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

Although this article is motivated by providing a broad new class of methods for improving clustering performance in practice, we were initially motivated by a particular application involving birth defects epidemiology. In this context, there are $N = 26$ different birth defects, which we can index using $i \in \{1, \ldots, N\}$, and for each defect i there is an highly variable number of observations. We are interested in clustering these birth defects into mechanistic groups, which may be useful, for example, in that birth defects in the same group may have similar coefficients in logistic regression analysis relating different exposures to risk of developing the defect. Investigators have provided us with an initial partition $\boldsymbol{c}_0$ of the defects $\{1, \ldots, N\}$ into groups. It is appealing to combine this prior knowledge with information in the data from a grouped logistic regression to produce a posterior distribution on clusters, which characterizes uncertainty. The motivating question of this article is how to do this, with the resulting method ideally having broader impact to other types of *centering* of priors for clustering; for example, we may want to center the prior based on information on the number of clusters or cluster sizes.

With these goals in mind, we start by reviewing the relevant literature on clustering priors. Most of these methods assume *exchangeability*, which means that the prior probability of a partition $\boldsymbol{c}$ of $\{1, \ldots, N\}$ into clusters depends only on the number of clusters and the cluster sizes; the indices on the clusters play no role. Under the exchangeability assumption, one can define what is referred to in the literature as an Exchangeable Partition Probability Function (EPPF) (Pitman, 1995). This EPPF provides a prior distribution on the random partition $\boldsymbol{c}$. One direction to obtain a specific form for the EPPF is to start with a nonparametric Bayesian discrete mixture model with a prior for the mixing measure P, and then marginalize over this prior to obtain an induced prior on partitions. Standard choices for P, such as the Dirichlet (Ferguson, 1973) and Pitman-Yor process (Pitman and Yor, 1997), lead to relatively simple analytic forms for the EPPF. There has been some recent literature studying extensions to broad classes of Gibbs-type processes (Gnedin and Pitman, 2006; De Blasi et al., 2015), mostly focused on improving flexibility while maintaining the ability to predict the number of new clusters in a future sample of data.

There is also a rich literature on relaxing exchangeability in various ways. Most of the emphasis has been on the case in which a vector of features $\mathbf{x}_i$ is available for index i, motivating feature-dependent random partitions models. Building on the stick-breaking representation of the DP (Sethuraman, 1994), MacEachern (1999, 2000) proposed a class of fixed weight dependent DP (DDP) priors. Applications of this DDP framework have been employed in ANOVA modeling (De Iorio et al., 2004), spatial data analysis (Gelfand et al., 2005), time series (Caron et al., 2006) and functional data analysis applications (Petrone et al., 2009; Scarpa and Dunson, 2009) among many others, with some theoretical properties highlighted in Barrientos et al. (2012).

However such fixed weight DDPs lack flexibility in feature-dependent clustering, as noted in MacEachern (2000). This has motivated alternative formulations which allow the mixing weights to change with the features, with some examples including the order-based dependent Dirichlet process (Griffin and Steel, 2006), kernel- (Dunson and Park, 2008), and probit- (Rodriguez and Dunson, 2011) stick breaking processes.

Alternative approaches build on random partition models (RPMs), working directly with the probability distribution $p(\boldsymbol{c})$ on the partition $\boldsymbol{c}$ of indices $\{1, \dots, N\}$ into clusters. Particular attention has been given to the class of product partition models (PPMs) (Barry and Hartigan, 1992; Hartigan, 1990) where $p(\boldsymbol{c})$ can be factorized into a product of cluster-dependent functions, known as *cohesion functions*. A common strategy modifies the cohesion function to allow feature-dependence; refer, for examples, to Park and Dunson (2010), Müller et al. (2011), Blei and Frazier (2011), Dahl et al. (2017) and Smith and Allenby (2019).

Our focus is fundamentally different. In particular, we do not have features $\mathbf{x}_i$ on indices i but have access to an informed prior guess $\boldsymbol{c}_0$ for the partition $\boldsymbol{c}$; other than this information it is plausible to rely on exchangeable priors. To address this problem, we propose a general strategy to modify a baseline EPPF to include centering on $\boldsymbol{c}_0$. In particular, our proposed Centered Partition (CP) process defines the partition prior as proportional to an EPPF multiplied by an exponential factor that depends on a distance function $d(\boldsymbol{c}, \boldsymbol{c}_0)$, measuring how far $\boldsymbol{c}$ is from $\boldsymbol{c}_0$. The proposed framework should be broadly useful in including extra information into EPPFs, which tend to face issues in lacking incorporation of real prior information from applications.

The paper is organized as follows. Section 2 introduces concepts and notation related to Bayesian nonparametric clustering. In Section 3 we illustrate the general CP process formulation and describe an approach to posterior computation relying on Markov chain Monte Carlo (MCMC). Section 4 proposes a general strategy for prior calibration building on a targeted Monte Carlo procedure. Simulation studies and application to the motivating birth defects epidemiology study are provided in Section 5, with technical details included in Paganin et al. (2020).

2 Clustering and Bayesian Models

This section introduces some concepts related to the representation of the clustering space from a combinatorial perspective, which will be useful to define the Centered Partition process, along with an introduction to Bayesian nonparametric clustering models.

2.1 Set Partitions

Let $\boldsymbol{c}$ be a generic clustering of indices $[N] = \{1, \dots, N\}$. It can be either represented as a vector of indices $\{c_1, \dots, c_N\}$, with $c_i \in \{1, \dots, K\}$ for $i = 1, \dots, N$ and $c_i = c_j$ when i and j belong to the same cluster, or as a collection of disjoint subsets (blocks) $\{B_1, B_2, \dots, B_K\}$ where B_k contains all the indices of data points in the k-th cluster and K is the number of clusters in the sample of size N. From a mathematical perspective $\boldsymbol{c} = \{B_1, \dots, B_K\}$ is a combinatorial object known as *set partition* of $[N]$. The collection of all possible set partitions of $[N]$, denoted with Π_N, is known as the *partition lattice*. We refer to Stanley (1997) and Davey and Priestley (2002) for an introduction to lattice theory, and to Meilă (2007) and Wade and Ghahramani (2018) for a review of the concepts from a more statistical perspective.

According to Knuth in Wilson and Watkins (2013), set partitions seem to have been studied first in Japan around A.D. 1500, due to a popular game in the upper class society known as *genji-ko*; five unknown incense sticks were burned and players were asked to identify which of the scents were the same, and which were different. Soon diagrams were developed to model all the 52 outcomes, which corresponds to all the possible set partitions of $N = 5$ elements. First results focused on enumerating the elements of the space. For example, for a fixed number of blocks K, the number of ways to assign N elements to K groups is described by the *Stirling number of the second kind*

$$\mathcal{S}_{N,K} = \frac{1}{K!}\sum_{j=0}^{K}(-1)^j\binom{K}{j}(K-j)^N,$$

while the *Bell number* $\mathcal{B}_N = \sum_{K=1}^{N}\mathcal{S}_{N,K}$ describes the number of all possible set partitions of N elements.

Interest progressively shifted towards characterizing the structure of the space of partitions using the notion of partial order. Consider Π_N endowed with the set containment relation $\leq$, meaning that for $\boldsymbol{c} = \{B_1, \dots, B_K\}$, $\boldsymbol{c}' = \{B'_1, \dots, B'_{K'}\}$ belonging to Π_N, $\boldsymbol{c} \leq \boldsymbol{c}'$ if for all $i = 1, \dots, K, B_i \subseteq B'_j$ for some $j \in \{1, \dots, K'\}$. Then the space $(\Pi_N, \leq)$ is a *partially ordered set* (poset), which satisfies the following properties:

1. Reflexivity: for every $\boldsymbol{c} \in \Pi_N$, $\boldsymbol{c} \leq \boldsymbol{c}$,
2. Antisymmetry: if $\boldsymbol{c} \leq \boldsymbol{c}'$ and $\boldsymbol{c}' \leq \boldsymbol{c}$ then $\boldsymbol{c} = \boldsymbol{c}'$,
3. Transitivity: if $\boldsymbol{c} \leq \boldsymbol{c}'$ and $\boldsymbol{c}' \leq \boldsymbol{c}''$, then $\boldsymbol{c} \leq \boldsymbol{c}''$.

Moreover, for any $\boldsymbol{c}, \boldsymbol{c}' \in \Pi_N$, it is said that $\boldsymbol{c}$ is *covered* (or refined) by $\boldsymbol{c}'$ if $\boldsymbol{c} \leq \boldsymbol{c}'$ and there is no $\boldsymbol{c}''$ such that $\boldsymbol{c} < \boldsymbol{c}'' < \boldsymbol{c}'$. Such a relation is indicated by $\boldsymbol{c} \prec \boldsymbol{c}'$. This covering relation allows one to represent the space of partitions using a *Hasse diagram*, in which the elements of Π_N correspond to nodes in a graph and a line is drawn from $\boldsymbol{c}$ to $\boldsymbol{c}'$ when $\boldsymbol{c} \prec \boldsymbol{c}'$; there is a connection from a partition $\boldsymbol{c}$ to another one when the second can be obtained by splitting or merging one of the blocks in $\boldsymbol{c}$. See Figure 1 for an example of the Hasse diagram of Π_4. Conventionally, the partition with just one cluster is represented at the top of the diagram and denoted as $\mathbf{1}$, while the partition having every observation in its own cluster is at the bottom and indicated with $\mathbf{0}$.

This representation of the set partitions space Π_N as a partially ordered set provides a useful framework to characterize its elements. As already mentioned, two partitions connected in the Hasse diagram can be obtained from one another by means of a single operation of split or merge; a sequence of connections is a *path*, linking the two extreme partitions $\mathbf{0}$ and $\mathbf{1}$. A path starting from $\mathbf{0}$ connects partitions with an increasing *rank*, which is related to the number of blocks through $r(\boldsymbol{c}) = N - |\boldsymbol{c}|$. Set partitions with the same rank may differ in terms of their *configuration* $\boldsymbol{\Lambda}(\boldsymbol{c})$, the sequence of block cardinalities $\{|B_1|, \dots, |B_K|\}$, which corresponds to another combinatorial object known as an *integer partition* of N. In combinatorics, an integer partition is defined as the multiset of positive integers $\{\lambda_1 \dots \lambda_K\}$, listed in decreasing order by convention, such

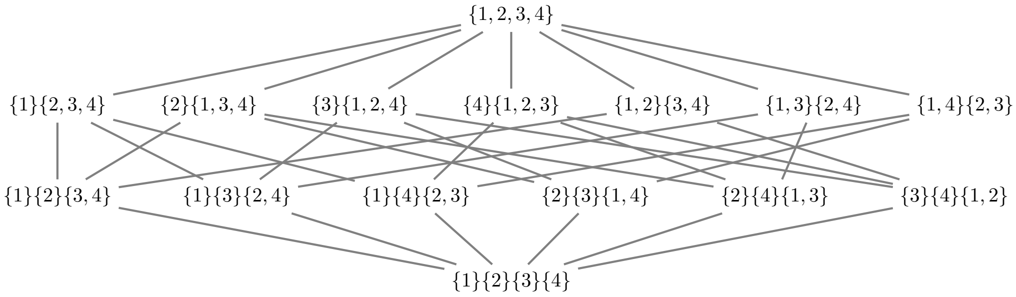


Figure 1: Hasse diagram for the lattice of set partitions of 4 elements. A line is drawn when two partitions have a covering relation. For example $\{1\}\{2, 3, 4\}$ is connected with 3 partitions obtained by splitting the block $\{2, 3, 4\}$ in every possible way, and with partition $\mathbf{1}$, obtained by merging the two clusters.

that $\sum_{i=1}^{K} \lambda_i = N$. Also the associated space of all possible integer partitions I_N is a partially ordered set, making the definition of configuration a poset mapping $\boldsymbol{\Lambda}(\cdot) : \boldsymbol{c} \in \Pi_N \to \boldsymbol{\lambda} \in I_N$.

Finally, the space Π_N is a *lattice*, based on the fact that every pair of elements has a *greatest lower bound* (g.l.b.) and a *least upper bound* (l.u.b.) indicated with the "meet" $\wedge$ and the "join" $\vee$ operators, i.e. $\boldsymbol{c} \wedge \boldsymbol{c}' = \text{g.l.b.}(\boldsymbol{c}, \boldsymbol{c}')$ and $\boldsymbol{c} \vee \boldsymbol{c}' = \text{l.u.b.}(\boldsymbol{c}, \boldsymbol{c}')$ and equality holds under a permutation of the cluster labels. An element $\boldsymbol{c} \in \Pi_N$ is an upper bound for a subset $\boldsymbol{S} \subseteq \Pi_N$ if $\boldsymbol{s} \leq \boldsymbol{c}$ for all $\boldsymbol{s} \in \boldsymbol{S}$, and it is the least upper bound for a subset $\boldsymbol{S} \subseteq \Pi_N$ if $\boldsymbol{c}$ is an upper bound for $\boldsymbol{S}$ and $\boldsymbol{c} \leq \boldsymbol{c}'$ for all upper bounds $\boldsymbol{c}'$ of $\boldsymbol{S}$. The lower bound and the greatest lower bound are defined similarly, and the definition applies also to the elements of the space I_N. Consider, as an example, $\boldsymbol{c} = \{1\}\{2, 3, 4\}$ and $\boldsymbol{c}' = \{3\}\{1, 2, 4\}$; their greatest lower bound is $\boldsymbol{c} \wedge \boldsymbol{c}' = \{1\}\{3\}\{2, 4\}$ while the lowest upper bound is $\boldsymbol{c} \vee \boldsymbol{c}' = \{1, 2, 3, 4\}$. Considering the Hasse diagram in Figure 1 the g.l.b. and l.u.b. are the two partitions which reach both $\boldsymbol{c}$ and $\boldsymbol{c}'$ through the shortest path, respectively from below and from above.

2.2 Bayesian Mixture Models

From a statistical perspective, set partitions are key elements in a Bayesian mixture model framework. The main underlying assumption is that observations $y_1, \ldots, y_N$ are independent conditional on the partition $\boldsymbol{c}$, and their joint probability density can be expressed as

$$p(\mathbf{y}|\boldsymbol{c}, \boldsymbol{\theta}) = \prod_{k=1}^{K} \prod_{i \in B_k} p(y_i|\theta_k) = \prod_{k=1}^{K} p(\mathbf{y}_k|\theta_k), \tag{2.1}$$

with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ a vector of unknown parameters indexing the distribution of observations $\mathbf{y}_k = \{y_i\}_{i \in B_k}$ for each cluster $k = 1, \ldots, K$. In a Bayesian formulation, a prior distribution is assigned to each possible partition $\boldsymbol{c}$, leading to a posterior of the

Random probability measure	Parameters	$p(\boldsymbol{c}) =$
Dirichlet process	(α)	$\frac{\alpha^K}{(\alpha)_N}\prod_{j=1}^K(\lambda_j-1)!$
Pitman-Yor process	(α,σ)	$\frac{\prod_{j=1}^{K-1}(\alpha+j\sigma)}{(\alpha+1)_{(N-1)}}\prod_{j=1}^K(1-\sigma)_{(\lambda_j-1)}$
Symmetric Dirichlet	(κ,γ)	$\frac{\kappa!}{(\kappa-K)!}\prod_{j=1}^K\frac{\Gamma(\gamma/\kappa+\lambda_j)}{\Gamma(\gamma/\kappa)}$

Table 1: Exchangeable Partition Probability Function for Dirichlet, Pitman-Yor processes and Symmetric Dirichlet distribution; $\lambda_j = |B_j|$ is the cardinality of the clusters composing the partition, while $(x)_r = x(x+1)\cdots(x+r-1)$ denotes the rising factorial.

form

$$p(\boldsymbol{c}|\mathbf{y},\boldsymbol{\theta}) \propto p(\boldsymbol{c})\prod_{k=1}^{K} p(\mathbf{y}_k|\theta_k). \tag{2.2}$$

Hence the set partition $\boldsymbol{c}$ is conceived as a random object and elicitation of its prior distribution is a critical issue in Bayesian modeling.

The first distribution one may use, in the absence of prior information, is the uniform distribution, which gives the same probability to every partition with $p(\boldsymbol{c}) = 1/\mathcal{B}_N$; however, even for small values of N the Bell number $\mathcal{B}_N$ is very large, making computation of the posterior intractable even for simple choices of the likelihood. This motivated the definition of alternative prior distributions based on different concepts of uniformity, with the Jensen and Liu (2008) prior favoring uniform placement of new observations in one of the existing clusters, and Casella et al. (2014) proposing a hierarchical uniform prior, which gives equal probability to set partitions having the same configuration.

Usual Bayesian nonparametric procedures build instead on discrete nonparametric priors, i.e. priors that have discrete realization almost surely. Dirichlet and Pitman-Yor processes are well known to have this property, as does the broader class of Gibbs-type priors. Any discrete random probability measure $\tilde{p}$ can induce an exchangeable random partition. Due to the discreteness of the process, $\tilde{p}$ induces a partition of the observations $y_1, \ldots, y_N$ which can be characterized via an Exchangeable Probability Partition Function. For both Dirichlet and Pitman-Yor processes, the EPPF is available in closed form as reported in Table 1 along with the case of the finite mixture model with κ components and a symmetric Dirichlet prior with parameters $(\gamma/\kappa, \ldots, \gamma/\kappa)$. Notice that $\lambda_j = |B_j|$ is the cardinality of the clusters composing the partition, while notation $(x)_r$ is for the rising factorial $x(x+1)\cdots(x+r-1)$.

There is a strong connection with the exchangeable random partitions induced by Gibbs-type priors and product partition models. A product partition model assumes that the prior probability for the partition $\boldsymbol{c}$ has the following form

$$p(\boldsymbol{c} = \{B_1, \ldots, B_K\}) \propto \prod_{j=1}^{K} \rho(B_j), \tag{2.3}$$

with $\rho(\cdot)$ known as the cohesion function. The underlying assumption is that the prior distribution for the set partition $\boldsymbol{c}$ can be factorized as the product of functions that depend only on the blocks composing it. Such a definition, in conjunction with formulation (2.1) for the data likelihood, guarantees the property that the posterior distribution for $\boldsymbol{c}$ is still in the class of product partition models.

Distributions in Table 1 are all characterized by a cohesion function that depends on the blocks through their cardinality. Although the parameters can control the expected number of clusters, this assumption is too strict in many applied contexts in which prior information is available about the grouping. In particular, the same probability is given to partitions with the same configuration but having a totally different composition.

3 Centered Partition Processes

Our focus is on incorporating structured knowledge about clustering of the finite set of indices $[N] = \{1, \dots, N\}$ in the prior distribution within a Bayesian mixture model framework. We consider as a first source of information a given potential clustering, but our approach can also accommodate prior information on summary statistics such as the number of clusters and cluster sizes.

3.1 General Formulation

Assume that a potential clustering $\boldsymbol{c}_0$ is given and we wish to include this information in the prior distribution. To address this problem, we propose a general strategy to modify a baseline EPPF to shrink towards $\boldsymbol{c}_0$. In particular, our proposed CP process defines the prior on set partitions as proportional to a baseline EPPF multiplied by a penalization term of the type

$$p(\boldsymbol{c}|\boldsymbol{c}_0, \psi) \propto p_0(\boldsymbol{c}) e^{-\psi d(\boldsymbol{c}, \boldsymbol{c}_0)}, \tag{3.1}$$

with $\psi > 0$ a penalization parameter, $d(\boldsymbol{c}, \boldsymbol{c}_0)$ a suitable distance measuring how far $\boldsymbol{c}$ is from $\boldsymbol{c}_0$ and $p_0(\boldsymbol{c})$ indicates a baseline EPPF, that may depend on some parameters that are not of interest at the moment. For $\psi \to 0$, $p(\boldsymbol{c}|\boldsymbol{c}_0, \psi)$ corresponds to the baseline EPPF $p_0(\boldsymbol{c})$, while for $\psi \to \infty$, $p(\boldsymbol{c} = \boldsymbol{c}_0|\boldsymbol{c}_0, \psi) \to 1$.

Note that $d(\boldsymbol{c}, \boldsymbol{c}_0)$ takes a finite number of discrete values $\Delta = \{\delta_0, \dots, \delta_L\}$, with L depending on $\boldsymbol{c}_0$ and on the distance $d(\cdot, \cdot)$. We can define sets of partitions having the same fixed distance from $\boldsymbol{c}_0$ as

$$s_l(\boldsymbol{c}_0) = \{\boldsymbol{c} \in \Pi_N : d(\boldsymbol{c}, \boldsymbol{c}_0) = \delta_l\}, \quad l = 0, 1, \dots, L. \tag{3.2}$$

Hence, for $\delta_0 = 0$, $s_0(\boldsymbol{c}_0)$ denotes the set of partitions equal to the base one, meaning that they differ from $\boldsymbol{c}_0$ only by a permutation of the cluster labels. Then $s_1(\boldsymbol{c}_0)$ denotes the set of partitions with minimum distance δ_1 from $\boldsymbol{c}_0$, $s_2(\boldsymbol{c}_0)$ the set of partitions with the second minimum distance δ_2 from $\boldsymbol{c}_0$ and so on. The introduced exponential term penalizes equally partitions in the same set $s_l(\boldsymbol{c}_0)$ for a given δ_l, but the resulting probabilities may differ depending on the chosen baseline EPPF.

3.2 Choices of Distance Function

The proposed CP process modifies a baseline EPPF to include a distance-based penalization term, which aims to shrink the prior distribution towards a prior partition guess. The choice of distance plays a key role in determining the behavior of the prior distribution. A variety of different distances and indices have been employed in clustering procedures and comparisons. We consider in this paper the Variation of Information (VI), obtained axiomatically in Meilă (2007) using information theory, and shown to nicely characterize neighborhoods of a given partition by Wade and Ghahramani (2018). The Variation of Information is based on the Shannon entropy $H(\cdot)$, and can be computed as

$$\begin{aligned}\mathrm{VI}(\boldsymbol{c},\boldsymbol{c}') &= -H(\boldsymbol{c}) - H(\boldsymbol{c}_0) + 2H(\boldsymbol{c}\wedge\boldsymbol{c}_0)\\ &= \sum_{j=1}^{K}\frac{\lambda_j}{N}\log\left(\frac{\lambda_j}{N}\right) + \sum_{l=1}^{K'}\frac{\lambda_l'}{N}\log\left(\frac{\lambda_l'}{N}\right) - 2\sum_{j=1}^{K}\sum_{l=1}^{K'}\frac{\lambda_{jl}^{\wedge}}{N}\log\left(\frac{\lambda_{jl}^{\wedge}}{N}\right),\end{aligned}$$

where log denotes log base 2, and $\lambda_{jl}^{\wedge}$ the size of blocks of the intersection $\boldsymbol{c}\wedge\boldsymbol{c}'$ and hence the number of indices in block j under partition $\boldsymbol{c}$ and block l under $\boldsymbol{c}'$. Notice that VI ranges from 0 to $\log_2(N)$. Although normalized versions have been proposed (Vinh et al., 2010), some desirable properties are lost under normalization. We refer to Meilă (2007) and Wade and Ghahramani (2018) for additional properties and empirical evaluations.

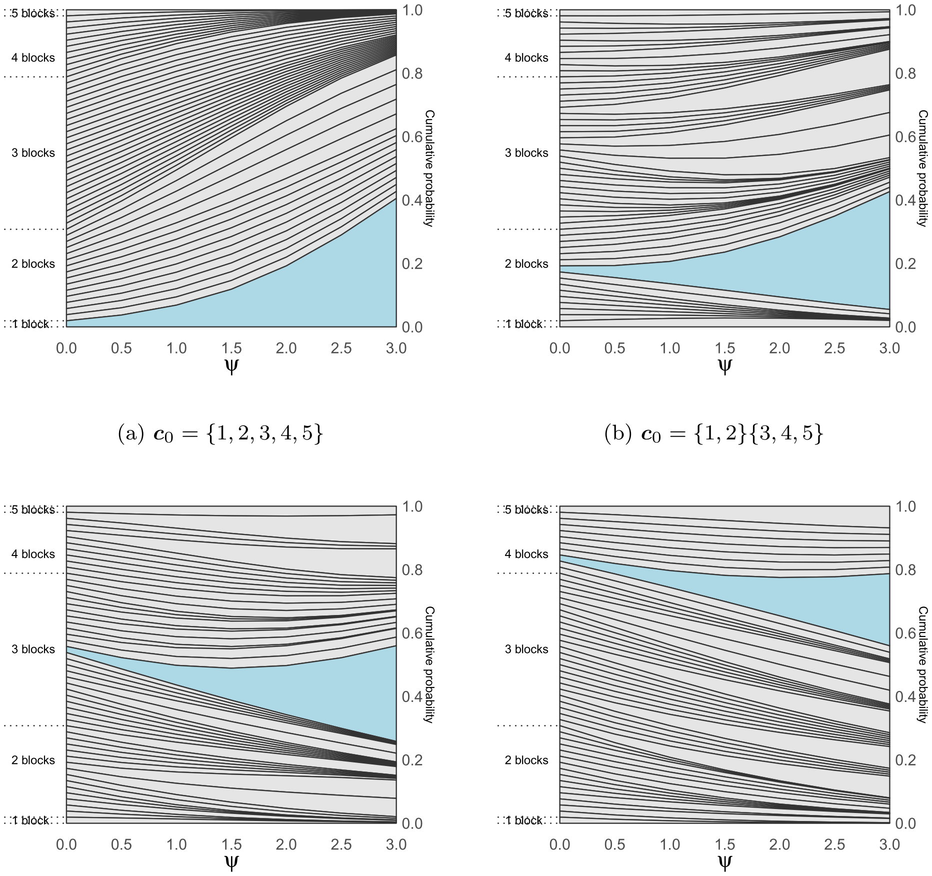
An alternative definition of the VI can be derived from lattice theory, exploiting the concepts provided in Section 2.1. We refer to Monjardet (1981) for general theory about metrics on lattices and ordered sets, and Rossi (2015) for a more recent review focused on set partitions. In general, a distance between two different partitions $\boldsymbol{c},\boldsymbol{c}' \in \Pi_N$ can be defined by means of the Hasse diagram via the minimum weighted path, which corresponds to the shortest path length when edges are equally weighted. Instead, when edges depend on the entropy function through $w(\boldsymbol{c},\boldsymbol{c}') = |H(\boldsymbol{c}) - H(\boldsymbol{c}')|$, the minimum weighted path between two partitions is the Variation of Information. Notice that two partitions are connected when in a covering relation, then $\boldsymbol{c}\wedge\boldsymbol{c}'$ is either equal to $\boldsymbol{c}$ or $\boldsymbol{c}'$ and $VI(\boldsymbol{c},\boldsymbol{c}') = w(\boldsymbol{c},\boldsymbol{c}')$. The minimum weight $w(\boldsymbol{c},\boldsymbol{c}')$ corresponds to $2/N$ which is attained when two singleton clusters are merged, or conversely, a cluster consisting of two points is split (see Meilă, 2007).

3.3 Effect of the Prior Penalization

We first consider the important special case in which the baseline EPPF is $p_0(\boldsymbol{c}) = 1/\mathcal{B}_N$ and the CP process reduces to $p(\boldsymbol{c}|\boldsymbol{c}_0,\psi) \propto \exp\{-\psi d(\boldsymbol{c},\boldsymbol{c}_0)\}$ with equation (3.1) simplifying to

$$p(\boldsymbol{c}|\boldsymbol{c}_0,\psi) = \frac{e^{-\psi\delta_l}}{\sum_{u=0}^{L}|s_u(\boldsymbol{c}_0)|e^{-\psi\delta_u}}, \quad \text{for } \boldsymbol{c}\in s_l(\boldsymbol{c}_0), \quad l = 0,1,\ldots,L, \tag{3.3}$$

where $|\cdot|$ indicates the cardinality and $s_l(\boldsymbol{c}_0)$ is defined in (3.2). Considering $N = 5$, there are 52 possible set partitions; Figure 2 shows the prior probabilities assigned to partitions



(a) $\boldsymbol{c}_0 = \{1, 2, 3, 4, 5\}$

(b) $\boldsymbol{c}_0 = \{1, 2\}\{3, 4, 5\}$

(c) $\boldsymbol{c}_0 = \{1, 2\}\{3, 4\}\{5\}$

(d) $\boldsymbol{c}_0 = \{1\}\{2\}\{3\}\{4, 5\}$

Figure 2: Prior probabilities of the 52 set partitions of $N = 5$ elements for the CP process with uniform base EPPF. In each graph the CP process is centered on a different partition $\boldsymbol{c}_0$ highlighted in blue. The cumulative probabilities across different values of the penalization parameter ψ are joined to form the curves, while the probability of a given partition corresponds to the area between the curves.

under the CP process for different values of $\psi \in (0, 3)$ with $\psi = 0$ corresponding to the uniform prior. Notice that base partitions with the same configuration (e.g. for $\boldsymbol{c}_0 = \{1, 2\}\{3, 4, 5\}$ all the partitions with blocks sizes $\{3, 2\}$), will behave in the same way, with the same probabilities assigned to partitions different in composition. Nonzero values of ψ increase the prior probability of partitions $\boldsymbol{c}$ that are relatively close to the chosen $\boldsymbol{c}_0$. However, the effect is not uniform but depends on the structure of both $\boldsymbol{c}$ and $\boldsymbol{c}_0$.

For example, consider the inflation that occurs in the blue region as ψ increases from 0 to 3. When $\boldsymbol{c}_0$ has 2 blocks (Figure 2a) versus 4 (Figure 2d) there is a bigger increase. This is because the space of set partitions Π_N is not "uniform", since given a fixed configuration there is a heterogeneous number of partitions. Expressing $\boldsymbol{\lambda} = \{\lambda_1, \dots, \lambda_K\}$ as $(1^{f_1}, 2^{f_2}, \dots, N^{f_N})$, with the notation indicating that there are f_i elements of $\boldsymbol{\lambda}$ equal to i, the number of set partitions with configuration $\boldsymbol{\lambda}$ is

$$\frac{N!}{\prod_{j=1}^{K} \lambda_j! \prod_{i=1}^{N} f_i!}.$$

For example, for $\{221\} = 1^1 2^2 3^0 4^0 5^0$, the number of corresponding set partitions is 15, while there are 10 set partitions of type $\{311\} = 1^2 2^0 3^1 4^0 5^0$.

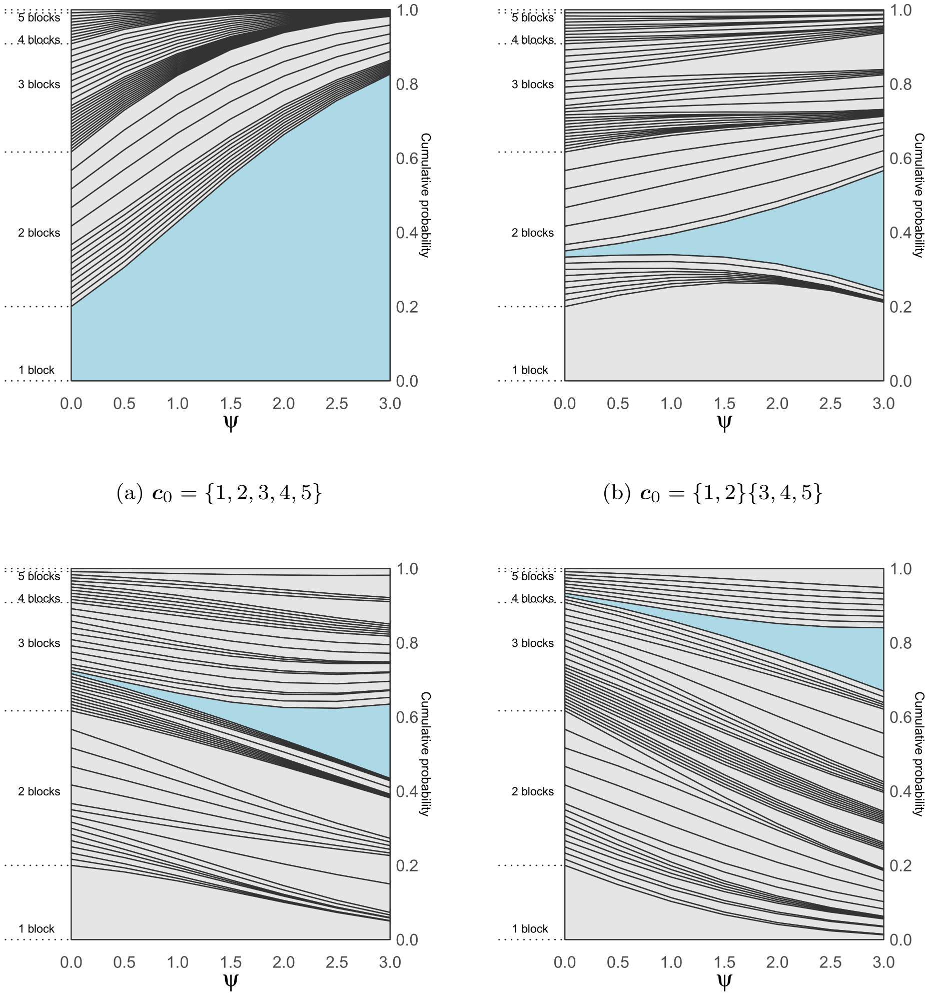
While the uniform distribution gives the same probability to each partition in the space, the EPPF induced by Gibbs-type priors distinguishes between different configurations, but not among partitions with the same configuration. We focus on the Dirichlet process case, being the most popular process employed in applications. Under the DP, the induced EPPF $p_0(\boldsymbol{c}) \propto \alpha^K \prod_{j=1}^{K} \Gamma(\lambda_j)$ is a function of the configuration $\boldsymbol{\Lambda}(\boldsymbol{c})$, which is one of $\{\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_M\}$ since the possible configurations are finite and correspond to the number of integer partitions. Letting $g(\boldsymbol{\Lambda}(\boldsymbol{c})) = \alpha^K \prod_{j=1}^{K} \Gamma(\lambda_j)$, the formulation in (3.1) can be written as

$$p(\boldsymbol{c}|\boldsymbol{c}_0, \psi) = \frac{g(\boldsymbol{\lambda}_m) e^{-\psi \delta_l}}{\sum_{u=0}^{L} \sum_{v=1}^{M} n_{uv} g(\boldsymbol{\lambda}_v) e^{-\psi \delta_u}}, \quad \text{for } \boldsymbol{c} \in s_{lm}(\boldsymbol{c}_0), \tag{3.4}$$

where $s_{lm}(\boldsymbol{c}_0) = \{\boldsymbol{c} \in \Pi_N : d(\boldsymbol{c}, \boldsymbol{c}_0) = \delta_l, \boldsymbol{\Lambda}(\boldsymbol{c}) = \boldsymbol{\lambda}_m\}$, the set of partitions with distance δ_l from $\boldsymbol{c}_0$ and configuration $\boldsymbol{\lambda}_m$ for $l = 0, 1, \dots, L$ and $m = 1, \dots, M$, with n_{lm} indicating the cardinality. The factorization (3.4) applies for the family of Gibbs-type priors in general, with different expressions of $g(\boldsymbol{\Lambda}(\boldsymbol{c}))$.

In Figure 3 we consider the prior distribution induced by the CP process when the baseline EPPF $p_0(\boldsymbol{c})$ comes from a Dirichlet process with concentration parameter $\alpha = 1$, considering the same base partitions and values for ψ as in Figure 2. For the same values of the parameter ψ, the behavior of the CP process changes significantly due to the effect of the base prior. In particular, in the top left panel the CP process is centered on $\boldsymbol{c}_0 = \{1, 2, 3, 4, 5\}$, the partition with only one cluster, which is *a priori* the most likely one for $\psi = 0$. In general, for small values of ψ the clustering process will most closely resemble that for a DP. As ψ increases, the DP prior probabilities decrease for partitions far from $\boldsymbol{c}_0$ while increase for partitions close to $\boldsymbol{c}_0$.

Finally we investigate in Figures 4–5 what happens to the prior partition probabilities of the CP process, when the baseline EPPF comes from a Pitman-Yor process. To allow comparison with the DP case, we choose the strength parameter α such that the a priori expected number of clusters matches the one under the DP case, $\log(5) \approx 1.6$. Choosing values of $\sigma = (0.25, 0.75)$ leads to values of the strength parameter α equal to $(-0.004, -0.691)$ respectively. It can be noticed that for the smaller value of the discount parameter $\sigma = 0.25$ (Figure 4) the graphs resemble more the ones related to the DP process, while for $\sigma = 0.75$ the prior probability mass concentrates more on partitions with number of clusters close to the expected one.



(a) $\boldsymbol{c}_0 = \{1, 2, 3, 4, 5\}$

(b) $\boldsymbol{c}_0 = \{1, 2\}\{3, 4, 5\}$

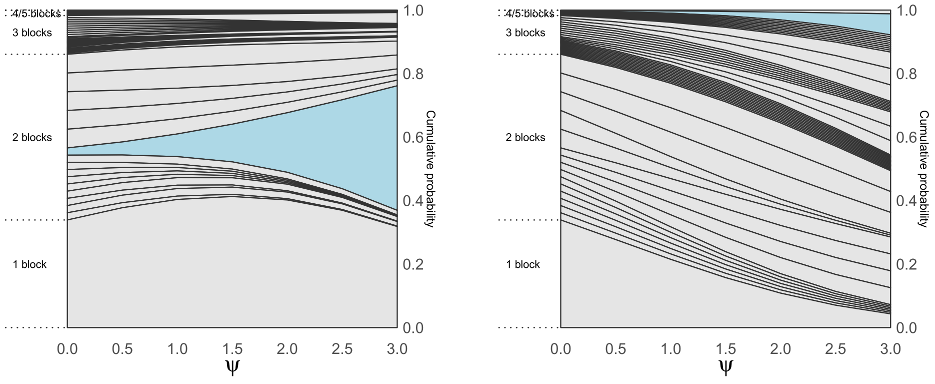
(c) $\boldsymbol{c}_0 = \{1, 2\}\{3, 4\}\{5\}$

(d) $\boldsymbol{c}_0 = \{1\}\{2\}\{3\}\{4, 5\}$

Figure 3: Prior probabilities of the 52 set partitions of $N = 5$ elements for the CP process with Dirichlet process of $\alpha = 1$ base EPPF. In each graph the CP process is centered on a different partition $\boldsymbol{c}_0$ highlighted in blue. The cumulative probabilities across different values of the penalization parameter ψ are joined to form the curves, while the probability of a given partition corresponds to the area between the curves.

3.4 Posterior Computation Under Gibbs-Type Priors

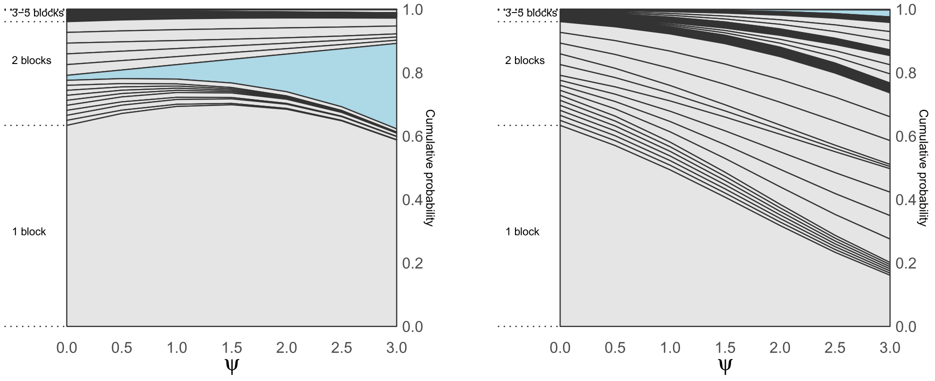
Certain MCMC algorithms for Bayesian nonparametric mixture models can be easily modified for posterior computation in CP process models. In particular, we adapt the so-



(a) $\boldsymbol{c}_0 = \{1, 2\}\{3, 4, 5\}$

(b) $\boldsymbol{c}_0 = \{1, 2\}\{3, 4\}\{5\}$

Figure 4: Prior probabilities of the 52 set partitions of $N = 5$ elements for the CP process with Pitman-Yor process base EPPF with $\sigma = 0.25$ and $\alpha \approx -0.004$, such that the expected number of clusters equal to $\log(5) \approx 1.6$. In each graph the CP process is centered on a different partition $\boldsymbol{c}_0$ highlighted in blue. The cumulative probabilities across different values of the penalization parameter ψ are joined to form the curves, while the probability of a given partition corresponds to the area between the curves.



(a) $\boldsymbol{c}_0 = \{1, 2\}\{3, 4, 5\}$

(b) $\boldsymbol{c}_0 = \{1, 2\}\{3, 4\}\{5\}$

Figure 5: Prior probabilities of the 52 set partitions of $N = 5$ elements for the CP process with Pitman-Yor process base EPPF with $\sigma = 0.75$ and $\alpha \approx -0.691$, such that the expected number of clusters equal to $\log(5) \approx 1.6$. In each graph the CP process is centered on a different partition $\boldsymbol{c}_0$ highlighted in blue. The cumulative probabilities across different values of the penalization parameter ψ are joined to form the curves, while the probability of a given partition corresponds to the area between the curves.

Random probability measure	Parameters	$p(c_i = k \mid \boldsymbol{c}^{-i}) \propto$
Dirichlet process	(α)	$\begin{cases} \frac{\lambda_k^{-i}}{\alpha+N-1} & k = 1, \dots, K^- \\ \frac{\alpha}{\alpha+N-1} & k = K^- + 1 \end{cases}$
Pitman-Yor process	(α, σ)	$\begin{cases} \frac{\lambda_k^{-i}-\sigma}{\alpha+N-1} & k = 1, \dots, K^- \\ \frac{\alpha+\sigma K^-}{\alpha+N-1} & k = K^- + 1 \end{cases}$
Symmetric Dirichlet	(κ, γ)	$\frac{\lambda_k^{-i}+\gamma/\kappa}{\alpha+N-1} \quad k = 1, \dots, \kappa$

Table 2: Conditional prior distribution for c_i given $\boldsymbol{c}^{-i}$ under different choices of the EPPF. With K^- we denote the total number of clusters after removing the ith observation while λ_k^{-i} is the corresponding size of cluster k.

called "marginal algorithms" developed for Dirichlet and Pitman-Yor processes. These methods are called marginal since the mixing measure P is integrated out of the model and the predictive distribution is used within a MCMC sampler. In the following, we recall Algorithm 2 in Neal (2000) and illustrate how it can be adapted to sample from the CP process posterior. We refer to Neal (2000) and references therein for an overview and discussion of methods for both conjugate and nonconjugate cases, and to Fall and Barat (2014) for adaptation to Pitman-Yor processes.

Let $\boldsymbol{c}$ be represented as an N-dimensional vector of indices $\{c_1, \dots, c_N\}$ encoding cluster allocation and let θ_k be the set of parameters currently associated to cluster k. The prior predictive distribution for a single c_i conditionally on $\boldsymbol{c}^{-i} = \{c_1, \dots, c_{i-1}, c_{i+1}, \dots, c_N\}$ is exploited to perform the Gibbs sampling step allocating observations to either a new cluster or one of the existing ones. Algorithm 2 in Neal (2000) updates each c_i sequentially for $i = 1, \dots, N$ via a reseating procedure, according to the conditional posterior distribution

$$p(c_i = k \mid \boldsymbol{c}^{-i}, \boldsymbol{\theta}, y_i) \propto \begin{cases} p(c_i = k \mid \boldsymbol{c}^{-i}) p(y_i \mid \theta_k) & k = 1, \dots, K^- \\ p(c_i = k \mid \boldsymbol{c}^{-i}) \int p(y_i \mid \theta) dG_0(\theta) & k = K^- + 1, \end{cases} \tag{3.5}$$

with K^- the number of clusters after removing observation i. The conditional distribution $p(c_i = k \mid \boldsymbol{c}^{-i})$ is reported in Table 2 for different choices of the prior EPPF. Notice that, for the case of finite Dirichlet prior, the update consists only in the first line of equation (3.5), since the number of classes is fixed. For Dirichlet and Pitman-Yor processes, when observation i is associated to a new cluster, a new value for θ is sampled from its posterior distribution based on the base measure G_0 and the observation y_i. This approach is straightforward when we can compute the integral $\int p(y_i \mid \theta) dG_0(\theta)$, as will generally be the case when G_0 is a conjugate prior.

Considering the proposed CP process, the conditional distribution for c_i given $\boldsymbol{c}^{-i}$ can still be computed, but it depends both on the base prior and the penalization term accounting for the distance between the base partition $\boldsymbol{c}_0$ and the one obtained by assigning the observation i to either one of the existing classes $k \in \{1, \dots, K^-\}$ or a new one. Hence, the step in equation (3.5) can be easily adapted by substituting the conditional distribution for $p(c_i = k|\boldsymbol{c}^{-i})$ with

$$p(c_i = k|\boldsymbol{c}^{-i}, \boldsymbol{c}_0, \psi) \propto p_0(c_i = k|\boldsymbol{c}^{-i}) \exp\{-\psi d(\boldsymbol{c}, \boldsymbol{c}_0)\} \quad k = 1, \ldots, K^-, K^- + 1,$$

with $\boldsymbol{c} = \{\boldsymbol{c}^{-i} \cup \{c_i = k\}\}$ the current state of the clustering and $p_0(c_i = k|\boldsymbol{c}^{-i})$ one of the conditional distributions in Table 2. Additional steps on the implementation using the variation of information as a distance are given in the Supplementary Material (Algorithm 2).

Extension to the non-conjugate context can be similarly handled exploiting Algorithm 8 in Neal (2000) based on auxiliary parameters, which avoids the computation of the integral $\int p(y_i|\theta)dG_0(\theta)$. The only difference is that, when c_i is updated, m temporary auxiliary variables are introduced to represent possible values of components parameters that are not associated with any other observations. Such variables are simply sampled from the base measure G_0, with the probabilities of a new cluster in Table 2 changing into $(\alpha/m)/(\alpha + N - 1)$ for the Dirichlet process and to $[(\alpha + \sigma K^-)/m]/(\alpha + N - 1)$ for the Pitman-Yor process, for $k = K^- + 1, \ldots, K^- + m$.

4 Prior Calibration

As the number of observations N increases, the number of partitions explodes, and higher values of ψ are needed to place non-negligible prior probability in small to moderate neighborhoods around $\boldsymbol{c}_0$. The prior concentration around $\boldsymbol{c}_0$ depends on three main factors: i) N through $\mathcal{B}_N$, i.e. the cardinality of the space of set partitions, ii) the baseline EPPF $p_0(\boldsymbol{c})$ and iii) where $\boldsymbol{c}_0$ is located in the space. We hence propose a general method to evaluate the prior behavior under different settings, while suggesting how to choose the parameter ψ.

One may evaluate the prior distribution for different values of ψ and check its behavior using graphs such as those in Section 3.3, but they become difficult to interpret as the space of partitions grows. We propose to evaluate the probability distribution of the distances $\delta = d(\boldsymbol{c}, \boldsymbol{c}_0)$ from the known partition $\boldsymbol{c}_0$. The probability assigned to different distances by the prior is

$$p(\delta = \delta_l) = \sum_{\boldsymbol{c} \in \Pi_N} p(\boldsymbol{c}) \mathcal{I}\{d(\boldsymbol{c}, \boldsymbol{c}_0) = \delta_l)\} = \sum_{\boldsymbol{c} \in s_l(\boldsymbol{c}_0)} p(\boldsymbol{c}) \quad l = 0, \ldots, L,$$

with $\mathcal{I}(\cdot)$ the indicator function and $s_l(\boldsymbol{c}_0)$ denoting the set of partitions distance δ_l from $\boldsymbol{c}_0$, as defined in (3.2). Considering the uniform distribution on set partitions, then $p(\delta = \delta_l) = |s_l(\boldsymbol{c}_0)|/\mathcal{B}_N$, the proportion of partitions distance δ_l from $\boldsymbol{c}_0$. Under the general definition of the CP process, the resulting distribution becomes

$$p(\delta = \delta_l) = \sum_{\boldsymbol{c} \in s_l(\boldsymbol{c}_0)} \frac{p_0(\boldsymbol{c}) e^{-\psi \delta_l}}{\sum_{u=0}^{L} \sum_{\boldsymbol{c}^* \in s_u(\boldsymbol{c}_0)} p_0(\boldsymbol{c}^*) e^{-\psi \delta_u}} \quad l = 0, \ldots, L, \tag{4.1}$$

with the case of Gibbs-type EPPF corresponding to

$$p(\delta = \delta_l) = \frac{\sum_{m=1}^{M} n_{lm} g(\boldsymbol{\lambda}_m) e^{-\psi \delta_l}}{\sum_{u=0}^{L} \sum_{v=1}^{M} n_{uv} g(\boldsymbol{\lambda}_v) e^{-\psi \delta_u}}, \quad l = 0, \ldots, L. \tag{4.2}$$

Notice that the uniform EPPF case is recovered when $g(\boldsymbol{\lambda}_m) = 1$ for $m = 1, \dots, M$, so that $\sum_{m=1}^{M} n_{lm} = n_l$. Hence the probability in (4.1) simplifies to

$$p(\delta = \delta_l) = \frac{n_l e^{-\psi \delta_l}}{\sum_{u=0}^{L} n_u e^{-\psi \delta_u}} \quad l = 0, \dots, L. \tag{4.3}$$

In general, since distances are naturally ordered, the corresponding cumulative distribution function can be simply defined as $F(\delta) = \sum_{\delta_l \leq \delta} p(\delta_l)$ for $\delta \in \{\delta_0, \dots, \delta_L\}$ and used to assess how much mass is placed in different size neighborhoods around $\boldsymbol{c}_0$ under different values of ψ. Hence we can choose ψ to place a specified probability q (e.g. $q = 0.9$) on partitions within a specified distance δ^* from $\boldsymbol{c}_0$. This would correspond to calibrating ψ so that $F(\delta^*) \approx q$, with $F(\delta^*) \geq q$. In other words, partitions generated from the prior would have at least probability q of being within distance δ^* from $\boldsymbol{c}_0$.

The main problem is in computing the probabilities in equations (4.2)–(4.3), which depend on all the set partitions in the space. In fact, one needs to count all the partitions having distance δ_l for $l = 0, \dots, L$ when the base EPPF is uniform, while taking account of configurations in the case of the Gibbs-type priors. Even if there are quite efficient algorithms to list all the possible set partitions of N (see Knuth, 2005; Nijenhuis and Wilf, 2014), it becomes computationally infeasible due to the extremely rapid growth of the space; for example from $N = 12$ to 13, the number of set partitions grows from $\mathcal{B}_{12} = 4,213,597$ to $\mathcal{B}_{13} = 27,644,437$.

Given that our motivating application involves a relatively small number of birth defects, we propose to directly approximate the prior probabilities assigned to different distances from $\boldsymbol{c}_0$. We focus on obtaining estimates of distance values and related counts, which are the sufficient quantities to compute (4.2)–(4.3) under different values of ψ. We propose a strategy based on a targeted Monte Carlo procedure which augments uniform sampling on the space of set partitions with a deterministic local search using the Hasse diagram to compute counts for small values of the distance. Although the procedure is generalizable to higher dimensions, the computational burden grows significantly with larger numbers of objects to cluster. Alternative computational directions are considered further in the Discussion.

4.1 Deterministic Local Search

Poset theory provides a nice representation of the space of set partitions by means of the Hasse diagram illustrated in Section 2.1, along with suitable definition of metrics. A known partition $\boldsymbol{c}_0$ can be characterized in terms of number of blocks K_0 and configuration $\boldsymbol{\Lambda}(\boldsymbol{c}_0)$. These elements allow one to locate $\boldsymbol{c}_0$ in the Hasse diagram, and hence explore connected partitions by means of split and merge operations on the clusters in $\boldsymbol{c}_0$.

As an illustrative example, consider the Hasse diagram of Π_4 in Figure 6 and $\boldsymbol{c}_0 = \{1\}\{2,3,4\}$, having 2 clusters and configuration $\boldsymbol{\Lambda}(\boldsymbol{c}_0) = \{31\}$. Let $\mathcal{N}_1(\boldsymbol{c}_0)$ denote the sets of partitions directly connected with $\boldsymbol{c}_0$, i.e. partitions covering $\boldsymbol{c}_0$ and those covered by $\boldsymbol{c}_0$. In general, a partition $\boldsymbol{c}_0$ with K_0 clusters is covered by $\binom{K_0}{2}$ partitions and covers

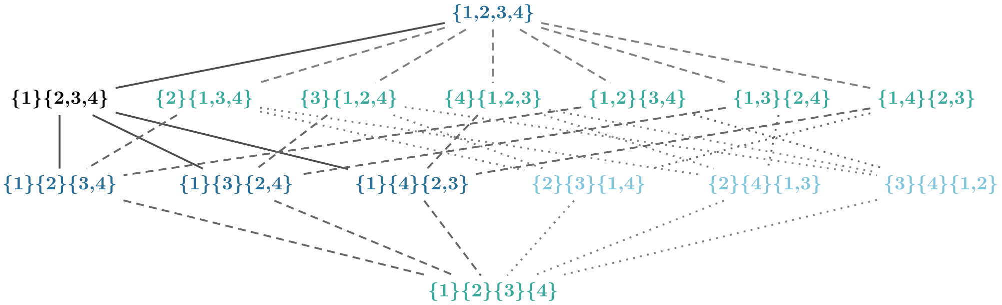


Figure 6: Illustration of results from the local search algorithm based on the Hasse diagram of Π_4 starting from $\boldsymbol{c}_0$ = **{1}{2,3,4}**. Partitions are colored according the exploration order following a dark-light gradient. Notice that after 3 iterations the space is entirely explored.

$\sum_{j=1}^{K_0} 2^{\lambda_j - 1} - 1$. In the example, $\mathcal{N}_1(\boldsymbol{c}_0)$ contains $\{1, 2, 3, 4\}$ obtained from $\boldsymbol{c}_0$ with a merge operation on the two clusters, and all the partitions obtained by splitting the cluster $\{2, 3, 4\}$ in any possible way. The idea underlying the proposed local search, consists in exploiting the Hasse diagram representation to find all the partitions in increasing distance neighborhoods of $\boldsymbol{c}_0$. One can list partitions at T connections from $\boldsymbol{c}_0$ starting from $\mathcal{N}_1(\boldsymbol{c}_0)$ by recursively applying split and merge operations on the set of partitions explored at each step. Potentially, with enough operations one can reach all the set partitions, since the space is finite with lower and upper bounds.

In practice, the space is too huge to be explored entirely, and a truncation is needed. From the example in Figure 6, $\mathcal{N}_1(\boldsymbol{c}_0)$ contains 3 partitions with distance 0.69 from $\boldsymbol{c}_0$ and one with distance 1.19. Although $\mathcal{N}_2(\boldsymbol{c}_0)$ may contain partitions closer to $\boldsymbol{c}_0$ than this last, the definition of distance in Section 3.2 guarantees that there are no other partitions with distance from $\boldsymbol{c}_0$ less than 0.69. Since the VI is the minimum weighted path between two partitions, all the partitions reached at the second exploration step add a nonzero weight to distance computation. This consideration extends to an arbitrary number of explorations T, with $\delta_{L^*} = \min\{d(\boldsymbol{c}^*, \boldsymbol{c}_0)\}_{\boldsymbol{c}^* \in \mathcal{N}_T(\boldsymbol{c}_0)}$ being the upper bound on the distance value. By discarding all partitions with distance greater that δ_{L^*}, one can compute exactly the counts in equations (4.2)–(4.3) related to distances $\delta_0, \dots, \delta_{L^*}$. Notice that $2/N$ is the minimum distance between two different partitions, and $2T/N$ is a general lower bound on the distances from $\boldsymbol{c}_0$ that can be reached in T iterations.

4.2 Monte Carlo Approximation

We pair the local exploration with a Monte Carlo procedure to estimate the counts and distances greater that δ_{L^*}, in order to obtain a more refined representation of the prior distance probabilities. Sampling uniformly from the space of partitions is not in general a trivial problem, but a nice strategy has been proposed in Stam (1983), in which the probability of a partition with K clusters is used to sample partitions via an urn model.

Derivation of the algorithm starts from the *Dobiński formula* (Dobiński, 1877) for the Bell numbers

$$\mathcal{B}_N = e^{-1} \sum_{k=1}^{\infty} \frac{k^N}{k!}, \tag{4.4}$$

which from a probabilistic perspective corresponds to the N-th moment of the Poisson distribution with expected value equal to 1. Then a probability distribution for the number of clusters $K \in \{1, 2, 3, \ldots\}$ of a set partition can be defined as

$$P(K = k) = e^{-1} \frac{k^N}{\mathcal{B}_N k!}, \tag{4.5}$$

which is a well defined law thanks to (4.4). To simulate a uniform law over Π_N, Stam (1983)'s algorithm first generates the number of clusters K according to (4.5) and, conditionally on the sampled value, it allocates observations to the clusters according a discrete uniform distribution over $\{1, \ldots, K\}$. We refer to Stam (1983) and Pitman (1997) (Proposition 2, Corollary 3) for derivations and proof of the validity of the algorithm.

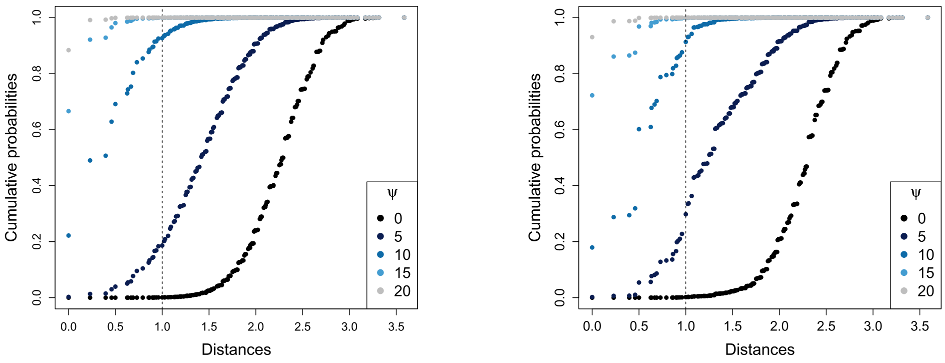
We adapt the uniform sampling to account for the values already computed by rejecting all the partitions with distance less that δ_{L^*}, restricting the space to $\big\{\Pi_N \setminus \{\mathcal{N}_t(\boldsymbol{c}_0)\}_{t=0}^T\big\}$. In practice, few samples are discarded since the probability to sample one such partition corresponds to $|\{\mathcal{N}_t(\boldsymbol{c}_0)\}_{t=0}^T|/\mathcal{B}_N$, which is negligible for small values of exploration steps T that are generally used in the local search. A sample of partitions $\boldsymbol{c}^{(1)}, \ldots, \boldsymbol{c}^{(R)}$, can be used to provide an estimate of the counts. Let R^* denote the number of accepted partitions and $\mathcal{B}^* = \mathcal{B}_N - |\{\mathcal{N}_t(\boldsymbol{c}_0)\}_{t=0}^T|$ be the number of partitions in the restricted space. Conditionally on the observed values of distances in the sample, $\hat{\delta}_{(L^*+1)}, \ldots, \hat{\delta}_{L'}$, an estimate of the number of partitions with distance $\hat{\delta}_l$ to use in the uniform EPPF case is

$$\hat{n}_l = \mathcal{B}^* \frac{1}{R^*} \sum_{r=1}^{R^*} \mathcal{I}\left\{ d(\boldsymbol{c}^{(r)}, \boldsymbol{c}_0) = \hat{\delta}_l \right\}, \tag{4.6}$$

obtained by multiplying the proportions of partitions in the sample by the total known number of partitions. For the Gibbs-type EPPF case one needs also to account for the configurations $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_M$ in a given orbital of the distance; hence, the estimates are

$$\hat{n}_{lm} = \mathcal{B}^* \frac{1}{R^*} \sum_{r=1}^{R^*} \mathcal{I}\left\{ d(\boldsymbol{c}^{(r)}, \boldsymbol{c}_0) = \hat{\delta}_l \right\} \mathcal{I}\left\{ \boldsymbol{\Lambda}(\boldsymbol{c}^{(r)}) = \boldsymbol{\lambda}_m \right\}. \tag{4.7}$$

Pairing these estimates with the counts obtained via the local search, one can evaluate the distributions in equations (4.2)–(4.3) for different values of ψ. The entire procedure is summarized in Algorithm 1 of the Supplementary Material. Although it requires a considerable number of steps, the procedure can be performed one single time providing information for different choices of ψ and EPPFs. Moreover the local search can be implemented in parallel to reduce computational costs.



ψ	0	5	10	15	20
$\min \delta : F(\delta) \geq 0.9$	2.68	1.97	0.90	0.5	0.22

Uniform EPPF

ψ	0	5	10	15	20
$\min \delta : F(\delta) \geq 0.9$	2.71	2.04	1.27	0.5	0

DP ($\alpha = 1$) EPPF

Figure 7: Estimate of the cumulative prior probabilities assigned to different distances from $\boldsymbol{c}_0$ for $N = 12$ and $\boldsymbol{c}_0$ with configuration $\{3, 3, 3, 3\}$, under the CP process with uniform prior on the left and Dirichlet Process on the right. Black dots correspond to the base prior with no penalization, while dots from bottom-to-top correspond to increasing values of $\psi \in \{5, 10, 15, 20\}$. Tables report the minimum distance values such that $F(\delta) \geq 0.9$.

We consider an example for $N = 12$ and $\boldsymbol{c}_0$ with configuration $\{3, 3, 3, 3\}$. Figure 7 shows the resulting cumulative probability estimates of the CP process under uniform and DP($\alpha = 1$) base distributions, estimated with $T = 4$ iterations of the local search and $20,000$ samples. Dots represent values of the cumulative probabilities, with different colors in correspondence to different values of the parameter ψ. Using these estimates one can assess how much probability is placed in different distance neighborhoods of $\boldsymbol{c}_0$; tables in Figure 7 show the distance values defining neighborhoods around $\boldsymbol{c}_0$ with 90% prior probability. If one wishes to place such probability mass on partitions within distance 1 from $\boldsymbol{c}_0$, a value of ψ around 10 and 15 is needed, respectively, under uniform and DP base EPPF prior.

5 The National Birth Defects Prevention Study

The National Birth Defects Prevention Study (NBDPS) is a multi-state population-based, case-control study of birth defects in the United States (Yoon et al., 2001). Infants were identified using birth defects surveillance systems in recruitment areas within ten US states (Arkansas, California, Georgia, Iowa, Massachusetts, New Jersey, New York, North Carolina, Texas, and Utah), which cover roughly 10% of US births. Diagnostic case information was obtained from medical records and verified by a standardized clinician review specific to the study (Rasmussen et al., 2003). Participants in the study included mothers with expected dates of delivery from 1997–2009. Controls were identified from birth certificates or hospital records and were live-born infants without any known birth defects. Each state site attempted to recruit 300 cases and 100 (unmatched) controls annually. A telephone interview was conducted with case and control mothers to solicit a wide range of demographic, lifestyle, medical, nutrition, occupational and environmental exposure history information.

Because birth defects are highly heterogeneous, a relatively large number of defects of unknown etiology are included in the NBDPS. We are particularly interested in congenital heart defects (CHD), the most common type of birth defect and the leading cause of infant death due to birth defects. Because some of these defects are relatively rare, in many cases we lack precision for investigating associations between potential risk factors and individual birth defects. For this reason, researchers typically lump embryologically distinct and potentially etiologically heterogeneous defects in order to increase power (e.g., grouping all heart defects together), even knowing the underlying mechanisms may differ substantially. In fact, how best to group defects is subject to uncertainty, despite a variety of proposed groupings available in the literature (Lin et al., 1999).

In this particular application, we consider 26 individual heart defects, which have been previously grouped into 6 categories by investigators (Botto et al., 2007). The prior grouping is shown in Table 3, along with basic summary statistics of the distribution of defects in the analyzed data. We are interested in evaluating the association between heart defects and about 90 potential risk factors related to mothers' health status, pregnancy experience, lifestyle and family history. We considered a subset of data from NBDPS, excluding observations with missing covariates, obtaining a dataset with 8,125 controls, while all heart defects together comprise 4,947 cases.

5.1 Modeling Birth Defects

Standard approaches assessing the impact of exposure factors on the risk to develop a birth defect often rely on logistic regression analysis. Let $i = 1, \dots, N$ index birth defects, while $j = 1, \dots, n_i$ indicates observations related to birth defect i, with $y_{ij} = 1$ if observation j has birth defect i and $y_{ij} = 0$ if observation j is a control, i.e. does not have any birth defect. Let $\mathbf{X}_i$ denote the data matrix associated to defect i, with each row $\mathbf{x}_{ij}^T = (x_{ij1}, \dots, x_{ijp})$ being the vector of the observed values of p categorical variables for the jth observation. At first one may consider N separate logistic regressions of the type

$$\log\left(\frac{\Pr(y_{ij} = 1|\mathbf{x}_{ij})}{\Pr(y_{ij} = 0|\mathbf{x}_{ij})}\right) = \operatorname{logit}(\boldsymbol{\pi}_{ij}) = \alpha_i + \mathbf{x}_{ij}^T\boldsymbol{\beta}_i, \tag{5.1}$$

with α_i denoting the defect-specific intercept, and $\boldsymbol{\beta}_i$ the $p \times 1$ vector of regression coefficients. However, Table 3 highlights the heterogeneity of heart defect prevalences, with some of them being so few as to preclude separate analyses.

A first step in introducing uncertainty about clustering of the defects may rely on a standard Bayesian nonparametric approach, placing a Dirichlet process prior on the distribution of regression coefficient vector $\boldsymbol{\beta}_i$ in order to borrow information across multiple defects while letting the data inform on the number and composition of the

Congenital Heart Defect	Abbreviation	Frequency	Percentage of cases
Septal			
Atrial septal defect	ASD	765	0.15
Perimembranous ventricular septal defect	VSDPM	552	0.11
Atrial septal defect, type not specified	ASDNOS	225	0.04
Muscular ventricular septal defect	VSDMUSC	68	0.02
Ventricular septal defect, otherwise specified	VSDOS	12	0.00
Ventricular septal defect, type not specified	VSDNOS	8	0.00
Atrial septal defect, otherwise specified	ASDOS	4	0.00
Conotruncal			
Tetralogy of Fallot	FALLOT	639	0.12
D-transposition of the great arteries	DTGA	406	0.08
Truncus arteriosus	COMMONTRUNCUS	61	0.01
Double outlet right ventricle	DORVTGA	35	0.01
Ventricular septal defect reported as conoventricular	VSDCONOV	32	0.01
D-transposition of the great arteries, other type	DORVOTHER	22	0.00
Interrupted aortic arch type B	IAATYPEB	13	0.00
Interrupted aortic arch, not otherwise specified	IAANOS	5	0.00
Left ventricular outflow			
Hypoplastic left heart syndrome	HLHS	389	0.08
Coarctation of the aorta	COARCT	358	0.07
Aortic stenosis	AORTICSTENOSIS	224	0.04
Interrupted aortic arch type A	IAATYPEA	12	0.00
Right ventricular outflow			
Pulmonary valve stenosis	PVS	678	0.13
Pulmonary atresia	PULMATRESIA	100	0.02
Ebstein anomaly	EBSTEIN	66	0.01
Tricuspid atresia	TRIATRESIA	46	0.01
Anomalous pulmonary venous return			
Total anomalous pulmonary venous return	TAPVR	163	0.03
Partial anomalous pulmonary venous return	PAPVR	21	0.01
Atrioventricular septal defect			
Atrioventricular septal defect	AVSD	112	0.02

Table 3: Summary statistics of the distribution of congenital heart defects among cases. Defects are divided according the grouping provided from investigators.

clusters. A similar approach has been previously proposed in MacLehose and Dunson (2010), with the aim being to shrink the coefficient estimates towards multiple unknown means. In our setting, an informed guess on the group structure is available through $\boldsymbol{c}_0$, reported in Table 3.

We consider a simple approach building on the Bayesian version of the model in (5.1), and allowing the exposure coefficients $\boldsymbol{\beta}_i$ for $i = 1, \dots, N$ to be shared across regressions, while accounting for $\boldsymbol{c}_0$. The model written in a hierarchical form is

$$\begin{aligned}
y_{ij} &\sim Ber(\pi_{ij}) & \operatorname{logit}(\pi_{ij}) &= \alpha_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}_{c_i}, \quad j = 1, \dots, n_i, \\
\alpha_i &\sim \mathcal{N}(a_0, \tau_0^{-1}) & \boldsymbol{\beta}_{c_i} | \boldsymbol{c} &\sim \mathcal{N}_p(\mathbf{b}, \mathbf{Q}) \quad i = 1, \dots, N, \\
p(\boldsymbol{c}) &\sim CP(\boldsymbol{c}_0, \psi, p_0(\boldsymbol{c})) & p_0(\boldsymbol{c}) &\propto \alpha^K \prod_{k=1}^{K} (\lambda_k - 1)!
\end{aligned} \tag{5.2}$$

where $CP(\boldsymbol{c}_0, \psi, p_0(\boldsymbol{c}))$ indicates the Centered Partition process, with base partition $\boldsymbol{c}_0$, tuning parameter ψ and baseline EPPF $p_0(\boldsymbol{c})$. We specify the baseline EPPF so that when $\psi = 0$ the prior distribution reduces to a Dirichlet Process with concentration parameter α. Instead, for $\psi \to \infty$ the model corresponds to K separate logistic regressions, one for each group composing $\boldsymbol{c}_0$. The model estimation can be performed by leveraging a Pòlya-Gamma data-augmentation strategy for Bayesian logistic regression (Polson et al., 2013), combined with the procedure illustrated in Section 3.4 for the clustering update step. The Gibbs sampler is detailed in the Supplementary Material (Algorithm 3), while code is available at https://github.com/salleuska/CPLogit.

5.2 Simulation Study

We conduct a simulation study to evaluate the performance of our approach in accurately estimating the impact of the covariates across regressions with common effects, under different prior guesses. In this section we choose a scenario mimicking the structure of our application. An additional simulation study under a continuous setting can be found in the Supplementary Material.

In simulating data we take a number of defects $N = 12$ equally partitioned in 4 groups and consider $p = 10$ dichotomous explanatory variables, under the assumption that defects in the same group have the same covariates effects. We take a different number of observations across defects, with $\{n_1, n_2, n_3\} = \{100, 600, 200\}$, $\{n_4, n_5, n_6\} = \{300, 100, 100\}$, $\{n_7, n_8, n_9\} = \{500, 100, 200\}$, $\{n_{10}, n_{11}, n_{12}\} = \{200, 200, 200\}$. For each defect i with $i = 1, \dots, 12$ we generate a data matrix $\mathbf{X}_i$ by sampling each of the variables from a Bernoulli distribution with probability of success equal to 0.5. We set most of coefficients $\beta_{i1}, \dots, \beta_{i10}$ to 0, while defining a challenging scenario with small to moderate changes across different groups. In particular we fix $\{\beta_1, \beta_2, \beta_3, \beta_4\} = \{0.7, -1.2, 0.5, 0.5\}$ for group 1, $\{\beta_4, \beta_5, \beta_6\} = \{0.7, -0.7, 0.7\}$ for group 2, $\{\beta_9, \beta_{10}\} = \{0.7, -1.2\}$ for group 3 and $\{\beta_1, \beta_2, \beta_9, \beta_{10}\} = \{0.7, -0.7, 0.7, -0.7\}$ for group 4. Finally, response variables $\mathbf{y}_i$ for $i = 1, \dots, 12$ are drawn from a Bernoulli distribution with probability of success $p_i = \operatorname{logit}(\mathbf{X}_i^T \boldsymbol{\beta}_i)$.

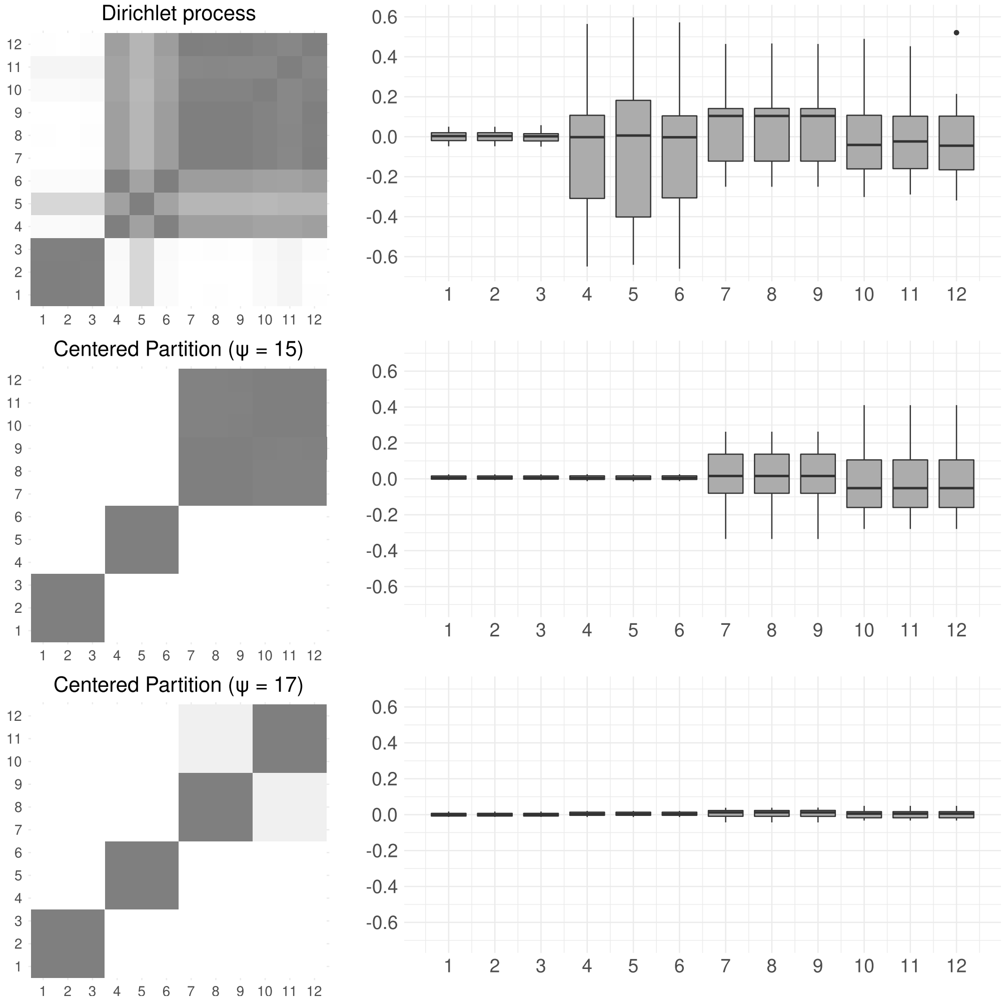


Figure 8: Results from grouped logistic regressions with DP($\alpha = 1$) prior and CP process prior with DP($\alpha = 1$) base EPPF for $\psi = \{15, 17\}$, centered on the true partition. Heatmaps on the left side show the posterior similarity matrix. On the right side, boxplots show the distribution of deviations from the maximum likelihood baseline coefficients and posterior mean estimates for each defect $i = 1, \dots, 12$.

We compare coefficients and partition estimates from a grouped logistic regression using a DP prior with $\alpha = 1$ and using a CP prior with DP base EPPF with $\alpha = 1$. In evaluating the CP prior performances, we consider both the true known partition and a wrong guess. Posterior estimates are obtained using the Gibbs sampler described in the Supplementary Material. We consider a multivariate normal distribution with zero mean vector and covariance matrix $\mathbf{Q} = \text{diag}_p(2)$ as base measure for the DP, while we assume the defect-specific intercepts $\alpha_i \sim N(0, 2)$ for $i = 1, \dots, 12$. We run the

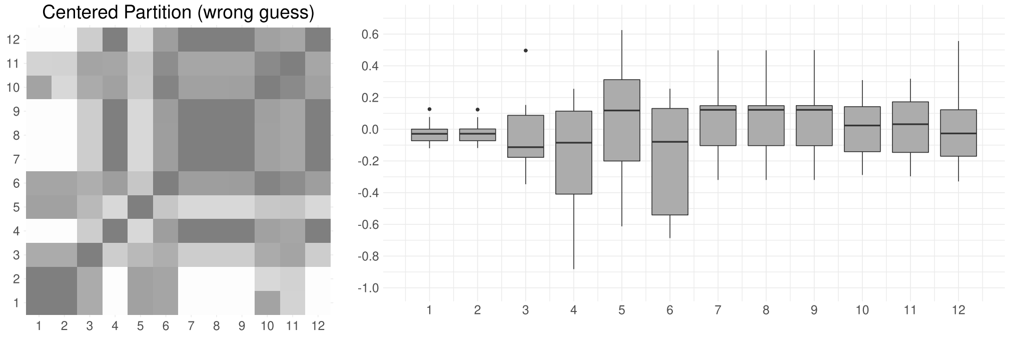


Figure 9: Results from grouped logistic regression using CP process prior with DP($\alpha = 1$) base EPPF for $\psi = 15$ centered on partition $\boldsymbol{c}_0' = \{1,5,9\}\{2,6,10\}\{3,7,11\}\{4,8,12\}$ which has distance 3.16 from the true one. Heatmaps on the left side show the posterior similarity matrix. On the right side, boxplots show the distribution of deviations from the maximum likelihood baseline coefficients and posterior mean estimates for each defect $i = 1,\dots,12$.

algorithm for 5,000 iterations discarding the first 1,000 as burn-in, with inspection of trace-plots suggesting convergence of the parameters.

In evaluating the resulting estimates under different settings, we take as baseline values for coefficients the maximum likelihood estimates obtained under the true grouping. Figure 8 shows the posterior similarity matrices obtained under the Dirichlet and Centered Partition processes, along with boxplots of the distribution of differences between the coefficients posterior mean estimates and their baseline values, for each of the 12 simulated defects. We first centered the CP prior on the true known grouping and, according to the considerations made in Section 4.2, we fixed the value of ψ to 15 for the CP process prior, founding the maximum a posteriori estimate of the partition almost recovering the true underlying grouping expect for merging together the third and fourth group. We also considered other values for ψ close to 15, and report the case for $\psi = 17$ in Figure 8, for which the true grouping is recovered, with resulting mean posterior estimates of the coefficients almost identical to the baseline. The Dirichlet process, although borrowing information across the defects, does not distinguish between all the groups but individuate only the first one, while the CP process recovers the true grouping, with better performances in estimating the coefficients.

Finally, we evaluate the CP prior performances when centered on a wrong guess $\boldsymbol{c}_0'$ of the base partition (Figure 9). In particular, we set $\boldsymbol{c}_0' = \{1,5,9\}\{2,6,10\}\{3,7,11\}\{4,8,12\}$. Despite having the same configuration of $\boldsymbol{c}_0$, it has distance from $\boldsymbol{c}_0$ of approximately 3.16, where the maximum possible distance is $\log_2(12) = 4.70$. Under such setting we estimate the partition $\hat{\boldsymbol{c}} = \{1,2,3,5\}\{4,6,7,8,9,10,11,12\}$ via maximum at posteriori, obtaining two clusters. Although we center the prior in $\boldsymbol{c}_0'$, the estimated partition results to be closer to the one induced by the DP (0.65) than $\boldsymbol{c}_0'$ (2.45), with also similar performances in the coefficient estimation, which may be interpreted as a suggestion that the chosen base partition is not supported by the data.

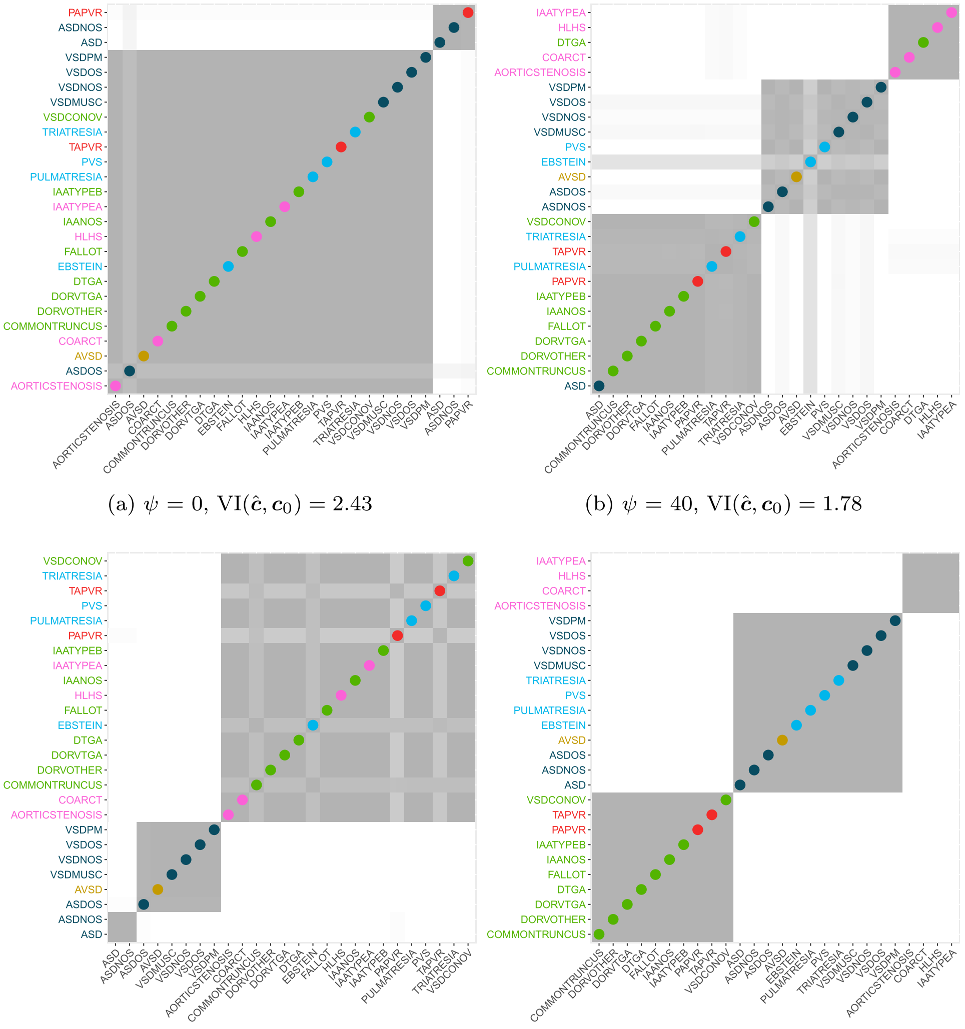
5.3 Application to NBDPS Data

We estimated the model in (5.2) on the NBDPS data, considering the controls as shared with the aim of grouping cases into informed groups on the basis of the available $\boldsymbol{c}_0$. In order to choose a value for the penalization parameter, we consider the prior calibration illustrated in Section 4, finding a value of $\psi = 40$ assigning a 90% probability to partitions within a distance around 0.8, where the maximum possible distance is equal to 4.70. In terms of moves on the Hasse diagram we are assigning 90% prior probability to partitions at most at 11 split/merge operations from $\boldsymbol{c}_0$, given that the minimum distance from $\boldsymbol{c}_0$ is $2/N \approx 0.07$. The R code is computationally intensive, running 2 days on a Linux cluster with 512 GB of RAM using a single core of a Intel Xeon (R) 2.10 GHz processor. Efficiency gains are expected by adapting our code through including precompiled C++ code or and/or adopting parallelization on a computing network. However, our current prior calibration algorithm is intrinsically very computational intensive in settings involving large numbers of objects to cluster. To assess sensitivity of the results, we performed the analysis under different values of $\psi \in \{0, 40, 80, 120, \infty\}$. In particular, for $\psi = 0$ the clustering behavior is governed by a Dirichlet process prior, while $\psi \to \infty$ corresponds to fixing the groups to $\boldsymbol{c}_0$.

In analyzing the data, we run the Gibbs sampler for 10,000 iterations and use a burn-in of 4,000, under the same prior settings as in Section 5.2. Figure 10 summarizes the posterior estimates of the allocation matrices under different values of ψ, with colored dots emphasizing differences with the base partition $\boldsymbol{c}_0$. Under the DP process ($\psi = 0$) the estimated partition differs substantially from the given prior clustering. Due to the immense space of the possible clusterings, this is likely reflective of limited information in the data, combined with the tendency of the DP to strongly favor certain types of partitions, typically characterized from few large clusters along many small ones. When increasing the value of the tuning parameter ψ the estimated clustering is closer to $\boldsymbol{c}_0$, with a tendency in favoring a total number of three clusters. In particular, for $\psi = 120$ one of the groups in $\boldsymbol{c}_0$ is recovered (left ventricular outflow), while the others are merged in two different groups. It is worth noticing that AVSD, which is placed in its own group under $\boldsymbol{c}_0$, is always grouped with other defects with a preference for ones in the septal group (blue color). Also two defects of this last class, ASD and ASDOS, are consistently lumped together across different values of ψ, and are in fact two closely related defects.

Details on the results for each of the estimated models are given in the Supplementary Material (Figures 3–7) and summarized here. Figure 11 shows a heatmap of the mean posterior log odds-ratios for increasing values of the penalization parameter ψ, with dots indicating significant values according to a 95% credibility interval. In general, the sign of the effects does not change for most of the exposure factors across the different clusterings. Figure 11 focuses on pharmaceutical use in the period from 1 month before the pregnancy and 3 months during, along with some exposures related to maternal behavior and health status.

We found consistent results for known risk factors for CHD in general, including diabetes (Correa et al., 2008) and obesity (Waller et al., 2007). The positive association between nausea and positive outcomes is likely due to the fact that nausea is indicative



(a) $\psi = 0$, $\mathrm{VI}(\hat{\boldsymbol{c}}, \boldsymbol{c}_0) = 2.43$

(b) $\psi = 40$, $\mathrm{VI}(\hat{\boldsymbol{c}}, \boldsymbol{c}_0) = 1.78$

(c) $\psi = 80$, $\mathrm{VI}(\hat{\boldsymbol{c}}, \boldsymbol{c}_0) = 1.65$

(d) $\psi = 120$, $\mathrm{VI}(\hat{\boldsymbol{c}}, \boldsymbol{c}_0) = 0.86$

Figure 10: Posterior allocation matrices obtained using the CP process with a DP ($\alpha = 1$) prior for different values of $\psi \in \{0, 40, 80, 120\}$. On the y-axis labels are colored according base grouping information $\boldsymbol{c}_0$, with dots on the diagonal highlighting differences between $\boldsymbol{c}_0$ and the estimated partition $\hat{\boldsymbol{c}}$.

of a healthy pregnancy, and is consistent with prior literature (Koren et al., 2014). The association between the use of SSRIs and pulmonary atresia was also reported in Reefhuis et al. (2015). It is worth noticing that estimates obtained under the DP prior are less consistent with prior work. In particular, there are apparent artifacts such as the protective effect of alcohol consumption related to defects in the bigger cluster, which is

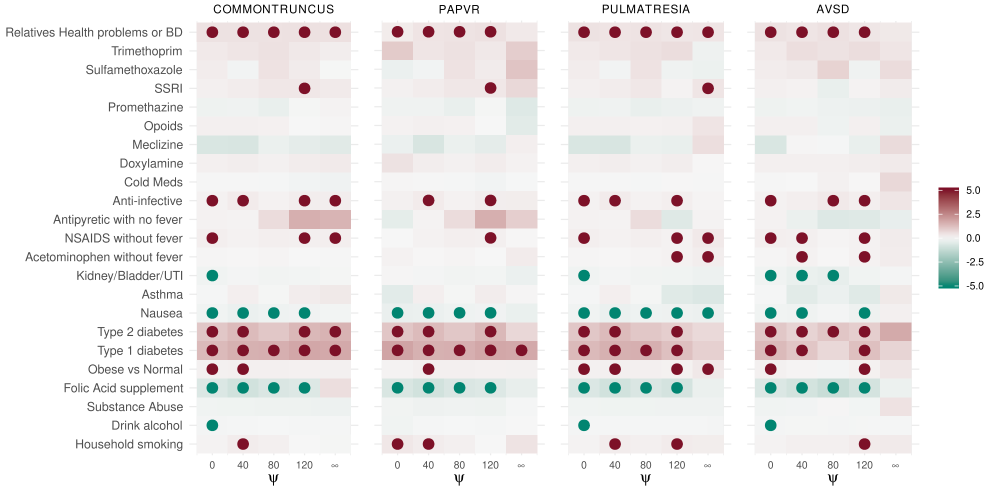


Figure 11: Comparison of significant odds ratio under $\psi \in \{0, 40, 80, 120, \infty\}$ for some exposure factors and 4 selected heart defects in 4 different groups under $\boldsymbol{c}_0$. Dots are in correspondence of significant mean posterior log-odds ratios (log-OR) at 95% with red encoding risk factors (log-OR > 0) and green protective factors (log-OR < 0).

mitigated from an informed borrowing across the defects. On the other side, estimates under separate models for AVSD or PAPVR, which corresponds to 0.02% and 0.01% of cases respectively, show how a separate analysis of cases with low prevalence misses even widely assessed risk factors, as for example diabetes.

Discussion

There is a very rich literature on priors for clustering, with almost all of the emphasis on exchangeable approaches, and a smaller literature focused on including dependence on known features (e.g. covariates, temporal or spatial structure). The main contribution of this article is to propose what is seemingly a first attempt at including prior information on an informed guess at the clustering structure. We were particularly motivated by a concrete application to a birth defects study in proposing our method, based on shrinking an initial clustering prior towards the prior guess.

Our approach is conceptually quite general and represents a first attempt to include this sort of prior information in clustering. However, we recognize that the proposed prior calibration does not allow a straightforward scaling when the number of objects is much larger than the $N = 26$ considered in the motivating birth defects application. This is due to a combinatorial explosion as N increases which leads to an inevitable deterioration of our prior calibration algorithm. For larger N, one can consider results from the prior calibration approach as providing a reasonable lower bound for ψ, with several higher values also considered in data analyses. An immediate direction of future research considers improving our prior calibration algorithm, by relying on more efficient sampling methods on discrete combinatorial spaces, with promising directions given in the recent works of Arratia and DeSalvo (2016) and DeSalvo (2017).

Although our proposed CP process may in principle accommodate hyperprior distributions for the Dirichlet and Pitman-Yor process parameters, a limitation is in that the prior calibration directly depends on such parameters, making the implementation difficult when hyperpriors are used. For example, if a prior is put on the hyperparameters of the baseline EPPF, then the calibration for ψ has to be performed at each MCMC step, conditionally on the value of the EPPF's hyperparameters, unless one integrates over the hyperparameters' distribution. We are considering the alternative of a prior distribution on ψ, although the corresponding posterior leads to an intractable normalizing constant. Possible options to address this issue may be to consider a direct approximation for the constant as in Vitelli et al. (2018), or to explore specialized MCMC algorithms for doubly intractable problems in which the likelihood involves an intractable normalizing constant (Murray et al., 2006; Møller et al., 2006; Rao et al., 2016).

There are many immediate interesting directions for future research. One thread pertains to developing better theoretical insight and analytical tractability into the new class of priors. For existing approaches, such as product partition models and Gibbs-type partitions, there is a substantial literature providing simple forms of prediction rules and other properties. It is an open question whether such properties can be modified to our new class. This may yield additional insight into the relative roles of the base prior, centering value and hyperparameters in controlling the behavior of the prior and its impact on the posterior. Another important thread relates to applications of the proposed framework beyond the setting in which we have an exact guess at the complete clustering structure. In many cases, we may have an informed guess or initial clustering in a subset of the objects under study, with the remaining objects (including future ones) completely unknown. Conceptually the proposed approach can be used directly in such cases, and also when one has different types of prior information on the clustering structure than simply which objects are clustered together.

Supplementary Material

Supplementary material for Centered Partition Processes: Informative Priors for Clustering (DOI: 10.1214/20-BA1197SUPP; .pdf).

References

Arratia, R. and DeSalvo, S. (2016). "Probabilistic divide-and-conquer: a new exact simulation method, with integer partitions as an example." *Combinatorics, Probability and Computing*, 25(3): 324–351. MR3482658. doi: https://doi.org/10.1017/S0963548315000358. 327

Barrientos, A. F., Jara, A., Quintana, F. A., et al. (2012). "On the support of MacEachern's dependent Dirichlet processes and extensions." *Bayesian Analysis*, 7(2): 277–310. MR2934952. doi: https://doi.org/10.1214/12-BA709. 302

Barry, D. and Hartigan, J. A. (1992). "Product partition models for change point problems." *The Annals of Statistics*, 260–279. MR1150343. doi: https://doi.org/10.1214/aos/1176348521. 303

Blei, D. M. and Frazier, P. I. (2011). "Distance dependent Chinese restaurant processes." *Journal of Machine Learning Research*, 12(Aug): 2461–2488. MR2834504. 303

Botto, L. D., Lin, A. E., Riehle-Colarusso, T., Malik, S., Correa, A., and Study, N. B. D. P. (2007). "Seeking causes: classifying and evaluating congenital hearth defects in etiologic studies." *Birth Defects Research Part A: Clinical and Molecular Teratology*, 79(10): 714–727. 319

Caron, F., Davy, M., Doucet, A., Duflos, E., and Vanheeghe, P. (2006). "Bayesian inference for dynamic models with Dirichlet process mixtures." In *International Conference on Information Fusion*. Florence, Italy. MR2439814. doi: https://doi.org/10.1109/TSP.2007.900167. 302

Casella, G., Moreno, E., Girón, F. J., et al. (2014). "Cluster analysis, model selection, and prior distributions on models." *Bayesian Analysis*, 9(3): 613–658. MR3256058. doi: https://doi.org/10.1214/14-BA869. 306

Correa, A., Gilboa, S. M., Besser, L. M., Botto, L. D., Moore, C. A., Hobbs, C. A., Cleves, M. A., Riehle-Colarusso, T. J., Waller, D. K., Reece, E. A., et al. (2008). "Diabetes mellitus and birth defects." *American Journal of Obstetrics and Gynecology*, 199(3): 237.e1–237.e9. 324

Dahl, D. B., Day, R., and Tsai, J. W. (2017). "Random partition distribution indexed by pairwise information." *Journal of the American Statistical Association*, 112(518): 721–732. MR3671765. doi: https://doi.org/10.1080/01621459.2016.1165103. 303

Davey, B. A. and Priestley, H. A. (2002). *Introduction to Lattices and Order*. Cambridge University Press. MR1902334. doi: https://doi.org/10.1017/CBO9780511809088. 303

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2): 212–229. 302

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). "An ANOVA model for dependent random measures." *Journal of the American Statistical Association*, 99(465): 205–215. MR2054299. doi: https://doi.org/10.1198/016214504000000205. 302

DeSalvo, S. (2017). "Improvements to exact Boltzmann sampling using probabilistic divide-and-conquer and the recursive method." *Pure Mathematics and Applications*, 26(1): 22–45. MR3674129. doi: https://doi.org/10.1515/puma-2015-0020. 327

Dobiński, G. (1877). "Summirung der Reihe $\sum \frac{n^m}{n!}$ für $m = 1, 2, 3, 4, 5, \ldots$" *Archiv der Mathematik und Physik*, 61: 333–336. 317

Dunson, D. B. and Park, J.-H. (2008). "Kernel stick-breaking processes." *Biometrika*, 95(2): 307–323. MR2521586. doi: https://doi.org/10.1093/biomet/asn012. 302

Fall, M. D. and Barat, É. (2014). "Gibbs sampling methods for Pitman-Yor mixture models." Working paper or preprint. URL https://hal.archives-ouvertes.fr/hal-00740770 313

Ferguson, T. S. (1973). "A Bayesian Analysis of Some Nonparametric Problems." *The Annals of Statistics*, 1(2): 209–230. MR0350949. 302

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). "Bayesian nonparametric spatial modeling with Dirichlet process mixing." *Journal of the American Statistical Association*, 100(471): 1021–1035. MR2201028. doi: https://doi.org/10.1198/016214504000002078. 302

Gnedin, A. and Pitman, J. (2006). "Exchangeable Gibbs partitions and Stirling triangles." *Journal of Mathematical Sciences*, 138(3): 5674–5685. MR2160320. doi: https://doi.org/10.1007/s10958-006-0335-z. 302

Griffin, J. E. and Steel, M. F. (2006). "Order-based dependent Dirichlet processes." *Journal of the American Statistical Association*, 101(473): 179–194. MR2268037. doi: https://doi.org/10.1198/016214505000000727. 302

Hartigan, J. (1990). "Partition models." *Communications in Statistics – Theory and Methods*, 19(8): 2745–2756. MR1088047. doi: https://doi.org/10.1080/03610929008830345. 303

Jensen, S. T. and Liu, J. S. (2008). "Bayesian clustering of transcription factor binding motifs." *Journal of the American Statistical Association*, 103(481): 188–200. MR2420226. doi: https://doi.org/10.1198/016214507000000365. 306

Knuth, D. E. (2005). *The Art of Computer Programming. Generating all combinations and partitions*. Addison-Wesley. MR0478692. 315

Koren, G., Madjunkova, S., and Maltepe, C. (2014). "The protective effects of nausea and vomiting of pregnancy against adverse fetal outcome. A systematic review." *Reproductive Toxicology*, 47: 77–80. 325

Lin, A. E., Herring, A. H., Amstutz, K. S., Westgate, M.-N., Lacro, R. V., Al-Jufan, M., Ryan, L., and Holmes, L. B. (1999). "Cardiovascular malformations: changes in prevalence and birth status, 1972–1990." *American Journal of Medical Genetics*, 84(2): 102–110. 319

MacEachern, S. N. (1999). "Dependent nonparametric processes." In *Proceedings of the Bayesian Section.*, 50–55. Alexandria, VA: American Statistical Association. 302

MacEachern, S. N. (2000). "Dependent nonparametric processes." Technical report, Department of Statistics, The Ohio State University. 302

MacLehose, R. F. and Dunson, D. B. (2010). "Bayesian semiparametric multiple shrinkage." *Biometrics*, 66(2): 455–462. MR2758825. doi: https://doi.org/10.1111/j.1541-0420.2009.01275.x. 319

Meilă, M. (2007). "Comparing clusterings – an information based distance." *Journal of Multivariate Analysis*, 98(5): 873–895. MR2325412. doi: https://doi.org/10.1016/j.jmva.2006.11.013. 303, 308

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). "An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants." *Biometrika*, 93(2): 451–458. MR2278096. doi: https://doi.org/10.1093/biomet/93.2.451. 327

Monjardet, B. (1981). "Metrics on partially ordered sets – A survey." *Discrete Mathematics*, 35(1): 173–184. Special Volume on Ordered Sets. MR0620670. doi: https://doi.org/10.1016/0012-365X(81)90206-5. 308

Müller, P., Quintana, F., and Rosner, G. L. (2011). "A product partition model with regression on covariates." *Journal of Computational and Graphical Statistics*, 20(1): 260–278. MR2816548. doi: https://doi.org/10.1198/jcgs.2011.09066. 303

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006). "MCMC for doubly-intractable distributions." In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, 359–366. AUAI Press. 327

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9: 249–265. MR1823804. doi: https://doi.org/10.2307/1390653. 313, 314

Nijenhuis, A. and Wilf, H. S. (2014). *Combinatorial Algorithms: for Computers and Calculators*. Elsevier. MR0510047. 315

Paganin, S., Herring, A. H., Olshan, A. F., Dunson, D. B., and The National Birth Defects Prevention Study (2020). "Centered Partition Processes: Informative Priors for Clustering – Supplementary Material." *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1197SUPP. 303

Park, J.-H. and Dunson, D. B. (2010). "Bayesian generalize product partition models." *Statistica Sinica*, 20: 1203–1226. MR2730180. 303

Petrone, S., Guindani, M., and Gelfand, A. E. (2009). "Hybrid Dirichlet mixture models for functional data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(4): 755–782. MR2750094. doi: https://doi.org/10.1111/j.1467-9868.2009.00708.x. 302

Pitman, J. (1995). "Exchangeable and partially exchangeable random partitions." *Probability Theory and Related Fields*, 102(2): 145–158. MR1337249. doi: https://doi.org/10.1007/BF01213386. 302

Pitman, J. (1997). "Some probabilistic aspects of set partitions." *The American Mathematical Monthly*, 104(3): 201–209. MR1436042. doi: https://doi.org/10.2307/2974785. 317

Pitman, J. and Yor, M. (1997). “The two-Parameter Poisson-Dirichlet distribution derived from a stable subordinator.” *The Annals of Probability*, 25(2): 855–900. MR1434129. doi: https://doi.org/10.1214/aop/1024404422. 302

Polson, N. G., Scott, J. G., and Windle, J. (2013). “Bayesian inference for logistic models using Pólya-Gamma latent variables.” *Journal of the American Statistical Association*, 108(504): 1339–1349. MR3174712. doi: https://doi.org/10.1080/01621459.2013.829001. 321

Rao, V., Lin, L., and Dunson, D. B. (2016). “Data augmentation for models based on rejection sampling.” *Biometrika*, 103(2): 319–335. MR3509889. doi: https://doi.org/10.1093/biomet/asw005. 327

Rasmussen, S. A., Olney, R. S., Holmes, L. B., Lin, A. E., Keppler-Noreuil, K. M., and Moore, C. A. (2003). “Guidelines for case classification for the National Birth Defects Prevention Study.” *Birth Defects Research Part A: Clinical and Molecular Teratology*, 67(3): 193–201. 318

Reefhuis, J., Devine, O., Friedman, J. M., Louik, C., and Honein, M. A. (2015). “Specific SSRIs and birth defects: Bayesian analysis to interpret new data in the context of previous reports.” *British Medical Journal*, 351. 325

Rodriguez, A. and Dunson, D. B. (2011). “Nonparametric Bayesian models through probit stick-breaking processes.” *Bayesian Analysis*, 6(1). MR2781811. doi: https://doi.org/10.1214/11-BA605. 302

Rossi, G. (2015). “Weighted paths between partitions.” *arXiv preprint*. URL https://arxiv.org/abs/1509.01852 308

Scarpa, B. and Dunson, D. B. (2009). “Bayesian Hierarchical Functional Data Analysis Via Contaminated Informative Priors.” *Biometrics*, 65(3): 772–780. MR2649850. doi: https://doi.org/10.1111/j.1541-0420.2008.01163.x. 302

Sethuraman, J. (1994). “A constructive definition of Dirichlet priors.” *Statistica Sinica*, 4(2): 639–650. MR1309433. 302

Smith, A. N. and Allenby, G. M. (2019). “Demand Models With Random Partitions.” *Journal of the American Statistical Association*. doi: https://doi.org/10.1080/01621459.2019.1604360. 303

Stam, A. (1983). “Generation of a random partition of a finite set by an urn model.” *Journal of Combinatorial Theory, Series A*, 35(2): 231–240. MR0712107. doi: https://doi.org/10.1016/0097-3165(83)90009-2. 316, 317

Stanley, R. P. (1997). *Enumerative combinatorics. Vol. 1*. Cambridge University Press. MR1442260. doi: https://doi.org/10.1017/CBO9780511805967. 303

Vinh, N. X., Epps, J., and Bailey, J. (2010). “Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance.” *Journal of Machine Learning Research*, 11(Oct): 2837–2854. MR2738784. 308

Vitelli, V., Øystein Sørensen, Crispino, M., Frigessi, A., and Arjas, E. (2018). “Probabilistic preference learning with the Mallows rank model." *Journal of Machine Learning Research*, 18(158): 1–49. MR3813807. 327

Wade, S. and Ghahramani, Z. (2018). "Bayesian cluster analysis: point estimation and credible balls (with Discussion)." *Bayesian Analysis*, 13(2): 559–626. MR3807860. doi: https://doi.org/10.1214/17-BA1073. 303, 308

Waller, D. K., Shaw, G. M., Rasmussen, S. A., Hobbs, C. A., Canfield, M. A., Siega-Riz, A.-M., Gallaway, M. S., and Correa, A. (2007). "Prepregnancy obesity as a risk factor for structural birth defects." *Archives of Pediatrics & Adolescent Medicine*, 161(8): 745–750. 324

Wilson, R. and Watkins, J. J. (2013). *Combinatorics: Ancient & Modern*. OUP Oxford. MR3204727. doi: https://doi.org/10.1093/acprof:oso/9780199656592.001.0001. 304

Yoon, P. W., Rasmussen, S. A., Lynberg, M. C., Moore, C. A., Anderka, M., Carmichael, S. L., Costa, P., Druschel, C., Hobbs, C. A., Romitti, P. A., Langlois, P. H., and Edmonds, L. D. (2001). "The National Birth Defects Prevention Study." *Public Health Reports*, 116: 32–40. 318

Acknowledgments

This work was supported through cooperative agreements under R01ES027498, PA 96043, PA 02081 and FOA DD09-001 from the Centers for Disease Control and Prevention (CDC) to the North Carolina Center for Birth Defects Research and Prevention at the University of North Carolina at Chapel Hill, and to other Centers for Birth Defects Research and Prevention participating in the National Birth Defects Prevention Study and/or the Birth Defects Study to Evaluate Pregnancy Exposures.

Invited Discussion

David B. Dahl[*,§], Richard L. Warr[†], and Thomas P. Jensen[‡]

We enthusiastically applaud Paganin et al. (2021) for a stimulating article. The emergent idea of developing a partition distribution that utilizes an *a priori* estimate of the partition itself is novel and intriguing. It is very natural in the Bayesian literature to use an *a priori* estimate of a parameter when placing a prior distribution on an unknown parameter. Yet, Smith and Allenby (2020) and Paganin et al. (2021) have recognized the need for such prior distributions in the context of random partition models, which presents unique challenges because of the vast size of the discrete space of partitions. These authors add to the utility and applicability of random partition models and they have laid a foundation for fruitful future work in this area.

The Centered Partition (CP) process proposed by Paganin et al. (2021) is a distribution over partitions that is composed of four primary ingredients: 1) a baseline exchangeable partition probability function (EPPF), 2) a centering partition $\boldsymbol{c}_0$, 3) a distance function between partitions $d(\boldsymbol{c}, \boldsymbol{c}_0)$ to measure departures from $\boldsymbol{c}_0$, and 4) a penalization parameter ψ which controls how much influence is given to $\boldsymbol{c}_0$. The probability mass function is provided in Equation (3.1) of their article. Their formulation is clever in that, when $\psi = 0$, the CP process reduces to the baseline EPPF. Although the default may be the Dirichlet process (DP) EPPF and the variation of information (VI) distance (Meilă, 2007; Wade and Ghahramani, 2018), Paganin et al. (2021) have developed a very general framework which allows for any EPPF or distance function. The resulting CP process is a non-exchangeable random partition distribution influenced by $\boldsymbol{c}_0$. The authors provide a compelling application that harnesses the power of the CP process.

In this discussion we raise a few ideas that have captured our attention as we have studied the CP process. First, we note that $\boldsymbol{c}_0$ is *not* the center of the CP process according to a conventional Bayesian definition and, instead, the CP process is a random partition distribution shrunk toward $\boldsymbol{c}_0$. Second, we suggest an alternative prior calibration algorithm for the penalization parameter ψ that scales beyond $N = 26$ which seems to be the limit in Paganin et al. (2021). This results in figures that reinforce the first point that $\boldsymbol{c}_0$ is not the center. Next, rather than fixing hyperparameters, we advocate for future research to allow priors to be placed on ψ and on hyperparameters of the baseline EPPF. Finally, we observe that a property of the VI distance may lead to problems using a Gibbs sampling scheme and also note the importance of congruity between $\boldsymbol{c}_0$ and the baseline EPPF. In sum, we are enthusiastic about the CP process and the new avenues for research which it opens.

[*] Brigham Young University, 2152 WVB, Provo, UT 84602, dahl@stat.byu.edu

[†] Brigham Young University, 2152 WVB, Provo, UT 84602

[‡] Brigham Young University, 2152 WVB, Provo, UT 84602

[§] Corresponding author.

1 "Center" Is a Misnomer

We believe that the term "center" in the proposed CP process is a misnomer that may lead to a mistaken understanding of its nature. The paper describes the CP process as a partition distribution "centered on $\boldsymbol{c}_0$" and one might therefore expect that $\boldsymbol{c}_0$ is, in some sense, the middle partition or the partition around which the CP process is concentrated. Of course, as the penalization parameter ψ goes to infinity, all probability mass collapses to $\boldsymbol{c}_0$ and there is no disagreement that $\boldsymbol{c}_0$ is the middle partition. But what is the middle partition of the CP process for finite ψ? The middle of a univariate distribution might be described as the mean, i.e., the value that minimizes the expected squared error loss. Likewise, the Bayesian analog of the Fréchet mean for a partition distribution is the partition that minimizes the expected loss for some partition metric. Here, let $\hat{\boldsymbol{c}}$ denote the partition that minimizes the expected value of the variation of information.

Consider, for example, the CP process with Pitman-Yor process base EPPF with discount $\sigma = 0.75$ and concentration parameter $\alpha = -0.691$ and $\boldsymbol{c}_0 = \{\{1,2\},\{3,4\},\{5\}\}$. (This is the distribution displayed in Figure 5(b) of Paganin et al. (2021).) For this CP process, $\hat{\boldsymbol{c}} = \boldsymbol{c}_0$ when $\psi \geq 2.74$, $\hat{\boldsymbol{c}} = \{\{1,2,3,4\},\{5\}\}$ when $2.56 \leq \psi \leq 2.73$, and $\hat{\boldsymbol{c}} = \{\{1,2,3,4,5\}\}$ when $\psi \leq 2.55$. The point is that, for $\psi \leq 2.73$, the "center" partition $\boldsymbol{c}_0$ is in fact *not* the middle partition, at least not according to a conventional Bayesian definition. As such, instead of thinking of $\boldsymbol{c}_0$ as a "center" partition, we suggest thinking of the CP process as a baseline partition distribution that is shrunk towards $\boldsymbol{c}_0$. That is, for nontrivial values of penalization parameter ψ, the CP process is a compromise between the baseline EPPF and $\boldsymbol{c}_0$.

2 A Scalable Prior Calibration Algorithm

Paganin et al. (2021) note that, as the number of observations N increases, larger values of the penalization parameter ψ are needed such that there is non-negligible prior probability around the center partition $\boldsymbol{c}_0$. Rather than picking ψ arbitrarily, they suggest finding ψ such that a random partition is within some chosen distance δ^* from $\boldsymbol{c}_0$ with a desired probability q. We think this is a practical and interpretable objective. To find ψ yielding the objective, Section 4 of Paganin et al. (2021) provides a creative algorithm involving a deterministic local search and uniform sampling of partitions from the partition space. As noticed in their Discussion section, however, this algorithm does not scale much beyond $N = 26$ used in their application.

Here we suggest a simple alternative approach to obtain the desired objective that easily scales in N. Simply tune ψ to obtain the objective by Monte Carlo estimation based on sampling from the CP process prior itself (rather than sampling partitions from the uniform distribution). Given samples $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_T$ from the CP process for some ψ, the Monte Carlo estimate $\hat{q}$ of the desired probability q for threshold δ^* is:

$$\hat{q} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{I}\left\{ d(\boldsymbol{c}_t, \boldsymbol{c}_0) \leq \delta^* \right\}.$$

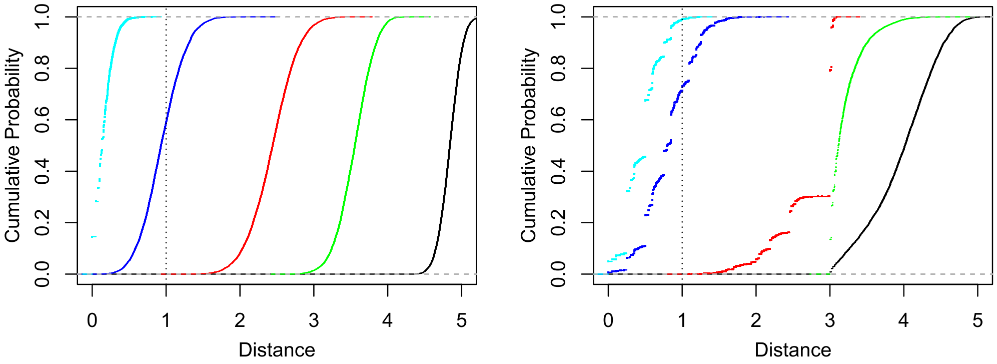


Figure 1: Estimate of the cumulative prior probabilities assigned to different distances from $\boldsymbol{c}_0$ for $N = 96$, where $\boldsymbol{c}_0$ consists of 8 clusters of 12 items each. The left plot shows the CP process using the uniform EPPF and $\psi \in \{0, 40, 55, 70, 90\}$ (from black to cyan). The right plot shows the DP($\alpha = 1$) EPPF and $\psi \in \{0, 5, 60, 70, 75\}$. The results are obtained using the scalable algorithm in our Section 2.

Of course, the definition of the CP process does not lend itself to straight Monte Carlo sampling, but the MCMC scheme for posterior simulation — detailed in Section 3.4 of Paganin et al. (2021) — is easily adapted to prior simulation by simply dropping the terms involving data y_i in their equation (3.5). We replicated their Figure 7 for $N = 12$ observations (not shown here). Now, consider $N = 96$ and $\boldsymbol{c}_0$ having 8 clusters of 12 items each. Our Figure 1 shows the distribution of distances for the CP process with $\psi \in \{0, 40, 55, 70, 90\}$ for the uniform EPPF and $\psi \in \{0, 5, 60, 70, 75\}$ for the DP($\alpha = 1$) EPPF. One might use, for example, $\psi = 70$ for both EPPFs to obtain a CP process prior with about 60-70% probability that a random partition is within $\delta^* = 1$ of $\boldsymbol{c}_0$. Refinements to our approach could be implemented, such as using samples for various values of ψ through importance sampling, starting with small T and increase it for more promising values of ψ, and sampling in parallel for a variety of ψ values.

We believe that the point that $\boldsymbol{c}_0$ is not the center, which we discussed in the previous section, is also made by inspecting Figure 7 of Paganin et al. (2021) and our Figure 1. Note that, for very large ψ, there is a large point mass at $\boldsymbol{c}_0$. But, for other values of ψ, there is virtually no probability for partitions close to $\boldsymbol{c}_0$, as evidenced by the slope of zero coming away from the origin. If $\boldsymbol{c}_0$ were indeed the center of the partition distribution, one would expect a non-zero slope. This is indeed the behavior of random samples from a Gaussian distribution centered at 0 with standard deviation $\exp\{-\psi\}$ and using Euclidean distance, as shown in Figure 2.

3 Posterior Inference on Hyperparameters

The lack of a tractable normalizing constant in the p.m.f. of the CP process — see (3.1) in Paganin et al. (2021) — is a major concern which motivated the development

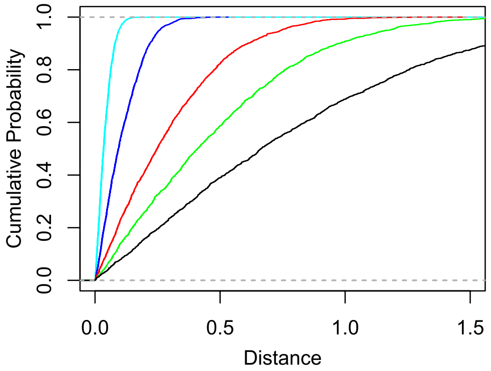


Figure 2: Cumulative distribution of distances from 0 for Gaussian distributions centered at 0 using Euclidean distance and standard deviation $\exp\{-\psi\}$ for $\psi \in \{0, 0.5, 1, 2, 3\}$ (black to cyan). In contrast to Figure 1, notice here the nonzero slope from the origin, suggesting that appreciable values are close to the center 0.

of the prior calibration algorithm in their Section 4. As suggested in our Section 2, the scalability in N of the calibration algorithm can be overcome and a suitable fixed value for the penalization parameter ψ can be obtained. Of course, the choice of q and δ^* in the prior calibration algorithm is somewhat arbitrary and sensitivity of the analysis among reasonable values of q and δ^* would need to be explored. More serious in our view, however, is that fixing ψ precludes the possibility of posterior inference on it. Although Bayesians often eventually fix hyperparameters that are far from the data and of little interest, the penalization parameter ψ is a key component of the CP process. It would seem that one would naturally be curious regarding how the data changes its value from a prior belief regarding it.

The last section of Paganin et al. (2021) discusses issues stemming from the lack of a tractable normalizing constant. They suggest that prior distributions could theoretically be placed on hyperparameters of the baseline EPPF (e.g., the concentration parameter α and the discount parameter σ from the Pitman-Yor process) but acknowledge that this would necessitate recalibration of ψ for every update of the other hyperparameters. We believe the issue is even more fundamental in that, even for a fixed value of ψ, the unknown normalizing constant potentially depends on these hyperparameters and, therefore, posterior inference on them would be challenging. We appreciate the ideas of estimating the normalizing constant or using MCMC algorithms for doubly intractable problems, but note that these approaches are nontrivial. Lacking that, in addition to fixing the penalization parameter ψ, these other hyperparameters must also be fixed (through an arbitrary choice or through some sort of calibration) and one cannot study the effect of the data on prior beliefs regarding these hyperparameters.

4 Local Modes Induced by Variation of Information

The one-at-a-time Gibbs sampling scheme described in Section 3.4 of Paganin et al. (2021) is a familiar and practical MCMC scheme for random partition models. When using the variation of information distance in the CP process, however, we came across an oddity. The issue is that the variation of information can induce local modes which may be difficult to escape using one-at-a-time updates. Consider, for example, the CP process with uniform base EPPF, center partition $\boldsymbol{c}_0 = \{\{1,2,3\},\{4,5,6\}\}$, and current state $\boldsymbol{c} = \{\{1,2,3,4,5,6\}\}$. With large penalization parameter ψ, the Markov chain should spend most of its time at $\boldsymbol{c}_0$. Moving towards $\boldsymbol{c}_0$ from $\boldsymbol{c}$, however, is unlikely since one-at-a-time updating requires going through a configuration with exactly one singleton cluster, e.g., $\boldsymbol{c}^* = \{\{1,2,3,4,5\},\{6\}\}$. To see the problem, note that $\mathrm{VI}(\boldsymbol{c}^*, \boldsymbol{c}_0) \approx 1.27$ is *greater* than $\mathrm{VI}(\boldsymbol{c}, \boldsymbol{c}_0) = 1$ (a difference that is accentuated by a large ψ) making it unlikely that a Gibbs update would move from $\boldsymbol{c}$ to $\boldsymbol{c}^*$, especially as ψ increases. It is hard to say how frequently this issue would occur in practical data analysis, but we note that Binder (1978) loss with equal costs may not suffer from this local mode issue, e.g., $\mathrm{Binder}(\boldsymbol{c}^*, \boldsymbol{c}_0) < \mathrm{Binder}(\boldsymbol{c}, \boldsymbol{c}_0)$. As such, we feel it may also be worth considering using Binder loss for the CP process.

5 Compatibility with the Baseline Distribution

Based on our explorations of the CP process, we suggest the users of the CP process should be careful that $\boldsymbol{c}_0$ and the baseline EPPF are congruous. We believe that conflicting choices for these important parameters leads to non-smooth behavior. Consider, for example, our Figure 1 which is smooth for the left plot and erratic for the right plot. Recall that $\boldsymbol{c}_0$ has eight clusters of equal size, which seems to be compatible with the uniform EPPF on the left, yet contradictory to the "rich get richer" property of the $\mathrm{DP}(\alpha = 1)$ based EPPF on the right. Further, for the right plot, it seems odd that the cumulative probabilities of $\psi = 60$ are closer to those of $\psi = 5$ than $\psi = 70$.

If an incongruity between $\boldsymbol{c}_0$ and the baseline EPPF exists, it may become difficult for a Markov chain to mix well in the partition space. In fact, in order to feel confident in Figure 1, we had to run 200,000 Gibbs scans — half of which were discarded and thinned by 1-in-100 — for Markov chains started from $\boldsymbol{c} = (1, \ldots, 1)$, $\boldsymbol{c} = (1, \ldots, N)$, and $\boldsymbol{c} = \boldsymbol{c}_0$. The difficulty in mixing well seems to be a consequence of a "tug of war" between $\boldsymbol{c}_0$ and the baseline EPPF, which we interpret as further evidence that $\boldsymbol{c}_0$ is, in fact, not the center of the CP process.

References

Binder, D. A. (1978). "Bayesian cluster analysis." *Biometrika*, 65(1): 31–38. MR0501592. doi: https://doi.org/10.1093/biomet/65.1.31. 337

Meilă, M. (2007). "Comparing clusterings—an information based distance." *Journal of Multivariate Analysis*, 98(5): 873–895. MR2325412. doi: https://doi.org/10.1016/j.jmva.2006.11.013. 333

Paganin, S., Herring, A. H., Olshan, A. F., and Dunson, D. B. (2021). "Centered partition processes: Informative priors for clustering." *Bayesian Analysis*. Advance publication. doi: https://doi.org/https://doi.org/10.1214/20-BA1197. 333, 334, 335, 336, 337

Smith, A. N. and Allenby, G. M. (2020). "Demand models with random partitions." *Journal of the American Statistical Association*, 115(529): 47–65. MR4078444. doi: https://doi.org/10.1080/01621459.2019.1604360. 333

Wade, S. and Ghahramani, Z. (2018). "Bayesian cluster analysis: Point estimation and credible balls (with discussion)." *Bayesian Analysis*, 13(2): 559–626. MR3807860. doi: https://doi.org/10.1214/17-BA1073. 333

Invited Discussion

Adam N. Smith[*]

The analysis of discrete random structures underlying Bayesian nonparametric models continues to be a growing area of research. Of particular interest is the way in which nonparametric priors can be used for model-based clustering. This paper makes an important and practically useful contribution to this literature by constructing a prior that can be "centered" around a pre-specified clustering. The elicitation of prior information is indeed at the core of the Bayesian paradigm and is often facilitated through the use of priors belonging to a location-scale family: a location parameter encodes what the *belief is* while a scale parameter encodes the *strength* of that belief. Constructing an analogous prior for a partition parameter is challenging given the complex topology on which partitions are defined. Consequently, researchers are often left resorting to default prior settings and lack the ability to bring substantive knowledge (or lack thereof) to bear on the analysis. This paper fills this gap and, in doing so, adds a nice tool to the Bayesian clustering toolkit.

The authors propose the centered partition (CP) process for a clustering parameter $\boldsymbol{c} \in \Pi_N$. The CP process consists of four components: (1) a baseline exchangeable partition probability function (EPPF) $p_0(\boldsymbol{c})$; (2) a pre-specified clustering $\boldsymbol{c}_0$; (3) a function $d(\boldsymbol{c}, \boldsymbol{c}_0)$ measuring the distance between $\boldsymbol{c}$ and $\boldsymbol{c}_0$; and (4) a penalty parameter $\psi \geq 0$. The CP process is written as: $p(\boldsymbol{c}|\boldsymbol{c}_0, \psi) \propto p_0(\boldsymbol{c})e^{-\psi d(\boldsymbol{c},\boldsymbol{c}_0)}$, where the limiting cases of $\psi = 0$ and $\psi = \infty$ reveal its location-scale flavor. The idea of a adding structure through a penalty that multiplies a baseline EPPF is quite parsimonious and is a point of departure from existing approaches that modify the EPPF directly (Park and Dunson, 2010; Müller and Quintana, 2011; Blei and Frazier, 2011; Dahl et al., 2017; Smith and Allenby, 2020).

In this discussion, I plan to first review the roles of the various model components and highlight the practical challenges of prior elicitation in the context of clustering. I will then comment on posterior computation and conclude with a few open questions and thoughts on fruitful areas for future work.

1 The Centering Partition and Domain Knowledge

Throughout the paper the authors assume that $\boldsymbol{c}_0$ is a single fixed clustering which represents the "location" component of the researcher's beliefs. The CP prior will assign higher probability to $\boldsymbol{c}_0$ and neighboring clusters as the penalty parameter increases. But given the complex nature of the space of partitions Π_N, do strong beliefs about $\boldsymbol{c}_0$ necessarily translate into strong beliefs about clusters within some small neighborhood of $\boldsymbol{c}_0$? For example, if I could enumerate all possible clusterings and then rank order them based on my prior beliefs, will the first two or three clusters always be "close" as

[*]UCL School of Management, University College London, a.smith@ucl.ac.uk

defined by an information-based distance metric? Or is it possible that clusters "close" to $\boldsymbol{c}_0$ (based on the distance metric) are actually less sensible a priori?

Consider the paper's empirical application to modeling congenital heart defects with a centered clustering $\boldsymbol{c}_0$ defined based on prior research (Botto et al., 2007). Specifically, the $N = 26$ individual heart defects are partitioned into $K = 6$ groups, where defects within a group are similar on the basis of various epidemiologic and anatomic factors. A CP prior with a large penalty term ψ will then place high probability on $\boldsymbol{c}_0$ and clusters close to $\boldsymbol{c}_0$. Now consider a new clustering $\boldsymbol{c}_0'$ which is equal to $\boldsymbol{c}_0$ but moves the "atrial septal defect" away from its original cluster ("Septal") and into another cluster, say "Conotruncal". Here $\boldsymbol{c}_0$ and $\boldsymbol{c}_0'$ have the same number of groups and differ only by one element so $d(\boldsymbol{c}_0, \boldsymbol{c}_0')$ will be small. But is it sensible, based on relevant epidemiologic or anatomic factors, that "atrial septal defect" is grouped assigned into "Conotruncal" while all other "Septal" defects are not? Perhaps a domain expert would place higher prior probability on clusterings that merge the "Conotruncal" and "Septal" groups than clusterings that merge individual defects across groups.

Another motivating example stems from the application of nested logit demand models (McFadden, 1978; Train, 2002) in fields like quantitative marketing and microeconometrics. Here, the goal is to model consumer choice among discrete alternatives such as products. The nested logit model is attractive because of its ability to accommodate correlated error structures across products, but it requires the researcher to first partition the set of products into groups (nests) such that products within a group are more similar than products across groups. One challenge is that products can have many attributes (e.g., brand name, size, flavor, package type) and so it is often unclear how to define this partitioning of goods a priori. In practice, researchers often resort to testing a few different grouping structures on the data. For example, Allenby (1989) compare clusters based on price tiers vs. size and Draganska and Jain (2006) compare clusters based on brand vs. flavor. In each of these examples, the researchers effectively place prior mass on only two points in the space of partitions. Moreover, while these clusterings are well-motivated by managerial/economic considerations, they are likely far away based on any information-based distance metric.

The examples described above demonstrate that domain knowledge may lead to prior beliefs that are spread across fairly disparate regions of Π_N, and so an application of the "vanilla" CP prior may be inconsistent with such beliefs. How can location-scale-type priors like the CP process better account for prior uncertainty around $\boldsymbol{c}_0$?

- *Point mass mixture priors.* One approach is to enlarge the space of possible centering partitions and directly model prior uncertainty in $\boldsymbol{c}_0$. For example, consider the following two-stage prior:

$$\boldsymbol{c}|\boldsymbol{c}_0, \psi \sim \mathrm{CP}(\boldsymbol{c}_0, \psi, p_0(\boldsymbol{c}))$$

$$\boldsymbol{c}_0 \sim \sum_{\ell=1}^{L} w_\ell \delta_{\bar{\boldsymbol{c}}_\ell}$$

  where $\bar{\boldsymbol{c}}_1, \dots, \bar{\boldsymbol{c}}_L$ are pre-specified partitions, $\delta_{\bar{\boldsymbol{c}}_\ell}$ is a point mass at $\bar{\boldsymbol{c}}_\ell$, and $w_1, \dots, w_L$ are weights satisfying $\sum_{\ell=1}^{L} w_\ell = 1$. This point mass mixture prior on $\boldsymbol{c}_0$ can induce a marginal prior $p(\boldsymbol{c})$ that exhibits a more global dispersion of probability mass across Π_N, while also retaining the ability to deviate locally around each fixed location $\bar{\boldsymbol{c}}_\ell$. This approach could also allow the researcher to incorporate information from a more general classification hierarchy, which can be common in clustering problems (including the three-level hierarchy presented in Botto et al., 2007). For example, one could define the set of partitions $\bar{\boldsymbol{c}}_1, \ldots, \bar{\boldsymbol{c}}_L$ to include an initial guess as well as variants that are derived by merging groups according to the next level in the hierarchy.

- *Pairwise information.* The distance function $d(\boldsymbol{c}, \boldsymbol{c}_0)$ inside the CP process is implicitly defined over N-vectors of group membership indices. One drawback with this measure of distance is that the domain knowledge driving prior co-clustering probabilities is reduced to whether the two items belong to the same group within $\boldsymbol{c}_0$. Another approach is to define distances over an $N \times N$ pairwise information matrix (Blei and Frazier, 2011; Dahl et al., 2017). The benefit is that prior co-clustering probabilities can depend on a more flexible measure of pairwise distance, including other item-level characteristics (e.g., the various epidemiologic and anatomic factors of heart defects). To see where this flexibility comes from, note that the information contained within a centering partition $\boldsymbol{c}_0$ can also be represented as a block-diagonal $N \times N$ matrix (after re-ordering items) with 1's within each block and 0's on the off-diagonals. A pairwise information approach will allow for richer sources of variation to enter the "within-group" and "across-group" elements of this matrix and thus more control over the spread of prior probability mass over Π_N.

2 The Penalization Parameter

The dispersion of probability mass under the CP process is largely governed by the penalization parameter ψ. All else equal, as $\psi \to \infty$, mass will concentrate on $\boldsymbol{c}_0$ and its close neighbors while as $\psi \to 0$, mass will be dispersed according the baseline EPPF. Given that ψ captures the "strength" of the prior belief and that the dimension of Π_N grows exponentially in the number of items N, care must be taken when choosing ψ across analyses with varying N. For example, choosing $\psi = 1$ will imply a very different strength of belief about $\boldsymbol{c}_0$ when $N = 5$ ($\mathcal{B}_5 = 52$) than it does when $N = 50$ ($\mathcal{B}_{50} > 1.8 \times 10^{47}$). The same issue is acknowledged by Smith and Allenby (2020) in the context of tuning random-walk Metropolis-Hastings proposals with their location-scale partition (LSP) distribution.

I appreciate that the authors address this point and propose a method that does not elicit ψ directly, but is instead based on choosing a probability q and a distance δ^* that together induce a penalty ψ. Their novel idea is to choose the pair (q, δ^*) such that the CP process places probability of at least q on partitions within distance δ^* from $\boldsymbol{c}_0$. The authors use the variation of information (VI) distance metric throughout, which has the key property of being N-invariant (Meilă, 2007). Therefore, eliciting a prior through q and δ^* is in principle more straightforward because the (q, δ^*) pair is invariant to the size of the clustering problem.

However, given the heavy computation involved with calibrating the CP prior (i.e., tracing out the values of ψ corresponding to different combinations of q and δ^*), I wonder what the trade-off is between investing time to get the prior "exactly right" vs. letting ψ be an estimated model parameter? Are there significant computational challenges associated with adding a step to the sampler which, say, cycles through a grid of possible ψ values? Within the context of the paper's empirical application, integrating over the uncertainty in ψ should lead to improved estimates of the regression coefficients and could even help guard against misspecification of $\boldsymbol{c}_0$.

3 Computation

The posterior sampling strategy for the CP process borrows from the usual suite of sampling methods for Dirichlet process mixture (DPM) models – specifically, Algorithm 2 of Neal (2000) where item-group indicators are iteratively sampled from their respective full conditional distributions $p(c_i = k|\boldsymbol{c}^{-i}, \text{else})$. One potential concern is that these "local moves" do not allow the sampler to sufficiently traverse the posterior and can lead to underestimated posterior uncertainty in estimates of $\boldsymbol{c}$. There is no real discussion of the sampler's mixing properties in the paper, and I wonder whether the imposition of strong prior information on $\boldsymbol{c}$ exacerbates this issue.

It is certainly true that more informative priors will lead to more concentrated posteriors. However, the real challenge is that the regions of high posterior probability may still be separated by sizable peaks and valleys due to the complex topology of Π_N, creating problems for samplers relying on incremental moves. As it becomes feasible to incorporate prior information on clustering problems, I believe it is also useful to ensure that this information does not mechanically lead to samplers getting stuck in small neighborhoods of high probability mass induced by the prior. To this end, more radical split-merge Metropolis-Hastings proposal mechanisms can be attractive (Dahl, 2003; Jain and Neal, 2004, 2007). Another option is to rely on the CP process itself to construct random-walk-style Metropolis-Hastings proposals (akin to Smith and Allenby, 2020), which would also have applicability beyond the class of DPM models.

4 Closing Thoughts

The CP process adds to a growing set of partitioning models designed to help researchers incorporate prior information in clustering problems (Park and Dunson, 2010; Müller and Quintana, 2011; Blei and Frazier, 2011; Dahl et al., 2017; Smith and Allenby, 2020). There are many nice features of the CP process – in particular, the user can directly input a "best guess" of the grouping structure and has the ability to control the dispersion of prior probability mass. However, the complex topology of the clustering space can create challenges in the prior elicitation process, especially relative to the more familiar case of location-scale priors with support over the real line. I conclude with a few closing thoughts, open questions, and ideas for future work.

- *On the role of directing shrinkage.* Many modern statical problems are high-dimensional in nature and so shrinkage estimators are becoming indispensable tools (especially for those working outside of the Bayesian paradigm!). Applied scientists often have prior information about these "shrinkage points" which can improve estimators that would otherwise rely on more ad-hoc default settings (for recent applications in economics, for example, see Fessler and Kasy 2019 or Smith et al. 2019). The paper's empirical application nicely highlights the often underappreciated role that model-based clustering can offer in this process.

- *What is the best way to compare and select models?* In the paper's empirical application, four different versions of the CP process are fit to the data with varying degrees of the penalty: $\psi \in \{0, 40, 80, 120\}$. The authors report distances from each model's MAP estimate $\hat{\boldsymbol{c}}$ and the centered clustering $\boldsymbol{c}_0$ and find that $d(\hat{\boldsymbol{c}}, \boldsymbol{c}_0)$ is monotonically decreasing in ψ. However, this seems to be driven by the mechanics of the prior itself and does not necessarily reflect which model is best supported by the data. I was left wondering how the inclusion of prior information here leads to improved measurements or insights? More generally, how should model fit should be assessed so that researchers can learn the extent to which the data supports or contradicts prior beliefs?

- *What happens for large N?* Many of the modeling decisions are motivated by the specific dimensions of the empirical application where $N = 26$. However, as the authors note, many aspects of their suggested prior elicitation and calibration processes become infeasible as N gets large. I am personally very excited about the opportunities to scale partitioning methods to much larger problems. For example, I work on applications in marketing and economics where the goal is measure competition between brands. The growth of e-commerce has led to massive product assortments and so in practice, retailers have a partitioning problem with N in the hundreds or thousands! One option for scaling existing methods in the short term is to impose more dogmatic prior assumptions. For example, we could impose the restriction that a subset of items must *always* be grouped together and so even if N is very large, the partitioning problem lives in a lower-dimensional space. I look forward to seeing the authors make future developments in this area.

In closing, I congratulate the authors for an exciting paper and a notable contribution to the field. I also thank the Editor-in-Chief of *Bayesian Analysis* for the opportunity to participate in this discussion.

References

Allenby, G. M. (1989). "A Unified Approach to Identifying, Estimating and Testing Demand Structures with Aggregate Scanner Data." *Marketing Science*, 8(3): 265–280. 340

Blei, D. M. and Frazier, P. I. (2011). "Distance Dependent Chinese Restaurant Processes." *Journal of Machine Learning Research*, 12(Aug): 2461–2488. MR2834504. 339, 341, 342

Botto, L. D., Lin, A. E., Riehle-Colarusso, T., Malik, S., Correa, A., and The National Birth Defects Prevention Study (2007). "Seeking Causes: Classifying and Evaluating Congenital Hearth Defects in Etiologic Studies." *Birth Defects Research Part A: Clinical and Molecular Teratology*, 79(10): 714–727. 340, 341

Dahl, D. B. (2003). "An Improved Merge-Split Sampler for Conjugate Dirichlet Process Mixture Models." Technical Report 1086, Department of Statistics, University of Wisconsin – Madison. MR2706330. 342

Dahl, D. B., Day, R., and Tsai, J. W. (2017). "Random Partition Distribution Indexed by Pairwise Information." *Journal of the American Statistical Association*, 112(518): 721–732. MR3671765. doi: https://doi.org/10.1080/01621459.2016.1165103. 339, 341, 342

Draganska, M. and Jain, D. C. (2006). "Consumer Preferences and Product-Line Pricing Strategies: An Empirical Analysis." *Marketing Science*, 25(2): 164–174. 340

Fessler, P. and Kasy, M. (2019). "How to Use Economic Theory to Improve Estimators: Shrinking Toward Theoretical Restrictions." *Review of Economics and Statistics*, 101(4): 681–698. 343

Jain, S. and Neal, R. M. (2004). "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model." *Journal of Computational and Graphical Statistics*, 13(1): 158–182. MR2044876. doi: https://doi.org/10.1198/1061860043001. 342

Jain, S. and Neal, R. M. (2007). "Splitting and Merging Components of a Nonconjugate Dirichlet Process Mixture Model." *Bayesian Analysis*, 2(3): 445–472. MR2342168. doi: https://doi.org/10.1214/07-BA219. 342

McFadden, D. (1978). *Modelling Choice of Residential Location*. Amsterdam: North-Holland. 340

Meilă, M. (2007). "Comparing Clusterings–An Information Based Distance." *Journal of Multivariate Analysis*, 98(5): 873–895. MR2325412. doi: https://doi.org/10.1016/j.jmva.2006.11.013. 341

Müller, P. and Quintana, F. A. (2011). "A Product Partition Model with Regression on Covariates." *Journal of Computational and Graphical Statistics*, 20(1): 260–278. MR2816548. doi: https://doi.org/10.1198/jcgs.2011.09066. 339, 342

Neal, R. M. (2000). "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. MR1823804. doi: https://doi.org/10.2307/1390653. 342

Park, J.-H. and Dunson, D. B. (2010). "Bayesian Generalized Product Partition Model." *Statistica Sinica*, 20: 1203–1226. MR2730180. 339, 342

Smith, A. N. and Allenby, G. M. (2020). "Demand Models With Random Partitions." *Journal of the American Statistical Association*, 115(529): 47–65. MR4078444. doi: https://doi.org/10.1080/01621459.2019.1604360. 339, 341, 342

Smith, A. N., Rossi, P. E., and Allenby, G. M. (2019). "Inference for Product Competition and Separable Demand." *Marketing Science*, 38(4): 690–710. 343

Train, K. (2002). *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edition. MR2003007. doi: https://doi.org/10.1017/CBO9780511753930. 340

Contributed Discussion

Isadora Antoniano-Villalobos[*], Cristiano Villa[†], and Sara Wade[‡]

We would like to congratulate the authors on a well written article containing innovative ideas and a compelling application. The authors propose to incorporate subjective prior information on the clustering structure by defining a centred partition process. The proposed family of processes combines well known priors for partitions (specifically exchangeable partition probability functions, EPPFs) with a measure of discrepancy from an initial partition $\boldsymbol{c}_0$, summarizing prior belief.

While it may be difficult to subjectively elicit $\boldsymbol{c}_0$ for large sample sizes, one limitation of the proposed methodology is that the prior calibration strategy is computationally expensive and therefore limited to small sample sizes. It also becomes too expensive to include hyperpriors on key parameters of the EPPF, such as the concentration parameter α of the Dirichlet process, which controls the number of clusters.

We discuss a simple, alternative idea to include the prior clustering information. Specifically, we include the initial partition $\boldsymbol{c}_0$ as a covariate, through dependent Dirichlet processes (e.g. MacEachern, 2000; Dunson and Park, 2008; Griffin and Steel, 2006; Rodriguez and Dunson, 2011) or more generally, dependent normalized random measures (e.g. Griffin and Leisen, 2017; Chen et al., 2013; Lijoi et al., 2014; Griffin et al., 2013). In this case, we view $\boldsymbol{c}_0$ as a categorical covariate, with each element $c_{0,i}$ indicating the cluster allocation of the ith data point in the initial partition. For example, we focus on the dependent normalized weights model proposed in Antoniano-Villalobos et al. (2014) and define:

$$\mathrm{p}(c_i = j|\boldsymbol{c}_0, \boldsymbol{w}, \boldsymbol{p}) = \frac{w_j \mathrm{Cat}(c_{0,i}|\boldsymbol{p}_j)}{\sum_{j'=1}^{\infty} w_{j'} \mathrm{Cat}(c_{0,i}|\boldsymbol{p}_{j'})} = \frac{w_j p_{j,c_{0,i}}}{\sum_{j'=1}^{\infty} w_{j'} p_{j',c_{0,i}}},$$

where $\boldsymbol{w} = (w_1, w_2, \ldots)$ and $\boldsymbol{p} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots)$ are the parameters defining the dependent weights. Specifically, we can assume $\boldsymbol{w}$ follow a stick-breaking construction with mass parameter α, and the $\boldsymbol{p}_j = (p_{j,1}, \ldots, p_{j,k_0})$ are iid with $\boldsymbol{p}_j \sim \mathrm{Dir}(\beta/k_0, \ldots, \beta/k_0)$.

In this construction, we can study the following limiting cases. On one hand, if $\beta \to \infty$, then

$$\boldsymbol{p}_j \to (1/k_0, \ldots, 1/k_0)$$

with probability one. Thus, the prior on the partition $\boldsymbol{c}$ converges to the EPPF induced by the DP with mass parameter α. On the other hand, if $\beta \to 0$, then base measure on $\boldsymbol{p}_j$ converges to a uniform discrete distribution over the vertices of the simplex. Moreover, when $\alpha \to 0$ and $\beta \to 0$, $\boldsymbol{c} = \boldsymbol{c}_0$ with probability one.

[*] Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Italy, isadora.antoniano@unive.it

[†] School of Mathematics, Statistics and Physics, Newcastle University, UK, Cristiano.Villa@newcastle.ac.uk

[‡] School of Mathematics, University of Edinburgh, UK, sara.wade@ed.ac.uk

While not as intuitive as the centred partition process, which includes a parameter ψ to control the tradeoff between the EPPF and subjective information, it is possible to fix the value of α to reflect prior belief in the number of clusters and β to reflect the strength of belief in $\boldsymbol{c}_0$. Moreover, an advantage of this alternative approach is the ability to place hyperpriors on α and β, thus avoiding the calibration issues and making the method more robust to misspecification. This also helps to scale to larger sample sizes.

We highlight an additional advantage of this alternative approach is the ability to simulate from the prior by using a reasonable computable truncation for the mixture weights. This can be exploited to investigate prior sensitivity, callibration or elicitation. Furthermore, prediction for new observations could incorporate prior information regarding clustering, without requiring a recalibration of the prior and subsequent recalculation of the posterior.

References

Antoniano-Villalobos, I., Wade, S., and Walker, S. (2014). "A Bayesian Nonparametric Regression Model With Normalized Weights: A Study of Hippocampal Atrophy in Alzheimer's Disease." *Journal of the American Statistical Association*, 109(506): 477–490. MR3223726. doi: https://doi.org/10.1080/01621459.2013.879061. 346

Chen, C., Rao, V., Buntine, W., and Teh, Y. W. (2013). "Dependent normalized random measures." In *International Conference on Machine Learning*, 969–977. PMLR. 346

Dunson, D. and Park, J. (2008). "Kernel Stick-Breaking Processes." *Biometrika*, 95: 307–323. MR2521586. doi: https://doi.org/10.1093/biomet/asn012. 346

Griffin, J. and Leisen, F. (2017). "Compound Random Measures and their use in Bayesian Nonparametrics." *JRSS B*, 79(2): 525–545. MR3611758. doi: https://doi.org/10.1111/rssb.12176. 346

Griffin, J. and Steel, M. (2006). "Order-Based Dependent Dirichlet Processes." *Journal of the American Statistical Association*, 10: 179–194. MR2268037. doi: https://doi.org/10.1198/016214505000000727. 346

Griffin, J. E., Kolossiatis, M., and Steel, M. F. (2013). "Comparing Distributions by Using Dependent Normalized Random-Measure Mixtures." *Journal of the Royal Statistical Society: SERIES B: Statistical Methodology*, 499–529. MR3065477. doi: https://doi.org/10.1111/rssb.12002. 346

Lijoi, A., Nipoti, B., and Prünster, I. (2014). "Bayesian Inference with Dependent Normalized Completely Random Measures." *Bernoulli*, 20(3): 1260–1291. MR3217444. doi: https://doi.org/10.3150/13-BEJ521. 346

MacEachern, S. (2000). "Dependent Dirichlet Processes." Technical report, Department of Statistics, Ohio State University. 346

Rodriguez, A. and Dunson, D. (2011). "Nonparametric Bayesian Models through Probit Stick-Breaking Processes." *Bayesian Analysis*, 6: 145–178. MR2781811. doi: https://doi.org/10.1214/11-BA605. 346

Contributed Discussion

Tommaso Rigon[*], Emanuele Aliverti[†], Massimiliano Russo[‡], and Bruno Scarpa[§]

We congratulate the authors on an interesting paper, which provides a concrete contribution in Bayesian nonparametric methods. The proposed centered partition (CP) process $p(\boldsymbol{c} \mid \boldsymbol{c}_0)$ is an exponential contamination of a baseline process $p_0(\boldsymbol{c})$ towards a fixed partition $\boldsymbol{c}_0$. The authors suggest a Gibbs-type specification for the baseline distribution $p_0(\boldsymbol{c})$, since this class displays a nice balance between flexibility and complexity (Lijoi et al., 2007). The CP informs the clustering process exploiting existing prior knowledge about the partition.

The CP process is defined as $p(\boldsymbol{c} \mid \boldsymbol{c}_0) \propto p_0(\boldsymbol{c}) \exp\{-\psi d(\boldsymbol{c}, \boldsymbol{c}_0)\}$, with $\psi > 0$ being a penalization parameter, and $d(\boldsymbol{c}, \boldsymbol{c}_0)$ being a metric between partitions, such as the Variation of Information (VI). The CP process can be also interpreted as a *generalized Bayes posterior*, in the sense of Bissiri et al. (2016). Within such a framework, the baseline distribution $p_0(\boldsymbol{c})$ represents the prior belief about an unknown partition, whereas $\boldsymbol{c}_0$ is regarded as a data point. Moreover, in the generalized Bayes terminology the distance $d(\boldsymbol{c}, \boldsymbol{c}_0)$ is the *loss function*, meaning that the parameter $\psi > 0$ balances the importance of the observations relative to the prior. This perspective leads to an alternative interpretation of CP processes, where $p(\boldsymbol{c} \mid \boldsymbol{c}_0)$ can be regarded as the posterior belief about the partition conditionally on the observation $\boldsymbol{c}_0$.

Such a generalized Bayes interpretation leads to interesting modeling extensions. In many practical contexts, it might be difficult to select a single $\boldsymbol{c}_0$ encapsulating our prior knowledge about the partition. Instead, it might be easier to identify several plausible partitions that well describe the phenomenon under consideration. For example, in the application considered by the authors, different investigators could provide equally plausible mechanistic groups of the birth defects $\boldsymbol{c}_{0,1}, \dots, \boldsymbol{c}_{0,S}$. Following Bissiri et al. (2016), it is natural to include all these representative partitions in an additive manner, namely

$$p(\boldsymbol{c} \mid \boldsymbol{c}_{0,1}, \dots, \boldsymbol{c}_{0,S}) \propto p_0(\boldsymbol{c}) \exp\left\{-\psi \sum_{s=1}^{S} d(\boldsymbol{c}, \boldsymbol{c}_{0,s})\right\}. \tag{1}$$

The above conditional distribution can be regarded as the posterior distribution of $\boldsymbol{c}$ given the observations $\boldsymbol{c}_{0,1}, \dots, \boldsymbol{c}_{0,S}$. As $\psi \to 0$ the distribution $p(\boldsymbol{c} \mid \boldsymbol{c}_{0,1}, \dots, \boldsymbol{c}_{0,S})$ converges to the baseline law $p_0(\boldsymbol{c})$. However, when $\psi \to \infty$ then $p(\boldsymbol{c} \mid \boldsymbol{c}_{0,1}, \dots, \boldsymbol{c}_{0,S})$ converges to a discrete distribution function placing mass over the set of partitions

[*] Department of Economics, Management & Statistics, University of Milano-Bicocca, Milano, Italy, tommaso.rigon@unimib.it

[†] Department of Economics, University Ca' Foscari, Venezia, Italy, emanuele.aliverti@unive.it

[‡] Harvard–MIT Center for Regulatory Science, Harvard Medical School and Department of Data Science Dana-Farber Cancer Institute, Boston, USA, m_russo@hms.harvard.edu

[§] Department of Statistical Sciences and Department of Mathematics "Tullio Levi-Civita", University of Padova, Padova, Italy, bruno.scarpa@unipd.it

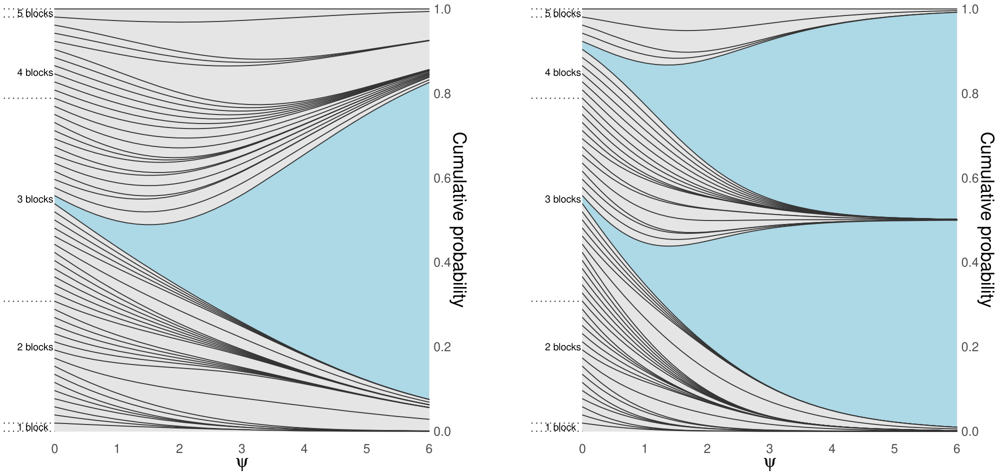


Figure 1: Prior probabilities of the 52 partitions of $N = 5$ elements for the CP process with $p_0(\boldsymbol{c}) \propto 1$. Left panel corresponds to the CP process centered on a single partition $\boldsymbol{c}_0 = \{1,2\}\{3,4\}\{5\}$. Right panel refers to a CP process centered on two partitions: $\boldsymbol{c}_{0,1} = \{1,2\}\{3,4\}\{5\}$ and $\boldsymbol{c}_{0,2} = \{1\}\{2\}\{3,4\}\{5\}$. The cumulative probabilities across different values of the penalization parameter ψ are joined to form the curves, so that the probability of a given partition corresponds to the area between the curves. Blue areas correspond to the centering partitions $\boldsymbol{c}_0$ (left plot), and $\boldsymbol{c}_{0,1}$, $\boldsymbol{c}_{0,2}$ (right plot).

$\hat{\boldsymbol{c}}_1, \dots, \hat{\boldsymbol{c}}_M$, corresponding to the minimizers of

$$\min_{\boldsymbol{c}} \sum_{s=1}^{S} d(\boldsymbol{c}, \boldsymbol{c}_{0,s}),$$

where M represents the number of solutions of the above minimization problem. Broadly speaking, each $\hat{\boldsymbol{c}}_m$, for $m = 1, \dots, M$, is an "average" partition summarizing the information contained in the observations $\boldsymbol{c}_{0,1}, \dots, \boldsymbol{c}_{0,S}$. Hence, the distribution $p(\boldsymbol{c} \mid \boldsymbol{c}_{0,1}, \dots, \boldsymbol{c}_{0,S})$ can be arguably regarded as a CP process with multiple centers $\hat{\boldsymbol{c}}_1, \dots, \hat{\boldsymbol{c}}_M$. Such a generalization of the CP is fairly straightforward and it might have useful practical implications, especially if there is uncertainty about the fixed partition $\boldsymbol{c}_0$. In addition, the Gibbs sampling devised by Paganin et al. (2021) can be easily modified to account for this extension.

In Figure 1 we reproduce Figure 2 of Paganin et al. (2021) and we illustrate the effect of our multi-centers extension. We compare the model of Paganin et al. (2021) when $p_0(\boldsymbol{c}) \propto 1$ and $\boldsymbol{c}_0 = \{1,2\}\{3,4\}\{5\}$, with the extension in (1) when $p_0(\boldsymbol{c}) \propto 1$, $S = 2$, and $\boldsymbol{c}_{0,1} = \{1,2\}\{3,4\}\{5\}$, $\boldsymbol{c}_{0,2} = \{1\}\{2\}\{3,4\}\{5\}$. Larger values of ψ increase the prior probability assigned to $\boldsymbol{c}_0$ in the left panel, and to each of the centers $\hat{\boldsymbol{c}}_1, \dots, \hat{\boldsymbol{c}}_M$ in the right panel. These centers represent the partitions that are more similar in terms of VI to $\boldsymbol{c}_{0,1}$ and $\boldsymbol{c}_{0,2}$. In this specific scenario, the centers $\hat{\boldsymbol{c}}_1, \dots, \hat{\boldsymbol{c}}_M$ actually coincide with the data points $\boldsymbol{c}_{0,1}, \boldsymbol{c}_{0,2}$ and $S = M = 2$, but this is not always the case.

References

Bissiri, P., Holmes, C., and Walker, S. (2016). “A general framework for updating belief distributions.” *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 78(5): 1103–1130. MR3557191. doi: https://doi.org/10.1111/rssb.12158. 348

Lijoi, A., Mena, R., and Prünster, I. (2007). “Controlling the reinforcement in Bayesian non-parametric mixture models.” *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 69(4): 715–740. MR2370077. doi: https://doi.org/10.1111/j.1467-9868.2007.00609.x. 348

Paganin, S., Herring, A. H., Olshan, A. F., Dunson, D. B., et al. (2021). “Centered partition processes: Informative priors for clustering.” *Bayesian Analysis*. 349

Contributed Discussion

Alejandra Avalos-Pacheco[*,†], Roberta De Vito[‡], and Sara Wade[§]

1 Introduction

We congratulate the authors on the development of a broad new class of Bayesian clustering models that allow for inclusion of prior information on the clustering structure, and result in improvement of model performance in practice when such prior information is available. The authors applied this method in an interesting and novel context where there are $N = 26$ different birth defects, and for each defect $i \in \{1, \ldots, N\}$, there is a highly variable number of observations. The prior knowledge of the initial partition $\mathbf{c}_0$ of the birth defects provided by experts is merged with information in the data through a grouped logistic regression model to produce a posterior distribution on the partition, which also characterizes uncertainty.

2 Biomedical Applications

In the following, we discuss two possible biomedical applications.

2.1 Clinical Trials

The model proposed by Paganin et al. (2020) is relevant in the clinical trial setting, in particular in novel trial designs such as *master protocols*. Master protocols are clinical designs that study in parallel multiple therapies, different sub-populations, multiple diseases and/or several targets (Woodcock and LaVange, 2017). Such innovative designs, when designed correctly, maximize the number of patients assigned to promising novel therapies, reduce the overall sample size of the trial, minimize downtime between trials, and reduce costs; overall, this assists in significantly speeding up drug discovery, in comparison with standard two-arm randomized controlled trials. Master protocols, can last for decades, allowing for incorporation of new therapies to an ongoing study at any time (Angus et al., 2019). This is of particular interest in case of pandemics, enabling the study to evaluate multiple novel treatments for critically ill patients, as was the case of the REMAP-CAP trial in the COVID-19 pandemic (Angus et al., 2020).

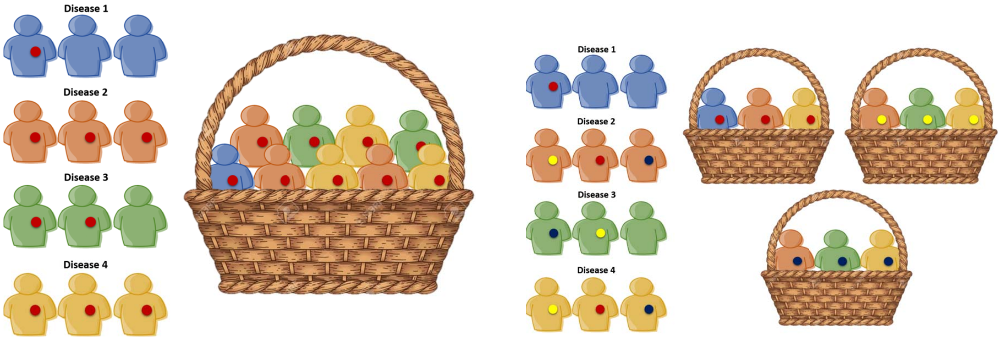
Basket or *bucket trials* are a particular type of master protocols that test the efficacy or safety of a novel treatment in a group of patients with different diseases that have the same mutation or biomarker, or with different subtypes of the same disease (Woodcock

[*] Harvard-MIT Center for Regulatory Science, Harvard Medical School, Boston, USA, avalos@hms.harvard.edu

[†] Department of Data Science, Dana-Farber Cancer Institute, Boston, USA

[‡] Department of Biostatistics and Data Science Initiative, Brown University, Providence, USA, roberta_devito@brown.edu

[§] School of Mathematics, University of Edinburgh, Edinburgh, UK, sara.wade@ed.ac.uk



(a) Basket trial

(b) Master protocol

Figure 1: Illustration of a basket trial (left) and a master protocol (right) for four different diseases. Biomarkers/mutations are illustrated with coloured circles.

and LaVange, 2017). Figure 1a provides an illustration of this clinical design. For example, in oncology, patients recruited in the basket study have the same genetic mutation but cancers located in different regions such as the lung, liver, prostate, etc.

Borrowing of information between disease groups in basket trials is usually done via Bayesian hierarchical models (Berry et al., 2013; Ventz et al., 2017). In this setting, all diseases are pooled together and are assumed to be exchangeable, which could lead to inflated type I error rates and reduced power (Freidlin and Korn, 2013). Thus, cluster-based models can be employed to borrow information, which are especially appealing when the differences between groups are large (Chen and Lee, 2019, 2020). However, in these methods, the number of clusters needs to be pre-specified and/or does not allow for incorporation of prior knowledge of the initial partition of the different groups.

Let $j = 1, \ldots, N$ index disease types and $i = 1, \ldots, n_j$ index patients with disease type j, with y_{ij} denoting the i^{th} response to the treatment of patient i with disease type j. Let $x_{ij} = (x_{ij1}, \ldots, x_{ijp})^\top$ be the p-dimensional vector of the observed pre-treatment characteristics for patient i with disease type j. The indicator $a_{ij} = 0$ or $a_{ij} = 1$ if patient i was assigned to the control group or the treatment group, respectively. Letting θ_j denote the treatment effect for disease type j, we test, for each disease type, the hypotheses

$$H_{0,j} : \theta_j \leq \delta_j \quad \text{vs} \quad H_{1,j} : \theta_j > \delta_j.$$

We model the data as:

$$\log\left(\frac{p(y_{ij} = 1 \mid x_{ij}, a_{ij})}{p(y_{ij} = 0 \mid x_{ij}, a_{ij})}\right) = \operatorname{logit}(\pi_{ij}) = \alpha_j + x_{ij}^\top \beta^*_{c_j} + \theta^*_{c_j} \mathrm{I}(a_{ij} = 1),$$

with α_j the disease-specific intercept, β^*_k the cluster-specific pre-treatment coefficients, and $\mathbf{c} = (c_1, \ldots, c_N)$ the cluster allocations. We propose to use the centered partition process to model $\mathbf{c}$ and incorporate expert knowledge on the grouping of disease types.

An attractive property of this model is that it can easily be adapted to cluster only the coefficients β_j, when there is no certainty about the exchangeability of the treatment effects θ_j within the diseases of the cluster, avoiding inflation/deflation of type I error/power. We could also model the data taking into account only the treatment effects, without including the pre-treatment patient characteristics. This setting can be seen as a traditional beta-binomial model, when borrowing of information can be done through θ_{c_j}. Finally, the design can also become more complex including more than one genetic mutation across multiple diseases (see e.g. Figure 1b), and the clustering structure can be adapted appropriately.

2.2 Nutritional Epidemiology

This application is motivated by nutritional epidemiological data analysis of single foods (Wirfalt and Jeffery, 1997; Hu, 2002). In this framework, models considering a single food can be difficult to interpret since foods are consumed in many different combinations, and studies of individual foods present strong inter-correlations (Hearty and Gibney, 2008). Factor analysis and cluster analysis are used to find groups of foods, referred to as dietary patterns (Edefonti et al., 2008; Brennan et al., 2010; Grosso et al., 2017; De Vito et al., 2019). The solution obtained by these dimension reduction techniques is used to predict disease risk. However, there is a discussion on how to group the foods, for example, based on their association with the disease or using a priori food groups provided by a nutritional expert; importantly, these different techniques can lead to quite different results. The proposed prior of Paganin et al. (2020) would allow for uncertainty in the food groups, while also incorporating prior information. Then including this in a logistic regression model to associate food consumption with disease risk, could improve cluster food methods and identification of dietary patterns.

We can adopt the model in this framework. Let $i = 1, \dots, n$ index the subject, $j = 1, \dots, p$ index the single foods, and y_i represents the case or control for a disease (e.g. cancer or cardiovascular disease). For each individual, let $x_i = (x_{i1}, \dots, x_{ip})^\top$ be the food level consumption by the subject i. Finally, let $z_i = (z_{i1}, \dots, z_{ir})^\top$ denote the vector of confounders (i.e., alcohol consumption or smoking status). The proposed model is

$$\log\left(\frac{p(y_i = 1 \mid x_i, z_i)}{p(y_i = 0 \mid x_i, z_i)}\right) = \alpha + z_i^\top \theta + \sum_{j=1}^{p} x_{ij}\beta^*_{c_j} = \alpha + z_i^\top \theta + \sum_{k=1}^{K} \beta^*_k (\sum_{j \in C_k} x_{ij}).$$

This allows us to cluster the foods which have similar effects, and we can employ the centred partition process to incorporate an initial partition $\mathbf{c}_0$ of food groups.

One other difference with the model studied in Paganin et al. (2020) is the inclusion of confounders in this logistic framework to detect food groups, independently from other covariates. The clustering framework is crucial to improve the estimation of β and interpretability (especially compared to dimension reduction techniques). The clustering itself is also interesting and could help nutritional epidemiologists detect groups of food more formally, along with measures of uncertainty.

3 Final Remarks

In this discussion, we have highlighted two possible applications of the proposed centered partition process, but this novel process is relevant and applicable in many other settings. However, the application in nutritional epidemiology highlights the need of alternative methods of prior calibration that can scale to higher dimensions. The proposed prior calibration does not scale well when the number of objects is much larger than the $N = 26$ considered in the motivating birth defects application. This is due to a combinatorial explosion of the partition space as N increases which leads to an inevitable deterioration of their prior calibration algorithm. Specifically, as the number of observations N increases, the number of partitions explodes, and higher values of the penalization parameter ψ are needed to place non-negligible prior probability in small to moderate neighborhoods around $\mathbf{c}_0$.

References

Angus, D. C., Alexander, B. M., Berry, S., and et al. (2019). "Adaptive platform trials: Definition, design, conduct and reporting considerations." *Nature Reviews. Drug Discovery*, 18(10): 797–807. 351

Angus, D. C., Derde, L., Al-Beidh, F., and et al. (2020). "Effect of hydrocortisone on mortality and organ support in patients with severe COVID-19: The REMAP-CAP COVID-19 corticosteroid domain randomized clinical trial." *Journal of the American Medical Association*, 324(13): 1317–1329. 351

Berry, S. M., Broglio, K. R., Groshen, S., and Berry, D. A. (2013). "Bayesian hierarchical modeling of patient subpopulations: Efficient designs of Phase II oncology clinical trials." *Clinical Trials*, 10(5): 720–734. 352

Brennan, S. F., Cantwell, M. M., Cardwell, C. R., Velentzis, L. S., and Woodside, J. V. (2010). "Dietary patterns and breast cancer risk: A systematic review and meta-analysis." *American Journal of Clinical Nutrition*, 91(5): 1294–1302. 353

Chen, N. and Lee, J. J. (2019). "Bayesian hierarchical classification and information sharing for clinical trials with subgroups and binary outcomes." *Biometrical Journal*, 61(5): 1219–1231. MR4013344. doi: https://doi.org/10.1002/bimj.201700275. 352

Chen, N. and Lee, J. J. (2020). "Bayesian cluster hierarchical model for subgroup borrowing in the design and analysis of basket trials with binary endpoints." *Statistical Methods in Medical Research*, 29(9): 2717–2732. MR4129440. doi: https://doi.org/10.1177/0962280220910186. 352

De Vito, R., Lee, Y. C. A., Parpinel, M., Serraino, D., Olshan, A. F., Zevallos, J. P., Levi, F., Zhang, Z. F., Morgenstern, H., Garavello, W., et al. (2019). "Shared and study-specific dietary patterns and head and neck cancer risk in an international consortium." *Epidemiology*, 30(1): 93. 353

Edefonti, V., Decarli, A., Vecchia, C. L., Bosetti, C., Randi, G., Franceschi, S., Maso,

L. D., and Ferraroni, M. (2008). "Nutrient dietary patterns and the risk of breast and ovarian cancers." *International Journal of Cancer*, 122(3): 609–613. 353

Freidlin, B. and Korn, E. L. (2013). "Borrowing information across subgroups in phase II trials: is it useful?" *Clinical Cancer Research*, 19(6): 1326–1334. 352

Grosso, G., Bella, F., Godos, J., Sciacca, S., Del Rio, D., Ray, S., Galvano, F., and Giovannucci, E. L. (2017). "Possible role of diet in cancer: Systematic review and multiple meta-analyses of dietary patterns, lifestyle factors, and cancer risk." *Nutrition Reviews*, 75(6): 405–419. 353

Hearty, A. P. and Gibney, M. J. (2008). "Comparison of cluster and principal component analysis techniques to derive dietary patterns in Irish adults." *British Journal of Nutrition*, 101(4): 598–608. 353

Hu, F. B. (2002). "Dietary pattern analysis: a new direction in nutritional epidemiology." *Current Opinion in Lipidology*, 13(1): 3–9. 353

Paganin, S., Herring, A. H., Olshan, A. F., and Dunson, D. B. (2020). "Centered partition processes: Informative priors for Clustering." *Bayesian Analysis*. 351, 353

Ventz, S., Barry, W. T., Parmigiani, G., and Trippa, L. (2017). "Bayesian response-adaptive designs for basket trials." *Biometrics*, 73(3): 905–915. MR3713124. doi: https://doi.org/10.1111/biom.12668. 352

Wirfalt, A. E. and Jeffery, R. W. (1997). "Using cluster analysis to examine dietary patterns: Nutrient intakes, gender, and weight status differ across food pattern clusters." *Journal of the American Dietetic Association*, 97(3): 272–279. 353

Woodcock, J. and LaVange, L. M. (2017). "Master protocols to study multiple therapies, multiple diseases, or both." *The New England Journal of Medicine*, 377(1): 62–70. 351

Contributed Discussion

Laura D'Angelo[*] and Antonio Canale[†]

We would like to congratulate the authors for their valuable contribution to Bayesian model-based clustering. We believe that one of the most appealing peculiarities of Bayesian statistics is the opportunity of introducing informative prior knowledge into the learning process. The centered partition processes (CPP) introduced by the authors have the merit of allowing an elaborate and rich informative prior elicitation in the context of some the most successful tools used in Bayesian model-based clustering, namely the Dirichlet process (DP, Ferguson, 1974), the Pitman-Yor process (Pitman and Yor, 1997), or, in general, the Gibbs-type priors (De Blasi et al., 2015).

Paganin et al. (2021) key idea is to penalize usual Exchangeable Partition Probability Functions (EPPF) of a general partition $\boldsymbol{c}$ with an exponential factor that depends on its distance $d(\boldsymbol{c}, \boldsymbol{c}_0)$ from $\boldsymbol{c}_0$, an informed prior guess for the partition of the dataset. The idea of using a suitable distance function in the space of partitions $d(\boldsymbol{c}, \boldsymbol{c}_0)$ is reasonable in many settings considering its symmetric nature. In some applications, however, we would like to avoid this intrinsic symmetry and differently penalize partitions with different characteristics even if they have the same distance from $\boldsymbol{c}_0$. The most trivial example, which we will discuss hereafter, consists in those situations when, in addition to an informed prior guess $\boldsymbol{c}_0$, we also seek a parsimonious clustering structure, thus preferring a partition with a small number of clusters. Application areas where interpretability and parsimony in the number of clusters are fundamental include genetics (Fu and Perry, 2020), market segmentation (Wagner et al., 2005), topic modeling (Yau et al., 2014), or network data analysis (Vu et al., 2013) among others. Consistently with this applied motivation in what follows we sketch a possible modification of the approach presented in the paper.

1 An asymmetric penalization for the CPP

We propose to generalize Equation 3.1 of the paper with

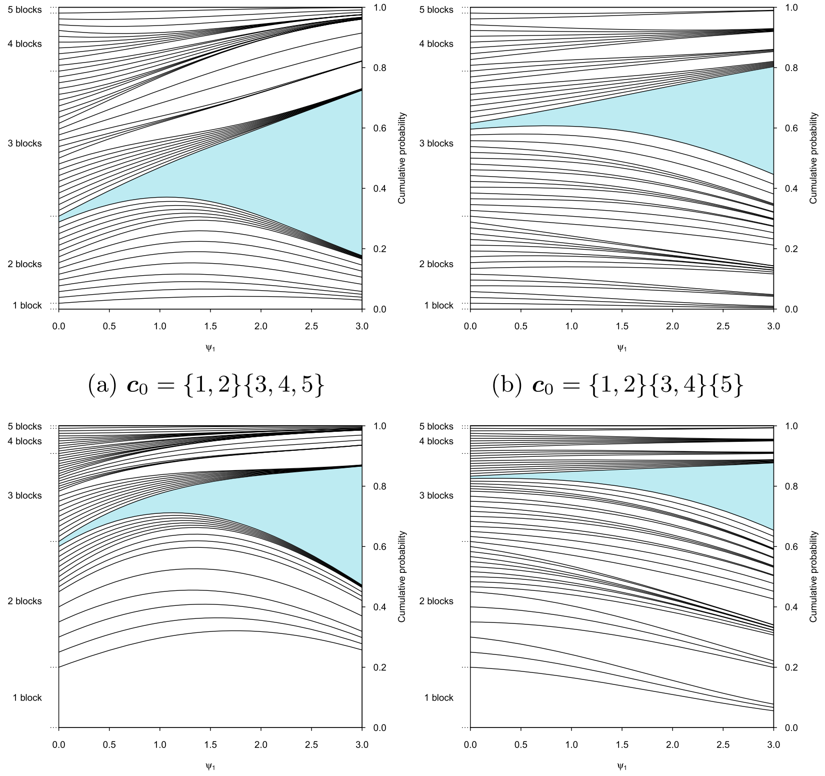
$$p(\boldsymbol{c} \mid \boldsymbol{c}_0, \varphi_d) \propto p_0(\boldsymbol{c})\, e^{-\varphi_d(\boldsymbol{c}, \boldsymbol{c}_0)}, \tag{1}$$

where $\varphi_d(\boldsymbol{c}, \boldsymbol{c}_0)$ is a general function of $\boldsymbol{c}$ and $\boldsymbol{c}_0$ depending on the distance function d. When $\varphi_d(\boldsymbol{c}, \boldsymbol{c}_0) = \psi d(\boldsymbol{c}, \boldsymbol{c}_0)$ and ψ is a positive scalar, we get the original formulation presented in Equation 3.1. Consistently with our applied motivation of penalizing partitions with many clusters, we let

$$\varphi_d(\boldsymbol{c}, \boldsymbol{c}_0) = \{\psi_1\, \mathcal{I}(K_{\boldsymbol{c}} \le K_0) + \psi_2\, \mathcal{I}(K_{\boldsymbol{c}} > K_0)\} d(\boldsymbol{c}, \boldsymbol{c}_0), \tag{2}$$

[*] Department of Statistical Sciences, University of Padova, Via C. Battisti 241, Padova, Italy, laura.dangelo.1@phd.unipd.it

[†] Department of Statistical Sciences, University of Padova, Via C. Battisti 241, Padova, Italy, antonio.canale@unipd.it



(a) $\boldsymbol{c}_0 = \{1,2\}\{3,4,5\}$ (b) $\boldsymbol{c}_0 = \{1,2\}\{3,4\}\{5\}$

(c) $\boldsymbol{c}_0 = \{1,2\}\{3,4,5\}$ (d) $\boldsymbol{c}_0 = \{1,2\}\{3,4\}\{5\}$

Figure 1: Prior probabilities of the 52 set partitions of $\mathrm{N} = 5$ elements for the CPP with uniform (top) and DP with $\alpha = 1$ (bottom) base EPPF. The asymmetric penalization follows (2) with $\psi_2 = 1.5\,\psi_1$ and different values of ψ_1.

where $K_{\boldsymbol{c}}$ and K_0 are the number of clusters of $\boldsymbol{c}$ and $\boldsymbol{c}_0$, respectively, ψ_1 and ψ_2 are two positive scalars, and $\mathcal{I}(\cdot)$ is the indicator function.

Figure 1, similarly to Figures 2–3 in the paper, shows the effect of this asymmetric penalization when the base EPPF is a uniform distribution (top panels) or a DP (bottom panels), $N = 5$, and $\psi_2 = 1.5\psi_1$. The plots depict the prior probabilities assigned to partitions for different values of the parameter $\psi_1 \in (0, 3)$: a quick visual comparison with the plots in panels (b) and (c) of Figures 2 and 3 of the paper clearly shows the effect of this penalization. The prior probability assigned to partitions with a small number of clusters are inflated proportionally to their distance from $\boldsymbol{c}_0$, while the probabilities for the finer partitions are shrunk. For example, for the case of a uniform prior and a centering partition with two clusters (Figure 1a), the cumulative probability assigned to partitions with two or less clusters using a constant penalization and $\psi = 3$ or using the proposed asymmetric penalization with $\psi_1 = 3$ and $\psi_2 = 4.5$ are equal to 0.491 and 0.725, respectively.

Prior calibration remains a delicate issue in general but for the specific formulation in (2), one can express $\psi_2 = k\,\psi_1$, and then adapt computation of the probability distribution of the distances (Section 4 of the paper) by splitting the procedure for finer and coarser partitions, keeping in mind that the asymmetry in the penalty potentially affects also the order of the local search in Section 4.2 of the paper.

References

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2): 212–229. 356

Ferguson, T. S. (1974). "Prior distributions on spaces of probability measures." *The Annals of Statistics*, 2(4): 615–629. MR0438568. 356

Fu, W. and Perry, P. O. (2020). "Estimating the number of clusters using cross-validation." *Journal of Computational and Graphical Statistics*, 29(1): 162–173. MR4085872. doi: https://doi.org/10.1080/10618600.2019.1647846. 356

Paganin, S., Herring, A. H., Olshan, A. F., and Dunson, D. B. (2021). "Centered partition processes: Informative priors for clustering." *Bayesian Analysis*, 1–32. 356

Pitman, J. and Yor, M. (1997). "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator." *The Annals of Probability*, 25(2): 855–900. MR1434129. doi: https://doi.org/10.1214/aop/1024404422. 356

Vu, D. Q., Hunter, D. R., and Schweinberger, M. (2013). "Model-based clustering of large networks." *The Annals of Applied Statistics*, 7(2): 1010—-1039. MR3113499. doi: https://doi.org/10.1214/12-AOAS617. 356

Wagner, R., Scholz, S., and Decker, R. (2005). "The Number of Clusters in Market Segmentation." In D., B., R., D., and L., S.-T. (eds.), *Data Analysis and Decision Support. Studies in Classification, Data Analysis, and Knowledge Organization*, 157–176. Berlin, Heidelberg: Springer. 356

Yau, C.-K., Porter, A., Newman, N., and Suominen, A. (2014). "Clustering scientific documents with topic modeling." *Scientometrics*, 100: 767–786. 356

Contributed Discussion

Christian Hennig[*]

The key idea of “Centered Partition Processes: Informative Priors for Clustering” is the use of a pre-specified clustering $\mathbf{c}_0$ to “center” the prior distribution of partitions. The prior probability of a partition is then governed by the Variation of Information distance to $\mathbf{c}_0$.

Generally, in Bayesian statistics the prior distribution is meant to represent prior information. Ideally the prior information comes as full probability distribution, but in reality this is rarely the case. In the motivating example, apparently the only information that is used is the partition given in Botto et al. (2007); no subject matter reasons are discussed or given for the way in which this was chosen to influence the prior construction. In particular, the parameter ψ seems to have a strong influence on the resulting clustering, but its choice has not been connected to any available information.

In Hennig (2015a,b) I have argued that there can be different legitimate clusterings on the same data and different concepts of what kind of clusters are of interest, depending on the aim of clustering, and that different methods and approaches imply different “cluster concepts”. The idea that the pre-specified $\mathbf{c}_0$ is aimed at the same “truth” as the clustering of the new data using the authors’ approach is debatable. Prior construction in Bayesian clustering can benefit from involving information about the aim of clustering rather than thinking in terms of trying to find a unique “true” clustering.

For the use of the method proposed by the authors with a pre-specified clustering it is important to think about how the concept and aim of the pre-specified clustering relates to what the authors try to achieve with their clustering. I have no expertise in birth defects, so I cannot discuss this competently, but it could be worthwhile to use more detailed information given in Botto et al. (2007) to this end. There are various possible principles to group birth defects. The cluster concept used by the authors is defined by (5.2) and the role of β_{c_i} in particular. The cluster concept of Botto et al. (2007) is apparently not related to fitting data, but rather based on specific characteristics of interest. To what extent these are related probably depends on how these characteristics are related to the variables in the matrices $\mathbf{X}_i$. I can imagine that it makes sense to favour to some extent clusters similar to $\mathbf{c}_0$, but I think that quite strong subject matter arguments would be required to make the case for a very large prior probability concentrating in a close neighbourhood of $\mathbf{c}_0$, i.e., effectively excluding clusterings that are substantially different. I also think that among clusterings that are very different from $\mathbf{c}_0$ the exact difference to $\mathbf{c}_0$ is no longer relevant (if it turns out that the data favour a substantially different clustering, relative closeness to $\mathbf{c}_0$ seems no longer informative). Therefore I would probably favour rather small values of ψ, and I am skeptical about prior probabilities going down exponentially with growing distance to $\mathbf{c}_0$.

[*] Dipartimento di Scienze Statistiche “Paolo Fortunati”, Universita di Bologna, Via delle belle arti 41, 40126 Bologna, Italy, christian.hennig@unibo.it

A major reason for clustering is given as lumping together rare defects in order to achieve better precision when estimating their association with risk factors. This suggests that a useful prior could put low probability on undesirable partitions in which rare defects are still isolated, regardless of their distance from $\mathbf{c}_0$.

My main point is however that for applying this approach properly, more thought should go into the role and meaning of $\mathbf{c}_0$ for clustering the new data, and this should involve how and to what extent the potentially different clustering aims and cluster concepts are related in the specific situation. In many applications a comparison of $\mathbf{c}_0$ with what is achieved without using it at all (i.e., $\psi = 0$) may be more informative than just choosing a specific compromise. In fact Figure 10 is quite informative, except that I would have preferred to see a smaller $\psi > 0$ involved.

References

Botto, L. D., Lin, A. E., Riehle-Colarusso, T., Malik, S., and Correa, A. (2007). “Seeking causes: Classifying and evaluating congenital heart defects in etiologic studies.” *Birth Defects Research Part A: Clinical and Molecular Teratology*, 79(10): 714–727. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/bdra.20403` 359

Hennig, C. (2015a). “Clustering Strategy and Method Selection.” In Hennig, C., Meila, M., Murtagh, F., and Rocci, R. (eds.), *Handbook of Cluster Analysis*, chapter 31, 703–730. Chapman & Hall/CRC, Boca Raton FL. 359

Hennig, C. (2015b). “What are the true clusters?” *Pattern Recognition Letters*, 64: 53–62. 359

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

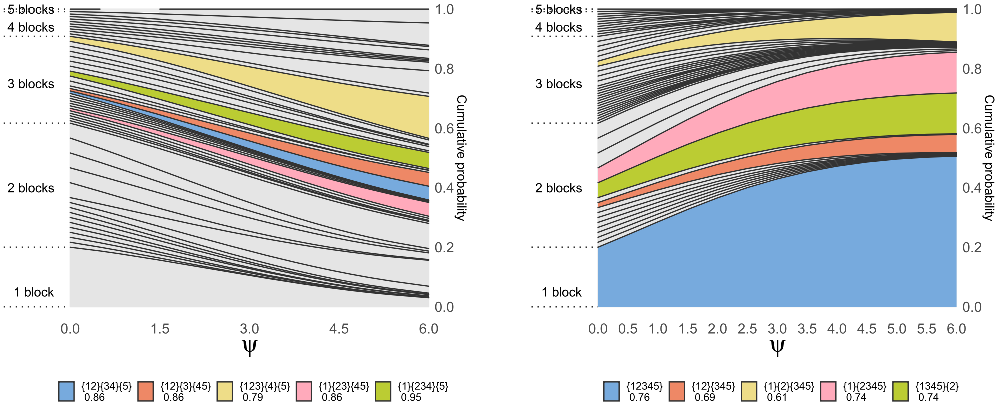


Figure 1: Prior probabilities of the 52 set partitions of $N = 5$ elements for the prior (1) with Dirichlet process of $\alpha = 1$ base EPFF. In each graph the modified CP process is centered on a different set of partitions $\mathcal{C}_0$ highlighted with different colors. The partitions in $\mathcal{C}_0$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

Rejoinder

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

We thank the editorial board of Bayesian Analysis and the editor-in-chief Michele Guindani for inviting this stimulating discussion. We are also grateful to the discussants for the thoughtful remarks and the very interesting extensions. It is exciting that the idea of a partition distribution that uses prior information on the partition itself has been recognized as valuable, and with such enthusiasm.

Our proposal is motivated by an epidemiological application involving data on birth defects, where prior information is available about a partition of these defects into groups, based on fetal developmental and epidemiologic considerations. The methodology we presented investigates how to include such information into a prior distribution over the set of possible clusterings, building on Bayesian nonparametric priors for model-based clustering, in particular Exchangeable Partition Probability Functions (EPPF). As highlighted by the discussants, there are a number of relevant extensions and open problems.

On how we interpreted the prior guess

One common theme concerns the meaning of the prior guess $\boldsymbol{c}_0$, and how we translated this information into a prior probability distribution. Our proposal considers a single partition $\boldsymbol{c}_0$, that we wish to account for when constructing a prior distribution over the space of set partitions. To do so, we start from a baseline EPPF and include an exponential penalization that depends on a distance from $\boldsymbol{c}_0$ and a penalization parameter ψ. Although we considered the Dirichlet Process (DP) and Variation of Information (VI) as default EPPF and distance, we developed a general framework that we name a Centered Partition (CP) process.

However, as **Dahl, Warr, and Jensen** argue, $\boldsymbol{c}_0$ is not necessarily the “center” of the CP process in general, where the center partition is defined as the partition that minimizes the expected loss for some partition metric. Under a uniform baseline EPPF, $\boldsymbol{c}_0$ is indeed the center of the CP process, but, as **Dahl, Warr, and Jensen** show, different combinations of the EPPF and the penalization parameter ψ lead to different center partitions. We agree with **Dahl, Warr, and Jensen**’s suggestion of thinking of the CP process as a baseline partition distribution shrunk towards $\boldsymbol{c}_0$.

The CP process induces shrinkage towards $\boldsymbol{c}_0$ via a distance dependent exponential penalty that includes a parameter ψ controlling for the amount of shrinkage. Depending

[*] Department of Biostatistics, Harvard School of Public Health, Harvard University, Boston, spaganin@hsph.harvard.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 on the value of ψ, partitions in a small to moderate neighborhood close to $\boldsymbol{c}_0$ have higher prior probability, with the notion of "closeness" defined by the distance.

In formulating our CP process, we implicitly make some assumptions that are mainly discussed in the contributions from **Smith** and **Hennig**. In particular, we assume that having a prior guess $\boldsymbol{c}_0$ translates into strong beliefs about clusters within some neighborhood of $\boldsymbol{c}_0$, and that prior probabilities decay exponentially with growing distance from $\boldsymbol{c}_0$.

In our particular application to birth defects we referred to one classification given in Botto et al. (2007), chosen in collaboration with epidemiologists involved in the National Birth Defects Prevention Study. This is a case in which there is some leading biological theory, and researchers wish to interpret results from data analysis with respect to the theory; in such settings, we support the choice of stronger and localized priors rather than weaker ones.

In general, one needs to take in account the clustering behavior induced by the baseline EPPF. **Dahl, Warr, and Jensen** warn about possible incompatibility between $\boldsymbol{c}_0$ and the EPPF, considering as an example a base partition with equal-sized clusters and the DP EPPF. In this setting, the penalty would favor partitions with similar cluster sizes, which seems contradictory to the "rich-get-richer" property of the DP with $\alpha = 1$. We agree that it is desirable that the EPPF is congruous with $\boldsymbol{c}_0$, and we stress that the choice of the hyperparameters of the DP and PY processes should be tuned accordingly. In practice, one can check that the a priori expected value and variance of the number of clusters are compatible with $\boldsymbol{c}_0$.

Finally, we definitely share **Hennig**'s perspective that different clustering concepts can apply to the same data depending on the objective of the analysis, and that these clustering concepts are often embedded in the choice of method. The implicit assumptions of our CP process may not be appropriate in all contexts. Domain knowledge and inferential goals can suggest different definitions of prior distributions over the set partitions space. However, some cases can still be easily addressed within our framework, and some of the discussions provided interesting examples and extensions.

Extensions and alternatives

We have been considering extending our proposal to account for more complex information about the data clustering, and we are excited that some of the discussions shed some light on interesting extensions of the CP process.

Particularly relevant to our birth defects application is the case of having multiple plausible prior partitions, i.e. $\mathcal{C}_0 = \{\boldsymbol{c}_{01}, \dots, \boldsymbol{c}_{0S}\}$. This situation is explored in **Smith**, **Rigon, Aliverti, Russo, and Scarpa** and **Casa, Fop, and Murphy**. While **Smith** formulates a two-stage prior, **Rigon, Aliverti, Russo, and Scarpa** and **Casa, Fop, and Murphy** extend our CP process formulation considering the sum of distances $\sum_{s=1}^{S} d(\boldsymbol{c}, \boldsymbol{c}_{0s})$ in the exponential penalty. The interpretation in **Rigon, Aliverti, Russo, and Scarpa** in light of the generalized Bayes perspective (Bissiri et al., 2016), seems to tie the two representations together, as the resulting $p(\boldsymbol{c}|\mathcal{C}_0)$ can be regarded as the posterior belief about the partition conditionally on observations in $\mathcal{C}_0$. **Smith**'s suggestion of hierarchical structure is particularly compelling in our application, in which it is possible to group birth defects more finely than shown in our manuscript.

Motivation in these discussions comes from different assumptions on the prior information available. For example $\mathcal{C}_0$ can include: a set of alternative partitions to a given classification (partitions in the same neighborhood); a set of very different partitions (far in the set partition space) coming from different leading theories or experts not agreeing with each other; or a relevant portion of the set partition space complying with partial information about the clustering, for example a set of cannot/must link constraints.

Another interesting extension of the CP process is given in **D'Angelo and Canale**, starting from considerations regarding the aim of the clustering, rather than the domain knowledge. The authors consider the case in which one wants to favor partitions in neighborhood of $\boldsymbol{c}_0$ that are coarser (i.e., parsimonious) rather than finer. Some asymmetric preference could be induced with the choice of different distances (e.g. the Binder Loss) or EPPFs (rich-get-richer behavior of the DP), however it can be hard to control for specific clustering characteristics of interest. Similar to the previous discussants, **D'Angelo and Canale** modify the exponential penalty, using different penalization parameters (ψ_1, ψ_2) depending on the number of clusters. In this formulation, preference for a parsimonious partition in a neighborhood of $\boldsymbol{c}_0$ is therefore made explicit, and one can accommodate alternative prior information on characteristics of the partitions.

The extensions discussed in **Casa, Fop, and Murphy**, **Rigon, Aliverti, Russo, and Scarpa** and **D'Angelo and Canale**, can be included in a generalized CP process defining as prior for $\boldsymbol{c} \in \Pi_N$,

$$p(\boldsymbol{c}|\mathcal{C}_0) \propto p_0(\boldsymbol{c}) \exp\left\{-\psi \sum_{s=1}^{S} \varphi_s(\boldsymbol{c}, \boldsymbol{c}_{0s}, d(\cdot))\right\},$$

where $\varphi_s(\cdot)$ is a general function that depends on the partition in $\mathcal{C}_0$ and the choice of the distance between partitions. This general formulation allows to account for multiple prior guesses, together with characteristics of interest of the partition. Considering for example the number of clusters as in **D'Angelo and Canale**, then $\varphi_s(\boldsymbol{c}, \boldsymbol{c}_{0s}) = \{\kappa_{1s}\mathbb{I}(|\boldsymbol{c}| \geq |\boldsymbol{c}_{0s}|) + \kappa_{2s}\mathbb{I}(|\boldsymbol{c}| < |\boldsymbol{c}_{0s}|)\}d(\boldsymbol{c}, \boldsymbol{c}_{0s})$, where $\{\kappa_{1s}, \kappa_{2s}\}$ are positive scalars that can change depending on the partition in $\mathcal{C}_0$.

Finally, **Antoniano-Villalobos, Villa, and Wade** build on Antoniano-Villalobos et al. (2014), interpreting $\boldsymbol{c}_0$ as a covariate, so that it can be included in a dependent normalized random measure. This formulation shares the same limiting behavior of our CP process, and some control on the prior probability mass is achieved via setting hyperparameters related to the number of clusters and strength of the influence of the covariate $\boldsymbol{c}_0$. Nevertheless, it seems hard to obtain an interpretation of where the prior probability mass is placed in the set partition space with respect $\boldsymbol{c}_0$. However we think this a possible scalable alternative, especially in situations where one wants to include mild dependence on $\boldsymbol{c}_0$ while not necessarily inflating the prior probability of partitions in a neighborhood of $\boldsymbol{c}_0$.

Computational challenges

As highlighted from different discussants, our CP process provides an intuitive representation for an informative prior distribution over the clustering space, but it comes with a computational price. Computational challenges and possible alternatives are considered in the discussions from **Dahl, Warr, and Jensen** and **Smith**.

Prior calibration. In the paper, we propose a prior calibration procedure to choose a value for the penalization parameter ψ. As noted by **Smith**, we bypass direct elicitation of ψ, focusing instead on the distribution of distances from $\boldsymbol{c}_0$ induced from our prior. Using the c.d.f. of this random variable, we can choose a probability q and a distance δ^*, and determine ψ such that the CP process places a probability of at least q on partitions within distance δ^*. We estimate the c.d.f. via Monte Carlo, relying on a greedy search algorithm paired with direct sampling on the partition space, a procedure difficult to scale with N.

Smith suggests avoiding exact computation of the parameter ψ via the prior calibration, considering a grid of plausible values to integrate out ψ during the posterior sampling. This option does not increase the computational burden of the posterior sampling, and it looks in line with our suggestion of considering other values of the penalization parameter besides the elicited one. However, we argue that choosing a plausible grid values for ψ still needs some sort of prior calibration procedure, as the same value of ψ can induce very different priors depending on the given partition $\boldsymbol{c}_0$, the type of distance, the chosen EPPF, and the number of objects to cluster.

An alternative strategy for prior calibration is given in the discussion from **Dahl, Warr, and Jensen**. They suggest to estimate the c.d.f. of the distance from $\boldsymbol{c}_0$ sampling from the prior itself, adapting the algorithm for posterior computation. We considered this strategy while developing our CP process, but found it not robust in some sense. In particular, we observed degenerate behaviors of the sampler, such as being stuck in local modes or collapsing towards $\boldsymbol{c}_0$. This behavior can potentially be related to situations where there is some incongruity between $\boldsymbol{c}_0$ and the baseline EPPF, as mentioned from **Dahl, Warr, and Jensen**. These issues motivated us to look into a solution that uses samples obtained independently from the prior. On the other hand, we recognize that improvements to the algorithm for posterior computation could benefit prior sampling. Some possible improvements are discussed in the next paragraphs.

Posterior computation. Our strategy for posterior sampling builds on traditional MCMC schemes for random partition models. In particular, we adapted Algorithm 2 in Neal (2000) which uses one-at-a-time moves, sampling group indicators for clustering objects iteratively.

As noted from **Smith**, a possible concern is that these one-at-a-time moves may be susceptible to getting stuck in local modes, not allowing the sampler to fully explore the posterior distribution. b Although we did not observe such problems in our applications, we definitely agree that our strategy for posterior sampling can be improved, for example adding split-and-merge moves (Dahl, 2003; Jain and Neal, 2004, 2007) or adapting the more recent proposal in Bouchard-Côté et al. (2017). We thank **Smith** for pointing out also the strategy adopted in Smith and Allenby (2020) of building random-walk style

M-H proposals relying on the CP process itself. This seems promising to extend our analysis of birth defect data beyond what we presented in the paper.

On the topic of posterior sampling, **Dahl, Warr, and Jensen** add that local modes could also have been induced by the use of the VI as a distance, suggesting the Binder (1978)'s loss as a more robust alternative for the distance. We chose the VI as it more closely represented our intuition of the neighborhood of a partition. For example, as noted in Wade and Ghahramani (2018), the Binder's loss seems to give more weight to partitions that differ from $\boldsymbol{c}_0$ because of a split rather than a merge, while the variation of information is more symmetric in this sense. However, we agree that it is worth investigating how the choice of distance impacts sampling efficiency.

Applications

We build our CP process starting from an epidemiological application to birth defects, and we are pleased that several discussants highlighted other relevant domains of application.

In particular, **Avalos-Pacheco, De Vito, and Wade** illustrate two interesting examples that can benefit from inclusion of prior information on the clustering: clinical trials and nutritional epidemiology. Some oncology clinical trials (e.g. basket trials) investigate drug effects across multiple cancer populations, divided into biomarker-based subgroups. Bayesian hierarchical models can be used to borrow information across biomarker and cancer subgroups. These groups are either estimated or prespecified, and relevant biological information is not accounted for. The other example from **Avalos-Pacheco, De Vito, and Wade** considers analysis of food intakes, where it is of interest to interpret results with respect to some dietary patterns. These patterns are typically estimated using dimension reduction techniques, ignoring available prior information, for example about food combinations related to some disease. On a related theme, **Casa, Fop, and Murphy** mention instead analysis of spectroscopy data representing food characteristics, and substantial information about grouping comes from knowledge of the chemical interactions. **Smith** brings an example from quantitative marketing, where partitions of products are typically motivated from managerial choices.

In considering wider applications of our method, there are still some open questions to investigate. For example, **Smith** mentions model assessment. In our application to birth defects data, we illustrated results from different models fit to the data with varying degrees of penalization parameter ψ, with extreme situations where the clustering behavior is governed by a Dirichlet process prior, or fixing the groups as $\boldsymbol{c}_0$. Model assessment can be done using standard techniques for Bayesian models, relying for example on Bayesian predictive checks (Gelman et al., 2013), or (calibrated) posterior predictive p-values (Meng et al., 1994; Hjort et al., 2006). Uncertainty about the clustering can be summarized via the recent proposal of Wade and Ghahramani (2018) using credible balls for the clustering estimate.

Finally, scalability with N remains a concern, although technical suggestions given in the discussions from **Dahl, Warr, and Jensen** and **Smith** look promising to address some of the computational issues. We agree with **Smith** that a practical approach is to impose more dogmatic prior assumptions, so that the partitioning problem lives in a lower dimensional space. Our application uses this concept already in that observations are collected at the individual level, and we used prior information for clustering group effects for individuals having the same birth defect. The use of more dogmatic assumptions seems a reasonable strategy, as it may be hard in practice to have substantial prior information on many many individuals.

References

Antoniano-Villalobos, I., Wade, S., and Walker, S. G. (2014). "A Bayesian nonparametric regression model with normalized weights: a study of hippocampal atrophy in Alzheimer's disease." *Journal of the American Statistical Association*, 109(506): 477–490. MR3223726. doi: https://doi.org/10.1080/01621459.2013.879061. 366

Binder, D. A. (1978). "Bayesian cluster analysis." *Biometrika*, 65(1): 31–38. URL https://doi.org/10.1093/biomet/65.1.31 MR0501592. doi: https://doi.org/10.1093/biomet/65.1.31. 368

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). "A general framework for updating belief distributions." *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 78(5): 1103. MR3557191. doi: https://doi.org/10.1111/rssb.12158. 365

Botto, L. D., Lin, A. E., Riehle-Colarusso, T., Malik, S., Correa, A., and Study, N. B. D. P. (2007). "Seeking causes: classifying and evaluating congenital hearth defects in etiologic studies." *Birth Defects Research Part A: Clinical and Molecular Teratology*, 79(10): 714–727. 365

Bouchard-Côté, A., Doucet, A., and Roth, A. (2017). "Particle Gibbs split-merge sampling for Bayesian inference in mixture models." *Journal of Machine Learning Research*, 18(28). MR3634895. 367

Dahl, D. (2003). "An improved merge-split sampler for conjugate Dirichlet process mixture models (Technical Report)." *University of Wisconsin*. MR2706330. 367

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press. MR3235677. 368

Hjort, N. L., Dahl, F. A., and Steinbakk, G. H. (2006). "Post-processing posterior predictive p values." *Journal of the American Statistical Association*, 101(475): 1157–1174. MR2324154. doi: https://doi.org/10.1198/016214505000001393. 368

Jain, S. and Neal, R. M. (2004). "A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model." *Journal of Computational and Graphical Statistics*, 13(1): 158–182. MR2044876. doi: https://doi.org/10.1198/1061860043001. 367

Jain, S. and Neal, R. M. (2007). "Splitting and merging components of a nonconjugate Dirichlet process mixture model." *Bayesian Analysis*, 2(3): 445–472. MR2342168. doi: https://doi.org/10.1214/07-BA219. 367

Meng, X.-L. et al. (1994). "Posterior predictive p-values." *The Annals of Statistics*, 22(3): 1142–1160. MR1311969. doi: https://doi.org/10.1214/aos/1176325622. 368

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9: 249–265. MR1823804. doi: https://doi.org/10.2307/1390653. 367

Smith, A. N. and Allenby, G. M. (2020). "Demand models with random partitions." *Journal of the American Statistical Association*, 115(529): 47–65. MR4078444. doi: https://doi.org/10.1080/01621459.2019.1604360. 367

Wade, S. and Ghahramani, Z. (2018). "Bayesian cluster analysis: point estimation and credible balls (with discussion)." *Bayesian Analysis*, 13(2): 559–626. URL https://doi.org/10.1214/17-BA1073 MR3807860. doi: https://doi.org/10.1214/17-BA1073. 368