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

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, Rottterdam
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, Universita 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 alphatbetical order:

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

Martina Mittlboeck
Angela Montanari
Irini Moustaki
Shu Ng
Ndeye Niang
Monique Noirhomme
Francisco A. Ocaña
Matej Oresic
Chongsun Park
Franceso 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 XVII. Supplementary Contributed Papers

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 $x_{n+1} = (ax_n + b) \mathrm{mod} m$, 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

Y. Lechevallier, G. Saporta (eds.), *Proceedings of COMPSTAT'2010*,
DOI 10.1007/978-3-7908-2604-3_1,

(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). Hoewever, 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