

Proceedings of COMPSTAT'2010

Yves Lechevallier · Gilbert Saporta
Editors

Proceedings of COMPSTAT'2010

19th International Conference on
Computational Statistics
Paris - France, August 22–27, 2010
Keynote, Invited and Contributed Papers



Physica-Verlag

Editors

Dr. Yves Lechevallier
INRIA Paris-Rocquencourt
Domaine de Voluceau
78153 Le Chesnay cedex
France
yves.lechevallier@inria.fr

Prof. Dr. Gilbert Saporta
CNAM
Chaire de Statistique Appliquée
292 rue Saint Martin
75141 Paris cedex 03
France
gilbert.saporta@cnam.fr

Additional material to this book can be downloaded from <http://extras.springer.com>

ISBN 978-3-7908-2603-6 e-ISBN 978-3-7908-2604-3
DOI 10.1007/978-3-7908-2604-3
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010934004

© Springer-Verlag Berlin Heidelberg 2010

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover design: WMXDesign GmbH, Heidelberg

Printed on acid-free paper

Physica-Verlag is a brand of Springer-Verlag Berlin Heidelberg
Springer-Verlag is part of Springer Science+Business Media (www.springer.com)

Preface

The 19th Conference of IASC-ERS, COMPSTAT'2010, is held in Paris, France, from August 22nd to August 27th 2010, locally organised by the Conservatoire National des Arts et Métiers (CNAM) and the French National Institute for Research in Computer Science and Control (INRIA).

COMPSTAT is an initiative of the European Regional Section of the International Association for Statistical Computing (IASC-ERS), a section of the International Statistical Institute (ISI). COMPSTAT conferences started in 1974 in Wien; previous editions of COMPSTAT were held in Berlin (2002), Prague (2004), Rome (2006) and Porto (2008). It is one of the most prestigious world conferences in Computational Statistics, regularly attracting hundreds of researchers and practitioners, and has gained a reputation as an ideal forum for presenting top quality theoretical and applied work, promoting interdisciplinary research and establishing contacts amongst researchers with common interests.

Keynote lectures are addressed by Luc Devroye (School of Computer Science, McGill University, Montreal), Lutz Edler (Division of Biostatistics, German Cancer Research Center, Heidelberg) and David Hand (Statistics section, Imperial College, London). The conference program includes three tutorials: "Statistical Approach for Complex data" by Lynne Billard (University of Georgia, United States), "Bayesian discrimination between embedded models" by Jean-Michel Marin (Université Montpellier II, France) and "Machine Learning and Association Rules" by Petr Berka and Jan Rauch (University of Economics, Prague, Czech Republic). Each COMPSTAT meeting is organised with a number of topics highlighted, which lead to Invited Sessions. The Conference program includes also contributed sessions and short communications (both oral communications and posters).

The Conference Scientific Program Committee chaired by Gilbert Saporta, CNAM, includes:

Ana Maria Aguilera, Universidad Granada
Avner Bar-Hen, Université René Descartes, Paris
Maria Paula Brito, University of Porto
Christophe Croux, Katholieke Universiteit Leuven
Michel Denuit, Université Catholique de Louvain
Gejza Dohnal, Technical University, Prag
Patrick J. F. Groenen, Erasmus University, Rotterdam
Georges Hébrail, TELECOM ParisTech
Henk Kiers, University of Groningen

Erricos Kontoghiorghes, University of Cyprus
Martina Mittlböck, Medical University of Vienna
Christian P. Robert, Université Paris-Dauphine
Maurizio Vichi, Università La Sapienza, Roma
Peter Winker, Universität Giessen
Moon Yul Huh, SungKyunKwan University, Seoul, Korea
Djamel Zighed, Université Lumière, Lyon

Due to space limitations, the Book of Proceedings includes keynote speakers' papers, invited sessions speakers' papers and a selection of the best contributed papers, while the e-book includes all accepted papers.

The papers included in this volume present new developments in topics of major interest for statistical computing, constituting a fine collection of methodological and application-oriented papers that characterize the current research in novel, developing areas. Combining new methodological advances with a wide variety of real applications, this volume is certainly of great value for researchers and practitioners of computational statistics alike.

First of all, the organisers of the Conference and the editors would like to thank all authors, both of invited and contributed papers and tutorial texts, for their cooperation and enthusiasm. We are specially grateful to all colleagues who served as reviewers, and whose work was crucial to the scientific quality of these proceedings. A special thanks to Hervé Abdi who took in charge the session on Brain Imaging. We also thank all those who have contributed to the design and production of this Book of Proceedings, Springer Verlag, in particular Dr. Martina Bihn and Dr. Niels Peter Thomas, for their help concerning all aspects of publication.

The organisers would like to express their gratitude to all people from CNAM and INRIA who contributed to the success of COMPSTAT'2010, and worked actively for its organisation. We are very grateful to all our sponsors, for their generous support. Finally, we thank all authors and participants, without whom the conference would not have been possible.

The organisers of COMPSTAT'2010 wish the best success to Erricos Kontoghiorghes, Chairman of the 20th edition of COMPSTAT, which will be held in Cyprus in Summer 2012. See you there!

Paris, August 2010

Yves Lechevallier
Gilbert Saporta

Stéphanie Aubin
Gérard Biau
Stéphanie Chaix
Marc Christine
Laurence de Crémiers
Séverine Demeyer
Thierry Despeyroux
Christian Derquenne
Vincenzo Esposito Vinzi
Ali Gannoun
Jean-Pierre Gauchi
Chantal Girodon
Pierre-Louis Gonzalez
Luan Jaupi
Ludovic Lebart
Ndeye Niang
Françoise Potier
Giorgio Russolillo
Julie Séguéla

Acknowledgements

The Editors are extremely grateful to the reviewers, whose work was determinant for the scientific quality of these proceeding. They were, in alphabetical order:

Hervé Abdi	Ali Gannoun
Ana Maria Aguilera	Bernard Garel
Massimo Aria	Cristian Gatu
Josef Arlt	Jean-Pierre Gauchi
Avner Bar-Hen	Pierre-Louis Gonzalez
Jean-Patrick Baudry	Gérard Govaert
Younès Bennani	Patrick Groenen
Petr Berka	Nistor Grozavu
Patrice Bertrand	Fabrice Guillet
Pierre Bertrand	Frederic Guilloux
Gerard Biau	Anne Gégout-Petit
Christophe Biernacki	Hakim Hacid
Lynne Billard	Peter Hall
Hans-Hermann Bock	André Hardy
Frank Bretz	Georges Hébrail
Henri Briand	Harald Heinzl
Maria Paula Brito	Marc Hoffman
Edgar Brunner	Moon Yul Huh
Stephane Canu	Alfonso Iodice d'Enza
Gilles Celeux	Antonio Irpino
Andrea Cerioli	Junling Ji
Roy Cerqueti	François-Xavier Jollois
Ka Chun Cheung	Henk A.L. Kiers
Marc Christine	Dong Kim
Guillaume Cleuziou	Christine Kiss
Claudio Conversano	Erricos Kontoghiorghes
Christophe Croux	Labiod Lazhar
Francisco de Assis De Carvalho	Ludovic Lebart
Michel Denuit	Mustapha Lebbah
Christian Derquenne	Yves Lechevallier
Thierry Despeyroux	Seung Lee
Gejza Dohnal	Guodong Li
Antonio D'Ambrosio	Olivier Lopez
Manuel Escabias	Maria Laura Maag
Vincenzo Esposito Vinzi	Jean-Michel Marin
Christian Francq	Claudia Marinica
Giuliano Galimberti	Roland Marion-Gallois
	Geoffrey McLachlan
	Bertrand Michel

Martina Mittlboeck
Angela Montanari
Irina Moustaki
Shu Ng
Ndeye Niang
Monique Noirhomme
Francisco A. Ocaña
Matej Oresic
Chongsun Park
Francesco Palumbo
Fabien Picarougne
Jean-Michel Poggi
Tommaso Proietti
Pierre Pudlo
Jan Rauch
Marco Riani
Christian Robert
Nicoleta Rogovschi
Rosaria Romano
Fabrice Rossi
Anthony Rossini
Judith Rousseau
Laurent Rouviere
Giorgio Russolillo
Lorenza Saitta
Ryan Skraba

Gilbert Saporta
Seisho Sato
Roberta Siciliano
Francoise Soulie Fogelman
Matthias Studer
Laura Trinchera
Brigitte Trousse
Mariano J. Valderrama
Stefan Van Aelst
Gilles Venturini
Rosanna Verde
Maurizio Vichi
Emmanuel Viennet
Cinzia Viroli
Michal Vrabec
François Wahl
William Wieczorek
Peter Winker
Jingyun Yang
In-Kwon Yeo
Kam Yuen
Daniela Zaharie
Djamel A. Zighed
Lihong Zhang
Xinyuan Zhao

Sponsors

We are extremely grateful to the following institutions whose support contributes to the success of COMPSTAT'2010:

- Conseil Régional Ile de France
- Mairie de Paris
- Société Française de Statistique
- Association EGC (Extraction et Gestion des Connaissances)
- Société Francophone de Classification
- Electricité de France
- Institut National de la Recherche Agronomique
- Institut National de la Statistique et des Etudes Economiques
- IPSOS
- Orange Labs
- SAS-Institute

Contents

Part I. Keynote

Complexity Questions in Non-Uniform Random Variate Generation	3
<i>Luc Devroye</i>	
Computational Statistics Solutions for Molecular Biomedical Research: A Challenge and Chance for Both	19
<i>Lutz Edler, Christina Wunder, Wiebke Werft, Axel Benner</i>	
The Laws of Coincidence	33
<i>David J. Hand</i>	

Part II. ABC Methods for Genetic Data

Choosing the Summary Statistics and the Acceptance Rate in Approximate Bayesian Computation	47
<i>Michael G.B. Blum</i>	
Integrating Approximate Bayesian Computation with Complex Agent-Based Models for Cancer Research	57
<i>Andrea Sottoriva, Simon Tavaré</i>	

Part III. Algorithms for Robust Statistics

Robust Model Selection with LARS Based on S-estimators ...	69
<i>Claudio Agostinelli, Matias Salibian-Barrera</i>	
Robust Methods for Compositional Data	79
<i>Peter Filzmoser, Karel Hron</i>	
Detecting Multivariate Outliers Using Projection Pursuit with Particle Swarm Optimization	89
<i>Anne Ruiz-Gazen, Souad Larabi Marie-Sainte, Alain Berro</i>	

Part IV. Brain Imaging

Imaging Genetics: Bio-Informatics and Bio-Statistics Challenges	101
<i>Jean-Baptiste Poline, Christophe Lalanne, Arthur Tenenhaus, Edouard Duchesnay, Bertrand Thirion, Vincent Frouin</i>	

The NPAIRS Computational Statistics Framework for Data Analysis in Neuroimaging 111
Stephen Strother, Anita Oder, Robyn Spring, Cheryl Grady

Part V. Computational Econometrics

Bootstrap Prediction in Unobserved Component Models 123
Alejandro F. Rodríguez, Esther Ruiz

Part VI. Computer-Intensive Actuarial Methods

A Numerical Approach to Ruin Models with Excess of Loss Reinsurance and Reinstatements 135
Hansjörg Albrecher, Sandra Haas

Computation of the Aggregate Claim Amount Distribution Using R and Actuar 145
Vincent Goulet

Applications of Multilevel Structured Additive Regression Models to Insurance Data 155
Stefan Lang, Nikolaus Umlauf

Part VII. Data Stream Mining

Temporally-Adaptive Linear Classification for Handling Population Drift in Credit Scoring 167
Niall M. Adams, Dimitris K. Tasoulis, Christoforos Anagnostopoulos, David J. Hand

Large-Scale Machine Learning with Stochastic Gradient Descent 177
Léon Bottou

Part VIII. Functional Data Analysis

Anticipated and Adaptive Prediction in Functional Discriminant Analysis 189
Cristian Preda, Gilbert Saporta, Mohamed Hadj Mbarek

Bootstrap Calibration in Functional Linear Regression Models with Applications 199
Wenceslao González-Manteiga, Adela Martínez-Calvo

Empirical Dynamics and Functional Data Analysis 209
Hans-Georg Müller

Part IX. Kernel Methods

Indefinite Kernel Discriminant Analysis 221
Bernard Haasdonk, Elżbieta Pełkalska

Data Dependent Priors in PAC-Bayes Bounds 231
John Shawe-Taylor, Emilio Parrado-Hernández, Amiran Ambroladze

Part X. Monte Carlo Methods in System Safety, Reliability and Risk Analysis

Some Algorithms to Fit some Reliability Mixture Models under Censoring 243
Laurent Bordes, Didier Chauveau

Computational and Monte-Carlo Aspects of Systems for Monitoring Reliability Data 253
Emmanuel Yashchin

Part XI. Optimization Heuristics in Statistical Modelling

Evolutionary Computation for Modelling and Optimization in Finance 265
Sandra Paterlini

Part XII. Spatial Statistics / Spatial Epidemiology

Examining the Association between Deprivation Profiles and Air Pollution in Greater London using Bayesian Dirichlet Process Mixture Models 277
John Molitor, Léa Fortunato, Nuoo-Ting Molitor, Sylvia Richardson

Assessing the Association between Environmental Exposures and Human Health 285
Linda J. Young, Carol A. Gotway, Kenneth K. Lopiano, Greg Kearney, Chris DuClos

Part XIII. ARS Session (Financial) Time Series

Semiparametric Seasonal Cointegrating Rank Selection 297
Byeongchan Seong, Sung K. Ahn, Sinsup Cho

**Estimating Factor Models for Multivariate Volatilities: An In-
novation Expansion Method** 305
Jiazhu Pan, Wolfgang Polonik, Qiwei Yao

Multivariate Stochastic Volatility Model with Cross Leverage. 315
Tsunehiro Ishihara, Yasuhiro Omori

Part XIV. KDD Session: Topological Learning

**Bag of Pursuits and Neural Gas for Improved Sparse
Coding** 327
Kai Labusch, Erhardt Barth, Thomas Martinetz

**On the Role and Impact of the Metaparameters in t-distributed
Stochastic Neighbor Embedding** 337
John A. Lee, Michel Verleysen

**Part XV. IFCS Session: New Developments in Two or Highermode
Clustering; Model Based Clustering and Reduction for High
Dimensional Data**

**Multiple Nested Reductions of Single Data Modes as a Tool
to Deal with Large Data Sets** 349
Iven Van Mechelen, Katrijn Van Deun

The Generic Subspace Clustering Model 359
Marieke E. Timmerman, Eva Ceulemans

Clustering Discrete Choice Data 369
Donatella Vicari, Marco Alfò

Part XVI. Selected Contributed Papers

**Application of Local Influence Diagnostics to the Buckley-
James Model** 381
Nazrina Aziz, Dong Qian Wang

Multiblock Method for Categorical Variables 389
Stéphanie Bougeard, El Mostafa Qannari, Claire Chauvin

A Flexible IRT Model for Health Questionnaire: an Application to HRQoL 397
Serena Broccoli, Giulia Cavrini

Multidimensional Exploratory Analysis of a Structural Model Using a Class of Generalized Covariance Criteria 405
Xavier Bry, Thomas Verron, Patrick Redont

Semiparametric Models with Functional Responses in a Model Assisted Survey Sampling Setting : Model Assisted Estimation of Electricity Consumption Curves 413
Hervé Cardot, Alain Dessertaine, Etienne Josserand

Stochastic Approximation for Multivariate and Functional Median 421
Hervé Cardot, Peggy Cénac, Mohamed Chaouch

A Markov Switching Re-evaluation of Event-Study Methodology 429
Rosella Castellano, Luisa Scaccia

Evaluation of DNA Mixtures Accounting for Sampling Variability 437
Yuk-Ka Chung, Yue-Qing Hu, De-Gang Zhu, Wing K. Fung

Monotone Graphical Multivariate Markov Chains 445
Roberto Colombi, Sabrina Giordano

Using Functional Data to Simulate a Stochastic Process via a Random Multiplicative Cascade Model 453
G. Damiana Costanzo, S. De Bartolo, F. Dell'Accio, G. Trombetta

A Clusterwise Center and Range Regression Model for Interval-Valued Data 461
Francisco de A. T. de Carvalho, Gilbert Saporta, Danilo N. Queiroz

Contributions to Bayesian Structural Equation Modeling 469
Séverine Demeyer, Nicolas Fischer, Gilbert Saporta

Some Examples of Statistical Computing in France During the 19th Century 477
Antoine de Falguerolles

Imputation by Gaussian Copula Model with an Application to Incomplete Customer Satisfaction Data	485
<i>Meelis Käärik, Ene Käärik</i>	
On Multiple-Case Diagnostics in Linear Subspace Method	493
<i>Kuniyoshi Hayashi, Hiroyuki Minami and Masahiro Mizuta</i>	
Fourier Methods for Sequential Change Point Analysis in Autoregressive Models	501
<i>Marie Hušková, Claudia Kirch, Simos G. Meintanis</i>	
Computational Treatment of the Error Distribution in Nonparametric Regression with Right-Censored and Selection-Biased Data	509
<i>Géraldine Laurent, Cédric Heuchenne</i>	
Mixtures of Weighted Distance-Based Models for Ranking Data	517
<i>Paul H. Lee, Philip L. H. Yu</i>	
Fourier Analysis and Swarm Intelligence for Stochastic Optimization of Discrete Functions	525
<i>Jin Rou New, Eldin Wee Chuan Lim</i>	
Global Hypothesis Test to Simultaneously Compare the Predictive Values of Two Binary Diagnostic Tests in Paired Designs: a Simulation Study	533
<i>J. A. Roldán Nofuentes, J. D. Luna del Castillo, M. A. Montero Alonso</i>	
Modeling Operational Risk: Estimation and Effects of Dependencies	541
<i>Stefan Mittnik, Sandra Paterlini, Tina Yener</i>	
Learning Hierarchical Bayesian Networks for Genome-Wide Association Studies	549
<i>Raphaël Mourad, Christine Sinoquet, Philippe Leray</i>	
Posterior Distribution over the Segmentation Space	557
<i>G. Rigaiil, E. Lebarbier, S. Robin</i>	
Parcellation Schemes and Statistical Tests to Detect Active Regions on the Cortical Surface	565
<i>Bertrand Thirion, Alan Tucholka, Jean-Baptiste Poline</i>	
Robust Principal Component Analysis Based on Pairwise Correlation Estimators	573
<i>Stefan Van Aelst, Ellen Vandervieren, Gert Willems</i>	

Ordinary Least Squares for Histogram Data Based on Wasserstein Distance 581
Rosanna Verde, Antonio Irpino

DetMCD in a Calibration Framework..... 589
Tim Verdonck, Mia Hubert, Peter J. Rousseeuw

Separable Two-Dimensional Linear Discriminant Analysis 597
Jianhua Zhao, Philip L.H. Yu, Shulan Li

List of Supplementary Contributed and Invited Papers Only Available on springerlink.com..... 605

Index 617

Part XVII. Supplementary Contributed Papers

Clustering of Waveforms-Data Based on FPCA Direction 625
Giada Adelfio, Marcello Chiodi, Antonino D’Alessandro, Dario Luzio

Symbolic Data Analysis of Complex Data: Application to nuclear power plant 633
Filipe Afonso, Edwin Diday, Norbert Badez, Yves Genest

Different P-spline Approaches for Smoothed Functional Principal Component Analysis 641
Ana M. Aguilera, M. Carmen Aguilera-Morillo, Manuel Escabias, Mariano J. Valderrama

Peak Detection in Mass Spectrometry Data Using Sparse Coding 649
Theodore Alexandrov, Klaus Steinhorst, Oliver Keszöcze, Stefan Schiffler

A Comparison between Beale Test and Some Heuristic Criteria to Establish Clusters Number 657
Angela Alibrandi, Massimiliano Giacalone

Estimating Population Proportions in Presence of Missing Data665
Encarnaciòn Álvarez-Verdejo, Antonio Arcos, Silvia González, Juan Francisco Muñoz, Maria Rueda

Sub-Quadratic Markov Tree Mixture Models for Probability Density Estimation 673
Sourour Ammar, Philippe Leray, Louis Wehenkel

Data Management in Symbolic Data Analysis	681
<i>Teh Amouh, Monique Noirhomme-Fraiture, Benoit Macq</i>	
Variable Selection for Semi-Functional Partial Linear Regression Models	689
<i>Germán Aneiros, Frédéric Ferraty, Philippe Vieu</i>	
Clustering Functional Data Using Wavelets	697
<i>Anestis Antoniadis, Xavier Brossat, Jairo Cugliari, Jean-Michel Poggi</i>	
Polynomial Methods in Time Series Analysis	705
<i>Félix Aparicio-Pérez</i>	
Cointegrated Lee-Carter Mortality Forecasting Method	713
<i>Josef Arlt, Markéta Arltová, Milan Bašta, Jitka Langhamrová</i>	
Empirical Analysis of the Climatic and Social-Economic Factors influence on the Suicide Development in the Czech Republic	721
<i>Markéta Arltová, Jitka Langhamrová, Jana Langhamrová</i>	
Yield Curve Predictability, Regimes, and Macroeconomic Information: A Data-Driven Approach	729
<i>Francesco Audrino, Kameliya Filipova</i>	
Socioeconomic Factors in Circulatory System Mortality in Europe: A Multilevel Analysis of Twenty Countries	737
<i>Sara Balduzzi, Lucio Balzani, Matteo Di Maso, Chiara Lambertini, Elena Toschi</i>	
Comparing ORF Length in DNA Code Observed in Sixteen Yeast Chromosomes	745
<i>Anna Bartkowiak, Adam Szustalewicz</i>	
Influence of the Calibration Weights on Results Obtained from Czech SILC Data	753
<i>Jitka Bartošová, Vladislav Bína</i>	
Continuous Wavelet Transform and the Annual Cycle in Temperature and the Number of Deaths	761
<i>Milan Bašta, Josef Arlt, Markéta Arltová, Karel Helman</i>	
EM-Like Algorithms for Nonparametric Estimation in Multivariate Mixtures	769
<i>Tatiana Benaglia, Didier Chauveau, David R. Hunter</i>	

On the use of Weighted Regression in Conjoint Analysis 777
Salwa Benammou, Besma Souissi, Gilbert Saporta

Wavelet-PLS Regression: Application to Oil Production Data. 785
Salwa Benammou, Kacem Zied, Hedi Kortas, Dhifaoui Zouhaier

Variable Selection and Parameter Tuning in High-Dimensional Prediction 793
Christoph Bernau, Anne-Laure Boulesteix

A Generative Model for Rank Data Based on Sorting Algorithm801
Christophe Biernacki, Julien Jacques

“Made in Italy” Firms Competitiveness: A Multilevel Longitudinal Model on Export Performance 809
Matilde Bini, Margherita Velucchi

Statistical Inference on Large Contingency Tables: Convergence, Testability, Stability 817
Marianna Bolla

A Class of Multivariate Type I Generalized Logistic Distributions 825
Salvatore Bologna

Adaptive Mixture Discriminant Analysis for Supervised Learning with Unobserved Classes. 831
Charles Bouveyron

Forecasting a Compound Cox Process by means of PCP 839
Paula R. Bouzas, Nuria Ruiz-Fuentes, Juan Eloy Ruiz-Castro

Cutting the Dendrogram through Permutation Tests 847
Dario Bruzzese, Domenico Vistocco

Design of Least-Squares Quadratic Estimators Based on Covariances from Interrupted Observations Transmitted by Different Sensors 855
R. Caballero-Águila, A. Hermoso-Carazo, J. Linares-Pérez

Pseudo-Bayes Factors 863
Stefano Cabras, Walter Racugno, Laura Ventura

Diagnostic Checking of Multivariate Normality Under Contamination. 871
Andrea Cerioli

On Computationally Complex Instances of the c-optimal Experimental Design Problem: Breaking RSA-based Cryptography via c-optimal Designs	879
<i>Michal Černý, Milan Hladík, Veronika Škočdopolová</i>	
Estimation and Detection of Outliers and Patches in Nonlinear Time Series Models	887
<i>Ping Chen</i>	
Two-way Classification of a Table with non-negative entries: Validation of an Approach based on Correspondence Analysis and Information Criteria	895
<i>Antonio Ciampi, Alina Dyachenko, Yves Lechevallier</i>	
A Mann-Whitney Spatial Scan Statistic for Continuous Data .	903
<i>Lionel Cucala</i>	
Quantile Regression for Group Effect Analysis	911
<i>Cristina Davino, Domenico Vistocco</i>	
Regularized Directions of Maximal Outlyingness	919
<i>Michiel Debruyne</i>	
A New Approach to Robust Clustering in \mathbb{R}^p	927
<i>Catherine Dehon, Kaveh Vakili</i>	
An Exploratory Segmentation Method for Time Series	935
<i>Christian Derquenne</i>	
Using Auxiliary Information Under a Generic Sampling Design	943
<i>Giancarlo Diana, Pier Francesco Perri</i>	
Improving Overlapping Clusters obtained by a Pyramidal Clustering	951
<i>Edwin Diday, Francisco de A. T. de Carvalho, Luciano D.S. Pacifico</i>	
Visualizing and Forecasting Complex Time Series: Beanplot Time Series	959
<i>Carlo Drago, Germana Scepi</i>	
M-estimation in INARCH Models with a Special Focus on Small Means	967
<i>Hanan El-Saied, Roland Fried</i>	
Score Moment Estimators	975
<i>Zdeněk Fabián</i>	

Testing the Number of Components in Poisson Mixture Regression Models 983
Susana Faria, Fátima Gonçalves

Support Vector Machines for Large Scale Text Mining in R .. 991
Ingo Feinerer, Alexandros Karatzoglou

Computation of the Projection of the Inhabitants of the Czech Republic by sex, age and the highest education level..... 999
Tomáš Fiala, Jitka Langhamrová

Two Kurtosis Measures in a Simulation Study1007
Anna Maria Fiori

Clustering of Czech Household Incomes Over Very Short Time Period1015
Marie Forbelská, Jitka Bartošová

Model-Based Nonparametric Variance Estimation for Systematic Sampling. An Application in a Forest Survey1023
Mario Francisco-Fernández, Jean Opsomer, Xiaoxi Li

Thresholding-Wavelet-Based Functional Estimation of Spatiotemporal Strong-Dependence in the Spectral Domain1031
María Pilar Frías, María Dolores Ruiz-Medina

Boolean Factor Analysis by the Expectation-Maximization Algorithm.....1039
Alexander. A. Frolov, Pavel. Y. Polyakov, Dusan Húšek

Modeling and Forecasting Electricity Prices and their Volatilities by Conditionally Heteroskedastic Seasonal Dynamic Factor Analysis1047
Carolina García-Martos, Julio Rodríguez, María Jesús Sánchez

Consensus Analysis Through Modal Symbolic Objects1055
Jose M. García-Santesmases, M. Carmen Bravo

Nonlinear Regression Model of Copper Bromide Laser Generation1063
Snezhana Georgieva Gocheva-Ilieva, Ilycho Petkov Iliev

Random Forests Based Feature Selection for Decoding fMRI Data.....1071
Robin Genuer, Vincent Michel, Evelyn Eger, Bertrand Thirion

Differentiation Tests for the Mean Shape and the Mean Variance of Renal Tumours appearing in early Childhood	1079
<i>Stefan Markus Giebel, Jens-Peter Schenk, Jang Schiltz</i>	
Local or Global Smoothing? A Bandwidth Selector for Dependent Data	1087
<i>Francesco Giordano, Maria Lucia Parrella</i>	
Panel Data Models for Productivity Analysis	1095
<i>Luigi Grossi, Giorgio Gozzi</i>	
A Stochastic Gamma Diffusion Model with Threshold Parameter. Computational Statistical Aspects and Application	1103
<i>Ramón Gutiérrez, Ramón Gutiérrez-Sánchez, Ahmed Nafidi, Eva Maria Ramos-Ábalos</i>	
On the Correlated Gamma Frailty Model for Bivariate Current Status Data	1111
<i>Niel Hens, Andreas Wienke</i>	
Evolutionary Stochastic Portfolio Optimization and Probabilistic Constraints	1119
<i>Ronald Hochreiter</i>	
Boosting a Generalised Poisson Hurdle Model	1127
<i>Vera Hofer</i>	
Fast and Robust Classifiers Adjusted for Skewness	1135
<i>Mia Hubert, Stephan Van der Veeken</i>	
Modelling the Andalusian Population by Means of a non-Homogeneous Stochastic Gompertz Process	1143
<i>Maria Dolores Huete Morales, Francisco Abad Montes</i>	
Neural Network Approach for Histopathological Diagnosis of Breast Diseases with Images	1151
<i>Yuichi Ishibashi, Atsuko Hara, Isao Okayasu, Koji Kurihara</i>	
Detection of Spatial Cluster for Suicide Data using Echelon Analysis	1159
<i>Fumio Ishioka, Makoto Tomita, Toshiharu Fujita</i>	
Time-Varying Coefficient Model with Linear Smoothing Function for Longitudinal Data in Clinical Trial	1167
<i>Masanori Ito, Toshihiro Misumi, Hideki Hirooka</i>	

Metropolis-Hastings Algorithm for Mixture Model and its Weak Convergence	1175
<i>Kengo Kamatani</i>	
A Method for Time Series Analysis Using Probability Distribution of Local Standard Fractal Dimension	1183
<i>Kenichi Kamijo, Akiko Yamanouchi</i>	
Assessment of Scoring Models Using Information Value	1191
<i>Jan Koláček, Martin Řezáč</i>	
The Moving Average Control Chart Based on the Sequence of Permutation Tests	1199
<i>Grzegorz Konczak</i>	
Depth Based Procedures for Estimation ARMA and GARCH Models	1207
<i>Daniel Kosiorowski</i>	
Half-Taxi Metric in Compositional Data Geometry Rcomp ..	1215
<i>Katarina Košmelj, Vesna Žabkar</i>	
LTPD Plans by Variables when the Remainder of Rejected Lots is Inspected	1223
<i>J. Klufa, L. Marek</i>	
A Comparison between Two Computing Methods for an Empirical Variogram in Geostatistical Data	1231
<i>Takafumi Kubota, Tomoyuki Tarumi</i>	
Improvement of Acceleration of the ALS Algorithm Using the Vector ε Algorithm	1239
<i>Masahiro Kuroda, Yuchi Mori, Masaya Izuka, Michio Sakakihara</i>	
Unsupervised Recall and Precision Measures: a Step towards New Efficient Clustering Quality Indexes	1247
<i>Jean-Charles Lamirel, Maha Ghribi, Pascal Cuxac</i>	
Performance Assessment of Optimal Allocation for Large Portfolios	1255
<i>Fabrizio Laurini, Luigi Grossi</i>	
Clustering of Multiple Dissimilarity Data Tables for Documents Categorization	1263
<i>Yves Lechevallier, Francisco de A. T. de Carvalho, Thierry Despeyroux, Filipe M. de Melo</i>	

Slimming Down a High-Dimensional Binary Datatable: relevant Eigen-Subspace and Substantial Content	1271
<i>Alain Lelu</i>	
Comparing Two Approaches to Testing Linearity against Markov-switching Type Non-linearity	1279
<i>Jana Lenčuchová, Anna Petričková, Magdaléna Komorníková</i>	
Numerical Error Analysis for Statistical Software on Multi-Core Systems	1287
<i>Wenbin Li, Sven Simon</i>	
Sparse Bayesian Hierarchical Model for Clustering Problems	1295
<i>Heng Lian</i>	
Data Mining and Multiple Correspondence Analysis via Polynomial Transformations	1303
<i>Rosaria Lombardo</i>	
Structural Modelling of Nonlinear Exposure-Response Relationships for Longitudinal Data	1311
<i>Xiaoshu Lu, Esa-Pekka Takala</i>	
Empirical Composite Likelihoods	1319
<i>Nicola Lunardon, Francesco Pauli, Laura Ventura</i>	
A Fast Parsimonious Maximum Likelihood Approach for Predicting Outcome Variables from a Large Number of Predictors	1327
<i>Jay Magidson</i>	
A Bootstrap Method to Improve Brain Subcortical Network Segregation in Resting-State fMRI Data	1335
<i>Caroline Malherbe, Eric Bardinnet, Arnaud Messé, Vincent Perlberg, Guillaume Marrelec, Mélanie Péligrini-Issac, Jérôme Yelnik, Stéphane Lehericy, Habib Benali</i>	
The Problem of Determining the Calibration Equations to Construct Model-calibration Estimators of the Distribution Function	1343
<i>Sergio Martínez, María Rueda, Antonio Arcos, Helena Martínez, Juan Francisco Muñoz</i>	
Dealing with Nonresponse in Survey Sampling: an Item Response Modeling Approach	1353
<i>Alina Matei</i>	

Estimation of the Bivariate Distribution Function for Censored Gap Times	1359
<i>Luís Meira-Machado, Ana Moreira</i>	
Two Measures of Dissimilarity for the Dendrogram Multi-Class SVM Model	1367
<i>Rafael Pino Mejías, María Dolores Cubiles de la Vega</i>	
Visualizing the Sampling Variability of Plots	1375
<i>Rajiv S. Menjoge, Roy E. Welsh</i>	
Empirical Mode Decomposition for Trend Extraction. Application to Electrical Data	1383
<i>Farouk Mhamdi, Mériem Jaïdane-Saïdane, Jean-Michel Poggi</i>	
The Evaluation of Non-centred Orthant Probabilities for Singular Multivariate Normal Distributions	1391
<i>Tetsuhisa Miwa</i>	
Variable Inclusion and Shrinkage Algorithm in High Dimension	1397
<i>Abdallah Mkhadri, Mohamed Ouhourane</i>	
Application of a Bayesian Approach for Analysing Disease Mapping Data: Modelling Spatially Correlated Small Area Counts	1405
<i>Mohammadreza Mohebbi, Rory Wolfe</i>	
Clusters of Gastrointestinal Tract Cancer in the Caspian Region of Iran: A Spatial Scan Analysis	1413
<i>Mohammadreza Mohebbi, Rory Wolfe</i>	
The Financial Crisis of 2008: Modelling the Transmission Mechanism Between the Markets	1421
<i>M. Pilar Muñoz Maria Dolores Márquez, Helena Chuliá</i>	
Determining the Direction of the Path Using a Bayesian Semi-parametric Model	1429
<i>Kei Miyazaki, Takahiro Hoshino, Kazuo Shigemasu</i>	
Data Visualization and Aggregation	1437
<i>Junji Nakano, Yoshikazu Yamamoto</i>	
Longitudinal Data Analysis Based on Ranks and its Performance	1445
<i>Takashi Nagakubo, Masashi Goto</i>	

Multiple Change Point Detection by Sparse Parameter Estimation1453
<i>Jiří Neubauer, Vítězslav Veselý</i>	
Quasi-Maximum Likelihood Estimators for Threshold ARMA Models: Theoretical Results and Computational Issues1461
<i>Marcella Niglio, Cosimo Damiano Vitale</i>	
A Case Study of Bank Branch Performance Using Linear Mixed Models1469
<i>Peggy Ng, Claudia Czado, Eike Christian Brechmann, Jon Kerr</i>	
Numerical Methods for some Classes of Matrices with Applications to Statistics and Optimization1477
<i>Juan M. Peña</i>	
Maximum Margin Learning of Gaussian Mixture Models with Application to Multipitch Tracking1485
<i>Franz Pernkopf, Michael Wohlmayr</i>	
Low-Pass Filter Design using Locally Weighted Polynomial Regression and Discrete Prolate Spheroidal Sequences1493
<i>Tommaso Proietti, Alessandra Luati</i>	
A Statistical Survival Model Based on Counting Processes1501
<i>Jose-Manuel Quesada-Rubio, Julia Garcia-Leal, Maria-Jose Del-Moral-Avila, Esteban Navarrete-Alvarez, Maria-Jesus Rosales-Moreno</i>	
Bootstrapping Additive Models in Presence of Missing Data1509
<i>Rocío Raya-Miranda, M. Dolores Martínez-Miranda, Andrés González-Carmona</i>	
On Aspects of Quality Indexes for Scoring Models1517
<i>Martin Řezáč, Jan Kolářek</i>	
Data Clustering with Mixed Type Variables and Cluster Number Determination1525
<i>Hana Řezanková, Dušan Húsek, Tomáš Löster</i>	
A General Strategy for Determining First-Passage-Time Densities Based on the First-Passage-Time Location Function1533
<i>Patricia Román-Román, Juan José Serrano-Pérez, Francisco Torres-Ruiz</i>	
Rplugin.Econometrics: R-GUI for Teaching Time Series Analysis1541
<i>Dedi Rosadi</i>	

Computational Statistics: the Symbolic Approach	1549
<i>Colin Rose</i>	
EOFs for Gap Filling in Multivariate Air Quality data: a FDA Approach	1557
<i>Mariantonietta Ruggieri, Francesca Di Salvo, Antonella Plaia, Gianna Agró</i>	
A Transient Analysis of a Complex Discrete k-out-of-n:G System with Multi-State Components	1565
<i>Juan Eloy Ruiz-Castro, Paula R. Bouzas</i>	
Using Logitboost for Stationary Signals Classification	1573
<i>Pedro Saavedra, Angelo Santana, Carmen Nieves Hernández, Juan Artiles, Juan-José González</i>	
Test of Mean Difference for Longitudinal Data Using Circular Block Bootstrap	1581
<i>Hirohito Sakurai, Masaaki Taguri</i>	
An Empirical Study of the Use of Nonparametric Regression Methods for Imputation	1589
<i>Ismael R. Sánchez-Borrego, Maria Rueda, Encarnación Álvarez-Verdejo</i>	
A Simulation Study of the Bayes Estimator of Parameters in an Extension of the Exponential Distribution	1597
<i>Samira Sadeghi</i>	
A Cluster-Target Similarity Based Principal Component Analysis for Interval-Valued Data	1605
<i>Mika Sato-Ilic</i>	
Bayesian Flexible Modelling of Mixed Logit Models	1613
<i>Luisa Scaccia, Edoardo Marcucci</i>	
A Decision Tree for Symbolic Data	1621
<i>Djamal Seck, Lynne Billard, Edwin Diday, Filipe Afonso</i>	
The Set of $3 \times 4 \times 4$ Contingency Tables has 3-Neighborhood Property	1629
<i>Toshio Sumi, Toshio Sakata</i>	
Visualization Techniques for the Integration of Rank Data ...	1637
<i>Michael G. Schimek, Eva Budinská</i>	

Comprehensive Assessment on Hierarchical Structures of DNA markers Using Echelon Analysis1645
Makoto Tomita, Koji Kurihara

Non-Hierarchical Clustering for Distribution-Valued Data1653
Yoshikazu Terada, Hiroshi Yadohisa

On Composite Pareto Models.....1661
Sandra Teodorescu, Raluca Vernic

Visualisation of Large Sized Data Sets : Constraints and Improvements for Graph Design.....1669
Jean-Paul Valois

Selecting Variables in Two-Group Robust Linear Discriminant Analysis1677
Stefan Van Aelst, Gert Willems

How to Take into Account the Discrete Parameters in the BIC Criterion?1685
Vincent Vandewalle

Analysis of Breath Alcohol Measurements Using Compartmental and Generalized Linear Models1693
Chi Ting Yang, Wing Kam Fung, Thomas Wai Ming Tam

Fisher Scoring for Some Univariate Discrete Distributions1701
Thomas W. Yee

Constructing Summary Indexes via Principal Curves1709
Mohammad Zayed, Jochen Einbeck

Censored Survival Data: Simulation and Kernel Estimates1717
Jiří Zelinka

Part XVIII. Supplementary Invited Papers

Heuristic Optimization for Model Selection and Estimation ..1727
Dietmar Maringer

General Index1737

Part I

Keynote

Complexity Questions in Non-Uniform Random Variate Generation

Luc Devroye

School of Computer Science
McGill University
Montreal, Canada H3A 2K6
lucdevroye@gmail.com

Abstract. In this short note, we recall the main developments in non-uniform random variate generation, and list some of the challenges ahead.

Keywords: random variate generation, Monte Carlo methods, simulation

1 The pioneers

World War II was a terrible event. But it can not be denied that it pushed science forward with a force never seen before. It was responsible for the quick development of the atomic bomb and led to the cold war, during which the United States and Russia set up many research labs and attracted the best and the brightest to run them. It was at Los Alamos and RAND that physicists and other scientists were involved in large-scale simulations. John von Neumann, Stan Ulam and Nick Metropolis developed the Monte Carlo Method in 1946: they suggested that we could compute and predict in ways never before considered. For example, the Metropolis chain method developed a few years later (Metropolis, Rosenbluth, Rosenbluth, Teller and Teller, 1953) can be used to simulate almost any distribution by setting up a Markov chain that has that distribution as a limit. At least asymptotically, that is. But it was feasible, because the computers were getting to be useful, with the creation of software and the FORTRAN compiler.

To drive the Markov chains and other processes, one would need large collections of uniform random numbers. That was a bit of a sore point, because no one knew where to get them. Still today, the discussion rages as to how one should secure a good source of uniform random numbers. The scientists eventually settled on something that a computer could generate, a sequence that looked random.

The early winner was the linear congruential generator, driven by $x_{n+1} = (ax_n + b) \bmod m$, which had several well-understood properties. Unfortunately, it is just a deterministic sequence, and many of its flaws have been exposed in the last three decades. The built-in linear-congruential generator in the early FORTRAN package for IBM computers was RANDU. Consecutive pairs

(x_n, x_{n+1}) produced by RANDU fall on just a few parallel lines, prompting Marsaglia (1968) to write a paper with the ominous title “Random numbers fall mainly in the plane”. But bad linear congruential or related generators have persisted until today—the generator in Wolfram’s Mathematica had a similar problem: their built-in generator Random uses the Marsaglia-Zaman subtract-with-borrow generator (1991), which has the amazing property that all consecutive triples (x_n, x_{n+1}, x_{n+2}) fall in only two hyperplanes of $[0, 1]^3$, a fact pointed out to me by Pierre Lecuyer. Many thousands of simulations with Mathematica are thus suspect—I was made aware of this due an inconsistency between simulation and theory brought to my attention by Jim Fill in 2010. The company has never apologized or offered a refund to its customers, but it has quietly started using other methods, including one based on a cellular automaton (the default). However, they are still offering linear congruential generators as an option. The story is far from over, and physical methods may well come back in force.

Information theorists and computer scientists have approached randomness from another angle. For them, random variables uniformly distributed on $[0, 1]$ do not and can not exist, because the binary expansions of such variables consist of infinitely many independent Bernoulli $(1/2)$ random bits. Each random bit has binary entropy equal to one, which means that its value or cost is one. A bit can store one unit of information, and vice versa, a random bit costs one unit of resources to produce. Binary entropy for a more complex random object can be measured in terms of how many random bits one needs to describe it. The binary entropy of a random vector of n independent fair coin flips is n , because we can describe it by n individual fair coins.

For the generation of discrete or integer-valued random variables, which includes the vast area of the generation of random combinatorial structures, one can adhere to a clean model, the pure bit model, in which each bit operation takes one time unit, and storage can be reported in terms of bits. In this model, one assumes that an i.i.d. sequence of independent perfect bits is available. This permits the development of an elegant information-theoretic theory. For example, Knuth and Yao (1976) showed that to generate a random integer X described by the probability distribution

$$\mathbf{P}\{X = n\} = p_n, n \geq 1,$$

any method must use an expected number of bits greater than the binary entropy of the distribution,

$$\sum_n p_n \log_2(1/p_n).$$

They also showed how to construct tree-based generators that can be implemented as finite or infinite automata to come within three bits of this lower bound for any distribution. While this theory is elegant and theoretically