

Proceedings e report

114

SIS 2017
Statistics and Data Science:
new challenges, new generations

28–30 June 2017
Florence (Italy)

Proceedings of the Conference
of the Italian Statistical Society

edited by
Alessandra Petrucci
Rosanna Verde

FIRENZE UNIVERSITY PRESS
2017

SIS 2017. Statistics and Data Science: new challenges, new generations : 28-30 June 2017 Florence (Italy) : proceedings of the Conference of the Italian Statistical Society / edited by Alessandra Petrucci, Rosanna Verde. – Firenze : Firenze University Press, 2017.
(Proceedings e report ; 114)

<http://digital.casalini.it/9788864535210>

ISBN 978-88-6453-521-0 (online)

Peer Review Process

All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Committees of the individual series. The works published in the FUP catalogue are evaluated and approved by the Editorial Board of the publishing house. For a more detailed description of the refereeing process we refer to the official documents published on the website and in the online catalogue of the FUP (www.fupress.com).

Firenze University Press Editorial Board

A. Dolfi (Editor-in-Chief), M. Boddi, A. Bucelli, R. Casalbuoni, M. Garzaniti, M.C. Grisolia, P. Guarnieri, R. Lanfredini, A. Lenzi, P. Lo Nostro, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, G. Nigro, A. Perulli, M.C. Torricelli.

This work is licensed under a Creative Commons Attribution 4.0 International License
(CC BY 4.0: <https://creativecommons.org/licenses/by/4.0/legalcode>)

CC 2017 Firenze University Press
Università degli Studi di Firenze
Firenze University Press
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

SOCIETÀ ITALIANA DI STATISTICA

Sede: Salita de' Crescenzi 26 - 00186 Roma

Tel +39-06-6869845 - Fax +39-06-68806742

email: sis@caspur.it web:<http://www.sis-statistica.it>

La Società Italiana di Statistica (SIS), fondata nel 1939, è una società scientifica eretta ad Ente morale ed inclusa tra gli Enti di particolare rilevanza scientifica. La SIS promuove lo sviluppo delle scienze statistiche e la loro applicazione in campo economico, sociale, sanitario, demografico, produttivo ed in molti altri settori di ricerca.

Organi della società:

Presidente:

- Prof.ssa Monica Pratesi, Università di Pisa

Segretario Generale:

- Prof.ssa Filomena Racioppi, Sapienza Università di Roma

Tesoriere:

- Prof.ssa Maria Felice Arezzo, Sapienza Università di Roma

Consiglieri:

- Prof. Giuseppe Arbia, Università Cattolica del Sacro Cuore

- Prof.ssa Maria Maddalena Barbieri, Università Roma Tre

- Prof.ssa Francesca Bassi, Università di Padova

- Prof. Eugenio Brentari, Università di Brescia

- Dott. Stefano Falorsi, ISTAT

- Prof. Alessio Pollice, Università di Bari

- Prof.ssa Rosanna Verde, Seconda Università di Napoli

- Prof. Daniele Vignoli, Università di Firenze

Collegio dei Revisori dei Conti:

- Prof. Francesco Campobasso, Prof. Michele Gallo, Prof. Francesco Sanna, Prof. Umberto Salinas (supplente)

SIS2017 Committees

Scientific Program Committee:

Rosanna Verde (chair), Università della Campania “Luigi Vanvitelli”
Maria Felice Arezzo, Sapienza Università di Roma
Antonino Mazzeo, Università di Napoli Federico II
Emanuele Baldacci, Eurostat
Pierpaolo Brutti, Sapienza Università di Roma
Marcello Chiodi, Università di Palermo
Corrado Crocetta, Università di Foggia
Giovanni De Luca, Università di Napoli Parthenope
Viviana Egidi, Sapienza Università di Roma
Giulio Ghellini, Università degli Studi di Siena
Ippoliti Luigi, Università di Chieti-Pescara “G. D’Annunzio”
Matteo Mazziotta, ISTAT
Lucia Paci, Università Cattolica del Sacro Cuore
Alessandra Petrucci, Università degli Studi di Firenze
Filomena Racioppi, Sapienza Università di Roma
Laura M. Sangalli, Politecnico di Milano
Bruno Scarpa, Università degli Studi di Padova
Cinzia Viroli, Università di Bologna

Local Organizing Committee:

Alessandra Petrucci (chair), Università degli Studi di Firenze
Gianni Betti, Università degli Studi di Siena
Fabrizio Cipollini, Università degli Studi di Firenze
Emanuela Dreassi, Università degli Studi di Firenze
Caterina Giusti, Università di Pisa
Leonardo Grilli, Università degli Studi di Firenze
Alessandra Mattei, Università degli Studi di Firenze
Elena Pirani, Università degli Studi di Firenze
Emilia Rocco, Università degli Studi di Firenze
Maria Cecilia Verri, Università degli Studi di Firenze

Supported by:

Università degli Studi di Firenze
Università di Pisa
Università degli Studi di Siena
ISTAT
Regione Toscana
Comune di Firenze
BITBANG srl

Index

Preface	XXV
Alexander Agapitov, Irina Lackman, Zoya Maksimenko <i>Determination of basis risk multiplier of a borrower default using survival analysis</i>	1
Tommaso Agasisti, Alex J. Bowers, Mara Soncin <i>School principals' leadership styles and students achievement: empirical results from a three-step Latent Class Analysis</i>	7
Tommaso Agasisti, Sergio Longobardi, Felice Russo <i>Poverty measures to analyse the educational inequality in the OECD Countries</i>	17
Mohamed-Salem Ahmed, Laurence Broze, Sophie Dabo-Niang, Zied Gharbi <i>Quasi-Maximum Likelihood Estimators For Functional Spatial Autoregressive Models</i>	23
Giacomo Aletti, Alessandra Micheletti <i>A clustering algorithm for multivariate big data with correlated components</i>	31
Emanuele Aliverti <i>A Bayesian semiparametric model for terrorist networks</i>	37

- Giorgio Alleva
Emerging challenges in official statistics: new sources, methods and skills 43
- Rémi André, Xavier Luciani and Eric Moreau
A fast algorithm for the canonical polyadic decomposition of large tensors 45
- Maria Simona Andreano, Roberto Benedetti, Paolo Postiglione, Giovanni Savio
On the use of Google Trend data as covariates in nowcasting: Sampling and modeling issues 53
- Francesco Andreoli, Mauro Mussini
A spatial decomposition of the change in urban poverty concentration 59
- Margaret Antonicelli, Vito Flavio Covella
How green advertising can impact on gender different approach towards sustainability 65
- Rosa Arboretti, Eleonora Carrozzo, Luigi Salmaso
Stratified data: a permutation approach for hypotheses testing 71
- Marika Arena, Anna Calissano, Simone Vantini
Crowd and Minorities: Is it possible to listen to both? Monitoring Rare Sentiment and Opinion Categories about Expo Milano 2015 79
- Maria Felice Arezzo, Giuseppina Guagnano
Using administrative data for statistical modeling: an application to tax evasion 83
- Monica Bailot, Rina Camporese, Silvia Da Valle, Sara Letardi, Susi Osti
Are Numbers too Large for Kids? Possible Answers in Probable Stories 89

Index	IX
Simona Balbi, Michelangelo Misuraca, Germana Scepti <i>A polarity-based strategy for ranking social media reviews</i>	95
A. Balzanella, S.A. Gattone, T. Di Battista, E. Romano, R. Verde <i>Monitoring the spatial correlation among functional data streams through Moran's Index</i>	103
Oumayma Banouar, Saïd Raghay <i>User query enrichment for personalized access to data through ontologies using matrix completion method</i>	109
Giulia Barbati, Francesca Ieva, Francesca Gasperoni, Annamaria Iorio, Gianfranco Sinagra, Andrea Di Lenarda <i>The Trieste Observatory of cardiovascular disease: an experience of administrative and clinical data integration at a regional level</i>	115
Francesco Bartolucci, Stefano Peluso, Antonietta Mira <i>Marginal modeling of multilateral relational events</i>	123
Francesca Bassi, Leonardo Grilli, Omar Paccagnella, Carla Rampichini, Roberta Varriale <i>New Insights on Students Evaluation of Teaching in Italy</i>	129
Mauro Bernardi, Marco Bottone, Lea Petrella <i>Bayesian Quantile Regression using the Skew Exponential Power Distribution</i>	135
Mauro Bernardi <i>Bayesian Factor-Augmented Dynamic Quantile Vector Autoregression</i>	141

- Bruno Bertaccini, Giulia Biagi, Antonio Giusti, Laura Grassini
Does data structure reflect monuments structure? Symbolic data analysis on Florence Brunelleschi Dome
149
- Gaia Bertarelli and Franca Crippa, Fulvia Mecatti
A latent markov model approach for measuring national gender inequality
157
- Agne Bikauskaite, Dario Buono
Eurostat's methodological network: Skills mapping for a collaborative statistical office
161
- Francesco C. Billari, Emilio Zagheni
Big Data and Population Processes: A Revolution?
167
- Monica Billio, Roberto Casarin, Matteo Iacopini
Bayesian Tensor Regression models
179
- Monica Billio, Roberto Casarin, Luca Rossini
Bayesian nonparametric sparse Vector Autoregressive models
187
- Chiara Bocci, Daniele Fadda, Lorenzo Gabrielli, Mirco Nanni, Leonardo Piccini
Using GPS Data to Understand Urban Mobility Patterns: An Application to the Florence Metropolitan Area
193
- Michele Boreale, Fabio Corradi
Relative privacy risks and learning from anonymized data
199
- Giacomo Bormetti, Roberto Casarin, Fulvio Corsi, Giulia Livieri
A stochastic volatility framework with analytical filtering
205

Index	XI
Alessandro Brunetti, Stefania Fatello, Federico Polidoro <i>Estimating Italian inflation using scanner data: results and perspectives</i>	211
Guénael Cabanes, Younès Bennani, Rosanna Verde, Antonio Irpino <i>Clustering of histogram data : a topological learning approach</i>	219
Renza Campagni, Lorenzo Gabrielli, Fosca Giannotti, Riccardo Guidotti, Filomena Maggino, Dino Pedreschi <i>Measuring Wellbeing by extracting Social Indicators from Big Data</i>	227
Maria Gabriella Campolo, Antonino Di Pino <i>Assessing Selectivity in the Estimation of the Causal Effects of Retirement on the Labour Division in the Italian Couples</i>	235
Stefania Capecchi, Rosaria Simone <i>Composite indicators for ordinal data: the impact of uncertainty</i>	241
Stefania Capecchi, Domenico Piccolo <i>The distribution of Net Promoter Score in socio-economic surveys</i>	247
Massimiliano Caporin, Francesco Poli <i>News, Volatility and Price Jumps</i>	253
Carmela Cappelli, Rosaria Simone, Francesca di Iorio <i>Growing happiness: a model-based tree</i>	261
Paolo Emilio Cardone <i>Inequalities in access to job-related learning among workers in Italy: evidence from Adult Education Survey (AES)</i>	267

- Alessandro Casa, Giovanna Menardi
Signal detection in high energy physics via a semisupervised nonparametric approach
 273
- Claudio Ceccarelli, Silvia Montagna, Francesca Petrarca
Employment study methodologies of Italian graduates through the data linkage of administrative archives and sample surveys
 279
- Ikram Chairi, Amina El Gonnouni, Sarah Zouinina, Abdelouahid Lyhyaoui
Prediction of Firm's Creditworthiness Risk using Feature Selection and Support Vector Machine
 285
- Sana Chakri, Said Raghay, Salah El Hadaj
Contribution of extracting meaningful patterns from semantic trajectories
 293
- Chieppa A., Ferrara R., Gallo G., Tomeo V.
Towards The Register-Based Statistical System: A New Valuable Source for Population Studies
 301
- Shirley Coleman
Consulting, knowledge transfer and impact case studies of statistics in practice
 305
- Michele Costa
The evaluation of the inequality between population subgroups
 313
- Michele Costola
Bayesian Non-Negative l_1 -Regularised Regression
 319
- Lisa Crosato, Caterina Liberati, Paolo Mariani, Biancamaria Zavarella
Industrial Production Index and the Web: an explorative cointegration analysis
 327

Index	XIII
Francesca Romana Crucinio, Roberto Fontana <i>Comparison of conditional tests on Poisson data</i>	333
Riccardo D'Alberto, Meri Raggi <i>Non-parametric micro Statistical Matching techniques: some developments</i>	339
Stefano De Cantis, Mauro Ferrante, Anna Maria Parroco <i>Measuring tourism from demand side</i>	345
Lucio De Capitani, Daniele De Martini <i>Optimal Ethical Balance for Phase III Trials Planning</i>	351
Claudia De Vitiis, Alessio Guandalini, Francesca Inglese, Marco D. Terribili <i>Sampling schemes using scanner data for the consumer price index</i>	357
Ermelinda Della Valle, Elena Scardovi, Andrea Iacobucci, Edoardo Tignone <i>Interactive machine learning prediction for budget allocation in digital marketing scenarios</i>	365
Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor <i>Nonparametric classification for directional data</i>	371
Edwin Diday <i>Introduction to Symbolic Data Analysis and application to post clustering for comparing and improving clustering methods by the Symbolic Data Table that they induce</i>	379
Carlo Drago <i>Identifying Meta Communities on Large Networks</i>	387

- Neska El Haouij, Jean-Michel Poggi, Raja Ghozi, Sylvie Sevestre Ghalila, Mériem Jaidane
Random Forest-Based Approach for Physiological Functional Variable Selection for Drivers Stress Level Classification
 393
- Silvia Facchinetti, Silvia A. Osmetti
A risk index to evaluate the criticality of a product defectiveness
 399
- Federico Ferraccioli, Livio Finos
Exponential family graphical models and penalizations
 405
- Mauro Ferrante, Giovanna Fantaci, Anna Maria Parroco, Anna Maria Milito, Salvatore Scondotto
Key-indicators for maternity hospitals and newborn readmission in Sicily
 411
- Ferretti Camilla, Ganugi Piero, Zammori Francesco
Change of Variables theorem to fit Bimodal Distributions
 417
- Francesco Finazzi, Lucia Paci
Space-time clustering for identifying population patterns from smartphone data
 423
- Annunziata Fiore, Antonella Simone, Antonino Virgillito
IT Solutions for Analyzing Large-Scale Statistical Datasets: Scanner Data for CPI
 429
- Michael Fop, Thomas Brendan Murphy, Luca Scrucca
Model-based Clustering with Sparse Covariance Matrices
 437
- Maria Franco-Villoria, Marian Scott
Quantile Regression for Functional Data
 441

Index	XV
Gallo M., Simonacci V., Di Palma M.A. <i>Three-way compositional data: a multi-stage trilinear decomposition algorithm</i>	445
Francesca Gasperoni, Francesca Ieva, Anna Maria Paganoni, Chris Jackson, Linda Sharples <i>Nonparametric shared frailty model for classification of survival data</i>	451
Stefano A. Gattone, Angela De Sanctis <i>Clustering landmark-based shapes using Information Geometry tools</i>	457
Alan E. Gelfand, Shinichiro Shirota <i>Space and circular time log Gaussian Cox processes with application to crime event data</i>	461
Abdelghani Ghazdali <i>Blind source separation</i>	469
Massimiliano Giacalone, Antonio Ruoto, Davide Liga, Maria Pilato, Vito Santarangelo <i>An innovative approach for Opinion Mining : the Plutchick analysis</i>	479
Massimiliano Giacalone, Demetrio Panarello <i>A G.E.D. method for market risk evaluation using a modified Gaussian Copula</i>	485
Chiara Gigliarano, Francesco Maria Chelli <i>Labour market dynamics and recent economic changes: the case of Italy</i>	491
Giuseppe Giordano, Giancarlo Ragozini, Maria Prosperina Vitale <i>On the use of DISTATIS to handle multiplex networks</i>	499

- Michela Gnaldi, Silvia Bacci, Samuel Greiff, Thiemo Kunze
Profiles of students on account of complex problem solving (CPS) strategies exploited via log-data
505
- Michela Gnaldi, Simone Del Sarto
Characterising Italian municipalities according to the annual report of the prevention-of-corruption supervisor: a Latent Class approach
513
- Silvia Golia
A proposal of a discretization method applicable to Rasch measures
519
- Anna Gottard
Tree-based Non-linear Graphical Models
525
- Sara Hbali, Youssef Hbali, Mohamed Sadgal, Abdelaziz El Fazziki
Sentiment Analysis for micro-blogging using LSTM Recurrent Neural Networks
531
- Stefano Maria Iacus, Giuseppe Porro, Silvia Salini, Elena Siletti
How to Exploit Big Data from Social Networks: a Subjective Well-being Indicator via Twitter
537
- Francesca Ieva
Network Analysis of Comorbidity Patterns in Heart Failure Patients using Administrative Data
543
- Antonio Irpino, Francisco de A.T. De Carvalho, Rosanna Verde
Automatic variable and components weighting systems for Fuzzy cmeans of distributional data
549
- Michael Jauch, Paolo Giordani, David Dunson
A Bayesian oblique factor model with extension to tensor data
553

Index	XVII
Johan Koskinen, Chiara Broccatelli, Peng Wang, Garry Robins <i>Statistical analysis for partially observed multilayered networks</i>	561
Francesco Lagona <i>Copula-based segmentation of environmental time series with linear and circular components</i>	569
Alessandro Lanteri, Mauro Maggioni <i>A Multiscale Approach to Manifold Estimation</i>	575
Tiziana Laureti, Carlo Ferrante, Barbara Dramis <i>Using scanner and CPI data to estimate Italian sub-national PPPs</i>	581
Antonio Lepore <i>Graphical approximation of Best Linear Unbiased Estimators for Extreme Value Distribution Parameters</i>	589
Antonio Lepore, Biagio Palumbo, Christian Capezza <i>Monitoring ship performance via multi-way partial least-squares analysis of functional data</i>	595
Caterina Liberati, Lisa Crosato, Paolo Mariani, Biancamaria Zavanella <i>Dynamic profiling of banking customers: a pseudo-panel study</i>	601
Giovanni L. Lo Magno, Mauro Ferrante, Stefano De Cantis <i>A comparison between seasonality indices deployed in evaluating unimodal and bimodal patterns</i>	607
Rosaria Lombardo, Eric J Beh <i>Three-way Correspondence Analysis for Ordinal-Nominal Variables</i>	613

- Monia Lupparelli, Alessandra Mattei
Log-mean linear models for causal inference
621
- Badiaa Lyoussi, Zineb Selihi, Mohamed Berraho, Karima El Rhazi, Youness El Achhab, Adiba El Marrakchi, Chakib Nejjari
Research on the Risk Factors accountable for the occurrence of degenerative complications of type 2 diabetes in Morocco: a prospective study
627
- Valentina Mameli, Debora Slanzi, Irene Poli
Bootstrap group penalty for high-dimensional regression models
633
- Stefano Marchetti, Monica Pratesi, Caterina Giusti
Improving small area estimates of households' share of food consumption expenditure in Italy by means of Twitter data
639
- Paolo Mariani, Andrea Marletta, Mariangela Zenga
Gross Annual Salary of a new graduate: is it a question of profile?
647
- Maria Francesca Marino, Marco Alfò
Dynamic random coefficient based drop-out models for longitudinal responses
653
- Antonello Maruotti, Jan Bulla
Hidden Markov models: dimensionality reduction, atypical observations and algorithms
659
- Chiara Masci, Geraint Johnes, Tommaso Agasisti
A flexible analysis of PISA 2015 data across countries, by means of multilevel trees and boosting
667

Index	XIX
Lucio Masserini, Matilde Bini <i>Impact of the 2008 and 2012 financial crises on the unemployment rate in Italy: an interrupted time series approach</i>	673
Angelo Mazza, Antonio Punzo, Salvatore Ingrassia <i>An R Package for Cluster-Weighted Models</i>	681
Antonino Mazzeo, Flora Amato <i>Methods and applications for the treatment of Big Data in strategic fields</i>	687
Letizia Mencarini, Viviana Patti, Mirko Lai, Emilio Sulis <i>Happy parents' tweets</i>	693
Rodolfo Metulini, Marica Manisera, Paola Zuccolotto <i>Space-Time Analysis of Movements in Basketball using Sensor Data</i>	701
Giorgio E. Montanari, Marco Doretto, Francesco Bartolucci <i>An ordinal Latent Markov model for the evaluation of health care services</i>	707
Isabella Morlini, Maristella Scorza <i>New fuzzy composite indicators for dyslexia</i>	713
Fionn Murtagh <i>Big Textual Data: Lessons and Challenges for Statistics</i>	719
Gaetano Musella, Gennaro Punzo <i>Workers' skills and wage inequality: A time-space comparison across European Mediterranean countries</i>	731

Marta Nai Ruscone <i>Exploratory factor analysis of ordinal variables: a copula approach</i>	737
Fausta Ongaro, Silvana Salvini <i>IPUMS Data for describing family and household structures in the world</i>	743
Tullia Padellini, Pierpaolo Brutti <i>Topological Summaries for Time-Varying Data</i>	747
Sally Paganin <i>Modeling of Complex Network Data for Targeted Marketing</i>	753
Francesco Palumbo, Giancarlo Ragozini <i>Statistical categorization through archetypal analysis</i>	759
Michela Eugenia Pasetto, Umberto Noè, Alessandra Luati, Dirk Husmeier <i>Inference with the Unscented Kalman Filter and optimization of sigma points</i>	767
Xanthi Pedeli, Cristiano Varin <i>Pairwise Likelihood Inference for Parameter-Driven Models</i>	773
Felicia Pelagalli, Francesca Greco, Enrico De Santis <i>Social emotional data analysis. The map of Europe</i>	779
Alessia Pini, Lorenzo Spreafico, Simone Vantini, Alessandro Vietti <i>Differential Interval-Wise Testing for the Inferential Analysis of Tongue Profiles</i>	785
Alessia Pini, Aymeric Stamm, Simone Vantini <i>Hotelling meets Hilbert: inference on the mean in functional Hilbert spaces</i>	791

Index	XXI
Silvia Poletini, Serena Arima <i>Accounting for measurement error in small area models: a study on generosity</i>	795
Gennaro Punzo, Mariateresa Ciommi <i>Structural changes in the employment composition and wage inequality: A comparison across European countries</i>	801
Walter J. Radermacher <i>Official Statistics 4.0 – learning from history for the challenges of the future</i>	809
Fabio Rapallo <i>Comparison of contingency tables under quasi-symmetry</i>	821
Valentina Raponi, Cesare Robotti, Paolo Zaffaroni <i>Testing Beta-Pricing Models Using Large Cross-Sections</i>	827
Marco Seabra dos Reis, Biagio Palumbo, Antonio Lepore, Ricardo Rendall, Christian Capezza <i>On the use of predictive methods for ship fuel consumption analysis from massive on-board operational data</i>	833
Alessandra Righi, Mauro Mario Gentile <i>Twitter as a Statistical Data Source: an Attempt of Profiling Italian Users Background Characteristics</i>	841
Paolo Righi, Giulio Barcaroli, Natalia Golini <i>Quality issues when using Big Data in Official Statistics</i>	847
Emilia Rocco <i>Indicators for the representativeness of survey response as well as convenience samples</i>	855

- Emilia Rocco, Bruno Bertaccini, Giulia Biagi, Andrea Giommi
A sampling design for the evaluation of earthquakes vulnerability of the residential buildings in Florence
861
- Elvira Romano, Jorge Mateu
A local regression technique for spatially dependent functional data: an heteroskedastic GWR model
867
- Eduardo Rossi, Paolo Santucci de Magistris
Models for jumps in trading volume
873
- Renata Rotondi, Elisa Varini
On a failure process driven by a self-correcting model in seismic hazard assessment
879
- M. Ruggieri, F. Di Salvo and A. Plaia
Functional principal component analysis of quantile curves
887
- Massimiliano Russo
Detecting group differences in multivariate categorical data
893
- Michele Scagliarini
A Sequential Test for the C_{pk} Index
899
- Steven L. Scott
Industrial Applications of Bayesian Structural Time Series
905
- Catia Scricciolo
Asymptotically Efficient Estimation in Measurement Error Models
913

Index	XXIII
Angela Serra, Pietro Coretto, Roberto Tagliaferri <i>On the noisy high-dimensional gene expression data analysis</i>	919
Mirko Signorelli <i>Variable selection for (realistic) stochastic blockmodels</i>	927
Marianna Siino, Francisco J. Rodriguez-Cortés, Jorge Mateu, Giada Adelfio <i>Detection of spatio-temporal local structure on seismic data</i>	935
A. Sottosanti, D. Bastieri, A. R. Brazzale <i>Bayesian Mixture Models for the Detection of High-Energy Astronomical Sources</i>	943
Federico Mattia Stefanini <i>Causal analysis of Cell Transformation Assays</i>	949
Paola Stolfi, Mauro Bernardi, Lea Petrella <i>Estimation and Inference of SkewStable distributions using the Multivariate Method of Simulated Quantiles</i>	955
Paola Stolfi, Mauro Bernardi, Lea Petrella <i>Sparse Indirect Inference</i>	961
Peter Struijs, Anke Consten, Piet Daas, Marc Debusschere, Maiki Ilves, Boro Nikic, Anna Nowicka, David Salgado, Monica Scannapieco, Nigel Swier <i>The ESSnet Big Data: Experimental Results</i>	969
Jérémie Sublime <i>Smart view selection in multi-view clustering</i>	977

- Emilio Sulis
Social Sensing and Official Statistics: call data records and social media sentiment analysis
985
- Matilde Trevisani, Arjuna Tuzzi
Knowledge mapping by a functional data analysis of scientific articles databases
993
- Amalia Vanacore, Maria Sole Pellegrino
Characterizing the extent of rater agreement via a non-parametric benchmarking procedure
999
- Maarten Vanhoof, Stephanie Combes, Marie-Pierre de Bellefon
Mining Mobile Phone Data to Detect Urban Areas
1005
- Viktoriya Voytsekhovska, Olivier Butzbach
Statistical methods in assessing the equality of income distribution, case study of Poland
1013
- Ernst C. Wit
Network inference in Genomics
1019
- Dilek Yildiz, Jo Munson, Agnese Vitali, Ramine Tinati, Jennifer Holland
Using Twitter data for Population Estimates
1025
- Marco Seabra dos Rei
Structured Approaches for High-Dimensional Predictive Modeling
1033

Preface

The 2017 SIS Conference aims to highlight the crucial role of the Statistics in Data Science. In this new domain of “meaning” extracted from the data, the increasing amount of produced and available data in databases, nowadays, has brought new challenges. That involves different fields of statistics, machine learning, information and computer science, optimization, pattern recognition. These afford together a considerable contribute in the analysis of “Big data”, open data, relational and complex data, structured and no-structured. The interest is to collect the contributes which provide from the different domains of Statistics, in the high dimensional data quality validation, sampling extraction, dimensional reduction, pattern selection, data modelling, testing hypotheses and confirming conclusions drawn from the data. In the mention that statistics is the “grammar of data science”, statistics has become a basic skill in data science: it gives right meaning to the data. Still, it isn’t replaced by newer techniques from machine learning and other disciplines but it complements them. The Conference is also addressed to the new challenges of the new generations: the native digital generations, who are called to develop professional skills as “data analyst”, one of the more request professionalism of the 21st Century, crossing the rigid disciplinary domains of competence. In this perspective, all the traditional statistical topics are admitted with an extension to the related machine learning and computer science ones. The present volume includes the short papers of the contributions that will be presented in the 4 invited speaker sessions; in the 19 specialized sessions; in the 11 solicited sessions; in the 6 foreign societies sessions and in the 17 contributed sessions as well as, in the panel session.

Rosanna Verde
President of the Scientific Programme Committee

Alessandra Petrucci
President of the Local Organizing Committee

Models for jumps in trading volume

Modelli per i salti nel trading volume

Eduardo Rossi and Paolo Santucci de Magistris

Abstract In finance theory the log-price is often supposed to follow an Ito semimartingale while no explicit assumptions are made on the dynamic evolution of trading volumes. Trading volume is a measure of the quantity of shares that change owners for a given security. The amount of daily volume on a security can fluctuate on any given day depending on the amount of new information available about the company. We assume that the dynamic evolution of trading volume is represented as a semimartingale. Analogously to stock prices, the stochastic process for trading volume might be characterized by jump components. We distinguish between two classes of widely used processes: Brownian semimartingales plus jumps and pure-jump models. The relative contribution of each of two components is estimated by means of alternative nonparametric methods. We also analyze if the jump component is a stochastic process of finite or infinite variation. Finally, alternative parametric models are estimated and compared.

Abstract *Nella teoria della finanza si assume che il processo stocastico del log-prezzo segua una semimartingala di Ito mentre non sono esplicitate le ipotesi sulla dinamica dei volumi scambiati (trading volume). Il trading volume di un titolo azionario è il numero di azioni scambiate. L'ammontare di volume giornaliero relativo al singolo titolo pu fluttuare ogni giorno in funzione delle nuove informazioni disponibili. Si assume che l'evoluzione dinamica del trading volume possa essere rappresentata da una semimartingala. Analogamente a quanto si suppone per i log-prezzi, il processo stocastico per il trading volume caratterizzato dalla presenza di una componente di salto. Nel lavoro si distingue tra due classi di processi: semimartingale browniane con salti e modelli di salto. Il contributo relativo di ognuna*

Eduardo Rossi

European Commission, Joint Research Centre, Directorate Innovation and Growth, B.01, Italy.
Dipartimento di Scienze Economiche ed Aziendali, University of Pavia, 27100 Pavia, Italy. e-mail: eduardo.rossi@unipv.it

Paolo Santucci de Magistris

Department of Economics and Business Economics and CREATES, Aarhus University, Denmark,
email: e-mail: psantucci@econ.au.dk

delle componenti stimato con tecniche non parametriche. Si indaga anche se la componente di salto un processo stocastico di variazione finita o infinita. Infine, sono stimati e comparati modelli parametrici alternativi.

Key words: Trading Volume, Jumps, Activity level, Infinite variation.

1 Introduction

For each market equilibrium we have an equilibrium price and quantity. In finance theory the price is often supposed to follow an Ito semimartingale while no explicit assumptions are made on the dynamic evolution of trading volumes. Trading volume is a measure of the quantity of shares that change owners for a given security. The amount of daily volume on a security can fluctuate on any given day depending on the amount of new information available about the company, whether options contracts are set to expire soon, whether the trading day is a full or half day, and many other possible factors. Of the many different elements affecting trading volume, the one which correlates the most to the fundamental valuation of the security is the new information provided. This information can be a press release or a regular earnings announcement provided by the company, or it can be a third party communication, such as a court ruling or a release by a regulatory agency pertaining to the company. The news release can generate large variations in the trading volume. The trading volume can be measured instantaneously for each trade or cumulated for a given time interval. This implies that for longer time intervals the trading volume is an increasing process. This is not the case for the price process.

As in the case of prices, we assume that the dynamic evolution of trading volume is represented as an Itô semimartingale (SM) defined on a filtered probability space $(\Omega; \mathcal{F}; (\mathcal{F})_{t \in [0, T]}; \mathcal{P})$ satisfying usual conditions, evolving as

$$X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s ds + \sum_{s \leq t} \Delta X_s$$

where

$$\Delta X_s = X_s - X_{s-}$$

is the size of the jump at time s . Even when the whole path of X is observed over $[0, T]$ one can infer neither the drift nor the Lévy measure. With a finite T we can only infer the behavior of the Lévy measure near 0. For a semimartingale the activity index takes values in the interval $[0, 2]$. For a Lévy process the jump activity index coincides with the Blumenthal-Gettoor index of the process [1, 2]. The index takes its values in $[0; 2]$ and allows to distinguish different classes of stochastic processes. The Blumenthal-Gettoor index is zero for finite activity jump processes (which have finite number of jumps in any finite interval) and it is equal to two for continuous (local) martingales. Stochastic processes with Blumenthal-Gettoor indices in $(0; 2)$

are infinitely active pure-jump processes, with paths of infinite variation if and only if the index is larger than unity. When the process is the result of the sum of a jump component and a continuous process driven by Brownian motion, its activity index will take a value of 2 independently from the activity of the jumps. In general, the jump activity of a superposition of different Ito semimartingales is equal to the Blumenthal-Gettoor index of the most active component. If X is a stable process β is also the stable index of the process. β captures the level of the activity: when β increases the (small) jumps tend to become more and more frequent.

The main research question of this paper is: which process best approximates the trading volume dynamics? In other words, we want to distinguish between two classes of widely used processes in modeling the dynamics of financial prices: Brownian semimartingales plus jumps (with Blumenthal-Gettoor index equal to 2) and pure-jump models (with Blumenthal-Gettoor index less than 2). The study of the trading volume (TV) dynamics allows to better understand the role played by small and large jumps in equilibrium and on the microstructure of financial markets.

2 Which jumps in trading volume?

We assume that the observations are collected at a discrete sampling interval Δ_n , which means that there are $[T/\Delta_n]$ observed increments of X on $[0, T]$, i.e.

$$\Delta_n^i X = X_{i\Delta_n} - X_{(i-1)\Delta_n}.$$

Let μ the jump measure of X and ν its predictable compensator, Lévy measure. Both positive measure on $\mathbb{R}_+ \times \mathbb{R}$.

$$\begin{aligned} \text{Small jumps} &= \int_0^t \int_{|x| \leq \varepsilon} x(\mu - \nu)(ds, dx) \\ \text{Big jumps} &= \int_0^t \int_{|x| > \varepsilon} x\mu(ds, dx) \end{aligned}$$

where the cutoff level $\varepsilon > 0$ is arbitrary, but fixed. A SM will always generate a finite number of big jumps on $[0, T]$ but it may give rise to either a finite or infinite number of small jumps, i.e.

$$\nu([0, t] \times (-\infty, -\varepsilon) \cup (\varepsilon, +\infty)) < \infty$$

whereas

$$\nu([0, t] \times [-\varepsilon, \varepsilon])$$

may be finite or infinite.

Using the methodology of power variation:

$$V(p) = \int_0^T |\sigma_s|^p ds$$

$$J(p) = \sum_{s \leq T} |\delta X_s|^p \quad p > 0.$$

1. $V(p)$ is finite $\forall p > 0$, and $V(p) > 0$ on Ω_T^W .
2. $J(p)$ is finite if $p \geq 2$ but often not when $p < 2$.

The realized power variations proposed by Ait-Sahalia & Jacod [2],

$$B(p, u_n, \Delta_n) = \sum_{i=1}^{[T/\Delta_n]} |\Delta_i^n X|^p \mathbf{1}_{\{|\Delta_i^n X| \leq u_n\}}$$

where u_n is a sequence of truncation levels. With T fixed, the asymptotics are all with respect to $\Delta_n \rightarrow 0$. Since u_n has to converge to 0, $u_n = \alpha \Delta_n^\varpi$, $\varpi \in (0, 1/2)$, and $\alpha > 0$. With $\varpi < 1/2$ we keep all the increments that mainly contain a Brownian contribution. The in-fill asymptotics:

$$p > 2, \forall X \implies B(p, \infty, \Delta_n) J(p)$$

$$\forall p, \text{ on } \Omega_T^c \implies \frac{\Delta_n^{1-p/2}}{m_p} B(p, \infty, \Delta_n) V(p)$$

m_p is the p th absolute moment of $z \sim N(0, 1)$. When $p > 2$

$$B(p, \infty, \Delta_n) \xrightarrow{P} J(p)$$

the jump component dominates. If there are jumps the limit $J(p)_t > 0$ is finite. If there are no jumps, X is continuous, then

$$J(p) = 0 \quad B(p, \infty, \Delta_n) \xrightarrow{P} 0$$

at rate $\Delta_n^{p/2-1}$. We can exploit the different asymptotic behavior of $B(p, u_n, \Delta_n)$ by varying the tuning parameters:

1. the power p : to isolate either the continuous or jump components or to keep both.
 - $p < 2$ emphasizes the continuous component
 - $p > 2$ accentuates the jump component
 - $p = 2$ equal treatment
2. the truncation level u_n . The assumption is that there exists a finite number of large jumps with fixed size. As $\Delta_n \rightarrow 0$, u_n becomes smaller than the large jumps which are thus no longer part of $B(p, u_n, \Delta_n)$. Alternatively, we can truncate to eliminate the Brownian component using the upward power variation

$$U(p, u_n, \Delta_n) = \sum_{i=1}^{[T/\Delta_n]} |\Delta_i^n X|^p \mathbf{1}_{\{|\Delta_i^n X| > u_n\}}$$

- the sampling frequency Δ_n . Sampling at different frequencies we can distinguish three cases based on the asymptotic behavior of the ratio

$$\frac{B(p, u_n, k\Delta_n)}{B(p, u_n, \Delta_n)} \quad k \geq 2$$

As $\Delta_n \rightarrow 0$, the limiting behavior can be

- $\frac{B(p, u_n, k\Delta_n)}{B(p, u_n, \Delta_n)} = 1$, $B(p, u_n, k\Delta_n)$ converges to a finite limit
- $\frac{B(p, u_n, k\Delta_n)}{B(p, u_n, \Delta_n)} < 1$, $B(p, u_n, k\Delta_n)$ diverges to infinity
- $\frac{B(p, u_n, k\Delta_n)}{B(p, u_n, \Delta_n)} > 1$, $B(p, u_n, k\Delta_n)$ converges to 0

The model includes three components: a continuous part, a small jumps part and a big jumps part. Accordingly we can describe the possible behavior by means of sets defined pathwise on $[0, T]$

- $\Omega_T^c = \{X \text{ is continuous in } [0, T]\}$
- $\Omega_T^j = \{X \text{ has jumps in } [0, T]\}$
- $\Omega_T^f = \{X \text{ has finitely many jumps in } [0, T]\}$
- $\Omega_T^i = \{X \text{ has infinitely many jumps in } [0, T]\}$
- $\Omega_W^i = \{X \text{ has a Wiener component in } [0, T]\}$
- $\Omega_{noW}^i = \{X \text{ has no Wiener component in } [0, T]\}$

We should also note that we observe a time series originating in a given unobserved path in Ω_T and wish to determine in which sets the path is likely to be. Any such time series can be obtained by discretization of a continuous path and also of a discontinuous one.

The jump activity index at time t is the random number (see [1])

$$\beta_t^i = \inf \left\{ r > 0 : \int_{\mathbb{R}} (|x|^r \wedge 1) F_s(dx) < \infty \right\}$$

Following [3] $u_n = \alpha \Delta_n^\omega$ and $u'_n = \alpha' \Delta_n^\omega$

$$\hat{\beta} = \frac{\log(U(0, u_n, \Delta_n)/U(0, \gamma u_n, \Delta_n))}{\log(\gamma)}$$

$$\gamma = \alpha' / \alpha$$

By using the statistic U , which simply counts the number of large increments, defined as those greater than $\alpha \Delta_n^\omega$, we are retaining only those increments of X that are not predominantly made of contributions from its continuous semimartingale part, which are $O_p(\Delta_n^{1/2})$, and instead are predominantly made of contributions due to a jump. When X has only finitely many jumps, the index is $\beta = 0$ and $U(p, u_n, \Delta_n)$ converges to the number of jumps between 0 and t , irrespective of the value of α , so $\hat{\beta} = 0$ for all $n = \lceil T/\Delta_n \rceil$ large enough.

The paper presents and discuss the results of the techniques shown above to high-frequency data of SPY and individual stocks traded on the NYSE.

References

1. Aït-Sahalia, Y., Jacod, J.: Estimating the degree of activity of jumps in high frequency data. *Annals of Statistics*, **37**(5A), 2202–2244 (2009)
2. Aït-Sahalia, Y., Jacod, J.: Analyzing the Spectrum of Asset returns: Jump and Volatility components in High Frequency Data. *Journal of Economic Literature*, **50**(4), 1007–1050 (2012)
3. Aït-Sahalia, Y., Jacod, J.: *High-frequency financial econometrics*. Princeton University Press, Princeton and Oxford (2014)