



Università
Bocconi
MILANO



POLITECNICO
MILANO 1863

UNIVERSITÀ DEGLI STUDI
DI MILANO
BICOCCA



UNIVERSITÀ DEGLI STUDI DI MILANO



Smart Statistics for Smart Applications

Book of Short Papers SIS2019



Editors: Giuseppe Arbia, Stefano Peluso,
Alessia Pini and Giulia Rivellini

Copyright © 2019

PUBLISHED BY PEARSON

WWW.PEARSON.COM

Giugno 2019 ISBN 9788891915108

Preface

Section 1. Plenary Sessions and Round Table

Preface	3
Shallow Learning for Data Science	7
<i>Antonio Canale</i>	
Smart Statistics: concept, technology and service	17
<i>David John Hand, Maurizio Vichi</i>	
Tavola rotonda “Smart ageing: lunga vita attiva, salute e nuove tecnologie”	19

Section 2. Invited Papers

Demography in the Digital Era: New Data Sources for Population Research	23
Demografia nell’era digitale: nuovi fonti di dati per gli studi di popolazione	23
<i>Diego Alburez-Gutierrez, Samin Aref, Sofia Gil-Clavel, André Grow, Daniela V. Negraia, Emilio Zagheni</i>	
Stationarity of a general class of observation driven models for discrete valued processes	31
Stazionarietà di una classe generale di modelli observation-driven per processi a valori discreti	
<i>Mirko Armillotta, Alessandra Luati and Monia Lupparelli</i>	
An extension of the censored gaussian lasso estimator	39
Un’estensione dello stimatore cglasso	
<i>Luigi Augugliaro and Gianluca Sottile and Veronica Vinciotti</i>	
A formal approach to data swapping and disclosure limitation techniques	47
Un approccio formale per tecniche di trasformazione dei dati in problemi di privacy	
<i>F. Ayed, M. Battiston and F. Camerlenghi</i>	
A new ordinary kriging predictor for histogram data in L2-Wasserstein space	55
Un nuovo predittore kriging per istogrammi nello spazio L2-Wasserstein	
<i>Antonio Balzanella and Antonio Irpino and Rosanna Verde</i>	
Keywords dynamics in online social networks: a case-study from Twitter	63
La dinamica delle parole chiave nelle reti sociali online: un esempio tratto da Twitter	
<i>Carolina Becatti, Irene Crimaldi and Fabio Saracco</i>	
Statistical Matching of HBS and ADL to analyse living conditions, poverty and happiness	71
Statistical Matching di HBS e ADL per l’analisi di condizioni di vita, povertà e felicità	
<i>Cristina Bernini, Silvia Emili, Maria Rosaria Ferrante</i>	
Statistical sources for cybersecurity and measurement issues	79
Fonti statistiche per la sicurezza cibernetica e problemi di misurazione	
<i>Claudia Biancotti, Riccardo Cristadoro, Raffaele Tartaglia Polcini</i>	
Use of GPS-enabled devices data to analyse commuting flows between Tuscan municipalities	89
Un’analisi dei flussi di pendolarismo sistematici tra i comuni toscani tramite l’utilizzo di dati GPS	
<i>Chiara Bocci, Leonardo Piccini and Emilia Rocco</i>	
Statistical calibration of the digital twin of a connected health object	97
Inversione statistica dei parametri di ingresso per il gemello digitale di un oggetto sanitario collegato	
<i>Nicolas Bousquet and Walid Dabachine</i>	
Time Series Forecasting: Is there a role for neural networks?	103
Le Reti Neurali nella Previsione di Serie Storiche	
<i>Giuseppe Bruno, Sabina Marchetti, Juri Marcucci, Diana Nicoletti</i>	

Modelling weighted signed networks.....	111
Modellazione di reti segnate pesate	
<i>Alberto Caimo and Isabella Gollini</i>	
Issues on Bayesian nonparametric measures of disclosure risk	119
Questioni su misure Bayesiane nonparametriche di rischio di "disclosure"	
<i>Federico Camerlenghi, Cinzia Carota and Stefano Favaro</i>	
Hierarchies of nonparametric priors.....	125
Gerarchie di distribuzioni iniziali nonparametriche	
<i>Federico Camerlenghi, Stefano Favaro and Lorenzo Masoero</i>	
Issues with Nonparametric Disclosure Risk Assessment.....	133
Questioni sull'Analisi Nonparametrica del Rischio di "Disclosure"	
<i>Federico Camerlenghi, Stefano Favaro, Zacharie Naulet and Francesca Panero</i>	
Technologies and data science for a better health both at individual and population level. ..	141
Two practical research cases.	
Tecnologie e data science per una salute migliore sia a livello individuale che di popolazione.	
<i>Stefano Campostrini and Lucia Zanotto</i>	
Temporal sentiment analysis with distributed lag models	149
Analisi temporale del "sentiment" con modelli a lag distribuiti	
<i>Carrannante M., Mattered R., Misuraca M., Scepti G., Spano M.</i>	
A statistical investigation on the relationships among financial disclosure, sociodemographic variables, financial literacy and retail investors' risk assessment ability	157
Indagine empirica sulle relazioni tra prospetti per la diffusione di informazioni finanziarie, variabili sociodemografiche, educazione finanziaria e abilità di valutazione del rischio	
<i>Rosella Castellano, Marco Mancinelli and Pasquale Samacchiaro</i>	
Bayesian Model Comparison based on Wasserstein Distances.....	167
Confronto di Modelli Bayesiani tramite Distanze di Wasserstein	
<i>Marta Catalano, Antonio Lijoi and Igor Prünster</i>	
Hierarchical Clustering and Dimensionality Reduction for Big Data	173
Clustering e Riduzione Dimensionale Gerarchici per Dati di Grandi Dimensioni	
<i>Carlo Cavicchia, Maurizio Vichi and Giorgia Zaccaria</i>	
ICOs success drivers: a textual and statistical analysis.....	181
Fattori di successo nelle ICOs: un'analisi testuale e statistica	
<i>Paola Cerchiello and Anca Mirela Toma</i>	
Small area estimators with linked data.....	189
Stimatori per piccole aree nel caso di dati ottenuti attraverso il record linkage	
<i>Chambers Raymond and Fabrizi Enrico and Salvati Nicola</i>	
Optimal Portfolio Selection via network theory in banking and insurance sector.....	197
<i>Gian Paolo Clemente, Rosanna Grassi and Asmerilda Hitaj</i>	
Matching error(s) and quality of statistical matching in complex surveys.....	205
Errori di matching e qualità del matching statistico in indagini complesse	
<i>Pier Luigi Conti and Daniela Marella</i>	
Hotel search engine architecture based on online reviews' content.....	213
Un motore di ricerca per gli hotel basato sulle recensioni online	
<i>Claudio Conversano, Maurizio Romano and Francesco Mola</i>	
Economic Crisis and Earnings Management: a Statistical Analysis	219
Crisi Economica e Gestione degli Utili: un'Analisi Statistica	
<i>C. Cusatelli, A.M. D'Uggento, M. Giacalone, F. Grimaldi</i>	
A Comparison of Nonparametric Bivariate Survival Functions.....	227
Confronto tra stimatori non-parametrici della funzione di sopravvivenza bivariata	
<i>Hongsheng Dai and Marialuca Restaino</i>	
Predictive Algorithms in Criminal Justice.....	237
Algoritmi predittivi e giustizia penale	
<i>Francesco D'Alessandro</i>	

A proposal for an integrated approach between sentiment analysis and social network analysis.....	247
Una proposta per un approccio integrato tra analisi del sentimento e analisi delle reti sociali	
<i>Domenico De Stefano and Francesco Santelli</i>	
A meta-tissue non-parametric factor analysis model for gene co-expression	255
Meta-analisi fattoriale non parametrica per lo studio di espressioni genetiche in diversi tessuti	
<i>Roberta De Vito and Barbara Engelhardt</i>	
Bayesian estimate of population count with false captures: a latent class approach.....	261
Stima Bayesiana della popolazione con false catture: un approccio basato sulle classi latenti	
<i>Davide Di Cecco, Marco Di Zio and Brunero Liseo</i>	
Spherical regression with local rotations and implementation in R	269
Regressione sferica con rotazioni locali ed implementazione in R	
<i>Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor</i>	
A clustering method for network data to analyse association football playing styles	277
Un metodo di raggruppamento per dati di rete finalizzato all'analisi degli schemi di gioco nel calcio	
<i>Jacopo Diqigiovanni</i>	
Big data in longitudinal observational studies: how to deal with non-probability samples and technological changes.....	285
I Big data negli studi longitudinali: come trattare campioni non probabilistici e cambi di tecnologia	
<i>Clelia Di Serio, Luca Del Core, Eugenio Montini and Andrea Calabria</i>	
Smart Data For Smart Health.....	293
Smart Data Per Smart Health	
<i>Clelia Di Serio, Ernst C. Wit, Elena Bottinelli and Roberto Buccione</i>	
Detecting and classifying moments in basketball matches using sensor tracked data.....	297
Una procedura per identificare e classificare momenti di gioco in pallacanestro con l'uso di dati sensori.	
<i>Tullio Facchinetti and Rodolfo Metulini and Paola Zuccolotto</i>	
Ordered response models for cyber risk	305
Modelli a risposta ordinale per la valutazione del cyber risk	
<i>Silvia Facchinetti and Claudia Tarantola</i>	
Functional data analysis-based sensitivity analysis of integrated assessment Models for climate change modelling	313
Analisi di sensibilità basata sull'analisi di dati funzionali per modelli di valutazione integrata dei cambiamenti climatici	
<i>Matteo Fontana, Massimo Tavoni and Simone Vantini</i>	
Coupled Gaussian Processes for Functional Data Analysis.....	319
Processi gaussiani per l'analisi dei dati funzionali	
<i>L. Fontanella, S. Fontanella, R. Ignaccolo, L. Ippoliti, P. Valentini</i>	
Two-fold data streams dimensionality reduction approach via FDA	323
Un approccio a due fasi per la riduzione di dimensionalità di data streams via FDA	
<i>F. Fortuna, T. Di Battista and S.A. Gattone</i>	
Statistical analysis of Sylt's coastal profiles using a spatiotemporal functional model	331
<i>Rik Gijsman, Philipp Otto, Torsten Schlurmann, Jan Visscher</i>	
Bootstrap prediction intervals for weighted TAR predictors	339
Intervalli di previsione bootstrap per previsori ponderati per modelli TAR	
<i>Francesco Giordano and Marcella Niglio</i>	
A rank graduation index to prioritise cyber risks	347
Un indice di graduazione per assegnare livelli di priorità ai rischi informatici	
<i>Paolo Giudici and Emanuela Raffinetti</i>	
Vector Error Correction models to measure connectedness of bitcoin exchange markets	355
Modelli di Vector Error Correction per misurare la connessione delle piattaforme di scambio di bitcoin	
<i>Paolo Giudici and Paolo Pagnottoni</i>	
Estimation of lineup efficiency effects in Basketball using play-by-play data.....	363
L'uso dei dati del play-by-play per la stima degli effetti di quintetto nella pallacanestro	
<i>Luca Grassetti, Ruggero Bellio, Giovanni Fonseca and Paolo Vidoni</i>	
Trajectory clustering using adaptive squared distances	371
Clustering di traiettorie attraverso distanze adattative quadratiche	
<i>Antonio Irpino</i>	

Bayesian Analysis of Privacy Attacks on GPS Trajectories	379
<i>Analisi Bayesiana degli Attacchi alla Privacy su Traiettorie GPS</i>	
<i>Sirio Legramanti</i>	
Data Analytics in the Insurance Industry: Market trends and lessons from a use case customer predictive modelling	387
<i>Data Analytics nel settore assicurativo: principali trend e considerazioni da un caso d'uso applicato alla predizione del comportamento degli assicurati</i>	
<i>Cristian Losito and Francesco Pantisano</i>	
BasketballAnalyzeR: the R package for basketball analytics	395
<i>BasketballAnalyzeR: il pacchetto R per l'analisi dei dati nella pallacanestro</i>	
<i>Marica Manisera, Marco Sandri and Paola Zuccolotto</i>	
Data Integration by Graphical Models	403
<i>Utilizzo dei modelli grafici per l'integrazione dei dati</i>	
<i>Daniela Marella and Paola Vicard and Vincenzina Vitale</i>	
A two-part finite mixture quantile regression model for semi-continuous longitudinal data	409
<i>Maruotti Antonello, Merlo Luca and Petrella Lea</i>	
Multivariate change-point analysis for climate time series	415
<i>Analisi di change-point multivariati per serie storiche climatiche</i>	
<i>Gianluca Mastrantonio, Giovanna Jona Lasinio, Alessio Pollice, Giulia Capotorti, Lorenzo Teodonio and Carlo Blasi</i>	
A divide-et-impera approach for the spatial prediction of object data over complex regions	423
<i>Un approccio divide-et-impera per la previsione spaziale di dati oggetto su regioni complesse</i>	
<i>Alessandra Menafoglio e Piercesare Secchi</i>	
A strategy for the matching of mobile phone signals with census data.....	427
<i>Una strategia per l'abbinamento di segnali di telefonia mobile con dati censuari</i>	
<i>Rodolfo Metulini and Maurizio Carpita</i>	
Risk-based analyses for non-proportional reinsurance pricing	435
<i>Analisi Risk-based per il pricing nella riassicurazione di trattati non proporzionali</i>	
<i>Fabio Moraldi and Nino Savelli</i>	
A Simplified Efficient and Direct Unequal Probability Resampling	441
<i>Un semplice Ricampionamento, efficiente e diretto per campioni a probabilità variabili</i>	
<i>Federica Nicolussi, Fulvia Mecatti and Pier Luigi Conti</i>	
Labour Law: Machine vs. Employer Powers Diritto del lavoro: Macchina vs. Poteri datoriali	449
<i>Antonella Occhino – Michele Faioli</i>	
Domain knowledge based priors for clustering.....	455
<i>Distribuzioni a priori per l'analisi di raggruppamento basate sulla conoscenza di settore</i>	
<i>Sally Paganin</i>	
Clustering of Behavioral Spatial Trajectories in Neuropsychological Assessment	463
<i>Analisi dei gruppi di traiettorie spaziali nella valutazione neuropsicologica</i>	
<i>Francesco Palumbo, Antonio Cerrato, Michela Ponticorvo, Onofrio Gigliotta, Paolo Bartolomeo, Orazio Miglino</i>	
What is wrong in the debate about smart contracts.....	471
<i>Smart contract e diritto: riflessioni critiche su un dualismo fuorviante</i>	
<i>Roberto Pardolesi and Antonio Davola</i>	
Financial Transaction Data for the Nowcasting in Official Statistics	485
<i>Transazioni elettroniche di pagamento per le previsioni a breve nella Statistica ufficiale</i>	
<i>Righi A., Ardizzi G., Gambini A., Iannaccone R., Moauro F., Renzi N. and Zurlo D.</i>	
On the examination of a criticality measure for a complex system in a forecasting perspective	493
<i>Esame di una misura di criticità per un sistema complesso in una prospettiva previsiva</i>	
<i>Renata Rotondi and Elisa Varini</i>	
Knowledge discovery for dynamic textual data: temporal patterns of topics and word clusters in corpora of scientific literature	501
<i>Estrazione della conoscenza da dati testuali dinamici: evoluzione temporale di argomenti e gruppi di parole in corpora di letteratura scientifica</i>	
<i>Stefano Sbalchiero, Matilde Trevisani and Arjuna Tuzzi</i>	

Classifying the Willingness to Act in Social Media Data: Supervised Machine Learning for U.N. 2030 Agenda	509
<i>Classificare la volontà di agire nei dati dei Social Media: Supervised Machine Learning per l'Agenda 2030 delle Nazioni Unite</i>	
<i>Andrea Sciandra, Alessio Surian and Livio Finos</i>	
Classification of spatio-temporal point pattern in the presence of clutter using K-th nearest neighbour distances.....	517
<i>Classificazione dei processi puntuali spazio-temporali basata sulla distanza dal K-mo vicino più vicino</i>	
<i>Siino Marianna, Francisco J. Rodríguez-Cortés, Jorge Mateu, Giada Adelfio</i>	
Modelling properties of high-dimensional molecular systems	525
<i>La modellazione di sistemi molecolari ad alta dimensionalità</i>	
<i>Debora Slanzi, Valentina Mameli and Irene Poli</i>	
Non-crossing parametric quantile functions: an application to extreme temperatures	533
<i>Il problema del crossing con funzioni quantiliche parametriche: un'applicazione alle temperature estreme</i>	
<i>Gianluca Sottile and Paolo Frumento</i>	
A new tuning parameter selector in lasso regression.....	541
<i>Un nuovo criterio di selezione per il parametro di penalizzazione nella regressione lasso</i>	
<i>Gianluca Sottile and Vito MR Muggeo</i>	
Similarity patterns, topological information and credit scoring models	549
<i>Strutture di similarità, informazioni topologiche e modelli di credit scoring</i>	
<i>Alessandro Spelta, Branka Hadji-Misheva and Paolo Giudici</i>	
Between hawks and doves: measuring central bank communication	557
<i>Fra falchi e colombe: valutazione delle comunicazioni di Banca Centrale</i>	
<i>Ellen Tobback, Stefano Nardelli, David Martens</i>	
New methods and data sources for the population census	561
<i>Nuovi metodi e fonti per il censimento della popolazione</i>	
<i>Paolo Valente</i>	
FinTech and the Search for "Smart" Regulation	569
<i>Fintech e la ricerca di una regolamentazione "smart"</i>	
<i>Silvia Vanon</i>	
An anisotropic model for global climate data	577
<i>Un modello anisotropico per i dati climatici globali</i>	
<i>Nil Venet and Alessandro Fassò</i>	
Analysis of the financial performance in Italian football championship clubs via GEE and diagnostic measures.....	585
<i>Analisi delle performance finanziaria delle squadre di calcio di serie A via GEE e misure di diagnostica</i>	
<i>Maria Kelly Venezuela, Anna Crisci, Luigi D'Ambrà, D'Ambrà Antonello</i>	
A statistical space-time functional model for air quality analysis and mapping.....	593
<i>Un modello statistico spazio-tempo funzionale per l'analisi e la mappatura della qualità dell'aria</i>	
<i>Yaqiong Wang, Alessandro Fassò and Francesco Finazzi</i>	
Tempering and computational efficiency of Bayesian variable selection.....	599
<i>Tempering e l'efficienza computazionale della selezione bayesiana delle variabili</i>	
<i>Giacomo Zanella and Gareth O. Roberts</i>	
Dimensions and links for Hate Speech in the social media	607
<i>Dimensioni e legami per i discorsi di odio nei social media</i>	
<i>Emma Zavarrone, Guido Ferilli</i>	

Section 3. Contributed Papers

Density-based Algorithm and Network Analysis for GPS Data.....	617
<i>Algoritmi di Cluster e Reti per lo studio di dati GPS</i>	
<i>Antonino Abbruzzo, Mauro Ferrante, Stefano De Cantis</i>	
Local inference on functional data based on the control of the family-wise error rate	623
<i>Inferenza locale per dati funzionali basata sul controllo del family-wise error rate</i>	
<i>Konrad Abramowicz, Alessia Pini, Lina Schelin, Sara Sjöstedt de Luna, Aymeric Stamm, and Simone Vantini</i>	

Application and validation of dynamic Poisson models to measure credit contagion	629
<i>Applicazione e validazione di modelli di Poisson dinamici per misurare il contagio nel credito</i>	
<i>Arianna Agosto and Emanuela Raffinetti</i>	
Monitoring SDGs at territorial level: the case of Lombardy.....	637
<i>Il monitoraggio degli SDGs a livello territoriale: il caso della Lombardia</i>	
<i>Leonardo Alaimo, Livia Celardo, Filomena Maggino, Adolfo Morrone, Federico Olivieri</i>	
The Experts Method for the prediction of periodic multivariate time series of high dimension.....	643
<i>Il Metodo degli Esperti per la previsione di serie temporali multivariate e periodiche, di dimensione elevata</i>	
<i>Giacomo Aletti, Marco Bellan and Alessandra Micheletti</i>	
Regression with time-dependent PDE regularization for the analysis of spatio-temporal data	649
<i>Regressione con regolarizzazione di PDE tempo dipendenti per modellizzare dati spatio-temporali</i>	
<i>Eleonora Arnone, Laura Azzimonti, Fabio Nobile, Laura M. Sangalli</i>	
A network analysis of museum preferences: the Firenzecard experience.....	653
<i>Un'analisi di rete delle preferenze museali: l'esperienza della Firenzecard</i>	
<i>Silvia Bacci, Bruno Bertaccini, Roberto Dinelli, Antonio Giusti, and Alessandra Petrucci</i>	
A statistical learning approach to group response categories in questionnaires.....	659
<i>Un approccio basato sull'apprendimento statistico per raggruppare le categorie di risposta nei questionari</i>	
<i>Michela Battauz</i>	
Tree-based Functional Data Analysis for Classification and Regression.....	665
<i>Alberi di Classificazione e Regressione per dati Funzionali</i>	
<i>Edoardo Belli, Enrico Ragaini, Simone Vantini</i>	
PDE-regularized regression for anisotropic	669
<i>spatial fields Regressione con regolarizzazione differenziale per campi spaziali anisotropi</i>	
<i>Mara S. Bernardi, Michelle Carey, James O. Ramsay and Laura M. Sangalli</i>	
A Bayesian model for network flow data: an application to BikeMi trips	673
<i>Giulia Bissoli, Celeste Principi, Gian Matteo Rinaldi, Mario Beraha and Alessandra Guglielmi</i>	
Statistical classics in the big data era. When (astro-physical) models are nonregular.....	679
<i>Statistica classica nell'era dei big data. Verosimiglianza e modelli non regolari</i>	
<i>Alessandra R. Brazzale and Valentina Mameli</i>	
Bayesian Variable Selection for High Dimensional Logistic Regression	685
<i>Selezione bayesiana delle variabili nel modello di regressione logistica ad alta dimensionalita</i>	
<i>Claudio Busatto, Andrea Sottosanti and Mauro Bernardi</i>	
Bayesian modeling for large spatio-temporal data: an application to mobile networks	691
<i>Modelli bayesiani per grandi dataset spatio-temporali: un'applicazione a dati di telefonia mobile</i>	
<i>Annalisa Cadonna, Andrea Cremaschi, Alessandra Guglielmi</i>	
A Mathematical Framework for Population of Networks: Comparing Public Transport of Different Cities.	697
<i>Un approccio matematico all'analisi di una popolazione di networks: come confrontare il sistema di trasporto pubblico di diverse città.</i>	
<i>Anna Calissano, Aasa Feragen, Simone Vantini</i>	
How Important Discrimination is for the Job Satisfaction of Immigrants in Italy: A Counterfactual Approach	703
<i>Quanto influisce la discriminazione sulla soddisfazione lavorativa degli immigrati in Italia: un approccio controfattuale</i>	
<i>Maria Gabriella Campolo, Antonino Di Pino and Michele Limosani</i>	
Unfolding the SEcrets of LongEvity: Current Trends and future prospects (SELECT)	709
<i>A path through morbidity, disability and mortality in Italy and Europe</i>	
<i>Stefano Campostrini, Daniele Durante, Fabrizio Faggiano and Stefano Mazzucco</i>	
Galaxy color distribution estimation via dependent nonparametric mixtures	713
<i>Stima della distribuzione del colore delle galassie via misture nonparametriche dipendenti</i>	
<i>Antonio Canale, Riccardo Corradin and Bernardo Nipoti</i>	
A case for order optimal matching: a salary gap study.....	719
<i>Un algoritmo di matching ottimale ordinato per un studio sulle differenze salariali</i>	
<i>Massimo Cannas</i>	

A Prediction Method for Ordinal Consistent Partial Least Squares	725
Un Metodo di Previsione per l'Algoritmo Ordinal Consistent Partial Least Squares	
<i>Gabriele Cantaluppi and Florian Schuberth</i>	
Functional control charts for monitoring ship operating conditions and CO2 emissions based on scalar-on-function linear model	731
Carte di controllo funzionali per il monitoraggio delle condizioni operative e delle emissioni di CO2 di navi da carico e passeggeri mediante modello di regressione funzionale con risposta scalare	
<i>Christian Capezza, Antonio Lepore, Alessandra Menafoglio, Biagio Palumbo, and Simone Vantini</i>	
Predicting and improving smart mobility: a robust model-based approach to the BikeMi BSS	737
Prevedere e migliorare la mobilità smart: un approccio robusto di classificazione applicato a BikeMi	
<i>Andrea Cappozzo, Francesca Greselin and Giancarlo Manzi</i>	
Public support for an EU-wide social benefit scheme: evidence from Round 8 of the European Social Survey (ESS)	743
Sostegno pubblico a un sistema di prestazioni sociali a livello dell'Unione Europea: i risultati del Round 8 della European Social Survey (ESS)	
<i>Paolo Emilio Cardone</i>	
Revenue management strategies and Booking.com ghost rates: a statistical analysis	751
Strategie di revenue management e Booking.com ghost rates: un'analisi statistica	
<i>Cinzia Carota, Consuelo R. Nava, Marco Alderighi</i>	
Analysing international migration flows: a Bayesian network approach	757
Analisi dei flussi migratori internazionali attraverso l'impiego di modelli grafici	
<i>Federico Castelletti and Emanuela Furfaro</i>	
A sparse estimator for the function-on-function linear regression model	763
Uno stimatore sparso per il modello di regressione lineare con regressore e risposta funzionali	
<i>Fabio Centofanti, Matteo Fontana, Antonio Lepore, and Simone Vantini</i>	
Robustness and fuzzy multidimensional poverty indicators: a simulation study.....	769
Robustezza ed indicatori fuzzy multidimensionali della povertà: uno studio di simulazione	
<i>Michele Costa</i>	
Text Based Pricing Modelling: an Application to the Fashion Industry	775
Modellazione dei prezzi basata su dati testuali: un'applicazione all'industria fashion	
<i>Federico Crescenzi, Marzia Freo and Alessandra Luati</i>	
Model based clustering in group life insurance via Bayesian nonparametric mixtures	781
Raggruppamento basato sul modello nel settore assicurativo: un approccio bayesiano nonparametrico	
<i>Laura D'Angelo</i>	
Smart Tools for Academic Submission Decisions: Waiting Times Modeling	787
Strumenti "Smart" per sottoporre i manoscritti accademici: modelli per i tempi di attesa	
<i>Francesca De Battisti - Giancarlo Manzi</i>	
On the Use of Control Variables in PLS-SEM	793
Sull'Uso delle Variabili di Controllo nei PLS-SEM	
<i>Francesca De Battisti and Elena Siletti</i>	
Partial dependence with copula and financial applications	799
Dipendenza parziale con funzioni copula e applicazioni finanziarie	
<i>Giovanni De Luca, Marta Nai Ruscone and Giorgia Riveccio</i>	
Exploring the relationship between fertility and well-being: What is smart?.....	805
Esplorando la relazione tra fecondità e benessere: cosa c'è di smart?	
<i>Alessandra De Rose, Filomena Racioppi, Maria Rita Sebastiani</i>	
Web-Based Data Collection and Quality Issues in Co-Authorship Network Analysis	811
Qualità dei dati bibliografici raccolti via web per l'analisi di reti di collaborazione scientifica	
<i>Domenico De Stefano, Vittorio Fuccella, Susanna Zaccarin</i>	
A new regression model for bounded multivariate responses.....	817
Un nuovo modello di regressione per risposte multivariate limitate	
<i>Agnese Maria Di Brisco, Roberto Ascarì, Sonia Migliorati and Andrea Ongaro</i>	
Turning big data into smart data: two examples based on the analysis of the Mappa dei Rischi dei Comuni Italiani.....	823
Trasformare i big data in smart data: due esempi di analisi della Mappa dei Rischi dei Comuni Italiani	
<i>Oleksandr Didkovskiy, Alessandra Menafoglio, Piercesare Secchi, Giovanni Azzone</i>	

Hidden Markov Model estimation via Particle Gibbs	829
Stima di Hidden Markov Model tramite Particle Gibbs	
<i>Pierfrancesco Alaimo Di Loro, Enrico Ciminello and Luca Tardella</i>	
A note on marginal effects in logistic regression with independent covariates	837
Una nota sugli effetti marginali nella regressione logistica con covariate indipendenti	
<i>Marco Doretti</i>	
DNA mixtures: a case study involving a Romani reference population	843
Misure di DNA: un caso di studio riguardante una popolazione di riferimento dei Rom	
<i>Francesco Dotto, Julia Mortera and Vincenzo Pascali</i>	
Pivotal seeding for K-means based on clustering ensembles	849
Inizializzazione pivotale dell'algoritmo delle K-medie tramite raggruppamento con metodi di insieme	
<i>Leonardo Egidi, Roberta Pappadà, Francesco Pauli, Nicola Torelli</i>	
Optimal scoring of partially ordered data, with an application to the ranking of smart cities	855
Scoring ottimale di dati parzialmente ordinati, con un'applicazione al ranking delle smart city	
<i>Marco Fattore, Alberto Arcagni, Filomena Maggino</i>	
Bounded Domain Density Estimation	861
Stima della densità non-parametrica su domini bidimensionali limitati	
<i>Federico Ferraccioli, Laura M. Sangalli and Livio Finos</i>	
Polarization and long-run mobility: yearly wages comparison in three southern European countries	867
Polarizzazione e mobilità sul lungo periodo: un confronto fra salari annuali in tre Paesi sud-Europei	
<i>Ferretti C., Crosato L., Cipollini F., Ganugi P.</i>	
Design of Experiments, aberration and Market Basket Analysis	873
Pianificazione degli esperimenti, aberrazione e Market Basket Analysis	
<i>Roberto Fontana and Fabio Rapall</i>	
Generalized Procrustes Analysis for Multilingual Studies	879
Analisi Procrustiana Generalizzata per studi Multilingue	
<i>Alessia Forciniti, Michelangelo Misuraca, Germana Scepti, Maria Spano</i>	
Prior specification in flexible models	885
Specificazione delle prior in modelli flessibili	
<i>Maria Franco-Villoria, Massimo Ventrucci and Haavard Rue</i>	
Modeling Cyclists' Itinerary Choices: Evidence from a Docking Station-Based Bike-Sharing System	889
Un modello per gli itinerari dei ciclisti: risultati da un bike-sharing a stazioni fisse	
<i>S. T. Gaito - G. Manzi - G. Saibene - S. Salini - M. Zignani</i>	
A PARAFAC-ALS variant for fitting large data sets	895
Una variante del PARAFAC-ALS per approssimare data set di grandi dimensioni	
<i>Michele Gallo, Violetta Simonacci and Massimo Guarino</i>	
A Convex Mixture Model for Binomial Regression	901
Un modello mistura convessa per la Regressione Binomiale	
<i>Luisa Galtarossa and Antonio Canale</i>	
Blockchain as a universal tool for business improvement	907
Blockchain come strumento universale per il miglioramento del business	
<i>Massimiliano Giacalone, Diego Carmine Sinitò, Emilio Massa, Federica Oddo, Enrico Medda, Vito Santarcangelo</i>	
Seasonality in tourist flows: a decomposition of the change in seasonal concentration	913
La stagionalità nei flussi turistici: una scomposizione della variazione nella concentrazione stagionale	
<i>Luigi Grossi and Mauro Mussini</i>	
Are Real World Data the smart way of doing Health Analytics?	919
Real World Data: la base di una nuova ricerca clinica?	
<i>Francesca Ieva</i>	
Internet use and leisure activities: are all young people equal?	925
Internet e tempo libero: i giovani sono uguali tra loro?	
<i>Giuseppe Lamberti, Jordi Lopez Sintas and Pilar Lopez Belbeze</i>	
On a Family of Transformed Stochastic Orders	931
Su una famiglia di ordinamenti stocastici trasformati	
<i>Tommaso Lando and Lucio Bertoli-Barsotti</i>	

Bayesian stochastic search for Ising chain graph models	935
Ricerca stocastica Bayesiana per modelli grafici a catena Ising	
<i>Andrea Lazerini · Monia Lupporelli · Francesco C. Stingo</i>	
On the statistical design of parameters for variables sampling plans based on process capability index Cpk	941
Progettazione statistica dei parametri per il piano di campionamento per variabili basate sull'indice di capacità di processo Cpk	
<i>Antonio Lepore, Biagio Palumbo and Philippe Castagliola</i>	
Nowcasting foreign tourist arrivals using Google Trends: an application to the city of Florence, Italy	947
Nowcasting degli arrivi turistici stranieri usando Google Trends: un'applicazione nella città di Firenze, Italia	
<i>Alessandro Magrini</i>	
Inclusive growth in European countries: a cointegration analysis	953
La crescita inclusiva nei paesi europei: un'analisi di cointegrazione	
<i>Paolo Mariani, Andrea Marletta, Alessandra Michelangeli</i>	
ESCO- the European Labour Language: a conceptual and operational asset in support of labour governance in complex environments	959
ESCO il linguaggio europeo del lavoro: uno strumento concettuale ed operativo per le politiche del lavoro in contesti complessi	
<i>Cristilla Martelli, Laura Grassini, Adham Kahlawi, Maria Flora Salvatori, Lucia Buzzigoli</i>	
Hidden Markov Models for High Dimensional Data	965
Hidden Markov Models per dati ad alta dimensionalità	
<i>Martino, A., Guatteri, G., Paganoni, A.M.</i>	
Classification of Italian classes via bivariate semi parametric multilevel models	971
Classificazione delle classi italiane per mezzo di modelli bivariati a effetti misti semi parametrici	
<i>Chiara Masci, Francesca Ieva, Tommaso Agasisti and Anna Maria Paganoni</i>	
Data Mining Application to Healthcare Fraud Detection: Two-Step Unsupervised Clustering Method for Outlier Detection with Administrative Databases	977
Data Mining Applicato al Riconoscimento Frodi in Sanità: Algoritmo a Due Step per l'Identificazione di Outliers con Database Amministrativi	
<i>Massi Michela C., Ieva Francesca, Lettieri Emanuele</i>	
Multivariate analysis and biodiversity partitioning of a demersal fish community: an application to Lazio coast	985
Analisi multivariata e partizione della biodiversità di una comunità di specie demersali: un'applicazione alla costa laziale	
<i>M. Mingione, G. Jona Lasinio, S. Martino, F. Colloca</i>	
Latent Markov models with discrete separate cluster random effects on initial and transition probabilities	991
Modelli Latent Markov ad effetti casuali discreti e separati per le probabilità iniziali e di transizione	
<i>Giorgio E. Montanari and Marco Doretti</i>	
Unsuitability of likelihood-based asymptotic confidence intervals for Response-Adaptive designs in normal homoscedastic trials	997
Inadeguatezza degli intervalli di confidenza asintotici basati sulla verosimiglianza per disegni Response-Adaptive in caso di risposte normali omoschedastiche	
<i>Marco Novelli and Maroussa Zagoraïou</i>	
Local Hypothesis Testing for Functional Data: Extending False Discovery Rate to the Functional Framework	1003
Verifica locale delle ipotesi nell'ambito dei dati funzionali: estensione della nozione di False Discovery Rate al contesto funzionale	
<i>Niels Asken Lundtorp Olsen, Alessia Pini, and Simone Vantini</i>	
Educational mismatch and attitudes towards migration in Europe	1009
Disallineamento fra formazione e lavoro e atteggiamenti verso le migrazioni in Europa	
<i>Marco Guido Palladino and Emiliano Sironi</i>	
Soft thresholding Bayesian variable selection for compositional data analysis	1015
Selezione di Variabili Bayesiana con funzioni di soglia per l'analisi di dati di composizione	
<i>Matteo Pedone, Francesco C. Stingo</i>	
Sentiment-driven investment strategies: a practical example of AI-powered engines in a corporate setting	1021
Strategie d'investimento guidate dal sentiment: un esempio pratico di Intelligenza Artificiale in contesto aziendale	
<i>Mattia Pedrini, Sebastian Donoso, Enrico Deusebio, Nicola Donelli, Gabriele Arici, Andrea Cosentini, Paola Mosconi, Diego Ostinelli and Claudio Cocchis</i>	

Betting on football: a model to predict match outcomes	1027
Scommettere sul calcio: un nuovo modello per prevedere l'esito delle partite	
<i>Marco Petretta, Lorenzo Schiavon and Jacopo Diquigiovanni</i>	
Estimation of dynamic quantile models via the MM algorithm	1033
Stima di modelli Quantilici Dinamici con algoritmo MM	
<i>Fabrizio Poggioni, Mauro Bernardi, Lea Petrella</i>	
The decomposition by subpopulations of the Pietra index: an application to the professional football teams in Italy	1039
La scomposizione per sottopopolazioni dell'indice di Pietra: un'applicazione alle squadre professionistiche di calcio in Italia	
<i>Francesco Porro and Mariangela Zenga</i>	
An Object Oriented Data Analysis of Tweets: the Case of Queen Elizabeth Olympic Park .	1045
Object Oriented Data Analysis di Tweet: il caso del Queen Elizabeth Olympic Park	
<i>Paola Riva, Paola Sturla, Anna Calissano and Simone Vantini</i>	
Bias reduced estimation of a fixed effects model for Expected Goals in association football	1051
Stima non distorta di un modello Expected Goal con effetti fissi nel calcio	
<i>Lorenzo Schiavon and Nicola Sartori</i>	
Looking for Efficient Methods to Collect and Geolocalise Tweets	1057
Alla ricerca di metodi efficienti per raccogliere e geolocalizzare tweet	
<i>Stephan Schlosser, Daniele Toninelli and Silvia Fabris</i>	
Principal ranking profiles	1063
Principal ranking profiles	
<i>Mariangela Sciandra, Antonella Plaia</i>	
A statistical model for voting probabilities	1069
Un modello statistico per le probabilità di voto	
<i>Rosaria Simone, Stefania Capecchi</i>	
How Citizen Science and smartphones can help to produce timely and reliable information? Evidence from the "Food Price Crowdsourcing in Africa" (FPCA) project in Nigeria	1075
Citizen Science e smartphone posso aiutare nella raccolta di dati tempestivi e affidabili? Testimonianze del progetto "Food Price Crowdsourcing in Africa" (FPCA) condotto in Nigeria	
<i>Gloria Solano-Hermosilla, Fabio Micale, Vincenzo Nardelli, Julius Adewopo, Celso Gorrín González</i>	
Dealing with uncertainty in automated test assembly problems	1083
La gestione dell'incertezza nei problemi di assemblaggio automatizzato dei test	
<i>Giada Spaccapanico Proietti, Mariagiulia Matteucci and Stefania Mignani</i>	
Joint Models: a smart way to include functional data in healthcare analytics	1089
Modelli congiunti: un metodo per includere i dati funzionali nelle analisi in ambito sanitario	
<i>Marta Spreafico, Francesca Ieva</i>	
Bayesian multiscale mixture of Gaussian kernels for density estimation	1095
Stima di densità tramite misture bayesiane multiscala di kernel gaussiani	
<i>Marco Stefanucci and Antonio Canale</i>	
Dynamic Bayesian clustering of running activities	1101
Clustering Bayesiano dinamico di attività di corsa	
<i>Mattia Stival and Mauro Bernardi</i>	
Employment and fertility in couples: whose employment uncertainty matter most?	1107
Lavoro e fecondità in coppia: il ruolo dell'incertezza lavorativa secondo una prospettiva di genere	
<i>Valentina Tocchioni, Daniele Vignoli, Alessandra Mattei, Bruno Arpino</i>	
A Functional Data Analysis Approach to Study a Bike Sharing Mobility Network in the City of Milan	1113
<i>Agostino Torti, Alessia Pini and Simone Vantini</i>	
Multiresolution Topological Data Analysis for Robust Activity Tracking	1119
<i>Giovanni Trappolini, Tullia Padellini, and Pierpaolo Brutti</i>	
Semilinear regression trees	1125
Alberi di regressione semilineari	
<i>Giulia Vannucci and Anna Gottard</i>	

A models selection criterion for evaluation of heat wave hazard: a case study of the city of Prato.....	1131
Un criterio di selezione dei modelli per la valutazione della pericolosità delle ondate di calore: un caso studio della città di Prato	
<i>Veronica Villani, Giuliana Barbato, Elvira Romano and Paola Mercogliano</i>	
Digital Inequalities and ICT Devices: The ambiguous Role of Smartphones.....	1139
<i>Laura Zannella, Marina Zannella</i>	

Section 4. Posters

Modelling Hedonic Price using semiparametric M-quantile regression	1147
Regressione m-quantilica semiparametrica per la modellizzazione dei prezzi edonici	
<i>Riccardo Borgoni, Antonella Carcagni, Alessandra Michelangeli, Nicola Salvati</i>	
Bayesian mixed latent factor model for multi-response marine litter data with multi-source auxiliary information	1153
Modello bayesiano misto a fattori latenti per l'abbondanza di rifiuti marini con informazioni ausiliarie di diversa provenienza	
<i>Crescenza Calculli, Alessio Pollice, Marco V. Guglielmi and Porzia Maiorano</i>	
Official statistics to support the projects of A Scuola di OpenCoesione	1159
L'esperienza di monitoraggio civico in Lombardia nell'anno scolastico 2018-19	
<i>del Vicario G. and Di Gennaro L. and Ferrazza D. and Spinella V. and Viviano L.</i>	
Spatial Logistic Regression for Events Lying on a Network: Car Crashes in Milan.....	1165
Regressione logistica per eventi su network: gli incidenti automobilistici nel comune di Milano	
<i>Andrea Gilardi, Riccardo Borgoni and Diego Zappa</i>	
Variable selection and classification by the GRID procedure	1171
Selezione e classificazione delle variabili attraverso il metodo GRID	
<i>Francesco Giordano, Soumendra Nath Lahiri and Maria Lucia Parrella</i>	
Joint VaR and ES forecasting in a multiple quantile regression framework.....	1177
Stima congiunta del VaR e dell'ES attraverso la regressione quantilica multipla	
<i>Merlo Luca, Petrella Lea and Raponi Valentina</i>	
Approximate Bayesian Computation methods to model Multistage Carcinogenesis	1183
Metodi di Approximate Bayesian Computation per modellare la Cancerogenesi Multistadiale	
<i>Consuelo R. Nava, Cinzia Carota, Jordy Bollon, Corrado Magnani, Francesco Barone-Adesi</i>	
Co-clustering TripAdvisor data for personalized recommendations	1189
Co-clustering di dati TripAdvisor per un sistema di raccomandazioni personalizzato	
<i>Giulia Pascali, Alessandro Casa and Giovanna Menardi</i>	
Latent class analysis of endoreduplicated nuclei in confocal microscopy.....	1195
Analisi di classi latenti per dati di nuclei endoreduplicati tramite microscopia confocale	
<i>Ivan Sciascia ivan.sciascia@unito.it, Gennaro Carotenuto gennaro.carotenuto@unito.it, Andrea Genre andrea.genre@unito.it, Università di Torino Dipartimento di Scienze della vita e biologia dei sistemi, viale Mattioli 25, 10125 Torino</i>	

An extension of the censored gaussian lasso estimator

Un'estensione dello stimatore clgasso

Luigi Augugliaro and Gianluca Sottile and Veronica Vinciotti

Abstract The conditional glasso is one of the most used estimators for inferring genetic networks. Despite its diffusion, there are several fields in applied research where the limits of detection of modern measurement technologies make the use of this estimator theoretically unfounded, even when the assumption of a multivariate Gaussian distribution is satisfied. In this paper we propose an extension to censored data.

Abstract *Il conditional graphical lasso è uno degli stimatori più utilizzati per fare inferenza sulle reti genetiche. Nonostante la sua elevata diffusione, esistono parecchi campi applicativi dove i limiti degli strumenti di misurazione ne rendono teoricamente ingiustificato l'utilizzo, anche quando l'assunzione relativa alla distribuzione normale multivariata è soddisfatta.*

Key words: Censored data, Censored glasso estimator, Gaussian graphical model, glasso estimator.

1 Introduction

An important aim in genomics is to understand interactions among genes, characterized by the regulation and synthesis of proteins under internal and external signals. These relationships can be represented by a genetic network, i.e., a graph where

Luigi Augugliaro
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: luigi.augugliaro@unipa.it

Gianluca Sottile
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: gianluca.sottile@unipa.it

Veronica Vinciotti
Department of Mathematics, Brunel University, UK, e-mail: veronica.vinciotti@brunel.ac.uk

nodes represent genes and edges describe the interactions among them. Gaussian graphical models [3] have been widely used for reconstructing a genetic network from expression data. The reason of such diffusion relies on the statistical properties of the multivariate Gaussian distribution which allow the topological structure of a network to be related with the non-zero elements of the concentration matrix, i.e., the inverse of the covariance matrix. Thus, the problem of network inference can be recast as the problem of estimating a concentration matrix. The conditional glasso estimator [8] is a popular method for estimating a sparse concentration matrix, based on the idea of adding an ℓ_1 -penalty function to the likelihood function of the multivariate Gaussian distribution.

Despite the widespread literature on the conditional glasso estimator, there is a great number of fields in applied research where modern measurement technologies make the use of this graphical model theoretically unfounded, even when the assumption of a multivariate Gaussian distribution is satisfied. A first example of this is Reverse Transcription quantitative Polymerase Chain Reaction (RT-qPCR), a popular technology for gene expression profiling. This technique relies on fluorescence-based detection of amplicon DNA and allows the kinetics of PCR amplification to be monitored in real time, making it possible to quantify nucleic acids with extraordinary ease and precision. The analysis of the raw RT-qPCR profiles is based on the cycle-threshold, defined as the fractional cycle number in the log-linear region of PCR amplification in which the reaction reaches fixed amounts of amplicon DNA. If a target is not expressed or the amplification step fails, the threshold is not reached after the maximum number of cycles (limit of detection) and the corresponding cycle-threshold is undetermined. For this reason, the resulting data is naturally right-censored data. In this paper we propose an extension of the conditional glasso estimator that takes into account the censoring mechanism of the data explicitly.

2 The conditional censored Gaussian graphical model

Let $\mathbf{Y} = (Y_1, \dots, Y_p)^\top$ be a p -dimensional random vector. Graphical models allow to represent the set of conditional independencies among these random variables by a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where \mathcal{V} is the set of nodes associated to \mathbf{Y} and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of ordered pairs, called edges, representing the conditional dependencies among the p random variables [3]. The conditional Gaussian graphical model is an extension of the classical Gaussian graphical model based on the assumption that the conditional distribution of \mathbf{Y} given a q -dimensional vector of predictors, say $\mathbf{X} = (X_1, \dots, X_q)^\top$, follows a multivariate Gaussian distribution with expected value:

$$\boldsymbol{\mu}(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \mathbf{x},$$

where $\boldsymbol{\beta} = (\beta_{hk})$ is a matrix $q \times p$ coefficient matrix, and covariance matrix denoted by $\boldsymbol{\Sigma} = (\sigma_{hk})$. Denoting with $\boldsymbol{\Theta} = (\theta_{hk})$ the concentration matrix, i.e., the inverse

of the covariance matrix, the conditional density function of \mathbf{Y} can be written as follows:

$$\phi(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\beta}, \Theta) = (2\pi)^{-p/2} |\Theta|^{1/2} \exp[-1/2 \{\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\}^\top \Theta \{\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\}]. \quad (1)$$

As shown in [3], the off-diagonal elements of the concentration matrix are the parametric tools relating the pairwise Markov property to the factorization of the density (1). Formally, two random variables, say Y_h and Y_k , are conditionally independent given all the remaining variables if and only if θ_{hk} is equal to zero. This result provides a simple way to relate the topological structure of the graph \mathcal{G} to the pairwise Markov property, i.e., the undirected edge (h, k) is an element of the edge set \mathcal{E} if and only if $\theta_{hk} \neq 0$,

As done in [1], we assume that \mathbf{Y} is a (partially) latent random vector with density function (1). In order to include the censoring mechanism inside our framework, let us denote by $\mathbf{l} = (l_1, \dots, l_p)^\top$ and $\mathbf{u} = (u_1, \dots, u_p)^\top$, with $l_h < u_h$ for $h = 1, \dots, p$, the vectors of known left and right censoring values. Thus, Y_h is observed only if it is inside the interval $[l_h, u_h]$ otherwise it is censored from below if $Y_h < l_h$ or censored from above if $Y_h > u_h$. Under this setting, a rigorous definition of the joint distribution of the observed data can be obtained using the approach for missing data with nonignorable mechanism [4]. This requires the specification of the distribution of a p -dimensional random vector, denoted by $R(\mathbf{Y}; \mathbf{l}, \mathbf{u})$, used to encode the censoring patterns. Formally, the h th element of $R(\mathbf{Y}; \mathbf{l}, \mathbf{u})$ is defined as $R(Y_h; l_h, u_h) = I(Y_h > u_h) - I(Y_h < l_h)$, where $I(\cdot)$ denotes the indicator function. By construction $R(\mathbf{Y}; \mathbf{l}, \mathbf{u})$ is a discrete random vector with support the set $\{-1, 0, 1\}^p$ and probability function $\Pr\{R(\mathbf{Y}; \mathbf{l}, \mathbf{u}) = \mathbf{r}\} = \int_{D_{\mathbf{r}}} \phi(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\beta}, \Theta) d\mathbf{y}$, where $D_{\mathbf{r}} = \{\mathbf{y} \in \mathbb{R}^p : R(\mathbf{y}; \mathbf{l}, \mathbf{u}) = \mathbf{r}\}$.

Given a censoring pattern, we can simplify our notation by partitioning the set $\mathcal{S} = \{1, \dots, p\}$ into $o = \{h \in \mathcal{S} : r_h = 0\}$, $c^- = \{h \in \mathcal{S} : r_h = -1\}$ and $c^+ = \{h \in \mathcal{S} : r_h = +1\}$ and, in the following of this paper, we shall use the convention that a vector indexed by a set of indices denotes the corresponding subvector. For example, the subvector of observed elements in \mathbf{y} is denoted by $\mathbf{y}_o = (y_h)_{h \in o}$ and, consequently, the observed data is the vector $(\mathbf{y}_o^\top, \mathbf{x}^\top, \mathbf{r}^\top)^\top$. As done in [1], the probability distribution of the observed data, denoted by $\varphi(\{\mathbf{y}_o, \mathbf{r}\} \mid \mathbf{x}; \boldsymbol{\beta}, \Theta)$, can be defined as follows:

$$\varphi(\{\mathbf{y}_o, \mathbf{r}\} \mid \mathbf{x}; \boldsymbol{\beta}, \Theta) = \int \phi(\{\mathbf{y}_o, \mathbf{y}_c\} \mid \mathbf{x}; \boldsymbol{\beta}, \Theta) \Pr\{R(\mathbf{Y}; \mathbf{l}, \mathbf{u}) = \mathbf{r} \mid \mathbf{Y} = \mathbf{y}\} d\mathbf{y}_c, \quad (2)$$

where $c = c^- \cup c^+$.

Density (2) can be simplified by observing that $\Pr\{R(\mathbf{Y}; \mathbf{l}, \mathbf{u}) = \mathbf{r} \mid \mathbf{Y} = \mathbf{y}\}$ is equal to one if the censoring pattern encoded in \mathbf{r} is equal to the pattern observed in \mathbf{y} , otherwise it is equal to zero, i.e.,

$$\Pr\{R(\mathbf{Y}; \mathbf{l}, \mathbf{u}) = \mathbf{r} \mid \mathbf{Y} = \mathbf{y}\} = I(\mathbf{y}_{c^-} < \mathbf{l}_{c^-}) I(\mathbf{l}_o \leq \mathbf{y}_o \leq \mathbf{u}_o) I(\mathbf{u}_{c^+} < \mathbf{y}_{c^+}),$$

where the inequalities in the previous expressions are intended elementwise. From this, $\varphi(\{\mathbf{y}_o, \mathbf{r}\} | \mathbf{x}; \boldsymbol{\beta}, \Theta)$ can be rewritten as

$$\varphi(\{\mathbf{y}_o, \mathbf{r}\} | \mathbf{x}; \boldsymbol{\beta}, \Theta) = \int_{D_c} \phi(\{\mathbf{y}_o, \mathbf{y}_c\} | \mathbf{x}; \boldsymbol{\beta}, \Theta) d\mathbf{y}_c I(\mathbf{l}_o \leq \mathbf{y}_o \leq \mathbf{u}_o), \quad (3)$$

where $D_c = (-\infty, \mathbf{l}_{c^-}) \times (\mathbf{u}_{c^+}, +\infty)$. Using density (3), the conditional censored Gaussian graphical model is defined as the set $\{\mathbf{Y}, R(\mathbf{Y}; \mathbf{l}, \mathbf{u}), \varphi(\{\mathbf{y}_o, \mathbf{r}\} | \mathbf{x}; \boldsymbol{\beta}, \Theta), \mathcal{G}\}$, where $\varphi(\{\mathbf{y}_o, \mathbf{r}\} | \mathbf{x}; \boldsymbol{\beta}, \Theta)$ factorizes according to the undirected graph \mathcal{G} .

3 The conditional censored glasso estimator

Suppose we have a sample of size n independent observations drawn from a conditional censored Gaussian graphical model. For ease of exposition, we shall assume that \mathbf{l} and \mathbf{u} are fixed across the n observations, but the extension to the cases where the censoring vectors are specific to each observation is straightforward and does not require a specific treatment. To simplify our notation the set of indices of the variables observed in the i th observation is denoted by $o_i = \{h \in \mathcal{J} : r_{ih} = 0\}$, while $c_i^- = \{h \in \mathcal{J} : r_{ih} = -1\}$ and $c_i^+ = \{h \in \mathcal{J} : r_{ih} = +1\}$ denote the sets of indices associated to the left and right-censored data, respectively. Denoting by \mathbf{r}_i the realization of the random vector $R(\mathbf{Y}_i; \mathbf{l}, \mathbf{u})$, the i th observed data is the vector $(\mathbf{y}_{io_i}^\top, \mathbf{x}_i^\top, \mathbf{r}_i^\top)^\top$. Using the density function (3), the observed log-likelihood function can be written as

$$\ell(\boldsymbol{\beta}, \Theta) = \sum_{i=1}^n \log \int_{D_{c_i}} \phi(\{\mathbf{y}_{io_i}, \mathbf{y}_{ic_i}\} | \mathbf{x}_i; \boldsymbol{\beta}, \Theta) d\mathbf{y}_{ic_i} = \sum_{i=1}^n \log \varphi(\{\mathbf{y}_{io_i}, \mathbf{r}_i\} | \mathbf{x}_i; \boldsymbol{\beta}, \Theta), \quad (4)$$

where $D_{c_i} = (-\infty, \mathbf{l}_{c_i^-}) \times (\mathbf{u}_{c_i^+}, +\infty)$ and $c_i = c_i^- \cup c_i^+$. Although inference about the parameters of this model can be carried out via the maximum likelihood method, the application of this inferential procedure to real datasets is limited for three main reasons. Firstly, the number of measured variables is often larger than the sample size and this implies the non-existence of the maximum likelihood estimator even when the dataset is fully observed. Secondly, even when the sample size is large enough, the maximum likelihood estimator will exhibit a very high variance [5, 7]. Thirdly, empirical evidence suggests that gene networks or more general biochemical networks are not fully connected [2]. In terms of conditional Gaussian graphical models this evidence translates in the assumption that $\boldsymbol{\beta}$ and Θ have a sparse structure, i.e., only few regression coefficients and few θ_{hk} are different from zero.

All that considered, we propose to estimate the parameters of the conditional censored Gaussian graphical model by generalizing the approach proposed in [8], i.e., by maximizing a new objective function defined by adding two lasso-type penalty functions to the observed log-likelihood (4). The resulting estimator, called conditional censored glasso estimator, is formally defined as

An extension of the censored gaussian lasso estimator

$$\{\hat{\boldsymbol{\beta}}^\lambda, \hat{\Theta}^\rho\} = \arg \max_{\boldsymbol{\beta}, \Theta > 0} \frac{1}{n} \sum_{i=1}^n \log \varphi(\{\mathbf{y}_{i\circ_i}, \mathbf{r}_i\} | \mathbf{x}_i; \boldsymbol{\beta}, \Theta) - \lambda \sum_{h,k} |\beta_{hk}| - \rho \sum_{h \neq k} |\theta_{hk}|, \quad (5)$$

where λ and ρ are two non-negative tuning parameters. The lasso penalty on $\boldsymbol{\beta}$ introduces sparsity in $\hat{\boldsymbol{\beta}}^\lambda$, in other words by varying λ we can select the relevant predictors for \mathbf{Y} . Like in the standard glasso estimator, the tuning parameter ρ controls the amount of sparsity in the estimated concentration matrix $\hat{\Theta}^\rho = (\hat{\theta}_{hk}^\rho)$ and, consequently, in the corresponding estimated graph $\hat{\mathcal{G}}^\rho = \{\mathcal{V}, \hat{\mathcal{E}}^\rho\}$, where $\hat{\mathcal{E}}^\rho = \{(h, k) : \hat{\theta}_{hk}^\rho \neq 0\}$. When ρ is large enough, some $\hat{\theta}_{hk}^\rho$ are shrunk to zero resulting in the removal of the corresponding link in $\hat{\mathcal{G}}^\rho$; on the other hand, when ρ is equal to zero and the sample size is large enough the estimator $\hat{\Theta}^\rho$ coincides with the maximum likelihood estimator of the concentration matrix, which implies a fully connected estimated concentration graph.

4 Simulation study

In this section, we compare our proposed estimator with MissGlasso [6], which performs ℓ_1 -penalized estimation under the assumption that the censored data are missing at random, and with the conditional glasso estimator [8], where the empirical covariance matrix is calculated by imputing the missing values with the censoring values. These estimators are evaluated in terms of both recovering the structure of the true graph and the mean squared error. We use the method implemented in the R package huge [9], to simulate a sparse concentration matrix with a random structure for \mathbf{Y} . In particular, we set the probability of observing a link between two nodes to k/p , where p is the number of responses and k is used to control the amount of sparsity in Θ . Moreover, we set the right censoring value to 40 for any variable and the sample size n to 100. The predictors matrix \mathbf{X} is sampled from a multivariate gaussian distribution with zero expected value and sparse covariance matrix simulated as done for \mathbf{Y} . Each column the true matrix of predictors $\boldsymbol{\beta}$ contains only two non-zero regression coefficients, the values are sampled from a uniform distribution on the interval $[0.3, 0.7]$. The values of the intercepts are chosen in such a way that H response variables are right censored with probability equal to 0.40. The quantities k, p, q and H are used to specify the different scenarios used to analyze the behavior of the considered estimators. In particular, we consider the following cases:

- **Scenario 1:** $k = 3, p = 50, q = 10$ and $H = 25$. This setting is used to evaluate the effects of the number of censored variables on the behavior of the proposed estimators when $n > p$.
- **Scenario 2:** $k = 3, p = 150, q = 10$ and $H = 75$. This setting is used to evaluate the impact of the high dimensionality on the estimators ($p > n$).

For each scenario, we simulate 50 samples and in each simulation, we compute the coefficients path using cglasso, MissGlasso, and glasso. Each path is computed

using an equally spaced sequence of ρ and λ -values. However, the two scenarios differ also on the length of the two sequences, that is 20 for the Scenario 1 and 10 for the Scenario 2. Moreover, the precision-recall curves and the area under the curves (AUCs) are computed for each Scenarios.

The curves report the relationship between precision and recall for any ρ and λ -value, which are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

where TP, FP, and FN are quantities defined as the number correctly selected non-null items, the number of wrong selected non-null items and the number of wrong selected null item, respectively. Table 1 shows how cglasso gives a better estimate of the concentration and coefficient matrices in terms of AUCs, for any given value of the tuning parameters. We report only five evenly spaced values of λ and ρ . Figures 1 and 2 show the averages of the quantities $\min_{\rho} \text{MSE}(\hat{\Theta}_{\rho})$ and $\min_{\lambda} \text{MSE}(\hat{\beta}_{\lambda})$, which gives the minimum value of the mean squared error attained along the path of solutions. These plots emphasize that cglasso has also a mean squared error much smaller than the considered competitors.

Table 1 Mean area under the curves across the sequence of ρ and λ -values under the specification of the two Scenarios. The first column block refers to the concentration matrix (Θ) when λ is fixed and the second refers to the coefficient matrix (β) when ρ is fixed.

	λ/λ_{\max}					ρ/ρ_{\max}				
	0.00	0.25	0.50	0.75	1.00	0.00	0.25	0.50	0.75	1.00
Model 1 cglasso	0.546	0.429	0.139	0.103	0.101	0.844	0.877	0.883	0.882	0.885
MissGlasso	0.239	0.199	0.086	0.073	0.073	0.745	0.764	0.766	0.767	0.768
glasso	0.414	0.218	0.097	0.092	0.091	0.813	0.847	0.864	0.866	0.866
Model 2 cglasso	0.418	0.094	0.037	0.035	0.035	0.794	0.930	0.931	0.929	0.933
MissGlasso	0.329	0.098	0.033	0.031	0.030	0.753	0.830	0.831	0.830	0.831
glasso	0.321	0.040	0.033	0.032	0.031	0.751	0.902	0.906	0.907	0.907

5 Conclusions

In this paper, we have proposed an extension of the conditional glasso estimator to multivariate censored data. A simulation study showed that the proposed estimator overcomes the existing estimators both in terms of parameter estimation and of network recovery.

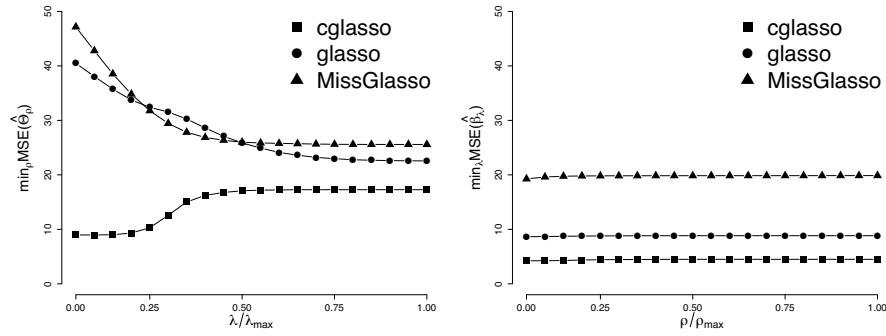


Fig. 1 Average of the minimum mean squared error attained along the path of solutions under the specification of Scenario 1. Left panel refers to the precision matrix for each fixed value of the tuning parameter λ ; right panel refers to the coefficient matrix for each fixed value of the tuning parameter ρ . The square refers to the cglasso estimator, the circle to the glasso estimator and the triangle to MissGlasso estimator.

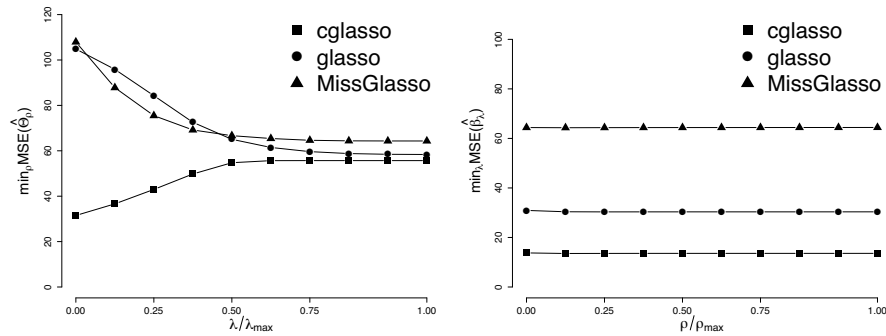


Fig. 2 Average of the minimum mean squared error attained, along the path of solutions under the specification of Model 2.

References

1. Augugliaro L., Abbruzzo A., Vinciotti V.: ℓ_1 -Penalized censored Gaussian graphical model. *Biostatistics*. <https://doi.org/10.1093/biostatistics/kxy043>
2. Gardner T. S., di Bernardo D., Lorenz D., Collins J. J.: Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*. **301**, 102–105 (2003)
3. Lauritzen S. L.: *Graphical Models*. Oxford University Press, Oxford (1996)
4. Little R. J. A., Rubin D. B.: *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., Hoboken (2002)
5. Schäfer J., Strimmer, K.: A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*. **4**(1). (2005)
6. Städler, N., Bühlmann, P.: Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Stat. Comput.* **22**(1), 219–235 (2012)

7. Uhler C.: Geometry of maximum likelihood estimation in Gaussian graphical models. *Ann. Statist.* **40**(12), 238–261 (2012)
8. Yin J., Li H.: A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Statist.* **5**(4), 2630–2650 (2011)
9. Zhao T., Li X., Liu H., Roeder K., Lafferty J., Wasserman L.: huge: High-Dimensional Undirected Graph Estimation. R package version 1.2.7 (2015). <https://CRAN.R-project.org/package=huge>