H_0, under certain regularity conditions, the GEE estimator $\hat{\boldsymbol{\beta}}$ has the following asymptotic property:

$$\sqrt{N} \boldsymbol{L}' (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d N(0, \boldsymbol{L}' \boldsymbol{\Sigma} \boldsymbol{L}),$$

where $\boldsymbol{\Sigma}$ has consistent estimator $\widehat{\boldsymbol{\Sigma}}_N^1 = N \boldsymbol{M}_N^{-1} \Big\{ \sum_{n=1}^{N} \boldsymbol{X}_n' (\boldsymbol{R}_{n,\hat{\boldsymbol{\lambda}}}/N - \boldsymbol{X}_n \hat{\boldsymbol{\beta}})(\boldsymbol{R}_{n,\hat{\boldsymbol{\lambda}}}/N - \boldsymbol{X}_n \hat{\boldsymbol{\beta}})' \boldsymbol{X}_n \Big\} \boldsymbol{M}_N^{-1}$, $\boldsymbol{R}_{n,\hat{\boldsymbol{\lambda}}} = \boldsymbol{R}_n - \boldsymbol{X}_n^* \boldsymbol{\lambda}$, and $\boldsymbol{M}_N = \sum_{n=1}^{N} \boldsymbol{X}_n' \boldsymbol{X}_n$.
The Wald-type test χ_1 and the ANOVA type test Q_1 can then be derived.
More details about the derivation process can be found in [2].

3 Example

The method was used to analyze the seizure count data. For the null hypothesis of no treatment effect, the p-values for the ANCOVA Wald-type and ANOVA-type tests are both 0.0174; as a comparison, the ANOVA (without covariate) Wald-type and ANOVA-type tests both give a p-value of 0.2974.

References

[1] Bathke, A., Brunner, E. (2003). A nonparametric alternative to analysis of covariance. In Akritas,M.G.,Politis, D.N. eds. *Recent Advances and Trends in Nonparametric Statistics* (pp. 109–120), Amsterdam: Elsevier.

[2] Fan, C., Zhang, D. (2017). Rank repeated measures analysis of covariance. *Communications in Statistics - Theory and Methods,* 46, 1158-1183.

Full-information Best Choice Problem with Unknown Change Point in Value Distribution of Options

Session: **Optimal Decision in Changepoint Models**

Aiko Kurushima *Department of Economics, Sophia University, Japan*

Abstract: We discuss a generalization of the full-information best choice problem with change point. We assume that the value of the options are distributed by two distributions. The distribution would be changed from one to another at some unknown moment in the selection process. Our objentive is to find the optimal strategy to maximiza the probability of accepting the best option.

1 Introduction

We discuss a generalization of the full-information best choice problem with the case that the distribution of the option values changes in the process of selection, which is called a change point or a disorder. Best choice problem is originally introduced and solved by Gilbert and Mosteller (1966), and best choice problem with disorder have some generalizations, such as Yoshida (1984) which dealt with the full-information case, Mazalov and Ivashko (2012) which considered the problem with imperfect observations and the optimal stopping rule in the class of Bayes' strategies. These papers discussed the problem with unknown change point.
Like the problems above, we consider the problem with unknown moment of disorder. We assume that the time θ when a disorder arises and the total number n of options are known. Here θ means the number of options which arrive before the change, and the distribution of θ is assumed to be a uniform between 1 and n. Furthermore, we assume that the distribution of value changes as follows: if $i = 1, 2, \ldots, \theta$, then it follows a uniform distribution between $(0, 1)$, $U(0, 1)$, otherwise, that is, if $i = \theta + 1, \ldots, n$, then it follows the other uniform distribution $U(0, b)$, $0 < b < 1$. To maximize the probability of accepting the best option is the objective of the problem.

2 OLA stopping rule and its optimality

Based on the assumptions stated in the previous section, we first consider the OLA (one-stage look-ahead) stopping rule which is shown to be optimal in monotone case [3].
In the selection process, the state of each moment is devided into two cases. The first case is when the maximum value of the options arrived so far is greater than or equal to b, which we could know the disorder has not arised yet. And the second case is when the maximum value is less than b, that is, we have no information about the current state, we do not know whether it is before or after the change point.
For the first case, we could reformulate the state which is the same as the best choice problem with random horizon solved in [2]. So we can easily shown the OLA stopping rule is optimal. For the second case, we do not have any information whether the disorder arises or not. To find the OLA stopping rule, we compare the expected value of the probability of accepting the current option with the on of accepting the next first option which has the maximum so far. To show the optimality of the OLA stopping rule, we show the monotonicity of the problem. The result of the problem is detailed in the presentation.

References

[1] Mazalov, V. and Ivashko, E. (2012). "Bayes' model of the best-choice problem with disorder," *Int. J. of Stoch. Analysis*, **2012**, Article ID 697458, 8 pages.

[2] Presman, E. L. and Sonin, I. M. (1972). "The best choice problem for a random number of objects," *Theory Prob. Appl.*, **17**, 657-668.

[3] Ross, S. M. (1970). *Applied Probability Models and Optimization Applications,* Holden-Day, San Francisco.

[4] Yoshida, M. (1984). "Probability maximizing approach to a secretary problem with random change-point of the distribution of the observed process," *J. Appl. Prob.*, **21**, 98–107.

Bayesian game on disordered process

Session: **Optimal Decision in Changepoint Models**

Krzysztof Szajowski *Faculty of Pure and Applied Mathematics, Wrocław University of Technology, Poland*

Abstract: The topic of this talk is devoted to games with two players who observe a sequence of objects. Before some random moment θ the objects value are represented by independent identically distributed random variables $\xi_1, \xi_2, \ldots, \xi_{\theta-1}$ with common distribution function $\mu(dx)$. After the moment θ, they are independent random variables $\xi_\theta, \xi_{\theta+1}, \ldots$ having another common distribution function $f(x)\mu(dx)$. The players information about θ can be constructed only by successively observed values of the ξ's. Both players are interested in choosing the largest of all $\theta + m - 1$ quantities for a given integer observation which maximize the probability that the quantity associated with the stopping time is the largest of all $\theta + m - 1$ quantities for a given integer m (cf. [Yoshida(1984)], [Mazalov & Ivashko(2007)]). The aim is to construct a Nash equilibrium in the game where the players strategies are stopping times. The various additional constrains concerning the knowledge of the players and their aims lead to unequal strategical knowledge of the players. The investigation extends former results by [Ferguson(2005)], [Mazalov(1996)] and [Szajowski(2009)].

References

[Ferguson(2005)] Thomas S. Ferguson. Selection by committee. In A.S. Nowak and K. Szajowski, editors, *Advances in dynamic games*, volume 7 of *Ann. Internat. Soc. Dynam. Games*, pages 203–209. Birkhäuser Boston, Boston, MA, 2005.

[Mazalov(1996)] Vladimir V. Mazalov. A game related to optimal stopping of two sequences of independent random variables having different distributions. *Math. Japon.*, 43(1):121–128, 1996.

[Mazalov & Ivashko(2007)] V.V. Mazalov and E.E. Ivashko. The best choice problem with complete information and disordering. *Obozr. Prikl. Prom. Mat.*, 14(2):215–224, 2007.

[Szajowski(2009)] K. Szajowski. Comparison among some optimal policies in rank-based selection problems. In Leon A. Petrosjan and Nikolay A. Zenkevich, editors, *Contribution to Game Theory and Management*, volume III of *The Third International Conference Game Theory and Management June 24-26, 2009, St. Petersburg, Russia*, pages 409–420. Graduate School of Management, St. Petersburg University, St. Petersburg,Russia.

[Yoshida(1984)] M. Yoshida. Probability maximizing approach to a secretary problem with random change-point of the distribution law of the observed process. *J. Appl. Probab.*, 21:98–107, 1984.

A Quality Control Chart design based on optimal stopping rules

Session: **Optimal Decision in Changepoint Models**

Marek Skarupski *Faculty of Pure and Applied Mathematics, Wrocław University of Science and Technology, Poland*

Abstract: In this talk we consider the Shewhart quality control charts (QCC) and propose a new design in case of external failure. In classical Statistical Process Control (SPC) it is assumed that the process will be out of control because of the internal problems. In this presentation we propose the method in case when the failure occurs because of the external event (eg. the failure of the machine used in manufacturing of the process) and we propose the method how to recognize this situation basis on the observation of the process. We introduce in the design of the QCC new control lines: Optimal Control Lines, and Optimal Warning Lines which are helpful in quality control process.

References

[1] Shewhart, W. A. (1932). *Economic Control of Quality of Manufactured Product*, D. Van Nostrand Company, Inc.

[2] Montgomery, D. C. (2005). *Introduction to Statistical Quality Control*, 6th ed., Wiley: New York.

[3] Porosiński, Z. and Skarupski, M. and Szajowski, K. (2016). *Duration problem: basic concept and some extensions*, Mathematica Applicanda, Vol. 44, No. 1 (2016), p. 87-112.

[4] Skarupski, M. (2016) *On a duration problem with unbounded geometrical horizon*, Preprint, Wrocław University of Technology

Registration to Low-Dimensional Function Spaces

Session: **Functional Data Analysis**

Alois Kneip *Department of Economics, University of Bonn, Mathematics, Germany*
Heiko Wagner *Department of Economics, University of Bonn, Mathematics, Germany*

Abstract: Registration aims to decompose amplitude and phase variation of samples of curves. Phase variation is captured by warping functions which monotonically transform the domains. Resulting registered curves should then only exhibit amplitude variation. Most existing registration method rely on aligning typical shape features like peaks or valleys to be found in each sample function. It is shown that this is not necessarily an optimal strategy for subsequent statistical data exploration and inference. In this context a major goal is to identify low dimensional linear subspaces of functions that are able to provide accurate approximations of the observed functional data. In this paper we present a registration method where warping functions are defined in such a way that the resulting registered curves span a low dimensional linear function space. Problems of identifiability are discussed in detail, and connections to established registration procedures are analyzed. The method is applied to real and simulated data.

General Eigenexpansions

Session: **Functional Data Analysis**

Moritz Jirak *TU Braunschweig, Germany*
Martin Wahl *Humboldt-Universität zu Berlin, Germany*

Abstract: For PCA the (long-run) covariance operator and its spectral decomposition is of fundamental interest. Based on a fairly general method, new results on the corresponding empirical eigenfunctions and values are presented. Despite its generality, the results are essentially optimal (up to logarithmic factors) in terms of moment, spectral gap and dependency assumptions. In particular, we are also able to handle heavy tailed cases where less than four moments are available. The method allows to handle both high-dimensional or functional data, that is, the trace of the operator may diverge as the dimension and/or sample size increases. Among other things, this includes a number of popular high-dimensional factor models, but also time series in general infinite dimensional Hilbert spaces.

Nonparametric density estimation for intentionally corrupted functional data

Session: **Functional Data Analysis**

Alexander Meister *University of Rostock, Germany*

Abstract: We consider statistical models in which functional data are artificially contaminated by independent Wiener processes in order to satisfy privacy constraints. We show that the corrupted observations have a Wiener density, which determines distribution of the original functional data uniquely. We construct a nonparametric estimator for the functional density and study its asymptotic properties. We discuss applications in the fields of classification and goodness-of-fit testing. This talk is based on a joint work with A. Delaigle (University of Melbourne).

Clustering electricity customers

Session: **Energy Statistics**

Tamsin Lee *Mathematical Institute, University of Oxford, UK*
Stephen Haben *Mathematical Institute, University of Oxford, UK*

Abstract: Predicting electricity behaviour on the low voltage network is of growing importance as our electricity consumption evolves. Smart meters provide distribution network operators with the required insight. However smart meter data in the UK is a propitiatory, and thus it is preferable to generalise the information gained by smart meters to properties without smart meters. We have clustered domestic customers and small-to-medium enterprises. This presentation will discuss our methods, and show that for small-to-medium enterprises, we can recreate an approximate smart meter profile (the electricity load every 30 minutes).

Short Term Load Forecasts of Low Voltage Level Networks

Session: **Energy Statistics**

Stephen Haben *Mathematical Institute, University of Oxford, U.K.*
Siddharth Arora *Mathematical Institute, University of Oxford, U.K.*
Tamsin Lee *Mathematical Institute, University of Oxford, U.K.*
Georgios Giasemidis *Dept. Mathematics and Statistics, CountingLab Ltd., Reading, U.K.*

Abstract: Low Voltage networks are increasingly at risk with the expected extra uptake of low carbon technologies. There are a number of potential solutions which can help to alleviate the strain on the security of supply of such networks such as battery storage devices, demand response etc. However to optimally implement such strategies will require accurate estimates of the potential demand on the network.
In this talk I discuss a number of methodologies for short term forecasts of low voltage networks. I also consider the role of weather forecasts and their inclusion in the model. I present both point and probabilistic methods. In addition, low voltage networks are quite variable due to the different types and numbers of customers which are connected. Hence we also investigate the potential change in accuracy with the different size of networks.

Modeling Electricity Load with Inhomogeneous Markov Switching Models

Session: **Energy Statistics**

Kevin Berk *Department of Mathematics, University of Siegen, Germany*
Alfred Müller *Department of Mathematics, University of Siegen, Germany*

Abstract: Probabilistic modeling of electricity load is a rising topic in energy markets, since the uncertainty of customer demand and its relation to the electricity price is a major risk factor. Therefore, accurate load forecasts are fundamental for many planning and decision-making processes of the market participants. There exists a wide diversity of load profiles of industrial companies, each of them having its own structure and special characteristics. We focus particularly on such profiles that exhibit two regimes, i.e. a production and a standby regime, for which the switches between them appear stochastically. Based on Hartigan's dip test for unimodality, we develop an automated procedure to classify regime-switching load profiles amongst others. The model we propose is a two regime process, the transitions are described through an underlying Markov chain with time-varying transition probabilities. The demand during the production regime is modeled by a seasonal pattern combined with an AR process, whereas we assume the standby demand to be i.i.d. random distributed. The parameter estimation via ML is implemented with a Differential Evolution algorithm. We use the proper scoring rule CRPS to score the goodness-of-fit of probabilistic forecasts to real observations out-of-sample, through which we compare our model to some benchmark models using a database of four customers from different industry sectors. The results show that classical additive time series approaches are barely appropriable for regime-switching load profiles and that the Markov switching model performs very well. Additionally, the scores show that our model with time-varying transition probabilities is preferable to homogeneous ones, while the improvement in forecast accuracy is significant in the sense of the Diebold-Mariano test.

Load Forecasting Using Lasso Based Time Series Methods

Session: **Energy Statistics**

Florian Ziel *European University Viadrina, Frankfurt (Oder), Germany*

Abstract: A new methodology for probabilistic electric load forecasting using high-dimensional time series techniques is presented. The approach is based on a bivariate autoregressive process for electric load and temperature. The heart of the methodology is the considered lasso (least absolute shrinkage and selection operator) estimation method. It allows to estimate models with very large parametrizations without having over-fitting concerns. The large parametrization allows to capture several stylized facts like annual, weekly and daily seasonalities, non-linear effects (esp. from temperature on load) and public holidays effects. Applications to GEFCom2014 data are shown [1].

References

[1] Ziel, F., & Liu, B. (2016). Lasso estimation for GEFCom2014 probabilistic electric load forecasting. *International Journal of Forecasting*, 32(3), 1029-1037.

On a LASSO-type estimator in errors-in-variables models

Session: **Errors-in-Variables Models**

Silvelyn Zwanzig *Department of Mathematics, Uppsala University, Sweden*
M. Rauf Ahmad *Department of Statistics, Uppsala University, Sweden*

Abstract: We introduce a LASSO-type estimator as a generalization of the classical LASSO estimator. The generalization relies on a SVD of the observed design matrix. We derive expressions for the risk of this LASSO-type estimators in case of high dimensional error-in-variables models and compare it with the ridge estimator and the TLS estimator.

On the risk of LASSO-type estimators in errors-in-variables models

Session: **Errors-in-Variables Models**

M. Rauf Ahmad *Department of Statistics, Uppsala University, Sweden*
Silvelyn Zwanzig *Department of Mathematics, Uppsala University, Sweden*

Abstract: Based on SvD of the design matrix, [1] introduce a LASSO-type estimator as a generalization of the classical LASSO estimator, thus named SvD-LASSO, and evaluate its risk performance for high-dimensional linear models and compare it with the corresponding risk of ridge estimator. The SvD-LASSO distinguishes itself by resting on no stringent assumptions or conditions, particularly sparsity. The present work pertains to an extension of the SvD-LASSO for errors-in-variables models. The risk of the extended estimator is studied and compared with that of the corresponding ridge and total least-squares estimators.

References

[1] Zwanzig, S., Ahmad, M. R. (2017). On the behavior of the risk of a LASSO- type estimator, In: *Analytical Methods in Statistics*, Antoch, J., Jurečková, J., Maciak, M., Pesta, M., eds., Springer-Verlag. To appear.

Towards a flexible statistical modelling by latent factors for evaluation of simulated climate forcing effects

Session: **Errors-in-Variables Models**

Katarina Fetisova *PhD student, Department of Mathematics, Division of Statistics, Stockholm University, Sweden*

Abstract: Evaluation of climate models is a key issue within climate research. Based on the statistical framework proposed by [4], we suggest several latent factor models of different complexity for evaluating climate model simulations against climate proxy data from the last millennium. To evaluate the performance of the factor models, a pseudo-proxy experiment is employed, in which the true unobservable temperature series is replaced by selected realizations of a climate simulation model. In the current analysis, the focus is placed on the factor model developed for evaluation of climate models driven by five (reconstructed) external climate forcings. The pseudo-proxy experiment is in progress, therefore conclusions will come later.

1 Introduction

The main components of the statistical framework [4] are:

- x_{ft} - a simulated temperature generated by a climate model driven by a particular forcing f for the region and time period of interest, $t = 1, 2, \ldots, n$.
- τ_t - a true unobservable temperature corresponding to x_{ft}.
- y_t - a measured temperature, intended to represent τ_t.
- z_t - a reconstructed temperature, derived from climate proxy data.

`Forcings`: *Solar*, *Greenhouse gases*, *Orbital*, *Land use*, and *Volcanic* (see [2]).

2 The method: Confirmatory Factor Analysis

For a detailed description of confirmatory factor analysis (CFA) see, for example, [3]. In short, CFA assumes that the researcher has hypotheses about which factors are to be involved and which restrictions on the parameter space it implies. This leads to factor models with a structure specified in advance. The main question becomes: How well do the empirical data conform to the hypothesized factor model? In other words, CFA can be viewed as a technique for theory testing. **Formulating factor models (τ is available)**

The main idea is that x_f and τ contain a common latent factor, say ξ_f.

• **The simulated forcing *f* is a single forcing**

The 2-indicator 1-factor model, abbrv. FA(2,1), is formulated:

$$\begin{cases} x_{ft} = a_{11} \cdot \xi_{ft} + \delta_{ft} \\ \tau_t = a_{21} \cdot \xi_{ft} + \tilde{\nu}_{1t} \end{cases} \tag{1}$$

Assumptions: $E[\xi_{\mathsf{f}t}] = 0$, $\mathrm{Var}(\xi_{\mathsf{f}t}) = \sigma^2_{\xi_{\mathsf{f}}} = 1$, $(\delta_{ft}, \tilde{\nu}_{1t})' \sim N\left(0, diag(\sigma^2_{\delta_f}, \sigma^2_{\tilde{\nu}_1})\right)$, where $\sigma^2_{\delta_f}$ is known a priori to achieve the model (just-)identifiability.

• **The simulated forcing *f* is a combination of two forcings, f1 and f2**

The 4-indicator 3-factor model, FA(4,3), is formulated:

$$\begin{cases} x_{f1\,t} &= a_{31}\cdot\xi_{f1\,t} &+ & 0\cdot\xi_{f2\,t} &+ & 0\cdot\xi_{f1f2_{interact}t} &+\delta_{f1\,t} \\ x_{f2\,t} &= 0\cdot\xi_{f1\,t} &+ & a_{32}\cdot\xi_{f2\,t} &+ & 0\cdot\xi_{f1f2_{interact}t} &+\delta_{f2\,t} \\ x_{f\,t} &= a_{31}\cdot\xi_{f1\,t} &+ & a_{32}\cdot\xi_{f2\,t} &+ & a_{33}\cdot\xi_{f1f2_{interact}t} &+\delta_{f\,t} \\ \tau_t &= a_{41}\cdot\xi_{f1\,t} &+ & a_{42}\cdot\xi_{f2\,t} &+ & a_{43}\cdot\xi_{f1f2_{interact}t} &+\tilde{\nu}_{2\,t}. \end{cases} \tag{2}$$

Assumptions: similar to those in Case 1. In addition, the latent factor are correlated. Various hypotheses can be tested. For example:

- the influence of ξ_{fi}, $i = 1, 2$, is not significant;
- the influence of the interaction effect is not significant;
- the latent factors are uncorrelated.

Under each hypothesis, the model becomes overidentified, thus making it possible to assess the fit of the model to the data. It can be done both statistically and heuristically (will be discussed later).

3 Pseudo-proxy experiment

The FA(7,6)-model, developed for evaluation of five-forcing climate model simulations, is studied in a pseudo-proxy experiment. The whole analysis is performed in the *R* package `sem` (see [1]). The results will come soon.

References

[1] Fox, J.: *Structural equation modeling with the* sem package in R, Structural Equation Modeling, 13(3), pp. 465-486, 2006.

[2] Goosse, H., et al.: *Introduction to climate dynamics and climate modeling*, 2010. Online textbook available at http://www.climate.be/textbook.

[3] Mulaik, S. A.: *Foundations of Factor Analysis*, 2nd edition, Chapman&Hall/CRC, 2010.

[4] Sundberg, R., Moberg, A., and Hind, A.: *Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium- Part 1: Theory*, Clim Past, **8**, pp. 1399 – 1353, 2012, doi: 10.5194/cp-8-1339-2012.

Optimal String Clustering Based on a Statistical Theory on a Topological Monoid of Strings

Session: **Miscellaneous**

Hitoshi Koyano *Quantitative Biology Center, Institute for Physical and Chemical Research, Japan*
Morihiro Hayashida *Bioinformatics Center, Kyoto University, Japan*
Tatsuya Akutsu *Bioinformatics Center, Kyoto University, Japan*

Abstract: Recently, the amount of string data generated has increased dramatically. Consequently, statistics for strings are required in many fields. To infer about populations from samples, statistics for numerical data were rigorously constructed based on probability theory on a set of numbers and numerical vector space. However, statistics for strings based on probability theory on a set of strings to infer about string populations has not been constructed. In this study, we address the problem of clustering string data in an unsupervised manner by developing a theory of a mixture model and an EM algorithm for string data based on probability theory on a topological monoid of strings developed in our previous studies. We begin with introducing a parametric probability distribution on a set of strings and finally derive a procedure for unsupervised string clustering that is asymptotically optimal in the sense that the posterior probability of making correct classifications is maximized.

1 Background and motivation

Numbers and numerical vectors account for a large portion of data. However, in recent years, the amount of string data such as biological sequences and texts has increased dramatically. Consequently, statistical methods for analyzing string data are required in many fields, including computer science and the life sciences. To infer about a population based on a sample, i.e., a part of the population observed according to a probability law, statistical methods for numerical data were rigorously constructed based on probability theory on a set $\mathbb{R}$ of real numbers and a real vector space $\mathbb{R}^p$. Similarly, statistical methods for string data should be constructed based on probability theory on a set of strings. Mathematicians have conducted detailed examinations of a large number of objects, such as numbers, manifolds, equations, and functions, throughout the long history of mathematics, but they have not studied strings in detail. A string is an object that computer scientists have addressed in depth. Stringology, a field of computer science, has thoroughly investigated algorithms and data structures for string processing. However, computer scientists have not studied strings using a mathematical approach; for example, functions, operators, and probabilities on a set of strings provided with topological and algebraic structures have not been investigated. From the viewpoint of mathematics, a set A^* of strings composed of letters of an alphabet A is a monoid with respect to concatenation and a metric space with various edit distances, and it forms a noncommutative topological monoid. Therefore, in [3, 4, 2], we constructed probability theory on the space A^* with these topological and algebraic structures to treat random strings that randomly generate strings according to a probability law in the motif of the theories on $\mathbb{R}$ and the Hilbert space L^2 of square-integrable functions for treating random variables and stochastic processes that randomly generate numbers and functions, respectively. Furthermore, we developed statistical methods for string data based on the theory and analyzed biological sequences applying the methods.

2 Results

Here we address the problem of clustering string data in an unsupervised manner by developing a theory of a mixture model and an EM algorithm for string data based on probability theory on a noncommutative topological monoid of strings constructed in our previous studies. We first construct a parametric distribution on a topological monoid of strings in the motif of the

Laplace distribution on a set of real numbers and reveal its basic properties. This Laplace-like distribution has two parameters: a string that represents the location of the distribution and a positive real number that represents the dispersion. It is difficult to explicitly write maximum likelihood estimators of the parameters because their log likelihood function is a complex function, the variables of which include a string; however, we construct estimators that almost surely converge to the maximum likelihood estimators as the number of observed strings increases and demonstrate that the estimators strongly consistently estimate the parameters. Next, we develop an iteration algorithm for estimating the parameters of the mixture model of the Laplace-like distributions and demonstrate that the algorithm almost surely converges to the EM algorithm for the Laplace-like mixture and strongly consistently estimates its parameters as the numbers of observed strings and iterations increase. Finally, we derive a procedure for unsupervised string clustering from the Laplace-like mixture that is asymptotically optimal in the sense that the posterior probability of making correct classifications is maximized. This presentation is based on [1].

References

[1] H. Koyano, M. Hayashida, and T. Akutsu. Optimal string clustering based on a Laplace-like mixture and EM algorithm on a set of strings. arXiv:1411.6471[math.ST].

[2] H. Koyano, M. Hayashida, and T. Akutsu. Maximum margin classifier working in a set of strings. *Proceedings of the Royal Society A*, 2016.

[3] H. Koyano and H. Kishino. Quantifying biodiversity and asymptotics for a sequence of random strings. *Physical Review E*, 81(6):061912, 2010.

[4] H. Koyano, T. Tsubouchi, H. Kishino, and T. Akutsu. Archaeal β diversity patterns under the seafloor along geochemical gradients. *Journal of Geophysical Research G: Biogeosciences*, 119(9):1770–1788, 2014.

Parameter estimation via failure times of coherent systems based on sequential order statistics

Session: **Miscellaneous**

Christoph Stahr *Department of Applied Statistics, RWTH Aachen University, Germany*
Marco Burkschat *Department of Applied Statistics, RWTH Aachen University, Germany*

Abstract: The assumption that component failure times in technical systems are independent may be violated if non-failed components have to compensate for failed ones and therefore failures lead to changes in the distribution of their remaining lifetime. A stochastic model should be able to incorporate such characteristics. For modeling the ordered component failure times in a given coherent system, we apply sequential order statistics based on a proportional hazard rate assumption, i.e. $F_i = 1 - (1 - F)^{\alpha_i}$ for a known baseline distribution function F and positive, unknown parameters $\alpha_1, \ldots, \alpha_n$. For estimating these parameters using system lifetime data, we consider different approaches and compare their performance in a simulation study.

1 Introduction

In classical modeling of coherent system's lifetimes, i.i.d. random variables have been applied for describing the component failure times (see e.g. [1]). Then, the cumulative distribution function of the lifetime of the coherent system is given by a mixture of the distribution functions of ordinary order statistics. If, in order to allow for possible effects of component losses (see e.g. [2]), the successive component failure times are modeled with sequential order statistics, it can be shown (see [4]) that the following similar representation holds:

$$F^T(x) = \sum_{i=1}^{n} s_i G_i(x),$$

where F^T is the c.d.f. of the entire system, G_i, $i = 1, \ldots, n$ are the c.d.f.s of the sequential order statistics and s_i, $i = 1, \ldots, n$ are the components of the signature vector that belongs to the system (see e.g. [3]).

2 The method

We are investigating possibilities to carry out inference based on system lifetime data. We assume a proportional hazard rate model, that means $F_i = 1 - (1 - F)^{\alpha_i}$, $i = 1, \ldots, n$, with positive but unknown parameters α_i and a known baseline c.d.f. F. The goal is to estimate the parameters α_i.
Beyond some established strategies such as numerically deriving the MLE, we propose a SEM-type algorithm. For every new parameter update, this algorithm simulates appropriate hidden variables which facilitate the calculation of the MLEs for the parameters α_i. The algorithm creates a markov chain whose stationary distribution will be used to estimate the MLE, thus providing a potentially fast alternative to the otherwise often complicated task of solving the likelihood equations.

References

[1] Barlow, R. E., Proschan, F. (1974). *Statistical Theory of Reliability and Life Testing: Probability Models*, Holt, Rinehart and Winston, New York.

[2] Kamps, U. (1995). A concept of generalized order statistics, pp. 1-23, Journal of Statistical Planning and Inference, Vol., 48.

[3] Samaniego, F. J. (2007). *System Signatures and Their Applications in Engineering Reliability*, Springer International series in operations research & management science, Vol., 110, New York.

[4] Navarro, J., Burkschat, M. (2011). Coherent systems based on sequential order statistics, pp. 123-135, Naval Research Logistics, Vol., 58.

Asymptotic Optimality of an Adaptive Wynn Algorithm in Binary Response Models

Session: **Miscellaneous**

Fritjof Freise *Faculty of Mathematics, Otto-von-Guericke University Magdeburg, Germany*
Norbert Gaffke *Faculty of Mathematics, Otto-von-Guericke University Magdeburg, Germany*

Abstract: Optimal designs for binary response models depend on the unknown parameter, since the corresponding information matrices do. One strategy to cope with this problem is to use sequential designs, where new design points are determined using estimates based on observations from prior stages. The dependencies of the sequences of estimators and designs can be quite complicated and standard results for their asymptotic properties cannot be applied. Especially if only one observation is taken in each step, the question arises if the information matrix of the generated design can be asymptotically singular.
A particular method for choosing the design points is an adaptive extension of the Wynn algorithm, which in its original version (see [2]) is a procedure to calculate D-optimal designs. For the case of a design region with finitely many points results on the asymptotic properties of this particular method were given in [1]. We present an extension for two parameter binary response models when the design region is a compact interval.

References

[1] Pronzato L. (2010). One-step ahead adaptive D-optimal design on a finite design space is asymptotically optimal, *Metrika 71*, pp. 219-238

[2] Wynn H.P. (1970). The sequential generation of D-optimum experimental designs. *The Annals of Mathematical Statistics 41*, pp. 1655-1664

Probabilistic investigation of complex dynamic systems in the presence of stochastic events

Session: **Miscellaneous**

Nadine Berner *Gesellschaft für Anlagen- und Reaktorsicherheit gGmbH, Garching, Germany*
Martina Kloos *Gesellschaft für Anlagen- und Reaktorsicherheit gGmbH, Garching, Germany*
Josef Scheuer *Gesellschaft für Anlagen- und Reaktorsicherheit gGmbH, Garching, Germany*
Jörg Peschke *Gesellschaft für Anlagen- und Reaktorsicherheit gGmbH, Garching, Germany*

Abstract: The probabilistic investigation of a complex dynamic process behavior in the presence of critical events, such as accident scenarios in nuclear power plants, is of paramount interest for a realistic and comprehensive safety assessment. The MCDET analysis, as a combination of Monte Carlo (MC) simulation and the Dynamic Event Tree (DET) approach, enables to study the dynamic evolution of a complex system under the consideration of aleatory (stochastic) and epistemic (knowledge-based) uncertainties [1, 2].

A considerable challenge of the MCDET analysis arises by the need to organize the computation of a massive amount of simulation sequences initiated as alternative system evolution paths due to the sampled random events. In order to efficiently perform the required multitude of simulation processes a generic approach has been developed that intuitively maps the DET topology onto a hierarchical structure of simulation processes and a congruent data structure.

The principles of the MCDET approach and the concept of the scheduling architecture are described using a classical hold-up tank system. Moreover, the potential multivariate analysis approaches for the probabilistic evaluation are discussed in the context of a fire scenario within a compartment of a nuclear power plant.

References

[1] Berner, N., Scheuer, J. (2016). MCDET analysis: From input specification of uncertainties to the probabilistic evaluation of critical variables of dynamic systems, *Risk, Reliability and Safety: Innovating Theory and Practice*, Walls, Revie & Bedford (Eds), Proceedings of the 26th Safety and Reliability Conference 2016, Glasgow, UK.

[2] Kloos, M., Berner, N., Peschke, J., Scheuer, J. (in press). MCDET: A Tool for Integrated Deterministic Probabilistic Safety Analyses, chapter in Aldemir (Ed), *Advanced Concepts in Nuclear Energy Risk Assessment and Management*, World Scientific Publishing.

Tracing Dependencies in Multivariate Networks

Session: **Graphical Models and Network Analysis**

Termeh Shafie *Department of Computer & Information Science, University of Konstanz, Germany*

Abstract: A new way of using statistical entropies to capture interdependencies among vertex and edge variables in multivariate networks was introduced in [1]. Extensions of using these entropy tools include analysing multilevel networks and selecting good summary measures of network structure.

1 Introduction

The interdependencies between vertex and edge attributes in multivariate networks are conveniently assessed by various measures based on statistical entropies. For example, joint entropy of two random variables is a measure of association and provides dependence graphs suggesting structural relationships of interest. Further, divergence measures obtained from multivariate entropies can be used to indicate and test different structural models. Some general references can be found in [3] and [4], and for multivariate entropies in [1] and [2]. An important use of the presented entropy analysis is to apply it for the selection of good summary measures of network structure, e.g. in order to obtain the centrality measure most relevant for a network property of interest. Another important use of the presented analysis is to trace tendencies of intra-dependency between variables at micro and macro level in multilevel networks.

2 Univariate, Bivariate and Higher Order Entropies

The univariate entropy for a discrete random variable X with r_X possible values x is given by

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)} \ , \tag{1}$$

where $p(x) > 0$ and $\sum_x p(x) = 1$. For two discrete random variables X and Y, the bivariate entropy is given by

$$H(X,Y) = \sum_x \sum_y p(x,y) \log_2 \frac{1}{p(x,y)} \ , \tag{2}$$

and bounded according to

$$H(X) \leq H(X,Y) \leq H(X) + H(Y) \ . \tag{3}$$

There is equality to the left if and only if X explains Y in the sense that there is a single outcome of Y for any outcome of X. There is equality to the right if and only if X and Y are independent. The two increments around $H(X,Y)$ are given by

$$EH(Y|X) = H(X,Y) - H(X) \ \text{ and } \ J(X,Y) = H(X) + H(Y) - H(X,Y) \ . \tag{4}$$

The first increment is the expected conditional entropy of Y given X and can be considered a measure of prediction strength. The second increment is the joint entropy and can be considered a measure of dependence or association between X and Y. Similarly, trivariate and higher order entropies can be used to check functional relationships and stochastic independence between several variables in a multidimensional data set.

3 Divergence Tests of Goodness of Fit

In order to judge the fit of different structural models, they can be compared and tested by using divergencies which are goodness-of-fit measures. The divergence D is the expected log

likelihood ratio according to

$$D = \sum_x \sum_y p(x,y) \log_2 \left(\frac{p(x,y)}{p_0(x,y)} \right) \ , \tag{5}$$

where p is a general model for (X,Y) and p_0 is a specified probability model for (X,Y). It is shown in [1] that $2nD/\log_2(e)$ is approximately equal to the standard Pearson χ^2 measure of goodness of fit with $(r_X - 1)(r_Y - 1)$ degrees of freedom.

4 Running Example

The well-known data set of Padgett's Florentine families[2] are used as an empirical example to illustrate the entropy tools. This data consists of 87 families with vertex attributes specifying economic, social and political influence, and two undirected relations specifying marriage and business alliances between pairs of families.

References

[1] Frank, O. and Shafie, T. (2016). Multivariate Entropy Analysis of Network Data. *Bulletin of Sociological Methodology*, 129(1):45-63.

[2] Frank, O. (2011). Statistical information tools for multivariate discrete data. In: Pardo L, Balakrishnan N and Àngeles Gil M (eds) *Modern Mathematical Tools and Techniques in Capturing Complexity*. Berlin: Springer Verlag, pp. 177–190.

[3] Gray, R. M. (2011). *Entropy and Information Theory*, New York: Springer Verlag.

[4] Kullback, S. (1968). *Information Theory and Statistics*, New York: Dover Publications.

[2] data provided by John Padgett and Ronald Breiger

Dynamic graphical models for samples of network time series

Session: **Graphical Models and Network Analysis**

Osvaldo Anacleto *Roslin Institute, University of Edinburgh, UK*
Lilia Costa *Department of Statistics, Federal University of Bahia, Brazil*
Jim Smith *Department of Statistics, University of Warwick, UK*
Catriona Queen *Department of Mathematics and Statistics, The Open University, UK*

1 Introduction

Multivariate time series are often observed on networks and can be available for a sample of observational units. For example, time series of brain activity measurements in regions forming a network in the brain are often analysed using a sample of individuals. In this case, there is great interest in estimating how the strength of connections between brain regions varies over time and across individuals. However, few models can pool information across individuals in order to obtain more precise and robust estimates.
It will be shown in this talk how samples of multivariate time series can be modelled using chain graphs. Computation in this new model is simplified and parallelisable since the conditional independence structure induced by the graph enables a multivariate problem to decomposed into separate, simpler sub-problems of lower dimensions. The advantages of the new model will be illustrated with network time series of brain activity and gene expression.

2 Representing network time series with chain graphs

Let $\boldsymbol{Y}_{1t}, \ldots, \boldsymbol{Y}_{qt}$ be a sample of n-dimensional time series of size q. For example, these time series can represent fMRI measurements of brain activity over time in n regions that forms a network in the brain, for a sample of q individuals. Suppose that, for each time $t \in \mathbb{N}$, there is a causal drive through the network with conditional independence structure (which is the same structure for each $\boldsymbol{Y}_{qt}$ and fixed over time) represented by a chain graph $\mathcal{G}$.
For example, consider q observations of the 5-dimensional time series ($n = 5$) with chain graph represented by white nodes in Figure 17. Each time series $\boldsymbol{Y}_{jt}^{\mathrm{T}} = (Y_{jt}(1), Y_{jt}(2), \boldsymbol{Y}_{jt}(3)^{\mathrm{T}}, Y_{jt}(4))$ has the *chain components* $Y_{jt}(1)$, $Y_{jt}(2)$, $\boldsymbol{Y}_{jt}(3) = (Y_{jt}^{(1)}(3), Y_{jt}^{(2)}(3))^{\mathrm{T}}$ and $Y_{jt}(4)$, $j = 1, \ldots, q$,. Also, in each of the q multivariate time series, the chain graphs at each time $t \in \mathbb{N}$ can be linked together by assuming inter time-slice dependencies, so that a *dynamic chain graph* for each $\boldsymbol{Y}_{jt}$ is obtained (see [1] for details).

3 The dynamic chain graph hierarchical model

If q multivariate time series $\boldsymbol{Y}_{1t}, \ldots, \boldsymbol{Y}_{qt}$ are represented by a dynamic chain graph as described above, and conditional state space models are defined for the set $\{\boldsymbol{Y}_{1t}(i), \ldots, \boldsymbol{Y}_{qt}(i)\}$ for each component $i = 1, \ldots N$, it can be proved that inference can be done separately for state vectors and parameters of the model from each set $\{\boldsymbol{Y}_{1t}(i), \ldots, \boldsymbol{Y}_{qt}(i)\}$. This approach therefore breaks a potential high-dimensional problem into N problems of lower dimensions.
As an example, consider again the q observations of the 5-dimensional time series represented by the chain graph in Figure 17. For $k = 1, 2$, the set $\{Y_{1t}(k), \ldots, Y_{qt}(k)\}$ follows a state space model with respective state vectors $\beta_{1t}(k), \ldots, \beta_{qt}(k)$. To accommodate the variation across observations with respect to the chain component k, it is assumed that $\beta_{1t}(k), \ldots, \beta_{qt}(k)$ are exchangeable with common mean $\mu_t(k)$, whose evolution over time must be defined in the model. A similar model is defined for $\{Y_{1t}(4), \ldots, Y_{qt}(4)\}$, considering their parents (induced by the chain graph) as regressors. Additionally, a multivariate state space model is defined for $\boldsymbol{Y}_{jt}(3) = (Y_{jt}^{(1)}(3), Y_{jt}^{(2)}(3))^{\mathrm{T}}$, also conditional on their parents. Under certain assumptions,

conjugate Bayesian inference for state vectors and parameters can be done for the models associated with the univariate chain components $Y_{jt}(1)$, $Y_{jt}(2)$ and $Y_{jt}(4)$ (see [2] for details). The resulting *dynamic chain graph hierarchical model* is represented in Figure 17.
This class of models is currently being applied to network time series of gene expression and brain activity, and preliminary results will be also presented in the talk.

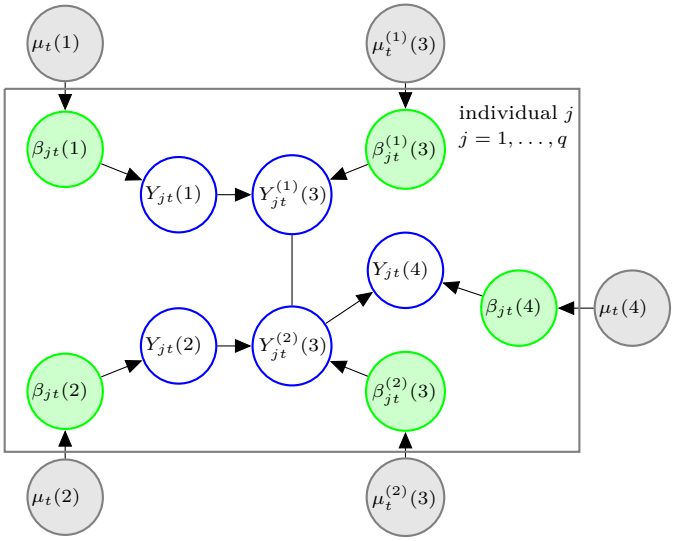


Figure 17: Structure of a dynamic chain graph hierarchical model

References

[1] Anacleto, O., & Queen, C. (2016). Dynamic chain graph models for time series network data. *Bayesian Analysis*, In-Press.

[2] Gamerman, D., & Migon, H. S. (1993). Dynamic hierarchical models. *Journal of the Royal Statistical Society*. Series B (Methodological), 629-642.

Causal reasoning for events in continuous time

Session: **Graphical Models and Network Analysis**

Vanessa Didelez *Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany*

Abstract:
Dynamic associations among different types of events in continuous time can be represented by local independence graphs as developed by Didelez (2008). Intuitively we say that a process is locally independent of another one if its short-term prediction is not improved by using the past of the other process, similar to Granger-non-causality; the graphical representation uses nodes for processes or events and the absence of a directed edge for local independence. Important independence properties can be read o these -possibly cyclic- graphs using delta-separation (Didelez, 2006) which generalises d-separation from DAGs. In related work, Roysland (2011, 2012) showed how causal inference based on inverse probability weighting (IPW), well known for longitudinal data (Robins et al., 2000), can be extended to the continuous-time situation using a martingale approach.
In this joint work (with Kjetil Rysland, Odd Aalen and Theis Lange), we start by defining causal validity of local independence graphs in terms of interventions, which in the context of events in time take the form of modifications to the intensities of specific processes, e.g. a treatment process; causal validity is given if the specification of the dynamic system is rich enough to model such an intervention. (similar to what is known as 'modularity' for causal DAGs). We then combine the above previous developments to give graphical rules for the identifiability of the effect of such interventions via IPW; these rules can be regarded as characterising 'unobserved confounding'. Re-weighting then simply replaces the observed intensity by the one given by the intervention of interest. For this to be meaningful, causal validity and identi ability are crucial assumptions.
Our approach can be regarded as the time-continuous version of Dawid & Didelez (2010), who develop a decision theoretic approach for sequential decisions in longitudinal settings and use a graphical representation with influence diagrams that include decision nodes; specifically causal validity is analogous to the extended stability of Dawid & Didelez (2010). As an aside, we find that it is helpful to also use causal reasoning when faced with censoring as the target of inference can often be regarded as the population in which censoring is prevented, i.e. its intensity is set to zero. We apply our theoretical results to the example of cancer screening in Norway.

Learning non-linear graphical models

Session: **Graphical Models and Network Analysis**

Anna Gottard *Department of Statistics, Computer Science, Applications, University of Florence, Italy*

Abstract: A graphical model is a probabilistic model associated to a graph, whose nodes represent random variables and missing edges correspond to conditional dependencies. This work proposes a class a Bayesian ensemble methods to learn the structure of undirected graphical models. Interesting applications of the proposal is metabolomics, where assumptions of a joint Gaussian distribution is typically considered unrealistic.

1 Introduction

Graphical models (see, for e.x. [5]) are models for a collection of random variables X_v, $v \in V = \{1, \ldots, p\}$ whose conditional independence structure is depicted in a graph $\mathcal{G}(V, E)$. The set V collects the nodes of the graph, each one associated with a random variable, and the set E collects the edges connecting the nodes. The rules regulating conditional independence statements are called Markov properties and are specific for each kind of graphs, undirected, directed and mixed. In particular, for undirected graphs, also called *Markov random fields* or simply *Networks*, all the edges are not oriented. The local Markov property for this kind of graph assesses that each variable X_v in the graph is conditionally independent of all the other nodes given its neighbours. The set of *neighbours* of X_v, denoted with $ne(v)$, collects all the variables whose nodes have an edge connecting them to v. Moreover, the set of v and its neighbours is called *closure* and denoted by $cl(v)$. Utilizing Dawid's notation [3], local Markov property can be written as $X_v \perp\!\!\!\perp X_{V \setminus cl(v)} \mid X_{ne(v)}$. The conditional independence structure of an undirected graphical model can be therefore characterized by a p by p adjacency matrix $\mathcal{A}$, with entry $\mathcal{A}_{ij} = 1$ when X_i and X_j are neighbours and $\mathcal{A}_{ij} = 0$ otherwise. When the joint distribution of $\mathbf{X}_V$ is Gaussian, say $N_p(0, \Sigma)$, the zeros in the adjacency matrix exactly match the zeros in the inverse Σ^{-1}. Typically the adjacency matrix in unknown and it is of interest to learn it from the data. Extensive research efforts have been conducted for learning the graph from data, for example using multiple testing procedures such as in [4], or via regularised estimators as in [9] and [8].
Most of this research efforts were developed for learning the graph structure in the case of joint Gaussian distribution, with linear relationships. As both linear and non-linear relationships may appear in many applications, take as an example metabolomics, in this work, I am going to briefly introduce a graphical model based on the Bayesian Additive Regression Trees as an alternative of what proposed by [7] and [6], to uncover non-linear relationships.

2 The proposal

This work proposes learning the adjacency matrix of an undirected graph via node-wise regression and local Markov property. Specifically, for each node $v \in V$, we assume that the dependence structure of X_v on its neighbours can be described by a Bayesian Additive Regression Trees (BART) [1], that is the sum of *trees*, as follow

$$\mathbb{E}(X_v \mid X_{ne(v)}) = \sum_{j=1}^{m} \mathcal{T}_j(X_{ne(v)}; T_j, M_j),$$

where T_j and M_j are the model parameters representing the structure of each tree $\mathcal{T}_j$ and the mean vectors over its terminal nodes respectively. BART models have shown a great ability in detecting non-linear dependencies and can be successfully applied also in case high dimensional data. The adjacency matrix can be then estimated as a function of the Importance Vector of each node-wise BART regression. The proposed models will be applied to metabolomics data.

References

[1] Chipman H.A., George E.I., & McCulloch R.E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1) 266-298

[2] Dahinden C., Kalisch M., & Bühlmann P. (2010). Decomposition and model selection for large contingency tables. *Biometrical Journal*, 52(2), 233-252.

[3] Dawid A.P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41, 1-31.

[4] Drton M., & Perlman M.D. (2007). Multiple testing and error control in Gaussian graphical model selection. *Statistical Science*, 22, 430-449.

[5] Lauritzen S. L (1996). *Graphical models*, Clarendon Press.

[6] Liu Y., Zayas-Castro J.L., Fabri P., & Huang S. (2014). Learning high-dimensional networks with nonlinear interactions by a novel tree-embedded graphical model. *Pattern Recognition Letters*, 49, 207-213.

[7] Fellinghauer, B., Bühlmann P., Ryffel M., Von Rhein M. and Reinhardt J. D. (2013). Stable graphical model estimation with random forests for discrete, continuous, and mixed variables, *Computational Statistics & Data Analysis*, 64, 132–152.

[8] Friedman J., Hastie T., & Tibshirani R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432-441.

[9] Meinshausen N., & Bühlmann P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3), 1436-1462.

Estimating divergences through weighted bootstrap

Session: **Distance-based Statistical Methods**

Michel Broniatowski *Université Pierre et Marie Curie, Paris, France*

Abstract: Sanov type results hold for some weighted versions of empirical measures, and the rates for those Large Deviation principles can be identified as divergences between measures, which in turn characterize the form of the weights. This correspondence is considered within the range of the Cressie-Read family of statistical divergences, which covers most of the usual statistical criterions. We propose a weighted bootstrap procedure in order to estimate these rates. To any such rate we produce an explicit procedure which defines the weights, therefore replacing a variational problem in the space of measures by a simple Monte Carlo procedure. Importance Sampling may be used in order to provide accurate estimators in reasonable run time; some insight on these issues will be discussed.
Keywords: Divergence, optimization, bootstrap, Monte Carlo, large deviation, weighted empirical measure, conditional Sanov Theorem

References

[1] Broniatowski, M. Weighted sampling, maximum likelihood and minimum divergence estimators. Geometric science of information, 467–478, Lecture Notes in Comput. Sci., 8085, Springer, Heidelberg, 2013

[2] Csiszár, I. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. (German) Magyar Tud. Akad. Mat. Kutató Int. Közl. 8 1963 85–108.

[3] Read, T. R. C., Cressie, N. A. C. Goodness-of-fit statistics for discrete multivariate data. Springer Series in Statistics. Springer-Verlag, New York, 1988. xii+211 pp. ISBN: 0-387-96682-X

[4] Letac, G., Mora, M. Natural real exponential families with cubic variance functions. Ann. Statist. 18 (1990), no. 1, 1–37. ..

[5] Barbe, P., Bertail, P., "The Weighted Bootstrap," Lecture Notes in Statistics (1995), Springer-Verlag, New York.

[6] Csiszár, I. On topology properties of f-divergences. Studia Sci. Math. Hungar. 2 1967 329–339.

[7] Najim, J., "A Cramer type theorem for weighted random variables," *Electron. J. Probab.,* Vol. 7 (2002).

[8] Liese, F., Vajda, I. Convex statistical distances. With German, French and Russian summaries. Teubner-Texte zur Mathematik [Teubner Texts in Mathematics], 95. BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1987. 224 pp. ISBN: 3-322-00428-7

[9] Trashorras, J; Wintenberger, O Large deviations for bootstrapped empirical measures. Bernoulli 20 (2014), no. 4, 1845–1878.

A New Method of Robust Statistics

Session: **Distance-based Statistical Methods**

Wolfgang Stummer *Department of Mathematics, University of Erlangen-Nürnberg, Germany*

Abstract: We introduce a new method for the goal-oriented design of outlier- and inlier-robust statistical inference tools. In particular, this includes the tasks of parameter estimation, testing for goodness-of-fit resp. homogenity resp. independence, clustering, change-point detection, exploratory model search, and some Bayesian decision procedures.

In order to achieve this goal, we adapt the concept of *scaled Bregman distances between two distributions*, which was introduced in Stummer (2007), Stummer & Vajda (2012), and which generalizes the widely-used (non-robust) concepts of Kullback-Leibler information distance/relative entropy, Pearson's chisquare distance, Hellinger distance, Csiszar-Ali-Slivey divergences, etc. The classical (i.e., unscaled) Bregman distances – such as the L^2-distance and the more general density power divergences – are covered as well.

In order to visualize effectively and transparently the corresponding robustness properties, we present 3D-plots of the newly developed *density-pair adjustment functions*. Numerous special cases will be illustrated. For the discrete case, some universally applicable results on the asymptotics of the underlying scaled-Bregman-distance test statistics are derived as well.

This talk is mainly based on several joint works with A.-L. Kißlinger (Erlangen-Nürnberg).

- Kißlinger AL, Stummer W (2013) Some decision procedures based on scaled Bregman distance surfaces. In: Nielsen F, Barbaresco F (eds) GSI 2013, Lecture Notes in Computer Science LNCS 8085, Springer, pp 479 - 486.
- Kißlinger AL, Stummer W (2015a) New model search for nonlinear recursive models, regressions and autoregressions. In: Nielsen F, Barbaresco F (eds) GSI 2015, Lecture Notes in Computer Science 9389, Springer, pp 693 - 701.
- Kißlinger AL, Stummer W (2015b) A new information-geometric method of change detection. Preprint.
- Kißlinger AL, Stummer W (2016) Robust statistical engineering by means of scaled Bregman distances. In: C. Agostinelli, A. Basu, P. Filzmoser and D. Mukherjee (eds.), Recent Advances in Robust Statistics: Theory and Applications, Springer 2016, pp. 81-113.
- Stummer W (2007) Some Bregman distances between financial diffusion processes. Proc Appl Math Mech (PAMM) 7:1050,503 - 1050,504.
- Stummer W, Vajda I (2012) On Bregman distances and divergences of probability measures. IEEE Trans Inf Theory 58(3):1277 - 1288.

A New Information-Geometric Method of Change Detection

Session: **Distance-based Statistical Methods**

Wolfgang Stummer *Department of Mathematics, University of Erlangen-Nürnberg, Germany*
Anna-Lena Kißlinger *Chair of Statistics and Econometric, University of Erlangen-Nürnberg, Germany*

Abstract: We present a new toolbox for detecting distributional changes in sets (e.g. streams, clouds) of random data, i.e. in data of partially non-deterministic (uncertain, random, noisy, risk-prone) nature or interpretation by means of scaled Bregman distances as introduced by Stummer (2007), Stummer & Vajda (2012). The range of applications for such kind of change (point) detection procedures is very vast, including for instance the identification of instabilities in industrial processes, tumor progression or anomalies in medical observations, DNA disorders or disease outbreaks in biology, structural breaks (switching of "regimes") in economics, substantial variations of volatility fluctuations in financial markets, essential climate changes in environmetrics, etc.

After situation-based data extraction resp. calculation of the corresponding data-derived distributions, we look for goal-oriented situation-based relevant changes between those distributions of the extracted subsamplings by using scaled Bregman distances. By using particular scaling measures – so called *scale connectors* – the sensitivity of the scaled Bregman distance to changes in the considered data-derived distributions can be controlled (see Kißlinger & Stummer (2016)).

For the purpose of statistical decisions on changes in terms of estimates, $p-$values, tests, etc. we will investigate some sample-size asymptotics on the distribution of scaled Bregman distances between two data-derived distributions. From this, we derive furthermore a 3D computer-graphical change-score decision-procedure.

This talk is mainly based on:

- Kißlinger AL., Stummer W. (2013). Some decision procedures based on scaled Bregman distance surfaces. In: Nielsen F, Barbaresco F (eds) GSI 2013, Lecture Notes in Computer Science LNCS 8085, Springer, pp 479 - 486.
- Kißlinger AL, Stummer W (2016) Robust statistical engineering by means of scaled Bregman distances. In: C. Agostinelli, A. Basu, P. Filzmoser and D. Mukherjee (eds.), Recent Advances in Robust Statistics: Theory and Applications, Springer 2016, pp. 81-113.
- Stummer W (2007) Some Bregman distances between financial diffusion processes. Proc Appl Math Mech (PAMM) 7:1050,503 - 1050,504.
- Stummer W, Vajda I (2012) On Bregman distances and divergences of probability measures. IEEE Trans Inf Theory 58(3):1277 - 1288.

Appendix

List of Authors

Overview of Adlershof: