

Italian Statistical Society Series on Advances in Statistics

Enrico di Bella
Vincenzo Gioia
Corrado Lagazio
Susanna Zaccarin *Editors*

Statistics for Innovation III

SIS 2025, Short Papers,
Contributed Sessions 2



**Italian Statistical Society Series on
Advances in Statistics**

This book series publishes volumes of peer-reviewed short papers presented at the scientific events such as conferences, seminars and workshops organized by the Italian Statistical Society (SIS) and its sections.

The Italian Statistical Society (Società Italiana di Statistica, SIS) was established in 1939 and ranks among the institutions of particular scientific relevance. SIS aims to promote scientific activities for the development of statistical sciences and carries out this task organizing scientific meetings and conferences, and by means of publications and partnerships at a national and international level.

Enrico di Bella · Vincenzo Gioia ·
Corrado Lagazio · Susanna Zaccarin
Editors

Statistics for Innovation III

SIS 2025, Short Papers, Contributed Sessions 2

Editors

Enrico di Bella
Department of Political and International
Sciences
University of Genoa
Genoa, Italy

Vincenzo Gioia
Department of Economics, Business,
Mathematics and Statistics
University of Trieste
Trieste, Italy

Corrado Lagazio
Department of Economics
University of Genoa
Genoa, Italy

Susanna Zaccarin
Department of Economics, Business,
Mathematics and Statistics
University of Trieste
Trieste, Italy

ISSN 3059-2135

ISSN 3059-2143 (electronic)

Italian Statistical Society Series on Advances in Statistics

ISBN 978-3-031-95994-3

ISBN 978-3-031-95995-0 (eBook)

<https://doi.org/10.1007/978-3-031-95995-0>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

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

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

The promotion of scientific activities for the advancement of statistical sciences represents the defining element of the Italian Statistical Society's (SIS) actions. Since its inception, the society has grown in tandem with the rapid dissemination of quantitative methods for data analysis across various fields of scientific research and social life. Since the beginning, in the odd years the SIS promotes an International Specialized Statistical Conference aimed at discussing the state of the art on a specific topic and to provide a forum where academics, researchers, and practitioners can share experiences and discuss present and future challenges of the topic.

The 2025 SIS Conference topic is Statistics for Innovation. Statistics is fundamental to innovation, offering the tools to analyse data and uncover insights needed for informed decision-making and strategic planning. Innovations in statistical methods and applications enhance our ability to gather and process vast amounts of data, thus effectively leveraging knowledge. This leads to more informed strategies, fostering advancements and breakthroughs across diverse fields and domains. Academics, researchers, doctoral students, and professionals interested in deepening the role of statistics and data science in fostering innovation in diverse fields (such as business, industry and finance, technology, environment, health and medicine, official statistics, public policy, welfare and social issues, etc.) have shared their activities during the Conference held at the University of Genova from June 16 to June 18, 2025.

The event featured over 350 presentations, including 3 keynotes in the 3 plenary sessions, 46 contributions in the 15 specialized sessions, 68 contributions in the 17 invited sessions, and more than 240 spontaneous communications and posters, all addressing specific themes related on Statistics for Innovations. The high number of contributions and the significant participation in the conference demonstrate the desire to participate in scientific events as venues for exchange and discussion on the new developments in our field, a vibrant characteristic of our scientific community. A round table on the challenges for statistical research from artificial intelligence and the constraint of data protection was also organized, along with satellite events such as the 8th edition of the hackathon "Stats Under the Stars" during the night between June 15 and 16 and the Sustainable, Accurate, Fair and Explainable Machine Learning (SAFE-ML) meeting on June 19 in Pavia.

We would also like to thank for their presence the Royal Statistical Society (RSS), the Institute of Mathematical Statistics (IMS), the European Network for Business and Industry Statistic (ENBIS), and the Italian region of the International Biometric Society (SIB – Società Italiana di Biometria). Special thanks goes to the University of Genova for endorsing and supporting the event, and to all the sponsors for their contributions: Società Italiana di Statistica, Università degli Studi di Genova, Scuola di Scienze Sociali (Università degli Studi di Genova), Dipartimento di Scienze Politiche e Internazionali (Università degli Studi di Genova), Dipartimento di Economia (Università degli Studi di Genova), Dipartimento di Matematica (Università degli Studi di Genova), Camera

di Commercio di Genova, Fondazione CARIGE, RULEX, LegaCoop Liguria, Tatabox Officina Studio, CNA Artigiani Imprenditori d'Italia - Liguria, Demetra Opinioni.net. Furthermore, marking a first for SIS conferences, childcare services were provided throughout the event thanks to the sponsorship of Fondazione Carige. This initiative was implemented to enable colleagues with young children to fully participate in both the conference sessions and the social events.

The high number of contributions has led to the production of four volumes based on the types of sessions outlined in the scientific program. The first volume includes the plenary, specialized and solicited sessions, and the other three volumes include the spontaneous communications and posters. These volumes contain works, as usual subjected to peer review, that the authors have agreed to publish, providing an updated and comprehensive overview of innovations in statistical methods and applications.

The publication of these contributions was only possible thanks to the invaluable efforts of several groups: the invited speakers for their valuable presentations, the authors of the contributed papers, the program committee for the careful selection and structuring of the program, the local organizing committee for the logistical management, and all the reviewers for their fundamental evaluation work, which ensured the high quality of the contributions.

This third volume presents a selection of contributions from authors who delivered presentations during the contributed and poster sessions.

June 2025

Enrico di Bella
Editor and Chair of the Local Organizing
Committee, SIS 2025

Corrado Lagazio
Vincenzo Gioia
Editor and Member of the Local
Organizing Committee, SIS 2025

Susanna Zaccarin
Editor and Chair of the Scientific Program
Committee, SIS 2025

Organization

SIS 2025 is organized by the Department of Political and International Sciences, University of Genova, in cooperation with the Italian Statistical Society.

Editors

Susanna Zaccarin	Università degli Studi di Trieste
Enrico di Bella	Università degli Studi di Genova
Corrado Lagazio	Università degli Studi di Genova
Vincenzo Gioia	Università degli Studi di Trieste

Program Committee

Susanna Zaccarin (Chair)	Università degli Studi di Trieste
Giammarco Alderotti	Università degli Studi di Firenze
Viviana Amati	Università degli Studi di Milano-Bicocca
Elena Ambrosetti	Università di Roma “La Sapienza”
Pietro Amenta	Università degli Studi del Sannio di Benevento
Annamaria Bianchi	Università di Bergamo
Isadora Antoniano Villalobos	Università Ca’Foscari di Venezia
Marco Bonetti	Università Bocconi
Pierluigi Conti	Università di Roma “La Sapienza”
Aldo Corbellini	Università degli Studi di Parma
Enrico di Bella	Università degli Studi di Genova
Concetta Ferruzzi	Istituto Nazionale di Statistica
Luigi Ippoliti	Università degli Studi “G. d’Annunzio” di Chieti-Pescara
Paolo Mariani	Università degli Studi di Milano-Bicocca
Marina Marino	Università di Napoli Federico II
Marco Mingione	Università di Roma Tre
Antonella Plaia	Università degli Studi di Palermo
Venera Tomaselli	Università degli Studi di Catania

Local Organizing Committee

Enrico di Bella (Chair)	Università degli Studi di Genova
Giulia Bertagnolli	Università degli Studi di Genova
Casper Cogen	Università degli Studi di Genova
Paolo Giudici	Università degli Studi di Pavia
Vincenzo Gioia	Università degli Studi di Trieste
Corrado Lagazio	Università degli Studi di Genova
Marta Nai Ruscone	Università degli Studi di Genova
Francesco Porro	Università degli Studi di Genova
Sara Preti	Università degli Studi di Genova
Fabio Rapallo	Università degli Studi di Genova
Eva Riccomagno	Università degli Studi di Genova
Sara Sommariva	Università degli Studi di Genova
Davide Valenzona	Università degli Studi di Genova

Referees

G. Adelfio	P. L. Conti
L. Aiello	A. Corbellini
P. Alaimo Di Loro	A. Crisci
A. Albano	C. Crocetta
G. Alderotti	D. Cucina
E. Aliverti	A. D'Agostino
V. Amati	N. D'Angelo
E. Ambrosetti	A. D'Ambra
P. Amenta	E. di Bella
I. Antoniano Villalobos	G. Di Credico
L. Augugliaro	C. Di Maria
E. Barbiano di Belgiojoso	M. Di Zio
R. Berni	D. De Stefano
P. Berta	N. Deliu
A. Bianchi	F. Dotto
M. Bini	L. Egidi
M. Bonetti	C. Ferruzzi
G. G. Cantone	A. S. Gattone
G. Capitoli	V. Gioia
C. Castagnaro	D. Girolimetto
F. Castelletti	P. Giudici
R. Cataldo	G. Giungato
M. Cipriani	E. Grammatica
C. L. Comolli	L. Greco

R. Guetto	F. Pauli
F. Iafrate	R. Pappadà
F. Ieva	T. G. Pereiro
R. Ignaccolo	D. Pirotta
L. Ippoliti	A. Plaia
F. Lagona	S. Preti
A. Lucadamo	N. Pronello
T. Laureti	F. Santelli
R. Y. Liu	M. Sciandra
F. Maggino	C. Seghieri
P. Maranzano	G. Sindoni
F. Mariani	I. Screpante
P. Mariani	G. Sottile
M. Marino	A. Stavolo
A. Marletta	M. Stefanucci
F. Mecatti	V. Tomaselli
L. Mencarini	E. Trappolini
L. Merlo	P. Valentini
A. Mineo	M. Ventrucchi
M. Mingione	R. Verde
J. Myftiu	V. Zarulli
M. Nai Ruscone	C. Zaccardi
L. Paci	S. Zaccarin
G. Panunzi	D. Zugna

Sponsoring Institutions

Società Italiana di Statistica, Roma, Italy
Università degli Studi di Genova, Genova, Italy
Scuola di Scienze Sociali, Università degli Studi di Genova, Genova, Italy
Dipartimento di Scienze Politiche e Internazionali, Università degli Studi di Genova, Genova, Italy
Dipartimento di Economia, Università degli Studi di Genova, Genova, Italy
Dipartimento di Matematica, Università degli Studi di Genova, Genova, Italy
Camera di Commercio di Genova
Fondazione CARIGE
RULEX
LegaCoop Liguria
Tatabox Officina Studio
CNA Artigiani Imprenditori d'Italia - Liguria
Demetra Opinioni.net

Contents

Digital Preservation of Data Assets. An Archival Perspective for Official Statistics in the Era of Big Data and AI	1
<i>Barbara Corvisieri</i>	
Navigating Economic Uncertainty: The Impact of Social Micro-Cooperatives in Italy, 2012–2022	7
<i>Michele Costa and Flavio Delbono</i>	
Mixed Effect Models and Dynamic Bayesian Networks: An Application to Local BES Indicators	12
<i>Federica Cugnata, Clelia Di Serio, Silvia Salini, and Elena Siletti</i>	
Caregivers’ Mental Health: Gender Differences Across Europe	19
<i>Sabbiana Cunsolo, Elena Pirani, and Valentina Tocchioni</i>	
A Bayesian Latent Factor Model for Functional Data	25
<i>Xuanye Dai, Anna Gottard, Michele Guindani, and Marina Vannucci</i>	
Italian Fuel Prices ‘23-‘24 and the Mandatory Display of the Regional Average Fuel Price	31
<i>Riccardo D’Alberto</i>	
Importance of Life Domains: A Compositional Analysis	36
<i>Elena Dalla Chiara and Eleonora Matteazzi</i>	
High-Dimensional Gaussian Mixtures with Random Projection Based Covariance Estimates	42
<i>Silvia Dallari, Laura Anderlucci, and Angela Montanari</i>	
Mobile Application as an Innovative Tool for Data Collection in Istat’s Household Budget Survey	47
<i>Gabriella D’Ambrosio, Elena Cezza, Loredana De Gaetano, and Edoardo Raimondi</i>	
Multimorbidity Patterns and the Risk of Cardiovascular Event: Results from an Italian Population-Based Study	52
<i>Cecilia Damiano, Benedetta Marcozzi, Chiara Donfrancesco, Luigi Palmieri, Cinzia Lo Noce, Graziano Onder, and Davide Vetrano Liborio</i>	

Three-Way Clustering of RT-QuIC for Prion Disease Diagnosis	60
<i>Gabriele D'Andrea, Guido Gigante, Anna Poleggi, and Maurizio Vichi</i>	
Change Points Detection in EU Inflation Rates	67
<i>Luca Danese, Riccardo Corradin, and Andrea Ongaro</i>	
Modeling Related Survival Samples via Dependent Nonparametric Mixtures	73
<i>Laura D'Angelo, Bernardo Nipoti, and Andrea Ongaro</i>	
Tessellated Spatial Poisson Point Process Models	79
<i>Nicoletta D'Angelo</i>	
Measuring Bureaucracy: A New Survey on Citizen Sentiment to Overcome Distortions and Biases in the Perception of Services	86
<i>Maria Letizia D'Autilia and Siriana Salvi</i>	
Baseball and Finance: An Application to Atlanta Braves Stock	92
<i>Manuel Delfino</i>	
Security and Privacy By Design for Official Statistics	98
<i>Giuseppe De Marco and Alessia Abbate</i>	
Concentration Indexes from the Geometric Perspective	104
<i>Daniele De Martini</i>	
Tuning Social Performance Index On Training-Data	110
<i>Giulio D'Epifanio</i>	
A Global Evaluation of the Probability of Success of an Experiment	116
<i>Fulvio De Santis, Stefania Gubbiotti, and Francesco Mariani</i>	
Closed-Based Testing When Multiple Quantile Regressions are Fitted	122
<i>Riccardo De Santis, Anna Vesely, and Angela Andreella</i>	
Exploring Dependencies in Compositional Data with Graphical Models	128
<i>Agnese M. Di Brisco, Roberto Ascari, Anna M. Fiori, and Federica Nicolussi</i>	
Multi-state Markov-Switching Model to Detect Regional Heat Wave Patterns	135
<i>Gioia Di Credico, Vincenzo Gioia, and Francesco Pauli</i>	
Causal Mediation Analysis with Spatial Interference	141
<i>Chiara Di Maria and Giada Adelfio</i>	

Split-and-Merge Sampling Algorithm for Hamming-Mixture Models of Categorical Data	147
<i>Sara Di Marino, Filippo Galli, Raffaele Argiento, Andrea Cremaschi, and Lucia Paci</i>	
A Note on Density Estimation for Proportion Data Observed with Errors	153
<i>Marco Di Marzio, Stefania Fensore, Agnese Panzera, and Chiara Passamonti</i>	
A Candecomp/Parafac Approach for Visualizing Flows of Asymmetric Proximity Matrices	159
<i>Cinzia Di Nuzzo, Donatella Vicari, and Alessia Vitozzi</i>	
Fuzzy Spectral Clustering	165
<i>Cinzia Di Nuzzo and Giorgia Zaccaria</i>	
Estimating the Over-Consumption Risk in European Countries	171
<i>Filippo Domma and Francesca Condino</i>	
The Perception of Gender Inequalities Among Italian Citizens: An Insight from Puglia	176
<i>Angela Maria D’Uggento, Ilaria Pepe, Rosy Bianco, and Ernesto Toma</i>	
Negative Correlations in Ising Models of Credit Risk	182
<i>Chiara Emonti and Roberto Fontana</i>	
Nothing Lasts Forever – Only Item Administration: An Item Response Theory Algorithm to Shorten Tests	188
<i>Ottavia M. Epifania and Livio Finos</i>	
Modeling Natural Disaster via Unbalanced Regularized Optimal Transport	194
<i>Angelo Famà, Paolo Pagnottoni, and Alessandro Spelta</i>	
Empirical Bayes for the Ridge Penalty in Probit Models	200
<i>Augusto Fasano, Giovanni Rebaudo, and Lorenzo Rimella</i>	
Hierarchical Clustering of Multidimensional Ordinal Data	206
<i>Marco Fattore, Paolo Gotti, and Lucio De Capitani</i>	
Healthy Working Life Expectancy: Measurements and Socio-Economic Inequalities	213
<i>Alessandro Feraldi and Christian Dudel</i>	
Going Beyond Gaussian Muon Tomography	219
<i>Marta Ferrari, Alessandra R. Brazzale, and Giovanna Menardi</i>	

Evaluating the Impact of a School-Based Padel Program on Students' Well-Being and Sports Participation	225
<i>Giacomo Ferrato, Luca Mo Costabella, and Chiara Pronzato</i>	
Distance and Covariance Functions on Periodic Generalised Time-Evolving Graphs	231
<i>Tobia Filosi, Claudio Agostinelli, and Emilio Porcu</i>	
Optimizing Sentiment Classification in Non-structured Text Data Using Ensemble Method	236
<i>Najada Firza, Domenico Viola, and Giuseppe Arbia</i>	
Assessing the Quality of Services on Sustainable Accommodation Facilities Through Sentiment Analysis: A Random Forest Approach	242
<i>Najada Firza and Dante Mazzitelli</i>	
Fostering Drug Repurposing Using Observational Data by Integrating Functional Data Analysis and Causal Inference	248
<i>Nicole Fontana, Piercesare Secchi, Carlo Andrea Pivato, Emanuele Di Angelantonio, and Francesca Ieva</i>	
Spatial Quantile Random Forests for the Analysis of Upward Mobility in Texas	255
<i>Sabrina Forte, Beatrice Foroni, Luca Merlo, and Lea Petrella</i>	
University Careers and Their Determinants: A Study at the University of Padova	261
<i>Claudia Franceschini, Anna Giraldo, Silvia Meggiolaro, and Lorenza Da Re</i>	
Profiling Mixed Unions' Assortative Mating in European Countries	268
<i>Giuseppe Gabrielli, Thaís García-Pereiro, and Anna Paterno</i>	
Modeling Multiple Node-Colored Networks with Partial Exchangeability	274
<i>Francesco Gaffi</i>	
Multichannel ECG Spectral Analysis via Functional Data Methods: A Structured Approach to Dynamic Signal Dependencies	281
<i>Antonino Gagliano, Chiara Di Maria, Gianluca Sottile, Sarah Beutler-Traktovenko, Luigi Augugliaro, and Valeria Vitelli</i>	

Measuring Public Value Through Composite Indices: Critical Issues and Preliminary Proposals in the Integrated Planning (PIAO) of Italian Public Administrations	288
<i>Enrico Deidda Gagliardo, Alessandra Dentini, Riccardo Ievoli, and Matteo Mazziotta</i>	
Bayesian Blended Landmark Model for Alignment of Functional Data	294
<i>Jacopo Gardella, Alessandro Casa, Raffaele Argiento, and Alessia Pini</i>	
Sample Size Requirements to Control Bayesian Error Probabilities in Superiority Trials	300
<i>Susanna Gentile and Valeria Sambucini</i>	
Income, Lifestyles and Subjective Well-Being in Italy	307
<i>Stefano Gerosa and Daniela Lo Castro</i>	
Approximate Bayesian Model Averaging and Forecasting of Autoregressive Models	314
<i>Riccardo Ghioni, Cecilia Viscardi, and Monia Lupparelli</i>	
Mapping University Students with Migrant Background: Statistical Challenges of an Italian Case Study	320
<i>Lorenzo Giammei, Laura Terzera, and Fulvia Mecatti</i>	
Joint Modelling of Particulates: An Application to Urban Areas in Friuli Venezia Giulia, Italy	326
<i>Vincenzo Gioia and Roberta Pappadà</i>	
Bayesian Inference for Initial Heat States with Gaussian Series Priors	332
<i>Matteo Giordano</i>	
Energy Load Forecasting Using Terna Public Data: A Free Lunch Multi-task Combination Approach	338
<i>Daniele Girolimetto and Tommaso Di Fonzo</i>	
Tackling Noise in Ranking Models with the Boosting Paradigm	345
<i>Giuseppe Gismondi, Luca Coraggio, and Antonio D'Ambrosio</i>	
The Role of Citizen Generated Data (CGD) in Measuring Violence Against Women (VAW) in Italy: Innovative Approaches and Progress in Enhancing Official Statistical Methods and Data Collection	350
<i>Francesco Gosetti, Federica Pellizzaro, and Lucilla Scarnicchia</i>	

Modeling Soft Skills Development During School-to-Work Programme: A Structural Equation Approach	356
<i>Grazia Graziosi, Donatella Poliandri, and Graziana Epifani</i>	
Biogeographical Ancestry Prediction via an Innovative Panel: Difficult Task or Complex Phenomenon?	363
<i>Cosimo Grazzini, Giorgia Spera, Daniele Castellana, Stefania Morelli, Elena Pilli, Michela Baccini, and Giulia Cereda</i>	
The Use of the BEPS Country by Country Report for the Purpose of Representing the Structure of MNEs and Recording the Economic Flows Between Units for Estimating GDP and GNI	369
<i>Gianna Greca</i>	
A Cross-Sectional Model Implementing Intersectional Analysis to Identify Extreme Profiles in Invalsi Data	373
<i>Enrico Contin and Leonardo Grilli</i>	
Climate Change and Banks' Operational Risk: a Research Road-Map	379
<i>Elena Grinza, Parisa Madhooshiarzanagh, and Consuelo Rubina Nava</i>	
Bayesian Synthetic Control with Spatially Augmented Priors	385
<i>Giulio Grossi, Leo Vanciu, and Falco J. Stoffi Bargagli</i>	
Sustainable and Innovative Management in a Circular Economy Perspective: An Application to the Italian Waste Sector	391
<i>Aamir Javed and Agnese Rapposelli</i>	
GAM-Based Variable Importance for Understanding Obesity	396
<i>Amir Khorrami Chokami</i>	
Gender Disparity in STEM Enrolment Across Italian Provinces	402
<i>Giuliana La Mantia and Giovanni Boscaino</i>	
Hierarchical Forecasting of Italian Mortality Data with Gender and Cause of Death Reconciliation	408
<i>Giacomo Lanfiuti Baldi, Raffaele Mattera, Andrea Nigri, and Hanlin Shang</i>	
Exploring 'Deaths of Despair' in Italy: Pattern in Alcohol, Drug, and Suicide Mortality	415
<i>Giacomo Lanfiuti Baldi, Andrea Nigri, Sergi Trias-Llimós, and Elisabetta Barbi</i>	

Model-Based Clustering to Detect Patterns in the Italian Parliament Decrees Issuing	420
<i>Alessandro Lanteri, Martina Dattilo, and Fabio Padovano</i>	
Functional Marked Hawkes Processes, with a View to Model Hurricanes Data	427
<i>Giada Lo Galbo, Giada Adelfio, Marcello Chiodi, and Mohammad Ghorbani</i>	
A Specification Test for the Matrix Form of Error Correction Models	434
<i>Emanuele Lopetuso and Massimiliano Caporin</i>	
Discontinuous Hamiltonian Monte Carlo for Non-continuous Probability Spaces: Patient Reaction to a Treatment	439
<i>Paolo Manildo and Bruno Scarpa</i>	
Robust Differential Network Analysis in Proteomics: Confounder Adjustment, Biologically Informed Thresholding, and Bootstrap Inference	445
<i>Alessia Mapelli, Solène Cadiou, Emanuele Di Angelantonio, and Francesca Ieva</i>	
A Convolution Process for Spatial Statistical Models on Directed Linear Networks	451
<i>Leonardo Marchesin, Alessandra Menafoglio, and Piercesare Secchi</i>	
EGNAN: An Enhanced Method for Interpretable Graph Neural Networks	457
<i>Chiara Marciano, Brian D. Bernhardt, Lucia Maddalena, and Mario R. Guarracino</i>	
Time Series Bootstrap for Renewable Energy Analysis in Power Grid Optimization	463
<i>Giulia Marcon, Andrea Marletta, Salar Moradi, Gaetano Zizzo, Salvatore Favuzza, and Gianluca Sottile</i>	
The Relationship Between Work, Health, and Well-Being: A Comparative Study of Two Italian Cohorts	469
<i>Benedetta Marcozzi, Chiara Donfrancesco, Cinzia Lo Noce, Serena Vannucchi, Luigi Palmieri, and Eleonora Trappolini</i>	
Balancing Covariates in Comparative Experiments with the Simulated Annealing	475
<i>Francesco Mariani, Rosamarie Frieri, and Marco Novelli</i>	

Simulating Joint Models for Recurrent and Terminal Events with Discrete Frailty 481
 Chiara Masci, Marta Spreafico, and Francesca Ieva

Using Vine Copulas for Estimating Highest-Density Regions in Multivariate Data 488
 Emanuele Masillo and Nina Deliu

Author Index 495



Robust Differential Network Analysis in Proteomics: Confounder Adjustment, Biologically Informed Thresholding, and Bootstrap Inference

Alessia Mapelli^{1,2}(✉), Solène Cadiou¹, Emanuele Di Angelantonio^{1,3},
and Francesca Ieva^{1,2}

¹ Fondazione Human Technopole, Milan, Italy
alessia.mapelli@polimi.it

² Politecnico di Milano, Milan, Italy

³ University of Cambridge, Cambridge, UK

Abstract. Differential network analysis has emerged as a powerful tool for identifying condition-specific changes in molecular interactions, particularly in high-dimensional omics data. Traditional approaches, such as Fisher's z-test and fixed-threshold correlation methods, often struggle with noise, confounding effects, and selecting an appropriate significance threshold. In this study, we present a novel framework that integrates confounder adjustment, bootstrapped correlation inference, and biologically informed threshold selection to improve the robustness of differential network analysis. We first validate our method using simulated data, comparing its performance against existing approaches. Our results demonstrate that our method better balances precision and recall, outperforming established benchmarks such as *DiffCorr*. We then apply our framework to a case study investigating sex-specific differences in protein interaction networks. Our results highlight the abilities of our approach to uncover meaningful molecular differential interactions in complex biological systems.

Keywords: Differential network analysis · Biologically informed thresholding · Noise-Resilient Correlation Estimation

1 Introduction

The advent of large-scale omics technologies, such as proteomics, has provided powerful tools for understanding the underlying cellular regulatory networks that govern complex biological systems. By leveraging these high-dimensional datasets, researchers can infer molecular interactions, uncover disease mechanisms, and identify key biomarkers. Among various network-based approaches, correlation networks have been widely used to reconstruct biological networks from high-throughput omics data [1]. In correlation-based networks, biomolecules (e.g., proteins, genes) are represented as nodes, and their

pairwise relationships, often measured via correlation coefficients, are depicted as undirected edges. Such networks have been successfully applied to gene co-expression analysis, protein interaction mapping, and metabolic pathway reconstruction [2]. Despite their utility, population-level network models often fail to capture context-specific variations in biological systems, such as those occurring in disease states. To address this limitation, differential network analysis has gained increasing attention to investigate alterations in connectivity patterns between two conditions (e.g., diseased vs. healthy) [3]. By examining changes in correlation structures, these methods can identify dysregulated pathways and pinpoint disease-specific regulatory mechanisms. Several computational approaches have been proposed to identify differential correlations in large-scale omics datasets, with one of the most common methods being Fisher's z-test for comparing correlation coefficients between two groups [4]. The construction of differential correlation-based networks generally follows a systematic pipeline. First, the dataset is divided into two subgroups based on the condition of interest (e.g., disease vs. control). Next, the pairwise correlation coefficients between variables are computed using various statistical methods. A hypothesis test, such as Fisher's z-test, is then applied to assess whether the correlation differs significantly between the two groups. The resulting p-values are then compared to a predefined significance threshold (e.g., $\alpha = 0.01$ or 0.05) with appropriate multiple-testing correction [4]. Despite its conceptual simplicity, this approach suffers from two major pitfalls that impact the robustness and reproducibility of correlation-based differential analysis. First, high-throughput omics datasets are inherently noisy, introducing biases in correlation estimates that can obscure true biological relationships. Spurious correlations may inflate the number of false positives, thereby reducing the statistical power of differential methods when multiple-testing correction is applied. To mitigate this issue, an additional preprocessing step, where only significant correlations are retained, is required to extract biologically meaningful relationships within each group. A crucial aspect of this step is the selection of an appropriate correlation threshold, as the chosen cutoff directly influences the network topology. Striking a balance between retaining meaningful interactions and eliminating noise is essential to performing reliable differential network analysis [5]. Second, systematic differences between the two groups (e.g., age, sex, batch effects) can confound correlation estimates, leading to spurious network structures that do not accurately reflect the effect of the condition of interest. Without proper adjustment for these confounders, the inferred differential network may capture their effects rather than accurately representing true differences between conditions.

This work presents a novel framework for differential network analysis that integrates robust non-parametric statistical tests, confounder adjustment strategies, and adaptive thresholding techniques. Our method leverages the vast knowledge embedded in biological databases, aligning inferred networks with prior biological information. By combining data-driven network inference with existing biological knowledge, we enhance the robustness and accuracy of the differential networks. We evaluate our approach using both simulated and real proteomics datasets, demonstrating its effectiveness in identifying biologically meaningful network perturbations and its applicability to high-dimensional settings.

2 Method

This study presents a novel approach to overcoming key challenges in correlation-based differential network inference. To mitigate the impact of confounding factors, we introduce a preprocessing step prior to the correlation computation. First, we examine the association between the proteomics data's principal components (PCs) and known covariates commonly considered in the literature (e.g., age, sex, batch effects) to identify the most relevant confounding factors. We focus on covariates that explain the majority of the variation in the data and adjust for them using linear modeling. The residuals from this adjustment are then split into two groups based on the condition of interest (e.g., disease vs. control). These residuals serve as the input for correlation estimation, ensuring that the inferred relationships are not biased by confounding factors. To guide the selection of a biologically meaningful correlation threshold, we begin by extracting the subset of proteins that are present in a reference database from the computed correlation matrix. We then evaluate a range of correlation thresholds, converting the correlation matrix into an adjacency matrix where proteins are connected by an edge if their correlation exceeds the selected threshold. We assess how well the resulting network aligns with the reference network for each threshold using Fisher's exact test. The threshold that yields the lowest p-value (indicating the highest agreement with the reference network) is selected as the optimal cutoff for the correlation. This approach ensures that the inferred network structure reflects established molecular interactions as much as possible, thereby enhancing the robustness and reliability of the results [8]. Following the approach in [7], we further refine the data by removing the final principal components, which typically represent low-variance noise. This step helps prevent the introduction of spurious correlations, ensuring that differential network inference captures biologically meaningful variation rather than technical artifacts. To enhance the stability of the correlation estimates, we implement a bootstrap procedure in which the correlation is computed multiple times on random subsamples of the datasets. For each iteration of the bootstrap algorithm, the pairwise correlations are stored, so that each edge in the network is associated with a distribution of correlation values across subsamples. We perform a Wilcoxon signed-rank test to determine whether the edges' correlations exceed the bootstrap optimal threshold, computed as the median of those stored. Considering a multiple-test correct threshold of 0.05, only significant edges are retained, yielding a robust correlation network for each condition.

Once we have separately inferred the correlation network in the two conditions, the final differential network is derived by comparing the distribution of correlation values for each edge between the two conditions of interest. A two-sided Wilcoxon test is used to assess significant differences in correlation distributions. Edges significantly different between the two groups, either present in only one condition or exhibiting a strong shift in correlation strength, are retained in the differential network.

3 Simulation Analysis

We assessed our method's performance through simulations, comparing it to DiffCorr [4], an R package based on Fisher's z-test, and more straightforward fixed-threshold correlation methods ($|r| > 0.7$, $|r| > 0.5$, $|r| > 0.3$) [9].

Data were generated with 100 variables and 500 samples per condition, simulating scale-free networks to mimic protein interactions, with added Gaussian noise to reflect proteomics variability. Shared edges were retained as prior biological knowledge, while true differential connections were defined as those differing between conditions. The results of our simulation analysis are summarized in Table 1. Our approach outperforms DiffCorr with a significantly improved precision (1 vs. 0.003) with only a modest trade-off in recall (0.375 vs. 0.53). This leads to a substantially higher F1 score (0.545 vs 0.006), the method’s ability to detect meaningful differential correlations. Compared to fixed-threshold correlation methods, our approach offers a better balance between precision and recall, avoiding the extremes of overly permissive or overly conservative cut-offs. Our method optimally balances sensitivity and specificity by filtering noise, adapting thresholds based on biological knowledge, and accounting for confounders. This highlights its advantage over existing differential network approaches and conventional threshold-based correlation methods.

Table 1. Results of the simulation analysis.

Model	Precision	Recall	F1
Proposed method	1	0.375	0.545
DiffCorr	0.003	0.53	0.006
Corr – 0.7	0.37	0.541	0.44
Corr – 0.5	0.014	0.833	0.028
Corr – 0.3	0.008	0.833	0.0153
Corr – 0	0.008	0.833	0.016

4 Case Study

We applied our methodology to the INTERVAL cohort data to analyze differential protein interaction networks between males and females. The INTERVAL study [6] is a large-scale cohort of approximately 50,000 whole blood donors recruited between 2012 and 2014 as part of an open randomized pragmatic trial of varying blood donation intervals. A case-cohort design selected 9,679 participants for plasma protein quantification using the SomaScan platform V4.1, measuring 7,289 proteins.

Within this case study, we aimed to identify key regulatory proteins linked to sex-specific biological processes and diseases. Using principal component analysis (PCA), we examined protein abundance in relation to various demographic and anthropometric measurements (such as age, BMI, and ethnicity) and technical factors (such as the time between blood draw and processing). To control for confounders, we adjusted for the most influential of them via the following linear regression model: protein ~ age + time of processing + season of blood drawn + genetic principal components. Residuals from this model were used for further analysis.

Previous biological knowledge on protein-protein interaction was extracted from the STRING database [7], which systematically collects and integrates data from multiple sources, including automated text mining of the scientific literature and databases of interaction experiments to define robust physical and functional interactions knowledge. To ensure reliability, we filtered the interactions at a high-confidence threshold (0.70) and mapped them to our dataset, resulting in a set of 5,189 proteins shared between the two. Network inference was conducted separately for males and females using 11 bootstrap samples of 300 individuals each. We then performed differential network analysis to identify diverse interactions between the two groups. The differential network comprises 13,041,076 edges, representing 48% of all possible connections and indicating substantial differences in protein interactions between females and males. This finding aligns with existing literature reporting sex-specific variation in protein abundance; in our dataset specifically, nearly 20% of proteins showed statistically significant associations with sex. To identify the most influential proteins within the differential network, we computed multiple centrality metrics (degree, eigenvector centrality, and PageRank) and selected the top 1% of nodes with the highest average of these measures. We then merged this set with the top 1% of nodes ranked by hubness, a measure of a node's importance for maintaining network connectivity. This approach yielded a final list of 63 proteins that appear central to the sex-specific interactome differences. When used as predictors in a logistic regression model, these proteins achieved an AUC of 71% in classifying patient sex. This result suggests that differential network analysis can reveal potential biomarkers and key regulators of sex-specific protein interaction patterns.

5 Discussion

High-throughput proteomics data present unique challenges for differential network analysis due to inherent noise and confounding variable effects. Traditional multiple-testing approaches struggle to maintain statistical power in this setting, leading to inflated false positives and reduced sensitivity. We developed a robust differential network inference framework to address these limitations that integrates biologically informed correlation thresholding, confounder adjustment, and non-parametric hypothesis testing. Additionally, our method leverages bootstrap resampling to enhance the stability and reproducibility of network inference. Through a simple simulation study, we demonstrated that our approach outperforms benchmark methods, improving specificity while maintaining a higher true positive rate. Moreover, the application to real proteomics data showcases its practicality in extracting meaningful biological insights from high-dimensional datasets. Despite these advancements, several challenges remain. Our method still depends on setting the correlation threshold based on general biological knowledge, which may introduce bias if the groups under study are not well represented in the broader population. Our case study was designed to mitigate this using a balanced common scenario. Additionally, the model has limitations in capturing complex conditional dependencies between proteins, which may obscure subtle but biologically relevant interactions. A promising direction for future work are Gaussian graphical models, which account for conditional correlations rather than marginal associations. However, these models introduce additional computational complexity and may struggle with the high dimensionality of proteomics data. Balancing model interpretability, statistical power, and

computational efficiency will be crucial in advancing differential network analysis for high-throughput omics studies.

References

1. Albert, R.: Network inference, analysis, and modeling in systems biology. *Plant Cell. American Society of Plant Biologists*, vol. 19, pp. 3327–3338, 2007
2. Barabási, A.L., Gulbahce, N., Loscalzo, J.: Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **12** (2011)
3. de la Fuente, A.: From differential expression to differential networking—Identification of dysfunctional regulatory networks in diseases. *Trends Genet.* **26**, 326–333 (2010)
4. Fukushima, A.: DiffCorr: an R package to analyze and visualize differential correlations in biological networks. *Gene* **518**(1), 209–214 (2013)
5. Benedetti, E., Pučić-Baković, M., Keser, T., Gerstner, N., Büyüközkcan, M., Štam-buk, T., et al.: A strategy to incorporate prior knowledge into correlation net-work cutoff selection. *Nat Commun.* **11**(1) (2020)
6. Kaptoge, S., Di Angelantonio, E., Moore, C., Walker, M., Armitage, J., Ouwehand, W.H., et al.: Longer-term efficiency and safety of increasing the frequency of whole blood donation (INTERVAL): extension study of a randomised trial of 20 757 blood donors. *Lancet Haematol.* **6**(10), e510–e520 (2019)
7. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., et al.: The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucl. Acids Res.* **51**(1 D), D638–D646 (2023)
8. Bakker, O.B., Claringbould, A., Westra, H.J., Wiersma, H., Boulogne, F., Vösa, U., et al.: Linking common and rare disease genetics through gene regulatory net-works. *medRxiv*. 2021
9. Hung, T.I., Hsieh, Y.J., Lu, W.L., Wu, K.P., Chia-en, A.C.: What strengthens protein-protein interactions: analysis and applications of residue correlation net-works. *J. Mol. Biol.* **435**(24), 168337 (2023)