



CLADAG 2021

BOOK OF ABSTRACTS AND SHORT PAPERS
13th Scientific Meeting of the Classification and Data Analysis Group
Firenze, September 9-11, 2021

edited by

Giovanni C. Porzio

Carla Rampichini

Chiara Bocci



PROCEEDINGS E REPORT

ISSN 2704-601X (PRINT) - ISSN 2704-5846 (ONLINE)

SCIENTIFIC PROGRAM COMMITTEE

Giovanni C. Porzio (chair) (University of Cassino and Southern Lazio - Italy)

Silvia Bianconcini (University of Bologna - Italy)

Christophe Biernacki (University of Lille - France)

Paula Brito (University of Porto - Portugal)

Francesca Marta Lilja Di Lascio (Free University of Bozen-Bolzano - Italy)

Marco Di Marzio ("Gabriele d'Annunzio" University of Chieti-Pescara - Italy)

Alessio Farcomeni ("Tor Vergata" University of Rome - Italy)

Luca Frigau (University of Cagliari - Italy)

Luis Ángel García Escudero (University of Valladolid - Spain)

Bettina Grün (Vienna University of Economics and Business - Austria)

Salvatore Ingrassia (University of Catania - Italy)

Volodymyr Melnykov (University of Alabama - USA)

Brendan Murphy (University College Dublin - Ireland)

Maria Lucia Parrella (University of Salerno - Italy)

Carla Rampichini (University of Florence - Italy)

Monia Ranalli (Sapienza University of Rome - Italy)

J. Sunil Rao (University of Miami - USA)

Marco Riani (University of di Parma - Italy)

Nicola Salvati (University of Pisa - Italy)

Laura Maria Sangalli (Polytechnic University of Milan - Italy)

Bruno Scarpa (University of Padua - Italy)

Mariangela Sciandra (University of Palermo - Italy)

Luca Scrucca (University of Perugia - Italy)

Domenico Vistocco (Federico II University of Naples - Italy)

Mariangela Zenga (University of Milan-Bicocca - Italy)

LOCAL PROGRAM COMMITTEE

Carla Rampichini (chair) (University of Florence - Italy)

Chiara Bocci (University of Florence - Italy)

Anna Gottard (University of Florence - Italy)

Leonardo Grilli (University of Florence - Italy)

Monia Lupparelli (University of Florence - Italy)

Maria Francesca Marino (University of Florence - Italy)

Agnese Panzera (University of Florence - Italy)

Emilia Rocco (University of Florence - Italy)

Domenico Vistocco (Federico II University of Naples - Italy)

CLADAG 2021
BOOK OF ABSTRACTS
AND SHORT PAPERS

13th Scientific Meeting of the Classification
and Data Analysis Group
Firenze, September 9-11, 2021

edited by
Giovanni C. Porzio
Carla Rampichini
Chiara Bocci

FIRENZE UNIVERSITY PRESS
2021

CLADAG 2021 BOOK OF ABSTRACTS AND SHORT PAPERS : 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021/ edited by Giovanni C. Porzio, Carla Rampichini, Chiara Bocci. — Firenze : Firenze University Press, 2021.
(Proceedings e report ; 128)

<https://www.fupress.com/isbn/9788855183406>

ISSN 2704-601X (print)

ISSN 2704-5846 (online)

ISBN 978-88-5518-340-6 (PDF)

ISBN 978-88-5518-341-3 (XML)

DOI 10.36253/978-88-5518-340-6

Graphic design: Alberto Pizarro Fernández, Lettera Meccanica SRLs

Front cover: Illustration of the statue by Giambologna, *Appennino* (1579-1580) by Anna Gottard



Classification and Data
Analysis Group (CLADAG)
of the Italian Statistical
Society (SIS)

FUP Best Practice in Scholarly Publishing (DOI https://doi.org/10.36253/fup_best_practice)

All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Boards of the series. The works published are evaluated and approved by the Editorial Board of the publishing house, and must be compliant with the Peer review policy, the Open Access, Copyright and Licensing policy and the Publication Ethics and Complaint policy.

Firenze University Press Editorial Board

M. Garzaniti (Editor-in-Chief), M.E. Alberti, F. Vittorio Arrigoni, E. Castellani, F. Ciampi, D. D'Andrea, A. Dolfi, R. Ferrise, A. Lambertini, R. Lanfredini, D. Lippi, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, A. Orlandi, I. Palchetti, A. Perulli, G. Pratesi, S. Scaramuzzi, I. Stolzi.

📖 The online digital edition is published in Open Access on www.fupress.com.

Content license: except where otherwise noted, the present work is released under Creative Commons Attribution 4.0 International license (CC BY 4.0: <http://creativecommons.org/licenses/by/4.0/legalcode>). This license allows you to share any part of the work by any means and format, modify it for any purpose, including commercial, as long as appropriate credit is given to the author, any changes made to the work are indicated and a URL link is provided to the license.

Metadata license: all the metadata are released under the Public Domain Dedication license (CC0 1.0 Universal: <https://creativecommons.org/publicdomain/zero/1.0/legalcode>).

© 2021 Author(s)

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

*This book is printed on acid-free paper
Printed in Italy*

INDEX

Preface	1
----------------	----------

Keynote Speakers

<i>Jean-Michel Loubes</i> Optimal transport methods for fairness in machine learning	5
<i>Peter Rousseeuw, Jakob Raymaekers and Mia Hubert</i> Class maps for visualizing classification results	6
<i>Robert Tibshirani, Stephen Bates and Trevor Hastie</i> Understanding cross-validation and prediction error	7
<i>Cinzia Viroli</i> Quantile-based classification	8
<i>Bin Yu</i> Veridical data science for responsible AI: characterizing V4 neurons through deepTune	9

Plenary Session

<i>Daniel Diaz</i> A simple correction for COVID-19 sampling bias	14
<i>Jeffrey S. Morris</i> A seat at the table: the key role of biostatistics and data science in the COVID-19 pandemic	15
<i>Bhramar Mukherjee</i> Predictions, role of interventions and the crisis of virus in India: a data science call to arms	16
<i>Danny Pfeffermann</i> Contributions of Israel's CBS to rout COVID-19	17

Invited Papers

<i>Claudio Agostinelli, Giovanni Saraceno and Luca Greco</i> Robust issues in estimating modes for multivariate torus data	21
<i>Emanuele Aliverti</i> Bayesian nonparametric dynamic modeling of psychological traits	25

<i>Andres M. Alonso, Carolina Gamboa and Daniel Peña</i> Clustering financial time series using generalized cross correlations	27
<i>Raffaele Argiento, Edoardo Filippi-Mazzola and Lucia Paci</i> Model-based clustering for categorical data via Hamming distance	31
<i>Antonio Balzanella, Antonio Irpino and Francisco de A.T. De Carvalho</i> Mining multiple time sequences through co-clustering algorithms for distributional data	32
<i>Francesco Bartolucci, Fulvia Pennoni and Federico Cortese</i> Hidden Markov and regime switching copula models for state allocation in multiple time-series	36
<i>Michela Battauz and Paolo Vidoni</i> Boosting multidimensional IRT models	40
<i>Matteo Bottai</i> Understanding and estimating conditional parametric quantile models	44
<i>Niklas Bussmann, Roman Enzmann, Paolo Giudici and Emanuela Raffinetti</i> Shapley Lorenz methods for eXplainable artificial intelligence	45
<i>Andrea Cappelozzo, Ludovic Duponchel, Francesca Greselin and Brendan Murphy</i> Robust classification of spectroscopic data in agri-food: first analysis on the stability of results	49
<i>Andrea Cerasa, Enrico Checchi, Domenico Perrotta and Francesca Torti</i> Issues in monitoring the EU trade of critical COVID-19 commodities	53
<i>Marcello Chiodi</i> Smoothed non linear PCA for multivariate data	54
<i>Roberto Colombi, Sabrina Giordano and Maria Kateri</i> Accounting for response behavior in longitudinal rating data	58
<i>Claudio Conversano, Giulia Contu, Luca Frigau and Carmela Cappelli</i> Network-based semi-supervised clustering of time series data	62
<i>Federica Cugnata, Chiara Brombin, Pietro Cippà, Alessandro Ceschi, Paolo Ferrari and Clelia Di Serio</i> Characterising longitudinal trajectories of COVID-19 biomarkers within a latent class framework	64
<i>Silvia D'Angelo</i> Sender and receiver effects in latent space models for multiplex data	68
<i>Anna Denkowska and Stanisław Wanat</i> DTW-based assessment of the predictive power of the copula-DCC-GARCH-MST model developed for European insurance institutions	71
<i>Roberto Di Mari, Zsuzsa Bakk, Jennifer Oser and Jouni Kuha</i> Two-step estimation of multilevel latent class models with covariates	75
<i>Marie Du Roy de Chaumaray and Matthieu Marbac</i> Clustering data with non-ignorable missingness using semi-parametric mixture models	79

<i>Pierpaolo D'Urso, Livia De Giovanni and Vincenzina Vitale</i> Spatial-temporal clustering based on B-splines: robust models with applications to COVID-19 pandemic	83
<i>Leonardo Egidi, Roberta Pappadà, Francesco Pauli and Nicola Torelli</i> PIVMET: pivotal methods for Bayesian relabelling in finite mixture models	87
<i>Tahir Ekin and Claudio Conversano</i> Cluster validity by random forests	91
<i>Luis Angel García-Escudero, Agustín Mayo-Iscar and Marco Riani</i> Robust estimation of parsimonious finite mixture of Gaussian models	92
<i>Silvia Facchinetti and Silvia Angela Osmetti</i> A risk indicator for categorical data	93
<i>Matteo Fasiolo</i> Additive quantile regression via the qgam R package	97
<i>Michael Fop, Dimitris Karlis, Ioannis Kosmidis, Adrian O'Hagan, Caitriona Ryan and Isobel Claire Gormley</i> Gaussian mixture models for high dimensional data using composite likelihood	98
<i>Carlo Gaetan, Paolo Girardi and Victor Muthama Musau</i> On model-based clustering using quantile regression	102
<i>Carlotta Galeone</i> Socioeconomic inequalities and cancer risk: myth or reality?	106
<i>Michael Gallagher, Christophe Biernacki and Paul McNicholas</i> Parameter-wise co-clustering for high dimensional data	107
<i>Francesca Greselin and Alina Jędrzejczak</i> Quantifying the impact of covariates on the gender gap measurement: an analysis based on EU-SILC data from Poland and Italy	108
<i>Alessandra Guglielmi, Mario Beraha, Matteo Giannella, Matteo Pegoraro and Riccardo Peli</i> A transdimensional MCMC sampler for spatially dependent mixture models	112
<i>Christian Hennig and Pietro Coretto</i> Non-parametric consistency for the Gaussian mixture maximum likelihood estimator	116
<i>Yinxuan Huang and Natalie Shlomo</i> Improving the reliability of a nonprobability web survey	120
<i>Maria Iannario and Claudia Tarantola</i> A semi-Bayesian approach for the analysis of scale effects in ordinal regression models	124
<i>Jayant Jha</i> Best approach direction for spherical random variables	128

<i>Maria Kateri</i>	
Simple effect measures for interpreting generalized binary regression models	129
<i>Shogo Kato, Kota Nagasaki and Wataru Nakanishi</i>	
Mixtures of Kato–Jones distributions on the circle, with an application to traffic count data	133
<i>John Kent</i>	
How to design a directional distribution	137
<i>Simona Korenjak-Černe and Nataša Kejžar</i>	
Identifying mortality patterns of main causes of death among young EU population using SDA approaches	141
<i>Fabrizio Laurini and Gianluca Morelli</i>	
Robust supervised clustering: some practical issues	142
<i>Daniela Marella and Danny Pfeffermann</i>	
A nonparametric approach for statistical matching under informative sampling and nonresponse	146
<i>Mariagiulia Matteucci and Stefania Mignani</i>	
Investigating model fit in item response models with the Hellinger distance	150
<i>Matteo Mazziotta and Adriano Pareto</i>	
PCA-based composite indices and measurement model	154
<i>Marcella Mazzoleni, Angiola Pollastri and Vanda Tulli</i>	
Gender inequalities from an income perspective	158
<i>Yana Melnykov, Xuwen Zhu and Volodymyr Melnykov</i>	
Transformation mixture modeling for skewed data groups with heavy tails and scatter	162
<i>Luca Merlo, Lea Petrella and Nikos Tzavidis</i>	
Unconditional M-quantile regression	163
<i>Jesper Møller, Mario Beraha, Raffaele Argiento and Alessandra Guglielmi</i>	
MCMC computations for Bayesian mixture models using repulsive point processes	167
<i>Keefe Murphy, Cinzia Viroli and Isobel Claire Gormley</i>	
Infinite mixtures of infinite factor analysers	168
<i>Stanislav Nagy, Petra Laketa and Rainer Dyckerhoff</i>	
Angular halfspace depth: computation	169
<i>Yarema Okhrin, Gazi Salah Uddin and Muhammad Yahya</i>	
Nonlinear Interconnectedness of crude oil and financial markets	173
<i>M. Rosário Oliveira, Ana Subtil and Lina Oliveira</i>	
Detection of internet attacks with histogram principal component analysis	174
<i>Sally Paganin</i>	
Semiparametric IRT models for non-normal latent traits	178

<i>Giuseppe Pandolfo</i>	
A graphical depth-based aid to detect deviation from unimodality on hyperspheres	182
<i>Panos Pardalos</i>	
Networks of networks	186
<i>Xanthi Pedeli and Cristiano Varin</i>	
Pairwise likelihood estimation of latent autoregressive count models	187
<i>Mark Reiser and Maduranga Dassanayake</i>	
A study of lack-of-fit diagnostics for models fit to cross-classified binary variables	191
<i>Giorgia Riveccio, Jean-Paul Chavas, Giovanni De Luca, Salvatore Di Falco and Fabian Capitanio</i>	
Assessing food security issues in Italy: a quantile copula approach	195
<i>Nicoleta Rogovschi</i>	
Co-clustering for high dimensional sparse data	199
<i>Massimiliano Russo</i>	
Malaria risk detection via mixed membership models	203
<i>Paula Saavedra-Nieves and Rosa M. Crujeiras</i>	
Nonparametric estimation of the number of clusters for directional data	207
<i>Shuchismita Sarkar, Volodymyr Melnykov and Xuwen Zhu</i>	
Tensor-variate finite mixture model for the analysis of university professor remuneration	208
<i>Florian Schuberth</i>	
Specifying composites in structural equation modeling: the Henseler-Ogasawara specification	209
<i>Jarod Smith, Mohammad Arashi and Andriette Bekker</i>	
Network analysis implementing a mixture distribution from Bayesian viewpoint	210
<i>Paul Smith, Peter van der Heijden and Maarten Cruyff</i>	
Measurement errors in multiple systems estimation	211
<i>Valentin Todorov and Peter Filzmoser</i>	
Robust classification in high dimensions using regularized covariance estimates	215
<i>Salvatore Daniele Tomarchio, Luca Bagnato and Antonio Punzo</i>	
Clustering via new parsimonious mixtures of heavy tailed distributions	216
<i>Agostino Torti, Marta Galvani, Alessandra Menafoglio, Piercesare Secchi and Simone Vantini</i>	
A general bi-clustering technique for functional data	217
<i>Laura Trinchera</i>	
Developing a multidimensional and hierarchical index following a composite-based approach	220

<i>Rosanna Verde, Francisco T. de A. De Carvalho and Antonio Balzanella</i> A generalised clusterwise regression for distributional data	223
<i>Marika Vezzoli, Francesco Doglietto, Stefano Renzetti, Marco Fontanella and Stefano Calza</i> A machine learning approach for evaluating anxiety in neurosurgical patients during the COVID-19 pandemic	227
<i>Isadora Antoniano Villalobos, Simone Padoan and Boris Beranger</i> Prediction of large observations via Bayesian inference for extreme-value theory	231
<i>Maria Prosperina Vitale, Vincenzo Giuseppe Genova, Giuseppe Giordano and Giancarlo Ragozini</i> Community detection in tripartite networks of university student mobility flows	232
<i>Ernst Wit and Lucas Kania</i> Causal regularization	236
<i>Qiuyi Wu and David Banks</i> Minimizing conflicts of interest: optimizing the JSM program	240

Contributed Papers

<i>Antonino Abbruzzo, Maria Francesca Cracolici and Furio Urso</i> Model selection procedure for mixture hidden Markov models	243
<i>Roberto Ascari and Sonia Migliorati</i> A full mixture of experts model to classify constrained data	247
<i>Luigi Augugliaro, Gianluca Sottile and Angelo Mineo</i> Sparse inference in covariate adjusted censored Gaussian graphical models	251
<i>Simona Balzano, Mario Rosario Guarracino and Giovanni Camillo Porzio</i> Semi-supervised learning through depth functions	255
<i>Lucio Barabesi, Andrea Cerasa, Andrea Cerioli and Domenico Perrotta</i> A combined test of the Benford hypothesis with anti-fraud applications	256
<i>Chiara Bardelli</i> Unbalanced classification of electronic invoicing	260
<i>Claudia Berloco, Raffaele Argiento and Silvia Montagna</i> Predictive power of Bayesian CAR models on scale free networks: an application for credit risk	264
<i>Marco Berrettini, Giuliano Galimberti and Saverio Ranciati</i> Semiparametric finite mixture of regression models with Bayesian P-splines	268

<i>Giuseppe Bove</i>	
A subject-specific measure of interrater agreement based on the homogeneity index	272
<i>Antonio Calcagni</i>	
Estimating latent linear correlations from fuzzy contingency tables	276
<i>Andrea Cappozzo, Alessandro Casa and Michael Fop</i>	
Model-based clustering with sparse matrix mixture models	280
<i>Andrea Cappozzo, Luis Angel Garcia Escudero, Francesca Greselin and Agustín Mayo-Iscar</i>	
Exploring solutions via monitoring for cluster weighted robust models	284
<i>Maurizio Carpita and Silvia Golia</i>	
Categorical classifiers in multi-class classification problems	288
<i>Gianmarco Caruso, Greta Panunzi, Marco Mingione, Pierfrancesco Alaimo Di Loro, Stefano Moro, Edoardo Bompiani, Caterina Lanfredi, Daniela Silvia Pace, Luca Tardella and Giovanna Jona Lasinio</i>	
Model-based clustering for estimating cetaceans site-fidelity and abundance	292
<i>Carlo Cavicchia, Maurizio Vichi and Giorgia Zaccaria</i>	
Model-based clustering with parsimonious covariance structure	296
<i>Francesca Condino</i>	
Clustering income data based on share densities	300
<i>Paula Costa Fontichiarì, Miriam Giuliani, Raffaele Argiento and Lucia Paci</i>	
Group-dependent finite mixture model	304
<i>Salvatore Cuomo, Federico Gatta, Fabio Giampaolo, Carmela Iorio and Francesco Piccialli</i>	
A machine learning approach in stock risk management	308
<i>Cristina Davino and Giuseppe Lamberti</i>	
Pathmix segmentation trees to compare linear regression models	312
<i>Houyem Demni, Davide Buttarazzi, Stanislav Nagy and Giovanni Camillo Porzio</i>	
Angular halfspace depth: classification using spherical bagdistances	316
<i>Agostino Di Ciaccio</i>	
Neural networks for high cardinality categorical data	320
<i>F. Marta L. Di Lascio, Andrea Menapace and Roberta Pappadà</i>	
Ali-Mikhail-Haq copula to detect low correlations in hierarchical clustering	324
<i>Maria Veronica Dorgali, Silvia Bacci, Bruno Bertaccini and Alessandra Petrucci</i>	
Higher education and employability: insights from the mandatory notices of the ministry of labour	328
<i>Lorenzo Focardi Olmi and Anna Gottard</i>	
An alternative to joint graphical lasso for learning multiple Gaussian graphical models	332

<i>Francesca Fortuna, Alessia Naccarato and Silvia Terzi</i>	
Functional cluster analysis of HDI evolution in European countries	336
<i>Sylvia Frühwirth-Schnatter, Bettina Grün and Gertraud Malsiner-Walli</i>	
Estimating Bayesian mixtures of finite mixtures with telescoping sampling	340
<i>Chiara Galimberti, Federico Castelletti and Stefano Peluso</i>	
A Bayesian framework for structural learning of mixed graphical models	344
<i>Andrea Gilardi, Riccardo Borgoni, Luca Presicce and Jorge Mateu</i>	
Measurement error models on spatial network lattices: car crashes in Leeds	348
<i>Carmela Iorio, Giuseppe Pandolfo, Michele Staiano, Massimo Aria and Roberta Siciliano</i>	
The L^P data depth and its application to multivariate process control charts	352
<i>Petra Laketa and Stanislav Nagy</i>	
Angular halfspace depth: central regions	356
<i>Michele La Rocca, Francesco Giordano and Cira Perna</i>	
Clustering production indexes for construction with forecast distributions	360
<i>Maria Mannone, Veronica Distefano, Claudio Silvestri and Irene Poli</i>	
Clustering longitudinal data with category theory for diabetic kidney disease	364
<i>Laura Marcis, Maria Chiara Pagliarella and Renato Salvatore</i>	
A redundancy analysis with multivariate random-coefficients linear models	368
<i>Paolo Mariani, Andrea Marletta and Matteo Locci</i>	
The use of multiple imputation techniques for social media data	372
<i>Federico Marotta, Paolo Provero and Silvia Montagna</i>	
Prediction of gene expression from transcription factors affinities: an application of Bayesian non-linear modelling	376
<i>Francesca Martella, Fabio Attorre, Michele De Sanctis and Giuliano Fanelli</i>	
High dimensional model-based clustering of European georeferenced vegetation plots	380
<i>Ana Martins, Paula Brito, Sónia Dias and Peter Filzmoser</i>	
Multivariate outlier detection for histogram-valued variables	384
<i>Giovanna Menardi and Federico Ferraccioli</i>	
A nonparametric test for mode significance	388
<i>Massimo Mucciardi, Giovanni Pirrotta, Andrea Briglia and Arnaud Sallaberry</i>	
Visualizing cluster of words: a graphical approach to grammar acquisition	392

<i>Marta Nai Ruscone and Dimitris Karlis</i> Robustness methods for modelling count data with general dependence structures	396
<i>Roberta Paroli, Luigi Spezia, Marc Stutter and Andy Vinten</i> Bayesian analysis of a water quality high-frequency time series through Markov switching autoregressive models	400
<i>Mariano Porcu, Isabella Sulis and Cristian Usala</i> Detecting the effect of secondary school in higher education university choices	404
<i>Roberto Rocci and Monia Ranalli</i> Semi-constrained model-based clustering of mixed-type data using a composite likelihood approach	408
<i>Annalina Sarra, Adelia Evangelista, Tonio Di Battista and Damiana Pieragostino</i> Antibodies to SARS-CoV-2: an exploratory analysis carried out through the Bayesian profile regression	412
<i>Theresa Scharl and Bettina Grün</i> Modelling three-way RNA sequencing data with mixture of multivariate Poisson-lognormal distribution	416
<i>Luca Scrucca</i> Stacking ensemble of Gaussian mixtures	420
<i>Rosaria Simone, Cristina Davino, Domenico Vistocco and Gerhard Tutz</i> A robust quantile approach to ordinal trees	424
<i>Venera Tomaselli, Giulio Giacomo Cantone and Valeria Mazzeo</i> The detection of spam behaviour in review bomb	428
<i>Donatella Vicari and Paolo Giordani</i> Clustering models for three-way data	432
<i>Gianpaolo Zammarchi and Jaromir Antoch</i> Using eye-tracking data to create a weighted dictionary for sentiment analysis: the eye dictionary	436

GROUP-DEPENDENT FINITE MIXTURE MODEL

Paula Costa Fontichiarì¹, Miriam Giuliani¹, Raffaele Argiento¹ and Lucia Paci¹

¹ Department of Statistical Sciences, Università Cattolica del Sacro Cuore, (e-mail: paula.costafontichiarì01@icatt.it, miriam.giuliani01@icatt.it, raffaele.argiento@unicatt.it, lucia.paci@unicatt.it)

ABSTRACT: We present a Bayesian nonparametric group-dependent mixture model for clustering. This is achieved by building a hierarchical structure, where the discreteness of the shared base measure is exploited to cluster the data, between and within groups. We study the properties of the group-dependent clustering structure based on the latent parameters of the model. Furthermore, we obtain the joint distribution of the clustering induced by the hierarchical mixture model and define the complete posterior characterization of interest. We construct a Gibbs sampler to perform Bayesian inference and measure performances on simulated and a real data.

KEYWORDS: Bayesian analysis, clustering, Gibbs sampling, EPPF.

1 Introduction

In several statistical settings there is the need to model data organized in groups, allowing for sharing of information across them. In the Bayesian framework, this is achieved by hierarchical modeling, where the joint distribution of group-specific parameters accounts for such dependence. For instance, in Bayesian nonparametrics, the seminal work of Teh *et al.*, 2006 considered a mixture model within each group j , where the group-specific parameter is the mixing measure P_j and whose joint law is defined by an extra layer of hierarchy, yielding to the hierarchical Dirichlet process. This approach has been extended to the class of NRMI (Regazzini *et al.*, 2003) by Camerlenghi *et al.*, 2019 and Argiento *et al.*, 2020. In the cited works, the mixing measure is infinite dimensional.

In this work, we propose a hierarchical model where the group-specific mixing distribution belongs to the class of almost surely finite dimensional distributions introduced by Argiento & Iorio, 2019. We assign the joint law of the group-specific parameter such that the random measures within each group share the same support. In this framework, it is possible to define a

group-dependent clustering as follows. First, a latent parameter $\boldsymbol{\theta}_{j,i} \sim P_j$ for individual i and group j is introduced. Second, since P_j is almost surely discrete, ties within are expected, leading to a group-specific clustering. Finally, since the P_j 's share same support, we expect also ties between groups, providing a global clustering. We are able to derive the joint law of the group-specific clustering as well as the one of the global clustering. Such results allows to build up a posterior sampling strategy based on the Gibbs sampler.

2 Model developments

Let y_{ji} be the observed variable for group j , $j = 1, \dots, d$, and individual i , $i = 1, \dots, n_j$. We assume that the data in each group j come from a mixture of M components, that is

$$y_{j1}, \dots, y_{jn_j} \mid w_{jl}, \boldsymbol{\tau}_l, M \sim \sum_{l=1}^M w_{jl} f(y_{ji} \mid \boldsymbol{\tau}_l), \quad (1)$$

where $f(y_{ji} \mid \boldsymbol{\tau}_l)$ is called kernel and is a parametric density over the sampling space, w_{jm} are the group-specific mixing weights and $\boldsymbol{\tau}_l$ are the kernel parameters that are shared across groups. We assign a prior distribution on the mixing weights by normalization, namely we define $w_{jl} = \frac{S_{jl}}{T_j}$, where $T_j = \sum_{l=1}^M S_{jl}$. Also, we assume a prior distribution on the number of components, i.e., $M \sim q(m)$. Conditionally on M , S_{jl} are independent positive random variables with distribution $h_j(s)$, while $\boldsymbol{\tau}_l$ follows a prior distribution over Θ , the parameter space of the kernel, that we denote $p_0(\boldsymbol{\tau})$.

As in Argiento & Iorio, 2019, the model can be framed in a Bayesian non-parametric fashion. Indeed, $q(M)$, $h_j(s)$ and $p_0(\boldsymbol{\tau})$ define the joint distribution of a vector of almost sure discrete random measures P_1, \dots, P_d with support Θ , where

$$P_j = \sum_{l=1}^M \frac{S_{jl}}{T_j} \delta_{\boldsymbol{\tau}_l}(\boldsymbol{\theta}), \quad j = 1, \dots, d \quad (2)$$

with $\boldsymbol{\theta} \in \Theta$. We refer the joint distribution of P_1, \dots, P_d to as the Vector Normalized Independent weights, i.e., $V - NIw(q, h_j, p_0)$. Model (1) and the priors described above can be rewritten in a hierarchical form as follows:

$$\begin{aligned} y_{ji} \mid \boldsymbol{\theta}_{ji} &\stackrel{\text{ind}}{\sim} f(y_{ji} \mid \boldsymbol{\theta}_{ji}) \\ \boldsymbol{\theta}_{j1}, \dots, \boldsymbol{\theta}_{jn_j} \mid P_j &\stackrel{\text{iid}}{\sim} P_j \\ P_1, \dots, P_d \mid q, h, p_0 &\sim V - NIw(q, h_j, p_0). \end{aligned} \quad (3)$$

In this work, the kernel $f(y | \boldsymbol{\theta})$ represents the density of a univariate normal distribution with parameter $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$. We assume $q(m)$ to be the p.m.f. of a 1-shifted Poisson distribution with parameter Λ and $h_j(s)$ is the density of a gamma distribution with shape parameter γ_i and rate equal to 1. Finally, $p_0(\boldsymbol{\tau})$ is the density of a conjugate normal inverse gamma prior with parameters μ_0 , κ_0 , ν_0 and σ_0^2 .

3 Group-dependent clustering

The hierarchical model in (3) allows to define a group-dependent clustering based on the latent variables $\boldsymbol{\theta}_{ji}$. First, we introduce latent allocation variables c_{ji} such that $c_{ji} = m$ if $\boldsymbol{\theta}_{ji} = \boldsymbol{\tau}_m$. Then, we denote $\mathcal{M}^{(a)}$ the set of couples (j, m) such that $\exists i$ for which $c_{ij} = m$ and we define the number of *allocated columns* as

$$M^{(a)} = \# \left\{ m : \text{there exists one couple } (j, m) \in \mathcal{M}^{(a)}, j = 1, \dots, d \right\}.$$

We denote $\mathcal{M}^{(na)}$ the complement of $\mathcal{M}^{(a)}$. Hence, for every pair (j, m) , we define $n_{jm} = \#\{(j, i) : c_{ji} = m\}$. Note that

$$(j, m) \in \mathcal{M}^{(na)} \Rightarrow n_{jm} = 0$$

$$(j, m) \in \mathcal{M}^{(a)} \Rightarrow n_{jm} \geq 0.$$

Finally, let $c_1^*, \dots, c_{M^{(a)}}^*$ be the allocated columns, that is, the indexes within $\{1, \dots, M\}$ such that $(j, c_k^*) \in \mathcal{M}^{(a)}$.

We are now ready to define, for each group j , the clustering $\rho_j = \{A_{j1}, \dots, A_{jM^{(a)}}\}$, where $A_{jk} = \{(j, i) : (j, c_{ki}^*) \in \mathcal{M}^{(a)}\}$ and $k = 1, \dots, M^{(a)}$. In other words, A_{jk} is the set of data points of group j belonging to the k -th cluster. Note that, a distinctive feature of our setting, is that A_{jk} can be an empty set. Nevertheless, if $A_{jk} = \emptyset$ appears in ρ_j , it means that there is at least another group \tilde{j} such that $A_{\tilde{j}k}$ is not empty.

We build upon the work Argiento & Iorio, 2019 and James *et al.*, 2009 to derive the joint distribution of the clustering ρ_1, \dots, ρ_d , induced by the hierar-

chical mixture model (3). This turns out to be:

$$\begin{aligned} \pi(\rho_1, \dots, \rho_d, M^{(a)}) &= \int_0^\infty \dots \int_0^\infty \prod_{j=1}^d \frac{1}{\Gamma(n_j)} u_j^{n_j-1} \prod_{k=1}^{M^{(a)}} \kappa_{\gamma_j}(n_{jk}, u_j) \\ &\quad \exp \left[-\Lambda \left(\prod_{j=1}^d \Psi_{\gamma_j}(u_j) - 1 \right) \right] \\ &\quad \Lambda^{M^{(a)}-1} \left[\Lambda \prod_{j=1}^d \Psi_{\gamma_j}(u_j) + M^{(a)} \right] du_1 \dots du_d, \end{aligned} \quad (4)$$

where $\Psi_{\gamma_j}(u_j) = \frac{1}{(u_j+1)^{\gamma_j}}$ is the Laplace transform of a gamma distribution with shape γ_j and rate equal to 1, while $\kappa_{\gamma_j}(n_{jk}, u_j) = \frac{\Gamma(\gamma_j+n_{jk})}{\Gamma(\gamma_j)} \frac{1}{(u_j+1)^{n_{jk}+\gamma_j}}$ is its relative cumulant function. The joint distribution in (4) enables us to build a Gibbs sampler for sampling from the full posterior distribution. We omit here the details for brevity. We will illustrate the performance of our model over a set of simulated and real data.

References

- ARGIENTO, RAFFAELE, & IORIO, MARIA DE. 2019. Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *arXiv: Methodology*.
- ARGIENTO, RAFFAELE, CREMASCHI, ANDREA, & VANNUCCI, MARINA. 2020. Hierarchical normalized completely random measures to cluster grouped data. *Journal of the American Statistical Association*, **115**(529), 318–333.
- CAMERLENGHI, FEDERICO, LIJOI, ANTONIO, ORBANZ, PETER, PRÜNSTER, IGOR, *et al.* . 2019. Distribution theory for hierarchical processes. *Annals of Statistics*, **47**(1), 67–92.
- JAMES, LANCELOT F, LIJOI, ANTONIO, & PRÜNSTER, IGOR. 2009. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, **36**(1), 76–97.
- REGAZZINI, EUGENIO, LIJOI, ANTONIO, & PRÜNSTER, IGOR. 2003. Distributional results for means of normalized random measures with independent increments. *Annals of Statistics*, 560–585.
- TEH, YEE WHYE, JORDAN, MICHAEL I, BEAL, MATTHEW J, & BLEI, DAVID M. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association*, **101**(476), 1566–1581.