The Annals of Applied Statistics
2023, Vol. 17, No. 1, 313–332
https://doi.org/10.1214/22-AOAS1628

MODEL SELECTION FOR MATERNAL HYPERTENSIVE DISORDERS WITH SYMMETRIC HIERARCHICAL DIRICHLET PROCESSES

BY BEATRICE FRANZOLINI[1,a], ANTONIO LIJOI[2,b] AND IGOR PRÜNSTER[2,c]

[1]*Singapore Institute for Clinical Sciences (SICS), Agency for Science, Technology and Research (A*STAR),* [a]*Beatrice_Franzolini@sics.a-star.edu.sg*

[2]*Department of Decision Sciences, Bocconi University,* [b]*antonio.lijoi@unibocconi.it,* [c]*igor.pruenster@unibocconi.it*

Hypertensive disorders of pregnancy occur in about 10% of pregnant women around the world. Though there is evidence that hypertension impacts maternal cardiac functions, the relation between hypertension and cardiac dysfunctions is only partially understood. The study of this relationship can be framed as a joint inferential problem on multiple populations, each corresponding to a different hypertensive disorder diagnosis, that combines multivariate information provided by a collection of cardiac function indexes. A Bayesian nonparametric approach seems particularly suited for this setup, and we demonstrate it on a dataset consisting of transthoracic echocardiography results of a cohort of Indian pregnant women. We are able to perform model selection, provide density estimates of cardiac function indexes and a latent clustering of patients: these readily interpretable inferential outputs allow to single out modified cardiac functions in hypertensive patients, compared to healthy subjects, and progressively increased alterations with the severity of the disorder. The analysis is based on a Bayesian nonparametric model that relies on a novel hierarchical structure, called symmetric hierarchical Dirichlet process. This is suitably designed so that the mean parameters are identified and used for model selection across populations, a penalization for multiplicity is enforced, and the presence of unobserved relevant factors is investigated through a latent clustering of subjects. Posterior inference relies on a suitable Markov chain Monte Carlo algorithm, and the model behaviour is also showcased on simulated data.

1. Introduction. Hypertensive disorders of pregnancy are a class of high blood pressure disorders that occur during the second half of pregnancy, which include gestational hypertension, preeclampsia and severe preeclampsia. They are characterized by a diastolic blood pressure higher than 90 mm Hg and/or a systolic blood pressure higher than 140 mm Hg, and they are often accompanied by proteinuria. These disorders affect about 10% of pregnant women around the world, with preeclampsia occurring in 2–8% of all pregnancies (Timokhina et al. (2019)). These disorders represent one of the leading causes of maternal and fetal morbidity and mortality, contributing to 7–8% of maternal death worldwide (Dolea and AbouZahr (2003), McClure et al. (2009), Shah et al. (2009)). The World Health Organization estimates that the incidence of preeclampsia is seven times higher in developing countries than in developed countries. However, the occurrence of these diseases appears underreported in low and middle income countries, implying that the true incidence is unknown (Igberase and Ebeigbe (2006), Malik, Jee and Gupta (2019)). While there is evidence that hypertensive disorders of pregnancy are related with the development of cardiac dysfunctions, both in the mother and in the child (Bellamy et al. (2007), Davis et al. (2012), Ambrožic et al. (2020), Garcia-Gonzalez et al. (2020), Aksu et al. (2021), deMartelly et al. (2021)), there is no common agreement on the relation between the severity of hypertension and cardiac dysfunction

Received May 2021; revised January 2022.

Key words and phrases. Bayesian nonparametrics, clustering populations, Dirichlet process, hierarchical partitions, hierarchical process, hypertensive disorders of pregnancy, model based clustering.

(Tatapudi and Pasumarthy (2017b)) and echocardiography is not included in baseline evaluation of hypertensive disorders of pregnancy. Further investigations on these disorders are needed, especially for developing countries, where women often give birth at a younger age with respect to developed countries.

The goal of this work is to detect which cardiac function is altered and under which hypertensive disorders by relying on a principled Bayesian nonparametric approach. An interesting case-control study to explore the relation between cardiac dysfunction and hypertensive disorders is provided by Tatapudi and Pasumarthy (2017a), where the measures of 10 different cardiac function indexes were recorded in four groups of pregnant women in India. Groups of women are characterized by different hypertensive disorder diagnoses that are naturally ordered based on their severity: healthy (C), gestational hypertension (G), mild preeclampsia (M) and severe preeclampsia (S). Hypertensive diagnoses are used as identifiers for what we call populations of patients and we refer to cardiac function indexes also with the term response variables. For each response variable we want to determine a partition of the four populations of patients. This amounts to identifying similarities between different hypertensive disorders with respect to each cardiac index. Supposing, for instance, that the selected partition assigns all the populations to the same cluster, one can conclude that no alteration is shown for the corresponding cardiac index across different hypertensive diseases.

Our goal of identifying a partition of the four patients' populations for each of the 10 responses can be rephrased as a problem of multiple model selection: we want to select the most plausible partition for each cardiac index. Frequentist hypothesis testing does not allow to deal with more than two populations in a straightforward way, and pairwise comparisons may lead to conflicting conclusions. Conversely, a Bayesian approach yields the posterior distribution on the space of partitions, which can be used for simultaneous comparisons. Moreover, the presence of $M = 10$ jointly tested cardiac indexes requires to perform model selection repeatedly 10 times. Once again, a Bayesian approach seems to be preferred, because, as observed for instance by Scott and Berger (2006), it does not require the introduction of a penalty term for multiple comparison, thanks to the prior distribution built-in penalty.

Here, we design a Bayesian nonparametric model that is tailored to deal with both a collection of ordered populations and the multivariate information of the response variables, while preserving the typical flexibility of nonparametric models and producing easily interpretable results. When applied to the dataset on transthoracic echocardiography results for a cohort of Indian pregnant women in Section 5, our model effectively identifies modified cardiac functions in hypertensive patients, compared to healthy subjects, and progressively increased alterations with the severity of the disorder, in addition to other more subtle findings. The observed data $X_{i,j,m}$ represent the measurement of the mth response variable (cardiac index) on the ith individual (pregnant woman) in the jth population (hypertensive disorder) and, as in standard univariate ANOVA models, they are assumed to be partially exchangeable across disorders. This means that, for every $m \in \{1, \ldots, M\}$, the law of $((X_{i,1,m})_{i\geq 1}, \ldots, (X_{i,J,m})_{i\geq 1})$ is invariant with respect to permutations within each sequence of random variables, namely, for any positive integers $n_1, \ldots, n_J$

$$\big((X_{i,1,m})_{i=1}^{n_1}, \ldots, (X_{i,J,m})_{i=1}^{n_J}\big) \stackrel{\mathrm{d}}{=} \big((X_{\sigma_1(i),1,m})_{i=1}^{n_1}, \ldots, (X_{\sigma_J(i),J,m})_{i=1}^{n_J}\big)$$

for all permutations σ_j of $(1, \ldots, n_j)$, with $j = 1, \ldots, J$. This is a natural generalization of exchangeability to tackle heterogeneous data and, by de Finetti's representation theorem, it amounts to assuming the existence of a collection of (possibly dependent) random probability measures $\{\pi_{j,m} : j = 1, \ldots, J m = 1, \ldots, M\}$ such that

$$X_{i,j,m} \mid \pi_{j,m} \stackrel{\text{iid}}{\sim} \pi_{j,m} \quad i = 1, \ldots, n_j.$$

Hence, for any two populations $j \neq j'$, homogeneity corresponds to $\pi_{j,m} = \pi_{j',m}$ (almost surely). However, a reliable assessment of this type of homogeneity is troublesome when having just few patients per diagnosis, as it happens in the mild preeclampsia subsample. Without relying on simplifying parametric assumptions, a small subsample size may not be sufficiently informative to infer equality of entire unknown distributions. To overcome this issue, without introducing parametric assumptions, we resort to an alternative weaker notion of homogeneity between populations j and j': we only require the conditional means of the two populations to (almost surely) coincide

$$\mathbb{E}(X_{i,j,m} \mid \pi_{j,m}) = \mathbb{E}(X_{i,j',m} \mid \pi_{j',m}). \tag{1}$$

According to this definition, the detection of heterogeneities in cardiac function indexes amounts to inferring which cardiac indexes have means that differ across diagnoses, as it is done in standard parametric ANOVA models. Besides clustering populations according to (1), it is also of interest to cluster patients, both within and across different groups, once the effect of the specific hypertensive disorder is taken into account. This task may be achieved by assuming a model that decomposes the observations as

$$X_{i,j,m} = \theta_{j,m} + \varepsilon_{i,j,m} \qquad \varepsilon_{i,j,m} | (\xi_{i,j,m}, \sigma^2_{i,j,m}) \overset{\text{ind}}{\sim} \mathrm{N}(\xi_{i,j,m}, \sigma^2_{i,j,m}), \tag{2}$$

and the $\xi_{i,j,m}$ have a symmetric distribution around the origin, in order to ensure $E(\xi_{i,j,m}) = 0$. In view of this decomposition, we will let $\theta_{j,m}$ govern the clustering of populations, while the $(\xi_{i,j,m}, \sigma^2_{i,j,m})$'s determine the clustering of individuals, namely patients, after removing the effect of the specific hypertensive disorder. In order to pursue this, for each cardiac index m, we will specify a hierarchical process prior for $(\xi_{i,j,m}, \sigma^2_{i,j,m})$ that is suited to infer the clustering structure both within and across different hypertensive disorders for a specific cardiac index. In particular, we will deploy a novel instance of hierarchical Dirichlet process, introduced in Teh et al. (2006), that we name *symmetric* to highlight its centering in 0.

Early examples of Bayesian nonparametric models for ANOVA can be found in Cifarelli and Regazzini (1978) and Muliere and Petrone (1993), while the first popular proposal, due to De Iorio et al. (2004), uses the dependent Dirichlet process (DDP) (MacEachern (2000)) and is, therefore, termed ANOVA-DDP. This model is mainly tailored to estimate populations' probability distributions, while we draw inferences over clusters of populations' means and obtain estimates of the unknown distributions as a by-product. Moreover, the ANOVA-DDP of De Iorio et al. (2004) was not introduced as a model selection procedure. A popular Bayesian nonparametric model that does cluster populations and can be used for model selection, is the nested Dirichlet process of Rodríguez, Dunson and Gelfand (2008). As shown in Camerlenghi et al. (2019a), such a prior is biased toward homogeneity, in the sense that even a single tie between populations j and j', namely, $X_{i,j,m} = X_{i',j',m}$ for some i and i', entails $\pi_{j,m} = \pi_{j'm}$ (almost surely). In order to overcome such a drawback, a novel class of nested and more flexible priors has been proposed in Camerlenghi et al. (2019a); see also Soriano and Ma (2017) for related work. Interesting alternatives that extend the analysis to more than two populations can be found in Christensen and Ma (2020), Lijoi, Prünster and Rebaudo (2022) and in Beraha, Guglielmi and Quintana (2021). Another similar proposal is the one by Gutiérrez et al. (2019), whose model identifies differences over cases' distributions and the control group. These models imply that two populations belong to the same cluster if they share the entire distribution. However, as already mentioned, distribution-based clustering is not ideal when dealing with scenarios as the one of hypertensive dataset. Further evidence will be provided in Section 5.1 through simulation studies. In addition, note that all these contributions deal with only one response variable and would need to be suitably generalized to fit the setup of this paper. As far as the contributions treating multiple response variables are concerned, uses of nonparametric priors for multiple testing can be found, for instance, in

Gopalan and Berry (1998), Do, Müller and Tang (2005), Dahl and Newton (2007), Guindani, Müller and Zhang (2009), Martin and Tokdar (2012) and, more recently, in Cipolli, Hanson and McLain (2016), who propose an approximate finite Pólya tree multiple testing procedure to compare two-samples' locations, and in Denti et al. (2021). However, in all these contributions, models are developed directly over summaries of the original data (e.g., averages, z-scores) and, as such, do not allow to draw any inference on the entire distributions and clusters of subjects.

The outline of the paper is as follows. In Section 2 we introduce the model which makes use of an original hierarchical prior structure for symmetric distributions (Section 2.2). In Section 3 we derive the prior law of the random partitions induced by the model, key ingredient for the Gibbs sampling scheme devised in Section 4. In Section 5 we first present a series of simulation studies that highlight the behaviour of the model before applying it to obtain our results on cardiac dysfunction in hypertensive disorders. Section 6 contains some concluding remarks. As Supplementary Material we provide the datasets and Python codes, some further background material and details about the derivation of the posterior sampling scheme as well as additional simulation studies and results on the application, including an analysis of prior sensitivity.

2. The Bayesian nonparametric model. The use of discrete nonparametric priors for Bayesian model-based clustering has become standard practice. The Dirichlet process (DP) (Ferguson (1973)) is the most popular instance, and clustering is typically addressed by resorting to a mixture model which, with our data structure, amounts to

$$X_{i,j,m}|\psi_{i,j,m} \overset{\text{ind}}{\sim} k(X_{i,j,m};\psi_{i,j,m}), \qquad \psi_{i,j,m}|\tilde{p}_{j,m} \overset{\text{ind}}{\sim} \tilde{p}_{j,m}$$

for $m = 1,\dots,M$, $j = 1,\dots,J$ and $i = 1,\dots,n_j$. Here, $k(\cdot;\cdot)$ is some kernel, and the $\tilde{p}_{j,m}$'s are discrete random probability measures. Hence, the $\psi_{i,j,m}$'s may exhibit ties. The model specification for $\tilde{p}_{j,m}$ will be tailored to address the following goals: (i) cluster the J probability distributions based on their means; (ii) cluster the observations $X_{i,j,m}$ according to the ties induced on the $\psi_{i,j,m}$'s by the $\tilde{p}_{j,m}$'s for a given fixed j and across different j's. These two issues will be targeted separately: we first design a clustering scheme for the populations through the specification of a prior on the means of the $X_{i,j,m}$'s, and, then, we cluster the data using a hierarchical DP having a specific invariance structure that is ideally suited to the application at hand.

2.1. *The prior on disease-specific locations.* As a model for the observations, we consider a nonparametric mixture of Gaussian distributions specified as

$$X_{i,j,m}|(\boldsymbol{\theta}_m,\boldsymbol{\xi}_m,\boldsymbol{\sigma}^2_m) \overset{\text{ind}}{\sim} \mathcal{N}(\theta_{j,m}+\xi_{i,j,m},\sigma^2_{i,j,m}), \tag{3}$$

where $\boldsymbol{\theta}_m = (\theta_{1,m},\dots,\theta_{J,m})$, $\boldsymbol{\xi}_m = (\xi_{1,1,m},\dots,\xi_{1,n_1,m},\xi_{2,1,m},\dots,\xi_{n_J,J,m})$, with a similar definition for the vector $\boldsymbol{\sigma}^2_m$, and $\mathcal{N}(\mu,\sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. The assumption in (3) clearly reflects (2). Moreover, in order to account for the two levels of clustering we are interested in, we will assume that

$$(\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_M) \sim P, \qquad (\xi_{i,j,m},\sigma^2_{i,j,m})|\tilde{q}_{j,m} \overset{\text{iid}}{\sim} \tilde{q}_{j,m} \quad (i = 1,\dots,n_j), \tag{4}$$

where $\tilde{q}_{1,m},\dots,\tilde{q}_{J,m}$ are discrete random probability measures independent from $(\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_M)$. Thus, the likelihood corresponds to

$$\prod_{m=1}^{M}\prod_{j=1}^{J}\prod_{i=1}^{n_j}\frac{1}{\sigma_{i,j,m}}\varphi\Big(\frac{x_{i,j,m}-\theta_{j,m}-\xi_{i,j,m}}{\sigma_{i,j,m}}\Big)\tilde{q}_{j,m}(\mathrm{d}\xi_{i,j,m},\mathrm{d}\sigma_{i,j,m}) \tag{5}$$

with φ denoting the standard Gaussian density. Relevant inferences can be carried out if one is able to marginalize this expression with respect to both $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ and $(\tilde{q}_{1,m}, \ldots, \tilde{q}_{J,m})$ for each $m = 1, \ldots, M$.

This specification allows to address the model selection problem in the following way. If $\mathcal{M}^m$ stands for the space of all partitions of the J populations for the mth cardiac function index, then $\mathcal{M}^m = \{M_b^m : b = 1, \ldots, \text{card}(\mathcal{P}_J)\}$ where $\mathcal{P}_J$ is the collection of all possible partitions of $[J] = \{1, \ldots, J\}$. In our specific case, $J = 4$ and $\text{card}(\mathcal{P}_J) = 15$; thus, we have 15 competing models per cardiac index. Each competing model corresponds to a specific partition in $\mathcal{M}^m$. In particular, the partition arises from ties between the population specific means in $\boldsymbol{\theta}_m$, and hence, the distribution P in (4) needs to associate positive probabilities to ties between the parameters within the vector $\boldsymbol{\theta}_m$, for each $m = 1, \ldots, M$.

Let us start considering as distribution P a well-known effective clustering prior, that is, a mixture of DPs in the spirit of Antoniak (1974), namely,

$$\begin{aligned} \theta_{j,m} \mid \tilde{p}_m &\overset{\text{iid}}{\sim} \tilde{p}_m \quad j = 1, \ldots, J, \\ \tilde{p}_m \mid \omega &\overset{\text{iid}}{\sim} \text{DP}(\omega, G_m) \quad m = 1, \ldots, M, \\ \omega &\sim p_\omega, \end{aligned} \tag{6}$$

where $\text{DP}(\omega, G_m)$ denotes the DP with concentration parameter ω and nonatomic baseline probability measure G_m and p_ω is a probability measure on $\mathbb{R}^+$. The discreteness of the DP implies the presence (with positive probability) of ties within the vector of locations $\boldsymbol{\theta}_m$ associated to a certain cardiac index m, as desired. The ties give rise to a random partition: as shown in Antoniak (1974), the probability of observing a specific partition of the elements in $\boldsymbol{\theta}_m$ consisting of $k \leq J$ distinct values with respective frequencies $n_1, \ldots, n_k$ coincides with

$$\Pi_k^{(J)}(n_1, \ldots, n_k) = \frac{\omega^k}{(\omega)_J} \prod_{i=1}^{k} (n_i - 1)!, \tag{7}$$

where $(\omega)_J = \Gamma(\omega + J)/\Gamma(\omega)$. The use of a shared concentration parameter over (7) to address multiple model selection has been already successfully employed in Moser, Rodríguez and Lofland (2021), where they cluster parameters in a probit model. When there is no pre-experimental information available on competing partitions, the use of (7) as a prior for model selection has some relevant benefits. Indeed, it induces borrowing of strength across diagnoses, and, being ω random, it generates borrowing of information also across cardiac indexes, thus improving the Bayesian learning mechanism. These two features can also be given a frequentist interpretation in terms of desirable penalties. As a matter of fact, the procedure penalizes for the multiplicity of the model selections that are performed. The penalty has to be meant in the following way: while J and/or M increase, the prior odds change in favor of less complex models. For more details on this, see Scott and Berger (2010). Summing up, the mixture of DPs automatically induces a prior distribution on $\{\mathcal{M}^m : m = 1, \ldots, M\}$ that arises from (7) combined with the prior p_ω on ω, and it presents desirable properties for model selection that can be interpreted either in terms of borrowing of information or in terms of penalties.

However, in the analysis of hypertensive disorders, some prior information on competing models is available, and this is not yet incorporated in (7). In fact, as already mentioned, there is a natural order of the diagnoses, which is given by the severity of the disorders, that is, C, G, M, S. Partitions that do not comply with this ordering, for example, $\{\{C, S\}\{G\}, \{M\}\}$, should be excluded from the support of the prior. Thus, we consider a prior over $\mathcal{M}^m$ that

associates zero probability to partitions that do not respect the natural order of the diagnoses and a probability proportional to that in (7) for the remaining partitions, that is,

$$\mathbb{P}\big(M_b^m \mid \omega\big) \propto \begin{cases} \Pi_k^{(J)}(n_1, \ldots, n_k) & \text{if } M_b^m \text{ is compatible with the natural order,} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

This amounts to a distribution P for $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$, given by

$$\begin{aligned} (\theta_{1,m}, \ldots, \theta_{J,m}) \mid \omega &\overset{\text{ind}}{\sim} P_{\omega, G_m} \quad m = 1, \ldots, M, \\ \omega &\sim p_\omega, \end{aligned} \tag{9}$$

where P_{ω,G_m} is the distribution obtained sampling a partition, according to (8), and associating to each cluster a unique value sampled from G_m. Using (9) as a prior for the disease-specific locations, we preserve the desirable properties of the mixture of DPs mentioned before, while incorporating prior information on the severity of the diseases.

As detailed in the next section, we further define random probability measures $\tilde{q}_{j,m}$ that satisfy the symmetry condition

$$\tilde{q}_{j,m}(A \times B) = \tilde{q}_{j,m}\big((-A) \times B\big) \quad \text{a.s.} \tag{10}$$

for any A and B. This condition ensures that the parameters $\theta_{j,m}$, for $j = 1, \ldots, J$ and $m = 1, \ldots, M$, in (3) are identified, namely, $\mathbb{E}(X_{i,j,m} \mid \boldsymbol{\theta}_m, \tilde{q}_{j,m}) = \theta_{j,m}$ with probability one. This identifiability property is crucial to make inference over the location parameters $\boldsymbol{\theta}_m$'s. Similar model specifications for discrete exchangeable data have been proposed and studied in Dalal (1979b), Doss (1984), Diaconis and Freedman (1986) and Ghosal, Ghosh and Ramamoorthi (1999) of which (5) represents a generalization to density functions and partially exchangeable data.

2.2. *The prior for the error terms.* While the clustering of populations is governed by (8), we use a mixture of hierarchical discrete processes for the error terms. This has the advantage of modeling the clustering of the observations, both within and across different samples, once the disease-specific effects are accounted for. This clustering structure allows to model heterogeneity across patients in a much more realistic way with respect to standard ANOVA models based on assumption of normality. Cardiac indexes may be influenced by a number of factors that are not directly observed in the study, such as preexisting conditions (Hall, George and Granger (2018)) and psychosocial factors (Pedersen et al. (2017)). These unobserved relevant factors may be shared across patients with the same or a different diagnosis and may also result in outliers. To take into account this latent heterogeneity of the data, we introduce the hierarchical symmetric DP that satisfies the symmetry condition in (10) and, moreover, allows to model heterogeneous data similarly to the hugely popular hierarchical DP (Teh et al. (2006)).

The basic building block of the proposed prior is the invariant Dirichlet process which was introduced for a single population ($J = 1$) in an exchangeable framework by Dalal (1979a). Such a modification of the DP satisfies a symmetry condition, in the sense that it is a random probability measure that is invariant with respect to a chosen group of transformations $\mathcal{G}$. A more formal definition and detailed description of the invariant DP can be found in Section A of the Supplementary Material (Franzolini, Lijoi and Prünster (2023)). For our purposes it is enough to consider the specific case of the symmetric Dirichlet process, which can be constructed through a symmetrization of a Dirichlet process. Consider a nonatomic probability measure P_0 on $\mathbb{R}$, and let $\tilde{Q}_0 \sim \text{DP}(\alpha, P_0)$. If

$$\tilde{Q}(A) = \frac{\tilde{Q}_0(A) + \tilde{Q}_0(-A)}{2} \quad \forall A \in \mathcal{B}(\mathbb{R}), \tag{11}$$

where $-A = \{x \in \mathbb{R} : -x \in A\}$, then $\tilde{Q}$ is symmetric about 0 (almost surely) and termed symmetric DP, in symbols $\tilde{Q} \sim \text{s-DP}(\alpha, P_0)$. For convenience and without loss of generality, we assume that P_0 is symmetric: this implies that P_0 is the expected value of $\tilde{Q}$ making it an interpretable parameter. The random probability measure $\tilde{Q}$ is the basic building block of the hierarchical process that we use to model the heterogeneity of the error terms across different populations, $j = 1, \ldots, J$, in such a way that clusters identified by the unique values can be shared within and across populations. This prior is termed *symmetric hierarchical Dirichlet process* (s-HDP) and described as

$$\begin{aligned} \tilde{q}_{j,m} \mid \gamma_{j,m}, \tilde{q}_{0,m} &\overset{\text{ind}}{\sim} \text{s-DP}(\gamma_{j,m}, \tilde{q}_{0,m}) \\ \tilde{q}_{0,m} \mid \alpha_m &\overset{\text{ind}}{\sim} \text{s-DP}(\alpha_m, P_{0,m}) \end{aligned}, \tag{12}$$

where $\gamma_{j,m}$ and α_m are positive parameters and $P_{0,m}$ is a nonatomic probability distribution symmetric about 0. We use the notation $(\tilde{q}_{1,m}, \ldots, \tilde{q}_{J,m}) \sim \text{s-HDP}(\boldsymbol{\gamma}_m, \alpha_m, P_{0,m})$, where $\boldsymbol{\gamma}_m = (\gamma_{1,m}, \ldots, \gamma_{j,m})$. This definition clearly ensures the validity of (10). A graphical model representation of the overall proposed model is displayed in Figure 1.

Still referring to the decomposition of the observations into disease-specific locations and an error term, that is, $X_{i,j,m} = \theta_{j,m} + \varepsilon_{i,j,m}$, it turns out that the $\varepsilon_{i,j,m}$'s are from a symmetric hierarchical DP mixture (s-HDP mixture) with a normal kernel. Hence, the patients' clusters are identified through the $\varepsilon_{i,j,m}$, which, according to (3), are conditionally independent from a $\mathcal{N}(\xi_{i,j,m}, \sigma^2_{i,j,m})$, given $(\xi_{i,j,m}, \sigma^2_{i,j,m})$. The choice of the specific invariant DP is aimed at ensuring that $\mathbb{E}(\varepsilon_{i,j,m} | \tilde{q}_{j,m}) = 0$. The clusters identified by the s-HDP mixture can be interpreted as representing common unobserved factors across patients, once the disease-specific locations have been accounted for. Indeed, for any pair of patients, we may consider the decomposition $X_{i,j,m} - X_{i',j',m} = \Delta^{(m)}_\theta + \Delta^{(m)}_\xi + (e_{i,j,m} - e_{i',j',m})$, where $\Delta^{(m)}_\theta = \theta_{j,m} - \theta_{j',m}$, $\Delta^{(m)}_\xi = \xi_{i,j,m} - \xi_{i',j',m}$ and $e_{i,j,m}$ and $e_{i',j',m}$ are independent and normally distributed random variables with zero mean and variances $\sigma^2_{i,j,m}$ and $\sigma^2_{i',j',m}$, respectively.

Hence, patients' clustering reflects the residual heterogeneity that is not captured by the disease-specific component $\Delta^{(m)}_\theta$ and are related to the subject-specific locations $\Delta^{(m)}_\xi$ and

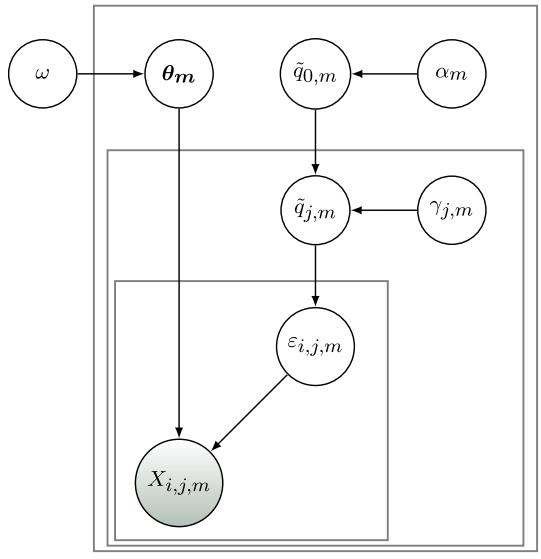


FIG. 1. *Graphical representation of the model. Each node represents a random variable and each rectangle denotes conditional i.i.d. replications of the model within the rectangle.*

to the zero-mean error component $(e_{i,j,m} - e_{i',j',m})$. In view of this interpretation, using a s-HDP mixture over error terms offers a three-fold advantage. First, the presence of clearly separated clusters of patients within and across populations will indicate the presence of unobserved relevant factors which affect the cardiac response variables. Second, single patients with very low probabilities of co-clustering with all other subjects will have to be interpreted as outliers. Finally, the estimated clustering structure can also be used to check whether the relative effect of a certain disease (with respect to another) is fully explained by the corresponding $\Delta_\theta^{(m)}$. To clarify this last point consider two diseases: if the posterior co-clustering probabilities among patients sharing the same disease are different between the two populations, this will indicate that different diagnoses not only have an influence on disease-specific locations (which is measured by $\Delta_\theta^{(m)}$), but they also have an impact on the shape of the distribution of the corresponding cardiac index. More details on this can be found in Section D of the Supplementary Material (Franzolini, Lijoi and Prünster (2023)).

3. Marginal distributions and random partitions. As emphasized in the previous sections, ties among the $\theta_{j,m}$'s and the $(\xi_{i,j,m}, \sigma^2_{i,j,m})$'s are relevant for inferring the clustering structure both among the populations (hypertensive diseases) and among the individual units (patients). Indeed, for each m (cardiac index) they induce a random partition that emerges as a composition of two partitions generated, respectively, by the prior in (9) and the s-HDP. The laws of these random partitions are not only crucial to understand the clustering mechanism but also necessary in order to derive posterior sampling schemes. In this section such a law is derived and used to compute the predictive distributions that, jointly with the likelihood, determine the full conditionals of the Gibbs sampler devised in Section 4. To reduce the notational burden, in this and the following section, we remove the dependence of observations and parameters on the specific response variable m and denote with $\phi_{i,j}$ the pair $(\xi_{i,j}, \sigma^2_{i,j})$ and with $\boldsymbol{\phi}$ the collection $(\phi_{1,1}, \ldots, \phi_{1,n_1}, \phi_{2,1}, \ldots \phi_{n_J,J})$.

Conditionally on ω, the law of the partition in (8) leads to the following predictive distribution for the disease-specific locations:

$$\theta_j | \omega, \theta_1, \ldots, \theta_{j-1} \sim a_j(\omega, \theta_1, \ldots, \theta_{j-1}) \delta_{\theta_{j-1}} + [1 - a_j(\omega, \theta_1, \ldots, \theta_{j-1})] G,$$

where

$$a_j(\omega, \theta_1, \ldots, \theta_{j-1}) = \frac{\sum_{(*_j)} \Pi_k^{(J)}(n_1, \ldots, n_k)}{\sum_{(\Delta_j)} \Pi_k^{(J)}(n_1, \ldots, n_k)}, \tag{13}$$

where the sum at the denominator runs over the set of partitions consistent with the one generated by $(\theta_1, \ldots, \theta_{j-1})$ and the one at the numerator runs over a subset of those partitions where one further has $\theta_j = \theta_{j-1}$. For $j = 4$, the predictive equals

$$\theta_4 \mid \omega, \theta_1, \theta_2, \theta_3 \begin{cases} \dfrac{3}{\omega+3}\delta_{\theta_3} + \dfrac{\omega}{\omega+3} G & \text{if } \theta_1 = \theta_2 = \theta_3, \\ \dfrac{2}{\omega+2}\delta_{\theta_3} + \dfrac{\omega}{\omega+2} G & \text{if } \theta_1 \neq \theta_2 = \theta_3, \\ \dfrac{1}{\omega+1}\delta_{\theta_3} + \dfrac{\omega}{\omega+1} G & \text{otherwise.} \end{cases}$$

Explicit expressions for the function a, for $j = 1, 2, 3$, can be easily computed, using (13) and (8), and are provided in Section B of the Supplementary Material (Franzolini, Lijoi and Prünster (2023)).

Moving to second-level partitions induced by the s-HDP, we recall that the key concept for studying random partitions on multisample data is the *partially exchangeable partition probability function* (pEPPF); see, for example, Lijoi, Nipoti and Prünster (2014) and Camerlenghi

et al. (2019b). The pEPPF returns the probability of a specific multisample partition and represents the appropriate generalization of the well-known single-sample EPPF, which in the DP case corresponds to (7). Discreteness of the s-HDP $(\tilde{q}_1, \ldots, \tilde{q}_m)$ in (12) induces a partition of the elements of $\boldsymbol{\phi}$ into equivalence classes identified by the distinct values. Taking into account the underlying partially exchangeable structure, such a random partition is characterized by the pEPPF

$$\tilde{\Pi}_k^{(N)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J) = \mathbb{E}\left(\int_{\Phi^k} \prod_{j=1}^{J} \prod_{h=1}^{k} \tilde{q}_{j,m}^{n_{j,h}}(\mathrm{d}\phi_i)\right), \tag{14}$$

where $\boldsymbol{n}_j = (n_{j,1}, \ldots, n_{j,k})$ are nonnegative integers, for any $j = 1, \ldots, J$, such that $n_{j,h}$ is the number of elements in $\boldsymbol{\phi}$ corresponding to population j and belonging to cluster h. Thus, $\sum_{j=1}^{J} n_{j,h} \geq 1$ for any $h = 1, \ldots, k$, $\sum_{h=1}^{k} n_{j,h} = n_j$ and $\sum_{h=1}^{k} \sum_{j=1}^{J} n_{j,h} = N$. The determination of probability distributions of this type is challenging, and, only recently, the first explicit instances have appeared in the literature; see, for example, Lijoi, Nipoti and Prünster (2014), Camerlenghi et al. (2019a) and Camerlenghi et al. (2019b). With respect to the hierarchical case considered in Camerlenghi et al. (2019b), the main difference is that here we have to take into account the specific structure (11) of the $\tilde{q}_{j,m}$. The almost sure symmetry of the process generates a natural random matching between sets in the induced partition. Therefore, instead of studying the marginal law in (14), we derive the joint law of the partition and of the random matching. Formally, consider a specific partition $\{A_1^+, A_1^-, \ldots, A_k^+, A_k^-\}$ of $\boldsymbol{\phi}$, such that, for $h = 1, \ldots, k$, all the elements in A_h^+ belong to $\mathbb{R}^+ \times \mathbb{R}^+$, all the elements in A_h^- belong to $\mathbb{R}^- \times \mathbb{R}^+$ and, if $\phi_{i,j} \in A_h^+$ and $\phi_{i',j'} \in A_h^-$, then the elementwise absolute values of $\phi_{i,j}$ and $\phi_{i',j'}$ are equal. Denote with $n_{j,h}^+$ the number of elements in $A_h^+ \cap \{\phi_{i,j}, i = 1, \ldots, n_j\}$ and with $n_{j,h}^-$ the number of elements in $A_h^- \cap \{\phi_{i,j}, i = 1, \ldots, n_j\}$. The probability of observing $\{A_1^+, A_1^-, \ldots, A_k^+, A_k^-\}$ is

$$\tilde{\tilde{\Pi}}_k^{(N)}(\boldsymbol{n_1}^+, \boldsymbol{n_1}^-, \ldots, \boldsymbol{n_J}^+, \boldsymbol{n_J}^-) = \mathbb{E}\left(\int_{\Phi^k} \prod_{j=1}^{J} \prod_{h=1}^{k} \tilde{q}_{j,m}{}^{n_{j,h}^+ + n_{j,h}^-}(d\phi)\right) \tag{15}$$

with $\boldsymbol{n_j^+} = (n_{j,1}^+, \ldots, n_{j,k}^+)$. As for the determination of (15), a more intuitive understanding may be gained if one considers its corresponding Chinese restaurant franchise (CRF) metaphor which displays a variation of both the standard Chinese restaurant franchise of Teh et al. (2006) and the skewed Chinese restaurant process of Iglesias, Orellana and Quintana (2009). Figure 2 provides a graphical representation. The scheme is as follows: there are J restaurants sharing the same menu, and the customers are identified by their choice of $\phi_{i,j}$, but, unlike in the usual CRF, at each table two *symmetric dishes* are served. Denote with $\phi_{t,j}^* = (\xi_{t,j}^*, \sigma_{t,j}^{2*})$ and $-\phi_{t,j}^* = (-\xi_{t,j}^*, \sigma_{t,j}^{2*})$ the two dishes served at table t in restaurant j,

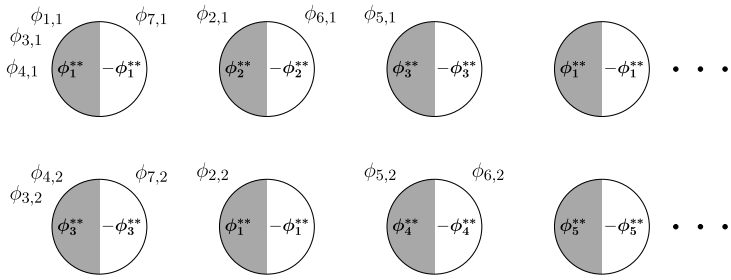


FIG. 2. *Chinese restaurant franchise representation of the symmetric hierarchical DP for $J = 2$ populations. Each circle represents a table.*

with $\phi_h^{**} = (\xi_h^{**}, \sigma_h^{**2})$ and $-\phi_h^{**} = (-\xi_h, \sigma_h^{**2})$ the hth pair of dishes in the menu and with $n_{j,h}^+$ and $n_{j,h}^-$ the number of customers in restaurant j eating dish ϕ_h^{**} and $-\phi_h^{**}$, respectively. This means that two options are available to a customer entering restaurant j: she/he will either sit at an already occupied table, with probability proportional to the number of customers at that table or will sit at a new table with probability proportional to the concentration parameter γ_j. In the former case, the customer will choose the dish $\phi_{t,j}^*$ with probability $1/2$ and $-\phi_{t,j}^*$ otherwise. In the latter case the customer will eat a dish served at another table of the franchise with probability proportional to half the number of tables that serve that dish or will make a new order with probability proportional to the concentration parameter α. In view of this scheme, the probability in (15) turns out to be

$$\tilde{\bar{\Pi}}_k^{(N)}(\boldsymbol{n}_1^+, \ldots, \boldsymbol{n}_J^-) = 2^{-N} \bar{\Pi}_k^{(N)}(\boldsymbol{n}_1^+ + \boldsymbol{n}_1^-, \ldots, \boldsymbol{n}_J^+ + \boldsymbol{n}_J^-),$$

and $\bar{\Pi}_k^{(N)}$ on the right-hand side is the pEPPF of the hierarchical DP, derived in Camerlenghi et al. (2019b), namely,

$$\bar{\Pi}_k^{(N)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_k) = \left(\prod_{j=1}^J \frac{\prod_{i=1}^k (\gamma_j)_{n_{j,h}}}{(\gamma_j)_{n_j}} \right) \sum_{\boldsymbol{\ell}} \frac{\alpha^k}{(\alpha)_{|\boldsymbol{\ell}|}} \prod_{h=1}^k (\ell_{\bullet,h} - 1)! \prod_{j=1}^J P(K_{n_{j,h}} = \ell_{j,h}),$$

where each sums runs over all $\ell_{j,h}$ in $\{1, \ldots, n_{j,h}\}$, if $n_{j,h} \geq 1$ and equals 1 if $n_{j,h} = 0$, whereas $\ell_{\bullet,h} = \sum_{j=1}^J \ell_{j,h}$ and $|\boldsymbol{\ell}| = \sum_{j=1}^J \sum_{h=1}^k \ell_{j,h}$. Note that the latent variable $\ell_{j,h}$ is the number of tables in restaurant j serving the hth pair of dishes. Moreover, $K_{n_{j,h}}$ is a random variable denoting the number of distinct clusters, out of $n_{j,h}$ observations generated by a DP with parameter γ_j and diffuse baseline P_0, and it is well known that

$$\mathbb{P}(K_{n_{j,h}} = \ell_{j,h}) = \frac{\gamma_j^{\ell_{j,h}}}{(\gamma_j)_{n_{j,h}}} |\mathfrak{s}(n_{j,h}, \ell_{j,h})|,$$

where $|\mathfrak{s}(n_{j,h}, \ell_{j,h})|$ is the signless Stirling number of the first kind. In view of this, one can deduce the predictive distribution

$$\begin{aligned}\mathbb{P}(\phi_{n_j+1,j} \in \cdot \,|\boldsymbol{\phi}) = {} & \frac{\gamma_j}{i - 1 + \gamma_j} \sum_{\boldsymbol{\ell}} \frac{\alpha}{|\boldsymbol{\ell}| + \alpha} \pi(\boldsymbol{\ell}|\boldsymbol{\phi}) P_0(\cdot) \\ & + \sum_{h=1}^k \left[\frac{n_{j,h}^+ + n_{j,h}^-}{n_j + \gamma_j} + \frac{\gamma_j}{n_j + \gamma_j} \sum_{\boldsymbol{\ell}} \frac{\ell_{\bullet,h}}{|\boldsymbol{\ell}| + \alpha} \pi(\boldsymbol{\ell}|\boldsymbol{\phi}) \right] \\ & \times \left(\frac{\delta_{\phi_h^{**}}(\cdot) + \delta_{-\phi_h^{**}}(\cdot)}{2} \right),\end{aligned}$$

where

$$\pi(\boldsymbol{\ell}|\boldsymbol{\phi}) \propto \frac{\alpha^k}{(\alpha)_{|\boldsymbol{\ell}|}} \prod_{h=1}^k (\ell_{\bullet,h} - 1)! \prod_{j=1}^J \frac{\gamma_j^{\ell_{j,h}}}{(\gamma_j)_{n_{j,h}^+ + n_{j,h}^-}} |\mathfrak{s}(n_{j,h}^+ + n_{j,h}^-, \ell_{j,h})| \mathbb{1}_{\{1,\ldots,n_{j,h}^+ + n_{j,h}^-\}}(\ell_{j,h})$$

is the posterior distribution of the latent variables $\ell_{j,h}$'s and $\mathbb{1}_A$ is the indicator function of set A.

4. Posterior inference. The findings of the previous section are the key ingredients to perform posterior inference with a marginal Gibbs sampler. The output of the sampler is structured into three levels: the first produces posterior probabilities on partitions of disease-specific locations; the second generates density estimates; the third provides clusters of patients. For notational simplicity we omit the dependence on m, except for the description of

the sampling step that generates ω. Recall that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ and $\boldsymbol{\phi} = \{(\phi_{1,j}, \ldots, \phi_{n_j,j}) : j = 1, \ldots, J\}$, with $\phi_{i,j} = (\xi_{i,j}, \sigma^2_{i,j})$. The target distribution of the sampler is the joint distribution of $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ and ω, conditionally on the observed data $\boldsymbol{X}$.

Sampling ϕ. In view of the CRF representation of the s-HDP, $t_{i,j}$ stands for the label of the table where the ith customer in restaurant j sits and $h_{t,j}$ for the dish label served at table t in restaurant j and with $\boldsymbol{t}$ and $\boldsymbol{h}$ we denote the corresponding arrays. Moreover, define the assignment variable $s_{i,j} = \mathbb{1}(\phi_{i,j} = \phi^*_{t_{i,j},j}) - \mathbb{1}(\phi_{i,j} = -\phi^*_{t_{i,j},j})$ and $\boldsymbol{s}$ is the corresponding arrays. In order to generate $\boldsymbol{\phi}$, we need to sample:

(i) $(t_{i,j}, s_{i,j})$ for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$;
(ii) $h_{t,j}$ for $t \in \boldsymbol{t}$ and $j = 1, \ldots, J$;
(ii) ϕ^{**}_h for $h \in \boldsymbol{h}$.

Note that, using the latent allocation indicators in $\boldsymbol{t}$ and $\boldsymbol{h}$, the sampling scheme is more efficient than sampling directly from the full conditional of each $\phi_{i,j}$, since the algorithm can update more than one parameter simultaneously (Neal (2000)). Define $\varepsilon_{i,j} = X_{i,j} - \theta_j$, and denote with $h(\varepsilon_{i,j,m}|\phi^*)$ the conditional normal density of $\varepsilon_{i,j}$, given $\phi^* = (\xi^*, \sigma^{2*})$, while the marginal density is

$$\bar{h}(\varepsilon_{i,j}) = \int h(\varepsilon_{i,j}|\phi) P_0(d\phi).$$

To sample $(t_{i,j}, s_{i,j})$ from their joint full conditional, we first sample $t_{i,j}$ from

$$P\big(t_{i,j} = t \mid \boldsymbol{t}^{-(i,j)}, \boldsymbol{h}^{-(i,j)}, \boldsymbol{\phi}^{*-(i,j)}, \boldsymbol{\phi}^{**-(i,j)}, \varepsilon_{i,j}\big)$$
$$\propto \begin{cases} n^{-(i,j)}_{t,j}\, p_{\text{old}}\big(\varepsilon_{i,j}|\phi^*_{t,j}\big) & \text{if } t \in \boldsymbol{t}^{-(i,j)}, \\ \gamma_j\, p_{\text{new}}\big(\varepsilon_{i,j}|\boldsymbol{\phi}^{**-(i,j)}\big) & \text{if } t = t^{\text{new}}, \end{cases}$$

where $\boldsymbol{t}^{-(i,j)}$, $\boldsymbol{h}^{-(i,j)}$ $\boldsymbol{\phi}^{*-(i,j)}$, $\boldsymbol{\phi}^{**-(i,j)}$ coincide with the arrays $\boldsymbol{t}$, $\boldsymbol{h}$ $\boldsymbol{\phi}^*$, $\boldsymbol{\phi}^{**}$ after having removed the entries corresponding to the ith customer in restaurant j. Moreover,

$$p_{\text{old}}\big(\varepsilon_{i,j}|\phi^*_{t,j}\big) = \frac{1}{2} h\big(\varepsilon_{i,j}|\phi^*_{t,j}\big) + \frac{1}{2} h\big(\varepsilon_{i,j}| - \phi^*_{t,j}\big),$$

and

$$p_{\text{new}}\big(\varepsilon_{i,j}|\boldsymbol{\phi}^{**-(i,j)}\big) = \sum_{h=1}^{k^{-(i,j)}} \frac{\ell_{\bullet,h}}{|\boldsymbol{\ell}| + \alpha} \left\{ \frac{1}{2} h\big(\varepsilon_{i,j}|\phi^{**}_h\big) + \frac{1}{2} h\big(\varepsilon_{i,j}| - \phi^{**}_h\big) \right\} + \frac{\alpha}{|\boldsymbol{\ell}| + \alpha} \bar{h}(\varepsilon_{i,j}).$$

Then, we sample $s_{i,j}$ from its full conditional

$$p\big(s_{i,j} = s \mid \boldsymbol{\phi}^*, t_{i,j}, \epsilon_{i,j}\big) \propto \begin{cases} h\big(\varepsilon_{i,j}|\phi^*_{t_{i,j}}\big) & \text{if } s = 1, \\ h\big(\varepsilon_{i,j}| - \phi^*_{t_{i,j}}\big) & \text{if } s = -1. \end{cases}$$

The conditional distribution of $h_{t,j}$ is

$$p\big(h_{t,j} = h \mid \boldsymbol{t}, \boldsymbol{h}^{-(t,j)}, \boldsymbol{\phi}^{**-(t,j)}, \boldsymbol{s}, \boldsymbol{\varepsilon}\big)$$
$$\propto \begin{cases} \ell^{-(t,j)}_{\bullet,h} \displaystyle\prod_{\{(i,j):t_{i,j}=t\}} h(s_{i,j}\varepsilon_{i,j}|\phi_h) & \text{if } h \in \boldsymbol{h}^{-(t,j)}, \\ \alpha \displaystyle\int \prod_{\{(i,j):t_{i,j}=t\}} h(s_{i,j}\varepsilon_{i,j}|\phi) P_0(d\phi) & \text{if } h = h^{\text{new}}. \end{cases}$$

Finally, when P_0 is conjugate with respect to the Gaussian kernel, the full conditional distribution of ϕ^{**}_h is obtained in closed form as posterior distribution of a Gaussian model, using as observations the collection $\{(s_{i,j}\epsilon_{i,j}) : h_{t_{i,j},j} = h\}$.

Sampling θ. When sampling the disease-specific location parameters, one can rely on a Chinese restaurant process restricted to those partitions that are consistent with the ordering of the diseases. Thus, in order to generate $\boldsymbol{\theta}$, we first sample the labels $\boldsymbol{t}_\theta = \{t_1, \ldots, t_J\}$, where t_j is the label of the table where the jth customer sits. Then, we sample the dish θ_t^* associated to table t for all $t \in \boldsymbol{t}_\theta$. If $z_{i,j} = X_{i,j} - \xi_{i,j}$, the conditional density of $\boldsymbol{z}_j = (z_{1,j}, \ldots, z_{n_j,j})$ associated to the location parameter θ^*, given $\boldsymbol{\sigma}_j = (\sigma_{1,j}, \ldots, \sigma_{n_j,j})$, is

$$f_{\theta^*}(\boldsymbol{z}_j|\boldsymbol{\sigma}_j) = \frac{1}{\sqrt{2\pi}\prod_{i=1}^{n_j}\sigma_{i,j}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n_j}\frac{(z_{i,j}-\theta^*)^2}{\sigma_{i,j}^2}\right\}.$$

Under the prior in (9), the full conditional distribution of $\boldsymbol{t}_\theta$ is provided by

$$p(t_j = t \mid t_1, \ldots, t_{j-1}, \theta_{j-1}, \boldsymbol{z}_j, \boldsymbol{\sigma}_j) \propto \begin{cases} a(\omega, \theta_1, \ldots, \theta_{j-1}) f_{\theta_{j-1}}(\boldsymbol{z}_j|\boldsymbol{\sigma}_j) & \text{if } t = t_j, \\ \big[1 - a(\omega, \theta_1, \ldots, \theta_{j-1})\big] \displaystyle\int f_\theta(\boldsymbol{z}_j|\boldsymbol{\sigma}_j) G(d\theta) & \text{if } t = t^{\text{new}}, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, when G is conjugate with respect to the Gaussian kernel, the full conditional distribution of θ_t^*, given $\{\boldsymbol{z}_j : t_j = t\}$, is obtained in closed form using conjugacy of the Normal-Normal model.

Sampling the concentration parameter. Finally, the concentration parameter ω can be sampled through an importance sampling step using as importance distribution the prior p_ω over ω. Denoting with M_m the selected partition for $\boldsymbol{\theta}_m$ and with T_m the number of clusters in M_m, we have

$$p(\omega|M_m : m = 1, \ldots, M) \propto p_\omega(\omega) \frac{\omega^{\sum_{m=1}^M T_m - M}}{(\omega+2)^M(\omega^2+\omega+3)^M}.$$

5. Results.

5.1. *Simulation studies.* We perform a series of simulation studies with two main goals. First, we aim to highlight the drawbacks of clustering based on the entire distribution with respect to our proposal in the context of small sample sizes. Second, we check the model's ability of detecting the presence of underlying relevant factors in the sense described in Section 2.2.

To accomplish the first goal, we compare the results obtained using our model with the nested Dirichlet process (NDP) (Rodríguez, Dunson and Gelfand (2008)), arguably the most popular Bayesian model to cluster populations. Mimicking the real hypertensive dataset, we simulate data for four samples, ideally corresponding to four diseases, with respective sample sizes of 50, 19, 9 and 22, which correspond to the sample sizes of the real data investigated in Section 5.2. Since the NDP does not allow to treat jointly multiple response variables, we consider only one response variable to ensure a fair comparison. The observations are sampled from the following distributions and 100 simulation studies are performed:

$$X_{i,1} \overset{\text{iid}}{\sim} 0.5\mathcal{N}(0, 0.5) + 0.5\mathcal{N}(2, 0.5) \quad \text{for } i = 1, \ldots, n_1;$$

$$X_{i,2} \overset{\text{iid}}{\sim} 0.5\mathcal{N}(2, 0.5) + 0.5\mathcal{N}(4, 0.5) \quad \text{for } i = 1, \ldots, n_2;$$

$$X_{i,3} \overset{\text{iid}}{\sim} 0.5\mathcal{N}(4, 0.5) + 0.5\mathcal{N}(6, 0.5) \quad \text{for } i = 1, \ldots, n_3;$$

$$X_{i,4} \overset{\text{iid}}{\sim} 0.5\mathcal{N}(6, 0.5) + 0.5\mathcal{N}(8, 0.5) \quad \text{for } i = 1, \ldots, n_4.$$

Note that the true data generating process corresponds to samples from distinct distributions with pairwise sharing of a mixture component. Alternative scenarios are considered in the additional simulation studies that can be found in Section D of the Supplementary Material (Franzolini, Lijoi and Prünster (2023)).

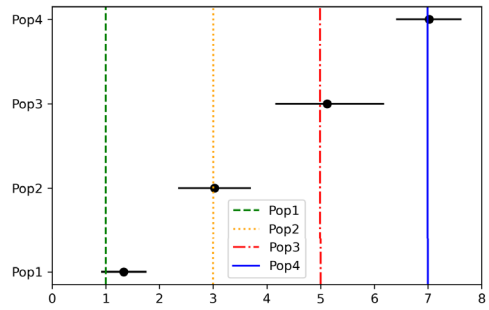
The implementation of the NDP was carried out through the marginal sampling scheme, proposed in Zuanetti et al. (2018), which is suitably extended to accommodate hyperpriors on the concentration parameters of the NDP. To simplify the choice of the hyperparameters, as suggested by Gelman et al. (2013, p. 535 and p. 551–554), we estimate both models over standardized data. For our model we set $G_m = \mathcal{N}(0, 1)$ and $P_{0,m} = \text{NIG}(\mu = 0, \tau = 1, \alpha = 2, \beta = 4)$. Here, $\text{NIG}(\mu, \tau, \alpha, \beta)$ indicates a normal inverse gamma distribution. The base distribution for the NDP is $\text{NIG}(\mu = 0, \tau = 0.01, \alpha = 3, \beta = 3)$, as in Rodríguez, Dunson and Gelfand (2008). Finally, we use gamma priors with shape 3 and rate 3 for all concentration parameters, which is a common choice. For each simulation study, we perform 10,000 iterations of the MCMC algorithms, with the first 5000 used as burn-in.

Table 1 displays summaries of the results on population clustering, darker rows correspond to partitions that are not consistent with the natural ordering of the diseases. The true clustering structure is given by the finest partition. As already observed in Rodríguez, Dunson and Gelfand (2008), the NDP tends to identify fewer rather than more clusters, due to the presence of small sample sizes. Using the *maximum* a posteriori estimate, our model correctly identifies the partition in 99 out of 100 simulation studies and a partition with three elements or more in 100 out of 100 simulation studies. The same counts for the NDP are, respectively, zero out of 100 and 21 out of 100. Analogous conclusions can be drawn looking at posterior probability averages and medians across the 100 simulation studies (see Table 1) leaving no doubt about the model to be preferred under this scenario.

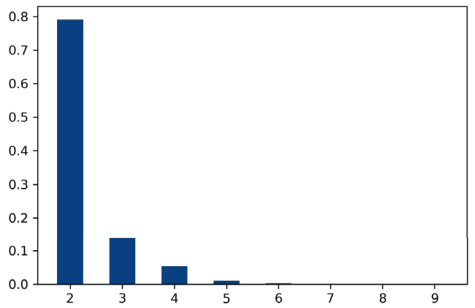
Finally, we randomly select three simulation studies among the 100 to better understand the performance in estimating the other model parameters. Here, we comment on one of the studies; the other two, leading to similar results, are reported in Section D.1.1 of the Supplementary Material (Franzolini, Lijoi and Prünster (2023)). Figure 3(a) shows point estimates and credible intervals for the population-specific location parameters θ_1, θ_2, θ_3, θ_4. The true means belong to the 95% credible intervals.

TABLE 1
Simulation studies summaries

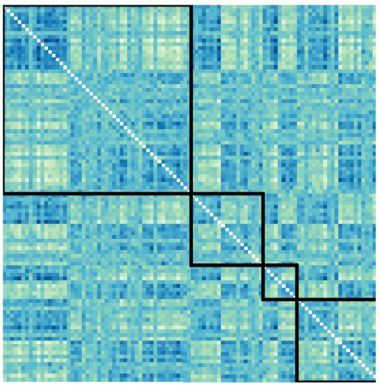
	sHDP			**NDP**		
Partitions	MAP count	Average post. prob.	Median post. prob.	MAP count	Average post. prob.	Median post. prob.
{1,2,3,4}	0	0.000	0.000	0	0.000	0.000
{1}{2,3,4}	0	0.000	0.000	2	0.020	0.000
{1,2}{3,4}	0	0.000	0.000	**72**	**0.695**	**0.860**
{1,3,4}{2}	0	0.000	0.000	0	0.000	0.000
{1}{2}{3,4}	0	0.027	0.007	3	0.035	0.000
{1,2,3}{4}	0	0.000	0.000	5	0.061	0.000
{1,4}{2,3}	0	0.000	0.000	0	0.000	0.000
{1}{2,3}{4}	1	0.054	0.015	0	0.014	0.000
{1,3}{2,4}	0	0.000	0.000	0	0.000	0.000
{1,2,4}{3}	0	0.000	0.000	0	0.000	0.000
{1}{2,4}{3}	0	0.000	0.000	0	0.000	0.000
{1,2}{3}{4}	0	0.004	0.000	18	0.175	0.032
{1,3}{2}{4}	0	0.000	0.000	0	0.000	0.000
{1,4}{2}{3}	0	0.000	0.000	0	0.000	0.000
{1}{2}{3}{4}	**99**	**0.915**	**0.954**	0	0.000	0.000



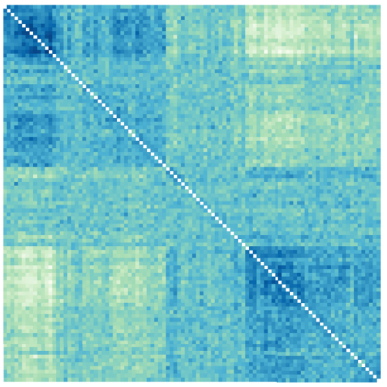
(a) 95% credible intervals for population-specific locations



(b) Number of second-level clusters.



(c) Coclustering.



(d) Coclustering.

FIG. 3. *Panel* (*a*): *Mean point estimates and* 95% *credible intervals for the four populations, vertical lines correspond to true values. Panel* (*b*): *Posterior distribution on the number of second-level clusters. Panels* (*c*) *and* (*d*): *Heatmaps of second-level clustering, darker colors correspond to higher probability of co-clustering*; *in* (*c*) *patients are ordered based on the diagnosis and the four black squares highlight the within-sample probabilities, and in* (*d*) *patients are reordered based on co-clustering probabilities.*

Moreover, it turns out that the model is able to detect the presence of two clusters of subjects leading to a posterior distribution for the number of clusters that is rather concentrated on the true value; see Figure 3(b)–3(d). Moreover, the point estimate for the subject partition, obtained minimizing the Binder loss function, also contains two clusters, proving the ability of the model to detect the underlying relevant factor. In Section D of the Supplementary Material (Franzolini, Lijoi and Prünster (2023)), a number of additional simulation studies are conducted, both using alternative specifications over the disorder-specific parameters and different data generating mechanisms: the results highlight a good performance of the model, which appears also able to detect outliers, identify other effects of the disorders than those affecting the location and produce reliable outputs even under deviation from symmetry.

5.2. *Impact of hypertensive disorders on maternal cardiac dysfunction.* Our analysis is based on the dataset of Tatapudi and Pasumarthy (2017a), which can be obtained from https://doi.org/10.17632/d72zr4xggx.1. The dataset contains observations for 10 cardiac function measurements, collected through a prospective case-control study on women in the third semester of pregnancy, divided in $n_1 = 50$ control cases (C), $n_2 = 19$ patients with gestational hypertension (G), $n_3 = 9$ patients with mild preeclampsia (M) and $n_4 = 22$ patients with severe preeclampsia (S). The cases are women admitted from 2012 to 2014 to the King George Hospital in Visakhapatnam, India. The healthy sample is composed by normotensive pregnant women. All women with hypertension were on antihypertensive treatment with

TABLE 2
Posterior probabilities over partitions of means. Maximum a posteriori probabilities are in bold

partitions	CI	CWI	LVMI	IVST	LVPW	EF	FS	EW	AW	E/A
{C,G,M,S}	0.021	0.000	0.000	0.000	0.000	**0.365**	**0.303**	0.096	0.000	0.000
{C}{G,M,S}	0.002	**0.546**	0.001	0.083	0.016	0.078	0.190	0.021	0.036	0.000
{C,G}{M,S}	0.002	0.000	0.001	0.000	0.000	0.037	0.038	0.072	0.076	0.049
{C,M,S}{G}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
{C}{G}{M,S}	0.001	0.139	0.001	0.019	0.024	0.028	0.078	0.042	0.232	0.055
{C,G,M}{S}	**0.463**	0.000	**0.595**	0.000	0.000	0.276	0.045	**0.498**	0.020	0.002
{C,S}{G,M}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
{C}{G,M}{S}	0.146	0.099	0.188	**0.551**	**0.672**	0.074	0.164	0.092	0.260	0.033
{C,M}{G,S}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
{C,G,S}{M}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
{C}{G,S}{M}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
{C,G}{M}{S}	0.233	0.000	0.107	0.000	0.000	0.083	0.062	0.114	0.091	0.371
{C,M}{G}{S}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
{C,S}{G}{M}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
{C}{G}{M}{S}	0.133	0.216	0.108	0.347	0.288	0.060	0.121	0.065	**0.287**	**0.491**
$\sum \log_{15}\left(p_i^{-p_i}\right)$	0.501	0.430	0.415	0.361	0.289	0.632	0.688	0.598	0.613	0.424

oral Labetalol or Nifedipine. Women with severe hypertension were treated with either oral nifedipine and parenteral labetalol or a combination. For more details on the dataset, we refer to Tatapudi and Pasumarthy (2017b). The prior specification is the same as in the previous section. Sections E.2 and E.3 of the Supplementary Material (Franzolini, Lijoi and Prünster (2023)) contain a prior-sensitivity analysis and show rather robust results w.r.t. different prior specifications. Inference is based on 10,000 MCMC iterations, with the first half used as burn-in.

Table 2 displays the posterior distributions for the partitions of unknown disease-specific means along with the corresponding entropy measurements that can be used as measures of uncertainty. First, note that if one takes also the ordering among distinct disease-specific locations into account, the posterior partition probabilities are, as desired, concentrated on specific orders of the associated unique values for all 10 cardiac indexes. For instance, we have $\mathbb{P}(\{\theta_{C,CI} = \theta_{G,CI} = \theta_{M,CI}\}\{\theta_{S,CI}\} \mid X) = \mathbb{P}(\theta_{C,CI} = \theta_{G,CI} = \theta_{M,CI} > \theta_{S,CI} \mid X) = 0.463$. The ordered partitions with the highest posterior probability are displayed in Table 3.

Considering the posterior probabilities summarized in Table 2 and in Table 3, we find that the cardiac index (CI) is reduced in severe preeclampsia, compared to all other patients, indicating reduced myocardial contractility in the presence of the most severe disorder. The cardiac work index (CWI) is a good indicator to distinguish between cases and control but

TABLE 3
Posterior probabilities over ordered partitions of means

cardiac index	ordered partition with highest posterior probability	posterior prob
CI	{C,G,M}>{S}	0.463
CWI	{C}<{G,M,S}	0.546
LVMI	{C,G,M}<{S}	0.595
IVST	{C}<{G,M}<{S}	0.548
LVPW	{C}<{G,M}<{S}	0.671
EF	{C,G,M,S}	0.365
FS	{C,G,M,S}	0.303
EW	{C,G,M}>{S}	0.497
AW	{C}<{G,M}<{S}	0.256
E/A	{C}>{G}>{M}>{S}	0.466

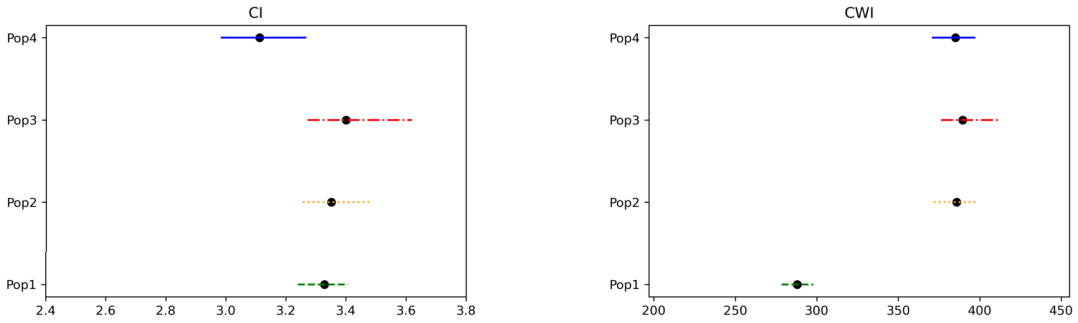


FIG. 4. *95% credible intervals for population-specific locations for CI and CWI.*

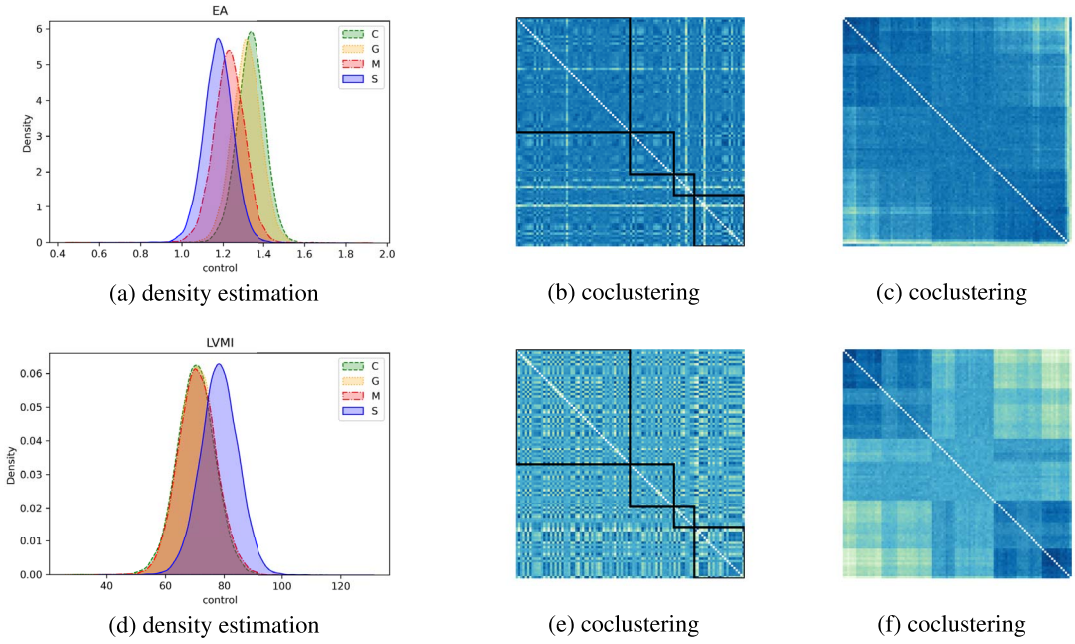
not among cases. The left ventricular mass index (LVMI) is increased in severe preeclampsia patients, compared to other pregnant women, indicating ventricular remodelling. Moreover, interventricular septal thickness (IVST) and left ventricular posterior wall thickness (LVPW) display a similar behaviour as they differ both between cases and controls and between severe preeclampsia and other disorders: this suggests a progressive increase in the indexes with the severity of the disorder. The posterior probabilities associated to indexes of systolic function, such as ejection fraction (EF) and fraction shortening (FS), are relatively concentrated on the partition of complete homogeneity, letting us to conclude that no differences are present among patients. As for the parameters of the diastolic function, the posterior distribution for the E-wave indicator identifies a modified index in severe preeclampsia patients, while the mean E/A ratio indicates a decreasing diastolic function with the severity of the disorder. The posterior for the A-wave index is actually concentrated on three distinct partitions, leaving a relatively high uncertainty regarding the modifications of the index. However, considering jointly the three partitions with the highest posterior probability, differences are detected between control and cases with a total posterior probability equal to 0.779. Figure 4 shows point estimates and credible intervals for disorder-specific location parameters for the first two cardiac indexes. Analogous plots for all cardiac indexes can be found in Section E.1 of the Supplementary Material (Franzolini, Lijoi and Prünster (2023)).

Table 4 shows the results obtained using the prior in (7), instead of (8). We remark that for all 10 cardiac indexes, the posterior associates negligible probabilities to partitions that

TABLE 4

Posterior probabilities over partitions of means. Maximum a posteriori probabilities are in bold

partitions	CI	CWI	LVMI	IVST	LVPW	EF	FS	EW	AW	E/A
{C,G,M,S}	0.019	0.000	0.000	0.000	0.000	**0.332**	**0.247**	0.078	0.000	0.000
{C}{G,M,S}	0.002	**0.643**	0.001	0.114	0.031	0.065	0.130	0.048	0.080	0.000
{C,G}{M,S}	0.004	0.000	0.003	0.000	0.000	0.044	0.019	0.152	0.073	0.103
{C,M,S}{G}	0.004	0.000	0.000	0.000	0.000	0.037	0.105	0.013	0.000	0.000
{C}{G}{M,S}	0.002	0.065	0.002	0.047	0.078	0.027	0.036	0.063	**0.424**	0.167
{C,G,M}{S}	**0.316**	0.000	**0.527**	0.000	0.000	0.178	0.032	**0.288**	0.002	0.000
{C,S}{G,M}	0.023	0.000	0.000	0.000	0.000	0.019	0.103	0.006	0.000	0.000
{C}{G,M}{S}	0.173	0.089	0.124	**0.472**	**0.594**	0.033	0.054	0.064	0.140	0.042
{C,M}{G,S}	0.002	0.000	0.001	0.003	0.000	0.044	0.031	0.017	0.000	0.000
{C,G,S}{M}	0.018	0.000	0.000	0.000	0.000	0.061	0.067	0.016	0.000	0.000
{C}{G,S}{M}	0.005	0.163	0.001	0.095	0.006	0.028	0.040	0.015	0.016	0.000
{C,G}{M}{S}	0.213	0.000	0.124	0.000	0.000	0.052	0.014	0.121	0.036	0.241
{C,M}{G}{S}	0.074	0.000	0.137	0.003	0.000	0.041	0.022	0.055	0.001	0.000
{C,S}{G}{M}	0.014	0.000	0.000	0.000	0.000	0.011	0.067	0.004	0.000	0.000
{C}{G}{M}{S}	0.133	0.040	0.079	0.265	0.291	0.029	0.033	0.059	0.229	**0.448**
$\sum \log_{15}\left(p_i^{-p_i}\right)$	0.687	0.407	0.509	0.501	0.371	0.828	0.886	0.823	0.582	0.505



(a) density estimation (b) coclustering (c) coclustering

(d) density estimation (e) coclustering (f) coclustering

FIG. 5. *Panels* (*a*) *and* (*d*): *Density estimates*. *Panels* (*b*)–(*c*) *and* (*e*)–(*f*): *heatmaps of the posterior probabilities of co-clustering*; *in* (*b*) *and* (*e*) *patients are ordered based on the diagnosis*, *and the four black squares highlight the within-sample probabilities*; *in* (*c*) *and* (*f*), *patients are reordered based on co-clustering probabilities*.

are in contrast with the natural order of the diagnoses. This is particularly reassuring in that the model, even without imposing such an order a priori, is able to single it out systematically across cardiac indexes. Moreover, we observe how the partitions identified by MAP are the same of Table 2 for all cardiac index, except AW. However, even under this alternative prior, the A-wave index is concentrated on the same three distinct partitions, leading to the conclusion that there exists a difference between cases and control.

As far as prediction and second-level clustering are concerned, Figure 5 displays the density estimates and the heatmap of co-clustering probabilities between pairs of patients for the E/A ratio and LVMI. Figure 5(b) shows that co-clustering probabilities are similar within and across diagnoses, indicating that the effect of the diseases on the distribution of the cardiac index is mostly explained through shifts between disease-specific locations. Moreover, Figure 5(b) suggests the presence of three outliers that have low probability of co-clustering with all the other subjects and that would be ignored by the model using a more traditional ANOVA structure. On the other hand, Figure 5(e) shows a slightly different pattern for co-clustering probabilities in the fourth square, which suggests that the heterogeneity between severe preeclampsia patients and the other patients is not entirely explained by shifts in disease-specific locations. Finally, Figure 5(f) suggests the presence of an underlying relevant factor. The corresponding figures for all 10 response variables are reported in Section E.1 of the Supplementary Material (Franzolini, Lijoi and Prünster (2023)) and can be used for prediction and for a graphical analysis aimed at controlling the presence of underlying relevant factors, outliers and differences across diseases distinct from shifts between disease-specific locations.

Our results are coherent with almost all of the findings in Tatapudi and Pasumarthy (2017b), where results were obtained through a series of independent frequentist tests. However, importantly, we are able to provide more insights thanks to the simultaneous comparison approach and the latent clustering of subjects. For instance, considering the response LVMI,

Tatapudi and Pasumarthy (2017b) detected a significant increase in cases compared to controls and an increase in severe preeclampsia compared to gestational hypertensive and mild preeclampsia patients. Such results do not clarify whether a modification exists between the control group and gestational hypertensive patients or between the latter and mild preeclampsia patients. Moreover, in contrast to our analysis, their results do not provide any information concerning the presence of underlying common factors, outliers or distributional effects (different from shifts in locations).

6. Concluding remarks. We designed a Bayesian nonparametric model to detect clusters of hypertensive disorders over different cardiac function indexes and found modified cardiac functions in hypertensive patients compared to healthy subjects as well as progressively increased alterations with the severity of the disorder. The proposed model has application potential also beyond the considered setup when the goal is to cluster populations according to multivariate information: it borrows strength across response variables, preserves the flexibility intrinsic to nonparametric models and correctly detects partitions of populations, even in presence of small sample sizes when alternative distribution-based clustering models tend to underestimate the number of clusters. The key component of the model is the s-HDP, a hierarchical nonparametric structure for the error terms that offers flexibility and serves as a tool to investigate the presence of unobserved factors, outliers and effects other than changes in locations. Interesting extensions of the model include generalizations to other types of invariances in order to accommodate identifiability in generalized linear models, for instance, with count data and a log link function as well as generalizations to other types of processes beyond the Dirichlet process.

Acknowledgements. The authors are grateful to the Editor, an Associate Editor and two anonymous referees for insightful comments and suggestions, which led to a substantial improvement of the manuscript. The authors are also affiliated to the Bocconi Institute for Data Science and Analytics (BIDSA). Most of the paper was completed while B. Franzolini was a Ph.D. student at the Bocconi University, Milan. A. Lijoi and I. Prünster are partially supported by MIUR, PRIN Project 2015SNS29B.

SUPPLEMENTARY MATERIAL

Supplementary Material to "Model selection for maternal hypertensive disorders with symmetric hierarchical Dirichlet processes" (DOI: 10.1214/22-AOAS1628SUPPA; .pdf). Review of the invariant Dirichlet process; derivation of predictive distributions, comparisons with alternative models; additional simulation studies; additional analysis and results on real data, including prior sensitivity.

Data and codes (DOI: 10.1214/22-AOAS1628SUPPB; .zip). Data and Python codes to reproduce all results.

REFERENCES

AKSU, E., CUGLAN, B., TOK, A., CELIK, E., DOGANER, A., SOKMEN, A. and SOKMEN, G. (2021). Cardiac electrical and structural alterations in preeclampsia. *J. Matern.-Fetal Neonatal Med.* 1–10.

AMBROŽIC, J., LUCOVNIK, M., PROKŠELJ, K., TOPLIŠEK, J. and CVIJIC, M. (2020). Dynamic changes in cardiac function before and early postdelivery in women with severe preeclampsia. *J. Hypertens.* **38** 1367–1374.

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. MR0365969

BELLAMY, L., CASAS, J.-P., HINGORANI, A. D. and WILLIAMS, D. J. (2007). Pre-eclampsia and risk of cardiovascular disease and cancer in later life: Systematic review and meta-analysis. *BMJ* **335** 974.

BERAHA, M., GUGLIELMI, A. and QUINTANA, F. A. (2021). The semi-hierarchical Dirichlet process and its application to clustering homogeneous distributions. *Bayesian Anal.* **16** 1187–1219. MR4381132 https://doi.org/10.1214/21-BA1278

CAMERLENGHI, F., DUNSON, D. B., LIJOI, A., PRÜNSTER, I. and RODRÍGUEZ, A. (2019a). Latent nested nonparametric priors (with discussion). *Bayesian Anal.* **14** 1303–1356. MR4044854 https://doi.org/10.1214/19-BA1169

CAMERLENGHI, F., LIJOI, A., ORBANZ, P. and PRÜNSTER, I. (2019b). Distribution theory for hierarchical processes. *Ann. Statist.* **47** 67–92. MR3909927 https://doi.org/10.1214/17-AOS1678

CHRISTENSEN, J. and MA, L. (2020). A Bayesian hierarchical model for related densities by using Pólya trees. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 127–153. MR4060979

CIFARELLI, D. M. and REGAZZINI, E. (1978). *Problemi Statistici Non Parametrici in Condizioni di Scambialbilita Parziale e Impiego di Medie Associative*. Quaderni Istituto Matematica Finanziaria, Torino.

CIPOLLI, W. III, HANSON, T. and MCLAIN, A. C. (2016). Bayesian nonparametric multiple testing. *Comput. Statist. Data Anal.* **101** 64–79. MR3504836 https://doi.org/10.1016/j.csda.2016.02.016

DAHL, D. B. and NEWTON, M. A. (2007). Multiple hypothesis testing by clustering treatment effects. *J. Amer. Statist. Assoc.* **102** 517–526. MR2325114 https://doi.org/10.1198/016214507000000211

DALAL, S. R. (1979a). Dirichlet invariant processes and applications to nonparametric estimation of symmetric distribution functions. *Stochastic Process. Appl.* **9** 99–107. MR0544719 https://doi.org/10.1016/0304-4149(79)90043-7

DALAL, S. R. (1979b). Nonparametric and robust Bayes estimation of location. In *Optimizing Methods in Statistics* (*Proc. Internat. Conf., Indian Inst. Tech., Bombay*, 1977) 141–166. Academic Press, New York. MR0541556

DAVIS, E. F., LAZDAM, M., LEWANDOWSKI, A. J., WORTON, S. A., KELLY, B., KENWORTHY, Y. et al. (2012). Cardiovascular risk factors in children and young adults born to preeclamptic pregnancies: A systematic review. *Pediatrics* **129** 1552–1561.

DE IORIO, M., MÜLLER, P., ROSNER, G. L. and MACEACHERN, S. N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.* **99** 205–215. MR2054299 https://doi.org/10.1198/016214504000000205

DEMARTELLY, V. A., DREIXLER, J., TUNG, A., MUELLER, A., HEIMBERGER, S., FAZAL, A. A., NASEEM, H., LANG, R., KRUSE, E. et al. (2021). Long-term postpartum cardiac function and its association with preeclampsia. *J. Am. Heart Assoc.* **10** e018526.

DENTI, F., GUINDANI, M., LEISEN, F., LIJOI, A., WADSWORTH, W. D. and VANNUCCI, M. (2021). Two-group Poisson–Dirichlet mixtures for multiple testing. *Biometrics* **77** 622–633. MR4307660 https://doi.org/10.1111/biom.13314

DIACONIS, P. and FREEDMAN, D. (1986). On inconsistent Bayes estimates of location. *Ann. Statist.* **14** 68–87. MR0829556 https://doi.org/10.1214/aos/1176349843

DO, K.-A., MÜLLER, P. and TANG, F. (2005). A Bayesian mixture model for differential gene expression. *J. Roy. Statist. Soc. Ser. C* **54** 627–644. MR2137258 https://doi.org/10.1111/j.1467-9876.2005.05593.x

DOLEA, C. and ABOUZAHR, C. (2003). Global burden of hypertensive disorders of pregnancy in the year 2000 Technical report, GBD 2000 Working Paper, World Health Organization, Geneva.

DOSS, H. (1984). Bayesian estimation in the symmetric location problem. *Z. Wahrsch. Verw. Gebiete* **68** 127–147. MR0767797 https://doi.org/10.1007/BF00531774

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

FRANZOLINI, B., LIJOI, A. and PRÜNSTER, I. (2023). Supplement to "Model Selection for Maternal Hypertensive Disorders with Symmetric Hierarchical Dirichlet Processes." https://doi.org/10.1214/22-AOAS1628SUPPA, https://doi.org/10.1214/22-AOAS1628SUPPB

GARCIA-GONZALEZ, C., GEORGIOPOULOS, G., AZIM, S. A., MACAYA, F., KAMETAS, N., NIHOYANNOPOULOS, P., NICOLAIDES, K. H. and CHARAKIDA, M. (2020). Maternal cardiac assessment at 35 to 37 weeks improves prediction of development of preeclampsia. *Hypertens.* **76** 514–522.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*. CRC Press, Boca Raton, FL.

GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1999). Consistent semiparametric Bayesian inference about a location parameter. *J. Statist. Plann. Inference* **77** 181–193. MR1687955 https://doi.org/10.1016/S0378-3758(98)00192-X

GOPALAN, R. and BERRY, D. A. (1998). Bayesian multiple comparisons using Dirichlet process priors. *J. Amer. Statist. Assoc.* **93** 1130–1139. MR1649207 https://doi.org/10.2307/2669856

GUINDANI, M., MÜLLER, P. and ZHANG, S. (2009). A Bayesian discovery procedure. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 905–925. MR2750250 https://doi.org/10.1111/j.1467-9868.2009.00714.x

GUTIÉRREZ, L., BARRIENTOS, A. F., GONZÁLEZ, J. and TAYLOR-RODRÍGUEZ, D. (2019). A Bayesian nonparametric multiple testing procedure for comparing several treatments against a control. *Bayesian Anal.* **14** 649–675. MR3959876 https://doi.org/10.1214/18-BA1122

HALL, M. E., GEORGE, E. M. and GRANGER, J. P. (2018). The heart during pregnancy. *Rev. Esp. Orientac.* **64** 1045–1050.

IGBERASE, G. and EBEIGBE, P. (2006). Eclampsia: Ten-years of experience in a rural tertiary hospital in the Niger Delta, Nigeria. *J. Obstet. Gynaecol.* **26** 414–417.

IGLESIAS, P. L., ORELLANA, Y. and QUINTANA, F. A. (2009). Nonparametric Bayesian modelling using skewed Dirichlet processes. *J. Statist. Plann. Inference* **139** 1203–1214. MR2479861 https://doi.org/10.1016/j.jspi.2008.07.009

LIJOI, A., NIPOTI, B. and PRÜNSTER, I. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20** 1260–1291. MR3217444 https://doi.org/10.3150/13-BEJ521

LIJOI, A., PRÜNSTER, I. and REBAUDO, G. (2022). Flexible clustering via hidden hierarchical Dirichlet priors. *Scand. J. Stat.*. https://doi.org/10.1111/sjos.12578

MACEACHERN, S. N. (2000). Dependent Dirichlet processes Technical Report, Department of Statistics, The Ohio State Univ.

MALIK, A., JEE, B. and GUPTA, S. K. (2019). Preeclampsia: Disease biology and burden, its management strategies with reference to India. *Pregnancy Hypertens.* **15** 23–31.

MARTIN, R. and TOKDAR, S. T. (2012). A nonparametric empirical Bayes framework for large-scale multiple testing. *Biostatistics* **13** 427–439.

MCCLURE, E. M., SALEEM, S., PASHA, O. and GOLDENBERG, R. L. (2009). Stillbirth in developing countries: A review of causes, risk factors and prevention strategies. *J. Matern.-Fetal Neonatal Med.* **22** 183–190.

MOSER, S., RODRÍGUEZ, A. and LOFLAND, C. L. (2021). Multiple ideal points: Revealed preferences in different domains. *Polit. Anal.* **29** 139–166.

MULIERE, P. and PETRONE, S. (1993). A Bayesian predictive approach to sequential search for an optimal dose: Parametric and nonparametric models. *J. Italian Stat. Soc.* **2** 349–364.

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804 https://doi.org/10.2307/1390653

PEDERSEN, S. S., VON KÄNEL, R., TULLY, P. J. and DENOLLET, J. (2017). Psychosocial perspectives in cardiovascular disease. *Eur. J. Prev. Cardiol.* **24** 108–115.

RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2008). The nested Dirichlet process. *J. Amer. Statist. Assoc.* **103** 1131–1144. MR2528831 https://doi.org/10.1198/016214508000000553

SCOTT, J. G. and BERGER, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plann. Inference* **136** 2144–2162. MR2235051 https://doi.org/10.1016/j.jspi.2005.08.031

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450 https://doi.org/10.1214/10-AOS792

SHAH, A., FAWOLE, B., M'IMUNYA, J. M., AMOKRANE, F., NAFIOU, I., WOLOMBY, J.-J. et al. (2009). Cesarean delivery outcomes from the WHO global survey on maternal and perinatal health in Africa. *Int. J. Gynecol. Obstet.* **107** 191–197.

SORIANO, J. and MA, L. (2017). Probabilistic multi-resolution scanning for two-sample differences. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 547–572. MR3611759 https://doi.org/10.1111/rssb.12180

TATAPUDI, R. and PASUMARTHY, L. R. (2017a). Data for: Maternal cardiac function in gestational hypertension, mild and severe preeclampsia and normal pregnancy: A comparative study. Available at https://data.mendeley.com/datasets/d72zr4xggx/1. https://doi.org/10.17632/d72zr4xggx.1 Licensed under a Creative Commons Attribution 4.0 International licence.

TATAPUDI, R. and PASUMARTHY, L. R. (2017b). Maternal cardiac function in gestational hypertension, mild and severe preeclampsia and normal pregnancy: A comparative study. *Pregnancy Hypertens.* **10** 238–241.

TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581. MR2279480 https://doi.org/10.1198/016214506000000302

TIMOKHINA, E., KUZMINA, T., STRIZHAKOV, A., PITSKHELAURI, E., IGNATKO, I. and BELOUSOVA, V. (2019). Maternal cardiac function after normal delivery, preeclampsia, and eclampsia: A prospective study. *J. Pregnancy* **2019** 2090–2727.

ZUANETTI, D. A., MÜLLER, P., ZHU, Y., YANG, S. and JI, Y. (2018). Clustering distributions with the marginalized nested Dirichlet process. *Biometrics* **74** 584–594. MR3825345 https://doi.org/10.1111/biom.12778