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.

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

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

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

Contributed Papers

GROUP-DEPENDENT FINITE MIXTURE MODEL

Paula Costa Fontichiari[1], Miriam Giuliani[1], Raffaele Argiento[1] and Lucia Paci[1]

[1] Department of Statistical Sciences, Università Cattolica del Sacro Cuore, (e-mail: paula.costafontichiari01@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, an 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

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

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 nonparametric 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 $\boldsymbol{\Theta}$, where

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

with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. 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:

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

In this work, the kernel $f(y \mid \boldsymbol{\theta})$ 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 $\boldsymbol{\theta}_{ji} = \tau_m$. 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

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

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

We are now ready to define, for each group j, the clustering $\rho_j = \{A_{j1}, \ldots$ $\ldots, A_{jM^{(a)}}\}$, where $A_{jk} = \{(j,i) : \ (j, c_{ki}^*) \in \mathcal{M}^{(a)}\}$ and $k = 1, ..., M^{(a)}$. 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 $\rho_1, \ldots, \rho_d$, induced by the hierar-

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

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

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.