ICORS LACSC 2019

GUAYAQUIL - ECUADOR

BOOK OF ABSTRACTS

ORGANIZER

ESPOL

Book of Abstracts

International Conference on Robust Statistics 2019

4th Latin American Conference on Statistical Computing

ICORS-LACSC 2019

ICORS LACSC 2019
GUAYAQUIL - ECUADOR

ORGANIZER
ESPOL

May 28-31, 2019

Escuela Superior Politécnica del Litoral (ESPOL)

Guayaquil-Ecuador

Contents

Preface

It is a great honor and privilege for Escuela Superior Politécnica del Litoral (ESPOL) to organize two important international conferences in statistics jointly: the *International Conference on Robust Statistics* (ICORS 2019) and the *4th Latin American Conference on Statistical Computing* (4th LACSC). ICORS-LACSC 2019 takes place from May 28-31, 2019 at ESPOL campus “Gustavo Galindo” in Guayaquil-Ecuador.

In 2019, ICORS celebrates its 19th edition, after the recent meetings in Sydney, Australia (2017) and Leuven, Belgium (2018). On the other hand, LACSC celebrates its 4th edition, following its previous meetings in Gramado, Brazil (2016), Valparaíso, Chile (2017) and San José, Costa Rica (2018). LACSC is the official conference of the *Latin American Regional Section* of the *International Association for Statistical Computing* (LARS-IASC). This will be an event without precedent as the communities of ICORS and LACSC will gather. ICORS meetings include contributions to practical and theoretical aspects of robust statistics, data analysis and related topics. The 4th LACSC includes contributions to theory and practice of computational statistics, bayesian statistics, optimization, functional data analysis, modeling, clustering methods, deep learning, parallel computing, time series and regression analysis, and other important themes. This meeting creates a forum to discuss recent progress and emerging ideas in these topics and encourage informal contacts and discussions among all the participants. Moreover, ICORS 2019 and the 4th LACSC can play an important role in maintaining cohesive these two communities, whose interactions transcend the meetings and endure year round.

This will also be a unique opportunity to visit the beautiful city of Guayaquil in Ecuador with all its natural attractions like the Guayas river, mangrove forests, the green iguanas, the green macaw, its canals, hills, parks, museums, history and the architecture of Spanish colonial houses.

This book contains the scientific programme of ICORS 2019 and of the 4th LACSC, and the abstracts of all the presentations. The abstracts are ordered according to the name of the presenting author. ICORS-LACSC 2019 features joint keynote talks by Stefan Van Aelst (KU Leuven, Belgium), Ana Bianco (University of Buenos Aires, Argentina), Dirk Eddelbuettel (University of Illinois, USA) and Gareth James (University of Southern California, USA), as well as invited talks for ICORS 2019 by Xuming He (University of Michigan, USA), Peter Rousseeuw (KU Leuven, Belgium), Marco Avella (Columbia University, USA), Maria-Pia Victoria-Feser (Université de Genève, Switzerland), Ruben Zamar (University of British Columbia, Canada), Daniela Rodríguez (University of Buenos Aires, Argentina) and Klaus Nordhausen (TU Wien, Austria). The 4th LACSC includes five invited sessions, one of them in collaboration with the International Society for Business and Industrial Statistics.

Finally, on behalf of the local organizing committee, we would like to express our gratitude to all persons and institutions who make this joint conference possible: the Rector of ESPOL Cecilia Paredes, the Dean of the Faculty of Natural Sciencies and Mathematics Marcos Mendoza, Peter Rousseeuw, Paulo Canas, Peter Filzmoser, the colleagues from the Scientific Programme Committee, the keynote and invited speakers, the sponsors, and all the people who contribute to the scientific programme. We also would like to thank all the participants for joining ICORS-LACSC 2019. We wish you a fruitful and stimulating conference and a pleasant stay in Guayaquil.

Guayaquil, May 2019

Holger Cevallos-Valdiviezo

Local Organizing Committee

- Holger Cevallos-Valdiviezo (Chair)
- Marcos Mendoza (Dean of Faculty of Natural Sciences and Math.)
- Francisca Flores (Vice-Dean of Faculty of Natural Sciences and Math.)
- Nereyda Espinoza
- María Pastuizaca
- Wendy Plata
- Francisco Moreira
- Fernando Tenesaca
- Ivanna Rodríguez
- Marcos Ruiz

ICORS Scientific Programme Committee

- Ana Bianco (University of Buenos Aires, Argentina)
- Graciela Boente (University of Buenos Aires, Argentina)
- Peter Filzmoser (TU Wien, Austria)
- Rosario Oliveira (Instituto Superior Técnico, Portugal)
- Elvezio Ronchetti (Université de Genève, Switzerland)
- Peter Rousseeuw (KU Leuven, Belgium)
- Anne Ruiz-Gazen (University Toulouse 1 Capitole, France)
- Alan Welsh (Australian National University, Australia)
- Ruben Zamar (University of British Columbia, Canada)

LACSC Scientific Programme Committee

- Alba Martínez-Ruiz (University of San Sebastián, Chile)
- David Muñoz (Instituto Tecnológico Autónomo de México, Mexico)
- Francisco Louzada (University of Sao Paulo, Brazil)
- Holger Cevallos-Valdiviezo (ESPOL, Ecuador)
- Javier Trejos (University of Costa Rica, Costa Rica)
- Luis Firinguetti (University of Bío-Bío, Chile)
- Martha Bohórquez (Universidad Nacional de Colombia, Colombia)
- Nikolai Kolev (University of Sao Paulo, Brazil)
- Paulo Canas-Rodrigues (Federal University of Bahia, Brazil)
- Verónica González-López (University of Campinas, Brazil)

Program Overview

Tuesday, 28 May 2019

Plenary Session ICORS-LACSC 2019 — 09:30 – 10:30

Keynote Presentation: P1 — Tuesday 28th
Chair: Holger Cevallos-Valdiviezo, Room: Auditorium 2

J. Derenski, Y. Fan, G. James
An Empirical Bayes Solution for Selection Bias in Functional Data (see p. 67)

Coffee Break: 10:30 – 11:00

Contributed Session ICORS — 11:00 – 12:00

ICORS Contributed Paper Session: ICORS.CPS1 — Tuesday 28th
Chair: Graciela Boente, Room: Auditorium 1

D. Kepplinger, E. Smucler
Robust Variable Selection via Adaptive Elastic Net S-Estimators for Linear Regression (see p. 69)
11:00 – 11:20

A. C. Garcia-Angulo, G. Claeskens
Post-selection confidence curves (see p. 57)
11:20 – 11:40

Feifang Hu, Yichen Qin, Yang Li, Wei Ma
Robust Designs of Big Comparative Studies (see p. 65)
11:40 – 12:00

Recent advances in Bayesian Statistics and Stochastic Processes 11:00 – 12:30

LACSC Contributed Paper Session: LACSC.CPS1 — Tuesday 28th
Chair: Marcelo Bourguignon, Room: Auditorium 3

E. E. Alvarez , M. L. Riddick
Bayesian Estimation in the Additive Hazard Model (see p. 97)
11:00 – 11:20

José Soto, Saba Infante, Franklín Camacho, Rafael Amaro
Inference in stochastic mixed-effect models (see p. 105)
11:20 – 11:40

R. Carvajal-Schiaffino, V. Cid-Ossandon, V. H. Salinas
Bayesian Estimation of the Limiting Availability in a Repairable Coherent System (see p. 37)
11:40 – 12:00

Saba Infante, Edward Gómez, Luis Sánchez, Aracelis Hernández
An estimation of the industrial production dynamic in the Mercosur countries using the Markov switching model (see p. 66)

12:00 – 12:20

ISBIS IPS: Robust and Clustering Methods in Time Series Analysis 11:00 – 12:30

LACSC Invited Paper Session: LACSC.IPS1 Tuesday 28th
Chair: Vanda Milheiro Lourenço, Room: Auditorium 2

P. C. Rodrigues
Robust and randomized singular spectrum analysis (see p. 36)
11:00 – 11:30

V. A. Reisen
Robust factor modeling for high-dimensional time series (see p. 95)
11:30 – 12:00

N. Ravishanker
Biclustering algorithms for high-frequency financial time series (see p. 93)
12:00 – 12:30

Lunch Break: 12:30 – 14:00

Invited Session ICORS 14:00 – 15:00

ICORS Invited Paper Session: ICORS.IPS1 Tuesday 28th
Chair: Klaus Nordhausen, Room: Auditorium 2

Y. Sun, X. He
Inference on Quantile Regions in Linear Models (see p. 62)
14:00 – 14:30

M. Avella-Medina
Privacy-preserving parametric inference: a case for robust statistics (see p. 26)
14:30 – 15:00

Stochastic and optimization processes 14:00 – 15:00

LACSC Contributed Paper Session: LACSC.CPS2 Tuesday 28th
Chair: Francisco Louzada-Neto, Room: Auditorium 1

X. Cabezas, S. García
A Two-Stage Stochastic Formulation for The Simple Plan Location Problem with Order (see p. 33)
14:00 – 14:20

I. Soria, H. A. Pérez
Lagrangian Relaxation for Design of a Soda Company Distribution System (see p. 104)
14:20 – 14:40

J. E. Fernandez, B. Silva
Probabilistic Constrained Optimization Using Bayesian Networks (see p. 53)
14:40 – 15:00

Functional Data Analysis and Applications **14:00 – 15:00**

LACSC Contributed Paper Session: LACSC.CPS3 Tuesday 28th
Chair: Ruben Carvajal-Schiaffino, Room: Auditorium 3

J. Olaya Ochoa, D. P. Ovalle
ANOVA test for correlated functional data applied to fine particulate matter measurements on air (see p. 88)
14:00 – 14:20

Julian A. A. Collazos, Adriano Z. Zambom, Ronaldo Dias
Variable Selection in Functional Linear Cox Regression Model via Regularization Methods Applied to Clinical Data (see p. 43)
14:20 – 14:40

A. Villamil, M. Bohorquez R. Giraldo, J. Mateu
Spatfd: An R package for functional kriging, functional cokriging and optimal spatial sampling of functional data (see p. 114)
14:40 – 15:00

Contributed Session ICORS **15:00 – 16:00**

ICORS Contributed Paper Session: ICORS.CPS2 Tuesday 28th
Chair: Maria-Pia Victoria-Feser, Room: Auditorium 1

A. N. Vidyashankar, L. Li
Robust Variational Inference via Divergences (see p. 113)
15:00 – 15:20

Lei Li, Anand N. Vidyashankar
Divergence Methods for Models with Latent Structure: Theory and Algorithms (see p. 73)
15:20 – 15:40

Rocío Maehara, Heleno Bolfarine Filidor Vilca, N. Balakrishnan
Sinh-skew-normal/Independent Regression Models (see p. 79)
15:40 – 16:00

Statistical Software and Parallel Computing **15:00 – 16:00**

LACSC Contributed Paper Session: LACSC.CPS4 Tuesday 28th
Chair: Juergen Symanzik, Room: Auditorium 3

L. Corain, R. Carvajal-Schiaffino, J-M. Graïc, E. Grisan, R. Luisetto, L. Salmaso, A. Peruffo
Efficient Permutation Inference by Computing Parallel Algorithms to Support Comparative Neuroanatomy (see p. 44)
15:00 – 15:20

R. Carvajal-Schiaffino, F. Novoa-Muñoz, C. González-Aguero
Implementation of a Parallel Algorithm with Shared/Private Memory for Parametric Booststrap (see p. 38)
15:20 – 15:40

O. Centeno-Mora
Predicting the public institutional budget: an application using shinydashboard (see p. 39)
15:40 – 16:00

Recent Advances in Regression and Inference **15:00 – 16:00**

LACSC Contributed Paper Session: LACSC.CPS5 Tuesday 28th
Chair: Francisco Plaza, Room: Auditorium 2

Marcelo Bourguignon, Manoel Santos-Neto, Mário de Castro
A new regression model for positive random variables with skewed and long tails (see p. 31)
15:00 – 15:20

E. Gonzalez-Estrada, W. Cosmes , J. A. Villasenor
Shapiro-Wilk test for skew normal distributions (see p. 59)
15:20 – 15:40

D. P. S. Bussola, J. A. Achcar, R. M. Souza
Linear regression models assuming a stable distribution for the response data (see p. 103)
15:40 – 16:00

Recent Advances in Multivariate Statistics **16:00 – 17:00**

LACSC Contributed Paper Session: LACSC.CPS6 Tuesday 28th
Chair: Higor Cotta, Room: Auditorium 2

M. Ramos-Barberán, M. V. Hinojosa-Ramos, J. Ascencio-Moreno, F. Vera, O. Ruiz-Barzola, M. P. Galindo-Villardón
Batch process control, monitoring: a Dual STATIS, Parallel Coordinates (DS-PC) approach (see p. 111)
16:00 – 16:20

Carlos Martin-Barreiro, John Ramirez-Figueroa
Disjoint orthogonal components in Tucker models (see p. 82)
16:20 – 16:40

John Ramirez-Figueroa, Carlos Martin-Barreiro
An alternative method for obtaining disjoint principal components by particle swarm optimization. (see p. 92)
16:40 – 17:00

Modelling Complex Data Structures with Applications in the Society and Environment I **16:00 – 17:00**

LACSC Contributed Paper Session: LACSC.CPS7 Tuesday 28th
Chair: Xavier Cabezas, Room: Auditorium 1

Maria Rodriguez C., Rafael España
Challenges in estimating individual/household level severity parameters with the Food Insecurity Scale (FIES) (see p. 100)
16:00 – 16:20

Anthony Villacís, Kenny Escobar, Juan Carlos Letechi, Laura Andrea López Rodríguez
The art of robust statistics for the analysis and improvement of a hospital's medical care (see p. 72)
16:20 – 16:40

J. Fernández Ledesma
A multivariate model to discovery knowledge of research groups from patents (see p. 52)
16:40 – 17:00

Welcome Reception: 17:00 – 19:00
(STEM main hall)

Wednesday, 29 May 2019

Deep Learning Applications to Data Science 08:30 – 10:00

LACSC Invited Paper Session: LACSC.IPS2 Wednesday 29th
Chair: Alba Martinez-Ruiz, Room: Auditorium 2

Javier Linkolk López-Gonzales, Cristian Ubal, Orietta Nicolis, Romina Torres, Rodrigo Salas Fuentes
Air Pollution prediction using Self-Organizing Long-Short Term Memory Networks (see p. 71)
08:30 – 08:50

G. DiGiorgi, R. Salas, M. Salinas, R. Torres, O. Nicolis
A Deep Neural Stochastic model for Cryptocurrency Volatility Prediction (see p. 49)
08:50 – 09:10

F. Plaza, R. Salas, O. Nicolis
Seismic activity forecast using Convolutional and LSTM Neural Networks (see p. 102)
09:10 – 09:30

A. Martínez-Ruiz, C. Montañola-Sales
Parallel Statistical Algorithms: Careful Design and Important Decisions (see p. 83)
09:30 – 09:50

Contributed Session ICORS 09:00 – 10:00

ICORS Contributed Paper Session: ICORS.CPS3 Wednesday 29th
Chair: Andreas Alfons, Room: Auditorium 1

M. Zhelonkin
Probabilistic Forecasting of Binary Outcomes in Presence of Outliers (see p. 118)
09:00 – 09:20

H. H. A. Cotta, V. A. Reisen, P. Bondon, C. Lévy-Leduc
A robust alternative to the sample autocovariance and autocorrelation functions (see p. 45)
09:20 – 09:40

T. Cipra, R. Hendrych
Estimation of volatility models by robustified recursive procedures (see p. 42)
09:40 – 10:00

Invited Session ICORS 10:00 – 11:00

ICORS Invited Paper Session: ICORS.IPS2 Wednesday 29th
Chair: Ana Bianco, Room: Auditorium 2

M. Hubert, P. J. Rousseeuw, W. Van den Bossche
MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers (see p. 101)
10:00 – 10:30

K. Nordhausen

Robust approaches for blind source separation (see p. 87)
10:30 – 11:00

Applications to Economics and Finance 10:00 – 11:00

LACSC Contributed Paper Session: LACSC.CPS8 Wednesday 29th
Chair: Saba Infante, Room: Auditorium 1

A. Alonzo , E. Avila, A. Matamoros
Bayesian Inference for the estimation of the touristic expenditure in Honduras (see p. 23)
10:00 – 10:20

L. Méndez, S. Ongena
"Finance, Growth" Re-Loaded (see p. 77)
10:20 – 10:40

J. Cruzatti Constantine, H. Cevallos-Valdiviezo
Multipolar Aid: A Human Development analysis with High-Resolution Data (see p. 46)
10:40 – 11:00

Clustering of Complex Data 10:00 – 11:00

LACSC Contributed Paper Session: LACSC.CPS9 Wednesday 29th
Chair: Cedric Heuchenne, Room: Auditorium 3

M. Choez
Q-means, a new quantile method to initial seed value selection in K-means algorithm for clustering. (see p. 41)
10:00 – 10:20

L. E. Amaya, J. Trejos
Comparison of Optimization Metaheuristics Based on Neighborhoods for Clustering Binary Data (see p. 24)
10:20 – 10:40

E. Acuña, W. Quispe, R. Trespalacios, V. Palomino, C. Vega, R. Mégret, J. Agosto[d]
Clustering using Functional Data Analysis for Honeybees Daily Activity Data (see p. 89)
10:40 – 11:00

Coffee Break: 11:00 – 11:15

Plenary Session ICORS-LACSC 2019 11:15 – 12:15

Keynote Presentation: P2 Wednesday 29th
Chair: Shoja Chenouri, Room: Auditorium 2

I. Kalogridis, S. Van Aelst
Robust estimation for functional and partially functional linear models (see p. 108)

Lunch Break: 12:15 – 14:00

Excursion: 14:00 – 19:00

Tickets still available at the Registration desk

Buses leave from ESPOL for the City Tour

Thursday, 30 May 2019

Statistical Computing: Software and Applications 08:30 – 10:00

LACSC Invited Paper Session: LACSC.IPS3 Thursday 30th
Chair: Verónica González-López, Room: Auditorium 2

D. F. Muñoz, H. Gardida, H. Velásquez, J. D. Ayala
Simulation models to support preliminary electoral results program for the Mexican Electoral Institute (see p. 85)
08:30 – 09:00

D. A. Medina, A. X. Jerves
A geometry-based algorithm for cloning real grains 2.0 (see p. 68)
09:00 – 09:30

J. Symanzik
Linked micromap plots: Design principles, past uses and new perspectives via the "rmapshaper" R package (see p. 106)
09:30 – 10:00

Contributed Session ICORS 09:00 – 10:00

ICORS Contributed Paper Session: ICORS.CPS4 Thursday 30th
Chair: Peter Filzmoser, Room: Auditorium 1

A. Alfons, N. Y. Ateş, P. J. F. Groenen
Mediation analysis via the fast and robust boostrap (see p. 22)
09:00 – 09:20

Kanchan Mukherjee, Hang Liu
R-estimators and their bootstrapped version for GARCH models (see p. 74)
09:20 – 09:40

A. Mozaffari, S. Chenouri, G. Rice
Multiple Change Point Detection Based on Standard and Wild Rank-CUSUM Binary segmentation (see p. 40)
09:40 – 10:00

Poster Session 10:00 – 10:30

ICORS and LACSC Poster Presentations Thursday 30th

Camelo Andres, Granada Jose Rodrigo, Ramirez Carlos
Comparison of different techniques of classification for the discrimination of patients with Parkinson pathologies (see p. 35)

Terezinha K. A. Ribeiro, Silvia L. P. Ferrari
Robust estimation in beta regression via maximum L_q-likelihood (see p. 96)

M. L. Pappaterra, S. M. Ojeda
Analysis and comparison of similarity measures and indices for image quality assessment (see p. 90)

G. M. Britos, S. M. Ojeda
Robust estimation for spatial autoregressive processes based on bounded innovation propagation models (see p. 32)

E. A. S. Lizzi, T. C. Cassiano
Space-temporal modeling using DAG's and generalized additive models: case study of tuberculosis data (see p. 47)

J. Ascencio-Moreno, M. V. Hinojosa-Ramos, F. Vera, O. Ruiz-Barzola, M. I. Jiménez-Feijoó, M. P. Galindo-Villardón, M. Ramos-Barberán
MPCA vs. DS-PC performance comparison: a case study of fungicide efficacy evaluation for controlling black sigatoka on Ecuadorian banana plantations (see p. 64)

J. L. Cabrera, M. Andrade Bejarano, C. Grenier
Modeling of experimental designs in the presence of spatial correlation applied to agricultural experiments (see p. 34)

Marin Erisbey, M. Fernando, Ramirez Carlos
Coordination of algorithm for the Lasso and Ridge techniques (see p. 81)

Milena Machado, V. A. Reisen, Jane Meri Santos , P. Bondon
Time series models and principal component analysis techniques to estimate the impact of particulate matter on health and quality of life (see p. 78)

J. V. S. Magri, E. A. S. Lizzi
Bayesian spatio-temporal modeling: case study domestic violence data against women in Brazil (see p. 48)

E. J. Vargas, R. D. Guevara, M. P Bohorquez, S. I. Villamizar
Classification and descriptive analysis for multivariate functional data of IMF signals (see p. 110)

E. Vargas, M. Bohorquez, R. Guevara, L. Sarmiento
Classification for georeferenced functional brain signals (see p. 109)

A. V. Navarrete, R. D. Guevara, M. P Bohorquez, J. Bacca
Classification for geostatistical functional data using depth (see p. 86)

Luis Benites, Rocío Maehara, Victor H. Lachos, Heleno Bolfarine
Linear regression models using finite mixtures of skew heavy-tailed distributions (see p. 27)

J. A. Villasenor, E. Gonzalez-Estrada
A test for variance equality (see p. 115)

Pablo Flores, Jordi Ocaña
Implications of assumptions verification in the means comparison tests (see p. 55)

Coffee Break: 10:30 – 10:45

Plenary Session ICORS-LACSC 2019 10:45 – 11:45

Keynote Presentation: P3 Thursday 30th
Chair: David Muñoz Negron, Room: Auditorium 2

D. Eddelbuettel
Extending R with C++: Motivation, Examples and Context (see p. 51)

Contributed Session ICORS 11:45 – 12:25

ICORS Contributed Paper Session: ICORS.CPS5 Thursday 30th
Chair: David Kepplinger, Room: Auditorium 1

P. Filzmoser, Š. Brodinová, T. Ortner, C. Breitender, M. Rohm
Robust k-means-based clustering for high-dimensional data (see p. 54)
11:45 – 12:05

Yingying Zhang, Huixia Judy Wang, Zhongyi Zhu
Quantile-regression-based clustering for panel data (see p. 116)
12:05 – 12:25

Modelling Complex Data Structures with Applications in the Society and Environment II 11:45 – 12:45

LACSC Contributed Paper Session: LACSC.CPS10 Thursday 30th
Chair: Alex Jerves Cobo, Room: Auditorium 3

D. Morán-Zuloaga, D. Hernick, J. I. Valdez-Hernández, M. H. Cornejo, J. Cáceres, K. Morán, P. Hernick
Air masses origins and sources from southern Ecuador using HYSPLIT analysis (see p. 84)
11:45 – 12:05

M. Doktor, W. Kurz, C. Redenbach, P. Ruckdeschel, J.-P. Stockis
Robust Approaches to Non-Destructive Testing in Civil Engineering (see p. 50)
12:05 – 12:25

D. Arévalo-Avecillas, C. Padilla-Lozano
The Personality Domains and their relationship with the Transformational Leadership Style (see p. 25)
12:25 – 12:45

Recent Advances in Distribution Theory 11:45 – 12:45

LACSC Contributed Paper Session: LACSC.CPS11 Thursday 30th
Chair: Tomas Cipra, Room: Auditorium 2

Moreno Bevilacqua, Tarik Faouzi, Igor Kondrashuk, Emilio Porcu
Linnik probability densities via integration over Hankel contours (see p. 70)
11:45 – 12:05

Luis Benites, Rocío Maehara Filidor Vilca, Fernando Marmolejo-Ramos

Finite Mixture of Birnbaum-Saunders distributions using the k-bumps algorithm (see p. 80)
12:05 – 12:25

Christian E. Galarza, Larissa A. Matos, Victor H. Lachos
On moments of folded and truncated multivariate extended skew-normal distributions (see p. 56)
12:25 – 12:45

Lunch Break: 12:45 – 14:15

Invited Session ICORS 14:15 – 15:15

ICORS Invited Paper Session: ICORS.IPS3 Thursday 30th
Chair: Stefan Van Aelst, Room: Auditorium 2

Maria-Pia Victoria-Feser, Stéphane Guerrier, Mucyo Karemera, Samuel Orso
Efficient Bias Reduced Simulation-Based Estimators in High Dimensions (see p. 112)
14:15 – 14:45

D. Rodriguez, A. Muñoz
Robust estimation in partially nonlinear models (see p. 98)
14:45 – 15:15

Modern Approaches for Time Series and Regression Analysis 14:15 – 15:15

LACSC Contributed Paper Session: LACSC.CPS12 Thursday 30th
Chair: Nalini Ravishanker, Room: Auditorium 1

P. R. Prezotti, V. A. Reisen, P. Bondon, M. Ispány
The $PINAR(1, 1_S)$ model (see p. 94)
14:15 – 14:35

C. Heuchenne, A. Jacquemin
Lorenz regression for single-index models with monotone link functions (see p. 63)
14:35 – 14:55

Carlos Trucíos, Joao H. G. Mazzeu, Mauricio Zevallos, Luiz K. Hotta, Pedro L. Valls Pereira, Marc Hallin
A general dynamic factor approach to forecast conditional covariance matrices in high-dimensional data (see p. 107)
14:55 – 15:15

Recent Topics in Functional Data Analysis: Analysis of Variance, Spatial Statistics and Quality Control 15:15 – 16:45

LACSC Invited Paper Session: LACSC.IPS4 Thursday 30th
Chair: Martha Bohórquez-Castañeda, Room: Auditorium 2

Jeimy Aristizabal-Rodríguez, Ramón Giraldo, Jorge Mateu
Analysis of variance for spatially correlated functional data: application to brain data (see p. 58)
15:15 – 15:45

Ruben Dario Guevara Gonzalez, Tania Alejandra Lopez Guasca, Jose Alberto Vargas Navas
Evaluation of process capability in nonlinear profiles (see p. 61)
15:45 – 16:15

F. J. Rodríguez-Cortés , E. Romano , J. A. González , J. Mateu
Spatial clustering based on the pair correlation LISA functions: A functional approach (see p. 99)
16:15 – 16:45

Conference Dinner: 18:00 – 21:30

Tickets still available at the Registration desk

Buses leave from ESPOL to Restaurante Casa Julián

Friday, 31 May 2019

Robust Methods: Advances and Applications 08:30 – 10:00

ICORS-LACSC Invited Paper Session: ICORSLACSC.IPS Friday 31th
Chair: Marco Avella, Room: Auditorium 2

V. Lourenço, J. Ogutu, H.-P. Piepho
Robust estimation in plant breeding: evaluation using simulation and empirical data (see p. 75)
08:30 – 09:00

A. Posekany
Gaining robustness and detecting outliers applying Mixtures of Gaussian and heavy-tailed distributions in Bayesian inference (see p. 91)
09:00 – 09:30

G. Boente, D. Rodriguez, P. Vena
Robust B-splines estimators in partly linear regression under monotony constraints (see p. 29)
09:30 – 10:00

Plenary Session ICORS-LACSC 2019 10:00 – 11:00

Keynote Presentation: P4 Friday 31th
Chair: Peter Rousseeuw, Room: Auditorium 2

A. M. Bianco
A robust approach to ROC curves with covariates (see p. 28)

Coffee Break: 11:00 – 11:15

Invited Session ICORS 11:15 – 11:45

ICORS Invited Paper Session: ICORS.IPS4 Friday 31th
Chair: Daniela Rodríguez, Room: Auditorium 2

L. Lakshnaman, E. Smucler, V. Yohai, R. H. Zamar
An alternative approach for testing hyphotesis (see p. 117)
11:15 – 11:45

Recent Advances in Statistical Computing 11:15 – 12:45

LACSC Invited Paper Session: LACSC.IPS5 Friday 31th
Chair: Paulo Canas, Room: Auditorium 1

F. Louzada-Neto
Zero-adjusted cure rate regression survival models (see p. 76)
11:15 – 11:45

M. T. A. Cordeiro, Jesús E. García, V. A. González-López, S. L. M. Londoño
Partition Markov Model for Multiple Processes (see p. 60)
11:45 – 12:15

M. Bohorquez, R. Guevara, L. Sarmiento, J. Mateu
Classification analysis for spatial functional random fields (see p. 30)
12:15 – 12:45

Closing ceremony 12:45 – 13:00

Auditorium 2

Abstracts ICORS-LACSC 2019

Mediation analysis via the fast and robust boostrap

A. Alfons[a], N. Y. Ateş[b,c] and P. J. F. Groenen[a]

[a] *Erasmus University Rotterdam,* [b] *Bilkent University,* [c] *Tilburg University*

Mediation analysis is one of the most widely used statistical techniques in the social and behavioral sciences. The mediation model in its simplest form allows to study how an independent variable (X) affects a dependent variable (Y) through an intervening variable that is called a mediator (M). Most often, such an analysis is performed through a pair of regression models $M = i_1 + aX + e_1$ and $Y = i_2 + bM + cX + e_2$, in which case the indirect effect of X on Y through M can be computed as the product of coefficients ab. In the social science literature, significance of the indirect effect is typically tested via a bootstrap test based on ordinary least squares (OLS) regressions [1]. Yet this test is highly sensitive to model deviations such as outliers or heavy tails, which poses a serious threat to empirical testing of theory about mediation mechanisms.

By estimating the above regression models via the MM-estimator [3] and applying the fast and robust bootstrap [2], we obtain a robust procedure for mediation analysis. Simulation results and empirical examples illustrate that this procedure yields reliable results for estimating the effect size and assessing its significance, even under deviations from the usual normality assumptions.

The proposed procedure is freely available in the R package robmed. The package funcionality is not limited to simple mediation models, and includes mediation models with multiple mediators as well as control variables. Furthermore, the standard bootstrap test and other proposals are included in the package as well.

Keywords: Mediation analysis, linear regression, bootstrap.

References

[1] K. J. Preacher, and A. F. Hayes (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Bahavior Research Methods, Instruments, & Computers*, **36**(4), 717–731.

[2] M. Salibián-Barrera, and R. H. Zamar (2002). Bootstrapping robust estimates of regression. *Annals of Statistics*, **30**(2), 556–582.

[3] V. J. Yohai (1987). High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics*, **15**(20), 642–656.

Bayesian Inference for the estimation of the touristic expenditure in Honduras

A. Alonzo[a], E. Avila[b], and A. Matamoros[c]

Universidad Nacional Autonoma de Honduras

Tourism is a social, cultural and economic phenomenon that has effects on the economy of a country, In Honduras, the contribution of tourism to the economy is measured using the tourism expenditure, and is estimated by applying two different surveys, the entrance survey that counts the total amount of tourists visiting the country, and has a section to determine the expected touristic expense. And the exit survey survey, which extracts the generated tourism expenditure.

So far, the exit survey is the only tool to estimate tourist spending, and the information from the entrance survey is omitted. In this study, we tested whether the entrance survey can be used to obtain better estimations. Many models were proposed, including a Bayesian model with hierarchical priors, where this priors were constructed from the entrance survey, and the posterioris were sampled using a Hamiltonian Monte Carlo with the NUTS algorithm (*No-U-Turn -Sampling*).

Of all the proposed models, the hierarchical and classic models obtained the best results, where the estimates were similar to each other, this is due to the large sample size in the exit survey. Although, the exit survey is enough to have good estimates, the hierarchical models are useful tools for estimating in specific touristic places or in short periods of time, were the sample sizes are not that large .

Keywords: Touristic expenditure, bayesian inference, Hamiltonian monte carlo.

References

[1] Y. Y. Sunm, D. J. Stynes (2006). A note on estimating visitor spending on a per-day/night basus. *Tourism Managment*, **27**, 721–725

[2] A. G. Assaf, H. Oh, and M. Tsionas (2016). Bayesian Approach for the measurement of tourism performance: A case of stochastic frontier models. *Journal of travel research*, 1–15.

[3] A. Gelman (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**(3), 515–534.

[4] M. D. Hoffman, A. Gelman (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of machine learnign research*, **15**, 1593–1629.

[5] M. Betancourt, and M. Girolami (2013). Hamiltonian Monte Carlo for hierarchical models. Department of statistical science, University College London.

Comparison of Optimization Metaheuristics Based on Neighborhoods for Clustering Binary Data

L. E. Amaya[a] and J. Trejos[b]

[a] *University of Costa Rica, Campus Guanacaste,* [b] *CIMPA,*

The methods of clustering consist of strategies which look for determining groups, under the generalized principle that objects (individuals) belonging to the same group, present characteristics of greater similarity to each other (with regard to some previously selected criterion), compared with the individuals that were assigned to other groups. Since the problem is NP-hard and methods of partitioning generally find local optima of the criteria to optimize, we have sought to improve them by using combinatorial optimization heuristics, techniques such as simulated annealing, tabu search, threshold accepting, genetic algorithms or ant colonies, among others, these methods have provided superior results to those obtained with "classical" methods such as k-means, dynamic clouds or hierarchical classification. In the present work, we focus on the classification in the presence of binary data by applying simulated annealing, tabu search and threshold accepting. The implemented algorithms were based on the concept of neighborhoods, this is, given a partition, we define a neighborhood by the change in the class assignment of a single object.

In this work, using two aggregation criteria of the Sum and L1, chosen among other studied, have the particularity of being summative: in the Sum, it is the simple sum between objects which belong to the same class, over all the classes, and the case of the criterion L1 it is defined a central or centroid object for each class as a vector of medians, and the criterion is the sum of distances L1 of objects to their centroid. The obtained results are very promising; they have been compared with those obtained with the k-means method and the ascending hierarchical classification, on real and simulated data, these last, with its different characteristics were generated by a Monte Carlo experiment. In no instance did classical methods achieve better results than those found with our methods.

Keywords: Clustering, simulated annealing aggregation criteria.

References

[1] G. Dueck, T. Scheuer (1990). Threshold Accepting: a general purpose optimization algorithm appearing superior to Simulated Annealing. *Journal of Computational Physics*, **90**, 161–175.

[2] F. Glover (1989). Tabu search - Part I. *ORSA J. Comput*, **1**, 190–206.

[3] S. Kirkpatrick, D. Gelatt, M. P. Vecchi (1983). Optimization by simulated annealing. *Science*, **220**, 671–680.

The Personality Domains and their relationship with the Transformational Leadership Style

D. Arévalo-Avecillas and C. Padilla-Lozano

Universidad Católica de Santiago de Guayaquil, km 0.5 vía Carlos Julio Arosemena, Guayaquil, Ecuador.

The purpose of the study was to evaluate the impact of personality domains and work experience in the transformational leadership style, providing new evidence of the causal relationship between both constructs. The research had a quantitative approach with a non-experimental design, cross - sectional and correlational - causal scope. A survey was carried out to 368 professionals who study Master of Business Administration programs in Ecuador. The instrument called Revised Personality Inventory NEO-PI-R was used to measure the five domains of personality: (a) extraversion, (b) agreeableness, (c) conscientiousness, (d) neuroticism, and (e) openness to experience. On the other hand, the Multifactor Leadership Questionnaire (MLQ) was the instrument used to measure the transformational leadership style. To analyze the data, it was used correlation analysis and a multivariate regression model. The results highlight that extraversion and conscientiousness domains proved to be the most important in the projection of transformational leadership style.

Keywords: personality, leadership, transformational

Privacy-preserving parametric inference: a case for robust statistics

M. Avella-Medina

Department of Statistics, Columbia University

Differential privacy is a cryptographically-motivated definition of privacy that has become a very active field of research over the last decade in theoretical computer science and machine learning. In this paradigm we assume there is a trusted curator who holds the data of individuals in a database and the goal of privacy is to simultaneously protect individual data while allowing statistical analysis of the database as a whole. In this setting we introduce a general framework for parametric inference with differential privacy guarantees. We first obtain differentially private estimators based on bounded influence M-estimators by leveraging their gross error sensitivity in the calibration of a noise term added to them in order to ensure privacy. We then we show how a similar construction can also be applied to construct differentially private test statistics analogous to the Wald, score and likelihood ratio tests. We provide statistical guarantees for all our proposals via an asymptotic analysis. An interesting consequence of our results is to further clarify the connection between differential privacy and robust statistics. In particular we demonstrate that differential privacy is a weaker requirement than infinitesimal robustness and show that robust M-estimators can be easily randomized in order to guarantee both differential privacy and robustness towards the presence of contaminated data. We illustrate our results both on simulated and real data.

Keywords: Differential privacy, M-estimators, Influence function, Robust tests, Linear regression

References

[1] M. Avella-Medina (2018). Privacy-preserving parametric inference: a case for robust statistics. *(manuscript)*.

[2] C. Dwork, and A. Roth (2014). The algorithmic foundations of differential privacy. *Foundations and Trends ® in Theoretical Computer Science*, **9**(3–4), 211–407.

Poster: May 30, 10:00 - 10:30, Posters Room

Linear regression models using finite mixtures of skew heavy-tailed distributions

Luis Benites^{a}, Rocío Maehara^{b} Victor H. Lachos^{c} and Heleno Bolfarine^{d}

^{a}Pontificia Universidad Católica del Perú, ^{b}Universidad del Pacífico, Perú, ^{c}University of Connecticut, ^{d}Universidade de São Paulo, Brazil

In this paper we extend the regression model based on the assumption that the error term follows a mixture of normals, by considering a finite mixture of scale mixtures of skew-normal distributions, a rich class of distributions that contains the skew-normal, skew-t, skew-slash and skew-contaminated normal distributions as proper elements. This approach allows us to model data with great flexibility, simultaneously accommodating multimodality, skewness and heavy tails. We develop a simple EM-type algorithm to perform maximum likelihood inference of the parameters of the proposed model with closed-form expressions for both E- and M-steps. Furthermore, the empirical information matrix is derived analytically to account for standard errors and a bootstrap procedure is implemented to test the number of components in the mixture. The practical utility of the new method is illustrated with the analysis of a real dataset and several simulation studies. The proposed algorithm and methods are implemented in the R package `FMsmsnReg`.

Keywords: ECME algorithm; Mixture model, Non-normal error distribution, Scale mixtures of skew-normal distributions

References

[1] F. Bartolucci, L. Scaccia (2005). The use of mixtures for dealing with non-normal regression errors. *Computational Statistics & Data Analysis* **28**(4), 821–834.

[2] L. Benites, R. Maehara, V. H. Lachos (2016). FMsmsnReg: Regression Models with Finite Mixtures of Skew Heavy-Tailed Errors. R package version 1.0. http://CRAN.R-project.org/package=FMsmsnReg

[3] M. D. Branco, D. K. Dey (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis* **79**, 99–113.

[4] A. Dempster, N. Laird, D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, **39**, 1–38.

A robust approach to ROC curves with covariates

A. M. Bianco

Instituto de Cálculo, Universidad de Buenos Aires and CONICET

Receiver Operating Characteristic (ROC) curves are a useful tool to measure the discriminating power of a continuous variable. They usually quantify the accuracy of a pharmaceutical or medical test to distinguish between two conditions or classes. ROC curves can also be extended to other general statistical situations such as classification or discrimination, where we typically have a set of individuals or items assigned to one of two classes on the basis of disposable information of that individual. Assignations are not perfect and may lead to classification errors. At this point, ROC curves become an interesting strategy either to evaluate the quality of a given assignment rule or to compare two available procedures.

In practical situations, the discriminatory effectiveness of the marker or test under study may be affected by several factors. When for each individual there is additional information contained in registered covariates, it is sensible to include them in the ROC analysis.

This talk aims to show the instability of the conditional ROC curve in presence of outliers and also to provide robust estimators for it when covariates are available. We focus on a semiparametric approach which, on one side, fits a location-scale regression model to the diagnostic variable and, on the other, considers empirical estimators of the regression residuals distributions. To this end, we combine robust parametric estimators with weighted empirical distribution estimators based on an adaptive procedure that downweights outliers.

We will discuss some aspects concerning consistency. Through a Monte Carlo study we compare the performance of the proposed estimators with the classical ones both, in clean and contaminated samples. The simulation results show that our robust procedure yields reliable results under different contaminated scenarios.

Keywords: ROC curves, Covariates, Robustness.

Robust B-splines estimators in partly linear regression under monotony constraints

G. Boente, D. Rodriguez and P. Vena

Universidad de Buenos Aires and CONICET, Argentina

In this talk, we will consider the situation in which the observations follow an isotonic generalized partly linear model. Under this model, the mean of the responses is modelled, through a link function, linearly on some covariates and nonparametrically on an univariate regressor in such a way that the nonparametric component is assumed to be a monotone function. A class of robust estimates for the monotone nonparametric component and for the regression parameter will be defined. The robust estimators are based on a spline approach combined with a score function which bounds large values of the deviance. Results regarding the asymptotic behaviour of the proposal will be presented. Through a Monte Carlo study and a real data set we will illustrate the advantages of the proposal. Finally, some extensions to the functional setting will be discussed.

Keywords: $B-$splines, Isotonic regression, Robust Estimation

Classification analysis for spatial functional random fields

M. Bohorquez[a], R. Guevara[a], L. Sarmiento[b] and J. Mateu[c]

[a] *Universidad Nacional de Colombia,*[b] *Universidad Pedagógica Nacional,*[c] *Universidad Jaume I*

We propose a methodology for classification of functional data taking into account its spatial autocorrelation. We use the representation of each function in terms of its empirical functional principal components, and the spatial covariance model among the associated score vectors, that are scalar random fields. This model leads to a structured and positive-definite matrix, that involves the autocovariance and cross-covariance between the score vectors. The group-covariance matrices are allowed to be unequal, and each matrix is modeled using the respective training-data group. The Mahalanobis distance is computed between the score vectors obtained for each realization in each group and the score vectors of the new observation. Finally, the new observation is assigned to the group with the smallest Mahalanobis distance. Our proposal is applied on meteorological data and on brain signals from the language area.

Keywords: Spatial functional data, Classification, Mahalanobis distance

References

[1] M. Bohorquez, R. Giraldo and J. Mateu. (2017) Multivariate functional random fields: prediction and optimal sampling. *Stochastic Environmental Research and Risk Assessment*, **31**(1), 53–70.

[2] A.C. Rencher and W.F. Christensen. (2012). *Multivariate analysis: Methods and applications.* John Wiley & Sons, NY.

[3] P. Galeano, J. Esdras and R. Lillo. (2015). The Mahalanobis distance for functional data with applications to classification. *Technometrics*, **57**(2), 281–291.

A new regression model for positive random variables with skewed and long tails

Marcelo Bourguignon[a], Manoel Santos-Neto[b] and Mário de Castro[c]

[a] *Universidade Federal do Rio Grande do Norte,* [b] *Universidade Federal de Campina Grande,* [c] *Universidade de São Paulo*

The main aim of this paper is to propose a regression model that is tailored for situations where the response variable is measured continuously on the positive real line that is in several aspects, like the generalized linear models. In particular, the proposed model is based on the assumption that the response is beta prime (BP) distributed. We considered a new parameterization of the BP distribution in terms of the mean and precision parameters. Under this parameterization, we propose a regression model, and we allow a regression structure for the mean and precision parameters by considering the mean and precision structure separately. The variance function of the proposed model assumes a quadratic form. The proposed regression model is convenient for modeling asymmetric data, and it is an alternative to the generalized linear models when the data presents skewness. Inference, diagnostic and selection tools for the proposed class of models will be presented. We summarize below the main contributions and advantages of the proposed BP model over the popular gamma model. With these contributions below, we provide a complete tool for modelling asymmetric data based on our BP regression.

- We allow a regression structure on the precision parameter; in a manner similar to the way the generalized linear models with dispersion covariates extend the generalized linear models.
- The variance function of proposed model assumes a quadratic form similar to the gamma distribution. However, the variance function of proposed model is larger than the variance function of gamma distribution, which may be more appropriate in certain practical situations.
- The BP hazard rate function can have an upside-down bathtub or increasing depending on the parameter values. Most classical two-parameter distributions such as Weibull and gamma distributions have monotone hazard rate functions.
- The skewness and kurtosis of the BP distribution can be much larger than the of the gamma distribution.

Keywords: Beta prime distribution, Data Analysis, Regression models.

Robust estimation for spatial autoregressive processes based on bounded innovation propagation models

G. M. Britos and S. M. Ojeda

Universidad Nacional de Córdoba - Facultad de Matemática, Astronomía, Física y Computación

Robust methods have been a successful approach for dealing with contamination and noise in the context of spatial statistics and, in particular, in image processing ([1], [2]). In this paper, we introduce a new robust method for spatial autoregressive models based on the work of [3] for time series. Our method, called BMM-2D, relies on representing a two-dimensional autoregressive process with an auxiliary model to attenuate the effect of contamination (outliers). We compare the performance of our method with existing robust estimators and the least squares estimator via a comprehensive Monte Carlo simulation study, which considers different levels of replacement contamination and window sizes. The results show that the new estimator is superior to the other estimators, both in accuracy and precision. An application to image filtering highlights the findings and illustrates how the estimator works in practical applications.

Keywords: AR-2D models, Robust estimators, Image processing, Spatial models

References

[1] O. Bustos, S. Ojeda, R. Vallejos (2009). Spatial ARMA models and its applications to image filtering. *Brazilian Journal of Probability and Statistics*, 141–165.

[2] X. Guyon (1995). *Random fields on a network: modeling, statistics, and applications.* Springer Science and Business Media.

[3] N. Muler, D. Peña, V. Yohai (2009). Robust estimation for ARMA models. *The Annals of Statistics*, **37**(2), 816–840.

A Two-Stage Stochastic Formulation for The Simple Plan Location Problem with Order

X. Cabezas[a,b] and S. García[b]

[a] *Escuela Superior Politécnica del Litoral,* [b] *The University of Edinburgh*

The simple plant location problem with order (SPLPO) is a variant of the simple plant location problem (SPLP) where the customers have preferences on the facilities that will serve them. The problem can be formulated as a mixed integer linear program (MILP) and some results about its strength can be found in the literature. In this paper we present a two-stage stochastic formulation for SPLPO where the preferences given by the costumers are considered random vectors. Furthermore, we carried out an experimental study to solve some instances by using a semi-Lagrangean relaxation methodology that exploits this structure.

Keywords: Optimization, Linear model.

Modeling of experimental designs in the presence of spatial correlation applied to agricultural experiments

J. L. Cabrera[a], M. Andrade Bejarano[a] and C. Grenier[b]

[a] *Universidad del Valle,* [b] *CIRAD*

Effect of spatial heterogeneity between plots has a great influence on the estimation of observations made in field experiments. In order to reduce this problem, the randomization of the experiments is used. But in some cases this is insufficient to neutralize the correlation effects between neighboring plots. One way to control this problem is to model the spatial correlation structure through the matrix of variances and covariances of errors. The objective of this work is to evaluate and determine the effect of the spatial correlation on the estimation in the modeling of the experimental design. For this purpose, the modeling of some trials carried out by CIAT's rice improvement program is carried out. In addition, simulated experiments are analyzed, based on a completely random block and augmented blocks design with and without presence of spatial correlation. Mixed linear models were used for the modeling as fixed and random effects were considered in the treatments and design structure. Finally, it can be concluded that the modeling of the error variances and covariances matrix when spatially correlated errors are considered presented better estimates compared to the model that assumes independence in errors.

Keywords: Modeling, Experimental Design, Spatial Correlation.

Comparison of different techniques of classification for the discrimination of patients with Parkinson pathologies

Camelo Andres[a], Granada Jose Rodrigo[b] and Ramirez Carlos[c]

[a] *Universidad Tecnologica de Pereira,* [b] *Universidad Tecnologica de Pereira,* [c] *Universidad Tecnologica de Pereira*

This document presents a brief description of the application of the technique of vector support machines (SVM) in order to classify patients with pathologies of Parkinson's disease from voice samples. The problem of linear classification and the case of non-linear classification is addressed, for which the Kernel functions will be used in order to transform the input space into a space of greater dimensionality where it can be classified by means of a hyperplane .

The problems of classification in recognition of patterns, have taken much consideration in recent years given the computational advance that is had for the execution of different techniques that allow to give a solution to this task. Its areas of application range from communications: voice recognition, image classification, biometric identification; analysis of bilogical samples for the detection of diseases; SPAM (junk mail recognition). The basic idea of the classification is to determine to which set of categories a new observation belongs from the analysis of a set of data or observations of which its membership is already known, that is, there is a training based on past data. This training can occur in two ways, supervised and unsupervised, however, the classification is considered to be machine learning as a supervised problem since it is based on a set of correctly identified observations, while unsupervised learning techniques do reference to groupings based on similarities of characteristics or the definition of some inherent distance or pattern.

In this document, the problem of classification is addressed by means of the algorithm of vector support machines, which is a highly used algorithm developed by Vapnik (1995) [1]. The operation of the SVM is based on the determination of a hyperplane that separates the classes by the greatest possible distance.

Keywords: Kernels,Principal Components Analysis, Hyperplane.

References

[1] C. Cortes, and V. Vapnik (1995). Support-vector networks. *Machine learning*, **20**(3), 273–297.

[2] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Biomedical engineering online*, **6**(1), 23.

[3] W. S. McCulloch, and W. Pitts (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, **5**(4), 115–133.

Robust and randomized singular spectrum analysis

P. C. Rodrigues

Department of Statistics, Federal University of Bahia, Salvador, Brazil

Singular spectrum analysis (SSA) is a non-parametric method for time series analysis and forecasting that incorporates elements of classical time series analysis, multivariate statistics, multivariate geometry, dynamical systems and signal processing. Although this technique has shown to be advantageous over traditional model based methods, in particular, one of the steps of the SSA algorithm, which refers to the singular value decomposition (SVD) of the trajectory matrix, is highly sensitive to data contamination and also time consuming.

In this talk we will present (i) the randomized SSA which is an alternative to SSA for long time series without losing the quality of the analysis; and (ii) a robust SSA algorithm, where a robust SVD procedure replaces the least-squares based SVD in the original SSA procedure, in order to reduce the effect of data contamination by outlying observations.

The SSA and the randomized SSA are compared in terms of quality of the model fit and forecasting accuracy, and computational time, via Monte Carlo simulations and real data about the daily prices of five of the major world commodities. The SSA and the robust SSA are compared in terms of the quality of the model fit via Monte Carlo simulations that contemplate both clean and noisy/contaminated time series, and considering a real data application.

Keywords: Tome series analysis, Robust singular value decomposition, Randomized singular value decomposition.

References

[1] P. C. Rodrigues, V. M. Lourenço, and R. Mahmoudvand (2018a). A robust approach to singular spectrum analysis. *Quality and Reliability Engineering International*, **34**, 1437–1447.

[2] P. C. Rodrigues, P.G.S.E. Tuy, and R. Mahmoudvand (2018b). Randomized singular spectrum analysis for long time series. *Journal of Statistical Computation and Simulation*, **88**, 1921–1935.

Bayesian Estimation of the Limiting Availability in a Repairable Coherent System

R. Carvajal-Schiaffino, V. Cid-Ossandon and V. H. Salinas

Departamento de Matemática y Ciencia de la Computación, Universidad de Santiago de Chile

This work presents a Bayesian approach for estimating the limiting availability of a repairable coherent system. A Bayesian analysis is developed considering an informative prior and a less informative prior distribution, for the failure and repairable times, respectively. Simulations are presented to study the performance of the Bayesian solutions, using distributed algorithms.

Keywords: Reliability distributions, MCMC Methods, Distributed algorithms.

References

[1] S. T. Román, J. S. Romeo and V. H. Salinas (2014). Bayesian Estimation of the Limiting Availability in the Presence of Right-Censored Data. *METRON* **72**(3), 247–267.

[2] V. H. Salinas., C. A. Vásquez, and J. S. Romeo (2019). Bayesian Estimation of the Limiting Availability in a Repairable One-Unit System. *Revista Colombiana de Estadística* **42**(1), 223–143.

[3] J. Bruck, D. Dolev, C. Ho, M- Rosu, and R. Strong. (1997). Efficient Message Passing Interface (MPI) for Parallel Computing on Clusters of Workstations. *Journal of Parallel and Distributed Computing* **40**(1), 19–34.

Implementation of a Parallel Algorithm with Shared/Private Memory for Parametric Booststrap

R. Carvajal-Schiaffino[a], F. Novoa-Muñoz[b] and C. González-Aguero[b]

[a] *Departamento de Matemática y Ciencia de la Computación - Universidad de Santiago de Chile,* [b] *Departamento de Estadística - Universidad del Bío-Bío, Chile*

In many cases the properties studied by a statistic describe the asymptotic behavior of the distribution of a statistical test, this is what is known as the study of a test for large samples, that is, virtually infinite samples. But the behavior of the test for small (finite) samples is unknown, this is where the bootstrap method, in particular the parametric bootstrap, plays a fundamental role, since it allows to approximate the unknown distribution of the statistic involved.

From a computational point of view the parametric bootstrap is a very expensive technique. We present a comparison between parallel implementations executed on a cluster of computers each one with several processors.

Keywords: Parametric Bootstrap, Parallel Algorithms.

References

[1] F. Novoa-Muñoz, and M. D. Jiménez-Gamero (2014). Testing for the Bivariate Poisson Distribution. *Metrika* **77**(6), 771–93.

[2] F. Novoa-Muñoz, and M. D. Jiménez-Gamero (2016). A Goodness-of-Fit Test for the Multivariate Poisson Distribution. *SORT* **40**, 113–38.

[3] P. Pacheco (2011). *An Introduction to Parallel Programming.* Morgan Kaufmann.

Predicting the public institutional budget: an application using shinydashboard

O. Centeno-Mora

Department of Statistics, University of Costa Rica.

In the framework of public budget approval, it is essential to understand, for each of the government institutions, the monetary forecast for the upcoming fiscal year, both in terms of expenditures and revenues. This paper proposes the use of a time series model for the approval of the institutional budget through a budgetary projection, as well as the application of a dashboard in order to handle all the information concerning the public institutional budget. The input information was supplied by the System of Planning and Budgeting of the Government Accountability Office of Costa Rica. ARIMA time series models are used to forecast monetary demand. Measures of goodness and adjustment are taken into account, as well as the projection of the budget through confidence intervals. This information is displayed in a web application using R Studio and shinydashboard. The results of the application allow to obtain the forecast of the monetary demand for any institution concerned. A system of indicators is created in form of alerts to identify where more attention should be paid. The application facilitates and supports the approval process for public institutional budget.

Keywords: time series forecasting, statistical computing, shinydashboard.

Multiple Change Point Detection Based on Standard and Wild Rank-CUSUM Binary segmentation

A. Mozaffari, S. Chenouri and G. Rice

University of Waterloo

In this paper, two non-parametric multiple change point detection methods are proposed based on standard and wild binary segmentation algorithms. The methods use rank-CUSUM statistic for the detection of change points in each segment. The asymptotic results pertaining to the consistency of the estimated change points as well as the number of detected change points are presented. Thereafter, a comparative analysis is done by adopting some well-known multiple change point detection methods from literature and using finite-sample Monte-Carlo simulation as well as real data. The obtained results endorse the theoretical findings and show the competence of the proposed methods.

Keywords: Multiple change point detection, Wild binary segmentation, Rank-CUSUM.

Q-means, a new quantile method to initial seed value selection in K-means algorithm for clustering.

M. Choez

ESPOL Polytechnic University

Clustering techniques have been applied in data analysis in a wide variety of fields like bioinformatic, medicine and computer science. One of the most popular and fastest clustering techniques used in these fields is K-means [1].

K-means, in few words is a clustering algorithm to classify data points based on their attributes into k predefined number of clusters. The algorithm works in two separate phases. The first phase selects k centroids from a dataset at random. The second phase assigns each data point to the nearest centroid. Euclidean distance is generally used to determine the distance between data points and the centroids. The algorithm is highly sensitive to the initial seed value selection of centroids [2]. This is the biggest disadvantage of K-means because the resulting clusters heavily depends on the selection of initial centroids.

The present study shows a new initialization method that guarantees unique clustering results instead of a random method that sometimes produces poorer results for several runnings on the same dataset. The main idea of the proposed algorithm Q-means was to use measurement of position quantiles to find out the initial centroids of the k-means. The new algorithm for initial selection called Q-means was proved comparing its performance against to K-means algorithm. The comparisons were done based on the following criteria iterations, sum squared errors, entropy and efficiency. The obtained results in this work proved that Q-means algorithm had better performance when was compared with K-means algorithm which used the random selection of the centroids. In the most of the comparison criteria Q-means was better than K-means, except in the the complexity running time when the dataset size is too large.

Keywords: K-means, Clustering, Data mining.

References

[1] N. Dhanachandra, K. Manglem, and Y. J. Chanu (2015). Image segmentation using $k-$means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, **54**, 764–771.

[2] M. E. Celebi, H. A. Kingravi, and P. A. Vela (2013). A comparative study of efficient initialization methods for the $k-$means clustering algorithm. *Expert Systems with Applications*, **40**(1), 200–210.

Estimation of volatility models by robustified recursive procedures

T. Cipra and R. Hendrych

Charles University, Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics, Sokolovská 83, 186 75 Prague, Czech Republic

The robust recursive algorithms for the parameter estimation and the volatility prediction are suggested. It seems to be useful for various financial time series, in particular for (high frequency) log returns contaminated by additive outliers. The proposed procedure can be effective in the risk control and regulation when the prediction of volatility is the main concern since the method is capable to distinguish and correct outlaid bursts of volatility. Another possible application concerns the dynamic approach to IBNR reserves in insurance companies.

Financial time series typically exhibit significant kurtosis and volatility clustering (the assets are usually stocks or stock indices or currencies). The GARCH models are applied commonly in such a context to model these typical properties with the aim to describe dynamics of conditional variances and forecast financial volatility. However, when fitted to real time series the residuals of the estimated models have frequently excess kurtosis explainable by the presence of outliers which are not captured by the given models (on the other hand, some authors argue that extreme observations are not outliers and they should be incorporated into the model).

The parameters of volatility models are routinely estimated by the (conditional) maximum likelihood but they are rarely calibrated recursively. Nevertheless, recursive estimates performed using recursive algorithms are undoubtedly advantageous (they are effective in terms of memory storage and computational complexity). This efficiency can be employed just in the framework of (high-frequency) financial time series data. Alternatively, it is possible to adopt these methods to monitor or forecast volatility on-line, to evaluate risk measures (e.g. Value at Risk or Expected Shortfall), to detect faults, to check model stability including detection of structural changes, etc. However, due to the previous arguments, the recursive estimation schemes should be robust to outliers. Therefore, the authors robustify their recursive prediction error estimation scheme which is suitable just in this context. The contribution documents it by means of simulations and various real data applications.

Keywords: GARCH model, outlier, robust recursive estimation.

References

[1] T. Cipra, and R. Hendrych (2018). Robust recursive estimation of GARCH models. *Kybernetika*, **54**, 1138–1155.

Variable Selection in Functional Linear Cox Regression Model via Regularization Methods Applied to Clinical Data

Julian A. Collazos[a], Adriano Z. Zambom[b] and Ronaldo Dias[c]

[a] *New Granada Military University, Colombia,* [b] *California State University, Northridge, USA,* [c] *Campinas State University, Brazil*

In this work is presented an alternative methodology to perform variable selection in functional linear Cox regression model to sparse functional covariates. In medical studies, longitudinal data is frequently recorded but often the profiles are measured in a different irregular and sparse set of time points across individuals (sparse longitudinal data). To deal with this data without loss of information, we use a mixed effect model framework to obtain the MLE of functional principal components by the EM algorithm that consider all information of the individuals in order to reconstruct the full profile. To perform variable selection of the sparse functional covariates, we use a group descent algorithm to penalized Cox regression on the scores of functional principal components using group nonconvex penalties such as SCAD and MCP which have several advantages over group lasso penalty. The proposed metholodogy selects the functional covariates that are related to the hazard function for the Cox regression model. An application to clinical data related to primary biliary cirrhosis data is introduced to illustrate the performance of the proposed methodology.

Keywords: Sparse functional covariate, functional principal components analysis, group penalized regression.

References

[1] D. R. Cox (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**(2), 187–202.

[2] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of statistical software*, **39**(5), 1.

[3] D. Kong, J. G. Ibrahim, E. Lee, and H. Zhu (2018). FLCRM: Functional linear cox regression model. *Biometrics*, **74**(1), 109–117.

[4] J. E. Gellar, E. Colantuoni, D. M. Needham, and C. M. Crainiceanu (2015). Cox regression models with functional covariates for survival data. *Statistical modelling*, **15**(3), 256–278.

[5] P. Breheny, and J. Huang (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and computing*, **25**(2), 173–187.

Efficient Permutation Inference by Computing Parallel Algorithms to Support Comparative Neuroanatomy

L. Corain[a], R. Carvajal-Schiaffino[b], J-M. Graïc[c], E. Grisan[d,e], R. Luisetto[f], L. Salmaso[a] and A. Peruffo[c]

[a] *Department of Management and Engineering, University of Padova, Italy*
[b] *Department of Mathematics and Computer Science, University of Santiago de Chile, Chile*
[c] *Department of Comparative Biomedicine and Food Science, University of Padova, Italy*
[d] *Department of Biomedical Engineering, King's College London, London, UK*
[e] *Department of Information Engineering, University of Padova, Italy*
[f] *Department of Surgery, Oncology and Gastroenterology, University of Padova, Italy*

The goal of the present paper is designing and evaluating novel efficient parallel-computing algorithms to optimize a new type of robust multivariate testing and ranking methods to support cell shape analysis in neuroanatomy studies. Our background refers to quantify fine structural differences in the brain cytoarchitecture, by analyzing cell morphology in digitalized Nissl stained sections as a function of factors of interest such as sex, age or pathologies. Towards this goal, we propose a novel nonparametric inferential multi-aspect testing and ranking approach as an extension of a well-established permutation and combination-based testing methodology. The main advantage of our proposal is in that it allows to analyse the set of cell morphometric descriptors, either from the univariate or the multivariate point of view, separately for each distributional aspect, i.e. mean (inference on location), and variance (inference on scatter). It is worth noting that, as soon as the number of populations to be compared and/or the set of different brain regions/layers to be considered increases to more that 4-5, the underlying computational complexity becomes dramatically overwhelming. Therefore, the support of efficient computing parallel algorithms makes the practical application of such permutation-based methods much more feasible in real-life biomedical applications. Finally, we applied the proposed procedure to two comparative neuroanatomy studies; the first one was aimed at quantifying structural differences in the brain cytoarchitecture of three sex-related populations, i.e. bovine male, female and natural intersex; the second study aimed at quantify structural differences in the brain cytoarchitecture of a mouse model for the *systemic lupus erythematosus* disease.

Keywords: p-value combination, multivariate ranking, union-intersection principle.

References

[1] R. Arboretti Giancristofaro, S. Bonnini, L. Corain and L. Salmaso (2014). A Permutation Approach for Ranking of Multivariate Populations. *Journal of Multivariate Analysis*, **132**, 39–57.

[2] S. Bonnini, L. Corain, M. Marozzi and L. Salmaso (2014). *Nonparametric Hypothesis Testing: Rank and Permutation Methods with Applications in R.*Wiley, Chichester.

A robust alternative to the sample autocovariance and autocorrelation functions

H. H. A. Cotta[a,b], V. A. Reisen[a,b], P. Bondon[b] and C. Lévy-Leduc[c]

[a] *Universidade Federal do Espírito Santo,* [b] *CentraleSupélec,* [c] *AgroParisTech*

The sample autocovariance and autocorrelation functions are sensitives to the occurrence of outliers[1, 2]. In this context, this paper proposes a robust estimation method for the autocovariance and autocorrelation functions in the presence of additive outliers. Time series analysis in the frequency domain is based on the study of the spectral density function from which the periodogram is an estimator. It is known that the periodogram may be also obtained from harmonic regression, that is, at each Fourier frequency, a sine and cosine coefficients are fitted in a procedure analogous of the discrete Fourier transform. Nonetheless, the periodogram is also affected by outliers[3]. Thus, the robustness property of the periodogram is achieved by replacing the standard Fourier transform by its robust version obtained by substituting the least square procedure in the harmonic regression by the non-linear M-regression. Recently, this approach has been considered by [4] and [5]. In consideration of the duality between the sample autocovariance function and the periodogram, the robust sample autocovariance function and robust sample autocorrelation function are obtained from the inverse diagonalization procedure of the matrix containing the estimated spectral density. Simulation experiments are conducted to assess the performance of the estimators under contaminated and non-contaminated scenarios. A real data set is also analyzed as an example of an application of the proposed methods.

Keywords: Robustness, Autocorrelation function, M-periodogram.

References

[1] W. Chan (1992). A note on time series model specification in the presence of outliers. *Journal of Applied Statistics*, **19**(1),117–124

[2] W. Chan (1995). Outliers and financial time series modelling: a cautionary note. *Mathematics and Computers in Simulation*, **39**(3), 425–430

[3] F.F. Molinares, V.A. Reisen, F. Cribari-Neto (2009). Robust estimation in long-memory processes under additive outliers. *Journal of Statistical Planning and Inference*, **139**(8), 2511–2525

[4] A. J. Q. Sarnaglia, V. A. Reisen, P. Bondon, C. Lévy-Leduc (2016). A robust estimation approach for fitting a PARMA model to real data. In *2016 IEEE Statistical Signal Processing Workshop (SSP).*

[5] V. Reisen, C. Lévy-Leduc, M. Taqqu (2017). An m-estimator for the long-memory parameter. *Journal of Statistical Planning and Inference*, **187**, 44– 55.

Multipolar Aid: A Human Development analysis with High-Resolution Data

J. Cruzatti Constantine[a,b], H. Cevallos-Valdiviezo[b]

[a]*Heidelberg University,* [b]*Escuela Superior Politécnica del Litoral (ESPOL)*

We investigate the effect of World Bank (WB) and Chinese Aid on different dimensions of Human Development at the sub-national level. We combine 144 demographic and health surveys (DHS) from 57 countries over the 1995-2013 period. Each survey compiles information on around 1700 socioeconomic variables, which can make it difficult to determine important variables. To tackle this problem, we use the Best Subset Selection technique. With a difference-in-difference strategy, our basic regressions focus on locations near aid projects, and compare the effect of projects that have disbursed aid to those that have not. We further address causality by instrumenting aid provided to a specific area with a variable that interacts the time-variant financial institution's liquidity and the space-variant probability of a particular location to receive aid. We exploit time-variation by the use of indicators that proxy World Bank and China's yearly liquidity, which in turn determines the potential for new loans. Spatial variation comes from the calculation of sub-national probability to receive aid. Controlled by this probability, in tandem with fixed effects for city-years and city-clustered standard errors, the interaction provides a powerful and excludable instrument. Our results show heterogeneous results. While WB's Aid seems to have a disrupting effect, China's Aid has negative-to-non significant effects.

Keywords: Human Development analysis, difference-in-difference.

Poster: May 30, 10:00 - 10:30, Posters Room

Space-temporal modeling using DAG‘s and generalized additive models: case study of tuberculosis data

E. A. S. Lizzi and T. C. Cassiano

Federal University of Technology - Parana

Introduction: Information on health and epidemiological studies are conducted in Brazil, considering information available in government health information systems. In this work we explore the combined use of epidemiology, applied mathematics and statistics, big data and computer science, in an unusual way. The case study refers to tuberculosis data, this disease has existed for thousands of years and is a global public health problem. The objective is to work with data mining and to relate social indicators to tuberculosis incidence rates in the southern region of Brazil, pondering this information in time and by geographical positioning. Methods: An ecological epidemiological study with data referring to the municipalities of residence of the southern region of Brazil for the years 2011 to 2017. The analysis was conducted in four stages: obtaining and structuring the data *(big data)*, reducing the size using networks Bayesian models and finally modeling using generalized space and time additive models. The Bayesian networks aid in the process of reducing the size of social indicators, since these are highly correlated, making it impossible for simple and usual techniques. Then, the modeling process relied on generalized space-time additive models to understand tuberculosis relations considering space and time with the covariates of interest that were selected by the results of the Bayesian networks. Results and Conclusion: The indicators showed that over time they have different behaviors, the HDI-education, income and longevity show that the municipalities that present higher values in these indices present a higher risk of tuberculosis cases, since in general they are municipalities and despite development have been unable to cope with the tuberculosis problem. On the other hand, for child mortality, places where there is a greater number of infant mortality rates have a lower risk of tuberculosis, showing that a combined strategy to monitor these two problems can lead to the adoption of efficient public policies, since infant mortality occurs in municipalities underdeveloped and, in these same municipalities, the risk of tuberculosis is lower. In addition, the thematic maps of tuberculosis rates smoothed by the model show that the disease is not stable in the state territory and may indicate the municipalities for priority actions and strategies of health surveillance.

Keywords: Generalized additive models, Big Data, Tuberculosis.

References

[1] I. Ben-Gal (2007). *Bayesian Networks.* In: Ruggeri F, Kenett R. Encyclopedia of Statistics in Quality and Reliability, Wiley & Sons.

[2] T. Hastie, R. Tibshirani, J. H. Friedman (2008). *The elements of statistical learning: data mining, inference, and predicition.* New York Springer.

[3] R. E. Neapolitan (2004). *Learning Bayesian Networks.* Upper Saddle River: Pearson.

Bayesian spatio-temporal modeling: case study domestic violence data against women in Brazil

J. V. S. Magri and E. A. S. Lizzi

Federal University Technology - Paraná

Introduction: Domestic violence transcends the home where it occurs, branching out across the population, thus affecting not only women and children. Besides this fact, this environment of violence generates ”sons” of the same, reflecting in aggressive behaviors on the part of the members of this society [1]. When one speaks of violence against women, Brazil stands out in the world scenario with this statistic. **Goals:** Perform a spatio temporal bayesian model based on cases of domestic violence, sexual and others violences in the municipalities of the state of São Paulo in Brazil, south american, in the period of 2009-2016. **Methods:** The data on number of cases the incidence of domestic violence,sexual and other violence, to select only the female sex in official government databases. In this scenario, a spatial bayesian model was proposed, where response variable follow poisson distribution and link function by logarithmic, using a spatial model with random effect BYM, where spatial dependence was modeled with adjacency matrix of order 1 (one) that represents the location of each municipality and its respective neighbors of first order [3]. The number of domestic violence, sexual and others violences per 100,000 inhabitants is response variable and the HDI and yours sub components, index of theil, index of gini and percentage of vulnerable to poverty entered as predictors in the model. The model was implemented with computational support by R software and used Laplace (INLA) approximation methods for estimations of the parameters. **Results and conclusion:** The model indicate that municipalities where there are higher levels of inequality and worse indicators of income distribution, are the places where there are higher rates of domestic violence. Regarding the indicators of IDH-income, IDH-Longevity and general HDI, the relationship shown was that the higher the indicators, the higher the number of domestic violence cases, both with growth over time. Besides, it is possible to infer that larger municipalities are more unequal, even controlling income and longevity, it is not possible to control the issue of violence against women. In this way, these results help us to propose decentralized social public policies for each municipality, evaluating the proposed indicators.

Keywords: Domestic violence, spatio temporal analysis, bayesian models

References

[1] D. Cerqueira, M. V. M. Matos, A. P. A. Martins, and J. P. Junior (2015). Avaliando a Efetividade da Lei Maria da Penha. *IPEA - Instituto de Pesquisa Econômica Aplicada, Brasília.*

[2] J. J. Waiselfisz (2015). *Mapa da violência 2015 homicídio de mulheres no Brasil.*

[3] E. A. S. Lizzi, A. A. Nunes, and E. Z. Martinez (2016). Current HIV Research. *Bentham Science Publishers*, **14**(10), 466–475.

A Deep Neural Stochastic model for Cryptocurrency Volatility Prediction

G. DiGiorgi[a], R. Salas[a], M. Salinas[a], R. Torres[b] and O. Nicolis[b]

[a] *Universidad de Valparaíso,* [b] *Universidad Andrés Bello*

Volatility is an important indicator of market risk. It is usually expressed as a percentage and calculated as the deviation recorded by an asset with respect to the average of its historical price on a given period. Classical mathematical methods to study volatility involve time series measurements whose behavior has a stochastic characteristic because both the mean and the variance are non-constant. The most outstanding models are the autorregressive conditional heteroskedasticity (ARCH) model and the generalized ARCH (GARCH), which incorporate the functional relationships that allow to relate the current conditional volatility with the past conditional volatilities assuming conditions and previous assumptions. Taking advantage of deep learning, Luo et al proposed a neural network of reformulated stochastic volatility in which the GARCH model and the Heston model take place. The objective of this study was to apply this neuronal network in a set of economic synthetic and real data to predict volatility. We simulated 2041 observations of volatility under t-student distribution for 3 degrees of freedom and data was multiplied by a smoothing function of the volatility with a window of 125 observations. Moreover, we took daily closing prices of cryptocurrencies to predict their volatility. Despite data had a high noise, the overall performance of NSVM showed a good prediction and high accuracy of the mean, standard deviation and its confidence bounds. This implies that neural network has a good training and it could be a useful tool to predict volatility.

Keywords: Deep learning, neural networks, volatility.

Robust Approaches to Non-Destructive Testing in Civil Engineering

M. Doktor[a,b], W. Kurz[b], C. Redenbach[b], P. Ruckdeschel[c] and J.-P. Stockis[b]

[a] *PwC Germany,* [b] *University of Kaiserslautern,* [c] *University of Oldenburg*

Non-destructive testing methods have become popular within the last decade, especially in mechanical and electrical engineering, whereas to date, in civil engineering one still mainly recurs to destructive testing. This is often not applicable, in particular when it comes to safely testing the yield limit as well as for determining the current stress level. We head for non-destructive alternatives in this framework based on ultrasonic and micro-magnetic tools to assess the steal beams incorporated in buildings. The rise of such tools for (non-destructive) measurements – which might still be error prone – makes them a promising starting point to characterize old steal buildings and bridges, enabling ecological and economical maintenance for ageing infrastructure. Mathematical tasks in this context a.o. are the classification of e.g. inbuilt material, further quantification of the global stress and successive estimation of occurring internal forces. The presented approach uses techniques from nonparametric statistics such as statistical classification based on support vector machines and on the other hand (robustied) sieve and partition estimators.

To illustrate the approch, we show applications determining internal forces in a steal beam from both simulated and real data and the benefit of robust diagnostics in this domain in identifying anomalies/deviations in the (possibly only partially known) statical system.

Keywords: Robust Regression, Robust Diagnostics, Mathematical and Mechanical Modeling

References

[1] M. Doktor, C. Fox, W. Kurz and J.-P. Stockis (2018). Characterization of steal buildings by means of non-destructive testing methods. *Journal of Mathematics in Industry*, 8:10.

Extending R with C++: Motivation, Examples, and Context

D. Eddelbuettel

University of Illinois at Urbana-Champaign

The R language and environment has become the *lingua franca* of statistical computing. Statistical research as well as statistical applications benefit greatly from its capabilities, both 'built-in' and via the excellent package system. And "interfaces to other software" are a key part of R as Chambers emphasized in his book "Extending R" [1]. Among the available extension mechanisms for R, the Rcpp package [2] has become the most popular and widely-used approach. In this talk, we briefly re-introduce Rcpp and its core characteristics, describe how it fits the R model, illustrate several key extensions implemented via Rcpp, and possibly speculate a little about future changes.

Keywords: Statistical Computing, R

References

[1] J. Chambers (2017). *Extending R.* Chapman&Hall / CRC: The R Series. CRC Press.

[2] D. Eddelbuettel, R. François, J. J. Allaire, K. Ushey, Q. Kou, N. Russel, J. Chambers, D. Bates (2019), Rcpp: Seamless R and C++ Integration, R package version 1.0.1, http://CRAN.R-Project.org/package=Rcpp.

A multivariate model to discovery knowledge of research groups from patents

J. Fernández Ledesma

Universidad Pontificia Bolivariana

The proposed research project is a descriptive and multivariate study with a mixed approach, where it is proposed to develop a methodology that allows the effective use of the present information of invention patents during the investigative process of some groups of Universities in Colombia, classified according to the National Research System (COLCIENCIAS) model and according to characteristics of the knowledge creation model of Nonaka and Takeuchi (1999). The proposed research process will be carried out in four stages, initially a characterization of the groups will be developed, as well as a description of how they create the knowledge according to the SECI model (Nonaka and Konno, 2000), using the patent information. From the data collection through structured surveys, in a sample of research groups that meet the inclusion criteria and that can be measured and intervened will be qualitative and quantitative analysis, using mainly multivariate techniques as PCA, Classification and Multivariate Regression that help define the elements of the methodology to be developed for its implementation.

Keywords: Multivariate Statistics, Applications

Probabilistic Constrained Optimization Using Bayesian Networks

J. E. Fernandez[a], B. Silva[b]

[a] *Pontificia Universidad Catolica del Ecuador,* [b] *Escuela Politecnica Nacional*

Linear optimization problems with probabilistic constraints have a wide variety of applications. In the specialized literature, there are approximate solution techniques for problems where each constraint is tied to a probability of occurrence. However, there is not the same level of advance when modeling the joint probability of simultaneous occurrence of the entire set of constraints in such problems. In this research, Bayesian networks and other graphical models are used to deal with the joint probability distribution of the constraints of a linear optimization problem with probabilistic constraints. Then, the optimization problem is solved using Monte Carlo simulations over the fitted graphical model. Finally, this approach is tested on a classic portfolio optimization problem and it is shown that the proposed methodology presents a better performance in comparison with solutions of traditional related methods.

Keywords: Bayesian networks, probabilistic optimization, probabilistic constraints.

Robust k-means-based clustering for high-dimensional data

P. Filzmoser[a], Š. Brodinová[a,b], T. Ortner[a], C. Breitender[a], and M. Rohm[a]

[a] *TU Wien, Austria,* [b] *Solvistas GmbH, Austria*

We introduce a robust k-means-based clustering method for high-dimensional data where not only outliers but also a large number of noise variables are very likely to be present. Although Kondo et al. [2] already addressed such an application scenario, our approach goes even further. Firstly, the introduced method is designed to identify clusters, informative variables, and outliers simultaneously. Secondly, the proposed clustering technique additionally aims at optimizing required parameters, e.g. the number of clusters. This is a great advantage over most existing methods. Moreover, the robustness aspect is achieved through a robust initialization [3] and a proposed weighting function using the Local Outlier Factor [1]. The weighting function provides a valuable source of information about the outlyingness of each observation for a subsequent outlier detection. In order to reveal both clusters and informative variables properly, the approach uses a lasso-type penalty [4]. The method has thoroughly been tested on simulated as well as on real high-dimensional datasets. The conducted experiments demonstrated a great ability of the clustering method to identify clusters, outliers, and informative variables.

Keywords: k-means, Outliers, High-dimensional data.

References

[1] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander (2000). LOF: Identifying density-based local outliers. In: *ACM Sigmod Record*, 29:93–104.

[2] Y. Kondo, M. Salibian-Barrera, and R. Zamar (2016). RSKC: An R package for a robust and sparse k-means clustering algorithm. *Journal of Statistical Software*, 72:1–26.

[3] A. H. Mohammad, C. Vineet, S. Saeed, and J. Z. Mohammed (2009). Robust partitional clustering by outlier and density insensitive seeding. *Pattern Recognition Letters*, **30**(11):994–1002.

[4] D. M. Witten, and R. Tibshirani (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, **105**(490):713–726.

Implications of assumptions verification in the means comparison tests

Pablo Flores[a,b,c] and Jordi Ocaña[c]

[a]*Escuela Superior Politécnica de Chimborazo,* [b]*Grupo de Investigación en Ciencia de Datos CIDED,* [c]*Universitat de Barcelona, Facultat de Biologia, Grup de Recerca en Bioestadística i Bioinformática GRBIO*

The effectiveness of $t-$student test to compare means is restricted to the fulfilment of normality and homocedasticity assumptions. Infringement of these assumptions implies the use of another tests such as Welch or Wilcoxon. Although, traditionally, these assumptions are pre-tested by another test (e.g Shapiro Wilk, $F-$test) on the same data, some studies [1] [2] show than this pretesting process alters the overall Type I Error Probability and it is not recommended to use it. However, we believe that problem is not pretesting, but the problem is the form than hypothesis are proposed [3], for example, on homoscedasticity, when null hypothesis of perfect equality is not rejected, we conclude than there is not enough evidence to conclude significant differences between the two compared variances, and this does not imply necessarily equality. For normality is nearly the same. There is a different approach called equivalence [4], this type of tests overcome the difficulty mentioned above, since the criterion of equivalence is established in the alternative hypothesis, in addition it does not prove a perfect equality but rather an equality except for irrelevant deviations. Using different sample sizes and different theoretical departures from normality and homoscedasicity, and pre-testing both assumptions in a combined way, this study shows through a stochastic simulation process that, when instead of the traditional approach, an equivalence approach with adequate irrelevance limits is used [4], the type I error is better controlled, around the significance level α. Finally it is observed that Welch is the test that remains more robust to deviations or departures from homoscedasticity and normality, it is even more robust than Wilcoxon test, which is supposed to be designed for these cases.

Keywords: Equivalence, pretesting assumptions, simulation.

References

[1] D. Rasch, K. D. Kubinger, and K. Moder (2011). The two-sample t test: pre-testing its assumptions does not pay off. *Statistical papers*, **52**(1), 219–231.

[2] D. W. Zimmerman (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, **57**(1), 173–181.

[3] D. G. Altman, and J. M. Bland (1995). Statistics notes: Absence of evidence is not evidence of absence. *Bmj*, **311**(7003), 485.

[4] S. Wellek (2010). *Testing statistical hypotheses of equivalence and noninferiority.* Chapman and Hall/CRC.

[5] P. Flores and J. Ocana (2018). Heteroscedasticity irrelevance when testing means difference. *SORT: statistics and operations research transactions*, **42**(1), 59–72.

On moments of folded and truncated multivariate extended skew-normal distributions

Christian E. Galarza[a], Larissa A. Matos[a] and Victor H. Lachos[b]

[a] *Department of Statistics, Campinas State University, Campinas, Brazil*
[b] *Department of Statistics, University of Connecticut, CT 06269, U.S.A.*

Following [3], this paper develops recurrence relations for integrals that involve the density of multivariate extended skew-normal distributions, which includes the well-known skew-normal distribution introduced by [1] and the popular multivariate normal distribution. These recursions offer fast computation of arbitrary order product moments of truncated multivariate extended skew-normal and folded multivariate extended skew-normal distributions with the product moments of the multivariate truncated skew-normal, folded skew-normal, truncated multivariate normal and folded normal distributions as a byproduct.

Finally, from the application point of view, these moments open the way to propose analytical expressions on the E-step of the Expectation-Maximization (EM) algorithm for complex data, such as, asymmetric longitudinal data with censored and/or missing observations. These new methods are provided to practitioners in the R MomTrunc package, an efficient R library incorporating C++ and FORTRAN subroutines through Rcpp [2].

Keywords: Product moments, Truncated distributions, Censored models.

References

[1] A. Azzalini, and A. Dalla-Valle (1996). The multivariate skew-normal distribution. *Biometrika*, **83**(4), 715–726.

[2] D. Eddelbuettel (2013). *Seamless R and C++ Integration with Rcpp.* Use R! 64. Springer-Verlag New York, first edition.

[3] R. Kan, and C. Robotti (2017). On moments of folded and truncated multivariate normal distributions. *Journal of Computational and Graphical Statistics*, **25**(1), 930–934.

Post-selection confidence curves

A. C. Garcia-Angulo, and G. Claeskens

ORSTAT, KU Leuven, Belgium

Post-selection inference requires novel methodology to produce inference that takes into account the extra variability added by model selection. Recent work has focused on obtaining valid p-values and confidence intervals for the parameters of interest.

We propose the use of the concept of confidence distributions (see, e.g., [1]) to get a more complete picture of the variable selection effect. A confidence distribution provides an inferential summary that can be explained in terms of p-values, confidence intervals and point estimators.

We show how to produce post-selection confidence distributions and curves for target parameters when the candidate models belong to exponential families. Our proposal applies to any selection method that can be rewritten as a set of inequalities forming truncation limits independently of the parameter value. This includes, for instance, all likelihood-based selection methods such as for example AIC, BIC, likelihood ratio tests and significance hunting via t-tests.

The model selection set might or not contain the true data generating model. For correctly specified selected models we make use of the sufficient statistics to obtain the uniformly most powerful post-selection confidence curves for continuous distributions. For misspecified models, conservative confidence curves are obtained by the use of robust covariance matrix estimators.

Keywords: Model selection, Inference, Post-selection effects.

References

[1] T. Schweder, and N. L. Hjort (2016). *Confidence, Likelihood, Probability.* Cambridge University Press, New York.

Analysis of variance for spatially correlated functional data: application to brain data

Jeimy Aristizabal-Rodríguez[a], Ramón Giraldo[a] and Jorge Mateu[b]

[a] *Universidad Nacional de Colombia,* [b] *Universidad Jaume I*

Functional data showing spatial dependence structure occur in many applied fields. For example, in meteorology when curves of temperature are obtained in a monitoring network, or in neurological studies when curves of the electrical activity are recorded in voxels of the brain. The statistical methods for modeling functional data must be adapted to this framework to provide valid inferential procedures. Recently, several works on functional (linear, generalized, additive or semiparametric) models considering correlated functional data have been proposed. Our contribution in this paper is framed in this scenario. Specifically, we show two approaches for carrying out analysis of variance of functional data (FANOVA) when the functions are spatially correlated. These are based on adaptations to the spatial context of two classical techniques given in the literature of FANOVA. To illustrate the methodologies proposed, we analyze a dataset from biomedicine corresponding to curves of evoked potentials obtained under a one-way experimental design.

Keywords: Analysis of Variance, Evoked potentials, Functional data analysis, Geostatistics

Shapiro-Wilk test for skew normal distributions

E. Gonzalez-Estrada, W. Cosmes and J. A. Villasenor

Programa de Estadística, Colegio de Postgraduados, México

The skew normal (SN) family of distributions [1] includes the normal distribution as well as a wide variety of skew densities. Because of its analytical tractability and flexibility for modeling both symmetric and skew data sets [2], the SN distribution has plenty of applications in finance, diverse engineering fields, medicine, etc. A probability property that connects the SN distribution with the normal distribution is used here for proposing a goodness of fit test for the composite null hypothesis that a random sample follows a SN distribution with unknown parameters. The random sample is transformed to approximately normal random variables and then Shapiro-Wilk test is used for testing normality. The implementation of this test does not require neither parametric bootstrap nor the use of tables for different values of the slant parameter. The results of an extensive power study conducted by Monte Carlo simulation shows some good properties of the proposed test in comparison to existing tests for the same problem, which are based on parametric bootstrap. The returns of daily closing prices of the German Stock Index (DAX) of the year 2018 are analyzed.

Keywords: Parametric bootstrap, Monte Carlo simulation, goodness of fit.

References

[1] A. Azzalini (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**(2), 171–178.

[2] C. Adcock, M. Eling, and N. Loperfido (2015). Skewed distributions in finance and actuarial science: a review. *The European Journal of Finance*, **21**(13-14), 1253–1281.

Partition Markov Model for Multiple Processes

M. T. A. Cordeiro[a], Jesús E. García[b], V. A. González-López[b] and S. L. M. Londoño[c]

[a] *Federal University of Technology, Ponta Grossa PR, Avenida Monteiro Lobato, s/n - Km 04, CEP: 84016-210, Brazil.*

[b] *Department of Statistics, University of Campinas, Sergio Buarque de Holanda, 651, Campinas, SP, CEP: 13083-859, Brazil.*

[c] *University of Campinas, Sergio Buarque de Holanda, 651, Campinas, SP, CEP: 13083-859, Brazil.*

In this paper we analyze the model proposed in [1] in which is considered a set of p independent samples of discrete time Markov chains, over a finite alphabet A and with finite order o. The model consists of identifying the states on the state space A^o where two or more samples share the same transition probabilities (see also [2]). This identification establishes a partition on $\{1,\ldots,p\} \times A^o$ the set of samples and the state space. We show that by means of the Bayesian Information Criterion (BIC) the partition can be estimated eventually almost surely. Also in [1] is given a notion, derived from the BIC, which serves to identify the proximity/discrepancy between elements of $\{1,\ldots,p\} \times A^o$ (see also [3]). In the present article we also prove that this notion is a metric in the space where the model is built and that it is statistically consistent to determine proximity/discrepancy between the elements of the space $\{1,\ldots,p\} \times A^o$.

Keywords: Stochastic Processes, Bayesian Information Criterion, Metric between Stochastic Processes.

References

[1] Jesús E. García and S. L. M. Londoño (2019). Optimal Model for a Set of Markov Processes. *AIP Conference Proceedings* (in press).

[2] Jesús E. García and V. A. González-López (2017). Consistent Estimation of Partition Markov Models. *Entropy*, **19**(4), 180.

[3] Jesús E. García, R. Gholizadeh and V.A. González-López (2018). A BIC - based consistent metric between Markovian processes. *Applied Stochastic Models in Business and Industry*, **34**(6), 868-878.

Evaluation of process capability in nonlinear profiles

Rubén Darío Guevara González, Tania Alejandra López Guasca, and José Alberto Vargas Navas

Universidad Nacional de Colombia

Process capability indices measure the ability of a process to provide products that meet certain specifications. Because the technological development, the quality of products or processes is more related to functional data. In particular, sometimes the quality of a process is best expressed by a relationship between a response variable and one or more explanatory variables, which is called profile. In many cases, the profiles are better described by a nonlinear function than by a linear function. In this section, initially, we present some methods to measure the capability of processes characterized by nonlinear functions, based on the concept of functional depth. After, we extend this methodology to processes characterized by multivariate nonlinear profiles.

Keywords: process capability analysis, nonlinear profiles, functional data, functional depth, multivariate functional principal component analysis

Inference on Quantile Regions in Linear Models

Y. Sun and X. He

University of Michigan

When certain quantile region of an outcome variable is of primary interest, it is natural to consider statistical inference directly on that quantile region after adjusting for any covariates or confounders. For example, we may wish to compare the risk of investment portfolios by focusing on a region of low quantiles of the return distributions; such a risk measure is often quantified by CVaR in financial risk management. In this talk we discuss a new inferential tool for hypothesis testing for regional quantile parameters in the linear quantile regression setting. We propose a rank-score-based test statistic and introduce a model-based resampling method to approximate its null distribution. The proposed test is robust and powerful.

Keywords: Bootstrap, Quantile regression, Rank-score.

Lorenz regression for single-index models with monotone link functions

C. Heuchenne[a,b] and A. Jacquemin[b]

[a] *University of Liege, Belgium,* [b] *Catholic University of Louvain, Belgium*

In many situations, parametric regression induces misspecified models while nonparametric regression techniques are less efficient and quickly deteriorate when the number of covariates increases. In addition, the last procedures involve sometimes not data-driven specific choices of smoothing parameters that make them more tedious and sensitive to numerical instabilities. To alleviate those flaws, we present a regression methodology that considers a model where the mean of the response given the covariates is a monotone link function of a linear combination of the covariates. The proposed techniques are based on the estimator developed in [1] and the Lorenz curve widely used in Economics to describe income inequalities. We treat the case of discrete covariates and solve the resulting constrained maximization problem with a genetic algorithm. We develop boostrap confidence intervals and hypothesis tests for the coefficients of the regression model as well as a goodness-of-fit measure evaluating the proportion of explained inequality. We assess the performance of our new procedures through a series of Monte-Carlo simulations and compare them to Ichimura's nonlinear least-squares related techniques [2]. Finally, we present and analyze a real data set.

Keywords: Single-index model, Rank estimator, Genetic algorithm

References

[1] C. Cavanagh and P. R. Sherman (1998). Rank Estimators for Monotonic Index Models. *Journal of Econometrics*, **84**, 351–381.

[2] H. Ichimura (1993). Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models. *Journal of Econometrics*, **58**, 71–120.

MPCA vs. DS-PC performance comparison: a case study of fungicide efficacy evaluation for controlling black sigatoka on Ecuadorian banana plantations

J. Ascencio-Moreno[a], M. V. Hinojosa-Ramos[a], F. Vera[b], O. Ruiz-Barzola[a], M. I. Jiménez-Feijoó[a], M. P. Galindo-Villardón[c] and M. Ramos-Barberán[b]

[a]*Facultad de Ciencias de la Vida, ESPOL, Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador,* [b]*Facultad de Ciencias Naturales y Matemáticas, ESPOL, Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador,* [c]*Departamento de Estadística, Campus Miguel de Unamuno, Universidad de Salamanca, Salamanca, España*

Nowadays, life scientists are surrounded by massive amounts of data derived from daily activity; extracting significant information from this data is often a complex task. As an agricultural case study, fungicide sensitivity of black sigatoka pathogen (*Pseudocercospora fijiensis*) was evaluated on bananas' plantations in Ecuador. Inhibition percentages were computed from laboratory assays creating several arrays of data, and the 50% inhibition dose (EC50) was the main reference for evaluating the efficacy of the products in every plantation. However, EC50, a univariate measurement in the multivariate spectrum of biology, is not the best representation for results in terms of products' performance. In this study, with few specific considerations, the industrial scope of statistical process control was adjusted to fungicide efficacy evaluation defining three-way arrays. The multivariate statistical control techniques applied were Multilinear Principal Component Analysis (MPCA) and Dual STATIS-Parallel Coordinates approach (DS-PC) based on an adaptation of singular value decomposition. A comparison was developed and showed that both methods discriminate correctly, by year, the normal and anomalous conditions within plantations, validating the ability of the novel method DS-PC for exhibiting better signaling of anomalous plantations and performing variable-wise analysis to find out possible causes of this behavior.

Keywords: Bananas, Fungicide efficacy, Multivariate control charts.

Robust Designs of Big Comparative Studies

Feifang Hu[a], Yichen Qin[b], Yang Li[c] and Wei Ma[c]

[a] *George Washington University,* [b] *University of Cincinnati,* [c] *Renmin University of China*

Covariate balance is one of the most important concerns for successful comparative studies, such as causal inference, online A/B testing and clinical trials, because it reduces bias and improves the accuracy of inference. In literature, chance imbalance still exists and is studied in traditional randomized experiments. The phenomenon of covariate imbalance even aggravates as the number of covariates p and the sample size n increase, which is almost ubiquitous in the era of big data. For example, suppose the probability of one particular covariate being unbalanced is $\alpha = 5\%$. For a study with 10 covariates, the chance of at least one covariate shows imbalance is $1 - (1 - \alpha)^p = 40\%$. To achieve better covariate imbalance in the framework of causal inference, [1] have proposed rerandomization (RR) procedure. However, their rerandomization (RR) procedure does not work well for large number of covariates p and the sample size n. It is important to develops new sequential covariate adaptive designs to address the issue.

In this talk, a family of robust designs are proposed by allocating the units sequentially and adaptively, using information on the current level of imbalance and the incoming unit's covariate. With a large number of covariates or a large number of units, the proposed method shows substantial advantages over the traditional methods in terms of the covariate balance and computational time, making it an ideal technique in the era of big data. Furthermore, the proposed method improves the estimated average treatment effect accuracy by achieving a minimum variance asymptotically. Numerical studies and real data analysis provide further evidence of the advantages of the proposed method.

The proposed method is following the similar spirit of the minimization methods in clinical trials [2], [3], [4]. However, the context for the minimization method is different from ours.

Keywords: Big Data, Covariate-Adaptive Design, Efficiency.

References

[1] K. L. Morgan, and D. B. Rubin (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, **40**, 1263-1282.

[2] D. R. Taves (1974). Minimization: A new method of assigning patients to treatment and control groups. *Clinical Pharmacology & Therapeutics*, **15**, 443–453.

[3] S. J. Pocock and R. Simon (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, **31**, 103–115.

[4] Y. Hu, and F. Hu (2012). Asymptotic properties of covariate-adaptive randomization. *Annals of Statistics*, **40**, 1794–1815.

An estimation of the industrial production dynamic in the Mercosur countries using the Markov switching model

Saba Infante[a,b], Edward Gómez[b], Luis Sánchez[c], Aracelis Hernández[a,b].

[a]*Escuela de Ciencias Matemáticas y Computacionales. Universidad de Investigación Yachay Tech. Ecuador,* [b]*Departamento de Matemáticas, Facultad de Ciencia y Tecnología, Universidad de Carabobo, Venezuela,* [c]*Departamento de Matemática y Estadística, Instituto de Ciencias Básicas, Universidad Técnica de Manabí, Ecuador.*

In this work is applied a methodology based on Bayesian statistical methods inspired by Markov Monte Carlo Chain sampling schemes, which simplifies the estimation and prediction process of the Markov switching model. The general objective of this study is to determine simultaneously: non-linearity, structural changes, asymmetries and outliers that are characteristics present in many financial series. The methodology is empirically illustrated using series that measure the annual growth rate of industrial production in the MERCOSUR countries. The sampling algorithm [1] is implemented to estimate the parameters of the model. The parameters were estimated in terms of expected values a posteriori and standard deviations a posteriori. Three criteria is used to assess the prediction model: square root of the mean quadratic error (RSME), the Diebold-Mariano (DM) test and the T^{RC} statistical test. These measures of goodness-of-adjustment are showed, that the estimates have small errors. The execution time of the algorithm is calculated, observing high performance. As a result of the analysis of the data, it is concluded, that there is no reduction of economic volatility and no reduction in the depth of economic cycles. At the breakpoints, atypical values and non-linearity in the data are observed.

Keywords: Markov Switching Models, Algorithms Monte Carlo Markov, Industrial Production in Mercosur.

References

[1] P. Giordani, R. Kohn and D. Van Dijk (2007). A unified approach to nonlinearity, structural change, and outliers. *Journal of Econometrics*, **137**, 112–133.

[2] C. Kim and C. Nelson (1999). *State-Space Models With Regime-Switching: Classical and Gibbs-Sampling Approaches with Applications*, MIT Press, Cambridge.

[3] R. Gerlach, C. Carter and R. Kohn (2000). Efficient bayesian inference for dynamic mixture model. *Journal of the American Statistical Association*, 819–828.

An Empirical Bayes Solution for Selection Bias in Functional Data

J. Derenski, Y. Fan and G. James

University of Southern California

Selection bias results from the selection of extreme observations and is a well-recognized issue for standard scalar or multivariate data. Numerous approaches have been proposed to address the issue, dating back at least as far as the James-Stein shrinkage estimator. However, the same potential issue arises, albeit with additional complications, for functional data. Given a set of observed functions, one may wish to select for further analysis those which are most extreme according to some metric such as the average, maximum, or minimum value of the function. However, given that the functions are often noisy realizations of some underlying mean process, these outliers are likely to generate biased estimates of the quantity of interest. In this talk I propose an Empirical Bayes approach, using Tweedie's formula, to adjust such functional data to generate approximately unbiased estimates of the true mean functions. The approach has several advantages. It is non-parametric in nature, but is capable of automatically shrinking back towards a James-Stein type estimator in low signal situations. It is also computationally efficient and possesses desirable theoretical properties. Furthermore, I demonstrate through extensive simulations that the approach can produce significant improvements in prediction accuracy relative to possible competitors.

Keywords: Functional Data, Empirical Bayes, Tweedie's formula.

A geometry-based algorithm for cloning real grains 2.0

D. A. Medina[b] and A. X. Jerves[a,b]

[a] *Facultad de Ciencias Naturales y Matemáticas, Escuela Superior Politécnica del Litoral, ESPOL, Campus Gustavo Galindo Velasco km 30.5 via Perimetral, Guayaquil, Guayas, Ecuador,* [b] *Fundación INSPIRE*

We introduce an improved version of a computational algorithm that “clones” / generates an arbitrary number of new digital grains from a sample of real digitalized granular material. Our improved algorithm produces “cloned” grains that more accurately approach the morphological features displayed by their parents. Now, the “cloned” grains were also included in a discrete element method simulation of a triaxial test and showed similar mechanical behavior compared to the one displayed by the original (parent) sample. Thus, the present work is divided in four parts. First, we compute multivariable probability density functions from the parents’ morphological parameters (morphological DNA), i.e., aspect ratio, roundness, volumesurface ratio, and particle diameter. Second, an improved, now parallelized and better tuned version of the geometric stochastic cloning algorithm [1], which is based on the aforementioned multivariable distributions and that, in the same way, introduces an enhanced radii sampling process, as well as a new quality control test based on the volume-surface ratio is discussed. Third, morphological DNA of the grains (i.e., aspect ratio, roundness, volume-surface ratio and particle diameter) is also extracted from the new “cloned” grains and compared to the one obtained from the parent sample. Fourth, clones and parents are subjected to a triaxial compression tests using a level set discrete element scheme (3DLS-DEM), and then, compared in terms of their mechanical response. Finally, the error of the “clones” in the morphology and mechanical behavior is analyzed and discussed for future improvements.

Keywords: Granular materials, Morphological parameters, Level sets

References

[1] A. Jerves, R. Y. Kawamoto, and J. E. Andrade (2017). A geometry-based algorithm for cloning real grains. *Granular Matter*, 19:30.

ICORS.CPS1: May 28, 11:00 - 12:00, Auditorium 1

Robust Variable Selection via Adaptive Elastic Net S-Estimators for Linear Regression

D. Kepplinger[a] and E. Smucler[b]

[a] *University of British Columbia,* [b] *Universidad Torcuato Di Tella*

Improving biomedical technology affords collecting increasingly large amounts of -omics data. Many applications, including biomarker discovery, aim to identify important predictors from a vast pool of candidates (e.g., protein levels) to predict a given response (e.g., disease status). In the early stages of biomarker discovery studies, it is crucial to identify all important predictors and limit the number of falsely included predictors to control the cost of future validation studies. Penalized regression methods are commonly used tools for these applications, but variable selection and parameter estimation is in practice often hampered by unusual observations (e.g., rare genetic profiles) and contamination.

In the presence of moderate to high correlations among predictors, as often found in -omics data, available robust regression methods with sparsity-inducing penalties show unsatisfactory variable selection performance and/or rely on a robust estimate of the residual scale. Except for low-dimensional settings, robust estimation of the residual scale is itself challenging. We tackle this issue by extending the theory of the recently proposed Penalized Elastic-Net S-Estimator (PENSE) [1], and improve variable selection properties by replacing the penalty with the adaptive elastic net, defined component-wise as $P(\beta; \tilde{\beta}, \lambda_1, \lambda_2, \zeta) = \frac{\lambda_1}{|\tilde{\beta}|^\zeta}|\beta| + \lambda_2|\beta|^2$. This estimator, called adaptive PENSE, leverages an initial PENSE estimate, $\tilde{\beta}$, to heavily penalize predictors with small coefficient values and thus eliminating false positives, while also reducing the bias for large coefficients. We show that under lenient conditions on the error term (e.g., no moment conditions) and the fixed-dimensional predictors, the adaptive PENSE possesses the oracle property, without relying on an auxiliary scale estimate.

In numerical experiments we demonstrate the usefulness of adaptive PENSE in finite samples. We show the theoretical variable selection properties translate to a high sensitivity and specificity in a wide range of settings, outperforming competing estimators. The simulation study further reveals that the improved variable selection and reduced bias lead to a better estimate of the residual scale compared with other methods. This improvement enables a more efficient coefficient estimate via a subsequent regularized M-step.

Keywords: Linear Regression, Regularized Estimation, Variable Selection.

References

[1] G. V. Cohen Freue, D. Kepplinger, M. Salibian-Barrera, and E. Smucler (2019+). Robust Elastic Net Estimators for Variable Selection and Identification of Proteomic Biomarkers. *Annals of Applied Statistics* (submitted).

Linnik probability densities via integration over Hankel contours

Moreno Bevilacqua[a], Tarik Faouzi[b], Igor Kondrashuk[c] and Emilio Porcu[d]

[a] *Department of Statistics, Universidad de Valparaiso, Chile and Millennium Nucleus Center for the Discovery of Structures in Complex Data, Chile,*
[b] *Department of Statistics, Applied Mathematics Research Group, University of Bio-Bio, Concepcion, Chile,*
[c] *Grupo de Matemática Aplicada, Departamento de Ciencias Básicas, Universidad del Bío-Bío, Chillán, Chile,*
[d] *School of Mathematics and Statistics, University of Newcastle, UK and Department of Mathematics, University of Atacama and Millennium Nucleus Center for the Discovery of Structures in Complex Data, Chile.*

We represent Linnik probability densities in a form of Hankel contour integral in a complex plane. We show that this contour integral representation is useful for analysis of the asymptotic behaviour of these densities. This contour integral may be written down in terms of infinite series in several different ways. We prove that all these at first sight different series are equivalent.

Keywords: Linnik probability densities, Hankel contour

References

[1] T. Faouzi, E. Porcu, M. Bevilacqua, and I. Kondrashuk (2018). Zastavnyi Operators and Positive Definite Radial Function, *arXiv:1811.09266[math.SP]*

Air Pollution prediction using Self-Organizing Long-Short Term Memory Networks

Javier Linkolk López-Gonzales[a,b], Cristian Ubal[b], Orietta Nicolis[c], Romina Torres[c] and Rodrigo Salas Fuentes[b]

[a] *Universidad Peruana Unión,* [b] *Universidad de Valparaíso,* [c] *Universidad Andrés Bello*

In highly populated and / or industrialized cities, they are affected by high levels of pollutant concentration in the air, which seriously affects the health of their inhabitants. In this work, we propose a new ensemble technique of Long-Short Term Memory networks based on self-organizing maps to enhance the estimation if PM2.5 concentration in urban Santiago, Chile. The research aims to clusterize scenarios and estimation of higher PM2.5 pollution in areas categorized as the zone with the highest pollution index in the Metropolitan Capital. For this, machine learning techniques will be used, such as self-organized maps (SOM) and long-short term memory networks (LSTM). The ensemble between SOM and LSTM, allowed grouping the time series according to the determined pattern. This showed neurons that have a group of similar time series that corresponds to the pollution record per hour in a day.

Keywords: Air pollution, Ensemble of Artificial Neural Networks, LSTM.

The art of robust statistics for the analysis and improvement of a hospital's medical care

Anthony Villacís[a], Kenny Escobar[a], Juan Carlos Letechi[a] and Laura Andrea López Rodríguez[b]

[a]*Escuela Superior Politécnica del Litoral, Guayaquil-Ecuador,* [b]*Pontificia Universidad Javeriana, Bogotá-Colombia*

Abstract

In the last decade most hospitals have modernized the assignment of medical consultations through agendas via call center or interconsultations made by the general practitioner, this research aims to improve the care of hospitals by reducing the allocation time of appointments for different areas due to poor management at the time of assigning them, causing problems to patients by waiting time for care. This study proposes through robust statistics to know the anomalies of the system of hospital appointments in a hospital of Guayaquil. With the help of statistical software, using as a method the analysis of monthly care data for each area, this way the type of distribution is known and the appropriate regression model is made for the information collected, then proceeds to develop the graph of the data with respect to standardized waste, this results in real time the presence of anomalies, which are the high waiting times for care to be treated by different robust criteria, know whether they are impressionable or not and then be highlighted to end up making a robust regression model. It is proposed as a solution to propose a model of queues, typical structure of a queue with multiple servers and FIFO discipline (first in first out). In order to know if there are changes, the value of the estimators vary considerably, allowing to know if the improvement of the time of attention to the patients in that hospital is carried out.

Keywords: Robust statistics, healthcare, database.

Divergence Methods for Models with Latent Structure: Theory and Algorithms

Lei Li[a] and Anand N. Vidyashankar[b]

[a] *George Mason University,* [b] *George Mason University*

Finite mixture models with random effects arise in several contemporary applications spanning various scientific disciplines. Existing algorithms for computing the estimates of parameters in such models include (i) EM algorithm, (ii) HMIX algorithm, and (iii) proximal point algorithm (see [1], [2], [3]). It is well-known that estimators obtained from an application of the EM algorithm are not robust. In this presentation, we propose a new algorithm, called the DivMin algorithm, which include as a special case the EM algorithm, HMIX algorithm, and other proximal point algorithms. We investigate the population version and the sample version of the algorithm and establish various theoretical properties including rates of convergence. Furthermore, when the divergence belongs to a subclass used in robust estimation, we show that the estimates are robust. Finally, using large deviation techniques, we provide a link between the number of steps of the algorithm that are required to achieve a pre-specified target for a given sample size. We illustrate our results with several simulation studies and real data examples. The software is also provided.

Keywords: DivMin Algorithm, Finite Mixture Models, Random Effects.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological*, 1–38.

[2] A. Cutler, and O. I. Cordero-Braña (1996). Minimum Hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association*, 1716–1723.

[3] P. Tseng (2004). An analysis of the EM algorithm and entropy-like proximal point methods. *Math. Oper. Res.*, 27–44.

R-estimators and their bootstrapped version for GARCH models

Kanchan Mukherjee and Hang Liu

Lancaster University

The quasi-maximum likelihood estimator (QMLE), which is a commonly used approach for estimating GARCH parameters, is sensitive to outliers and does require finite fourth moment of the error distribution. Therefore, a more robust estimator should be considered as financial time series usually have heavy-tails. In this paper, we propose a class of rank estimators, which requires only finite second moment and is highly efficient. The R-estimator converges to a unknown distribution at rate $O_p(n^{-1/2})$. The weighted bootstrap is considered to approximate its distribution. The algorithms for computing the R-estimator and bootstrap estimator are given. Both real data analysis and simulation show the great performance of the R-estimator under normal and heavy-tailed distributions. The bootstrap is also shown to have good coverage rates through extensive simulation. The GARCH $(2, 1)$ model is investigated through simulation. Applications of the R-estimator to predict the value at risk (VaR) and estimating parameters in the GJR model are considered.

Keywords: R-estimators, Bootstrap, GARCH.

ICORSLACSC.IPS: May 31, 08:30 - 10:00, Auditorium 2

Robust estimation in plant breeding: evaluation using simulation and empirical data

V. Lourenço[a], J. Ogutu[b] and H.-P. Piepho[b]

[a] *FCT and CMA, NOVA University of Lisbon, Portugal,* [b] *Institute of Crop Science, Biostatistics Unit, University of Hohenheim, Germany*

Genomic prediction (GP) is used in animal and plant breeding to help identify the best genotypes for selection. One of the most important measures of the effectiveness and reliability of GP in plant breeding is predictive accuracy. An accurate estimate of this measure is thus central to GP. Moreover, regression models are the models of choice for analyzing field trial data in plant breeding. However, models that use the classical likelihood typically perform poorly, often resulting in biased parameter estimates, when their underlying assumptions are violated. This typically happens when data are contaminated with outliers. These biases often translate into inaccurate estimates of heritability and predictive accuracy, compromising the performance of GP. Robust statistical methods provide an intuitively appealing and a theoretically well justified framework for overcoming some of the drawbacks of classical regression. We compare the performance of robust and classical approaches to two recently published methods for estimating heritability and predictive accuracy of GP using simulation of several plausible scenarios of random and block data contamination with outliers and commercial maize and rye breeding datasets. The proposed robust approach enhances the predictive accuracy of heritability and genomic prediction while alleviating the need for performing outlier detection for a broad range of simulation scenarios. Analyses of empirical maize and rye datasets further reinforce the stability and reliability of the robust approach in the presence of outliers or missing data.

Keywords: Outliers, Genomic prediction, Predictive accuracy, Heritability, Robust estimation.

Zero-adjusted cure rate regression survival models

F. Louzada-Neto

Institute for Mathematical Sciences and Computing, University of São Paulo, Brazil

Cure rate survival models are widely used in practice, where certain individuals are not susceptible to the occurrence of the event of interest. We extend the cure rate survival models by incorporating a proportion of early failures or zero-adjustment in the modeling, the so-called zero-adjusted cure rate regression survival models, providing a new general class of survival models with a strong practical appeal. Focusing in obstetrical studies, an issue that should be considered in the modeling is the inclusion of women for whom the duration of labor cannot be observed due to fetal death. In banking loans, it is observed the propensity to fraud in lending loans in the presence of straight-to-default customers that never pay their loans. In both cases, generating a proportion of lifetimes equal to zero. Maximum likelihood and Bayesian estimation procedures reach parameter estimation. A comprehensive simulation study is carried out to assess the performance of the estimation procedure. Our modeling is illustrated on real datasets. This work is co-authored by Gleici da Silva Castro Perodona, Mauro Ribeiro de Oliveira and Hayala Cristina Cavenague de Souza and Pedro Luis Ramos.

Keywords: Cure-Rate Models, Survival Analysis, Zero-Adjustment.

“Finance and Growth” Re-Loaded

L. Méndez[a] and S. Ongena[b]

[a] *Universidad Autónoma Metropolitana, Mexico City* [b] *University of Zurich, Siwiss Finance Institute, KU Leuven and CEPR*

We assess the relationship between finance and growth over the period 1980-2014. We estimate a cross-country growth regression for 48 countries during 20 periods of 15 years starting in 1980 (to 1995) and ending in 1999 (to 2014). We use OLS and IV estimations and we find that: 1) overall financial development had a positive effect on economic growth during all periods of our sample, i.e., we confirm that from 1980 to 2014 financial services provided by the various financial systems were significant (to various degrees) for firm creation, industrial expansion and economic growth; but that, 2) the structure of financial markets was particularly relevant for economic growth until the financial crisis; while 3) the structure of the banking sector played a major role since; and finally that, 4) the legal system is the primary determinant of the effectiveness of the overall financial system in facilitating innovation and growth in (almost) all of our sample period. Hence, overall our results suggest that the relationship between finance and growth matters but also that it varies over time in strength and in sector origination.

Keywords: Financial Structure, Economic Growth, Financial Development.

References

[1] T. Beck, A. Demirgüç-Kunt, and D. Singer (2013). Is Small Beautiful? Financial Structure, Size and Access to Finance, *World Development*, **52**, 19–33.

[2] A. Demirgüç-Kunt, and R. Levine (2001). *Bank-based and Market-based Financial Systems: Cross-country comparisons.* In A. Demirgüç-Kunt, and R. Levine (Eds.), Financial Structure and Economic Growth: A Cross-Country Comparison of Banks, Markets, and Development. MIT Press, Cambridge, MA, 82–140.

[3] R. Levine (2005). *Finance and Growth: Theory and Evidence.* In P. Aghion, and S. N. Durlauf (Eds.), Handbook of Economic Growth, Elsevier North Holland, 866–934.

[4] P. Wachtel (2018). Credit Deepening: Precursor to Growth or Crisis? In *Comparative Economic Studies*, **60**(1), 34–43.

Time series models and principal component analysis techniques to estimate the impact of particulate matter on health and quality of life

Milena Machado[a], V. A. Reisen[b], Jane Meri Santos[b] and P. Bondon[c]

[a] *Federal Institute of Espírito Santo (IFES),* [b] *University of Espírito Santo (UFES),* [c] *CentraleSupélec*

Adverse health effects, such as respiratory and cardiovascular diseases, caused by inhalable particles have been of great concern due to the high exposure risk even at relatively low concentrations of air pollutants, especially particulate matter [1]. Several studies for example [1,2], applied regression models to quantify the relationship between annoyance and air pollutants. All that studies have applied linear regression and logistic regression techniques however, they considered only one pollutant in the model as a single covariate. As pointed out by [3], this kind of analysis not considering this characteristic of the data may lead to serious consequences on the health of the population under study. The aim of this study is to estimate the risk between exposure to particulate matter concentrations (SPM, PM_{10} and TSP) and perceived annoyance, reported by respondents in Vitoria Region (Brazil). A survey by phone (panel survey) from 2011 to 2014 with a representative sample of the region was conducted. It was examined the behaviour of inhalable particulate matter, as PM_{10}, total suspended particles (TSP) and settleable particulate matter (SPM), monitored in the air quality stations. The data analysis showed time and cross correlation in and between pollutants data. Thus, the variables of interest were modelled using techniques of time series analysis (the multivariate VAR models) and multivariate analysis, namely principal component analysis (PCA) and logistic regression (LOG). This combination resulted in a hybrid model denoted here as LOG-PCA-VAR which allows to estimate the Relative Risk (RR) by handling multipollutant effects. The results showed the relative risk values significant for each pollutant demonstrating that there is a strong association between the observed concentration levels of PM_{10}, TSP, SPM and perceived annoyance reported by respondents. For comparison purpose, the estimate of the RR was also calculated using single pollutant models and ignoring the temporal correlation of the air pollutants series. The RR estimates were not significant which may lead to a false negative interpretation, that is, the test indicated that the perceived annoyance reported by respondents is not associated with air pollution which can be considered as a spurious result.

Keywords: time series, annoyance, relative risk.

References

[1] M. Machado, J.M. Santos, V. A. Reisen, N. C. Reis, I. Mavroidis, and A. T. Lima (2018). A new methodology to derive settleable particulate matter guidelines to assist policy-makers on reducing public nuisance. *Atmospheric Environment*, **182**, 242–251.

[2] A. H. Amundsen, R. Klaeboe, and A. Fyhri (2008). Annoyance from vehicular air pollution: Exposure-response relationships for Norway. *Atmospheric Environment*, **42**, 679–688.

[3] J. B. Souza, V. A. Reisen, V. A., G. C. Franco, M. Ispány, P. Bondon, J. M. Santos (2018). Generalized additive models with principal component analysis: an application to time series of respiratory disease and air pollution data. *JRSS Series C*, **67**, 453–480.

Sinh-skew-normal/Independent Regression Models

R. Maehara[a], H. Bolfarine[b], F. Vilca[c] and N. Balakrishnan[d]

[a] *Universidad del Pacífico,* [b] *Universidade Estadual de São Paulo,* [c] *Universidade Estadual de Campinas,* [d] *McMaster University*

Skew-normal/independent (SNI) distributions form an attractive class of asymmetric heavy-tailed distributions that also accommodate skewness. We use this class of distributions here to derive a generalization of sinh-normal distributions [1], called the sinh-skew-normal/independent (sinh-SNI) distribution. Based on this distribution, we then propose a general class of nonlinear regression models, generalizing the regression models of [2] that have been used extensively in Birnbaum-Saunders regression models. The proposed regression models have a nice hierarchical representation that facilitates easy implementation of an EM-algorithm for the maximum likelihood estimation of model parameters and provide a robust alternative to estimation of parameters. Simulation studies as well as applications to a real dataset are presented to illustrate the usefulness of the proposed model as well as all the inferential methods developed here.

Keywords: Nonlinear regression, Birnbaum-Saunders distribution, EM-algorithm, Robust estimation, Skew-normal/Independent distribution, Sinh-normal distribution.

References

[1] J. R. Rieck (1989). *Statistical Analysis for the Birnbaum-Saunders Fatigue Life Distribution.* Clemson University, South Carolina.

[2] J. R. Rieck, J. R. Nedelman (1991). A log-linear model for the Birnbaum-Saunders distribution. *Technometrics*, **33**, 51–60.

Finite Mixture of Birnbaum-Saunders distributions using the k-bumps algorithm

Luis Benites[a], Rocío Maehara[b], Filidor Vilca[c] and Fernando Marmolejo-Ramos[d]

[a] *Pontificia Universidad Católica del Perú,* [b] *Universidad del Pacífico,* [c] *Universidade de Campinas,* [d] *The University of Adelaide*

Mixture models have received a great deal of attention in statistics due to the wide range of applications found in recent years. This paper discusses a finite mixture model of Birnbaum-Saunders distributions with G components, as an important supplement of the work developed by Balakrishnan et al. (2011) [1], who only considered two components. Our proposal enables the modeling of proper multimodal scenarios with greater flexibility, where the identifiability of the model with G components is proven and an EM-algorithm for the maximum likelihood (ML) estimation of the mixture parameters is developed, in which the k-bumps algorithm is used as an initialization strategy in the EM algorithm. The performance of the k-bumps algorithm as an initialization tool is evaluated through simulation experiments. Moreover, the empirical information matrix is derived analytically to account for standard error, and bootstrap procedures for testing hypotheses about the number of components in the mixture are implemented. Finally, we perform simulation studies and analyze two real datasets to illustrate the usefulness of the proposed method.

Keywords: Birnbaum-Saunders distributio, EM algorithm, k-bumps algorithm, Maximum likelihood estimation, Finite mixture

References

[1] N. Balakrishnan, R. Gupta, D. Kundu, V. Leiva, and A. Sanhueza (2011). On some mixture models based on the birnbaum-saunders distribution and associated inference. *Journal of Statistical Planning and Inference*, **141**, 2175–2190.

Poster: May 30, 10:00 - 10:30, Posters Room

Coordination of algorithm for the Lasso and Ridge techniques

Marin Erisbey, M. Fernando and Ramirez Carlos

Universidad Tecnológica de Pereira

In this work the coordinate descent algorithm was implemented for the regularization type lasso and the analytical solution of Ridge Regression. Additionally, Kernels was used for the implementation of the L1VM. The previous algorithms were compared with Ridge regularization and with L2VM. The Lasso-type penalized regression is a linear regression technique proposed by Tibshirani [1], capable of selecting variables, a very important task when the number of predictors p exceeds the number of samples n. Lasso is a regularized linear regression technique, like the Ridge regression, with the slight difference in the penalty, since it makes use of the l1 rule instead of the l2 norm, which has important consequences.

Every problem of regularization raises two important questions, the first is, what is the most efficient method to minimize the objective function? The second question is how to choose the most appropriate value of the adjustment parameter? From the start, the second question could be answered, which has to do with cross-validation. The answer to the first question is not so obvious, because the standard methods of regression include the diagonalization of matrices, matrix inversion or at least the solution of large systems of linear equations that are the result of having many input variables (predictors), becoming intractable problems. In this work we will use the coordinate algorithm descent which is a very simple algorithm, with a high stability and speed of convergence.

Keywords: Kernels, Coordinate Descent, Regularization.

References

[1] R. Tibshirani (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 267–288.

[2] J. Fan, and R. Li (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association.*

[3] A .E. Hoerl, and R. W. Kennard (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, **12**(1), 55–67.

Disjoint orthogonal components in Tucker models

Carlos Martin-Barreiro[a,b] and John Ramirez-Figueroa[a,b]

[a] *ESPOL Polytechnic University, ESPOL, FCNM, Guayaquil, Ecuador,* [b] *University of Salamanca, Salamanca, Spain*

In this paper, 3 algorithms (CBPSO-TuckALS3, CBPSO-TuckALS2, and CBPSO-TuckALS1) are presented for the calculation of disjoint orthogonal components in the 3 Tucker models used to perform a PCA (Principal Component Analysis) in 3-way tables (tensors). The design of the algorithms is based on the matrix equations of frontal cuts, horizontal cuts and vertical cuts of the Tucker models for 3-way tables. In addition, all algorithms are based on the well-known ALS algorithm used for the classical calculation of components in Tucker models, and are also supported by a binary and constrained PSO (Particle Swarm Optimization) algorithm (CBPSO) that is used for the calculation of disjoint components in 2-way tables. The detail and the role that both the ALS and the CBPSO play in the proposed algorithms are explained. Computational experiments are then included, where the quality of the obtained solutions and their interpretability in the context of the problem are analyzed. Finally, some recommendations are made with the intention of guiding the analyst of the data in the scenarios in which the use of these 3 algorithms is an important alternative to consider.

Keywords: Disjoint components, Tensors, Tucker models.

References

[1] L. R. Tucker (1966). Some Mathematical Notes on Three-Mode Factor Analysis. *Psychometrika*, **31**, 279–311.

[2] H. Kiers and I. Van Mechelen (2001). Three-Way Component Analysis: Principles and Illustrative Applications. *Psychological Methods*, **6**, 84–110.

[3] P. Giordani, H. Kiers, and M. A. Del Ferraro (2014). Three-Way Component Analysis Using the R Package. *Journal of Statistical Software*, **57**(7), 1–23.

Parallel Statistical Algorithms: Careful Design and Important Decisions

A. Martínez-Ruiz[a] and C. Montañola-Sales[b]

[a] *Universidad San Sebastián,* [b] *Universidad Ramón Lull*

When statistical experiments are executed in High Performance Computing Systems, performance depends on many factors. A proper mathematical design of the statistical algorithm, the complexity of the implementation in a programming language and encoding, the task versus data distribution scheme, the degree of parallelism of the implementation and data movement are important decisions as well as the linear algebra library for processing dense matrix operations. In this work, we present a brief review on the approaches that are used to parallelize methods and run algorithms in a parallel computer environment and define a set of items to be considered to find the most suitable solution in this context. We illustrate with some simple examples using R-project, we especially consider data distribution and movement, and data generation process for simulations and reproducibility purposes.

Keywords: Parallel computing, statistical algorithms.

Air masses origins and sources from southern Ecuador using HYSPLIT analysis

D. Morán-Zuloaga[a], D. Hernick[b], J. I. Valdez-Hernández[c], M. H. Cornejo[a], J. Cáceres[a], K. Morán[d] and P. Hernick[d]

[a]*Escuela Superior Politécnica del Litoral (ESPOL),* [b]*Full Harvest Technologies Inc,* [c]*Colegio de Postgraduados México,* [d]*Master in Water Steward Freshwater Society*

The countryside usually considered less polluted than cities and the reason behind it correspond to the forest surrounded those ecosystems. The constant modification of land use and land cover around the world make those assumptions less unreliable. The current situation in south Manabí is uncertain. The aim of the present work is to unveil the importance of environmental atmospheric modeling to provide a clear view of events at in situ studies. The following study took a farmland as origin point a 1.67 S and 80.55 W at 400 m asl; by using a Hybrid Single Particle Lagrangian Integrated Trajectory Model HYSPLIT to characterized the air masses parcels 96 hours backward trajectories, arrival height of 1000 m. Data was processed by using HYSPLIT webtool (https://www.ready.noaa.gov/HYSPLIT.php, last visited: 2017-09-30) and then processed with r project 3.5.1 and rstudio version 1.1.463 and the packages: maps, openair, reshape; and igor pro from Wavemetric version 6.37. Furthermore, Aerosol Optical Depth, NO2 was collected by using Giovanni by Goddard Earth Science Data and Innovation web interface (https://giovanni.gsfc.nasa.gov/giovanni/, last visited 2018-04-01). We provided a close approximation of air masses origins and sources transported into the forest and farmlands in south Manabí; moreover, we infer about the possibility of local pollution incidences. The preliminary results revealed a dominant air masses transported at around ± 20 km from west and north that provides marine breeze from the Pacific Ocean during the wet season. In contrast, air masses came from the south west during the dry period. Therefore, the use of global tools, satellite data, for understanding the surrounding of the green cover.

Keywords: HYSPLIT, atmospheric models, tropical forest

Simulation models to support preliminary electoral results program for the Mexican Electoral Institute

D. F. Muñoz[a], H. Gardida[b], H. Velásquez[b] and J. D. Ayala[a]

[a] *Department of Industrial and Operations Engineering, Instituto Tecnológico Autónomo de México,* [b] *Technical Unit for Information Services, Instituto Nacional Electoral de México*

On July 1st, 2018, federal elections for president, senators and deputies took place in Mexico and, in most states, elections for state governors and representatives took also place in the same polling booths. The Technical Unit for Information Services (UNICOM) of the Instituto Nacional Electoral (INE) of Mexico has the responsibility for planning and implementation of the Preliminary Electoral Results Program (PREP) for federal elections and, for the 2018 elections UNICOM developed a forecasting model for the flow time of published booth results, based on simulation models that were developed using a special-purpose simulation software and C++ subroutines for fast simulation of queues.

The PREP is the program that publishes data and images from the scrutiny and computation forms (SCF) that were filled by booth representatives to summarize the election results from the corresponding booth and, according to [1], the objective is to timely inform the electoral results to interested parties, media and the public, under the principles of security, transparency, reliability and integrity. Before the PREP operational process (POP) begins, required tasks are performed by booth representatives, including scrutiny, computation and filling of operations notebooks and SCF. The main tasks for resource pools that executed the POP took place in booths (31,207 scrutiny and computation trainers), 186 Collection and Transportation Centers, 300 Collection and Data Transmission Centers, and 2 Capture and Verification Centers with more than 4,000 digitizers. The expected total number of booths was 156,840, with 3 federal SCF per booth. This large system was modeled using the special-purpose software Simio [2] because of its advantages for developing a detailed model in a relatively short time. Although main simulation runs using SIMIO were performed under parallel computing, in order to reduce large computational times required by this complex model, UNICOM also developed a less detailed model that uses C++ subroutines for fast simulation of queues. The development of this model was facilitated by using the SIMIO model for verification and fast development.

Keywords: Stochastic simulation, Elections results, Flow rate forecasting, Transparency.

References

[1] Cámara de Diputados del H. Congreso de la Unión (2017). Ley general de elecciones y procedimientos electorales. In *Diario Oficial de la Federación del 27/01/2017*, Secretaría de Servicios Parlamentarios (eds.), Mexico City, 1–217.

[2] J. S. Smith, D. T. Sturrock, and W. D. Kelton (2019). *Simio and Simulation, Modeling, Analysis, Applications.* Simio LLC, Sewickley, Pennsylvania.

Classification for geostatistical functional data using depth.

A. V. Navarrete[a], R. D. Guevara[a], M. P. Bohorquez[a] and J. Bacca[b]

[a] *Dept. of Statistics, Universidad Nacional de Colombia, Bogotá,* [b] *Dept. of Electrical and Electronic Engineering, Universidad Nacional de Colombia, Bogotá*

This paper presents a classification proposal for geostatistical functional data using depth functions, which takes into account spatial dependence. Concepts such as modified band depth (MBD), multivariate depth and distance are implemented in the supervised classification method, in order to obtain robust classification outcomes. The spatial dependence is incorporated in the computation of the depths through weights, which are based on the spatial covariance matrix. The covariance for every pair of functional observations is modelled using the score vectors associated with the empirical functional principal components. An application illustrates how to classify a set of signals generated by the brain of a person when thinking of a vowel, taking into account the location of the electrodes. To make the analysis, the signals collected by the electrodes are transformed through power spectral density (PSD) which shows the strength of the variations as a function of frequency. Finally, a comparison is shown between methods including or not the spatial dependence between the electrodes.

Keywords: Classification, Depth functions, Geostatistical functio nal data.

References

[1] A. Balzanella, and R. Elvira (2015). A depth function for geostatistical functional data. In *Advances in Statistical Models for Data Analysis*, Springer, Cham, 9–16.

[2] M. Bohorquez, R. Giraldo, and J. Mateu (2017). Multivariate functional random fields: prediction and optimal sampling. *Stochastic environmental research and risk assessment*, **31**(1), 53–70.

[3] S. López-Pintado and J. Romo (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, **104**, 718–734.

Robust approaches for blind source separation

K. Nordhausen

Vienna University of Technology

In blind source separation (BSS) it is assumed that the observable multivariate phenomenon is a linear mixture of unobservable latent factors. The goal of BSS is to estimate these latent factors based on the observations alone. A task clearly not possible without further assumptions. Many different BSS approaches were suggested in the literature and the most popular is undoubtedly independent component analysis (ICA) where it is assumed that the latent factors are independent and higher order moments are exploited in the separation. Other BSS approaches like second order source separation or nonstationary source separation exploit other features of the latent factors. One of the key ideas of many BSS approaches is to find scatter functionals which are diagonal for the latent factors and then find a transformation which makes them diagonal again when being computed for the observed data.

Multivariate robust scatter functionals are usually however developed having an elliptical model in mind and not BSS models. In this talk we discuss properties needed for robust functionals to be useful in BSS and review different approaches suggested so far to robustify different BSS methods.

Keywords: Scatter functionals, independent component analysis

References

[1] K. Nordhausen and D. E. Tyler (2015). A cautionary note on robust covariance plug-in methods. *Biometrika*, **102**, 573–588.

[2] P. Ilmonen, K. Nordhausen, H. Oja, and F. J. Theis (2015). An affine equivariant robust second order blind source separation Method. In *Latent Variable Analysis and Signal Separation*, LNCS 9237, 328–335.

[3] K. Nordhausen (2014). On robustifying some second order blind source separation methods for nonstationary time series. *Statistical Papers*, **55**, 141–156.

ANOVA test for correlated functional data applied to fine particulate matter measurements on air

J. Olaya Ochoa and D. P. Ovalle

Universidad del Valle, Cali, Colombia

Environmental authorities have defined fine particulate matter as particles suspended in air whose aerodynamic diameter is less than 2.5 μm (usually denoted as $\text{PM}_{2.5}$). We have daily information on this air pollutant coming from three surveillance stations and we want to check whether the levels of $\text{PM}_{2.5}$ are the same at the three places or not.

Datasets consist of daily records of as much as 24 observations per day, and so we have the typical framework on which Functional Data Analysis plays a key role. The reason is that $\text{PM}_{2.5}$ levels are originated from a continuous phenomenon and that we collect discrete observations from it. Then, using those discrete observations from a continuous phenomenon, we get a curve using smoothing techniques. This way, we have one curve, rather than one real number, per day.

We conducted the analysis using a Functional Analysis of Variance. Statistical comparison of the means from more than two populations is a very well-known problem if we are dealing with scalar values. However, it is not the case as soon as we move toward the observation of variables whose values are curves, rather than scalars. Finally, since data from these stations were likely not to be independent, we needed to get estimations of the functional correlation between stations. Then we introduced such correlation structure into the analysis. Final results indicate statistically significant differences among the three stations.

Keywords: $\text{PM}_{2.5}$, FANOVA, FDA.

References

[1] F. Ferraty (2013). *Recent Advances in Functional Data Analysis and Related Topics*, New York, NY: Physica-Verlag.

[2] T. Hastie, R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, 2nd. edn, Springer.

[3] P. Kokoszka and M. Reimherr (2017). *Introduction to Functional Data Analysis* , New York, NY: Chapman & Hall/CRC.

[4] J. O. Ramsay and B. W. Silverman (2005). *Functional Data Analysis*, 2nd. edn, Springer.

[5] J.-T. Zhang (2014). *Analysis of Variance for Functional Data*, 2nd. edn, Chapman & Hall/CRC.

Clustering using Functional Data Analysis for Honeybees Daily Activity Data

E. Acuña[a], W. Quispe[b], R. Trespalacios[b], V. Palomino[b], C. Vega[a], R. Mégret[c] and J. Agosto[d]

[a] *Department of Mathematical Science, University of Puerto Rico at Mayaguez* [b] *Doctoral Program in CISE, University of Puerto Rico at Mayaguez,* [c] *Department of Computer Science, University of Puerto Rico at Rio Piedras,*[d] *Department of Biology, University of Puerto Rico at Rio Piedras.*

In recent years, Functional data analysis (FDA) is being used to analyze, model and predict time series data. Main aspects of FDA include the choice of smoothing technique, data reduction, adjustment for clustering, functional linear regression and prediction methods.

Usually, time series data are treated as multivariate data because they are given as a finite discrete time series. This multivariate approach completely ignores important information about the smooth functional behavior of the generating process that underpins the data. It also suffers from issues associated with highly correlated measurements within each functional object. The basic idea behind FDA is to express discrete observations arising from time series in the form of a function (to create functional data) that represents the entire measured function as a single observation, and then to draw modeling and/or prediction information from a collection of functional data by applying statistical concepts from multivariate data analysis.

In this work, we present different modeling approaches for the clustering of a functional data set of honeybees activities. Our procedures are applied to data collected during experiments carried out in France and Puerto Rico.

Keywords: Functional Data Analysis, Clustering, Honeybees Behavior.

Analysis and comparison of similarity measures and indices for image quality assessment

M. L. Pappaterra and S. M. Ojeda

Facultad de Matemática, Astronomía, Física y Computación, Universidad Nacional de Córdoba

The amount of digital imagery is rapidly increasing every year to the extent that subjective assessment of image quality has become virtually impossible due to time and cost constraints. Furthermore, in digital image processing there is a need to compare the performance of different image processing algorithms by comparing the quality of their output images. To overcome these and similar problems many objective image quality indices were developed. Although many of these indices have been proposed, they stem from different theoretical frameworks, and thus their application scenarios are different, and they may serve different purposes. To the best of our knowledge, this is the first work to compare a large number of indexes, analyzing their differences in performance and assessing which index is more suitable in which application scenario. We select, analyze and compare several full-reference indices and measures: the mean square error (MSE) and root mean square error (RMSE), the signal to noise ratio (SNR), peak signal to noise ratio (PSNR) and weighted signal to noise ratio (WSNR), the noise quality measure (NQM) and visual information fidelity (VIF), the universal quality index (UQI), the structural similarity index (SSIM) and multi-scale structural similarity index (MSSIM), the gradient magnitude similarity mean (GMSM), the gradient magnitude similarity deviation (GMSD) and the codispersion coefficient based CQ-Index. Our Python implementation of all these indices can be found at https://github.com/lucia15/IQA-metrics. We use Kendall's Tau and Spearman's Rank Correlation Coefficient and other non-parametric correlation tests and methods in order to determine the best procedures for comparing digital images, for their mathematical and statistical properties and their ability to emulate the Human Visual System (HVS).

Keywords: Image Quality Assessment, Image analysis, Measures of association

References

[1] Z. Wang and A. Bovik (2009). Mean squared error: love it or leave it? - a new look at fidelity measures, *IEEE Signal Process Magazine*.

[2] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo (2014). Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, **30**, 57–77.

[3] M. Ojeda Silvia, O. R. Vallejos, and P. W. Lamberti (2012). Measure of similarity between images based on the codispersion coefficient. *Journal of Electronic Imaging*.

Gaining robustness and detecting outliers applying Mixtures of Gaussian and heavy-tailed distributions in Bayesian inference

A. Posekany

Department for Clinical Neurosciences and Preventive Medicine, Danube University Krems, Austria

For linear models, normal distributions are generally assumed which many data in fields like biology or economics do not fulfill even though the observation sizes are large. For robustification of Bayesian inference, we employ mixture models which simultaneously allow robust density fit and outlier detection. By mixing standard Gaussian components with Student's t distributed one can identify the over-dispersed part of data which is partially very noisy. As one application, we present data from microarrays, which show a complicated, over-dispersed noise behaviour. Microarrays have found their way from research into clinical practice which makes detecting problems in applied analyses even more relevant. Our goal is to improve the inference of differential expression for identifying extremely dispersed data and in addition to recognize whether subsets creating this data have a common origin granting a means of quality control.

Keywords: Outliers, Mixed models, Bayesian inference, Gene expression microarrays.

An alternative method for obtaining disjoint principal components by particle swarm optimization.

John Ramirez-Figueroa[a,b], Carlos Martin-Barreiro[a,b]

[a] *ESPOL Polytechnic University, ESPOL, FCNM, Guayaquil, Ecuador,* [b] *University of Salamanca, Salamanca, Spain*

The disjoint principal components analysis allows to determine components that are lin- ear combinations of subsets that constitute a partition of the set of variables considered in the problem. Each of the disjoint principal components has zeros in the positions of the variables not considered, which facilitates their interpretation in terms of the origi- nal variables. The proposed new method is named Constrained Binary Particle Swarm Optimization Disjoint Component (CBPSO DC). This method uses particles represented by binary stochastic matrices, within of discrete feasible solutions space. Through an evolutionary algorithm it implements a stochastic type optimization designed to find high quality solutions in situations of high computational complexity. The proposed algorithm initiates randomly generating a population of particles that iteratively evolve to reach the global optimum, minimizing the objective function that depends on the disjoint princi- pal components. Diferent types of topologies and other parameters of the particles are tested. Numerical results are provided confirming the quality of the solutions obtained by the new method.

Keywords: Disjoint Component, Particle Swarm Optimization, Singular Value Decomposition.

References

[1] C. Ferrara, F. Martella, and M. Vichi (2018). Probabilistic Disjoint Principal Component Analysis. *Multivariate Behavioral Research*, Published online.

[2] S. Lee, S. Soak, S. Oh, W. Pedrycz, and M. Jeon (2008). Modified binary particle swarm optimization. *Progress in Natural Science*, **18**(9), 1161–1166.

[3] M. Vichi and G. Saporta (2009). Clustering and disjoint principal component analysis. *Computational Statistics and Data Analysis*, **53**(8), 3194–3208.

Biclustering algorithms for high-frequency financial time series

N. Ravishanker

Department of Statistics, University of Connecticut, USA

As high-frequency transaction-by-transaction data are widely available, it is critical for researchers and investors to dynamically study patterns of co-movement over multiple trading days. Exploring high frequency transaction level financial data is of considerable interest to researchers and investors. To this end, we have developed a multiple day time series biclustering algorithm based on aggregating the transaction-by-transaction data to regular (one to five minute) time intervals within each trading day. We examine the robustness of co-movement probabilities of selected m-tuples of stocks to stay within the same bicluster over multiple trading days to a) the sampling aggregation frequency and b) the bliclustering metric. Additionally, we describe an approach to describe patterns and monitor the structure of high-dimensional daily or weekly time series that track linkages between any given m-tuple of stocks over a long time period. This is joint work with Jian Zou and Hiatao Liu from Worcester Polytechnic Institute.

The PINAR$(1, 1_S)$ model

P. R. Prezotti F.[a,b,c], V. A. Reisen[b], P. Bondon[c] and M. Ispány[d]

[a] *Federal Institute of Espírito Santo (IFES),* [b] *University of Espírito Santo (UFES),* [c] *CentraleSupélec,* [d] *University of Debrecen*

We introduce a new class of models based on the well known Integer Autoregressive (INAR) models ([1], [2], [3]) for count time series with Poisson and Geometric innovations which have a periodic and seasonal second-order autoregressive structure. Statistical properties of the model, such as mean, variance, marginal and joint distributions, are discussed. The Moments-based (Yule-Walker equations), the conditional least squares and quasi-maximum likelihood method of parameters estimation are presented. Their performances are investigated through Monte Carlo simulations, and we present a proof of consistency and asymptotically normality of the estimators. The usefulness of the Periodic INAR, the PINAR$(1, 1_S)$, model is verified in an application to a real data referring to the daily number of visits of children with respiratory problems (International Classification of Diseases ICD-10) to the emergency service of the public health care system of the region of Vitória, Espírito Santo, Brazil. A section is focused on the forecast purposes.

Keywords: INAR models, periodic stationarity, estimation methods.

References

[1] M. A. Al-Osh and A. A. Alzaid (1987). First-order integer-valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis*, **8**(3), 261–275.

[2] A. A. Alzaid and M. Al-Osh (1990). An integer-valued pth-order autoregressive structure (INAR(p)) process. *Journal of Applied Probability*, **27**(2), 314–324.

[3] D. Jin-Guan and L. Yuan (1991). The integer-valued autoregressive (INAR(p)) model. *Journal of time series analysis*, **12**(2), 129–142.

Robust factor modeling for high-dimensional time series

V. A. Reisen

Department of Statistics, Federal University of Espírito Santo, Vitória, Brazil

This paper considers the factor modeling for high-dimensional time series with short and long-memory properties and in the presence of additive outliers. The factor model studied by Lam and Yao (2012) is extended to the presence of additive outliers and series with long-memory. The estimators of the number of factors are obtained by the robust covariance matrix. The methodology is analyzed in terms of the convergence rate of the number factors and by means of Monte Carlo simulations. Application with the aiming to reduce the dimensionality of the data the pollutant PM10 in the Greater Vitória region (ES, Brazil).

Robust estimation in beta regression via maximum L_q-likelihood

Terezinha K. A. Ribeiro and Silvia L. P. Ferrari

Department of Statistics, University of São Paulo, Brazil

Beta regression models (Ferrari and Cribari-Neto, 2004; Smithson and Verkuilen, 2006) are widely used for modeling continuous data limited to the unit interval, such as proportions, fractions, and rates. The inference for the parameters of the beta regression model is commonly based on maximum likelihood estimation. However, it is known to be sensitive to discrepant observations. In some cases, one atypical observation can lead to severe bias and erroneous conclusions about the features of interest. In this work, we develop a robust estimation procedure for beta regression models based on the maximization of a tilting L_q-likelihood proposed by Ferrari and La Vecchia (2012). The new estimator offers a trade-off between robustness and efficiency through a tuning constant. The proposed estimator and the maximum likelihood estimator are compared through Monte Carlo simulations in the presence and absence of contamination in the data. The simulation suggests marked robustness of the new estimator with little loss of efficiency. An application to real data is presented and discussed to show the robustness and applicability of the new estimator. Finally, we also develop a diagnostic plot based on the robust estimator to assess the adequacy of the assumed model and still detect outlying observations.

Keywords: Beta regression, L_q-likelihood, Robustness.

References

[1] S. L. P. Ferrari and F. Cribari-Neto (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**, 799–815

[2] D. Ferrari and D. La Vecchia (2012). On robust estimation via pseudo-additive information. *Biometrika*, **99**, 238–244.

[3] M. Smithson and J. Verkuilen (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, **11**, 54–71.

Bayesian Estimation in the Additive Hazard Model

E. E. Alvarez[a,b] and M. L. Riddick[c]

[a]*FI-UNLP,* [b]*IC-FCEyN-UBA-CONICET,* [c]*DM-FCEx-UNLP-CONICET*

Suppose we have a sample of n individuals who may experience a terminal event over a window $[0, u]$. We denote by T_i^* the true, possibly latent, time to occurrence for the i-th individual. Because some individuals experience censoring at times C_i, their duration until the event is observed only when $C_i \geq T_i^*$. In classical Survival Analysis, it is of interest to study the soujourn times as related to observed individual covariates, which we assume time-independent and denote by Z_i. In the literature, models for survival data typically focus on the so-called hazard rate, which we assume takes the additive form $\lambda(t, \beta) = \lambda_0(t) + z'\beta$ due to Aalen (1980) where $\lambda_0(.)$ is the *baseline hazard function* and β is a vector of unknown coefficients. Alternative approaches abound in the literature, the most common being Cox's (1972)proportional hazards model and the Accelerated failure time model.

It is our goal in this study to propose a Bayesian method of estimation for the semiparametric Additive Hazards Model (AHM) under right-censoring. With this aim, we review the AHM and introduce the likelihood function, so that we comment on the challenges posed by estimation from the full likelihood. Thus we discuss an alternative approach based on a hybrid Bayesian method that exploits Lin and Ying (1994) estimating equation approach and simple tractable priors for the parameters. For the simplest case, we obtain the posterior distributions and we provide algorithms for the estimators. We illustrate our method with a simple dataset in the medical field.

Keywords: Additive Hazards Model, Survival Analysis, Bayesian Inference.

Robust estimation in partially nonlinear models

D. Rodriguez[a] and A. Muñoz[b]

[a] *Universidad de Buenos Aires and CONICET,* [b] *Instituto Tecnológico de Buenos Aires*

Statistical inference for multidimensional random variables commonly focuses on functionals of its distribution that are either purely parametric or purely nonparametric. A reasonable parametric model produces precise inferences, while a badly misspecifed model possibly leads to seriously misleading conclusions.

In the partially nonlinear regression model, one observes the response variable y obeying the model:

$$y = \eta(t) + g(x, \beta) + \varepsilon$$

where (x, t) is a vector of explanatory variables, g is a prespecified function, β is a vector of unknown true parameters to be estimated, η is an unknow true smooth function to be estimared and ε is a random error.

The partially nonlinear model, retains the flexibility of nonparametric models and the interpretability of nonlinear parametric models. As special cases of partially nonlinear models, partially linear models are popular in the literature.

In this talk, we introduce robust estimates for the parametric and nonparametric components for the partially nonlinear model. The proposed estimators are based on a three step procedure. We show some asymptotic properties such as that the consistent of both estimatores and the asymptotic distribution of the estimators of the parametric component. Also, we study the behaviour of the proposal, through a Monte Carlo study where we compare the performance of our estimators with that of the classical ones. We illustrated our proposal with a real dataset.

Keywords: Asymptotic Properties, Partly Nonlinear Models, Smoothing Techniques.

Spatial clustering based on the pair correlation LISA functions: A functional approach

F. J. Rodríguez-Cortés[a], E. Romano[c], J. A. González[b] and J. Mateu[b]

[a] *Escuela de Estadística, Universidad Nacional de Colombia, Medellín, Colombia,* [b] *Department of Mathematics, University Jaume I, Castellón, Spain,* [c] *Department of Mathematics and Physics, Universitá della Campania Luigi Vanvitelli, Caserta, Italy*

The second-order product density function provides a global measure of the covariance structure by summing up over the contributions from each event in a spatial point process. Furthermore, the pair correlation function can be interpreted as its standardised version and it can be used as a tool to discriminate among several spatial point process structures taking into account its intrinsic inhomogeneity. Individual contributions of the global estimator for the pair correlation function can be considered as a measure of clustering and can be used as an analytic exploratory data tool to examine individual points in a point pattern in terms of how they relate to their neighbouring points. Pair correlation LISA functions can then be grouped into bundles of similar functions using multivariate clustering techniques according to a particular distance between them. We propose a functional approach for classifying events with a similar local structure by means of pair correlation LISA functions. The main idea is to work with each event of a spatial point pattern by considering the LISA function as an attached functional attribute. We extend functional clustering method to the spatial case in order to classify correlated LISA functions by using an appropriated distance and compare this with the classical L^2 distance. The performance is evaluated through a simulation study and applied to a real earthquake data-set.

Keywords: Clustering, functional distance, functional marked data, local indicators of spatial association, spatial point processes.

Challenges in estimating individual/household level severity parameters with the Food Insecurity Scale (FIES)

Maria Rodriguez[a] and Rafael España[b]

[a] *Instituto de Estadística Aplicada, Universidad de los Andes, Mérida, Venezuela,* [b] *Universidad de los Llanos Ezequiel Zamora, Portuguesa, Venezuela*

The theory of item response provides a tool to approach latent traits such as concerns and stress on the home/individuals derived from the difficulties to get food, situations that would be very difficult to measure under other circumstances. FAO Statistics division and “Voices of Hunger” project have made some adaptations and extensions to the item-response theory to face these challenges; through the survey “Food Insecurity Experience Scale”, a metric is established to measure the food security condition of individuals/households based on the direct responses of people about access to food. The questions are designed to measure “safe access at all times to sufficient food”. The FIES module itself consists of 8 items with dichotomous response (yes/no) and where the person is asked if during a certain period of time has been worried about the ability to get food, if they have had to change their diet or finally to stay without eating at some time due to limited availability of money or other resources. When apply the item-response theory models to the measurement of food insecurity, this postulates that [1]: A) the severity of the food insecure situation of the respondent and the associated with each of the experiences can be located on the same unidimensional scale and B) while more severe is the severity food insecurity situation higher probability the person express experiences associated food insecurity experiences. To know the food security level for individual/household the item response theory proposes the estimation of item parameters (β) and individual parameters (Θ) through maximum likelihood. Individual parameters (severity of the food insecurity condition) is based on the raw score as a sufficient statistic. This raw score is understood as the sum of the affirmative responses of each person to the 8 questions of the FIES module. However, there are important challenges in estimating the severity of the food insecurity condition in extreme raw scores using maximum likelihood [1] and the implications this has on estimates of food insecurity especially for the extreme food insecurity level: severe food insecurity. Possible options and their important considerations are raised. The discussion about the potentialities of each inferential framework is vital for the definition of the indicator, in addition to the importance of justifying the most suitable form of estimation.

Key words: Item-Response Theory, Food Insecurity, FIES.

References

[1] FAO. (2016). *Métodos para la estimación de índices comparables de prevalencia de la inseguridad alimentaria experimentada por adultos en todo el mundo.* Voices of the Hungry project. FAO Roma.

[2] K. B. Christensen, S. Kreiner, and M. Mesbah, M. (Edits.) (2013). *Rasch Models in Health.* Great Britain & United States: ISTE Ltd and John Wiley & Sons, Inc, 65–66

MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers

M. Hubert, P. J. Rousseeuw and W. Van den Bossche

Department of Mathematics, KU Leuven, Belgium

Multivariate data are typically represented by a rectangular matrix (table) in which the rows are the objects (cases) and the columns are the variables (measurements). When there are many variables one often reduces the dimension by principal component analysis (PCA), which in its basic form is not robust to outliers. Much research has focused on handling rowwise outliers, i.e. rows that deviate from the majority of the rows in the data (for instance, they might belong to a different population). In recent years also cellwise outliers are receiving attention. These are suspicious cells (entries) that can occur anywhere in the table. Even a relatively small proportion of outlying cells can contaminate over half the rows, which causes rowwise robust methods to break down.

In this paper a new PCA method is constructed which combines the strengths of two existing robust methods in order to be robust against both cellwise and rowwise outliers. At the same time, the algorithm can cope with missing values. As of yet it is the only PCA method that can deal with all three problems simultaneously. Its name MacroPCA stands for **PCA** allowing for **M**issings **A**nd **C**ellwise & **R**owwise **O**utliers. Several simulations and real data sets illustrate its robustness. New residual maps are introduced, which help to determine which variables are responsible for the outlying behavior. The method is well-suited for online process control. The function MacroPCA has been incorporated in the R package *cellWise* [2] on CRAN, which also contains a vignette with real data examples.

Keywords: Detecting deviating cells, Outlier map, Residual map.

References

[1] M. Hubert, P. J. Rousseeuw, and W. Van den Bossche (2019). MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers. *Technometrics*, in press.

[2] J. Raymaekers, P. J. Rousseeuw, W. Van den Bossche, and M. Hubert (2019). *cellWise*: Analyzing Data with Cellwise Outliers. R package version 2.1.0, CRAN.

Seismic activity forecast using Convolutional and LSTM Neural Networks

F. Plaza[a,b], R. Salas[b] and O. Nicolis[c]

[a] *Instituto de Fomento Pesquero,* [b] *Universidad de Valparaíso,* [c] *Universidad Andrés Bello*

Earthquakes are one of the most destructive and hard to predict natural disasters. Chile remains as one of the most seismic countries in the planet, with an average of large-scale earthquake ($> 8°$ in Richter scale) every 10 years. The last major earthquake in Chile was registered in February 27 2010 (8.8° Richter) with a massive physical, psychological and economic impact in the population. Additionally, the physical system from which earthquakes result is highly complex, chaotic, or non-linear, and also, their occurrence depends on a multitude of variables, in most cases are yet unknown. Moreover, contemporary techniques are insufficiently sensitive to allow for precise modelling of future earthquake occurrences. In that sense, Deep Neural Networks (DNN) have state-of-art accuracy for most of the problems where statistical learning models are applied and where a precise mathematical formulation is hard to obtain. This work presents a Deep Learning approach for earthquake modelling and forecast in Chile, using seismological records from 2000 to 2018. Our approach implements Long Short Term Memory Networks (LSTM) for magnitude forecast and Convolutional Neural Networks (CNN) for location prediction. The results showed a good performance of the proposed models, for that matter, to have an approximation or additional information on where and when an event would occur, represent an invaluable tool for managing and designing public policies regarding natural disasters.

Keywords: Earthquakes, Deep Learning, Prediction.

Linear regression models assuming a stable distribution for the response data

D. P. S. Bussola[a], J. A. Achcar[a] and R. M. Souza[b]

[a] *University of Sao Paulo,* [b] *Federal Technological University of Parana*

There have been many studies on the effects of smoking on health, but for this study the data used comes from an observational study where subjects self-select which group they belong to - smoking or non-smoking group and smoking status. Using a spirometer, Forced Expiratory Volume (FEV) was recorded for each subject. The following variables are reported associated: age, height, sex and smoke. To contribute to the understanding of the variables associated with FEV (response), it is analysed using the stable distribution. The use of stable distributions allows skewness and heavy tails. It has been introduced as a generalization of the Gaussian distribution [1], described by four parameters α (index of stability), β (skewness), σ (scale) and μ(location) [2]. Due the lack of closed form for the probability density functions an alternative to the classical approach is the use of Bayesian methods given by [1] using Markov Chain Monte Carlo (MCMC) methods and latent variables [3]. We propose the use of the free available software OpenBugs, given a great computational simplification for determining posterior summaries of interest. The data analysis under classical approach relating the response with the covariates show significative effects on the response and the residual analysis it is observed that the needed assumptions are reasonable, but not totally accepted. Assuming a stable distribution under Bayesian approach with models using known and unknown parameters values α, β and σ, for the results it is observed that only the covariates height and age show significative effects on the response. Assuming the regression model with normal errors all covariates show significative effects on the response. Also, it is observed in this case very similar inference results when compared to the classical approach. Was also done a reanalysis in presence of an outlier, for the results it's observed that the inference results are strongly affected by the presence of the outlier assuming a linear regression model with normal distribution for the error, but the inference results are more similar in presence of outlier assuming the regression model with a stable distribution. That is, get more robust results assuming a regression model with a stable distribution in presence of outliers.

Keywords: Stable Distribution, Bayesian, OpenBugs.

References

[1] D. J. Buckle (1995). Bayesian inference for stable distributions. *Journal of the American Statistical Association*, **90**, 605–613.

[2] R. Casarin (2004). *Bayesian Inference for Mixtures of Stable Distributions.* Cahier du CEREMADE, No. 0428. Available at SSRN: https://ssrn.com/abstract=739791.

[3] P. Damien, J. Wakefield, and S. Walker (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society Series B*, **61**, 331–344.

Lagrangian Relaxation for Design of a Soda Company Distribution System

I. Soria and H. A. Pérez

Universidad Iberoamericana Campus Ciudad de México

This study focuses on the distribution problem of products of a soda company that counts with 4 plants, 5 warehouses and 17 regional centers of distribution [1]. The solution is generated constructing an optimization model and applying the technique of the Lagrangian Relaxation. The lagrangian dual problem is solved by the subgradient method, programmed in GAMS. The determination of a feasible solution to the optimization problem is obtained by means of a heuristic algorithm [2]. The results are compared with the optimal solution of the original problem found through a commercial program in order to determine the quality of the found solution. Finally, the described methodology is applied to four instances corresponding to years $2018 - 2021$ to give continuity to the growth of the drink business in Mexico.

Keywords: Mathematical Programming, Logistics, Optimization

References

[1] I. Soria (2008). *Rediseño de la cadena de abastecimiento de un grupo embotellador de bebidas.* M.S. thesis, ITESM Campus Toluca, México.

[2] J. A. Marmolejo, I. Soria, and H. A. Pérez (2015). A Decomposition Strategy for Optimal Design of a Soda Company Distribution System. *Mathematical Problems in Engineering*, Article ID 891204.

Inference in stochastic mixed-effect models

José Soto[a,b], Saba Infante[a], Franklín Camacho[a] and Rafael Amaro[a].

[a]*Escuela de Ciencias Matemáticas y Computacionales, Universidad de Investigación Yachay Tech, Ecuador,* [b]*Universidad de los Andes, Venezuela*

The biological processes that usually occur in the real world have complex dynamics. Mathematical models that try to describe these phenomena usually have unpredictable behaviors. To model them, systems of stochastic differential equations (SDE) are usually used, which require simultaneous estimation of solution states and parameters. When it comes to experimental studies that consist of repeated measurements on the same experimental unit, the variability between individuals can be modeled by introducing a random effect through the drift of the SDE, known in the literature as the mixed effects model driven by an SDE. The standard estimation method in these models is the maximum likelihood algorithm. Since plausibility does not have an explicit form, it is required to implement algorithms such as: Expectation and Maximization (EM), Newton Rapson or Euler-Maruyama. Other more sophisticated algorithms use Markov Chain Monte Carlo methods such as the Gibbs sampler or the Metropolis-Hastings algorithm. In this article it is proposed to use a methodology inspired by the paradigm of Sequential Monte Carlo techniques; specifically, it is proposed to implement two Bayesian algorithms known as the Kalman filter and the Kalman assembly filter, for the estimation of states-solutions and parameters of the mixed stochastic effect model. The methodology will be illustrated using synthetic and real models. To validate the robustness of the model estimates, some statistical measures of adjustment are implemented and the execution times of the algorithms are calculated. It is expected that the results obtained by this methodology will be of great use in the design of subsequent studies based on real-world biological processes.

Keywords: Mixed effect models stochastic, Kalman filter, Emsamble Kalman filter.

References

[1] G. Evensen (1994). Sequential data assimilation with a nonlinear quasi-geostrophicmodel using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, 10143–10162.

[2] L. Sánchez, S. Infante, V. Griffin, and D. Rey (2016). Spatio-temporal dynamic model and parallelized ensemble kalman filter for precipitation data. *Brazilian Journal of Probability and Statistics*, **30**(4), 653–675.

[3] L. Sánchez, S. Infante, J. Marcano, and V. Griffin (2015). Polinomial Chaos based on the parallelized ensamble Kalman filter to estimate precipitation states. *Statistics, Optimization and Information Computing,* **3**, 79–95.

Linked micromap plots: Design principles, past uses, and new perspectives via the "rmapshaper" R package

J. Symanzik

Department of Mathematics and Statistics, Utah State University, USA

Linked micromap (LM) plots have been in use since their creation in the mid-1990s. Initially, the underlying code was complex and the shapefiles used to represent the spatial boundaries were not easily obtained or efficient to use. Using modern software, the process of modifying and simplifying shapefiles has become more accessible, facilitating the ability to more easily create and analyze linked micromap plots – and doing so on a larger scale. In this talk, we will revisit the design principles and the past uses of LM plots. The rmapshaper R Package makes it relatively easy to modify boundary files for use in LM plots, even for difficult geographic regions such as Ecuador. For this country, the far-away location of the Galapagos Islands makes it necessary to shift these islands closer to the mainland for a meaningful visualization via LM plots. (Joint work with Braden Probst)

A general dynamic factor approach to forecast conditional covariance matrices in high-dimensional data

Carlos Trucíos[a], Joao H. G. Mazzeu[b], Mauricio Zevallos[b], Luiz K. Hotta[b], Pedro L. Valls Pereira[a], and Marc Hallin[c]

[a] *São Paulo School of Economics, FGV, Brazil,* [b] *Department of Statistics, University of Campinas, Brazil,* [c] *ECARES, Université Libre de Bruxelles, Belgium*

In this paper, we use the General Dynamic Factor Model with infinite-dimensional factor space ([1], [2]) to develop new estimation and forecasting procedures for conditional covariance matrices in high-dimensional data. Most of the approaches available in the literature to model and forecast conditional covariance in high-dimensional time series use static dimension reduction techniques, although, static approaches are not optimal in a time series context since do not exploits the dependence structure of the data [3]. To overcome this issue, we use a dynamic factor approach bases in one-sided filters, which are appropriate for forecasting purposes. The performance of our approach is evaluated via Monte Carlo experiments and yield excellent finite-sample properties. The new procedure is used to construct minimum variance portfolios in a high-dimensional real dataset. The results are shown to achieve better out-of-sample portfolio performance than alternative existing procedures.

Keywords: Dimension reduction, Minimum variance portfolio, Multivariate GARCH.

References

[1] M. Forni, M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and Statistics*, **82**, (4), 540–554.

[2] M. Forni, M. Hallin, M. Lippi, and P. Zaffaroni (2017). Dynamic factor models with infinite-dimensional factor space: asymptotic analysis. *Journal of Econometrics*, **199**, (1), 74–92.

[3] M. Hallin, S. Hörmann, and M. Lippi (2018). Optimal dimension reduction for high-dimensional and functional time series. *Statistical Inference for Stochastic Processes*, **21**, (2), 385–398.

Robust estimation for functional and partially functional linear models

I. Kalogridis and S. Van Aelst

KU Leuven

Functional data analysis is a fast evolving branch of modern statistics, yet despite the popularity of the functional linear model in recent years, current estimation procedures such as [3, 1, 5, 2, 4] either suffer from lack of robustness or are computationally burdensome. To address these drawbacks, we propose a flexible family of lower-rank smoothers that combines penalized splines and M-estimation.Our focus is on regression models with a functional predictor and scalar outcome. We show that, with an approprriate condition on the design matrix, these estimators exhibit the same asymptotic properties as the corresponding least-squares estimators, while being considerably more reliable in the presence of outliers. Further, the proposed methods easily generalize to functional models that also include scalar covariates in a linear or nonparametric term, thus providing a flexible framework of estimation. Simulation experiments show that the proposed estimators have a high efficiency and at the same time protect against outliers. Moreover, smooth estimates are produced which compare favorably with existing least squares and robust procedures.

Keywords: penalized splines, M-estimators

References

[1] C. Crambes, A. Kneip, and P. Sarda (2009). Smoothing splines estimators for functional linear regression. *The Annals of Statistics*, **37**(1), 35–72.

[2] R. A. Maronna, and V. J. Yohai (2013). Robust functional linear regression based on splines. *Computational Statistics & Data Analysis*, **65**, 46–55.

[3] P. T. Reiss, and R. T. Ogden (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, **102**(479), 984–996.

[4] H. Shin, and S. Lee (2016). An rkhs approach to robust functional linear regression. *Statistica Sinica*, **26**, 255–272.

[5] M. Yuan, and T. T. Cai (2010). A reproducing kernel hilbert space approach to functional linear regression. *The Annals of Statistics*, **38**(6), 3412–3444.

Classification for georeferenced functional brain signals

E. Vargas[a], M. Bohorquez[a], R. Guevara[a] and L. Sarmiento[b]

[a] *Universidad Nacional de Colombia,* [b] *Universidad Pedagógica Nacional*

We present a method to classify brain signals from the language area using empirical mode decomposition and functional data analysis. The functional data are built from the intrinsic mode functions (IMFs) obtained for a set of 250 frequencies. The data were obtained for vowels with silent speech. The 20 subjects carrying a neuroheadset EEG with 21 georeferenced electrodes, with the specific task of thinking a vowel during an interval of time. When the subject thinks a vowel, the 21 curves are observed, and the classification method decides what vowel has thought the subject. The speech imageries were captured using electroencephalography (EEG). While the light source is on, he (she) must think continuously the respective vowel with silent speech, and once the light source is off, he (she) must stop thinking about the vowel and move to a state of relaxation. We model the spatial auto-covariance of functional data for each vowel and classify based on Mahalanobis distance using the scores obtained from the functional principal empirical analysis. Our method shows a significant improvement in the apparent error rate comparing with the results obtained using classification methods that do not take into account the spatial autocorrelation.

Keywords: Functional data, empirical mode decomposition, brain signals

References

[1] M. Bohorquez, R. Giraldo, and J. Mateu (2017) Multivariate functional random fields: prediction and optimal sampling. *Stochastic Environmental Research and Risk Assessment*, **31**(1), 53–70.

[2] A. C. Rencher and W. F. Christensen (2012). *Multivariate analysis: Methods and applications.* John Wiley & Sons, NY.

[3] P. Galeano, J. Esdras, and R. Lillo (2015). The Mahalanobis distance for functional data with applications to classification. *Technometrics*, **57**(2), 281–291.

Classification and descriptive analysis for multivariate functional data of IMF signals

E. J. Vargas[a], R. D. Guevara[a], M. P. Bohórquez[a] and S. I. Villamizar[b]

[a] *Dept. of Statistics, Universidad Nacional de Colombia, Bogotá,* [b] *Dept. of Electrical and Electronic Engineering, Universidad Nacional de Colombia, Bogotá*

A new nonlinear technique called Empirical Mode Decomposition (EMD) decomposes any non-stationary time series in a sum of Intrinsic Mode Functions (IMF) that represent zero-mean amplitude, and frequency modulated components. The time series analysed here are the signal of each of the five vowels observed at 21 locations in the brain's languaje area. Thus, 21 spatially correlated time series are obtained. We propose a new methodology of classification based on the EMD of all time series observed. We combine techniques of multivariate functional data, functional principal component analysis and multivariate geostatistics to model the spatial auto-correlation of each vowel. The new observation is classified using the Mahalanobis Distance with the spatial autocorrelation structure found for each vowel.

Keywords: Empirical Mode Decomposition, Multivariate functional data, Spatial classification.

References

[1] M. P. Bohórquez, R. Giraldo, and J. Mateu (2016). Optimal sampling for spatial prediction of functional data. *Statistical Methods & Applications*, **25**(1), 39–54.

[2] M. P. Bohórquez, R. Giraldo, and J. Mateu (2017). Multivariate functional random fields: prediction and optimal sampling. *Stochastic environmental research and risk assessment*, **31**(1), 53–70.

[3] W. Dai and M. G. Genton (2019). Directional outlyingness for multivariate functional data. *Computational Statistics & Data Analysis*, **131**, 50–65.

[4] W. Dai and M. G. Genton (2018). Multivariate functional data visualization and outlier detection. *Journal of Computational and Graphical Statistics*, **27**(4), 923–934.

[5] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen , C. C. Tung, and H. H. Liu (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. In *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, **454**(1971), 903–995.

Batch process control and monitoring: a Dual STATIS and Parallel Coordinates (DS-PC) approach

M. Ramos-Barberán[a], M. V. Hinojosa-Ramos[b], J. Ascencio-Moreno[b], F. Vera[a], O. Ruiz-Barzola[b] and M. P. Galindo-Villardón[c]

[a] *Facultad de Ciencias Naturales y Matemáticas, ESPOL, Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador,* [b] *Facultad de Ciencias de la Vida, ESPOL, Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador,* [c] *Departamento de Estadística, Campus Miguel de Unamuno, Universidad de Salamanca, Salamanca, España*

Multivariate data collected from batches is usually monitored via control charts (CCs) based on MPCA and MPLS for batch to batch comparison. In addition, distribution free approaches include other dimensionality reduction methods for batch and time-wise analysis. However, techniques for multivariate data focused on variable-wise analysis haven't been widely developed. Here, we propose a nonparametric quality control strategy for off-line monitoring of batches and variables, besides visual clustering of observations within batches. In our approach, CCs based on Dual STATIS are created using robust bagplots to enhance signal detection in batch and variable-wise analysis, while parallel coordinate plots are used in identification of unusual observations' behavior per variable, regardless distributional assumptions. This proposed strategy poses the main advantage of detecting different type of changes through meaningful visualization tools, allowing easier interpretation of results in industrial settings. A singular value decomposition algorithm was adapted to the specifications of the Dual STATIS model.

Keywords: Batch process, Dual STATIS, Parallel Coordinates.

Efficient Bias Reduced Simulation-Based Estimators in High Dimensions

Maria-Pia Victoria-Feser[a], Stéphane Guerrier[a], Mucyo Karemera[b] and Samuel Orso[a]

[a] Research Center for Statistics, Geneva School of Economics and Management, University of Geneva, Switzerland, [b] Department of Statistics, Eberly College of Science, Pennsylvania State University, State College, PA, USA.

With the availability of large and complex data settings, an important challenge lies in the control of (finite-sample) bias of estimators associated to the models assumed for the data generating mechanism. The sources of bias are diverse, one being the slow convergence rate for the consistency property (fixed large p or increasing p), as is the case with the maximum likelihood estimator (MLE) outside Gaussian models, another being the presence of spurious outlying observations (outliers) that have a large influence on the value of the estimators. In both situations, separatelly or jointly, the resulting analysis can suffer from important inferential losses. We propose a general framework from which estimators can be derived in a computationally efficient manner for a wide class of models. It requires an initial estimator, possibly baised and non consistent, and the final estimator is obtained by using the Iterative Bootstrap [1, 2] which converges exponentially fast. This simulation-based framework is particularly well suited for bounded influence function estimators, large discrete and/or non Gaussian models with a large number of parameters, as well as with missing data. For example, a simple weighted MLE (non consistent) can be used as initial estimator, avoiding therefore, the approximation of multiple intergrals, a standard feature of consistent weighted MLE. Within this framework we obtain (finite sample and/or asymptotic) unbiased, consistent and (asymptotically) normal estimators when the number of parameters is allowed to increase with the sample size. In a simulation study covering several data and model settings, we find evidence of the advantages of our simulation based estimator in terms of finite sample mean squared error of estimation over other available estimators.

Keywords: Iterative bootstrap, Two-step estimators, Indirect inference.

References

[1] S. Guerrier, E. Dupuis-Lozeron, Y. Ma, and M.-P. Victoria-Feser (2018). Simulation-based bias correction methods for complex models. *Journal of the American Statistical Association*, **114**, 146–157.

[2] S. Guerrier, M. Karemera, S. Orso, and M.P. Victoria-Feser (2018). On the properties of simulation-based estimators in high dimensions. Research Center for Statistics, University of Geneva. Working paper available at: https://arXiv:1810.04443v1.

ICORS.CPS2: May 28, 15:30 - 16:30, Auditorium 1

Robust Variational Inference via Divergences

A. N. Vidyashankar and L. Li

George Mason University

Latent variable models arise in several areas of scientic investigations and variational methods are being increasingly used to reduce computational complexity. In such methods, it is often the case that the properties of the variational estimator depend critically on the variational distribution of the latent variables. In this presentation, we focus on the problem of model misspecification and consider variational inference using general divergences for generalized linear mixed models (GLMM). Specifically, we establish a variational approximation to the divergence minimzation problem and provide a useful correspondence with the stochastic optimization problem. We use this correpondence to study the asymptotic properties of the resulting minimum variational divergence estimators (MVDE) and establish that, within a class of distributions for the latent variables, the MVDE are "robust" to model misspecifications and asymptotically variationally efficient. We propose a divergence based bootstrap approach to estimate the resulting covariance matrix. Finally, we investigate the role of breakdown points as a measure of robustness in this context. We illustrate our methods using numerical experiments and data analysis.

Keywords: Stochastic Optimization, Divergence based bootstrap, Variationally efficient

Spatfd: An R package for functional kriging, functional cokriging and optimal spatial sampling of functional data

A. Villamil[a], M. Bohorquez[a], R. Giraldo[a] and J. Mateu[b]

[a] *National University of Colombia,* [b] *Jaume I University*

This package is based on the methodologies propose in Multivariate functional random fields: prediction and optimal sampling [2] and Optimal sampling for spatial prediction of functional data [1], that extend the framework of multivariate spatial prediction and optimal sampling for functional data. Our proposes presents univariate and multivariate predictors built from the representation of functional data using the Karhunen- loéve expansion. So, the functional auto-covariances and cross-covariances required for predictions and optimal sampling, are completely determined by the sum of the spatial auto-covariances and cross-covariances of the respective score components. This package has the following functions:

- SpatFD: Create an object of class SpatFD
- FKSK: Functional kriging using scalar simple kriging of the scores
- FKCK: Functional kriging using scalar simple cokriging of the scores
- FCOK: Cokriging with p functional random fields
- OSFKSK: Optimal spatial sampling for functional kriging using scalar simple kriging of the scores
- OSFKCK: Optimal spatial sampling for Functional kriging using scalar simple cokriging of the scores
- OSFCOK: Optimal spatial sampling for functional cokriging

Keywords: Functional kriging, Functional Cokriging, Optimal Spatial Sampling

References

[1] M. Bohorquez, R. Giraldo, and J. Mateu (2016). Optimal sampling for spatial prediction of functional data. *Statistical Methods & Applications*, **25**(1), 39–54.

[2] M. Bohorquez, R. Giraldo, and J. Mateu (2017). Multivariate functional random fields: prediction and optimal sampling. *Stochastic Environmental Research and Risk Assessment*, **31**(1), 53–70.

A test for variance equality

J. A. Villasenor and E. Gonzalez-Estrada

Programa de Estadistica, Colegio de Postgraduados, México

The best known procedure for testing variance equality under the assumption of normality is the F-test; however, it is also well-known that this test is not robust against preserving the nominal test size when the samples come from non-normal distributions. Therefore, the statistical analyses based on this test can produce unreliable inferences. The existence of a large number of tests for this problem provides evidence about the relevance of the problem. In this work we propose an asymptotic nonparametric test for the two-sample variance equality problem based on the sample covariance of $U = X + Y$ and $W = X - Y$, denoted by T_n, where $X \sim F_1$ and $Y \sim F_2$, due to the fact that $cov(U, W) = \sigma_1^2 - \sigma_2^2 = 0$ when the null hypothesis holds. It is shown here that under the null hypothesis, T_n has an asymptotic standard normal distribution which is used to obtain critical values for samples of size $n \geq 100$. Besides, whenever F_1 and F_2 are known, the null distribution of T_n, which turns out to be location-scale invariant, from which critical values are obtained, can be approximated by Monte Carlo simulation for sample sizes less than 100. A Monte Carlo simulation study provides evidence that this test is robust with respect to preserving the nominal test size for a large class of parental distributions and it is in general more powerful than existing tests for the same problem.

Keywords: ANOVA, F-test, homoscedasticity.

Quantile-regression-based clustering for panel data

Yingying Zhang[a], Huixia Judy Wang[b] and Zhongyi Zhu[a]

[a]*Fudan University,* [b]*The George Washington University*

In many applications such as economic and medical studies, it is important to identify subgroups of subjects who associate with covariates in different ways. In this talk I will introduce a new quantile-regression-based clustering method for panel data. We develop an iterative algorithm using a similar idea of k-means clustering to identify subgroups at a single quantile level or at multiple quantiles jointly. Even in cases where the group membership is the same across quantile levels, the signal differentiating subgroups may vary with quantiles. It remains unclear which quantile is preferable or should we combine information across multiple quantiles. To answer this question, we propose a new stability measure to choose among multiple quantiles that gives the most stable clustering results. The consistency of the proposed parameter and group membership estimation is established. The finite sample performance of the proposed method is assessed through simulation and the analysis of an economy growth data.

Keywords: Heterogeneity, panel data, subgroup identification.

An alternative approach for testing hyphotesis

L. Lakshnaman[a], E. Smucler[b], V. Yohai[c] and R. H. Zamar[a]

[a] *UBC,* [b] *U.T. Di Tella,* [c] *UBA*

Hypothesis testing is an important tool in academic research and applications. The approach of using $p-$values to test hypothesis became an standard procedure in science and industry. Statistical significant (a small $p-$value) is often a requirement for a scientific findings to be deemed worthy of publication. In recent times, however, $p-$values attracted considerable questioning and criticism. For example, [1] address the following issues regarding the use of p-values:

1. A $p-$value, or statistical significance, does not measure the size of an effect or the importance of a result.
2. By itself, a $p-$value does not provide a good measure of evidence regarding a model or hypothesis.

They then conclude that scientific findings and business or policy decisions should not be based only on whether a $p-$value passes a specific threshold and that proper inference requires full reporting and transparency.

We also add the following criticism:

3. Classical hypothesis testing often relies on questionable assumptions about the data generating process such as parametric probability models, homoscedasticity and symmetry.

We introduce a new approach for comparing hypothesis that addresses criticisms 1, 2 and 3 above. We present our ideas in the context of comparison of two variables. Comparison of pair of variables is an important building block in statistical inference. To fix ideas we assume that larger outcomes are preferable, and that the random variables are positive.

Keywords: $p-$value, nonparametric, relevance

References

[1] R. L. Wasserstein and N. A. Lazar (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, **70**(2), 129–133.

Probabilistic Forecasting of Binary Outcomes in Presence of Outliers

M. Zhelonkin

Erasmus University Rotterdam

The problem of forecasting of binary outcomes is of prominent importance in various fields including Economics, Management, Finance, Marketing, to mention a few. For instance, it can be a default of a company, a click on the online advertisement, or a churn of a client. The traditional approach is to use classification methods, which can be seen as point forecasts. However, from the perspective of a decision maker, it is valuable to have a probability forecast. The utilities or losses can be asymmetric and the price to pay for a mistake can be different for false positive or false negative, for instance it is safer not to give a good loan rather than to give a bad one. The reliably forecasted probabilities can help in optimization of decisions. The typical benchmark model for this problem is logistic regression. However, it is well known that it is very sensitive even to minor model misspecification and even a single outlier can break the maximum likelihood estimator down. In practice (Kuhn & Johnson 2013), the next solution is to switch to machine learning techniques (e.g. Random Forests or Support Vector Machines). However, their probabilistic forecasting performance is not well-studied. In this work we study the behaviour of several techniques in presence of contamination and the influence of outliers on traditional forecast evoluation metrics.

Keywords: Binary outcomes, Calibration, Probabilistic forecasting.

References

[1] M. Kuhn and K. Johnson (2013). *Applied Predictive Modeling.* Springer, New York.

ORGANIZER

ESPOL

@ICORSLACSC2019 WWW.ICORS-LACSC-2019.COM