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

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

Jean-Michel Loubes
Optimal transport methods for fairness in machine learning 5

Peter Rousseeuw, Jakob Raymaekers and Mia Hubert
Class maps for visualizing classification results 6

Robert Tibshirani, Stephen Bates and Trevor Hastie
Understanding cross-validation and prediction error 7

Cinzia Viroli
Quantile-based classification 8

Bin Yu
Veridical data science for responsible AI: characterizing V4 neurons through deepTune 9

Plenary Session

Daniel Diaz
A simple correction for COVID-19 sampling bias 14

Jeffrey S. Morris
A seat at the table: the key role of biostatistics and data science in the COVID-19 pandemic 15

Bhramar Mukherjee
Predictions, role of interventions and the crisis of virus in India: a data science call to arms 16

Danny Pfeffermann
Contributions of Israel's CBS to rout COVID-19 17

Invited Papers

Claudio Agostinelli, Giovanni Saraceno and Luca Greco
Robust issues in estimating modes for multivariate torus data 21

Emanuele Aliverti
Bayesian nonparametric dynamic modeling of psychological traits 25

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giovanni C. Porzio, Carla Rampichini, Chiara Bocci (edited by), *CLADAG 2021 Book of abstracts and short papers. 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-340-6 (PDF), DOI 10.36253/978-88-5518-340-6

Andres M. Alonso, Carolina Gamboa and Daniel Peña
Clustering financial time series using generalized cross correlations **27**

Raffaele Argiento, Edoardo Filippi-Mazzola and Lucia Paci
Model-based clustering for categorical data via Hamming distance **31**

Antonio Balzanella, Antonio Irpino and Francisco de A.T. De Carvalho
Mining multiple time sequences through co-clustering algorithms for distributional data **32**

Francesco Bartolucci, Fulvia Pennoni and Federico Cortese
Hidden Markov and regime switching copula models for state allocation in multiple time-series **36**

Michela Battauz and Paolo Vidoni
Boosting multidimensional IRT models **40**

Matteo Bottai
Understanding and estimating conditional parametric quantile models **44**

Niklas Bussmann, Roman Enzmann, Paolo Giudici and Emanuela Raffinetti
Shapley Lorenz methods for eXplainable artificial intelligence **45**

Andrea Cappozzo, Ludovic Duponchel, Francesca Greselin and Brendan Murphy
Robust classification of spectroscopic data in agri-food: first analysis on the stability of results **49**

Andrea Cerasa, Enrico Checchi, Domenico Perrotta and Francesca Torti
Issues in monitoring the EU trade of critical COVID-19 commodities **53**

Marcello Chiodi
Smoothed non linear PCA for multivariate data **54**

Roberto Colombi, Sabrina Giordano and Maria Kateri
Accounting for response behavior in longitudinal rating data **58**

Claudio Conversano, Giulia Contu, Luca Frigau and Carmela Cappelli
Netwok-based semi-supervised clustering of time series data **62**

Federica Cugnata, Chiara Brombin, Pietro Cippà, Alessandro Ceschi, Paolo Ferrari and Clelia Di Serio
Characterising longitudinal trajectories of COVID-19 biomarkers within a latent class framework **64**

Silvia D'Angelo
Sender and receiver effects in latent space models for multiplex data **68**

Anna Denkowska and Stanisław Wanat
DTW-based assessment of the predictive power of the copula-DCC-GARCH-MST model developed for European insurance institutions **71**

Roberto Di Mari, Zsuzsa Bakk, Jennifer Oser and Jouni Kuha
Two-step estimation of multilevel latent class models with covariates **75**

Marie Du Roy de Chaumaray and Matthieu Marbac
Clustering data with non-ignorable missingness using semi-parametric mixture models **79**

Pierpaolo D'Urso, Livia De Giovanni and Vincenzina Vitale
Spatial-temporal clustering based on B-splines: robust models with applications to COVID-19 pandemic **83**

Leonardo Egidi, Roberta Pappadà, Francesco Pauli and Nicola Torelli
PIVMET: pivotal methods for Bayesian relabelling in finite mixture models **87**

Tahir Ekin and Claudio Conversano
Cluster validity by random forests **91**

Luis Angel García-Escudero, Agustín Mayo-Iscar and Marco Riani
Robust estimation of parsimonious finite mixture of Gaussian models **92**

Silvia Facchinetti and Silvia Angela Osmetti
A risk indicator for categorical data **93**

Matteo Fasiolo
Additive quantile regression via the qgam R package **97**

Michael Fop, Dimitris Karlis, Ioannis Kosmidis, Adrian O'Hagan, Caitriona Ryan and Isobel Claire Gormley
Gaussian mixture models for high dimensional data using composite likelihood **98**

Carlo Gaetan, Paolo Girardi and Victor Muthama Musau
On model-based clustering using quantile regression **102**

Carlotta Galeone
Socioeconomic inequalities and cancer risk: myth or reality? **106**

Michael Gallaugher, Christophe Biernacki and Paul McNicholas
Parameter-wise co-clustering for high dimensional data **107**

Francesca Greselin and Alina Jędrzejczak
Quantifying the impact of covariates on the gender gap measurement: an analysis based on EU-SILC data from Poland and Italy **108**

Alessandra Guglielmi, Mario Beraha, Matteo Giannella, Matteo Pegoraro and Riccardo Peli
A transdimensional MCMC sampler for spatially dependent mixture models **112**

Christian Hennig and Pietro Coretto
Non-parametric consistency for the Gaussian mixture maximum likelihood estimator **116**

Yinxuan Huang and Natalie Shlomo
Improving the reliability of a nonprobability web survey **120**

Maria Iannario and Claudia Tarantola
A semi-Bayesian approach for the analysis of scale effects in ordinal regression models **124**

Jayant Jha
Best approach direction for spherical random variables **128**

Maria Kateri
Simple effect measures for interpreting generalized binary regression models **129**

Shogo Kato, Kota Nagasaki and Wataru Nakanishi
Mixtures of Kato–Jones distributions on the circle, with an application to traffic count data **133**

John Kent
How to design a directional distribution **137**

Simona Korenjak-Černe and Nataša Kejžar
Identifying mortality patterns of main causes of death among young EU population using SDA approaches **141**

Fabrizio Laurini and Gianluca Morelli
Robust supervised clustering: some practical issues **142**

Daniela Marella and Danny Pfeffermann
A nonparametric approach for statistical matching under informative sampling and nonresponse **146**

Mariagiulia Matteucci and Stefania Mignani
Investigating model fit in item response models with the Hellinger distance **150**

Matteo Mazziotta and Adriano Pareto
PCA-based composite indices and measurement model **154**

Marcella Mazzoleni, Angiola Pollastri and Vanda Tulli
Gender inequalities from an income perspective **158**

Yana Melnykov, Xuwen Zhu and Volodymyr Melnykov
Transformation mixture modeling for skewed data groups with heavy tails and scatter **162**

Luca Merlo, Lea Petrella and Nikos Tzavidis
Unconditional M-quantile regression **163**

Jesper Møller , Mario Beraha, Raffaele Argiento and Alessandra Guglielmi
MCMC computations for Bayesian mixture models using repulsive point processes **167**

Keefe Murphy, Cinzia Viroli and Isobel Claire Gormley
Infinite mixtures of infinite factor analysers **168**

Stanislav Nagy, Petra Laketa and Rainer Dyckerhoff
Angular halfspace depth: computation **169**

Yarema Okhrin, Gazi Salah Uddin and Muhammad Yahya
Nonlinear Interconnectedness of crude oil and financial markets **173**

M. Rosário Oliveira, Ana Subtil and Lina Oliveira
Detection of internet attacks with histogram principal component analysis **174**

Sally Paganin
Semiparametric IRT models for non-normal latent traits **178**

Giuseppe Pandolfo
A graphical depth-based aid to detect deviation from unimodality on hyperspheres **182**

Panos Pardalos
Networks of networks **186**

Xanthi Pedeli and Cristiano Varin
Pairwise likelihood estimation of latent autoregressive count models **187**

Mark Reiser and Maduranga Dassanayake
A study of lack-of-fit diagnostics for models fit to cross-classified binary variables **191**

Giorgia Rivieccio, Jean-Paul Chavas, Giovanni De Luca, Salvatore Di Falco and Fabian Capitanio
Assessing food security issues in Italy: a quantile copula approach **195**

Nicoleta Rogovschi
Co-clustering for high dimensional sparse data **199**

Massimiliano Russo
Malaria risk detection via mixed membership models **203**

Paula Saavedra-Nieves and Rosa M. Crujeiras
Nonparametric estimation of the number of clusters for directional data **207**

Shuchismita Sarkar, Volodymyr Melnykov and Xuwen Zhu
Tensor-variate finite mixture model for the analysis of university professor remuneration **208**

Florian Schuberth
Specifying composites in structural equation modeling: the Henseler-Ogasawara specification **209**

Jarod Smith, Mohammad Arashi and Andriette Bekker
Network analysis implementing a mixture distribution from Bayesian viewpoint **210**

Paul Smith, Peter van der Heijden and Maarten Cruyff
Measurement errors in multiple systems estimation **211**

Valentin Todorov and Peter Filzmoser
Robust classification in high dimensions using regularized covariance estimates **215**

Salvatore Daniele Tomarchio, Luca Bagnato and Antonio Punzo
Clustering via new parsimonious mixtures of heavy tailed distributions **216**

Agostino Torti, Marta Galvani, Alessandra Menafoglio, Piercesare Secchi and Simone Vantini
A general bi-clustering technique for functional data **217**

Laura Trinchera
Developing a multidimensional and hierarchical index following a composite-based approach **220**

Rosanna Verde, Francisco T. de A. De Carvalho and Antonio Balzanella
A generalised clusteriwise regression for distributional data **223**

Marika Vezzoli, Francesco Doglietto, Stefano Renzetti, Marco Fontanella and Stefano Calza
A machine learning approach for evaluating anxiety in neurosurgical patients during the COVID-19 pandemic **227**

Isadora Antoniano Villalobos, Simone Padoan and Boris Beranger
Prediction of large observations via Bayesian inference for extreme-value theory **231**

Maria Prosperina Vitale, Vincenzo Giuseppe Genova, Giuseppe Giordano and Giancarlo Ragozini
Community detection in tripartite networks of university student mobility flows **232**

Ernst Wit and Lucas Kania
Causal regularization **236**

Qiuyi Wu and David Banks
Minimizing conflicts of interest: optimizing the JSM program **240**

Contributed Papers

Antonino Abbruzzo, Maria Francesca Cracolici and Furio Urso
Model selection procedure for mixture hidden Markov models **243**

Roberto Ascari and Sonia Migliorati
A full mixture of experts model to classify constrained data **247**

Luigi Augugliaro, Gianluca Sottile and Angelo Mineo
Sparse inference in covariate adjusted censored Gaussian graphical models **251**

Simona Balzano, Mario Rosario Guarracino and Giovanni Camillo Porzio
Semi-supervised learning through depth functions **255**

Lucio Barabesi, Andrea Cerasa, Andrea Cerioli and Domenico Perrotta
A combined test of the Benford hypothesis with anti-fraud applications **256**

Chiara Bardelli
Unabalanced classfication of electronic invoicing **260**

Claudia Berloco, Raffaele Argiento and Silvia Montagna
Predictive power of Bayesian CAR models on scale free networks: an application for credit risk **264**

Marco Berrettini, Giuliano Galimberti and Saverio Ranciati
Semiparametric finite mixture of regression models with Bayesian P-splines **268**

Giuseppe Bove
A subject-specific measure of interrater agreement based on the homogeneity index **272**

Antonio Calcagni
Estimating latent linear correlations from fuzzy contingency tables **276**

Andrea Cappozzo, Alessandro Casa and Michael Fop
Model-based clustering with sparse matrix mixture models **280**

Andrea Cappozzo, Luis Angel Garcìa Escudero, Francesca Greselin and Agustìn Mayo-Iscar
Exploring solutions via monitoring for cluster weighted robust models **284**

Maurizio Carpita and Silvia Golia
Categorical classifiers in multi-class classification problems **288**

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
Model-based clustering for estimating cetaceans site-fidelity and abundance **292**

Carlo Cavicchia, Maurizio Vichi and Giorgia Zaccaria
Model-based clustering with parsimonious covariance structure **296**

Francesca Condino
Clustering income data based on share densities **300**

Paula Costa Fontichiari, Miriam Giuliani, Raffaele Argiento and Lucia Paci
Group-dependent finite mixture model **304**

Salvatore Cuomo, Federico Gatta, Fabio Giampaolo, Carmela Iorio and Francesco Piccialli
A machine learning approach in stock risk management **308**

Cristina Davino and Giuseppe Lamberti
Pathmox segmentation trees to compare linear regression models **312**

Houyem Demni, Davide Buttarazzi, Stanislav Nagy and Giovanni Camillo Porzio
Angular halfspace depth: classification using spherical bagdistances **316**

Agostino Di Ciaccio
Neural networks for high cardinality categorical data **320**

F. Marta L. Di Lascio, Andrea Menapace and Roberta Pappadà
Ali-Mikhail-Haq copula to detect low correlations in hierarchical clustering **324**

Maria Veronica Dorgali, Silvia Bacci, Bruno Bertaccini and Alessandra Petrucci
Higher education and employability: insights from the mandatory notices of the ministry of labour **328**

Lorenzo Focardi Olmi and Anna Gottard
An alternative to joint graphical lasso for learning multiple Gaussian graphical models **332**

Francesca Fortuna, Alessia Naccarato and Silvia Terzi
Functional cluster analysis of HDI evolution in European countries **336**

Sylvia Frühwirth-Schnatter, Bettina Grün and Gertraud Malsiner-Walli
Estimating Bayesian mixtures of finite mixtures with telescoping sampling **340**

Chiara Galimberti, Federico Castelletti and Stefano Peluso
A Bayesian framework for structural learning of mixed graphical models **344**

Andrea Gilardi, Riccardo Borgoni, Luca Presicce and Jorge Mateu
Measurement error models on spatial network lattices: car crashes in Leeds **348**

Carmela Iorio, Giuseppe Pandolfo, Michele Staiano, Massimo Aria and Roberta Siciliano
The L^p data depth and its application to multivariate process control charts **352**

Petra Laketa and Stanislav Nagy
Angular halfspace depth: central regions **356**

Michele La Rocca, Francesco Giordano and Cira Perna
Clustering production indexes for construction with forecast distributions **360**

Maria Mannone, Veronica Distefano, Claudio Silvestri and Irene Poli
Clustering longitudinal data with category theory for diabetic kidney disease **364**

Laura Marcis, Maria Chiara Pagliarella and Renato Salvatore
A redundancy analysis with multivariate random-coefficients linear models **368**

Paolo Mariani, Andrea Marletta and Matteo Locci
The use of multiple imputation techniques for social media data **372**

Federico Marotta, Paolo Provero and Silvia Montagna
Prediction of gene expression from transcription factors affinities: an application of Bayesian non-linear modelling **376**

Francesca Martella, Fabio Attorre, Michele De Sanctis and Giuliano Fanelli
High dimensional model-based clustering of European georeferenced vegetation plots **380**

Ana Martins, Paula Brito, Sónia Dias and Peter Filzmoser
Multivariate outlier detection for histogram-valued variables **384**

Giovanna Menardi and Federico Ferraccioli
A nonparametric test for mode significance **388**

Massimo Mucciardi, Giovanni Pirrotta, Andrea Briglia and Arnaud Sallaberry
Visualizing cluster of words: a graphical approach to grammar acquisition **392**

Marta Nai Ruscone and Dimitris Karlis
Robustness methods for modelling count data with general dependence structures **396**

Roberta Paroli, Luigi Spezia, Marc Stutter and Andy Vinten
Bayesian analysis of a water quality high-frequency time series through Markov switching autoregressive models **400**

Mariano Porcu, Isabella Sulis and Cristian Usala
Detecting the effect of secondary school in higher education university choices **404**

Roberto Rocci and Monia Ranalli
Semi-constrained model-based clustering of mixed-type data using a composite likelihood approach **408**

Annalina Sarra, Adelia Evangelista, Tonio Di Battista and Damiana Pieragostino
Antibodies to SARS-CoV-2: an exploratory analysis carried out through the Bayesian profile regression **412**

Theresa Scharl and Bettina Grün
Modelling three-way RNA sequencing data with mixture of multivariate Poisson-lognormal distribution **416**

Luca Scrucca
Stacking ensemble of Gaussian mixtures **420**

Rosaria Simone, Cristina Davino, Domenico Vistocco and Gerhard Tutz
A robust quantile approach to ordinal trees **424**

Venera Tomaselli, Giulio Giacomo Cantone and Valeria Mazzeo
The detection of spam behaviour in review bomb **428**

Donatella Vicari and Paolo Giordani
Clustering models for three-way data **432**

Gianpaolo Zammarchi and Jaromir Antoch
Using eye-traking data to create a weighted dictionary for sentiment analysis: the eye dictionary **436**

PIVMET: PIVOTAL METHODS FOR BAYESIAN RELABELLING IN FINITE MIXTURE MODELS

Leonardo Egidi [1], Roberta Pappadà [1], Francesco Pauli [1] and Nicola Torelli [1]

[1] Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche 'Bruno de Finetti', Università degli Studi di Trieste (e-mail: legidi@units.it, rpappada@units.it, francesco.pauli@deams.units.it, nicola.torelli@deams.units.it)

ABSTRACT: The identification of groups' prototypes, i.e. elements of a dataset that are representative of the group they belong to, is relevant to the tasks of clustering, classification and mixture modeling. The R package pivmet includes different methods for extracting pivotal units from a dataset, to be exploited for a Markov Chain Monte Carlo (MCMC) relabelling technique for dealing with label switching in Bayesian estimation of mixture models. Moreover, consensus clustering based on pivotal units may improve classical algorithms (e.g. *k*-means) by means of a careful seeding.

KEYWORDS: pivotal unit, mixture model, relabelling, consensus clustering.

1 Introduction

The identification of some units which may be representative of the group they belong to is often a matter of statistical importance and can help avoiding an extra amount of work when processing the data. The advantage of such pivotal units (hereafter called pivots) is that they are somehow chosen to be as far as possible from units in the other groups and as similar as possible to the units in the same group, and may be beneficial in many statistical frameworks, such as clustering, classification, and mixture modeling.

The pivmet R package (Egidi *et al.*, 2021) implements various pivotal selection criteria, graphical tools and the relabelling method (Papastamoulis, 2016) described in Egidi *et al.*, 2018 to deal with 'label switching' (Redner & Walker, 1984), a well-known phenomenon causing nonidentifiability of the mixture parameters during the MCMC sampling (Frühwirth-Schnatter, 2001). Compared to other packages, it allows the user to fit their own mixture model using data augmentation with component memberships either via the JAGS (Plummer, 2018) or the Stan (Carpenter *et al.*, 2017) software, by specifying suitable prior distributions. Pivotal units are detected via the similarity matrix

derived from the MCMC sample—whose elements are the estimated probabilities that any two units in the observed sample are drawn from the same component—and used to relabel the chains. Such units may be fruitfully used in Dirichlet process mixture models (DPMM) (Ferguson, 1973, Neal, 2000), a class of models that naturally sorts data into clusters, and in data clustering to guarantee a better final clustering solution starting from a careful seeding based on well-separated statistical units.

The aim of the paper is to provide a quick overview of the computational capabilities of our package in the field of Bayesian mixture models.

2 Finite mixtures of Gaussian distributions

Consider a multivariate mixture of Gaussian distributions, let $\boldsymbol{y}_i \in \mathbb{R}^d$ and assume that

$$\boldsymbol{y}_i \sim \sum_{j=1}^{k} \eta_j \mathcal{N}_d(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \;\; i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\mu}_j \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_j$ is a $d \times d$ positive definite covariance matrix. We assume the following prior specification for the parameters in (1):

$$\begin{aligned} \boldsymbol{\mu}_j \sim & \mathcal{N}_2(\boldsymbol{\mu}_0, S_2), \;\; \boldsymbol{\Sigma}_j^{-1} \sim \text{Wishart}(S_3, d+1) \\ \eta \sim & \text{Dirichlet}(\boldsymbol{\alpha}), \end{aligned} \tag{2}$$

where $\boldsymbol{\alpha}$ is a k-dimensional vector and S_2 and S_3 are positive definite matrices. We fix $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\alpha} = (1, \ldots, 1)$ and assume S_2 and S_3 are diagonal matrices, with diagonal elements equal to 10^5. We simulate a sample of size $n = 150$ from a bivariate Gaussian distribution with the function `piv_sim` and we fit the model using the JAGS option. From the bivariate traceplot chains for each mean component $\mu_{j,1}$ and $\mu_{j,2}$ in Figure 1 we clearly note that label switching has occurred and the relabelling algorithm fixed it, by isolating the four bivariate high-density regions.

3 Dirichlet Process Mixture Models

DPMM are useful tools for non-parametric density estimation, and, more generally, the choice of a Dirichlet process prior avoids the specification of an inappropriate parametric form. The DPMM has the following form:

$$\begin{aligned} & y_i \sim K(y_i|\theta_i), \; i = 1, \ldots, n, \\ & \theta_i \sim F, \;\; F \sim \text{DP}(\alpha, G), \end{aligned} \tag{3}$$

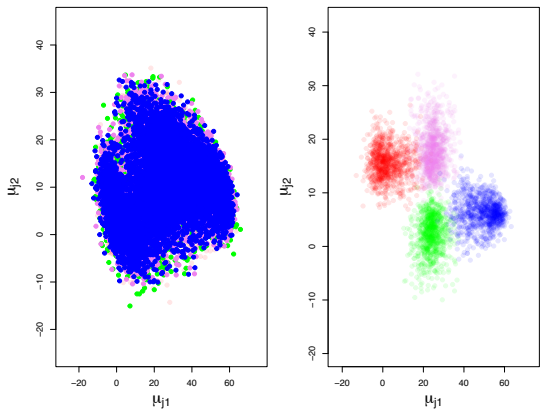

Figure 1. *Bivariate mixture data: scatterplot for the mean parameters obtained via JAGS sampling (left plot) and relabelled estimates (right plot) via the* `maxsumdiff` *pivotal criterion.*

where $K(\cdot)$ is a parametric kernel function which is usually continuous, F is an unknown probability distribution, DP is the nonparametric Dirichlet process prior with concentration parameter α and *base measure* G, which encapsulates any prior knowledge about F. A common choice for $K(\cdot)$ is a Gaussian mixture model, so that $K(y_i|\theta_i) = \mathcal{N}(\mu_i, \sigma_i^2)$. The DPMM sorts the data into clusters, corresponding to the mixture components. Thus, it may be seen as an infinite dimensional mixture model which generalizes finite mixture models. Thus, pivotal units detection may be quite relevant for this class of models in order to identify distinct groups characteristics. We generate $n = 200$ data from a student$-t$ distribution with 3 degrees of freedom and we draw posterior samples for $\mu_1, \mu_2, \ldots, \mu_k$ via the `dirichletprocess` package. Figure 2 represents posterior density estimation for the simulated dataset along with nine pivotal units (blue points). detected by `pivmet` via the `maxsumdiff` pivotal criterion.

References

CARPENTER, BOB, GELMAN, ANDREW, HOFFMAN, MATTHEW D, LEE, DANIEL, GOODRICH, BEN, BETANCOURT, MICHAEL, BRUBAKER, MARCUS A, GUO, JIQIANG, LI, PETER, & RIDDELL, ALLEN. 2017. Stan: a probabilistic programming language. *Journal of Statistical Software*, **76**(1), 1–32.

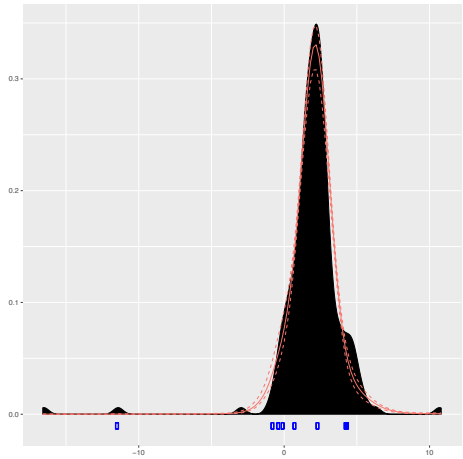

Figure 2. *Posterior density estimation (red line) for a sample of* $n = 200$ *data points from a student*$-t$ *distribution. Blue points below the x-axis denote the pivotal units.*

EGIDI, LEONARDO, PAPPADÀ, ROBERTA, PAULI, FRANCESCO, & TORELLI, NICOLA. 2018. Relabelling in Bayesian mixture models by pivotal units. *Statistics and Computing*, **28**(4), 957–969.

EGIDI, LEONARDO, PAPPADÀ, ROBERTA, PAULI, FRANCESCO, & TORELLI, NICOLA. 2021. pivmet: Pivotal methods for Bayesian relabelling and k-means clustering. *arXiv preprint arXiv:2103.16948*.

FERGUSON, THOMAS S. 1973. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, 209–230.

FRÜHWIRTH-SCHNATTER, SYLVIA. 2001. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, **96**(453), 194–209.

NEAL, RADFORD M. 2000. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, **9**(2), 249–265.

PAPASTAMOULIS, PANAGIOTIS. 2016. label.switching: An R package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software, Code Snippets*, **69**(1), 1–24.

PLUMMER, MARTYN. 2018. *rjags: Bayesian graphical models using MCMC*. R package version 4-8.

REDNER, RICHARD A., & WALKER, HOMER F. 1984. Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review*, **26**(2), 195–239.