Authors: Casa A., Fop M. & Murphy T.B.

Title: Discussion on: "Centered partition processes: Informative priors for clustering" by Paganin, S., Herring, A. H., Olshan, A. F., Dunson, D.B.

Year: 2021

DOI: 10.1214/20-BA1197

Journal: Bayesian Analysis

Journal ISSN: 1931-6690 (web)
1936-0975 (print)

Bayesian Analysis (2021) **16**, Number 1, pp. 301–370

Centered Partition Processes: Informative Priors for Clustering (with Discussion)

Sally Paganin[*], Amy H. Herring[†], Andrew F. Olshan[‡], David B. Dunson[§], and The National Birth Defects Prevention Study

Abstract. There is a very rich literature proposing Bayesian approaches for clustering starting with a prior probability distribution on partitions. Most approaches assume exchangeability, leading to simple representations in terms of Exchangeable Partition Probability Functions (EPPF). Gibbs-type priors encompass a broad class of such cases, including Dirichlet and Pitman-Yor processes. Even though there have been some proposals to relax the exchangeability assumption, allowing covariate-dependence and partial exchangeability, limited consideration has been given on how to include concrete prior knowledge on the partition. For example, we are motivated by an epidemiological application, in which we wish to cluster birth defects into groups and we have prior knowledge of an initial clustering provided by experts. As a general approach for including such prior knowledge, we propose a Centered Partition (CP) process that modifies the EPPF to favor partitions close to an initial one. Some properties of the CP prior are described, a general algorithm for posterior computation is developed, and we illustrate the methodology through simulation examples and an application to the motivating epidemiology study of birth defects.

Keywords: Bayesian clustering, Bayesian nonparametrics, centered process, Dirichlet Process, exchangeable probability partition function, mixture model, product partition model.

1 Introduction

Clustering is one of the canonical data analysis goals in statistics. There are two main strategies that have been used for clustering; namely, distance and model-based clustering. Distance-based methods leverage upon a distance metric between data points, and do not in general require a generative probability model of the data. Model-based methods rely on discrete mixture models, which model the data in different clusters as arising from kernels having different parameter values. The majority of the model-based literature uses maximum likelihood estimation, commonly relying on the EM algorithm. Bayesian approaches that aim to approximate a full posterior distribution on the clusters have advantages in terms of uncertainty quantification, while also having the ability to incorporate prior information.

[*]Department of Environmental Science, Policy, and Management, University of California, Berkeley, sally.paganin@berkeley.edu

[†]Department of Statistical Science, Duke University, Durham, amy.herring@duke.edu

[‡]Department of Epidemiology, The University of North Carolina at Chapel Hill, Chapel Hill, andy_olshan@unc.edu

[§]Department of Statistical Science, Duke University, Durham, dunson@duke.edu

 https://doi.org/10.1214/20-BA1197

Contributed Discussion

Alessandro Casa[*,§], Michael Fop[†], and Thomas Brendan Murphy[‡,¶]

We would like to congratulate the authors for their work, which represents a relevant contribution to the Bayesian cluster analysis framework. Prior elicitation is a critical issue and currently most people rely on the exchangeability assumption. To the best of our knowledge, this work is one of the first attempts to include concrete available prior information on the partition, and we hope it will serve as a stepping stone motivating further explorations of the topic.

The proposal is directly motivated by an epidemiological application where some experts provided an initial clustering $\mathbf{c}_0$ subsequently used to center the proposed prior. However, there could be cases where the experts do not agree on the classification of the objects to be clustered, thus resulting in a situation where a set of G initial clusterings $\mathcal{C}_0 = \{\mathbf{c}_0^1, \ldots, \mathbf{c}_0^G\}$ is available. As a consequence, it may be interesting to propose a suitable modification of the proposed prior, possibly able to encompass scenarios where multiple initial partitions are available, thus enlarging the applicability of the CP process. In our opinion, a reasonable and simple modification may be expressed as follows:

$$p(\mathbf{c}|\mathcal{C}_0) \propto p_0(\mathbf{c}) e^{-\psi \sum_{g=1}^{G} \omega_g d(\mathbf{c}, \mathbf{c_0}^g)} \tag{1}$$

where $\omega_g \geq 0$ for $g = 1, \ldots, G$ with $\sum_g \omega_g = 1$, while the other quantities are defined as in the original paper. The coefficients ω_g's allows to assign different weights to the initial partitions in $\mathcal{C}_0$.

Note that a wider range of situations may be framed in a multiple initial partitions scenario, namely all the ones where only partial information are available a priori. In fact, a similar problem appears in the recent work by Casa et al. (2021), involving searching for a partition of the wavelengths in a spectroscopy application. The prior (1) could be used to incorporate subject matter knowledge on those spectral regions influenced by the same chemical compounds, and likely to be clustered together. We believe that a broad set of issues arising in the semi-supervised clustering framework (see Melnykov et al., 2016, and reference therein) can be flexibly faced by considering the strategy outlined above. In fact, this approach would encompass restrictions on cluster membership, as well as cannot- or must-link among them, by simply populating $\mathcal{C}_0$ with those partitions complying with the restrictions themselves. Finally, note that the same reasoning applies when relevant prior mass has to be considered for partitions with specific cluster sizes or number of clusters.

[*]School of Mathematics and Statistics, University College Dublin, Ireland, alessandro.casa@ucd.ie
[†]School of Mathematics and Statistics, University College Dublin, Ireland, michael.fop@ucd.ie
[‡]School of Mathematics and Statistics, University College Dublin, Ireland, brendan.murphy@ucd.ie
[§]Insight Centre for Data Analytics and Vistamilk SFI Research Centre
[¶]Insight Centre for Data Analytics and Vistamilk SFI Research Centre

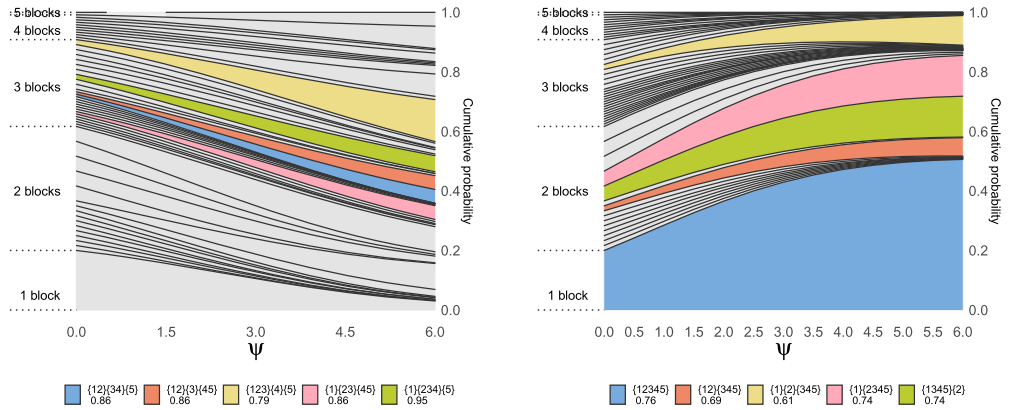

Figure 1: Prior probabilities of the 52 set partitions of $N = 5$ elements for the prior (1) with Dirichlet process of $\alpha = 1$ base EPFF. In each graph the modified CP process is centered on a different set of partitions $\mathcal{C}_0$ highlighted with different colors. The partitions in $\mathcal{C}_0$ are reported below the respective graph alongside with the mean of the pairwise Variation Information (VI) computed for the partitions in $\mathcal{C}_0$.

In the following, mimicking what the authors did in the paper, we study the behavior of the prior in (1) as a function of ψ. As a base EPFF $p_0(\mathbf{c})$ we use the Dirichlet process with $\alpha = 1$ while $\omega_g = 1/G$ for $g = 1, \ldots, G$. In the left plot the set of initial partitions $\mathcal{C}_0$ contains five partitions with 3 clusters. For increasing values of ψ, the prior (1) naturally tends to assign higher probabilities to the partitions in $\mathcal{C}_0$. Moreover a greater increase in the probability for the partition $\{1, 2, 3\}\{4\}\{5\}$, highlighted in yellow, being the one closer to the others in $\mathcal{C}_0$, is witnessed: this implies that the modified CP process tends to favor the partitions in $\mathcal{C}_0$ being more similar to the others in the same set. On the other hand, in the right plot, $\mathcal{C}_0$ contains all those partitions where the observations $\{3, 4, 5\}$ are clustered together; this scenario resembles the one in Casa et al. (2021) outlined above. It stands out even more clearly how, for increasing ψ, most of the mass is assigned to the partitions in $\mathcal{C}_0$.

An additional point, which might worth a reflection, consists in the potential changes to the prior calibration step and to the local search when the prior is not centered on a single node of the Hasse diagram but on multiple ones. We would like to hear authors' thoughts on this and, more generally, about our alternative prior formulation, encompassing the situation where multiple reference partitions and partial grouping information are available.

References

Casa, A., O'Callaghan, T. F., and Murphy, T. B. (2021). "Parsimonious Bayesian Factor Analysis for modelling latent structures in spectroscopy data." *arXiv preprint arXiv:2101.12499*. 361, 362

Melnykov, V., Melnykov, I., and Michael, S. (2016). "Semi-supervised model-based clustering with positive and negative constraints." *Advances in Data Analysis and Classification*, 10(3): 327–349. MR3541239. doi: https://doi.org/10.1007/s11634-015-0200-3. 361