Book of the Short Papers

Editors: Francesco Maria Chelli, Mariateresa Ciommi, Salvatore Ingrassia, Francesca Mariani, Maria Cristina Recchioni

Società Italiana di Statistica
UNIVERSITÀ POLITECNICA DELLE MARCHE
Dipartimento di Scienze Economiche e Sociali DISES
LIUC Università Cattaneo
LIUC Università Cattaneo | BUSINESS ANALYTICS AND DATA SCIENCE HUB
Pearson

CHAIRS

Salvatore Ingrassia (Chair of the Program Committee) - *Università degli Studi di Catania*
Maria Cristina Recchioni (Chair of the Local Organizing Committee) - *Università Politecnica delle Marche*

PROGRAM COMMITTEE

Salvatore Ingrassia (Chair), Elena Ambrosetti, Antonio Balzanella, Matilde Bini, Annalisa Busetta, Fabio Centofanti, Francesco M. Chelli, Simone Di Zio, Sabrina Giordano, Rosaria Ignaccolo, Filomena Maggino, Stefania Mignani, Lucia Paci, Monica Palma, Emilia Rocco.

LOCAL ORGANIZING COMMITTEE

Maria Cristina Recchioni (Chair), Chiara Capogrossi, Mariateresa Ciommi, Barbara Ermini, Chiara Gigliarano, Riccardo Lucchetti, Francesca Mariani, Gloria Polinesi, Giuseppe Ricciardo Lamonica, Barbara Zagaglia.

ORGANIZERS OF INVITED SESSIONS

Pierfrancesco Alaimo Di Loro, Laura Anderlucci, Luigi Augugliaro, Ilaria Benedetti, Rossella Berni, Mario Bolzan, Silvia Cagnone, Michela Cameletti, Federico Camerlenghi, Gabriella Campolo, Christian Capezza, Carlo Cavicchia, Mariateresa Ciommi, Guido Consonni, Giuseppe Ricciardo Lamonica, Regina Liu, Daniela Marella, Francesca Mariani, Matteo Mazziotta, Stefano Mazzuco, Raya Muttarak, Livia Elisa Ortensi, Edoardo Otranto, Ilaria Prosdocimi, Pasquale Sarnacchiaro, Manuela Stranges, Claudia Tarantola, Isabella Sulis, Roberta Varriale, Rosanna Verde.

FURTHER PEPOPLE OF LOCAL ORGANIZING COMMITTEE

Elisa D'Adamo, Christian Ferretti, Giada Gabbianelli, Elvina Merkaj, Luca Pedini, Alessandro Pionati, Marco Tedeschi, Francesco Valentini, Rostand Arland Yebetchou Tchounkeu

Technical support: Matteo Mercuri, Maila Ragni, Daniele Ripanti

PUBLISHED BY PEARSON
WWW.PEARSON.COM
ISBN 9788891935618AAVV

Contents

Preface XXII

1 Plenary Sessions 1

Inequality indices: accurate simulation-based inference 2
Maria-Pia Victoria-Feser

Examples from the Interface of Neural Models and Spatio-Temporal Statistics in Environmental Applications 7
Christopher K. Wikle, Likun Zhang, Myungsoo Yoo and Xiaoyu Ma

Demographic change and sustainability: novel approaches from digital and computational demography n.a.
Emilio Zagheni

2 Invited Sessions 14

Machine learning in the design, analysis and integration of sample surveys

Causal Discovery for complex survey data 15
Paola Vicard

Data Integration without conditional independence: a Bayesian Networks approach 21
Pier Luigi Conti, Paola Vicard and Vincenzina Vitale

Mass imputation through Machine Learning techniques in presence of multi-source data 27
Fabrizio De Fausti, Marco Di Zio, Romina Filippini and Simona Toti

Machine learning: different uses and perspectives

Evaluation of pollution containment policies in the US and the role of machine learning algorithms 32
Marco Di Cataldo, Margherita Gerolimetto, Stefano Magrini and Alessandro Spiganti

Machine Learning for Official Statistics: An Application on External Trade n.a.
Mauro Bruno, Maria Serena Causo, Alessio Guandalini, Francesco Ortame and Silvia Russo

Machine learning, data quality and official statistics: challenges and opportunities n.a.
Stefano Menghinello

Statistical Machine Learning for environmental applications

Gaussian Processes and Deep Neural Networks for Spatial Prediction 38
Alex Cucco, Luigi Ippoliti, Nicola Pronello, Pasquale Valentini and Carlo Zaccardi

How can we explain Random Forests in a spatial framework? 42
Natalia Golini, Luca Patelli and Xavier Barber

Recent approaches in coupling deep learning methods with the statistical analysis of spatial point patterns 48
Jorge Mateu and Abdollah Jalilian

Statistical Process Monitoring for Complex Data in Industry 4.0

A Kernel-based Nonparametric Multivariate CUSUM for Location Shifts 53
Konstantinos Bourazas, Konstantinos Fokianos, Christos Panayiotou and Marios Polycarpou

An Approach for Profile Monitoring via Mixture Regression Models 58
Davide Forcina, Antonio Lepore and Biagio Palumbo

Anomaly Detection in Circular Data 63
Houyem Demni and Giovanni C. Porzio

Advances in Data Science and Statistical Learning [IMS Invited Session]

Empirical Bayes approximation of Bayesian learning: understanding a common practice n.a.
Sonia Petrone

Generalized Fiducial Inference on Differentiable Manifolds - a geometric perspective n.a.
Jan Hannig

Model-free bootstrap and conformal prediction in regression n.a.
Dimitris Politis

ENBIS Session: System Maintenance, Boosting algorithms for regression, and Research Excellence

Boosting Diversity in Regression Ensembles 69
Mathias Bourel, Jairo Cugliari, Yannig Goude and Jean-Michel Poggi

How ENBIS has contributed to the UK Universities Research Excellence Framework 71
Shirley Coleman

Maintenance of degrading systems by dynamic programming or reinforcement learning 75
Antonio Pievatolo

Population Dynamics, Climate Change and Sustainability

Climate change impacts on fertility in low- and middle-income countries: An analysis based on global sub-national data n.a.
Côme Cheritel, Roman Hoffmann and Raya Muttarak

Environmental Exposures and Under-5 Mortality in India: A Survival Analysis of DHS data 79
Vinod Joseph Kannankeril Joseph

The impact of temperature on expressed sentiment by migration status: Evidence from geo-located Twitter data 84
Risto Conte Keivabu and Jisu Kim

Statistical Learning for health research and omics data

An alternative to the Dirichlet-multinomial regression model for microbiome data analysis 95
Roberto Ascari, Sonia Migliorati and Andrea Ongaro

Modelling ordinal response to treatment in a real-world cohort study 101
Marco Alfò, Maria Francesca Marino and Silvia D'Elia

On the application of the symmetric graphical lasso for paired data 105
Saverio Ranciati and Alberto Roverato

The Economic behaviour of Sustainability

Airports performances and sustainable practices. An empirical study on Italian data 110
Riccardo Gianluigi Serio, Maria Michela Dickson, Diego Giuliani and Giuseppe Espa

Sustainability: still an undefined concept for Italians 116
Raffaele Angelone and Andrea Marletta

Quasi-experimental evidence on COVID-19 lockdown effects on Italian household food shopping basket composition and its sustainability 122
Beatrice Biondi and Mario Mazzocchi

Advances in statistical methods for complex problems

Inferring multiple treatment effects from observational studies using confounder importance learning n.a.
Omiros Papaspiliopoulos

Path analysis in Ising models: an application to cyber-security risk assessment 127
Monia Lupparelli and Giovanni M. Marchetti

Causal Regularization n.a.
Lucas Kania and Ernst Wit

Explainable machine learning models

Enhancing Markowitz model: inspection of correlations and tail covariances 133
Gloria Polinesi

Objective and subjective dimension of economic well-being: an approach based on statistical matching 139
Daniela Marella, Vincenzina Vitale and Pierpaolo D'Urso

Sustainable, Accurate, Fair and Explainable Machine Learning Models n.a.
Paolo Giudici and Emanuela Raffinetti

Flexible Learning for Environmental Sustainability

Comparison of traffic flow data sources for air pollution modelling 145
Theresa Smith and Nick McCullen

Data analysis of photogrammetry-based mapping: the sea cucumbers in the Giglio Island as a case-study 150
Gianluca Mastrantonio, Daniele Ventura, Edoardo Casoli, Arnold Rakaj, Giovanna Jona Lasinio and Alessio Pollice

Understanding forest damage in Germany: Finding key drivers to help with future forest conversion of climate sensitive 156
Nicole Augustin, Heike Puhlmann and Simon Trust

Inequalities in higher education outcomes: learning from data

Inequalities in international students mobility 163
Kristijan Breznik, Giancarlo Ragozini and Marialuisa Restaino

Uncovering the interplay of territorial, socioeconomic, and demographic factors in high school to university transition 169
Vincenzo Giuseppe Genova, Andrea Priulla and Martina Vittorietti

Statistical Learning of demographic and health dynamics

Estimating the impact of a vaccine mandate: the case of measles in Italy n.a.
Chiara Chiavenna

Leveraging deep neural networks to estimate age-specific mortality from life expectancy at birth n.a.
Andrea Nigri

Nowcasting Daily Population Displacement in Ukraine through Social Media Advertising Data n.a.
Claire Dooley, Ridhi Kashyap, Douglas Leasure and Francesco Rampazzo

Challenges towards Fairness and Transparency for Data Processes, Algorithms and Decision-Support Models

Challenges on Ethics, and Privacy in AI Applications to Fintech 175
Catarina Silva, Joana Matos Dias and Bernardete Ribeiro

Uncertainty and fairness metrics 180
Anna Gottard

Educational Data mining: methods for complex data in students' assessment

Analysis of University Grades: An IRT Model for Responses and Response Times with Censoring 186
Michela Battauz

Predicting high schools' students performances with registry's data: a machine learning approach 191
Lidia Rossi, Marta Cannistrà and Tommaso Agasisti

Using response times to identify cheaters in CAT: A simulation study 195
Luca Bungaro, Bernard P. Veldkamp and Mariagiulia Matteucci

Spatial and Spatio-Temporal Modeling: Theory and Applications

A geostatistical investigation of the ammonia-livestock relationship in the Po Valley, Italy 200
Paolo Maranzano, Kelly McConville , Philipp Otto and Felicetta Carillo

Bayesian multi-species N-mixture models for large scale spatial data in community ecology 206
Michele Peruzzi

Minimum contrast for point processes' first-order intensity estimation 212
Nicoletta D'Angelo and Giada Adelfio

Statistical Framework for Measuring the Sustainability of Tourism

Data validity and statistical conformity with Benford's Law: the case of tourism in Sicily 217
Roy Cerqueti and Davide Provenzano

Exploring the level of digitalization of the Italian museums through a multilevel ordered logit model 228
Claudia Cappello, Sabrina Maggio and Sandra De Iaco

Functional Partial Least-Squares via Regression Splines. An application on Italian Sustainable Development Goals data 232
Ida Camminatiello, Rosaria Lombardo, Jean-Francois Durand and Leonardo S. Alaimo

Statistical learning for well-being analysis

Assessing multidimensional poverty of the Italian provinces during Covid-19: a small area estimation approach 238
Mariateresa Ciommi, Chiara Gigliarano, Francesca Mariani and Gloria Polinesi

The fuzzy set approach as statistical learning for the analysis of multidimensional well-being 244
Gianni Betti, Federico Crescenzi, Antonella D'Agostino and Laura Neri

What Makes a Satisfying Life? Prediction and Interpretation with Machine-Learning Algorithms n.a.
Conchita D'Ambrosio

Bayesian contributions to Statistical Learning

A Bayesian framework for early cancer screening 249
Sally Paganin and Jeff Miller

Imputing Synthetic Pseudo Data from Aggregate Data: Development and Validation for Precision Medicine n.a.
Cecilia Balocchi

Linear models with assumptions-free residuals: a Bayesian Nonparametric approach 254
Filippo Ascolani and Valentina Ghidini

Data Visualization for Smart Insights and Advanced Predictive Analytics

Applications of data visualization for industry 259
Martina Dossi, Stefano Sangaletti, Marilena Di Bari and Federica Bruschini

Some Notes on the Use of the Circular Boxplot n.a.
Giovanni Camillo Porzio and Davide Buttarazzi

TERRA: a smart visualization tool for international trade in goods statistics 265
Francesco Amato, Mauro Bruno and Maria Serena Causo

Methods for the analysis of distributional data

Clustering of Distributional Data based on LDQ transformation 271
Gianmarco Borrata and Rosanna Verde

Dynamic learning from data streams through the combined use of probability density functions and simplicial functional principal component analysis 276
Francesca Fortuna, Fabrizio Maturo and Tonio Di Battista

Multivariate Parametric Analysis of Distributional Data n.a.
Paula Brito

Migrants and Refugees in Europe: social, economic and health-related issues

Labor Market Return to Refugees' Human Capital Investment: A Natural Experiment in Sweden n.a.
Eleonora Mussino

Social networks and loneliness among older migrants in Italy 282
Viviana Amati, Eralba Cela and Elisa Barbiano di Belgiojoso

The Italian Decree on Security: An Analysis of the Impact on Asylum Applications 287
Giorgio Piccitto

Modelling and Forecasting High-dimensional time series

Adaptive combinations of tail-risk forecasts 293
Alessandra Amendola, Vincenzo Candila, Antonio Naimoli and Giuseppe Storti

Are Monetary Policy Announcements related to Volatility Jumps? 299
Giampiero Gallo, Demetrio Lacava and Edoardo Otranto

Regularized Estimation and Prediction of the El Nino/Southern Oscillation Cycle n.a.
Alessandro Giovannelli and Tommaso Proietti

3 Contributed Sessions 305

Bayesian nonparametric methods

Bayesian density estimation for modeling age-at-death distribution 306
Davide Agnoletto, Tommaso Rigon and Bruno Scarpa

Bayesian mixing distribution estimation in the Gaussian-smoothed 1-Wasserstein distance 311
Catia Scricciolo

Bayesian nonparametric estimation of heterogeneous intrinsic dimension via product partition models 316
Francesco Denti, Antonio Di Noia and Antonietta Mira

Bayesian nonparametric multiple change point detection for time series of compositional data 322
Edoardo Marchionni and Riccardo Corradin

Galton-Watson process: a non parametric prior for the offspring distribution 328
Massimo Cannas, Michele Guindani and Nicola Piras

Hierarchical processes in survival analysis 333
Riccardo Cogo, Federico Camerlenghi and Tommaso Rigon

Economics and Statistics

A regression analysis for count data to investigate the effectiveness of incentives on the adoption of 4.0 technologies 339
Stefano Bonnini and Michela Borghesi

Statistical analysis on SDGs indicators related to environmental sustainability 344
Najada Firza, Anisa Bakiu and Dante Mazzitelli

Empowering futures adopting a spatial convergence of opinions: a Real-Time Spatial Delphi approach 349
Yuri Calleo, Simone Di Zio and Francesco Pilla

Stocks price forecasts using Stochastic Differential Equations: an empirical assessment 355
Dario Frisardi and Matteo Spuri

The Added-Worker Effect within Italian Households 361
Donata Favaro and Anna Giraldo

Health statistics 1

A model for the natural history of breast cancer: application to a Norwegian screening dataset 365
Laura Bondi, Marco Bonetti and Solveig Hofvind

Generalized Bayesian Ensemble Survival Trees: an extension to categorical variables to apply it to real data 370
Elena Ballante

Joint modelling of hospitalizations and survival in Heart Failure patients: a discrete non parametric frailty approach 375
Chiara Masci, Marta Spreafico and Francesca Ieva

Mobility trends in Italy during the first wave of Covid-19 pandemic: analysis on Google data 381
Ilaria Bombelli and Daniele De Rocchi

Tracking attitudes towards COVID vaccines: A text mining analysis 387
Leonardo Scarso, Marco Novelli and Francesco Saverio Violante

Treatment effect assessment in observational studies with multi-level treatment and outcome 393
Federica Cugnata, Paola Vicard, Paola M.V. Rancoita, Fulvia Mecatti, Clelia Di Serio and Pier Luigi Conti

Indicators: composition, uses and limitations

Are European consumers willing to pay the true price for sustainable food? 399
Luca Secondi and Mengting Yu

Can the reliability of composite indexes be impacted by uncertainty of individual indicators? 406
Caterina Giusti, Stefano Marchetti and Vincenzo Mauro

Initial Coin Offerings and ESG: allies or enemies? 411
Alessandro Bitetto and Paola Cerchiello

On the impact of intraclass correlation in the ANVUR evaluation of academic departments 417
Giorgio Edoardo Montanari and Marco Doretti

Small area estimation of monetary poverty indicators with poverty lines adjusted using local price indexes 422
Luigi Biggeri, Stefano Marchetti, Caterina Giusti, Monica Pratesi, Francesco Schirripa Spagnolo and Gaia Bertarelli

Smart Composite Indicators Measuring Corporate Sustainability: A Sensitivity Analysis 428
Camilla Salvatore, Annamaria Bianchi and Silvia Biffignandi

Multivariate data analysis 1

A note on most powerful tests for right censored survival data 434
Maria Veronica Vinattieri and Marco Bonetti

Enhancing Principal Components by a Linear Predictor: an Application to Well-Being Italian Data 439
Laura Marcis, Maria Chiara Pagliarella and Renato Salvatore

Proper Bayesian Bootstrap for Bagging tree model in survival analysis with correlated data 445
Farah Naz and Elena Ballante

ROBOUT: a multi-step methodology for conditional outlier detection 450
Matteo Farnè and Angelos Vouldis

Robustness of the Efficient Covariate-Adaptive Design for balancing covariates in comparative experiments 456
Rosamarie Frieri, Alessandro Baldi Antognini, Maroussa Zagoraiou, and Marco Novelli

Separation scores: a new statistical tool for scoring and ranking partially ordered data 462
Marco Fattore

Statistics in Society 1

Community detection analysis with robin on hashtag network 468
Valeria Policastro, Francesco Santelli and Giancarlo Ragozini

Film Tourism Motivation through the lens of Trip Advisor data 474
Nicolò Biasetton, Marta Disegna, Girish Prayag and Elena Barzizza

Life satisfaction and social activities in later life in Italy: a focus on the Internet use 480
Claudia Furlan and Silvia Meggiolaro

Social capital endowment's role in the intergenerational transmission of education 485
Alessandra Trimarchi, Maria Gabriella Campolo and Antonino Di Pino Incognito

Streaming Data from Social Networks to Track Political Trends 490
Emiliano del Gobbo and Barbara Cafarelli

The scientific production on gender dysphoria: a bibliometric analysis 495
Maria Gabriella Grassia, Marina Marino, Massimo Aria, Rocco Mazza, Luca D'Aniello and Agostino Stavolo

Assessment and Education

A hierarchical modelling approach to explain differential functioning of mathematics items by student's gender 500
Clelia Cascella

A latent variable approach to Millennials' knowledge of green finance 506
Maria Iannario, Alessandra Tanda and Claudia Tarantola

Archetypal analysis and latent Markov models: A step-wise approach 512
Lucio Palazzo, Rosa Fabbricatore and Francesco Palumbo

From high school to university: academic intentions and enrolment of foreign students in Italy 518
Francesca Di Patrizio, Eleonora Trappolini and Cristina Giudici

Growth models for the progress test in Italian dentistry degree program 523
Giulio Biscardi, Leonardo Grilli, Carla Rampichini, Laura Antonucci and Corrado Crocetta

The COVID-19 pandemic and academic E-learning: Italian students and instructors' perceptions 527
Francesco Santelli, Teresa Gentile, Davide Bizjak and Lorenzo Fattori

Working Students and job market outcomes: Insights from the University of Florence 532
Gabriele Lombardi, Valentina Tocchioni and Alessandra Petrucci

Bayesian methods and applications 1

Analyzing RNA data with scVelo: identifiability issues and a Bayesian implementation 538
Elena Sabbioni, Enrico Bibbona, Gianluca Mastrantonio and Guido Sanguinetti

Approximate Bayesian Computation for Probabilistic Damage Identification 544
Cecilia Viscardi, Silvia Monchetti, Luisa Collodi, Gianni Bartoli, Michele Betti, Michele Boreale and Fabio Corradi

Estimation of scientific productivity with a hierarchical Bayesian model 550
Maura Mezzetti and Ilia Negri

Heat waves and free-knots splines 555
Gioia Di Credico and Francesco Pauli

The Hierarchical Beta-Bernoulli Process as Out-of-Scope Query Detector 560
Marco Dalla Pria and Silvia Montagna

Health and mortality

A novel definition of comorbidity based on the Global Burden of Diseases project weights 566
Angela Andreella, Lorenzo Monasta and Stefano Campostrini

An Age-Period-Cohort model of gender gap in youth mortality 572
Giacomo Lanfiuti Baldi and Andrea Nigri

Kinlessness in adult and old age across Europe 578
Marta Pittavino, Bruno Arpino and Elena Pirani

Parameter orthogonalization for Siler mortality model 584
Claudia Di Caterina and Lucia Zanotto

Pseudo-observations in survival analysis 590
Marta Cipriani, Alfonso Piciocchi, Valentina Arena and Marco Alfò

Sex Gap in Cancer-Free Life Expectancy: The Association with Smoking, Obesity and Physical Inactivity 595
Alessandro Feraldi, Cristina Giudici and Nicolas Brouard

Women's Exposure to HIV in Africa: the Role of Intimate Partner Violence 599
Micaela Arcaio and Anna Maria Parroco

Revealing the dynamic relations between traffic and crowding using big data from mobile phone network 691
Selene Perazzini, Rodolfo Metulini and Maurizio Carpita

SMaC: Spatial Matrix Completion method 697
Giulio Grossi, Alessandra Mattei and Georgia Papadogeorgou

The impact of traffic flow and road signs on road accidents: an approach based on spatiotemporal point pattern analysis on linear networks 702
Andrea Gilardi and Riccardo Borgoni

Clustering and classification 1

A clustering model for flow data: an application to international student mobility 708
Cinzia Di Nuzzo and Donatella Vicari

Contingency tables with structural zeros and discrete copulas 713
Roberto Fontana, Elisa Perrone and Fabio Rapallo

Levels Merging in the Latent Class Model 719
Christophe Biernacki

Model-based clustering of count processes with multiple change 725
Shuchismita Sarkar and Xuwen Zhu

Similarity Measures and Internal Evaluation Criteria in Hierarchical Clustering of Categorical Data 729
Jana Cibulková, Zdeněk Šulc, Hana Řezanková and Jaroslav Horníček

Spectral clustering of mixed data via association-based distance 735
Alfonso Iodice D'Enza, Francesco Palumbo and Cristina Tortora

Dynamic models and time series

A graph based convolution Neural Network approach for forecast reconciliation 741
Andrea Marcocchia and Pierpaolo Brutti

A multivariate hidden semi-Markov model for the analysis of multiple air pollutants 747
Marco Mingione, Pierfrancesco Alaimo Di Loro, Francesco Lagona and Antonello Maruotti

A smooth transition autoregressive model for matrix-variate time series 753
Andrea Bucci

Dynamic network models with time-varying nodes 759
Luca Gherardini, Mauro Bernardi and Monia Lupparelli

Time lapse analysis of nuclear calcium spiking in plant cells during symbiotic signaling 765
Ivan Sciascia, Andrea Crosino and Andrea Genre

Two-stage weighted least squares estimator of multivariate conditional mean observation-driven time series models 770
Mirko Armillotta

Environmental learning and indicators

Assessing the performance of nuclear norm-based matrix completion methods on CO_2 emissions data 776

Rodolfo Metulini, Francesco Biancalani, Giorgio Gnecco and Massimo Riccaboni

Deep Learning for smart and sustainable agriculture 782

Amalia Vanacore, Armando Ciardiello, Annalisa Izzo, Pierdomenico Zaffino, Carolina Vecchio, Gennaro Pio Auricchio and Luigi Uccelli

Do green transition, environmental taxes and renew-able energy promote ecological sustainability in G7 countries? Evidence from panel quantile regression 788

Aamir Javed, Agnese Rapposelli and Asif Javed

Doubly Robust DID for National Parks evaluation: "just" environmental benefits, or socioeconomics impacts as well? 795

Riccardo D'Alberto, Francesco Pagliacci and Matteo Zavalloni

On the gap between emitted and absorbed carbon dioxide. Are trees enough to save us? 801

Lorenzo Mori and Maria Rosaria Ferrante

Small scale analysis of energy vulnerability in the municipality of Palermo 806

Giuliana La Mantia

Health statistics 2

A test for non-differential misclassification error in database epidemiological studies 812

Giorgio Limoncella, Leonardo Grilli, Emanuela Dreassi, Carla Rampichini, Robert Platt and Rosa Gini

Is the COVID-19 'color code' of Italian regions subjected to political manipulation? 816

Giovanni Busetta and Fabio Fiorillo

Modelling multilevel ordinal response under endogeneity: application to DTC patients' outcome 822

Silvia D'Elia

Monitoring drugs-based diagnostic therapeutic paths in heart failure patients using state-sequence analysis techniques 827

Nicole Fontana, Laura Savaré and Francesca Ieva

Optimal two-stage design based on error rates under a Bayesian perspective 833

Susanna Gentile and Valeria Sambucini

Migrants in Italy and return migration

Comparing migrant and "native" Italian adolescents in risky behaviours from FSS and SHARE Corona surveys n.a.

Daniela Foresta

EU-Border crisis on Twitter: sentiments and misinformation analysis 839

Elena Ambrosetti, Cecilia Fortunato and Sara Miccoli

Graduates' interregional migration in times of crisis: the Italian case 843
Thaís García-Pereiro, Ivano Dileo and Anna Paterno

Intentions to stay: The experience of return migrants in Albania 848
Maria Carella, Thaís García-Pereiro, Roberta Pace and Anna Paterno

Return migration to home country: a systematic literature review with text mining and topic modelling 853
Cecilia Fortunato, Andrea Iacobucci and Elena Ambrosetti

The allocation of time within native and foreign couples living in Italy 860
Giovanni Busetta, Maria Gabriella Campolo and Antonino Di Pino Incognito

Eἰλείθυια comes from afar: The foreigners’ contribution to fertility by Italian provinces 866
Eleonora Miaci, Cristina Giudici, Eleonora Trappolini, Marina Attili, Cinzia Castagnaro and Antonella Guarneri

Sustainability assessment

ESG, sustainability and stock market risk 871
Michele Costa

Exploring the effect of consumer motivation and perception of sustainability on food choices with a Discrete Choice Experiment 875
Gloria Solano-Hermosilla, Jesus Barreiro-Hurle and Ilaria Amerise

Sustainability explained by ChatGPT artificial intelligence in a HITL perspective: innovative approaches 881
Vito Santarcangelo, Angelo Lamacchia, Emilio Massa, Saverio Gianluca Crisafulli, Massimiliano Giacalone and Vincenzo Basile

Measuring economic and ecological efficiency of urban waste systems in Italy: a comparison of SFA and DEA techniques 887
Massimo Gastaldi, Ginevra Virginia Lombardi, Agnese Rapposelli and Giulia Romano

Profile based latent distance association analysis for sparse tables. Application to the attitude of EU citizens towards sustainable tourism 893
Francesca Bassi, Josè Fernando Vera and Juan Antonio Marmolejo Martìn

Sustainable tourism: a survey on the propensity towards eco-friendly accommodations 899
Claudia Furlan and Giovanni Finocchiaro

Bayesian methods and applications 2

A comparison of computational approaches for posterior inference in Bayesian Poisson regression 903
Laura D’Angelo

Bias-reduction methods for Poisson regression models 908
Luca Presicce, Tommaso Rigon and Emanuele Aliverti

Finite Mixture Model for Multiple Sample Data 913
Alessandro Colombi, Raffaele Argiento, Federico Camerlenghi and Lucia Paci

On Bayesian power analysis in reliability 918
Fulvio De Santis, Stefania Gubbiotti and Francesco Mariani

Power priors elicitation through Bayes factors 923
Roberto Macrì Demartino, Leonardo Egidi and Nicola Torelli

Predictive Bayes factors 929
Leonardo Egidi and Ioannis Ntzoufras

Clustering and classification 2

A Clusterwise Regression Method for Distributional-Valued Data 935
Antonio Balzanella, Rosanna Verde and Francisco de A.T. de Carvalho

A novel statistical-significance based semi-parametric GLMM for clustering countries standing on their innumeracy levels 939
Alessandra Ragni, Chiara Masci, Francesca Ieva and Anna Maria Paganoni

Introducing a novel directional distribution depth function for supervised classification 945
Edoardo Redivo and Cinzia Viroli

Clustering alternatives in the preference-approval context 950
Alessandro Albano, José Luis Garcia-Lapresta , Mariangela Sciandra and Antonella Plaia

Computational assessment of k-means clustering on a Structural Equation Model based index 955
Mariaelena Bottazzi Schenone, Elena Grimaccia and Maurizio Vichi

Handling missing data in complex phenomena: an ultrametric model-based approach for clustering 961
Francesca Greselin and Giorgia Zaccaria

Economics and labour markets

A multivariate ranking analysis on the employability of young adults 967
Rosa Arboretti, Elena Barzizza, Nicolo Biasetton, Riccardo Ceccato, Monica Fedeli and Concetta Tino

Analysis of the Gender Pay Gap in the Italian Labour Market 973
Giulia Cappelletti and Daniele Toninelli

Evaluating the effect of home-based working employing causal Bayesian networks and potential outcomes 979
Lorenzo Giammei

Patterns of flexible employment careers. Does measurement error matter? 985
Mauricio Garnier-Villarreal, Dimitris Pavlopoulos and Roberta Varriale

Staying or leaving? A nonlinear framework to explore the role of employee well-being on retention 991
Ulpiani Kocollari, Fabio Demaria and Maddalena Cavicchioli

The CAP instruments impact on GVA and employment: a multivalued treatment approach 997
Montezuma Dumangane and Marzia Freo

The determinants of leaving the parental home in Italy: 2012-18 1003
Ilaria Rocco and Gianpiero Dalla Zuanna

Environmental modeling

A Bayesian weather-driven spatio-temporal model for PM10 in Lombardy 1109
Michela Frigeri, Alessandra Guglielmi and Giovanni Lonati

A preliminary study on shape descriptors for the characterization of microplastics ingested by fish 1015
Greta Panunzi, Tommaso Valente, Marco Matiddi and Giovanna Jona Lasinio

Artificial neural network in predicting odour concentrations: a case study 1021
Veronica Distefano and Gideon Mazuruse

Bayesian analysis of PM10 concentration by spatio-temporal ARIMA and STS models 1026
Michela Frigeri and Ilenia Epifani

Functional ANOVA to monitor yearly Adriatic sea temperature variations 1032
Annalina Sarra, Adelia Evangelista, Tonio Di Battista and Nicola Di Deo

New perspectives in the measurement of biodiversity 1038
Linda Altieri, Daniela Cocchi and Massimo Ventrucci

Multivariate data analysis 2

Feature Selection via anomaly detection autoencoders in radiogenomics studies 1044
Alessia Mapelli, Michela Carlotta Massi, Nicola Rares Franco, Francesca Ieva, Catharine West, Petra Seibold, Jenny Chang-Claude and the REQUITE and RADprecise Consortia

Further considerations on the Spectral Information Criterion 1050
Luca Martino

How to increase the power of the test in sparse contingency tables: a simulation study 1057
Federica Nicolussi and Manuela Cazzaro

Latent event history models for quasi-reaction systems 1063
Matteo Framba, VeronicaVinciotti and Ernst Wit

Quantile-based graphical models for continuous and discrete variables 1069
Luca Merlo, Marco Geraci and Lea Petrella

The logratio Student t distribution 1075
Gianna Monti and Gloria Mateu-Figueras

Statistics in Society 2

A decomposition of the changes in tourism demand in Tuscany over the 2019-2021 period 1079
Mauro Mussini

Bayesian networks as a territorial gender impact assessment tool 1084
Flaminia Musella, Lorenzo Giammei, Fulvia Mecatti and Paola Vicar

Can statistics be helpful in detecting electoral fraud? 1088
Massimo Attanasio, Vincenzo G. Genova and Michele Tumminello

Companies' sustainability disclosure and contrast to hunger: the role of social inclusion 1093
Chiara Di Maria and Rodolfo Damiano

Passing network-based performance indicator in football: evidence from UEFA Champions League 2016-2017 1099
Riccardo Ievoli, Lucio Palazzo and Giancarlo Ragozini

Topic Modeling for the travel and tourism industry: classical and innovative methods compared 1105
Fabrizio Di Mari

Bayesian methods and applications 3

An Importance Sampling Algorithm For Bayesian Logistic Regression with Independent Gaussian Scale Mixture Prior 1111
Paolo Onorati and Brunero Liseo

Bayesian analysis of Amazon's best-selling books via finite nested mixture model 1117
Laura D'Angelo and Francesco Denti

Binomial Extended Stochastic Block Model for Brain Networks 1121
Valentina Ghidini, Sirio Legramanti and Raffaele Argiento

Detecting latent spatial patterns in mass spectrometry brain imaging data via Bayesian mixtures 1127
Giulia Capitoli, Simone Colombara, Alessia Cotroneo, Francesco De Caro, Riccardo Morandi, Chiara Schembri, Alfredo G. Zapiola and Francesco Denti

Efficient expectation propagation for high-dimensional probit models 1133
Augusto Fasano, Niccolo Anceschi, Beatrice Franzolini and Giovanni Rebaudo

Model-based clustering of non-stationary time series with common historical change times 1139
Riccardo Corradin, Luca Danese, Wasiur KhudaBukhsh and Andrea Ongaro

Functional Data Analysis

A functional Ground Motion Model for Italy built with a weighted analysis of reconstructed seismic curves 1145
Teresa Bortolotti, Riccardo Peli, Giovanni Lanzano, Sara Sgobba and Alessandra Menafoglio

Conditional Gaussian Graphical Models for Functional Variables whit Partial Separable Operators 1149
Rita Fici, Gianluca Sottile and Luigi Augugliaro

Does the Inflation Factor need tuning? Simulation-based adjustment for Outlier Detection via the Functional Boxplot 1155
Annachiara Rossi, Andrea Cappozzo and Francesca Ieva

Functional Graphical Models to map Brexit debate on Twitter 1160
Nicola Pronello, Emiliano del Gobbo, Lara Fontanella, Rosaria Ignaccolo, Luigi Ippoliti and Sara Fontanella

Measuring Dependence in Multivariate Functional Datasets 1166
Francesca Ieva, Michael Ronzulli and Anna Maria Paganoni

Robust Statistical Process Monitoring of Multivariate Functional Data 1173
Christian Capezza, Fabio Centofanti, Antonio Lepore and Biagio Palumbo

The effects of mobility restrictions on public health: a functional data analysis for Italy over the years 2020 and 2021 1179
Veronica Mazzola, Giovanni Bonaccorsi, Piercesare Secchi and Francesca Ieva

Machine Learning and text mining

A vocabulary-based approach for risk detection in textual annotations of contracts of public procurement 1185
Giulio Giacomo Cantone, Simone Del Sarto and Michela Gnaldi

Explainable Machine Learning based on Group Equivariant Non-Expansive Operators (GENEOs). Protein pocket detection: a case study 1191
Giovanni Bocchi, Alessandra Micheletti, Patrizio Frosini, Alessandro Pedretti, Andrea R. Beccari, Filippo Lunghini, Carmine Talarico and Carmen Gratteri

Hedging global currency risk with factorial machine learning models 1197
Paolo Pagnottoni and Alessandro Spelta

InstanceSHAP: An instance-based estimation approach for Shapley values 1203
Golnoosh Babaei and Paolo Giudici

Networks & Nature Based Solutions: an application for Milan hydric resources 1209
Alessia Forciniti and Emma Zavarrone

The Roe v. Wade sentence: an analysis of tweets trough Symmetric Non-Negative Matrix Factorization 1215
Maria Gabriella Grassia, Marina Marino, Rocco Mazza and Agostino Stavolo

Multivariate data analysis 3

A comparison of different techniques for handling missing covariate values in propensity score methods 1219
Anna Zanovello, Alessandra R. Brazzale and Omar Paccagnella

A New Penalized Estimator for Sparse Inference in Gaussian Graphical Models: An Adaptive Non-Convex Approach 1224
Daniele Cuntrera, Vito M.R. Muggeo and Luigi Augugliaro

A tool for assessing weak identifiability of statistical models 1230
Antonio Di Noia, Francesco Denti and Antonietta Mira

Computing Highest Density Regions with Copulae 1235
Nina Deliu and Brunero Liseo

Parameter estimation via Indirect Inference for multivariate Wrapped Normal distributions 1241
Francesca Labanca and Anna Gottard

Sequential marginal likelihood selection for the estimation of sparse correlation matrices 1246
Claudia Di Caterina and Davide Ferrari

Nonparametric statistical methods

A Comparison of Distribution-Free Control Charts 1252
Michele Scagliarini

Characterizing Heterogeneity of Causal Effects in Air Pollution in Florida 1257
Dafne Zorzetto

Comparing three robust procedures for CANDECOMP/PARAFAC estimation 1262
Valentin Todorov, Violetta Simonacci, Michele Gallo and Nikolay Trendafilov

How active is a genetic pathway? Comparative analysis of post-hoc permutation-based methods 1268
Anna Vesely and Angela Andreella

Non Parametric Combination methodology: a literature review on recent developments 1274
Elena Barzizza, Nicolò Biasetton and Riccardo Ceccato

Regression modeling

A Quantile Regression Model to Evaluate the Performance of the Italian Courts of Law 1280
Carlo Cusatelli, Massimiliano Giacalone and Eugenia Nissi

A variable selection procedure based on predictive ability: a preliminary study on logistic regression 1285
Rosaria Simone and Mariarosaria Coppola

Comparison of binary regressions with asymmetric link function for imbalanced data 1291
Michele La Rocca, Marcella Niglio and Marialuisa Restaino

New advances in Regression Forests 1297
Mila Andreani, Lea Petrella and Nicola Salvati

On the Optimal Non-Convexity of Penalty in Sparse Regression Models 1303
Daniele Cuntrera, Vito M.R. Muggeo and Luigi Augugliaro

Using expectile regression with latent variables for digital assets 1309
Beatrice Foroni, Luca Merlo and Lea Petrella

4 Program 1315

Preface

This book includes the contributions presented at the Intermediate Meeting of the Italian Statistical Society (SIS) "SIS 2023 - Statistical Learning, Sustanaibility and Impact Evolution" held in Ancona at the Università Politecnica delle Marche, from June 21th to 23th of 2023.

The new challenges of digitalization, innovation and sustainability are showing the crucial role of data-driven approaches in supporting decision-making processes. Methodologies resulting from the integration of different know-how seem to be a reliable way to deal with the increasing need to measure the impact of the policies and to forecast scenarios. This meeting welcomed any attempt to face new challenges.

The conference registered more than 250 presentations, including 3 keynote speakers in 3 plenary sessions and 72 presentations in 24 invited sessions, all dealing with specific themes in methodological and/or applied statistics and demography. Furthermore, more than 180 contributions, with one or more authors, have been spontaneously submitted to the Program Committee and arranged in 30 contributed sessions.

The numerous participation of researchers in the conference shows how the challenges of sustainability, in its broadest sense, are of interest to both methodological and applied statistics.

With the publication of this book, we wish to offer to all members of the Italian Statistical Society, all international academics, researchers, Ph.D. students, and all interested practitioners, a good snapshot of the on-going research in the statistical and demographic fields.
We aim to provide all members of the Italian Statistical Society - as well as international academics, researchers, Ph.D. students, and interested practitioners - with a comprehensive overview of the ongoing research in the fields of statistics and demography.

We extend our heartfelt gratitude to all the contributors for submitting their works to the conference and to the researchers for their outstanding job in serving as referees and discussants with precision and timeliness.

A special appreciation goes to the Scientific and Organizational Committees for their tremendous efforts in managing all the organizational aspects, as well as to the Università Politecnica delle Marche and the Department of Economic and Social Science for making this event possible.

Finally, we wish to express our gratitude to the publisher Pearson Italia for all the support received.

Enhancing Principal Components by a Linear Predictor: an Application to Well-Being Italian Data

Laura Marcis[a], Maria Chiara Pagliarella[a], and Renato Salvatore[a]

[a]University of Cassino and Southern Lazio (Italy); laura.marcis@unicas.it, mc.pagliarella@unicas.it, rsalvatore@unicas.it

Abstract

We consider the case of a multivariate random vector that obeys a linear mixed model when the vector itself lies in a lower dimensional subspace. This situation suggests that this subspace can be modeled by the probabilistic (random-effects) principal components. By reason of this, the random vector follows at the same time two different models. We employ a linear predictor adjusted by the residual part of the probabilistic principal components that results not explained by the linear model. The new predictor can be considered as the vector of scores that comes from that principal components, enhanced by the linear mixed model. The application to the official Italian well-being data shows some features of the method.

Keywords: multivariate random vector, principal components, linear mixed model, well-being

1. Introduction

Principal component analysis (PCA) is recognized as one of the most employed methods to reduce dimensionality, by means of the projection of a set of variables in a subspace of them. By summarizing and allowing to visualize data, and, at the same time, minimizing the loss of information in the lower dimensional space, in many cases principal components (PCs) lead to a better assessment of the bundled statistical information, seized by the original variables [2]. Because PCs are linear combinations, the interpretation of the scores by these new "data-dependent" variables is hard to give some time. In particular situations, the contribution improve understanding of some case studies may be poor and may lead to misguided or unclear findings. Furthermore, in the great majority of cases, when the interpretation stills on loadings that exceed a threshold, the linking of the variables hardly fails to provide some explanation or a bit of insight. Because of using of the common practice of ignoring the PCs affected by lower loadings, the focus shifts to the first PCs, which arise from the most correlated variables in the original set. The issue of resting on the main linearly-dependent variables is particularly relevant and becomes crucial in several instances. We may come across the trade-off between considering the retained PCs as highlighting latent phenomena, or in reproducing similar information unnecessarily. One of the recurrent ways of approaching redundancy and, in general, the recursive informative content of multivariate sample data, is given by considering a common subset of covariates that the population obeys. Two main cases in the literature are deemed representative of the joint dependence on a multivariate vector, the PCs "with covariates" or Partial PCA, and the redundancy analysis. Given a subspace spanned by the sample vectors of predictor variables, both of them rely on the common baseline of splitting the sample variance as the sum of the variance "explained" by a multivariate linear regression model, and the variance due to the regression residuals. Although these last represent a very useful tool in some cases, the deployment of linear models to explain part of the sample variability has had a remarkable development in the last years. One of these studies brings into play the role of prediction by linear statistical models.

Tipping and Bishop [7] had already introduced the notion of prediction for PCs. They called "Probabilistic PCA" (PPCA) the model behind the PCA, which parameters are estimated with the Expectation - Maximization algorithm. The "noisy" PC model (nPC), proposed by Ulfarsson and Solo [8] has a quite similar formulation with respect to the PPC model, providing - in a similar way - the nPC prediction after giving the model estimates. Instead of a "fixed effects PCs", as the traditional linear regression PCA model, the PPC (or nPC) are random variables. This condition suggests, on the one hand, the Bayesian approach to handle the estimates for the PPC linear model and, on the other hand, to predict PCs under its meaning within random linear models theory [4]. The Bayesian approach to the estimation requires an expectation of some model parameters that are random, conditional to the observed data. Given normality of the error $\varepsilon \sim N(0, \sigma^2 I)$, for a linear model $\tau = B\lambda + \varepsilon$ - in case of the vector λ random - the likelihood is based on the conditional distribution $\lambda|\tau \sim N[E(\lambda|\tau), var(\lambda|\tau)]$.

Moreover, it is known [6] that $E(\lambda|\tau) = \widetilde{\lambda}$ is the "Best Prediction" (BP) estimate, with $var(\widetilde{\lambda} - \lambda) = E_\tau[var(\lambda|\tau)]$. This is somewhat different from the standard linear regression model prediction by $E(\tau|\lambda)$. Therefore, given a linear mixed model (LMM) [1] for τ, with $E(\tau|\lambda) = \lambda$, the model parameters are the realizations of random variables. The BP of a linear combination of the LMM fixed and random effects (i.e. linear in τ, with $E[E(\tau|\lambda)] = 0$) gives the "Best Linear Unbiased Prediction" (BLUP) estimates [5]. LMM's are particularly suitable for modeling with covariates (fixed and random) and for specifying model covariance structures [1]. They allow researchers to take in account special data, such as hierarchical, time-dependent, correlated, covariance-patterned models. Thus, given the BP estimates of the nPC λ, $\widetilde{\lambda} = E(\lambda|\tau)$, the vector $\widetilde{\tau} = B\widetilde{\lambda}$ represents the BP of the p-variate vector.

In the present paper, we introduce a multivariate LMM that considers the dependent random vector effectively represented by the subspace of the PPCs. The new predictor combines the linear model and the PPC's, carrying simultaneously the Best Linear Predictor, and the contribution given by the PPCs not "explained" by the linear predictor itself. An application to the official Well-being Italian indicators shows some of the features of the method.

2. Theory

In the sequel we report the following symbols, giving the model specification.

- $n =$ the number of subjects in the LMM model $(i = 1, ..., n)$;
- $N = \Sigma n_i =$ total sampling units considered;
- $p =$ the number of the response dependent variables;
- $l =$ the number of the linear model covariates;
- $j = 1, ..., n_i$ the within-subject (groups) units;
- $s =$ the dimension of the effective PC subspace.

Consider Θ as the $N \times p$ sample matrix of the p-variate $p \times 1$ random vector θ, with N as the total number of the units given by the sample. Moreover, consider that the vector θ obeys the linear model:

$$\theta = \beta' x + u', \tag{1}$$

where x is the $l \times 1$ vector of covariates, β is the $l \times p$ matrix of the regression effects, u is the vector of the p-variate random effect, with $u \sim N(0, \Sigma_u)$, $\Sigma_u = cov(u)$. Furthermore, we consider at the same time that the multivariate random vector θ obeys the following linear model:

$$\theta = Ab + \epsilon, \tag{2}$$

in which A is $p \times s$ a loading matrix of eigenvectors, b is the random vector of PPCs, and ϵ is a vector of isotropic error, with $\theta \sim N(\mu, A\Psi A' + \sigma_\epsilon^2 I)$, $b \sim N(0, \Psi)$, $\Psi = diag(\psi_1, \ldots, \psi_s)$, $s < p$, and $\epsilon \sim N(0, \sigma_\epsilon^2 I)$. When a sample of N observations is given, an $N \times p$ matrix Y of observations from the random vector θ is simply modeled as $Y = \Theta + E$, with the "sampling error" $Np \times Np$ covariance matrix

$cov(vec(E)) = (\Sigma_e) \otimes I_N$ ($\otimes$ is the Kronecker product), $e \sim N(0, \Sigma_e)$, $\Sigma_e = var(e)$. Thus, models (1) and (2) are rewritten as $Y = \Theta + E = X\beta + ZU + E$, with the (1) that becomes $\theta = \beta' x + u' + e'$, and $Y = \Theta + E = BA' + \Xi + E = BA' + \Gamma$, with the (2) is $\theta = Ab + \epsilon + e = Ab + \gamma$, respectively. The model errors u, ϵ, and e, are mutually independent. The matrix Z represents the $N \times n$ design matrix of random effects and E is the $N \times p$ matrix of the residual errors of the multivariate LMM, B is the $N \times s$ matrix of the PPCs that lie in the s-dimensional subspace, Ξ is the $N \times p$ matrix of the isotropic errors of the model (2). The models (1) and (2) have the following conditional expectations and variances:

$$\begin{aligned} E(\theta|y) &= \widetilde{\theta}_y = y - E(e|y) = y - cov(e,y)var(y)^{-1}y = y - var(e)Py, \\ var(\theta|y) &= var(\theta) - cov(\theta,y)var(y)^{-1}cov(y,\theta) \\ &= var(e) - var(e)Pvar(e), \end{aligned}$$

for the model in (1), where $P = \Sigma_y^{-1}(I - P_X)$, $\Sigma_y = var(y)$, and P_X is the projection matrix. For the model in (2):

$$\begin{aligned} E(b|\theta) &= E(b) + cov(b,\theta)var(\theta)^{-1}(\theta - \mu) \\ &= cov(b, \mu + Ab + \epsilon)C^{-1}(\theta - \mu) = \Psi A' C^{-1}(\theta - \mu), \\ var(b|\theta) &= var(b) - cov(b,\theta)var(\theta)^{-1}[cov(b,\theta)]' \\ &= \Psi - \Psi A' C^{-1} A \Psi \\ C &= A\Psi A' + \sigma_\epsilon^2 I. \end{aligned}$$

Based on some results on linear projections, i.e., given the random variable y, and the $1 \times j$, $1 \times k$ random vectors x, z, with positive definite covariance matrix of $(y, x, z)'$, then for the linear projection $L(y|x,z)$:

$$L(y|x,z) = L(y|x) + [z - L(z|x)]\gamma,$$

where $\gamma = var(z|x)^{-1}[cov(y,z|x)]'$, we get the following:

Proposition 1. *Given the model (2) for the p-dimensional random vector θ, with $b = \overline{F}'(\theta - \epsilon)$, and under the models in (1) and (2), the multivariate Best Predictor based on (y, b), $E(\theta|y,b)$, is:*

$$E(\theta|y,b) = \widetilde{\theta}_{y,b} = E(\theta|y) + cov(\theta,b|y)var(b|y)^{-1}\left\{\widetilde{b} - E(b|y)\right\}, \tag{3}$$

with $\widetilde{b} = E(b|\theta)$, $\overline{F}$ the $sN \times pN$ matrix $(\overline{A}'\overline{A})^{-1}\overline{A}$, and $\overline{A}$ is the $pN \times sN$ matrix $A \otimes I_N$. Then, $var(\theta|y,b) = var(\theta|y) - cov(\theta,b|y)var(b|y)^{-1}[cov(\theta,b|y)]'$.

The "hybrid" predictor $\widetilde{\theta}_{y,b}$ in (3) gives the Best Linear Unbiased Predictor $E(\theta|y)$, "embedding" the PPCs through an adjoint component. The last is due to knowing that the random vector θ lies in the s-dimensional subspace of the PPCs. In particular, the difference $\widetilde{b} - E(b|y)$ gives the multivariate vector of the PPCs "not explained" by the estimation of the linear model $E(\theta|y)$. The matrix $var(\theta|y,b)$ has rank s, and, consequently, there are $(p - s)$ linear combinations of θ for which their respective variances do not depend on the PPCs.

Proposition 2. *Given the p-dimensional random vector θ, under the models in (1) and (2), and the Best Predictor $E(\theta|y,b)$ in (3), we get:*

$$\begin{aligned} \overline{F}E(\theta|y,b) &= \overline{F}\widetilde{\theta}_{y,b} = \widetilde{b}^* \\ &= \overline{F}E(\theta|y) + \overline{F}cov(\theta,b|y)var(b|y)^{-1}\left\{\widetilde{b} - E(b|y)\right\}, \end{aligned} \tag{4}$$

where $\widetilde{b}^*$ is the s-dimensional vector of the PPCs "enhanced" (ePCs) by the linear predictor $E(\theta|y)$. As a particular case, when $\sigma_\epsilon^2 \longrightarrow 0$, $var(\epsilon) \longrightarrow 0$, $var(\gamma) \longrightarrow var(e)$, and $\widetilde{b} \longrightarrow \bar{b}$. Therefore, $\overline{F}\widetilde{\theta}_{y,b} = \widetilde{b}^* \longrightarrow \bar{b}$ with $\bar{b}$ the sample PCs $\bar{b} = A'\theta$.

The ePCs $\widetilde{b}^*$ are then the PPCs "adjusted" by the Linear BP $E(\theta|y)$, and $\widetilde{b}^*$ is then the vector of the ePC scores. Note that the ePC scores give a non-orthogonal matrix. Given $\sigma_\epsilon^2 = 0$, the vector θ in the model (2) lies in the s-dimensional subspace of the sample PCs $\bar{b}$. In fact, in this case $\widetilde{b} \equiv \bar{b}$, $cov(\theta,b|y) = var(\theta|y)\overline{F}'$, $var(b|y) = \overline{F}var(\theta|y)\overline{F}'$, $E(b|y) = \overline{F}E(\theta|y)$, and then $\overline{F}E(\theta|y,b) = \bar{b}$.

3. Application

In accordance with the recent law reforms in Italy, the Equitable and Sustainable Well-being indicators (in Italian, BES) [3] - annually provided by the Italian Statistical Institute (ISTAT) - are designed to define the economic policies which largely act on some fundamental aspects of the quality of life. In order to highlight the result of the proposed method we use 12 BES indicators relating to the years 2013-2016, collected at NUTS-2 (Nomenclature of Territorial Units for Statistics 2 level). The variables employed in the application study are in Table 1. We use the per capita adjusted disposable income variable (its logarithm, as is usually done in economics studies) - indicated with BE1 - as a unique covariate in the LMM model, while the remaining 11 variables are dependent variables (Table 1 reports the description and acronyms used for the variables). The application uses the Restricted Maximum Likelihood estimation, a Sas/IML code, and a sequence of Sas-HPMixed procedures. Table 2 shows the slope parameter estimates from the multivariate regression, with their significance level. Table 3 reports the MANOVA multivariate test statistics, based on the characteristic roots. These are the eigenvalues of the product of the sum-of-squares matrix of the regression model and the sum-of-squares matrix of the regression error. The null hypothesis for each of these tests is the same: the independent variable (LBE1) has no effect on any of the dependent variables. The four tests are all significant.

Figure 1 shows the application of Proposition 1, where all the measures are plotted in the space of the sample PCs. This plot reports simultaneously the factorial coordinates of the original variables, of the linear predictor $E(\theta|y)$, and of the hybrid predictor $\widetilde{\theta}_{y,b} = E(\theta|y, b)$. The dependent criterion variables in the application can be split into two main groups, starting from both an analysis of the plot and the correlation matrix between the sample PCs, the LMM predicted values, and the hybrid predictor values. Moreover, we have eight dependent variables for which there is accordance in terms of their mutual correlation inside the original variables, as well as the LMM predicted and the hybrid predictor. This means that the hybrid predictor $\widetilde{\theta}_{y,b}$ does not change significantly the mutual correlations, substantially because the component of the PPCs not explained by the linear predictor $E(\theta|y)$ is relatively small. For the remaining three variables - with the acronyms REL4, Q2, and BS3 - the correlation changes: in some cases it changes sign, going from positive correlation values by the sampling and predicted values, to negative correlation values between the predictor $\widetilde{\theta}_{y,b}$, and vice versa. Therefore the predictor (3) highlights the major influence of the component of the PPCs not explained by the linear predictor $E(\theta|y)$. The latter is in accordance with the sample PCs in the mutual correlations between these three criterion variables. Since the classical predictor matches the mutual correlation inside the original variables - meaning that these mutual relations in the sample are due to the disposable income (BE1), the covariate in the mixed model regression - then the different mutual correlation values of the hybrid predictor $\widetilde{\theta}_{y,b}$ can be interpreted as relationships not captured by the model. For instance, the correlation between $Q2_{E(\theta|y)}$ and $INN1_{E(\theta|y)}$ is -0.82, and becomes $+0.30$ between $Q2_{\widetilde{\theta}_{y,b}}$ and $INN1_{\widetilde{\theta}_{y,b}}$, meaning that conditionally to the model - thus looking at the correlation between the attendance of childhood services (Q2) and the percentage of R&D (INN1) in the space orthogonal to per capita income - the correlation is positive. It could be interpreted as saying that the Regions with a greater investment in R&D have a higher benefit of childhood services. In our opinion, this highlights how the "hybrid" predictor is able to grasp the relationship that actually exists between investment in research and development and the importance given to training starting from the earliest years of life, regardless of per capita income.

4. Discussion

The introduced predictor (3) can be viewed from two different perspectives. An "adjusted" linear predictor by the PPCs, and, by relation (4), as the enhanced PCs (ePCs) that modify the probabilistic PCs (PPCs) to accommodate the mixed model regression predicted values. The present work considers the probabilistic principal components like a "constraint" model, that links together the components of the multivariate random vector in a lower dimensional subspace.

While the estimation of the PPC model requests a quite simple procedure, one of the causes of concern in the estimation of the parameters of a multivariate mixed model is the number of covariance

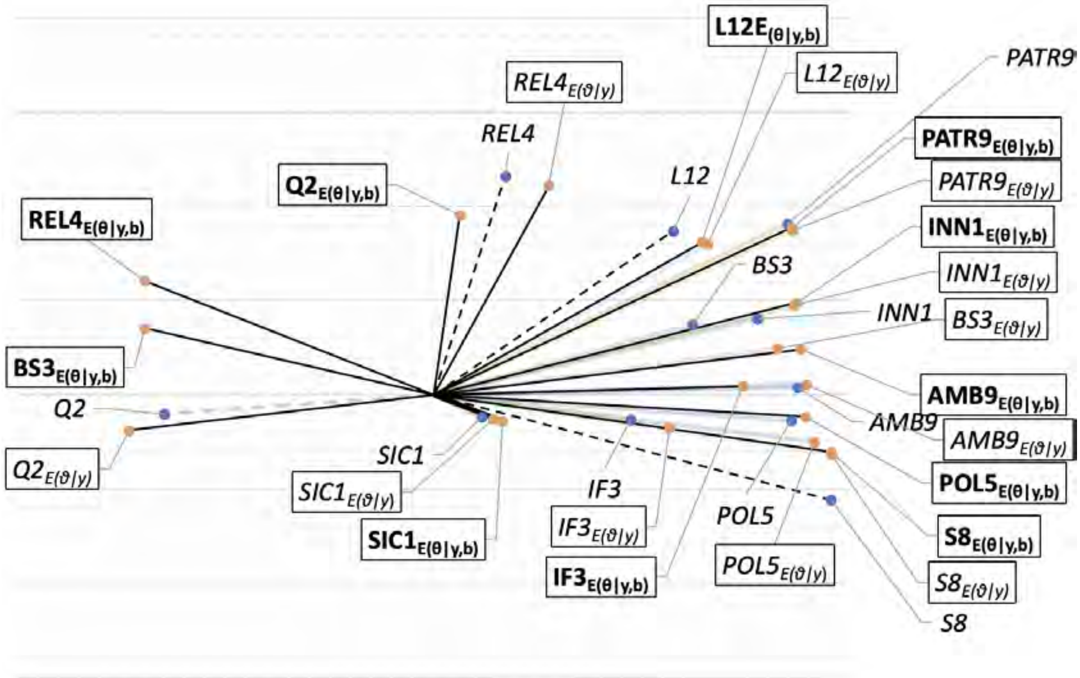

Figure 1: Plot of the application of the model (1) in the space of the sample PCs. There are represented the factorial coordinates of the original variables, of the linear predictor $E(\theta|y)$, and of the hybrid predictor $\widetilde{\theta}_{y,b} = E(\theta|y,b)$.

parameters, which may be too high for a speed software computation. Like the application presented, we suggest estimating the model covariance parameters under a uniform correlation structure among the multivariate components of the random effects. This structure is equivalent to the compound-symmetry covariance structure, with a better numerical property in terms of optimization. Indeed, some studies highlight that using uniform correlation matrices reduces the estimation noise. The model covariance matrix of random effects is then a generalized uniform correlation matrix, and works with two parameters. The "hybrid" multivariate linear predictor (3), by adjusting its standard formulation through the sample parameter vector scores in a convenient subspace, is designed to accommodate not only PPCs, but also factor models based on a random structure. In the present work, the components of the multivariate parameter among the subjects are linked by a principal components model. In order to overcome convergence problems, the method introduced can be extended to include multidimensional information from the data, through a reduced number of dependent variables in the linear model.

References

[1] Demidenko, E.: Mixed Models: Theory and Applications. Wiley, New York (2004)
[2] Hardle, W. K., Simar, L.: Principal Components Analysis. In Hardle, W. K., Simar, L. (eds.) Applied Multivariate Statistical Analysis, 319-358. Springer (2015)
[3] ISTAT: BES project www.istat.it/en/well-being-and-sustainability
[4] Longford N.T.: Random Coefficient Models. In: Lovric M. (eds.) International Encyclopedia of Statistical Science. Springer, Heidelberg (2011)
[5] McCulloch, C.E., Searle, S.R.: Generalized Linear and Mixed Models. Wiley, New York (2001)
[6] Timm, N. H.: Applied Multivariate Analysis. Springer, New York (2002)
[7] Tipping, M.E., Bishop C.M.: Probabilistic principal component analysis. J. R. Stat. Soc., Ser. B (Stat. Methodol.) **61**(3), 611–622 (1999)
[8] Ulfarsson, M.O., Solo, V.: Sparse variable PCA using geodesic steepest descent. IEEE Trans. Signal Process **56**(12), 5823–5832 (2008)

Table 1: Description of the variables used for the application

Variables	Description
S8	Age-standardised mortality rate for dementia and nervous system diseases
IF3	People having completed tertiary education (30-34 years old)
L12	Share of employed persons who feel satisfied with their work
REL4	Social participation
POL5	Trust in other institutions like the police and the fire brigade
SIC1	Homicide rate
BS3	Positive judgment for future perspectives
PATR9	Presence of Historic Parks/Gardens and other Urban Parks recognised of significant public interest
AMB9	Satisfaction for the environment - air, water, noise
INN1	Percentage of R&D expenditure on GDP
Q2	Children who benefited of early childhood services
BE1	Per capita adjusted disposable income
LBE1	Logarithm of Per capita adjusted disposable income

Table 2: The slope parameters by the multivariate regression with the LBE1 covariate

Dependent variable	Slope parameter (LBE1)	STD Error	t	Pr >t
AMB9	0.9802	0.3255	3.01	0.0035
BS3	0.9330	0.0891	10.47	0.0001
IF3	-0.3166	0.1673	-1.89	0.0621
INN1	-0.0433	0.0170	-2.54	0.0130
L12	0.0016	0.0107	0.15	0.8786
PATR9	0.0975	0.0756	1.29	0.2007
POL5	-0.0036	0.0085	-0.42	0.6775
Q2	0.2031	0.1762	1.15	0.2526
REL4	0.5602	0.1690	3.31	0.0014
S8	-0.0506	0.0293	-1.73	0.0879
SIC1	-0.0072	0.0150	-0.48	0.6314

Table 3: MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall LBE1 Effect

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks Lambda	0.0566	102.98	11	68	<.0001
Pillai's Trace	0.9434	102.98	11	68	<.0001
Hotelling-Lawley Trace	16.6590	102.98	11	68	<.0001
Roy's Largest Root	16.6590	102.98	11	68	<.0001