The Generalized Nested Common Atoms Model

Francesco Denti* and Laura D'Angelo†

Bayesian hierarchical nonparametric models offer a convenient framework for modeling nested data, where observations are organized into groups. These priors jointly accommodate the dependence among groups and among observations within the same group in a flexible way. Several recent instances of such models have combined nested levels of Dirichlet processes and a common sequence of atoms, a formulation that allows for multi-layered partitions, i.e., a simultaneous clustering of observations and groups. However, using a common set of atoms can lead to a forced high prior correlation between the generated random measures. This characteristic can cause shortcomings in the clustering results and even biased density estimation. Extending the nested process with more general stick-breaking specifications for the weights alleviates these issues. Specifically, the proposed generalized Common Atoms Model enhances the flexibility of the dependence structure and improves density estimation. Three notable instances, particularly useful for practical applications, are discussed, and an efficient Gibbs sampler algorithm for this novel nested mixture model is developed. Finally, posterior results are validated with simulation studies and a real data application.

Bayesian nonparametrics; mixture model; Pitman-Yor process; spike-and-slab; two-parameter beta process

Article published in `Econometrics and Statistics`
https://doi.org/10.1016/j.ecosta.2025.01.001

*Department of Statistical Sciences; University of Padua, Italy.
francesco.denti@unipd.it
†Department of Economics, Management and Statistics; University of Milano-Bicocca; Milan, Italy.
laura.dangelo@unimib.it

1. Introduction

Nested designs represent the standard experimental framework for analyzing multicenter data, where observations are organized into groups. Recently, Bayesian nonparametric nested models have emerged as an effective approach for modeling these data, thanks to their non-restrictive distributive assumptions, the possibility to borrow information across groups, and the ability to account for within- and between-group variability. Foundational contributions to this class of nonparametric priors are the Hierarchical Dirichlet process (HDP, Teh, Jordan, Beal, & Blei, 2006) and the nested Dirichlet Process (nDP, Rodríguez, Dunson, & Gelfand, 2008). These priors share a construction based on nested levels of discrete random distributions. Despite the similar structure, the way each level interacts with the others and the distributive assumptions of the latent measures profoundly affect the processes' realizations and, consequently, their clustering properties. The nDP, in particular, is defined as a Dirichlet process (DP) whose random atoms are discrete random probability measures, sampled independently from another DP. Such a construction entails a simultaneous partition of groups and observations, an appealing characteristic when modeling multicenter data. Successful applications of the nDP can be found, for example, in Graziani, Guindani, and Thall (2015); Rodríguez and Dunson (2014); Zuanetti, Müller, Zhu, Yang, and Ji (2018). However, recently, Camerlenghi, Dunson, Lijoi, Prünster, and Rodríguez (2019) showed how the nDP, by construction, does not allow observational clusters to be shared across different distributional clusters, forcing the model, in some cases, to create more clusters than needed or collapse different distributions together. To solve this issue, several extensions have been proposed (see, for example, Beraha, Guglielmi, & Quintana, 2021; D'Angelo, Canale, Yu, & Guindani, 2023; Lijoi, Prünster, & Rebaudo, 2023), including the Common Atoms Model (CAM, Denti, Camerlenghi, Guindani, & Mira, 2023). The CAM mimics the structure of the nDP but introduces a key difference: the random measures used as atoms are no longer independent realizations of a DP but rather random distributions based on a shared set of atoms. This modification enables clustering observations across groups, even when they are assigned to different distributional clusters. Nonetheless, because of such a structure, the CAM induces a high prior correlation between the generated random measures. This constraint can lead to borrowing *too much* information between groups, resulting in biased inference, especially of the posterior density estimates. A solution to this problem has been proposed in D'Angelo and Denti (2024), where the authors suggest a hybrid approach, adopting a finite (parametric) mixing measure for the observational layer of the model. On the contrary, in this paper, we investigate fully nonparametric specifications, exploring how the CAM's nested framework can be extended with beta stick-breaking (SB) process priors (Ishwaran & James, 2001), and how different distributional assumptions can dramatically impact the flexibility of the induced random measures.

The rest of this paper is organized as follows. In Section 2, we introduce the framework and provide a general expression for the correlation of nested random measures with common atoms. In Section 2.2, we derive closed-form expressions for three special cases: the Pitman-Yor process (Pitman & Yor, 1997), the two-parameter beta process (Ishwaran

& Zarepour, 2000), and the atom-skipping process (Bi & Ji, 2023), discussing the pros and cons of the different specifications. In Section 3, we illustrate the use of the proposed prior as a mixing measure in nonparametric mixtures, and we derive an efficient blocked Gibbs sampler algorithm to perform posterior inference. Then, in Section 4, we focus on models characterized by Dirichlet process distributional weights and investigate the posterior performances of different processes for the observational weights on simulated data. Finally, we compare the flexibility of our proposed model with the standard CAM on the Collaborative Perinatal Project study, a benchmark dataset for nested clustering. Section 5 discusses possible future research directions suggested by these findings. All the proofs are deferred to the Supplementary Material.

2. The generalized nested common atoms model

2.1. Definition and general results

Consider a nested design where the data are divided into J groups, each containing N_j measurements. Specifically, denote the data as $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_J)$, where $\boldsymbol{y}_j = (y_{1,j}, \ldots, y_{N_j,j})$ is the sample in the j-th group, for $j = 1, \ldots, J$. The generic observation $y_{i,j}$, for $i = 1, \ldots, N_j$ and $j = 1, \ldots, J$, takes values in a Polish space $\mathbb{X}$ endowed with the respective Borel σ-field $\mathcal{X}$. The generalized Common Atoms Model (geCAM) assumes that each sample is generated by a group-specific distribution G_j. The group-specific random measures G_j's are sampled from an almost surely discrete distribution Q, defined over the space of probability distributions on $\mathcal{X}$. In formulas, we write, for $j = 1, \ldots, J$,

$$\begin{aligned} y_{1,j}, \ldots, y_{N_j,j} \mid G_j \sim G_j, \quad G_j \mid Q \sim Q, \\ \text{and} \quad Q = \sum_{k \geq 1} \pi_k \delta_{G_k^*}. \end{aligned} \tag{1}$$

We assume a common atoms structure for the *distributional atoms* G_k^*, i.e.,

$$G_k^* = \sum_{l \geq 1} \omega_{l,k} \delta_{\theta_l^*}, \tag{2}$$

where, crucially, $\{\theta_l^*\}_{l \geq 1}$ is a sequence of i.i.d. *common observational atoms*, randomly sampled from a non-atomic base measure H defined on $(\mathbb{X}, \mathcal{X})$. The sequences $\boldsymbol{\pi} = \{\pi_k\}_{k \geq 1}$ and $\boldsymbol{\omega}_k = \{\omega_{l,k}\}_{l \geq 1}$, for $k \geq 1$, are independent, and they are referred to as distributional and observational weights, respectively. We assume a general formulation for the distributions of these weights, denoting with $\mathcal{L}(\boldsymbol{\pi})$ the law of $\boldsymbol{\pi}$, and with $\mathcal{L}(\boldsymbol{\omega}_k)$ the common law of the $\boldsymbol{\omega}_k$'s, for $k \geq 1$. We will refer to the model defined in Equations (1) and (2) as the geCAM with laws $\mathcal{L}(\boldsymbol{\pi})$ and $\mathcal{L}(\boldsymbol{\omega}_k)$, and write $Q \sim \text{geCAM}(\mathcal{L}(\boldsymbol{\pi}), \mathcal{L}(\boldsymbol{\omega}_k), H)$.

The discrete nature of Q and the G_k^*'s allows for the possibility of ties between distributions and between observations, hence inducing a distributional and an observational partition, respectively. The former is guaranteed by noticing that $\mathbb{P}\left[G_j = G_{j'}\right] > 0$, for $j, j' = 1, \ldots, J$; while the latter follows since $\mathbb{P}\left[y_{i,j} = y_{i',j'}\right] > 0$, for $i = 1, \ldots, N_j$,

$i' = 1, \dots, N_{j'}$ and $j, j' = 1, \dots, J$. The following proposition shows that, under this general construction, the correlation between two probability measures G_j and $G_{j'}$ is related to these two co-clustering probabilities and can be expressed as a function of the distributional and observational weights.

Proposition 2.1 *Let $G_j, G_{j'} \mid Q \sim Q$, with $j \neq j'$, and $Q \sim \text{geCAM}(\mathcal{L}(\boldsymbol{\pi}), \mathcal{L}(\boldsymbol{\omega}_k), H)$. Then, the correlation between G_j and $G_{j'}$ is given by*

$$\begin{aligned} \rho_{j,j'} &:= Corr(G_j, G_{j'}) \\ &= 1 - \left(1 - \mathbb{P}\left[G_j = G_{j'}\right]\right) \left(1 - \frac{\mathbb{P}\left[y_{i,j} = y_{i',j'} \mid G_j \neq G_{j'}\right]}{\mathbb{P}\left[y_{i,j} = y_{i',j'} \mid G_j = G_{j'}\right]}\right) \\ &= 1 - (1 - q_1)(1 - q_2), \end{aligned} \tag{3}$$

where $q_1 = \mathbb{P}\left[G_j = G_{j'}\right] = \sum_{k \geq 1} \mathbb{E}\left[\pi_k^2\right]$ and

$$q_2 = \frac{\mathbb{P}\left[y_{i,j} = y_{i',j'} \mid G_j \neq G_{j'}\right]}{\mathbb{P}\left[y_{i,j} = y_{i',j'} \mid G_j = G_{j'}\right]} = \frac{\sum_{l \geq 1} \mathbb{E}\left[\omega_{l,k}\right]^2}{\sum_{l \geq 1} \mathbb{E}\left[\omega_{l,k}^2\right]}.$$

Moreover, the correlation is always non-negative.

Notice that $\text{Corr}(G_j, G_{j'})$ is short for $\text{Corr}(G_j(A), G_{j'}(A))$, with A a generic Borel set $A \in \mathcal{X}$. However, we can suppress the dependence on the specific set since the correlation is constant for any choice of A. For this reason, we follow the indications of Lijoi et al. (2023) and focus on $\rho_{j,j'}$ as a measure of the overall dependence between G_j and $G_{j'}$. From Equation (3) we see that $\rho_{j,j'}$ can be expressed as a difference between the upper bound 1 and a product of two quantities. The first is $(1 - q_1) = \mathbb{P}\left[G_j \neq G_{j'}\right]$, whose interpretation is trivial. The second quantity, $(1 - q_2) = \sum_{l \geq 1} Var(\omega_{l,k}) / \sum_{l \geq 1} \mathbb{E}\left[w_{l,k}^2\right]$, is less straightforward, and the leading cause of the poor flexibility induced, for example, by the Dirichlet process prior used in the definition of the CAM proposed by Denti, Camerlenghi, et al. (2023) (see Section 2.2.1). Hence, in the following, we will study the effect on q_2 of different SB specifications for $\boldsymbol{\omega}_k$.

Before moving to the specific cases, one last remark is useful. Proposition 2.1 expresses the correlation as a function of the *conditional* probability of observational ties. However, it is immediate to explicitly link $\rho_{j,j'}$ to the marginal probability as well, as stated in the following corollary.

Corollary 2.2 *Let $y_{i,j} \mid G_j \sim G_j$ and $y_{i',j'} \mid G_{j'} \sim G_{j'}$, with $j \neq j'$, be two realizations from two random probability measures sampled from $Q \sim \text{geCAM}(\mathcal{L}(\boldsymbol{\pi}), \mathcal{L}(\boldsymbol{\omega}_k), H)$. Then,*

$$\rho_{j,j'} = \frac{\mathbb{P}\left[y_{i,j} = y_{i',j'}\right]}{\sum_{l \geq 1} \mathbb{E}\left[\omega_{l,k}^2\right]}. \tag{4}$$

Equations (3) and (4) hold for generic common atoms models, as long as the laws $\mathcal{L}(\boldsymbol{\pi})$ and $\mathcal{L}(\boldsymbol{\omega}_k)$ are valid distributions for modeling the mixture weights. In the following, we discuss the popular class of SB priors based on independent sequences of beta random variables (Ishwaran & James, 2001).

2.2. The impact of stick-breaking priors on the prior correlation

We begin by recalling the definition of SB process using the construction with generic beta random variables of Ishwaran and James (2001). As will be discussed later, the law $\mathcal{L}(\boldsymbol{\pi})$ has a limited impact on the nested prior, whose behavior is ultimately defined by $\mathcal{L}(\boldsymbol{\omega}_k)$. Hence, we focus on the observational weights and assume that all sequences $\boldsymbol{\omega}_k$'s are identically distributed according to a beta SB process,

$$\omega_{l,k} = u_{l,k} \prod_{q<l} (1 - u_{q,k}), \quad \text{with} \quad u_{l,k} \sim \text{Beta}(a_l, b_l) \quad \text{for } l, k \geq 1, \tag{5}$$

which is equivalent to writing $\boldsymbol{\omega}_k \sim \text{SB}(\boldsymbol{a}, \boldsymbol{b})$ for all $k \geq 1$. Here, both $\boldsymbol{a} = \{a_l\}_{l\geq 1}$ and $\boldsymbol{b} = \{b_l\}_{l\geq 1}$ are positive-valued sequences such that $\sum_{l\geq 1} \log(1 + a_l/b_l) = +\infty$ (see Lemma 1 in Ishwaran & James, 2001). Under the general specification in Equation (5), we have that

$$\mathbb{E}\left[\omega_{l,k}\right]^2 = \left(\frac{a_l \prod_{q=1}^{l-1} b_q}{\prod_{q=1}^{l} (a_q + b_q)} \right)^2 \quad \text{and} \quad \mathbb{E}\left[\omega_{l,k}^2\right] = \frac{a_l(a_l+1) \prod_{q=1}^{l-1} b_q(b_q+1)}{\prod_{q=1}^{l} (a_q + b_q)(a_q + b_q + 1)}.$$

This framework encompasses several popular nonparametric priors. In the next subsections, we explore specific instances of SB priors and discuss their impact on the correlation among nested random measures.

2.2.1. The Dirichlet process

The most popular SB prior was introduced as a constructive definition of the DP (Sethuraman, 1994). The weights of a DP with concentration parameter β are obtained by setting $a_l = 1$ and $b_l = \beta$ for $l \geq 1$ in Equation (5), i.e., assuming that $\boldsymbol{\omega}_k \sim \text{SB}(1, \beta)$ for all $k \geq 1$. The $\text{SB}(1, \beta)$ prior specification for the observational weights stands at the basis of the original CAM, and its effects have been studied in Denti, Camerlenghi, et al. (2023). Here, we recall their result in the following Corollary as a benchmark.

Corollary 2.3 *Let $G_j, G_{j'} \mid Q \sim Q$, with $j \neq j'$, and $Q \sim \text{geCAM}(\mathcal{L}(\boldsymbol{\pi}), SB(1, \beta), H)$. Then,*

$$\rho_{j,j'} = 1 - (1 - q_1) \frac{\beta}{1 + 2\beta}. \tag{6}$$

We remark that, in Equation (6), $(1 - q_2) = \beta/(1 + 2\beta) \in (0.5, 1)$, hence $\rho_{j,j'}$ falls within the same interval. This limited range of attainable correlations stems from the interaction between the stochastic ordering of the SB construction and the implicit

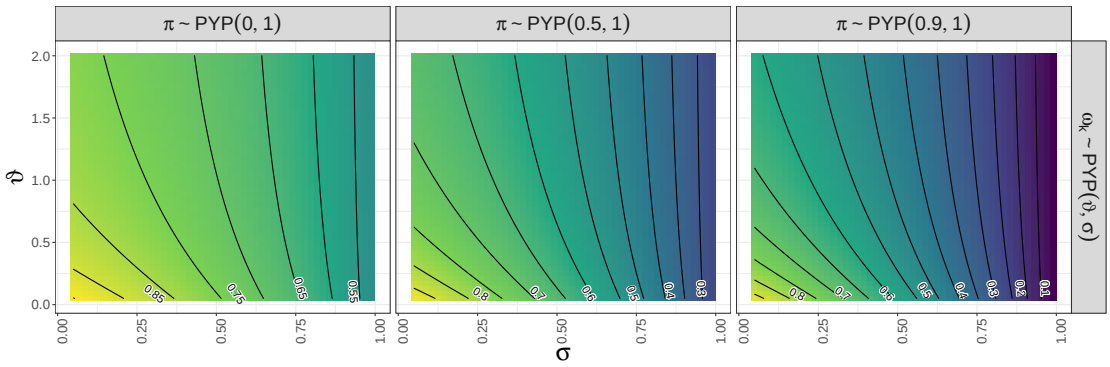

Figure 1: Correlation for varying parameters of the observational PYP prior. Each panel corresponds to a different PYP specification of the distributional weights.

ordering of the common atoms across the different latent distributions G_k^*'s. Specifically, $\mathbb{P}[y_{i,j} = \theta_l^*] = \mathbb{E}[\omega_{l,k}] = \beta^{l-1}/(1+\beta)^l$ is geometrically decreasing in l and, crucially, does not depend on k. In other words, atoms appearing earlier in the sequence $\{\theta_l^*\}_{l\geq 1}$ are highly favored in expectation, and this relation holds *across all the random measures*. Adopting a more general SB specification can alleviate the implicit ordering problem.

2.2.2. The Pitman-Yor process

An intuitive approach to address the implicit ordering problem is to adopt a process that exhibits a slower decay of the weights compared to the DP. One such process is the Pitman-Yor process (PYP, Pitman & Yor, 1997), which is renowned for its power-law tails behavior (Ghosal & van der Vaart, 2017). While $\mathbb{P}[y_{i,j} = \theta_l^*]$ is still decreasing in l, the heavy tails of the PYP lead to higher variance and more diverse realizations. To obtain a PYP from the general SB construction in Equation (5), we set $a_l = 1 - \sigma$ and $b_l = \vartheta + l\sigma$. In the following, this process will be denoted as $\text{PYP}(\vartheta, \sigma)$, with the parameter σ being referred to as the *discount parameter*. Here, we consider $\sigma \in (0, 1)$ and $\vartheta > 0$. Note that when $\sigma = 0$, we obtain the DP as a special case. If we specify Proposition 2.1 to the case of a PYP, we get the following corollary.

Corollary 2.4 *Let* $G_j, G_{j'} \mid Q \sim Q$, *with* $j \neq j'$, *and* $Q \sim \text{geCAM}(\mathcal{L}(\boldsymbol{\pi}), PYP(\vartheta, \sigma), H)$. *Then,*

$$\rho_{j,j'} = 1 - (1 - q_1)\left[1 - \frac{(1-\sigma)(1+\vartheta)}{\vartheta^2}\mathcal{S}^2_{\sigma,\vartheta}\right], \tag{7}$$

where $\mathcal{S}^2_{\sigma,\vartheta} = \sum_{l\geq 1}\zeta_l^2(\sigma,\vartheta)$ *and* $\zeta_l(\sigma,\vartheta) = \tilde{\vartheta}^{[l]}/\hat{\vartheta}^{[l]}$. *In the last expression,* $\tilde{\vartheta} = \vartheta/\sigma$ *and* $\hat{\vartheta} = (\vartheta+1)/\sigma$, *while* $a^{[l]} = \prod_{s=0}^{l-1}(a+s)$ *indicates the ascending factorial.*

The expression in Equation (7) does not have a simple interpretation, but it is straightforward to evaluate numerically. Figure 1 shows the resulting correlation for varying parameters of the prior on the observational weights. Each panel reflects a different PYP specification of the distributional weights $\boldsymbol{\pi}$. In the Supplementary Material, we

report additional heatmaps depicting the evolution of the term $(1 - q_2)$, independently of q_1. Since now $(1-q_2) \in (0,1)$ also $\rho_{j,j'}$ can attain values in $(0,1)$. The tail behavior of the process is governed by the discount parameter σ, and thus, the correlation is heavily affected by its value. Larger values of σ make ties less likely to occur *a priori*, redistributing the probability mass across numerous atoms. Conversely, the parameter ϑ has a negligible impact unless σ is close to zero. Upon these considerations, the PYP emerges as a valuable choice for the law of the observational weights in the geCAM, especially when considering large values of the discount parameter. However, it's worth noting that this choice may affect the practicality of posterior computation. For instance, as σ approaches one, many conditional samplers for nonparametric mixtures become unfeasible, necessitating reliance on more sophisticated sampling schemes (Canale, Corradin, & Nipoti, 2022).

2.2.3. The two-parameter beta process

A computationally simpler alternative is the two-parameter beta process (2PBP, Ishwaran & Zarepour, 2000), obtained by setting $a_l = s_1 > 0$ and $b_l = s_2 > 0$ in (5): we will denote this process as 2PBP(s_1, s_2). Here, the decay of the weights is again geometric, since $\mathbb{P}[y_{i,j} = \theta_l^*] = s_1 s_2^{l-1}/(s_1 + s_2)^l$, but, unlike the DP, there are two free parameters controlling the tail behavior. We can establish the following:

Corollary 2.5 *Let* $G_j, G_{j'} \mid Q \sim Q$, *with* $j \neq j'$, *and* $Q \sim \text{geCAM}(\mathcal{L}(\boldsymbol{\pi}), \text{2PBP}(s_1, s_2), H)$. *Then,*

$$\rho_{j,j'} = 1 - (1 - q_1)\left(1 - \frac{s_1(s_1 + 1 + 2s_2)}{(s_1 + 2s_2)(s_1 + 1)}\right).$$

Notably, a lower correlation is achieved when s_1 approaches zero: small values of s_1 lead to the creation of sticks with negligible mass, inducing a slow decay rate. Specifically, for $s_2 > 0$, $\lim_{s_1 \to 0} \rho_{j,j'} = q_1 \in (0,1)$, proving the 2PBP as a valid solution to improve the flexibility of the prior. In Figure 2, we present the correlation as a function of the parameters s_1 and s_2 at the observational level, assuming again different specifications for $\boldsymbol{\pi}$.

A special case of this process arises when we assume $s_1 = s_2 = s < 1$, which corresponds to a horseshoe-shaped beta random variable (Carvalho, Polson, & Scott, 2010). In this case, each $\omega_{l,k}$ assumes with high probability either a value very close to 0 (hence, assigning negligible mass to the atom θ_l^*) or very close to 1 (assigning most of the mass to the atom). The correlation is then given by $\rho_{j,j'} = 1 - 2(1 - q_1)/(3(s + 1))$, and $\lim_{s \to 0} \rho_{j,j'} = (1 + 2q_1)/3 \in (1/3, 1)$. Thus, with just one free parameter to control the tail behavior, we revert to a correlation that is lower-bounded by $1/3$, representing only a slight improvement compared to the DP. Nonetheless, the above reasoning suggests an ingenious way to make the model more flexible. Indeed, it emerges that a key factor for reducing the correlation lies in the ability to randomly "skip" some atoms in the common sequence by assigning null mass in a non-deterministic order across various random measures. This behavior can be achieved by the atom-skipping process of Bi and Ji (2023), which we discuss in the following paragraph.

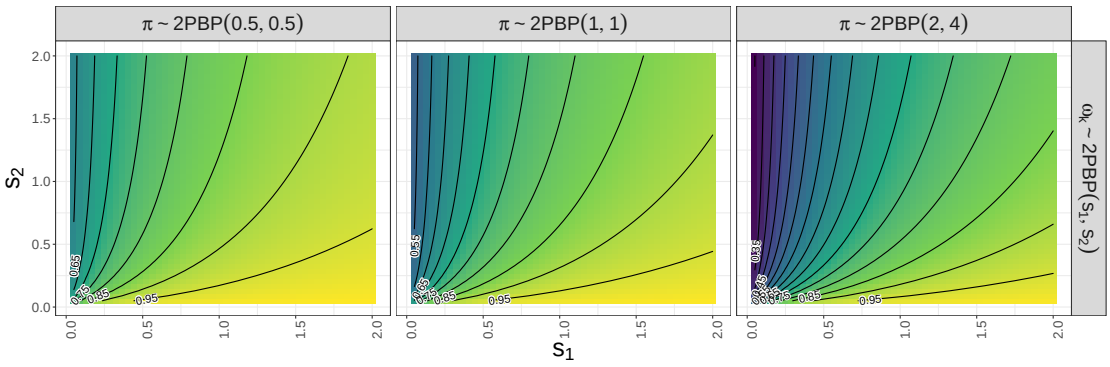

Figure 2: Correlation under the 2PBP prior. Each panel corresponds to a different 2PBP specification of the distributional weights.

2.2.4. The skip-breaking process

The atom-skipping process was recently proposed by Bi and Ji (2023) to extend the HDP to allow for the presence of group-specific observational clusters. This is achieved through a zero-augmented beta distribution for the SB variables, similar to the Quasi-Bernoulli SB process of Zeng, Miller, and Duan (2023). We apply the same idea to reduce correlation in the geCAM, extending the 2PBP with the following spike-and-slab formulation:

$$u_{l,k} \sim p\,\delta_0 + (1-p)\text{Beta}(s_1, s_2), \quad \text{for } l, k \geq 1. \tag{8}$$

With this definition, $u_{l,k}$ can assume the value zero with probability p, resulting in the omission of the l-th atom from the common sequence $\{\theta_l^*\}_{l\geq 1}$. To stress the meaning of this construction, in our nested setting, we refer to SB weights generated according to Equation (8) as distributed following a *skip-breaking process*, denoted with $\text{SKBP}(s_1, s_2, p)$. Notice that this specification is equivalent to a two-parameter beta process where the first hyperparameter s_1 randomly degenerates to zero with probability p since $\text{Beta}(0, s_2) \overset{(d)}{=} \delta_0$. In other words, we can rewrite Equation (8) in hierarchical form, for $l, k \geq 1$, as:

$$u_{l,k} \mid x_{l,k} \sim \text{Beta}(s_1 \cdot x_{l,k}, s_2) \quad \text{and} \quad x_{l,k} \sim \text{Bern}(1-p). \tag{9}$$

The additional parameter $p \in [0, 1)$ grants more control over the evolution of the weights, which is now given by

$$\mathbb{P}\left[y_{i,j} = \theta_l^*\right] = (1-p)s_1 \frac{(s_1 p + s_2)^{l-1}}{(s_1 + s_2)^l}.$$

Therefore, when $p = 0$, we revert to the 2PBP, while $p \to 1$ leads to a sequence of zeros. It is possible to prove the following:

Corollary 2.6 *Let* $G_j, G_{j'} \mid Q \sim Q$, *with* $j \neq j'$, *and* $Q \sim \text{geCAM}(\mathcal{L}(\boldsymbol{\pi}), \text{SKBP}(s_1, s_2, p), H)$.

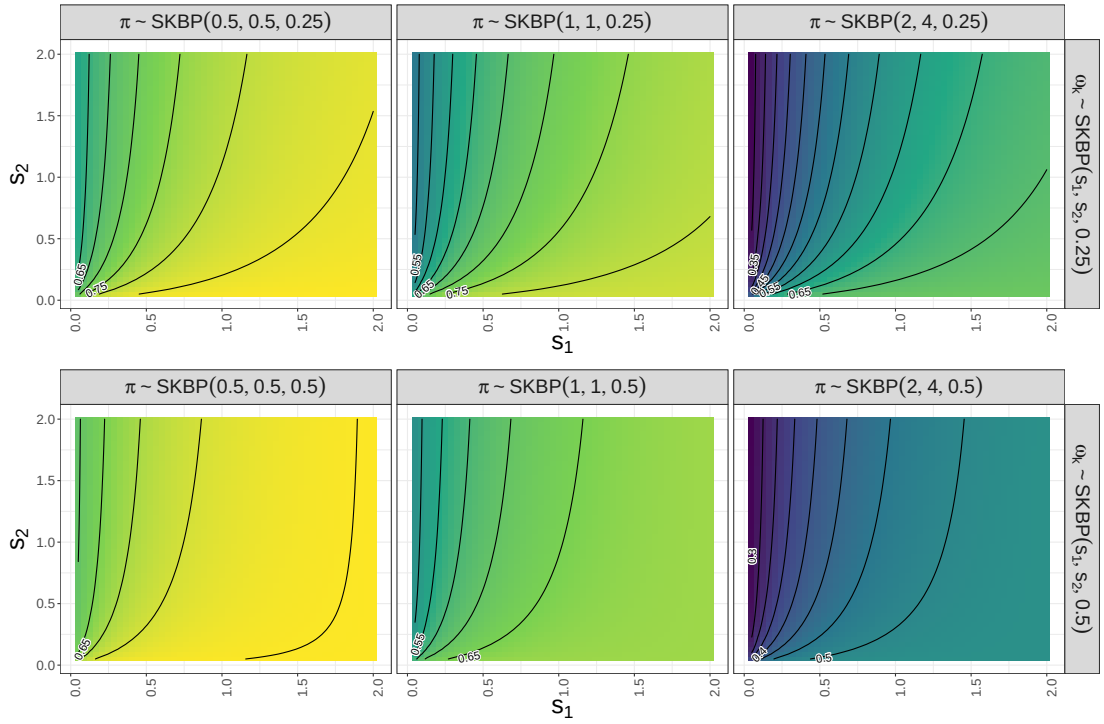

Figure 3: Correlation under the SKBP prior with $p = 0.25$ (top row) and $p = 0.50$ (bottom row). Each panel corresponds to a different SKBP specification of the distributional weights.

Then,

$$\rho_{j,j'} = 1 - (1 - q_1)\left(1 - \frac{(1-p)s_1(s_1 + 2s_2 + 1)}{((1+p)s_1 + 2s_2)(s_1 + 1)}\right).$$

Figure 3 displays the correlations for different specifications of $\boldsymbol{\pi}$ and p. As the probability of a skip approaches 1, the concentration parameters can assume a broader range of values while maintaining low correlation. In other words, the SKBP enhances the flexibility of the 2PBP, concentrating the mass only on active atoms and discarding the superfluous ones.

The improvement introduced by the random skipping is well demonstrated if we consider the special case where $s_1 = s_2 = s$, as we did in the previous paragraph. This choice leads to $(1 - q_2) = 2(1 + p + 2ps)/((3 + p)(s + 1))$, and $\lim_{s\to 0} \rho_{j,j'} = (1 - p + 2q_1(p + 1))/(3 + p) \in (0, 1)$. The additional parameter p in the latter quantity extends the 2PBP(s, s) and enables the prior to attain the full range of possible correlations. Therefore, the SKBP offers an effective integration of flexibility and computational tractability.

2.2.5. Combining different processes

Notice that, in Figures 1, 2, and 3, we have assumed the same process on both $\boldsymbol{\pi}$ and the $\boldsymbol{\omega}_k$'s. However, we remarked on how the flexibility of the overall process is essentially defined by the distribution assumed for the observational weights, affecting the term $(1 - q_2)$. The different roles of the two sequences of weights hence suggest that a mixed approach, where $\boldsymbol{\pi}$ and the $\boldsymbol{\omega}_k$ are assigned different prior distributions, could be a valid strategy. Figures S2, S3, S4, and S5 of the Supplementary Material investigate this solution, confirming how the process on the distributional weights has a limited impact on the correlation support. Thus, its choice can be driven by other considerations (e.g., the computational tractability).

We can conclude that one should avoid a DP prior on the observational weights unless it is strongly motivated by the specific problem under consideration (as, for example, in the application of Denti, Camerlenghi, et al., 2023). This choice indeed prevents the resulting correlation from escaping the interval $(0.5, 1)$, regardless of the process that drives the distributional weights. However, thanks to its computational tractability, the DP still represents a convenient prior for the distributional weights, as long as it is combined with a more flexible prior on the observational ones.

3. Posterior inference for generalized common atoms mixture models

With continuous measurements, using the discrete distributions G_j's to model the data is not appropriate, and it is customary to convolute them with a continuous kernel $f(\cdot \mid \theta)$. Hence, in the following, we assume that

$$y_{i,j} \mid p_j \sim p_j, \quad \text{with} \quad p_j = \int_\Theta f(\cdot \mid \theta) dG_j(\theta),$$

for $i = 1, \ldots, N_j$ and $j = 1, \ldots, J$. In this section, we devise a blocked Gibbs sampler algorithm for performing posterior inference in mixture models driven by a geCAM prior. In particular, we keep a general formulation on the distributional weights, while we explicitly state the case of a skip-breaking process on the observational weights. The other processes mentioned in Section 2 can be easily recovered by setting $p = 0$ and carefully choosing the appropriate values for the sequences $\{s_{1,l}\}_{l \geq 1}$ and $\{s_{2,l}\}_{l \geq 1}$.

Before detailing the algorithm, it is useful to specify the model with the augmented formulation involving the random variables $\{S_j\}_{j=1}^J$ and $\{\{M_{i,j}\}_{i=1}^{N_j}\}_{j=1}^J$ that indicate the distributional and observational cluster memberships, respectively. Hence, $(G_j \mid$

$S_j = k) = G_k^*$, and $(\theta_{i,j} \mid M_{i,j} = l) = \theta_l^*$. The model can be expressed as

$$
\begin{aligned}
y_{i,j} \mid M_{i,j}, \{\theta_l^*\}_{l\geq 1} &\sim f(y_{i,j}|\theta^*_{M_{i,j}}),\\
p(M_{i,j}|S_j, \boldsymbol{\omega}) = \sum_{l\geq 1} \omega_{l,S_j}\delta_l(\cdot), &\quad p(S_j|\boldsymbol{\pi}) = \sum_{k\geq 1}\pi_k\delta_k(\cdot),\\
\pi_k = v_k\prod_{g<k}(1-v_g), &\quad \text{with} \quad v_k \sim \text{Beta}(a_k^D, b_k^D),\\
\omega_{l,k} = u_{l,k}\prod_{q<l}(1-u_{q,k}), &\quad \text{with} \quad u_{l,k} \sim \text{Beta}(s_{1,l}\cdot x_{l,k}, s_{2,l}),\\
x_{l,k} \mid p \sim \text{Bern}(1-p), &\quad \text{and} \quad \theta_l^* \overset{i.i.d.}{\sim} H,
\end{aligned}
\tag{10}
$$

where the vectors $\{a_k^D\}_{k\geq 1}$ and $\{b_k^D\}_{k\geq 1}$ indicate the parameters of the beta laws governing the distributional stick-breaking processes. Finally, p can be set to a fixed value or assigned an additional prior level, $p \sim \text{Beta}(b_1^0, b_2^0)$. We implement a blocked Gibbs sampler, relying on truncating the processes to K distributional weights and L observational weights (Ishwaran & James, 2001; Lijoi et al., 2023; Rodríguez et al., 2008). The steps are outlined in Algorithm 1. Notice that we adopt a collapsed Gibbs sampler to improve mixing: in particular, the full conditional distribution in step 2 is obtained by marginalizing over $x_{l,k}$. Similarly, we integrate out $M_{i,j}$ in step 3.

Algorithm 1: Blocked Gibbs sampler for geCAM mixture models.

1. For each $k = 1, \dots, K$, compute $m_k = \sum_{j=1}^{J} \mathbb{1}_{\{S_j = k\}}$. Then, sample v_k from a $\text{Beta}(a_k + m_k, b_k + \sum_{g=m+1}^{K} m_g)$ distribution.
2. For each $k = 1, \dots, K$ and $l = 1, \dots, L$, compute $s_{1,l}^* = s_{1,l} + n_{l,k}$ and $s_{2,l}^* = s_{2,l} + \sum_{q=l+1}^{L} n_{q,k}$, where $n_{l,k} = \sum_{j=1}^{J}\sum_{i=1}^{N_j} \mathbb{1}_{\{S_j = k \,\cap\, M_{i,j} = l\}}$.
   Then, set $u_{l,k}$ equal to zero with probability

$$\tilde{p}_{l,k} = \frac{p\mathbb{1}_{\{n_{l,k}=0\}}}{p\mathbb{1}_{\{n_{l,k}=0\}} + (1-p)B(s_{1,l}^*, s_{2,l}^*)/B(s_{1,l}, s_{2,l})},$$

   where $B(\cdot,\cdot)$ is the beta function.
   Otherwise, sample a new value for $u_{l,k}$ from a $\text{Beta}(s_{1,l}^*, s_{2,l}^*)$ distribution.
3. For each $j = 1, \dots, J$, sample S_j from a categorical distribution with K levels, where

$$\mathbb{P}\left[S_j = k \mid \cdots\right] = \pi_k \prod_{i=1}^{N_j}\left(\sum_{l=1}^{L}\omega_{l,k} f(y_{i,j}; \theta_l^*)\right).$$

4. For each $j = 1, \dots, J$ and $i = 1, \dots, N_j$, sample $M_{i,j}$ from a categorical distribution with L levels, where

$$\mathbb{P}\left[M_{i,j} = l \mid \cdots\right] = \omega_{l,S_j} f(y_{i,j} \mid \theta_l^*).$$

5. For each $l = 1, \dots, L$, sample θ_l^* from the corresponding full conditional distribution $p(\theta_l^* \mid \cdots) \propto h(\theta_l^*)\prod_{i,j:M_{i,j}=l} f(y_{i,j} \mid \theta_l^*)$, where h is the density of the measure H.
6. If p is random, sample a new value from a $\text{Beta}(b_1^0 + n_0, b_2^0 + K\cdot L - n_0)$ distribution, where $n_0 = \sum_{l=1}^{L}\sum_{k=1}^{K}\mathbb{1}_{\{u_{l,k}=0\}}$.

Algorithm 1 is general and can be easily adjusted to all the discussed beta stick-breaking priors. However, it is useful to remark on some technical issues. The truncation on which the blocked Gibbs sampler relies should not negatively affect the accuracy of the posterior estimates. This aspect becomes critical when considering heavy-tailed processes, such as the PYP. Thus, it is important to carefully assess the adequacy of the truncation level (for example, by keeping track of the highest allocated component throughout the MCMC iterations) and, possibly, increase the truncation parameters. Of course, this does not come without costs: allocating larger matrices induces higher computing time and larger memory allocation. Hence, the choice of the prior process should carefully balance good theoretical properties and computational efficiency. For this reason, and in light of Section 2.2.5, in the simulation study, we focus on a geCAM specification driven by a DP on the distributional weights and investigate the 2PBP and SKBP on the observational weights.

4. Illustrations

We apply the geCAM mixture model introduced in the previous section to univariate grouped data to investigate its posterior properties. First, we apply the proposed model to a synthetic dataset to analyze the impact on the accuracy of the posterior density estimate compared to a standard CAM prior. We consider a simple data-generating mechanism to better investigate the limitations of the CAM and the impact of more flexible SB priors. In the Supplementary Material, we report a second experiment with more complicated subpopulation distributions, which resulted in similar conclusions. Then, we examine its performance on a subset of the data from the Collaborative Perinatal Project study (Hardy, 2003), comprising 2313 observations divided into 12 groups.

All the experiments were run on an Ubuntu system with Intel Core i7-14700K processor, with 32 gigabytes of memory. The code used to perform the experiments, written both in `R` and `C++`, is openly available at the GitHub repository `Fradenti/geCAM`.

4.1. Synthetic data

We consider a simple setting comprising data generated from two possible distributions. Each distribution is Gaussian with a different mean, specifically $\mathrm{N}(-5, 1)$ and $\mathrm{N}(5, 1)$, and the probability of being assigned to each of the subpopulations is 0.5. Let us denote the true density functions of subpopulation $k = 1, 2$ with $p_k^*(y)$. Then, we sample $J \in \{2, 4, 6\}$ groups comprising $N_j = n \in \{10, 25, 50\}$ independent observations. Considering all the possible combinations between the values of J and n results in 9 simulation scenarios.

We set $a_k^D = b_k^D = 1$ for $k \geq 1$ considering a DP prior with unitary concentration parameter for the distributional weights, and study the effect of different processes on the observational ones. Specifically, we consider two instances of SB priors: the skip-breaking process $\mathrm{SKBP}(1, 1, p)$ and the two-parameter beta process $\mathrm{2PBP}(s, s)$. To assess the impact of the prior parameters, for the SKBP we both fix $p \in \{0, 0.25, 0.5, 0.75\}$ and consider a further hyperprior, $p \sim \mathrm{Beta}(1, 1)$; for the 2PBP, we vary $s \in \{0.1, 0.5, 1\}$.

Notice that the cases $p = 0$ and $s = 1$ correspond to a standard CAM prior, which will serve as a benchmark. Finally, to complete the prior specification, we assume Normal kernels, $f(y \mid \theta_l^*) = \phi(y \mid \mu_l^*, {\sigma^2}_l^*)$, and a conjugate Normal-inverse Gamma base measure, i.e., $\mu_l^* \mid {\sigma^2}_l^* \sim \mathrm{N}\left(m_0, {\sigma^2}_l^*/\kappa_0\right)$ and ${\sigma^2}_l^* \sim \mathrm{IG}(\gamma_0, \lambda_0)$.

We simulate 50 independent datasets, and on each of them, we estimate the mixture models running 10,000 MCMC iterations, discarding the first half as burn-in. Given the simplicity of the data-generating mechanism, we truncate the processes at $L = 20$ and $K = 10$. Notice that the Gibbs sampler to perform posterior inference under a 2PBP prior can be obtained by setting $p = 0$ in Algorithm 1.

We compare the models by focusing on the accuracy of the posterior density estimate. Denote with T the number of MCMC iterations and consider a generic group j. For each of the 50 replications, we compute the point-wise posterior mean of the group-specific density on a grid of approximately 2000 points for $y \in [-10, 10]$ as

$$\hat{p}_j(y) = \sum_{t=1}^{T} \frac{1}{T} \sum_{l=1}^{L} \omega_{l,S_j^{(t)}}^{(t)} \phi(y \mid (\mu_l^*, {\sigma^2}_l^*)^{(t)}). \tag{11}$$

Thanks to the simplicity of the simulation scenario, the distributional partition is always correctly recovered, hence $\hat{p}_j$ is an estimate of the true corresponding distribution $p_k^* = p_{S_j}^*$. Moreover, the quantity in (11) sums over all the observational mixture components, hence it does not suffer from the label-switching problem that characterizes conditional algorithms. Figure 4 shows examples of such quantity computed for the two subpopulations (displayed by row) in the first simulation scenario ($n = 10$, $J = 2$). Each panel corresponds to a different prior process: the standard CAM, the skip-breaking process SKBP$(1, 1, 0.5)$ and the two-parameter beta process 2PBP$(0.5, 0.5)$. We report the posterior density estimate $\hat{p}_j(y)$ computed on the 50 datasets (gray lines), their overall average (blue line), and the ground truth (black line). The density estimate relative to the CAM clearly shows the presence of a spurious mode, which should not be detected. The issue is greatly alleviated by the 2PBP and, even more, by the SKBP. This behavior provides an intuition for the general results discussed in the next paragraph, obtained using formal measures of fit. Plots displaying all the estimated posterior densities under every scenario are reported in the Supplementary Material.

To formally assess the accuracy of the posterior estimate in (11) for the various models, we evaluate three different metrics of discrepancy between the true density of the k-th subpopulation p_k^*, and the posterior density estimate $\hat{p}_j$ of the corresponding j-th group. Specifically, we compute the Jensen-Shannon divergence (JS; Lin, 1991), the total variation (TV) distance, and the Kullback-Leibler (KL) divergence. Although they are all measures of discrepancy between distributions, they highlight different aspects. The JS divergence (defined as $\mathrm{JSD}_j(p_k^* \mid\mid \hat{p}_j) = 0.5\left(\mathrm{KL}(p_k^* \mid\mid R) + \mathrm{KL}(\hat{p}_j \mid\mid R)\right)$, with $R = 0.5(\hat{p}_j + p_k^*)$) and TV distance are symmetric measures, hence they provide an intuition about the overall discrepancy between the true and estimated density. Differently, the KL divergence weights unevenly different types of errors. Specifically, we consider the

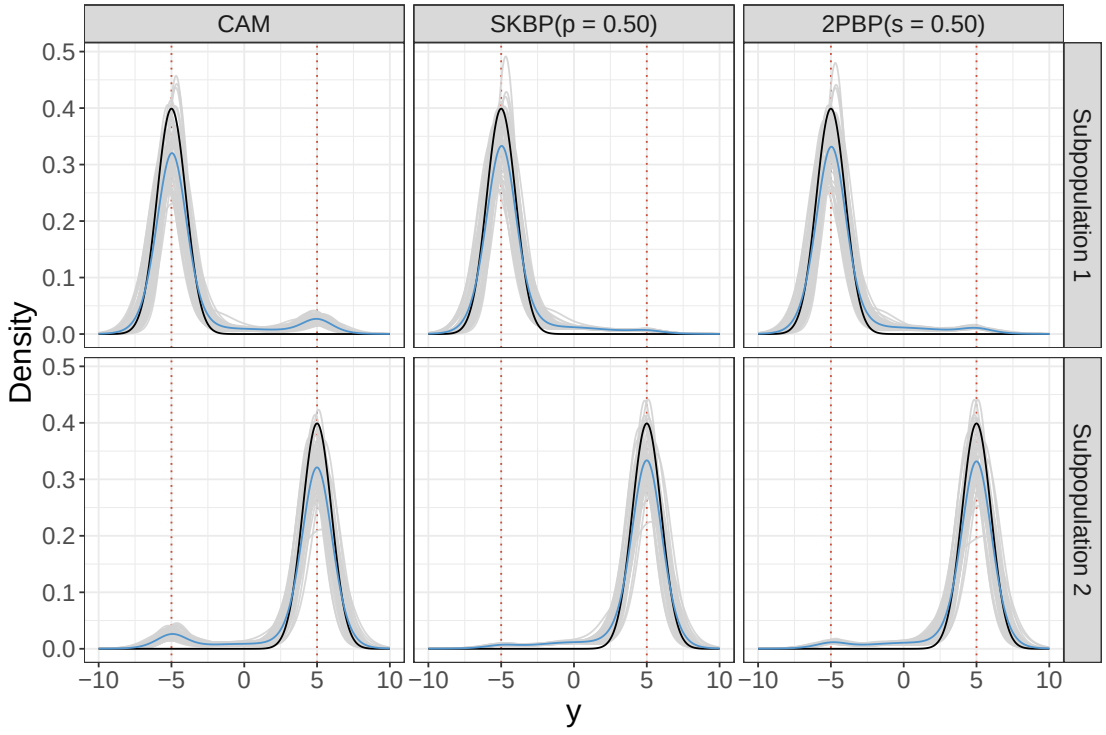

Figure 4: Posterior density estimate of group 1, when $J = 2$ and $n = 10$. The gray curves show the mean of the posterior densities in the 50 replications, and their pointwise average is marked in blue. The black lines display the true density. The left panels correspond to the CAM, the central and right panels correspond to geCAMs embedded with SKBP$(1, 1, 0.5)$ and 2PBP$(0.5, 0.5)$, respectively. The rows reports the results by subpopulations.

Kullback-Leibler divergence $\mathrm{KL}(p_k^* \,||\, \hat{p}_j)$, defined as

$$\mathrm{KL}(p_k^* \,||\, \hat{p}_j) = \int p_k^*(y) \log \frac{p_k^*(y)}{\hat{p}_j(y)} dy. \tag{12}$$

The divergence in (12) is carefully chosen so it quickly grows in all regions of the support in which $p_k^*(y)$ is near zero *unless* $\hat{p}_j(y)$ *is also close to zero* (Bishop, 2006). This characteristic is particularly compelling when evaluating mixture models, as it can be employed to emphasize the creation of spurious modes. The results for the JS divergence and TV distance criteria are reported in Figures S6 and S7 of the Supplementary Material. The discrepancy between the estimate and the ground truth is consistently larger for the CAM compared to more general specifications, although the difference is small. Hence, these plots suggest that the CAM has slightly more difficulties in estimating the posterior density compared to other formulations. Still, they do not provide an intuition about *why* this happens. We get a more complete picture by additionally looking at the KL divergence, presented in Figures 5 and 6. These figures show the results for the

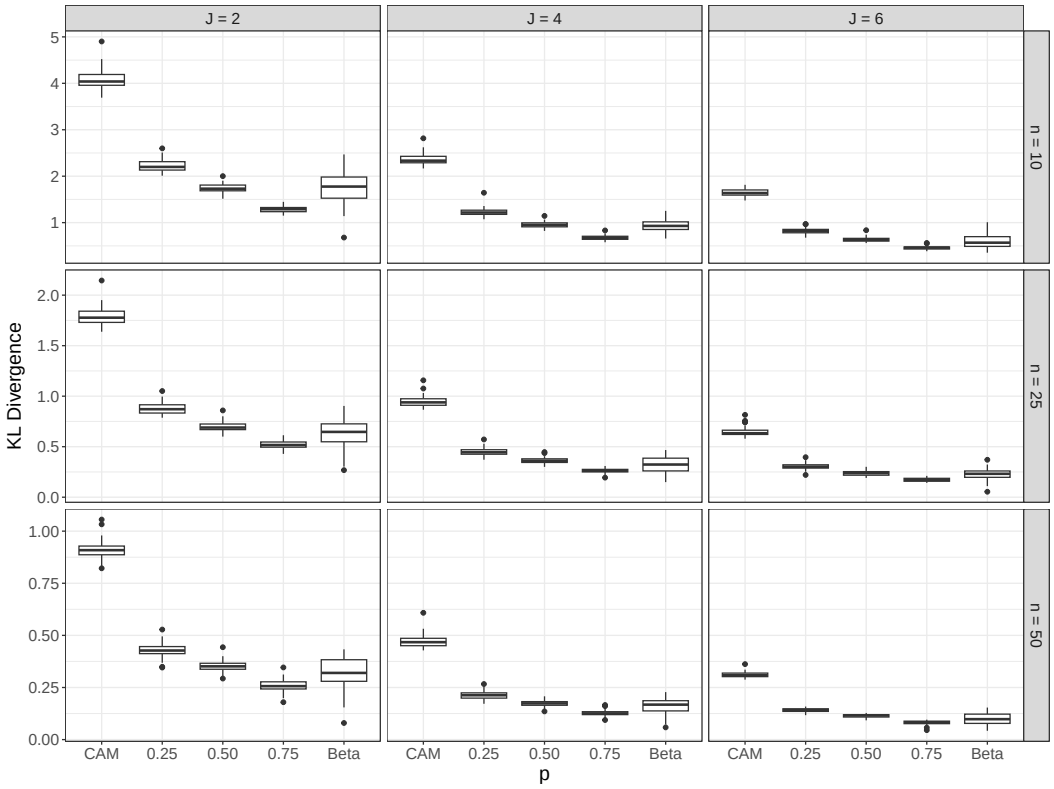

Figure 5: Distributions of the KL divergence over the 50 replications for a geCAM embedded with a SKBP$(1, 1, p)$ process, for different specifications of p. The standard CAM is obtained for $p = 0$, while the rightmost boxplot of each panel corresponds to a random $p \sim \text{Beta}(1, 1)$. Each panel corresponds to a simulation scenario.

SKBP and the 2PBP, respectively. Each panel corresponds to a simulation scenario, and we study the distribution of the KL divergence for varying prior parameters. We computed the KL divergences separately for each subpopulation and averaged them to obtain a single measure. Each boxplot represents the distribution of these averages over the 50 replications. Here, the performances of the CAM get significantly worse. Because of the peculiarities of the KL divergence, we have a confirmation that the cause of the CAM's less accurate posterior density is the creation of spurious modes (i.e., it estimates a large density in regions where the true density is actually close to zero). We remark that the two subpopulations are designed to be very simple and well-separated. Nevertheless, even in this straightforward case, the accuracy of the CAM is affected by its rigid structure. The results for the 2PBP show an improvement, especially for $s = 0.10$. However, the SKBP has the best performance overall, even for small p.

Lastly, to assess the accuracy of the proposed sampling algorithm, we investigate the approximation arising from truncating the process to a finite number of components.

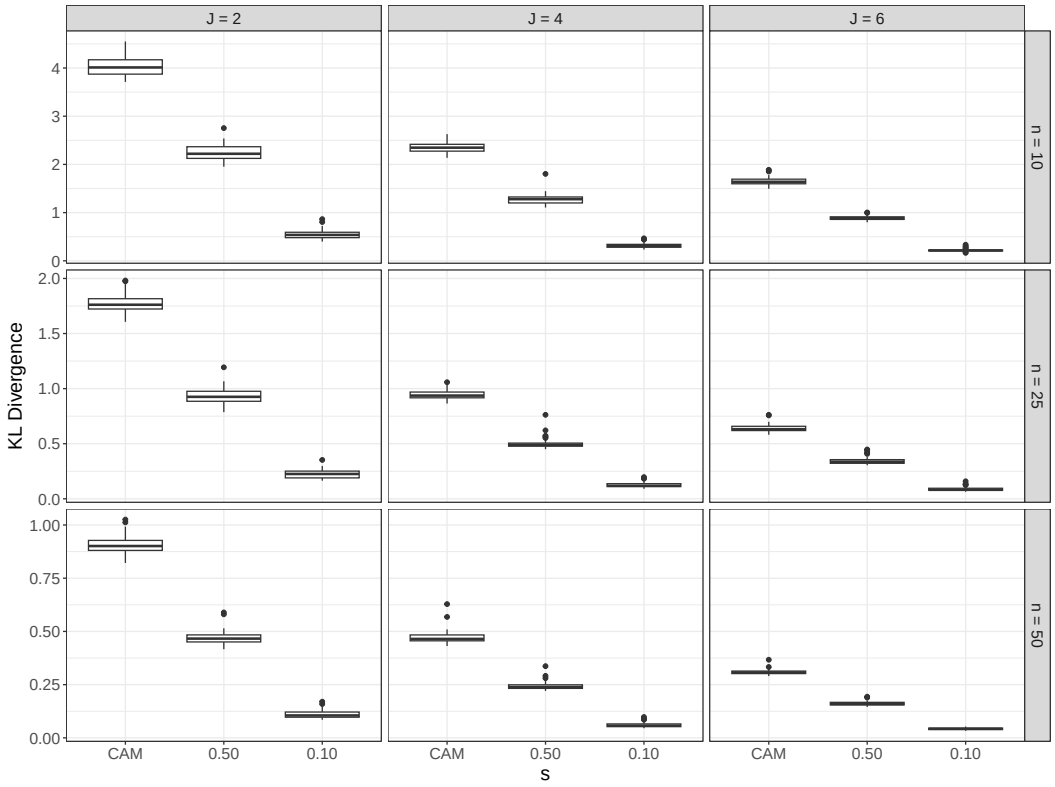

Figure 6: Distributions of the KL divergence over the 50 replications for a geCAM embedded with a 2PBP(s, s), for different specifications of s. The standard CAM is obtained for $s = 1$. Each panel corresponds to a simulation scenario.

Adopting heavy tails for the mixing weights can lead to allocating atoms that appear later in the sequence, as already discussed in Section 2.2 for the PYP case and at the end of Section 3. Hence, it is important to evaluate if truncating the process is feasible without causing a loss of accuracy. To monitor this aspect, we compute the number of estimated distributional (K^*) and observational (L^*) clusters, and the maximum values of the distributional $(\bar{S})$ and observational $(\bar{M})$ labels. Table 1 reports the average and the standard deviation of these metrics computed across the 50 replications for the extreme cases $n = 10$, $J = 2$, and $n = 50$, $J = 6$. The complete tables are available as Supplementary Material. The estimates of the number of clusters are satisfactory and consistent across the scenarios; the maxima of the cluster labels are never close to the upper bounds L and K, suggesting that such truncations are not affecting the posterior estimate of the process.

4.2. Application to the Collaborative Perinatal Project dataset

The Collaborative Perinatal Project (CPP) is a broad epidemiological study conducted in the U.S. from 1959 to 1974 to investigate complications during pregnancy and birth

SKBP(1, 1, p)						
p	n	J	K^*	L^*	$\bar{S}$	$\bar{M}$
0	10	2	2.000 (0.000)	2.564 (0.210)	3.453 (0.098)	2.998 (0.345)
0.25	10	2	2.000 (0.000)	2.669 (0.235)	3.488 (0.101)	3.919 (0.499)
0.5	10	2	2.000 (0.000)	2.729 (0.244)	3.488 (0.163)	5.623 (0.716)
0.75	10	2	2.000 (0.000)	2.698 (0.234)	3.453 (0.143)	9.011 (1.379)
Beta	10	2	2.000 (0.000)	2.677 (0.242)	3.471 (0.140)	6.770 (1.717)
0	50	6	2.053 (0.014)	2.922 (0.289)	2.494 (0.291)	3.371 (0.378)
0.25	50	6	2.067 (0.019)	3.094 (0.340)	2.645 (0.325)	4.371 (0.970)
0.5	50	6	2.043 (0.020)	3.397 (0.349)	2.625 (0.457)	7.355 (1.961)
0.75	50	6	2.021 (0.010)	3.281 (0.299)	2.528 (0.323)	10.349 (1.750)
Beta	50	6	2.030 (0.018)	3.141 (0.335)	2.551 (0.387)	9.395 (3.716)
2PBP(s, s)						
s	n	J	K^*	L^*	$\bar{S}$	$\bar{M}$
1.0	10	2	2.000 (0.000)	2.558 (0.207)	3.463 (0.094)	2.993 (0.340)
0.5	10	2	2.000 (0.000)	2.383 (0.175)	3.460 (0.096)	2.874 (0.330)
0.2	10	2	2.000 (0.000)	2.135 (0.087)	3.464 (0.081)	3.022 (0.607)
1.0	50	6	2.058 (0.013)	2.909 (0.270)	2.581 (0.321)	3.347 (0.371)
0.5	50	6	2.171 (0.030)	2.596 (0.282)	2.798 (0.232)	2.959 (0.420)
0.2	50	6	2.305 (0.179)	2.493 (0.290)	3.001 (0.326)	4.980 (1.805)

Table 1: Average and standard deviation (across the 50 replications) of the number of estimated distributional and observational clusters (K^*, L^*), and of the maximum value of the distributional and observational labels ($\bar{S}, \bar{M}$).

outcomes. We consider a subset of this large data repository, which comprises the gestational age and weight at birth of 2313 newborns collected from 12 different hospitals. Moreover, to assess the effect of possible risk factors on the health of babies, it contains information about the mothers' smoking habits and the concentration level in maternal serum of DDE, a toxic chemical compound of the pesticide DDT, known to adversely impact the health of the babies (Longnecker, Klebanoff, Zhou, & Brock, 2001). The data are available in the R package `BNPmix` (Corradin, Canale, & Nipoti, 2021) and have been analyzed, for example, in Canale et al. (2022) and Lijoi et al. (2023).

We focus on the distribution of the weight at birth for both smoking and non-smoking women, and we fit a Gaussian mixture driven by a geCAM prior embedded with a SKBP$(1, 1, 0.5)$ for the observational weights, and a DP with concentration parameter $\alpha = 1$ on the distributional ones. Again, we consider conjugate normal inverse-gamma prior distributions, and we set m_0 equal to the overall sample mean, $\kappa_0 = 0.1$, $\gamma_0 = 1$, and $\lambda_0 = 4$. We truncate the processes at $L = 50$ and $K = 30$. We run the MCMC algorithm outlined in Section 3 for 20,000 iterations, discarding the first half as burn-in. For comparison, we also fit the CAM to the data, under the same conditions. In the Supplementary Material, we report the MCMC traceplots for the quantities $K^*, L^*, \bar{S}$, and $\bar{M}$, as defined in the previous section. The plots suggest that, for both models, the truncation levels do not affect the accuracy of the estimates since the number of

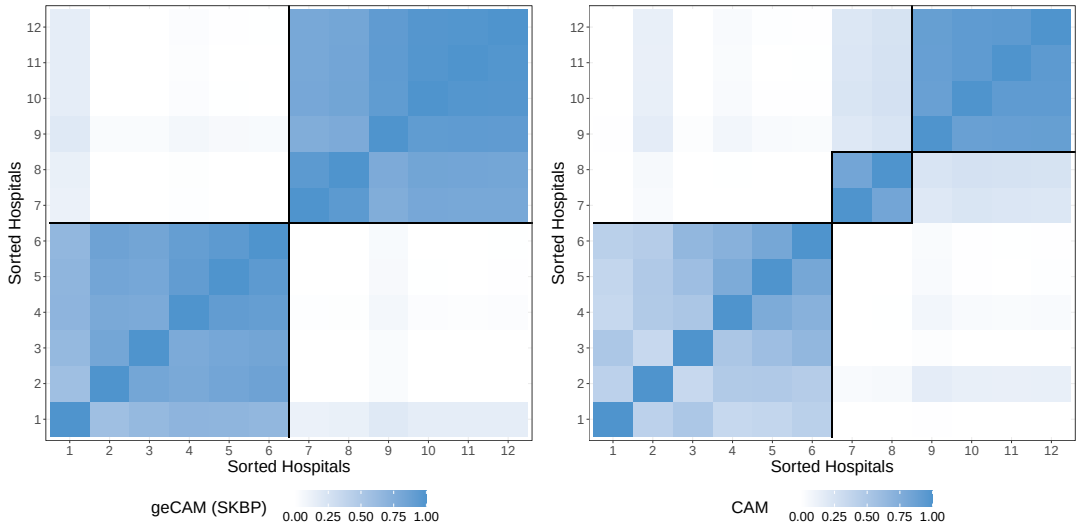

Figure 7: Posterior similarity matrices computed across the 12 groups under the geCAM with SKBP on the observational weights (left panel), and the standard CAM (right panel).

occupied clusters is always smaller than the upper bound of available components. Each run took approximately 3 minutes to complete.

We computed a point estimate of the partitions of observations and groups by minimizing the variation of information criterion (Dahl, Johnson, & Müller, 2022; Wade & Ghahramani, 2018). We estimated two distributional clusters (DCs) using the geCAM, while the CAM found three DCs, isolating two hospitals. The posterior similarity matrices between the 12 groups computed under the two model specifications are displayed in Figure 7. While the one obtained with the geCAM clearly shows the presence of two well-distinct clusters of hospitals, the CAM's partition is more nuanced.

In Figure 8 we show the histograms of each hospital, together with the posterior density estimates. The estimates under the CAM are shown with solid blue lines (in three different shades, corresponding to the three estimated DCs), while the ones obtained with geCAM are displayed with dashed red and orange lines (reflecting the two estimated DCs). It is interesting to notice that the differences in the CAM's clustering and density estimation are driven by an additional component centered around 50. While the presence of such a component allows a nice fit of the data in the third DC (in dark blue, very similar in shape to SKBP's second DC), it seems to strongly influence the other two DCs, over-inflating the density when not necessary (see, for example, hospitals 1, 3, and 5). This behavior is further highlighted in Figure 9, which shows the posterior density estimate of each group colored by DC for the geCAM and the CAM. Focusing on the density estimates obtained with the geCAM, we see that, despite the great overlap, the two distributions present different skewness, with DC2 containing the hospitals where mothers gave birth to lighter babies, potentially indicating centers suitable for high-risk

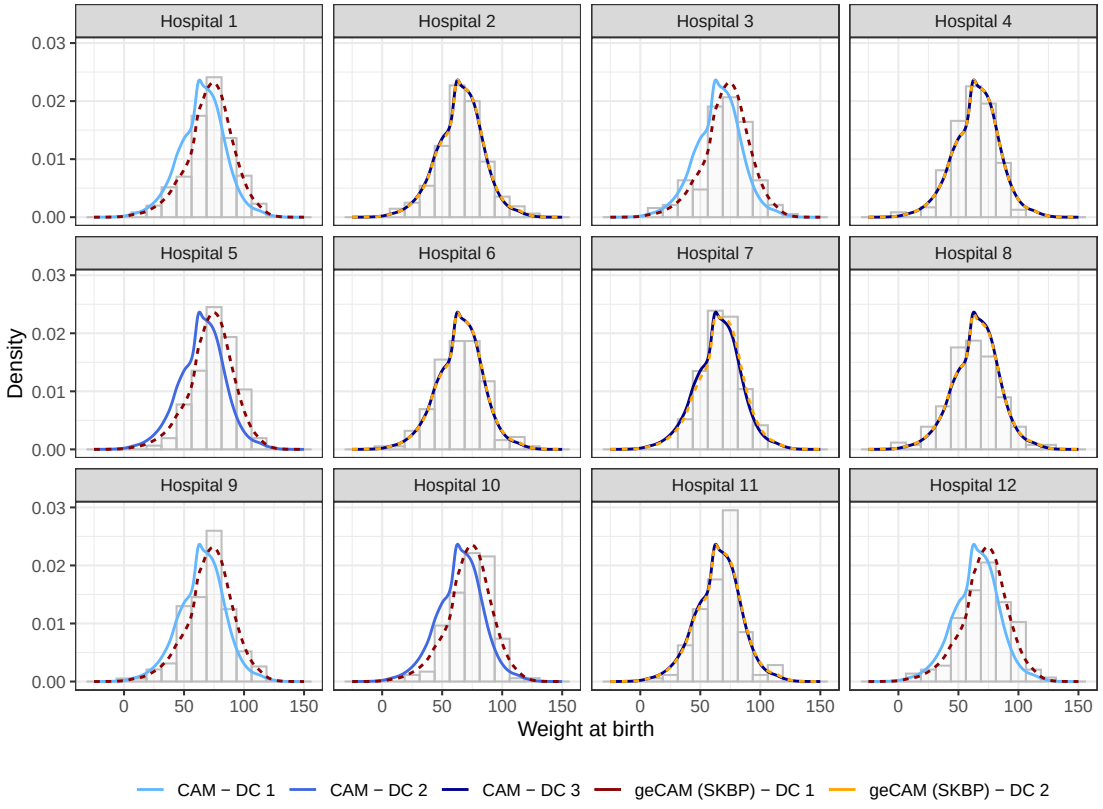

Figure 8: Posterior density estimates for the 12 groups (hospitals). The estimates under the CAM are shown with solid blue lines, while the ones obtained with geCAM are displayed with dashed red and orange lines.

pregnancies.

It is interesting to compare the distribution of the risk factor variables within the two DCs, as displayed in Figure 10. Both the gestational age and the concentration of DDE are linked to the weight at birth, and indeed, their distributions are quite different within the two clusters. Remarkably, DC2 is associated with shorter pregnancies and a higher dosage of DDE, hence, again, suggesting that this cluster comprises high-risk pregnancies. Lastly, the proportion of smokers in DC2 is almost 60%, while it decreases to 52.8% in DC1.

5. Conclusions

In this paper, we introduced the framework of the generalized CAM prior, designed to enhance the flexibility of the model of Denti, Camerlenghi, et al. (2023). In particular, we investigated how different SB specifications impact the prior correlation between random probability measures generated by the CAM. In our study, we demonstrated

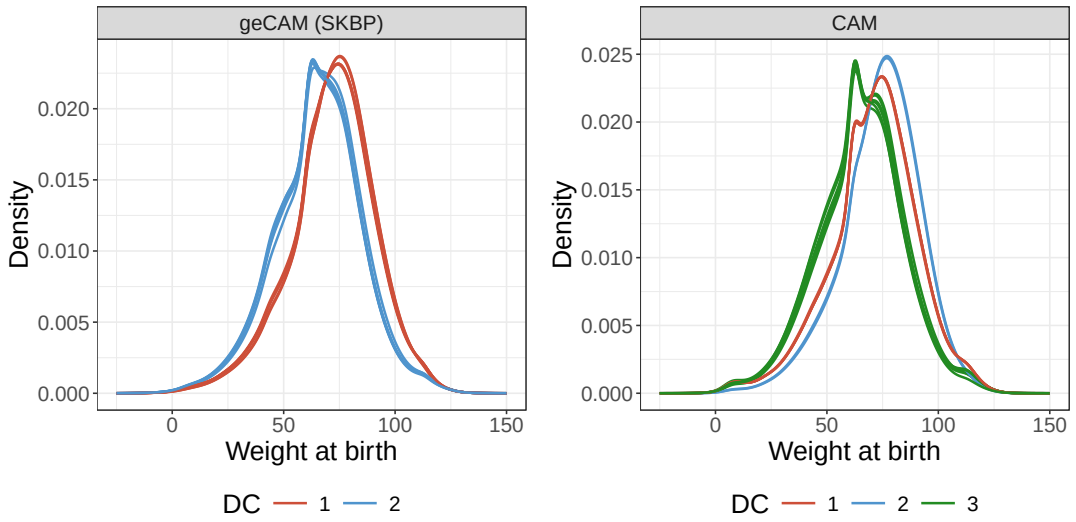

Figure 9: Posterior density estimates for each group under the geCAM with SKBP observational prior (left) and standard CAM (right).

that adopting a process with heavier tails than the Dirichlet process, such as the Pitman-Yor or the two-parameter beta process, enables the mass to be distributed over a larger number of atoms. This promotes more heterogeneous group-specific mixing measures and mitigates the strong prior correlations imposed by the CAM. Additionally, we explored an alternative approach using a spike-and-slab hierarchical prior to explicitly introduce group-specific atoms. We showed how the proposed prior can be conveniently used as a mixing measure in nonparametric nested mixture models, developing an efficient blocked Gibbs sampler for conducting posterior inference. On simulated and real data we displayed how the geCAM mitigates the original "excess of borrowed information" across groups, leading to more accurate posterior density estimates. Hence, the proposed framework is preferable to the model of Denti, Camerlenghi, et al. (2023) whenever there is no prior information about the similarity between the group-specific distributions.

The results of this work prove how different SB priors affect the flexibility of the CAM. However, more elaborate SB constructions can be leveraged to extend the prior in other possible directions. For example, covariate-dependent weights such as the probit SB process (Dunson & Rodríguez, 2011) can modulate the correlations between nested random measures according to additional variables. Another interesting direction is investigating the usage of shrinkage priors other than the horseshoe for modeling the "stick" variables in the SB process. Alternatively, clustering these random variables would rank the atoms of each random measure into tiers of relevance, similarly to Denti, Azevedo, et al. (2023). Alternatively, one could focus on solutions devised to impact the order in which the atoms appear in the sequence, proposing permutations to disrupt the correspondence with the SB weights. For instance, the approach proposed in Griffin and Steel (2006) could be leveraged to address this issue.

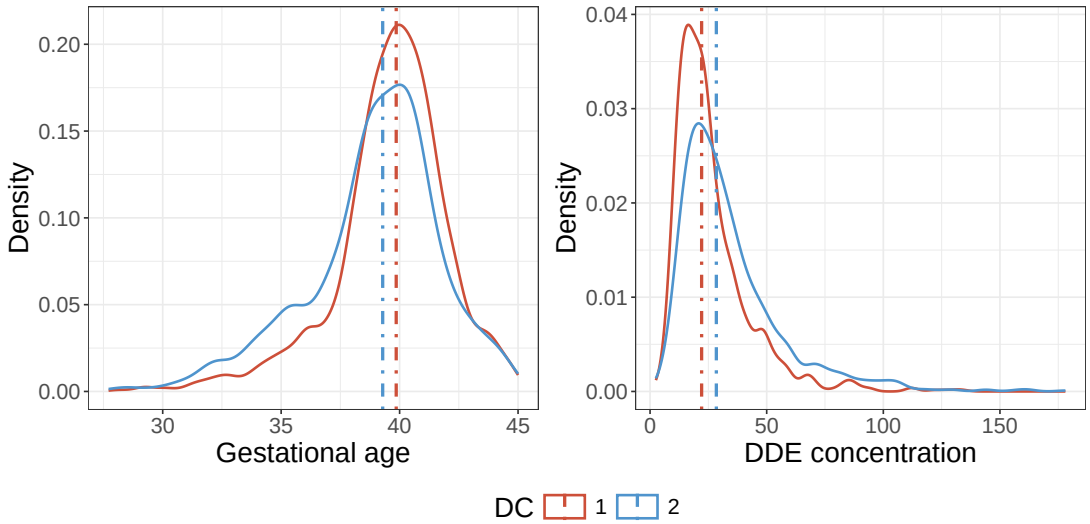

Figure 10: Kernel density estimates of the gestational age (left) and the DDE concentration (right) in the two DCs obtained with the geCAM (DC1: red, DC2: blue). The vertical lines denote the medians of the two distributions.

Another aspect worth investigating is the derivation of efficient variational inference algorithms for posterior inference (Blei, Kucukelbir, & McAuliffe, 2017; D'Angelo & Denti, 2024), which are especially useful when dealing with large datasets.

References

Beraha, M., Guglielmi, A., & Quintana, F. A. (2021). The Semi-Hierarchical Dirichlet Process and Its Application to Clustering Homogeneous Distributions. *Bayesian Analysis*, *16*(4), 1187–1219. doi: 10.1214/21-ba1278

Bi, D., & Ji, Y. (2023). A Class of Dependent Random Distributions Based on Atom Skipping. *Arxiv preprint, arXiv:2304.14954*.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning* (Vol. 4). Springer, New York.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. doi: 10.1080/01621459.2017.1285773

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., & Rodríguez, A. (2019). Latent Nested Nonparametric Priors (With Discussion). *Bayesian Analysis*, *14*(4), 1303–1356. doi: 10.1214/19-ba1169

Canale, A., Corradin, R., & Nipoti, B. (2022). Importance Conditional Sampling for Pitman–Yor Mixtures. *Statistics and Computing*, *32*(3), 1–40. doi: 10.1007/s11222-022-10096-0

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The Horseshoe Estimator for Sparse Signals. *Biometrika*, *97*(2), 465–480. doi: 10.1093/biomet/asq017

Corradin, R., Canale, A., & Nipoti, B. (2021). BNPmix: An R Package for Bayesian Nonparametric Modeling via Pitman-Yor Mixtures. *Journal of Statistical Software*, *100*, 1–33. doi: 10.18637/jss.v100.i15

Dahl, D. B., Johnson, D. J., & Müller, P. (2022, 10). Search Algorithms and Loss Functions for Bayesian Clustering. *Journal of Computational and Graphical Statistics*, *31*, 1189-1201. doi: 10.1080/10618600.2022.2069779

D'Angelo, L., Canale, A., Yu, Z., & Guindani, M. (2023). Bayesian Nonparametric Analysis for the Detection of Spikes in Noisy Calcium Imaging Data. *Biometrics*, *79*(2), 1370–1382. doi: 10.1111/biom.13626

D'Angelo, L., & Denti, F. (2024). A Finite-Infinite Shared Atoms Nested Model for the Bayesian Analysis of Large Grouped Data Sets. *Bayesian Analysis*, *In press*(), 1–34. doi: 10.1214/24-BA1458

Denti, F., Azevedo, R., Lo, C., Wheeler, D., Gandhi, S., Guindani, M., & Shahbaba, B. (2023). A horseshoe mixture model for bayesian screening with an application to light sheet fluorescence microscopy in brain imaging. *The Annals of Applied Statistics*, *17*, 2639–2658. doi: 10.1214/23-AOAS1736

Denti, F., Camerlenghi, F., Guindani, M., & Mira, A. (2023). A Common Atoms Model for the Bayesian Nonparametric Analysis of Nested Data. *Journal of the American Statistical Association*, *118*(541), 405–416. doi: 10.1080/01621459.2021.1933499

Dunson, D. B., & Rodríguez, A. (2011, 3). Nonparametric bayesian models through probit stick-breaking processes. *Bayesian Analysis*, *6*. doi: 10.1214/11-BA605

Ghosal, S., & van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press. doi: 10.1017/9781139029834

Graziani, R., Guindani, M., & Thall, P. F. (2015). Bayesian Nonparametric Estimation

of Targeted Agent Effects on Biomarker Change to Predict Clinical Outcome. *Biometrics*, *71*(1), 188–197. doi: 10.1111/biom.12250

Griffin, J. E., & Steel, M. F. (2006). Order-Based Dependent Dirichlet Processes. *Journal of the American Statistical Association*, *101*(473), 179–194. doi: 10.1198/016214505000000727

Hardy, J. B. (2003). The Collaborative Perinatal Project: Lessons and Legacy. *Annals of Epidemiology*, *13*(5), 303–311.

Ishwaran, H., & James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, *96*(453), 161–173. doi: 10.1198/016214501750332758

Ishwaran, H., & Zarepour, M. (2000). Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-Parameter Process Hierarchical Models. *Biometrika*, *87*(2), 371–390. doi: 10.1093/biomet/87.2.371

Lijoi, A., Prünster, I., & Rebaudo, G. (2023). Flexible Clustering Via Hidden Hierarchical Dirichlet Priors. *Scandinavian Journal of Statistics*, *50*(1), 213–234. doi: https://doi.org/10.1111/sjos.12578

Lin, J. (1991). Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, *37*, 145–151. doi: 10.1109/18.61115

Longnecker, M. P., Klebanoff, M. A., Zhou, H., & Brock, J. W. (2001, 7). Association Between Maternal Serum Concentration of the DDT Metabolite DDE and Preterm and Small-for-Gestational-Age Babies at Birth. *The Lancet*, *358*, 110–114. doi: 10.1016/s0140-6736(01)05329-6

Pitman, J., & Yor, M. (1997). The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator. *Annals of Probability*, *25*(2), 855–900. doi: 10.1214/aop/1024404422

Rodríguez, A., & Dunson, D. B. (2014). Functional Clustering in Nested Designs: Modeling Variability in Reproductive Epidemiology Studies. *Annals of Applied Statistics*, *8*(3), 1416–1442. doi: 10.1214/14-AOAS751

Rodríguez, A., Dunson, D. B., & Gelfand, A. E. (2008). The Nested Dirichlet Process. *Journal of the American Statistical Association*, *103*(483), 1131–1154. doi: 10.1198/016214508000000553

Sethuraman, J. (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, *4*, 639–650.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581. doi: 10.1198/016214506000000302

Wade, S., & Ghahramani, Z. (2018, 6). Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion). *Bayesian Analysis*, *13*, 559–626. doi: 10.1214/17-BA1073

Zeng, C., Miller, J. W., & Duan, L. L. (2023). Consistent model-based clustering using the quasi-bernoulli stick-breaking process. *Journal of Machine Learning Research*, *24*(153), 1–32.

Zuanetti, D. A., Müller, P., Zhu, Y., Yang, S., & Ji, Y. (2018). Clustering Distributions With the Marginalized Nested Dirichlet Process. *Biometrics*, *74*(2), 584–594.

Supplementary Material

A. Proofs of the theoretical results

In all the derivations, we assume that the vectors of distributional and observational weights are such that $\pi_k \in [0,1]$, $\omega_{l,k} \in [0,1]$ for all $l, k \geq 1$ and $\sum_{k\geq 1} \pi_k = 1$, $\sum_{l\geq 1} \omega_{l,k} = 1$ for all $k \geq 1$.

A.1. Proof of Proposition 2.1

Suppose that the G_j's are defined on a Polish space $(\mathbb{X}, \mathcal{X})$ and consider $A \in \mathcal{X}$. We also recall that $G_j, G_{j'} \overset{i.i.d.}{\sim} Q$, with $Q = \sum_{k\geq 1} \pi_k \delta_{G_k^*}$ and $G_k^* = \sum_{l\geq 1} \omega_{l,k} \delta_{\theta_l^*}$. Also, $\theta_l^* \overset{i.i.d.}{\sim} H$. To compute the correlation, we first derive the covariance between random measures:

$$\begin{aligned} Cov(G_j(A), G_{j'}(A)) &= \mathbb{E}\left[G_j(A) \cdot G_{j'}(A)\right] - \mathbb{E}\left[G_j(A)\right]\mathbb{E}\left[G_{j'}(A)\right] \\ &= \mathbb{E}\left[G_j(A) G_{j'}(A)\right] - H(A)^2. \end{aligned}$$

Let $q_1 = \mathbb{P}\left[G_j = G_{j'}\right] = \sum_{k\geq 1} \mathbb{P}\left[G_j = G_k^*, G_{j'} = G_k^*\right]$. Then, the first term can be rewritten as:

$$\mathbb{E}\left[G_j(A) G_{j'}(A)\right] = q_1 \mathbb{E}\left[G_k^*(A)^2\right] + (1 - q_1)\mathbb{E}\left[G_k^*(A) \cdot G_{k'}^*(A)\right].$$

Then, we have, for fixed k, k' such that $k \neq k'$

$$\begin{aligned} \mathbb{E}\left[G_k^*(A)^2\right] &= \mathbb{E}\left[\sum_{l\geq 1} \omega_{l,k}^2 \delta_{\theta_l^*}(A)\right] + \mathbb{E}\left[\sum_{l\geq 1}\sum_{l\neq l'} \omega_{l,k}\omega_{l',k} \delta_{\theta_l^*}(A)\delta_{\theta_{l'}^*}(A)\right] \\ &= \sum_{l\geq 1} \mathbb{E}\left[\omega_{l,k}^2\right] H(A) + \left(1 - \sum_{l\geq 1} \mathbb{E}\left[\omega_{l,k}^2\right]\right) H(A)^2 \\ &= \left(\sum_{l\geq 1} \mathbb{E}\left[\omega_{l,k}^2\right]\right) H(A)(1 - H(A)) + H(A)^2, \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}\left[G_k^*(A) G_{k'}^*(A)\right] &= \mathbb{E}\left[\sum_{l\geq 1} \omega_{l,k}\omega_{l,k'} \delta_{\theta_l^*}(A)\right] + \mathbb{E}\left[\sum_{l\geq 1}\sum_{l\neq l'} \omega_{l,k}\omega_{l',j'} \delta_{\theta_l^*}(A)\delta_{\theta_{l'}^*}(A)\right] \\ &= \sum_{l\geq 1} \mathbb{E}\left[\omega_{l,k}\right]\mathbb{E}\left[\omega_{l,k'}\right] H(A) + \left(1 - \sum_{l\geq 1} \mathbb{E}\left[\omega_{l,k}\right]\mathbb{E}\left[\omega_{l,k'}\right]\right) H(A)^2 \\ &= \left(\sum_{l\geq 1} \mathbb{E}\left[\omega_{l,k}\right]^2\right) H(A)(1 - H(A)) + H(A)^2. \end{aligned}$$

In the second derivation, we leveraged the fact that the first equality holds true in particular when $A = \mathbf{X}$, as noted in Denti, Camerlenghi, et al. (2023). In this specific case, $\mathbb{E}\left[G_k^*(\mathbb{X})G_{k'}^*(\mathbb{X})\right] = 1 = \mathbb{E}\left[\sum_{l\geq 1}\omega_{l,k}\omega_{l,k'}\right] + \mathbb{E}\left[\sum_{l\geq 1}\sum_{l\neq l'}\omega_{l,k}\omega_{l',k'}\right]$, so $1 - \mathbb{E}\left[\sum_{l\geq 1}\omega_{l,k}\omega_{l,k'}\right] = \mathbb{E}\left[\sum_{l\geq 1}\sum_{l\neq l'}\omega_{l,k}\omega_{l',k'}\right]$.

Now, let $\xi_1 = \sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}\right]^2$ and $\xi_2 = \sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}^2\right]$. Combining the previous two results, we obtain

$$\begin{aligned}\mathbb{E}\left[G_j(A)G_{j'}(A)\right] &= \left[q_1\xi_2 + (1-q_1)\xi_1\right]H(A)(1-H(A)) + H(A)^2, \quad \text{and}\\ Cov(G_j(A), G_{j'}(A)) &= \left[q_1\xi_2 + (1-q_1)\xi_1\right]H(A)(1-H(A)).\end{aligned}$$

The variance is immediately obtained by letting $j = j'$ in the covariance (hence $q_1 = 1$):

$$Var(G_j(A)) = \xi_2 H(A)(1-H(A)).$$

Therefore, we can conclude that

$$\begin{aligned}Corr(G_j(A), G_{j'}(A)) &= \frac{Cov(G_j(A), G_{j'}(A))}{Var(G_j(A))}\\ &= \frac{\left[q_1\xi_2 + (1-q_1)\xi_1\right]H(A)(1-H(A))}{\xi_2 H(A)(1-H(A))}\\ &= q_1 + (1-q_1)\xi_1/\xi_2\\ &= 1 - (1-q_1)(1-q_2),\end{aligned}$$

where we denoted $q_2 = \xi_1/\xi_2$. Note that all the terms involving $H(A)$ cancel out, making the correlation invariant w.r.t. the set A. Finally, it is immediate to verify that the term $(1-q_1) \in [0,1]$, since q_1 is a probability. Regarding the second term, we can show that $\mathbb{E}\left[\omega_{l,k}\right]^2 \leq \mathbb{E}\left[\omega_{l,k}^2\right]$ by Jensen inequality. Summing the elements of both sides leads to $\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}\right]^2 \leq \sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}^2\right]$. Therefore, in principle, the correlation is always non-negative, i.e., $\rho_{j,j'} \in (0,1)$.

We remark that this result does not guarantee that the correlation can attain every value on $(0,1)$ regardless of the weight specification. Indeed, even for standard choices of stochastic processes (e.g., DP, PYP, 2PBP), while there exists a combination of parameters for which $q_1 \to 0$ and $q_1 \to 1$, it is important to highlight that $(1-q_2)$ can potentially lie on a smaller subset, limiting the range of the correlation (this happens, for example, in the case of a DP, where $\rho_{j,j'} \in (0.5, 1)$).

A.2. Proof of Corollary 2.2

Let $y_{i,j} \mid G_j \sim G_j$ and $y_{i',j'} \mid G_{j'} \sim G_{j'}$ be two observations coming from two probability measures both sampled from Q with $j \neq j'$. Also, denote with $q_1 = \mathbb{P}\left[G_j = G_{j'}\right]$. We

have that

$$\begin{aligned}
\mathbb{P}\left[y_{i,j} = y_{i',j'}\right] &= \mathbb{E}\left[\mathbb{P}\left[y_{i,j} = y_{i',j'} \mid G_j, G_{j'}\right]\right] \\
&= \mathbb{E}\left[q_1 \mathbb{P}\left[y_{i,j} = y_{i',j'} \mid G_j = G_{j'}\right] + (1-q_1)\mathbb{P}\left[y_{i,j} = y_{i',j'} \mid G_j \neq G_{j'}\right]\right] \\
&= \mathbb{E}\left[q_1\left(\sum_{l\geq 1}\omega_{l,k}^2\right) + (1-q_1)\left(\sum_{l\geq 1}\omega_{l,k}\omega_{l,k'}\right)\right] \\
&= q_1\left(\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}^2\right]\right) + (1-q_1)\left(\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}\right]\mathbb{E}\left[\omega_{l,k'}\right]\right) \\
&= \left(\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}^2\right]\right)\left[q_1 + (1-q_1)\left(\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}\right]^2\right) / \left(\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}^2\right]\right)\right] \\
&= \left(\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}^2\right]\right)\left[1-(1-q_1)(1-q_2)\right] \\
&= \left(\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}^2\right]\right)\cdot \rho_{j,j'}.
\end{aligned}$$

A.3. Proof of Corollary 2.4

To compute $\rho_{j,j'}$ using a PYP(ϑ,σ) on the observational level, we need to compute $\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}^2\right]$ and $\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}\right]^2$ under the same process. The first quantity is immediate to compute as

$$\mathbb{E}\left[\sum_{l\geq 1}\omega_{l,k}^2\right] = \frac{1-\sigma}{1+\vartheta}, \tag{13}$$

since $\mathbb{E}\left[\sum_{l\geq 1}\omega_{l,k}^2\right]$ equals the probability of ties under a PYP (Ghosal & van der Vaart, 2017). The second quantity can be computed knowing that

$$\begin{aligned}
\mathbb{E}\left[\omega_{l,k}\right] &= \frac{1-\sigma}{\vartheta+1+(l-1)\sigma}\left(\prod_{q<l}\frac{\vartheta+q\sigma}{\vartheta+1+(q-1)\sigma}\right) \\
&= \frac{1-\sigma}{\sigma}\cdot\frac{(\tilde{\vartheta}+1)(\tilde{\vartheta}+2)\cdots(\tilde{\vartheta}+(l-1))}{(\hat{\vartheta})(\hat{\vartheta}+1)\cdots(\hat{\vartheta}+(l-1))} \\
&= \frac{1-\sigma}{\vartheta}\cdot\frac{\tilde{\vartheta}^{[l]}}{\hat{\vartheta}^{[l]}},
\end{aligned}$$

where we defined $\tilde{\vartheta} = \vartheta/\sigma$, $\hat{\vartheta} = (\vartheta+1)/\sigma$, and the notation $a^{[n]}$ denotes the ascending factorial (as in Ghosal & van der Vaart, 2017), i.e., $a^{[n]} = a(a+1)\cdots(a+n-1)$. Let

us call $\zeta_l(\sigma,\vartheta) = \tilde{\vartheta}^{[l]}/\hat{\vartheta}^{[l]}$. Then,

$$\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}\right]^2 = \left(\frac{1-\sigma}{\vartheta}\right)^2 \sum_{l\geq 1}\zeta_l^2(\sigma,\vartheta). \tag{14}$$

The term $\sum_{l\geq 1}\zeta_l^2(\sigma,\vartheta)$ can be easily evaluated numerically. Taking the ratios of the quantities in (13)-(14) gives the results.

A.4. Proof of Corollary 2.5

We want to compute $\rho_{j,j'}$ using a 2PBP(s_1,s_2) on the observational level. First, we need to compute $\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}^2\right]$ and $\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}\right]^2$. Under the 2PBP$(s_1,s_2)$, these quantities are easy to derive using the properties of beta random variables, as

$$\begin{aligned}
\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}\right]^2 &= \frac{s_1^2}{(s_1+s_2)^2}\sum_{l\geq 1}\left(\frac{s_2^2}{(s_1+s_2)^2}\right)^{l-1} = \frac{s_1}{s_1+2s_2},\\
\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}^2\right] &= \frac{s_1(s_1+1)}{(s_1+s_2)(s_1+s_2+1)}\sum_{l\geq 1}\left(\frac{s_2(s_2+1)}{(s_1+s_2)(s_1+s_2+1)}\right)^{l-1}\\
&= \frac{(s_1+1)}{(s_1+2s_2+1)}.
\end{aligned} \tag{15}$$

By considering the ratio of the quantities in (15), we obtain

$$q_2 = \frac{\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}\right]^2}{\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}^2\right]} = \frac{s_1(s_1+2s_2+1)}{(s_1+2s_2)(s_1+1)}. \tag{16}$$

A.5. Proof of Corollary 2.6

Now consider the following generic SKBP(s_1,s_2,p), with $s_1,s_2>0$ and $p\in(0,1)$, defined as

$$\omega_{l,k} = u_{l,k}\prod_{q<l}(1-u_{q,k}), \qquad u_{l,k}\sim p\delta_0 + (1-p)Beta(s_1,s_2).$$

The stick variables have the following moments:

$$\begin{aligned}
\mathbb{E}\left[u_{l,k}\right] &= \frac{(1-p)s_1}{s_1+s_2}, \qquad \mathbb{E}\left[(1-u_{l,k})\right] = \frac{ps_1+s_2}{s_1+s_2},\\
\mathbb{E}\left[u_{l,k}^2\right] &= \frac{(1-p)s_1(s_1+1)}{(s_1+s_2)(s_1+s_2+1)},\\
\mathbb{E}\left[(1-u_{l,k})^2\right] &= \frac{ps_1\left(s_1+2s_2+1\right)+s_2(s_2+1)}{(s_1+s_2)(s_1+s_2+1)}.
\end{aligned} \tag{17}$$

Then, let us denote with $f_B(x;s_1,s_2)$ the density of a $Beta(s_1,s_2)$ random variable. We can derive that $\mathbb{E}\left[\log(1-u_{l,k})\right] = (1-p)\int_0^1\log(1-x)f_B(x;s_1,s_2)dx = (1-p)\left(\psi(s_2)-\psi(s_1+s_2)\right)$,

where $\psi(\cdot)$ is the digamma function. We check that $\sum_{l\geq 1}\mathbb{E}\left[\log(1-u_{l,k})\right] = (1-p)\sum_{l\geq 1}\left(\psi(s_2)-\psi(s_1+s_2)\right) = -\infty$. Following Ishwaran and James (2001), this is a sufficient condition to state that SKBP(s_1,s_2,p) is a valid SB construction. An important quantity to calculate is $\sum_{l\geq 1}\left[\mathbb{E}\left[(1-u_{l,k})^2\right]\right]^l$. Since

$$\left|\mathbb{E}\left[(1-u_{l,k})^2\right]\right| < 1,$$

the series converges to the following value:

$$\sum_{l\geq 1}\left[\mathbb{E}\left[(1-u_{l,k})^2\right]\right]^{l-1} = \frac{(s_1+s_2)(s_1+s_2+1)}{(1-p)s_1(s_1+2s_2+1)}. \tag{18}$$

Finally, we can also deduce that

$$\mathbb{E}\left[\omega_{l,k}\right] = \frac{(1-p)s_1}{s_1+s_2}\left(\frac{ps_1+s_2}{s_1+s_2}\right)^{l-1}.$$

To compute $\rho_{j,j'}$ using a SKBP(s_1,s_2,p) on the observational level, we need to compute $\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}^2\right]$ and $\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}\right]^2$. As for the first quantity, using the results in (17) and exploiting the fact that the $u_{l,k}$'s are all i.i.d., we can write

$$\begin{aligned}\mathbb{E}\left[\omega_{l,k}^2\right] &= \mathbb{E}\left[u_{1,k}^2\right]\mathbb{E}\left[(1-u_{1,k})^2\right]^{l-1} \\ &= \frac{(1-p)s_1(s_1+1)}{(s_1+s_2)(s_1+s_2+1)}\cdot\left(\frac{ps_1\left(s_1+2s_2+1\right)+s_2(s_2+1)}{(s_1+s_2)(s_1+s_2+1)}\right)^{l-1}.\end{aligned}$$

Then,

$$\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}^2\right] = \frac{(1-p)s_1(s_1+1)}{(s_1+s_2)(s_1+s_2+1)}\sum_{l\geq 1}\left(\frac{ps_1\left(s_1+2s_2+1\right)+s_2(s_2+1)}{(s_1+s_2)(s_1+s_2+1)}\right)^{l-1},$$

which simplifies to

$$\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}^2\right] = \frac{s_1+1}{s_1+2s_2+1}. \tag{19}$$

Knowing that

$$\mathbb{E}\left[\omega_{l,k}\right]^2 = \frac{(1-p)^2s_1^2}{(s_1+s_2)^2}\left(\frac{(ps_1+s_2)^2}{(s_1+s_2)^2}\right)^{l-1},$$

the second quantity is given by

$$\begin{aligned}\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}\right]^2 &= \frac{(1-p)^2s_1^2}{(s_1+s_2)^2}\sum_{l\geq 1}\left(\frac{(ps_1+s_2)^2}{(s_1+s_2)^2}\right)^{l-1} \\ &= \frac{(1-p)s_1}{(1+p)s_1+2s_2}.\end{aligned} \tag{20}$$

Finally, by considering the ratio of the quantities in (19)-(20), we obtain

$$q_2 = \frac{\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}\right]^2}{\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}^2\right]} = \frac{(1-p)s_1(s_1+2s_2+1)}{((1+p)s_1+2s_2)(s_1+1)}. \tag{21}$$

A.6. Expressions for q_1 under the different SB specifications

Recall that q_1 is simply defined as $q_1 = \sum_{k\geq 1} \mathbb{E}\left[\pi_k^2\right]$. Therefore, it is immediate to derive the expressions of q_1 under the different distributional SB specifications explored:

- $\boldsymbol{\pi} \sim \text{DP}(\alpha_d) \implies q_1 = \frac{1}{1+\alpha_d}$
- $\boldsymbol{\pi} \sim \text{PYP}(\theta_d, \sigma_d) \implies q_1 = \frac{1-\sigma_d}{1+\theta_d}$
- $\boldsymbol{\pi} \sim \text{2PBP}(d_1, d_2) \implies q_1 = \frac{d_1+1}{d_1+2d_2+1}$
- $\boldsymbol{\pi} \sim \text{SKBP}(d_1, d_2, p) \implies q_1 = \frac{d_1+1}{d_1+2d_2+1}$

B. Additional results

B.1. Prior correlations

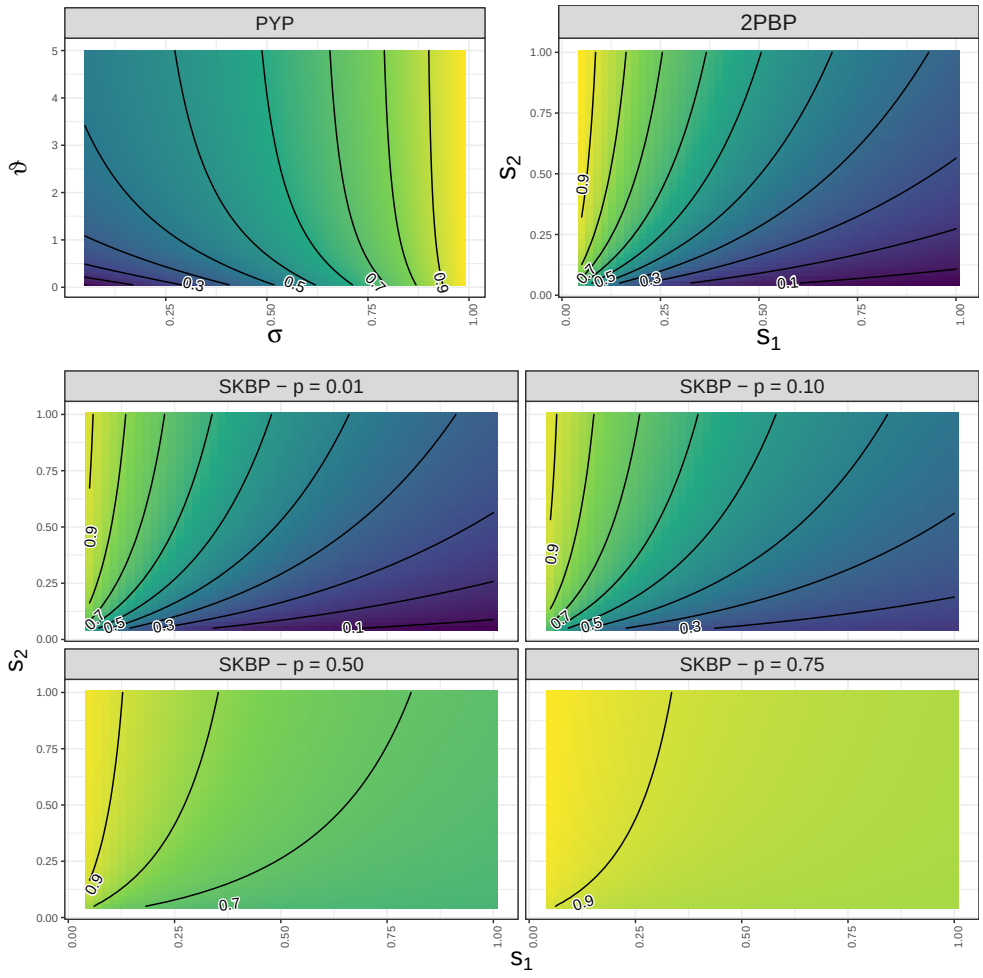

Figure S1: Evolution of the term $(1 - q_2)$ across the different prior specifications for the observational weights.

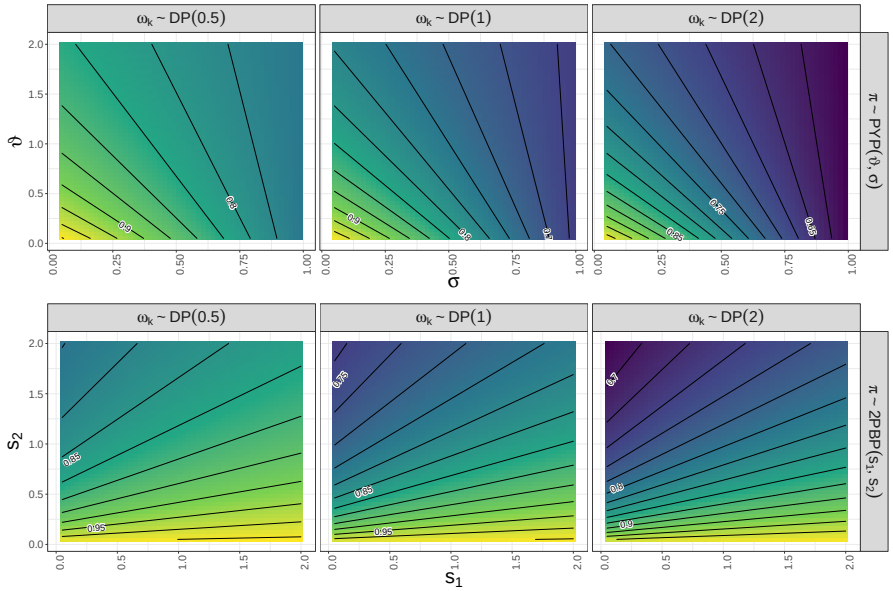

Figure S2: Prior correlations for different values of the parameters (ϑ, σ) (PYP, top row) and (s_1, s_2) (2PBP, bottom row) when assumed as priors for the distributional weights in a nested generalized common atoms model, with different DP specifications for the observational weights (over the columns).

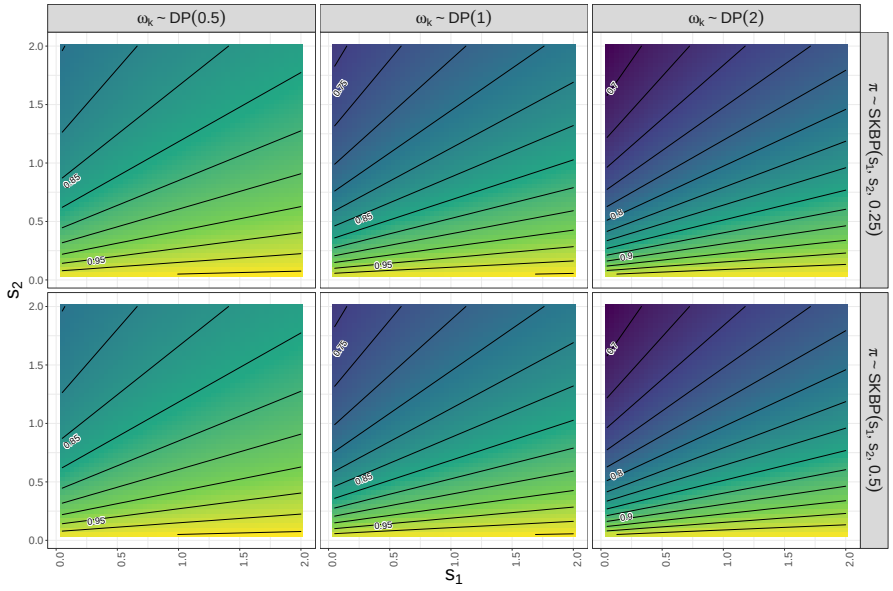

Figure S3: Prior correlations for different values of the parameters (s_1, s_2) (SKBP, with different values of p across rows) when assumed as priors for the distributional weights in a nested generalized common atoms model, with different DP specifications for the observational weights (over the columns).

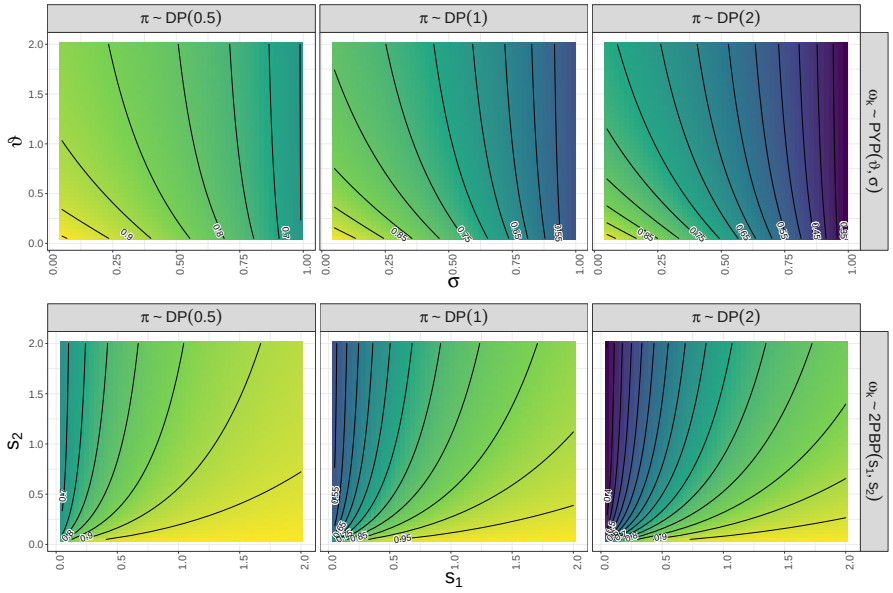

Figure S4: Prior correlations for different values of the parameters (ϑ, σ) (PYP, top row) and (s_1, s_2) (2PBP, bottom row) when assumed as priors for the observational weights in a nested generalized common atoms model, with different DP specifications for the distributional weights (over the columns).

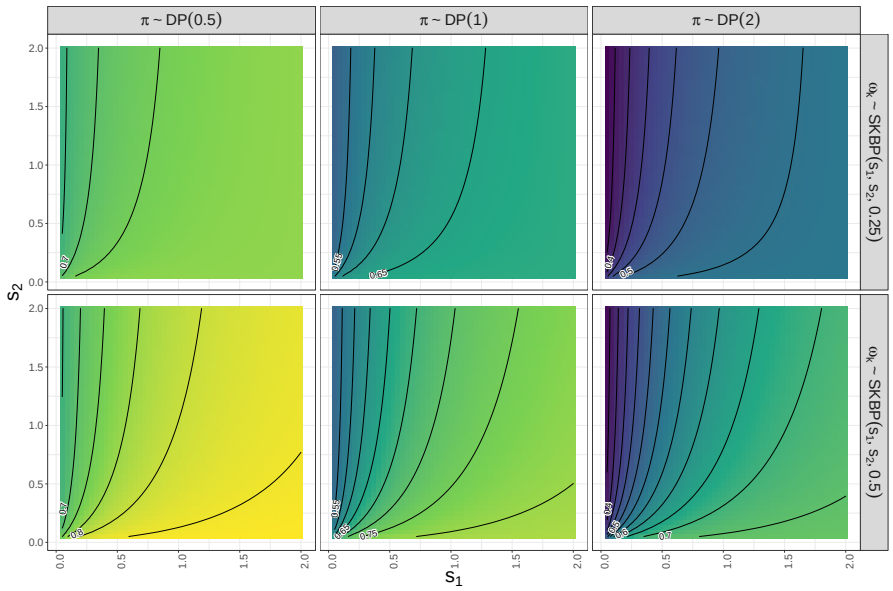

Figure S5: Prior correlations for different values of the parameters (s_1, s_2) (SKBP, with different values of p across rows) when assumed as priors for the observational weights in a nested generalized common atoms model, with different DP specifications for the distributional weights (over the columns).

B.2. Simulation study in the main paper

SKBP(1, 1, p)							
p	n	J	K^*	L^*	$\bar{S}$	$\bar{M}$	Seconds
0	10	2	2.000 (0.000)	2.100 (0.303)	3.458 (0.062)	3.028 (0.330)	0.835 (0.230)
0.25	10	2	2.000 (0.000)	2.240 (0.431)	3.458 (0.074)	3.947 (0.487)	0.874 (0.243)
0.5	10	2	2.000 (0.000)	2.320 (0.513)	3.474 (0.088)	5.564 (0.782)	0.851 (0.234)
0.75	10	2	2.000 (0.000)	2.320 (0.513)	3.477 (0.091)	8.919 (1.041)	0.817 (0.201)
Beta	10	2	2.000 (0.000)	2.320 (0.513)	3.456 (0.104)	6.752 (1.537)	0.841 (0.193)
0	10	4	2.000 (0.000)	2.440 (0.501)	3.135 (0.081)	3.151 (0.309)	1.121 (0.281)
0.25	10	4	2.000 (0.000)	2.520 (0.505)	3.082 (0.104)	4.122 (0.465)	1.165 (0.287)
0.5	10	4	2.000 (0.000)	2.640 (0.485)	3.000 (0.123)	5.744 (0.785)	1.134 (0.277)
0.75	10	4	2.000 (0.000)	2.540 (0.503)	2.906 (0.117)	9.185 (1.377)	1.060 (0.233)
Beta	10	4	2.000 (0.000)	2.440 (0.501)	2.990 (0.120)	6.898 (1.785)	1.120 (0.253)
0	10	6	2.000 (0.000)	2.300 (0.463)	2.953 (0.103)	3.171 (0.248)	1.365 (0.332)
0.25	10	6	2.000 (0.000)	2.580 (0.499)	2.929 (0.089)	4.015 (0.403)	1.399 (0.329)
0.5	10	6	2.000 (0.000)	2.680 (0.471)	2.816 (0.107)	5.614 (0.678)	1.368 (0.318)
0.75	10	6	2.000 (0.000)	2.700 (0.505)	2.663 (0.127)	9.055 (1.344)	1.307 (0.304)
Beta	10	6	2.000 (0.000)	2.600 (0.495)	2.786 (0.130)	6.882 (1.589)	1.341 (0.270)

Table S1: Average and standard deviation (across the 50 replications) of the number of estimated distributional and observational clusters (K^*, L^*), of the maximum value of the distributional and observational labels ($\bar{S}, \bar{M}$), and of the computational time (in seconds). Results for $n = 10$.

SKBP(1, 1, p)							
p	n	J	K^*	L^*	$\bar{S}$	$\bar{M}$	Seconds
0	25	2	2.000 (0.000)	2.400 (0.535)	3.480 (0.134)	3.184 (0.410)	1.158 (0.300)
0.25	25	2	2.000 (0.000)	2.480 (0.544)	3.465 (0.129)	4.067 (0.589)	1.200 (0.301)
0.5	25	2	2.000 (0.000)	2.600 (0.535)	3.455 (0.128)	5.935 (0.843)	1.174 (0.292)
0.75	25	2	2.000 (0.000)	2.580 (0.538)	3.488 (0.171)	9.158 (1.503)	1.122 (0.251)
Beta	25	2	2.000 (0.000)	2.500 (0.544)	3.478 (0.143)	7.334 (1.934)	1.148 (0.253)
0	25	4	2.000 (0.000)	2.400 (0.495)	2.887 (0.151)	3.188 (0.312)	1.729 (0.409)
0.25	25	4	2.000 (0.000)	2.560 (0.541)	2.895 (0.119)	4.149 (0.528)	1.774 (0.400)
0.5	25	4	2.000 (0.000)	2.780 (0.465)	2.896 (0.162)	5.971 (0.875)	1.746 (0.390)
0.75	25	4	2.000 (0.000)	2.860 (0.405)	2.757 (0.166)	9.418 (1.846)	1.675 (0.368)
Beta	25	4	2.000 (0.000)	2.620 (0.490)	2.884 (0.145)	7.181 (2.378)	1.721 (0.333)
0	25	6	2.000 (0.000)	2.660 (0.593)	2.677 (0.152)	3.396 (0.476)	2.303 (0.515)
0.25	25	6	2.000 (0.000)	2.860 (0.572)	2.715 (0.135)	4.455 (0.799)	2.335 (0.495)
0.5	25	6	2.000 (0.000)	3.020 (0.622)	2.650 (0.181)	6.569 (1.296)	2.293 (0.478)
0.75	25	6	2.000 (0.000)	3.040 (0.493)	2.568 (0.154)	9.939 (2.010)	2.223 (0.459)
Beta	25	6	2.000 (0.000)	2.880 (0.521)	2.612 (0.173)	7.871 (2.462)	2.289 (0.454)
0	50	2	2.000 (0.000)	2.420 (0.499)	3.488 (0.230)	3.206 (0.290)	1.689 (0.392)
0.25	50	2	2.000 (0.000)	2.580 (0.538)	3.520 (0.195)	4.177 (0.534)	1.738 (0.386)
0.5	50	2	2.000 (0.000)	2.760 (0.476)	3.425 (0.141)	5.924 (0.920)	1.683 (0.365)
0.75	50	2	2.000 (0.000)	2.820 (0.388)	3.449 (0.250)	9.907 (1.791)	1.624 (0.350)
Beta	50	2	2.000 (0.000)	2.560 (0.501)	3.432 (0.214)	7.289 (2.768)	1.665 (0.314)
0	50	4	2.000 (0.000)	2.600 (0.535)	2.852 (0.262)	3.364 (0.438)	2.752 (0.582)
0.25	50	4	2.000 (0.000)	2.700 (0.544)	2.839 (0.178)	4.461 (0.797)	2.795 (0.565)
0.5	50	4	2.000 (0.000)	2.940 (0.586)	2.797 (0.202)	6.35 (1.448)	2.747 (0.548)
0.75	50	4	2.000 (0.000)	2.920 (0.444)	2.762 (0.245)	9.83 (2.153)	2.675 (0.527)
Beta	50	4	2.000 (0.000)	2.820 (0.482)	2.823 (0.277)	7.802 (2.785)	2.761 (0.558)
0	50	6	2.000 (0.000)	2.580 (0.538)	2.671 (0.319)	3.391 (0.345)	3.872 (0.792)
0.25	50	6	2.000 (0.000)	2.900 (0.463)	2.608 (0.143)	4.241 (0.601)	3.885 (0.757)
0.5	50	6	2.000 (0.000)	3.020 (0.377)	2.532 (0.172)	6.613 (1.465)	3.845 (0.735)
0.75	50	6	2.000 (0.000)	2.960 (0.402)	2.596 (0.213)	10.409 (2.339)	3.759 (0.715)
Beta	50	6	2.000 (0.000)	2.880 (0.480)	2.514 (0.181)	8.524 (3.280)	3.825 (0.732)

Table S2: Average and standard deviation (across the 50 replications) of the number of estimated distributional and observational clusters (K^*, L^*), of the maximum value of the distributional and observational labels ($\bar{S}, \bar{M}$), and of the computational time (in seconds). Results for $n = 25, 50$.

2PBP(s, s)							
s	n	J	K^*	L^*	$\bar{S}$	$\bar{M}$	Seconds
1.000	10	2	2.000 (0.000)	2.200 (0.404)	3.456 (0.081)	3.015 (0.357)	0.804 (0.233)
0.500	10	2	2.000 (0.000)	2.060 (0.240)	3.460 (0.072)	2.862 (0.298)	0.815 (0.203)
0.100	10	2	2.000 (0.000)	2.000 (0.000)	3.452 (0.074)	3.013 (0.499)	0.830 (0.193)
1.000	10	4	2.000 (0.000)	2.440 (0.501)	3.130 (0.091)	3.17 (0.319)	1.088 (0.278)
0.500	10	4	2.000 (0.000)	2.100 (0.303)	3.237 (0.089)	2.887 (0.337)	1.128 (0.282)
0.100	10	4	2.000 (0.000)	2.000 (0.000)	3.428 (0.101)	3.013 (0.468)	1.106 (0.230)
1.000	10	6	2.000 (0.000)	2.420 (0.499)	2.949 (0.091)	3.181 (0.223)	1.335 (0.328)
0.500	10	6	2.000 (0.000)	2.060 (0.240)	3.150 (0.090)	2.892 (0.256)	1.377 (0.332)
0.100	10	6	2.000 (0.000)	2.000 (0.000)	3.502 (0.126)	2.774 (0.538)	1.357 (0.272)
1.000	25	2	2.000 (0.000)	2.300 (0.463)	3.483 (0.143)	3.136 (0.405)	1.143 (0.298)
0.500	25	2	2.000 (0.000)	2.100 (0.303)	3.431 (0.098)	2.882 (0.356)	1.177 (0.299)
0.100	25	2	2.000 (0.000)	2.020 (0.141)	3.460 (0.069)	3.072 (0.691)	1.151 (0.245)
1.000	25	4	2.000 (0.000)	2.340 (0.479)	2.921 (0.127)	3.217 (0.320)	1.714 (0.408)
0.500	25	4	2.000 (0.000)	2.160 (0.370)	3.049 (0.093)	2.901 (0.298)	1.748 (0.406)
0.100	25	4	2.000 (0.000)	2.000 (0.000)	3.351 (0.151)	3.017 (0.859)	1.750 (0.356)
1.000	25	6	2.000 (0.000)	2.580 (0.575)	2.700 (0.152)	3.38 (0.498)	2.298 (0.510)
0.500	25	6	2.000 (0.000)	2.200 (0.452)	2.913 (0.105)	2.98 (0.447)	2.322 (0.506)
0.100	25	6	2.000 (0.000)	2.060 (0.240)	3.351 (0.213)	3.432 (0.935)	2.323 (0.470)
1.000	50	2	2.000 (0.000)	2.380 (0.490)	3.488 (0.189)	3.198 (0.317)	1.759 (0.370)
0.500	50	2	2.000 (0.000)	2.040 (0.198)	3.465 (0.143)	2.835 (0.267)	1.771 (0.356)
0.100	50	2	2.000 (0.000)	2.000 (0.000)	3.460 (0.068)	2.879 (0.646)	1.775 (0.342)
1.000	50	4	2.000 (0.000)	2.580 (0.538)	2.853 (0.271)	3.339 (0.416)	2.751 (0.582)
0.500	50	4	2.000 (0.000)	2.160 (0.422)	2.956 (0.120)	2.904 (0.426)	2.791 (0.592)
0.100	50	4	2.000 (0.000)	2.040 (0.198)	3.270 (0.163)	3.406 (1.079)	2.799 (0.578)
1.000	50	6	2.000 (0.000)	2.600 (0.535)	2.592 (0.261)	3.376 (0.370)	3.850 (0.789)
0.500	50	6	2.000 (0.000)	2.220 (0.418)	2.769 (0.140)	3.025 (0.355)	3.867 (0.789)
0.100	50	6	2.000 (0.000)	2.040 (0.198)	3.190 (0.224)	3.674 (1.106)	3.886 (0.778)

Table S3: Average and standard deviation (across the 50 replications) of the number of estimated distributional and observational clusters (K^*, L^*), of the maximum value of the distributional and observational labels ($\bar{S}, \bar{M}$), and of the computational time (in seconds).

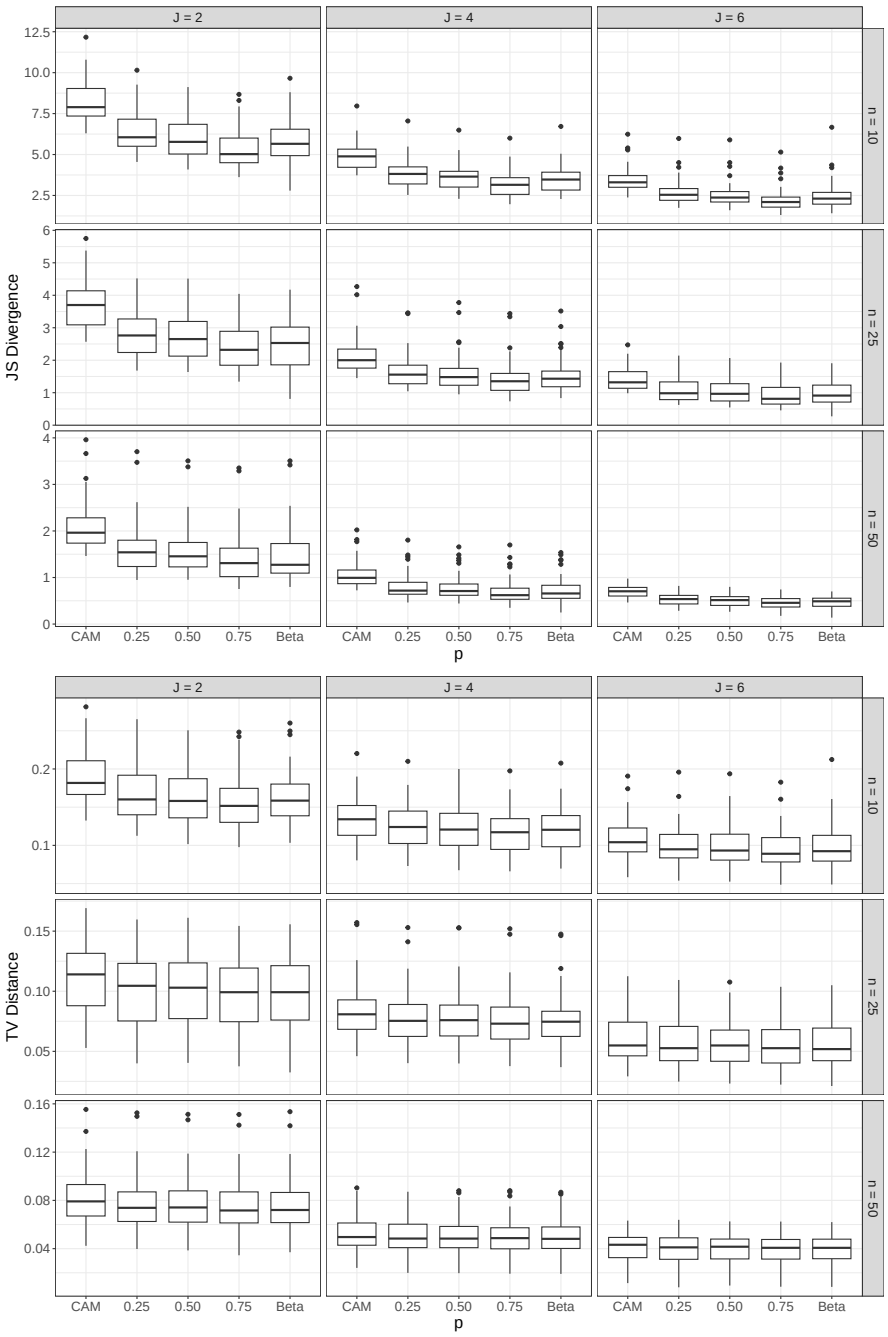

Figure S6: Distributions of the JS divergence (top panels) and TV distances (bottom panels) over the 50 replications for a geCAM embedded with a SKBP$(1, 1, p)$, for different specifications of p. The standard CAM is obtained for $p = 0$, while the last column corresponds to a random $p \sim \text{Beta}(1, 1)$. Each panel corresponds to a simulation scenario.

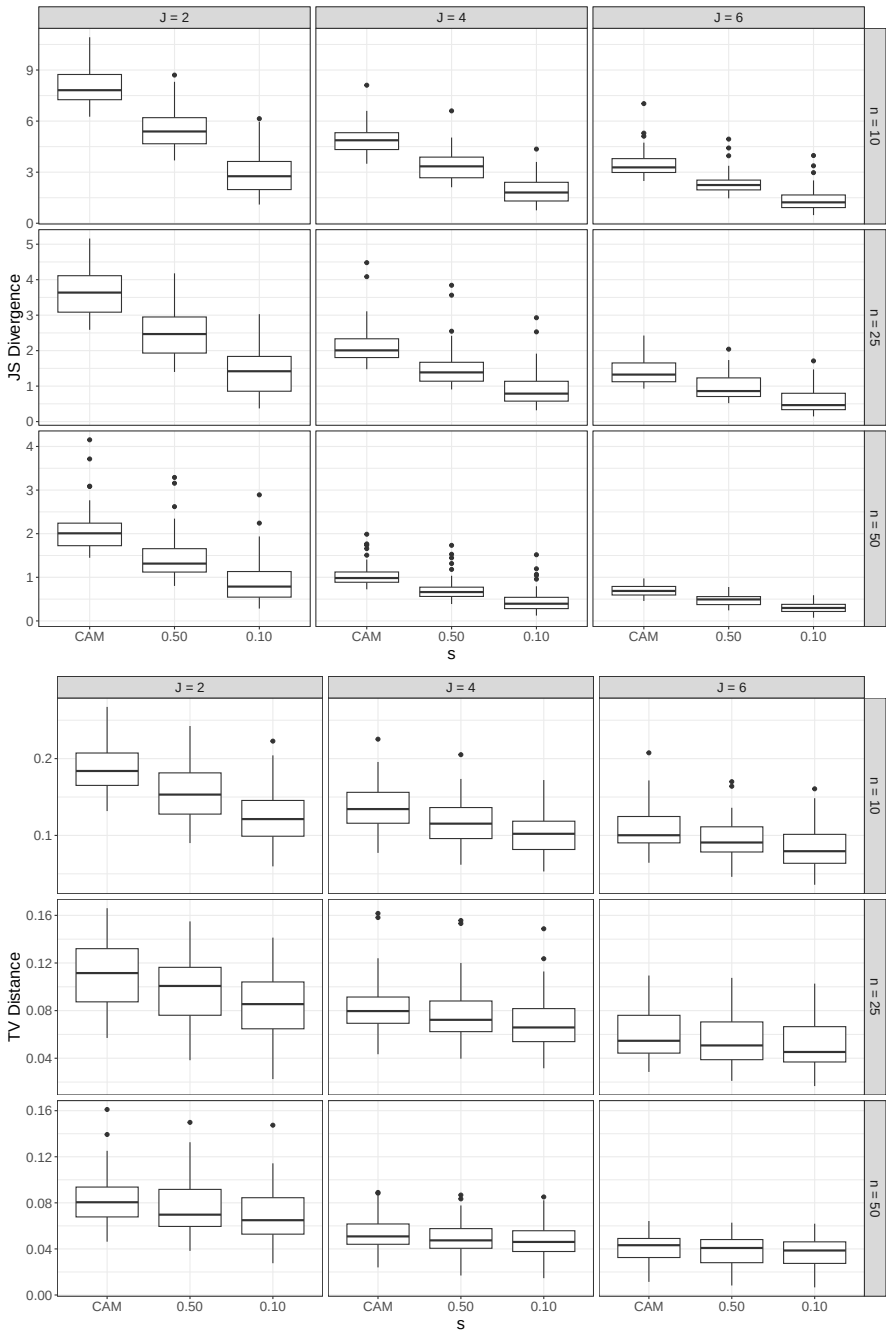

Figure S7: Distributions of the JS divergence (top panels) and TV distances (bottom panels) over the 50 replications for a geCAM embedded with a 2PBP(s, s), for different specifications of s. The standard CAM is obtained for $s = 1$. Each panel corresponds to a simulation scenario.

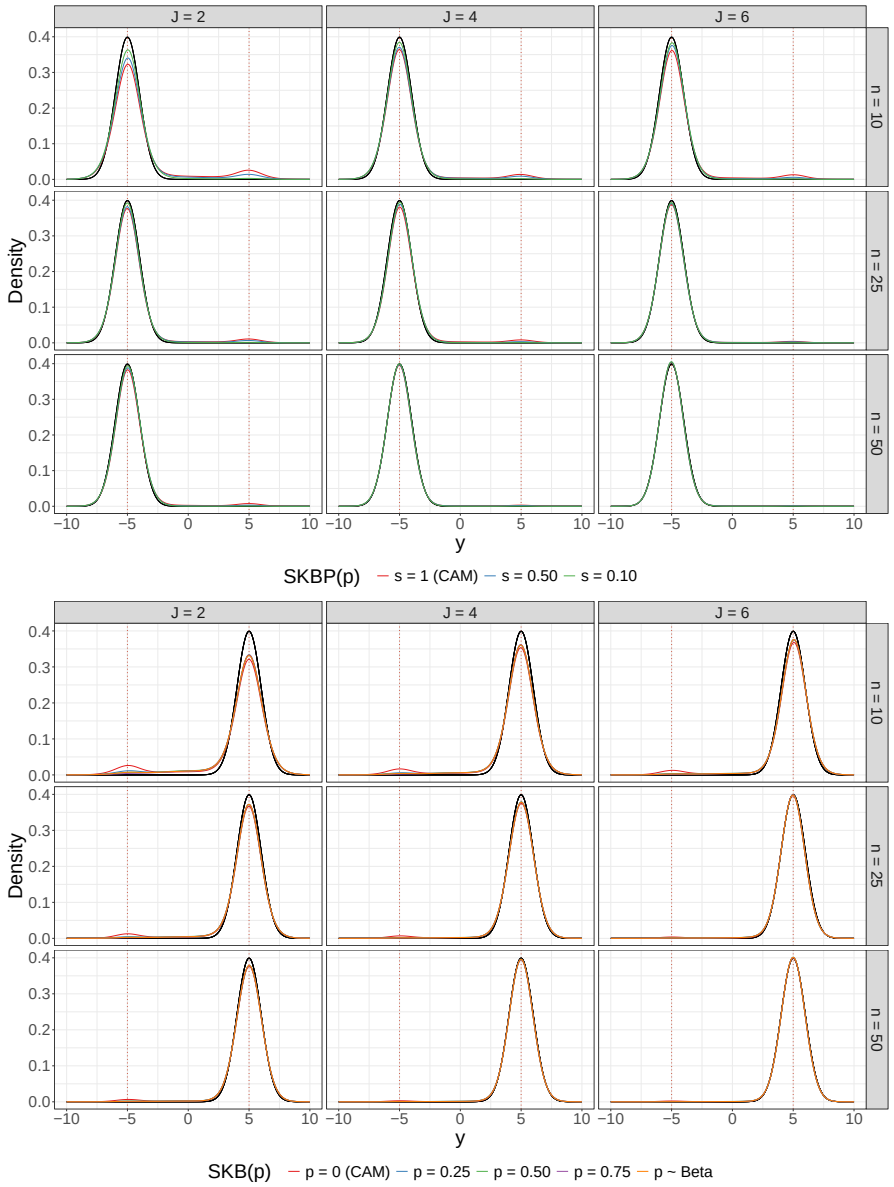

Figure S8: Ground truth (black line) and posterior density estimates of different SKBP$(1, 1, p)$ model specifications (different values of p correspond to different colors) for the first (top graphs) and second (bottom graphs) subpopulations used in the simulations study.

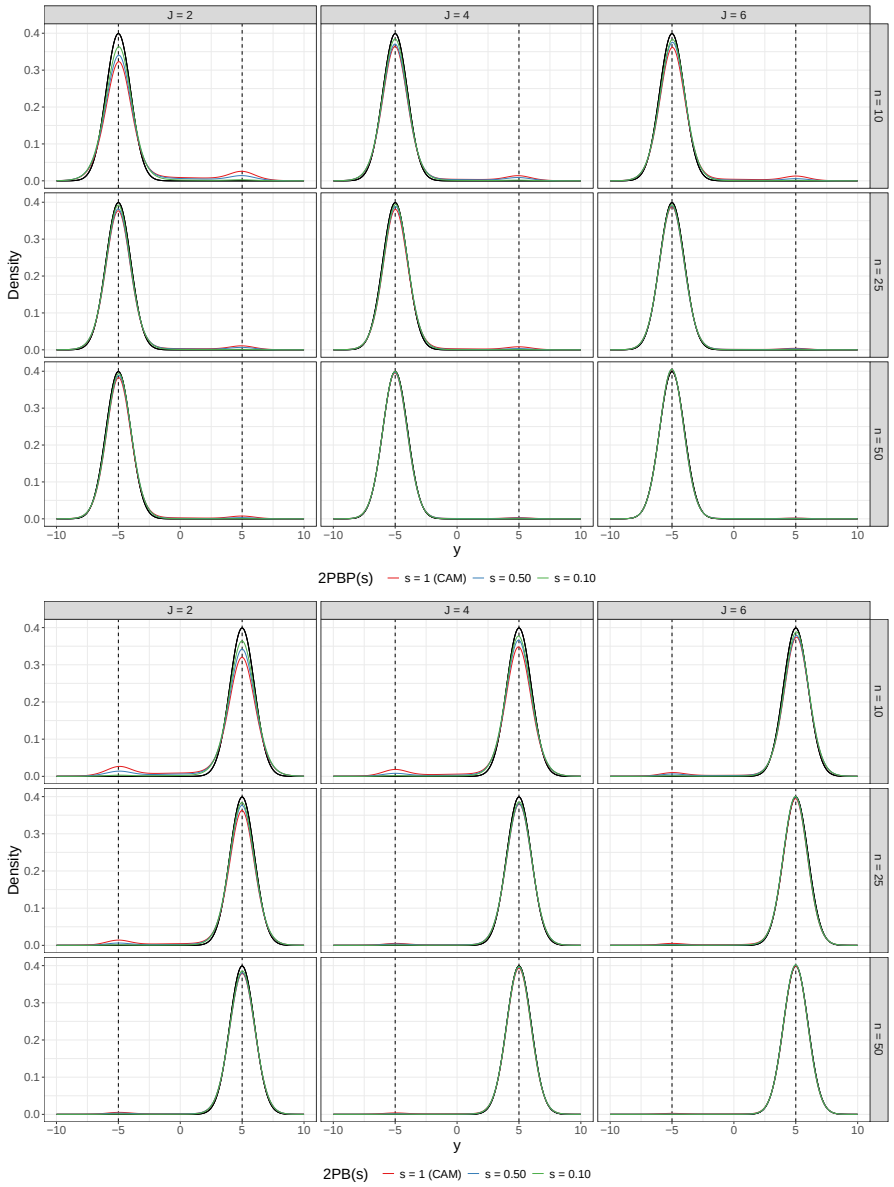

Figure S9: Ground truth (black line) and posterior density estimates of different 2PBP(s, s) model specifications (different values of s correspond to different colors) for the first (top graphs) and second (bottom graphs) subpopulations used in the simulations study.

B.3. Additional simulation study

Similarly to the case presented in the main text, here we consider a simulation study with more complicated subpopulation distributions generating the grouped data. In particular, we generate $J/2$ samples of size n from

$$p_1 = \frac{1}{3}N(-5, 1) + \frac{1}{3}N(0, 1) + \frac{1}{3}N(5, 1)$$

and another $J/2$ samples of size n from

$$p_2 = \frac{1}{2}N(0, 1) + \frac{1}{2}N(3, 1).$$

Then, we sample $J \in \{2, 4\}$ groups comprising $N_j = n \in \{12, 30\}$ independent observations. Again, considering all the possible combinations between the values of J and n results in 4 simulation scenarios. All the other simulation settings and the assessment of the results follow the same procedure delineated in the main paper. Figures S10, S11, and S12 display the results in terms of KL divergence, JS divergence, and TV distance, respectively.

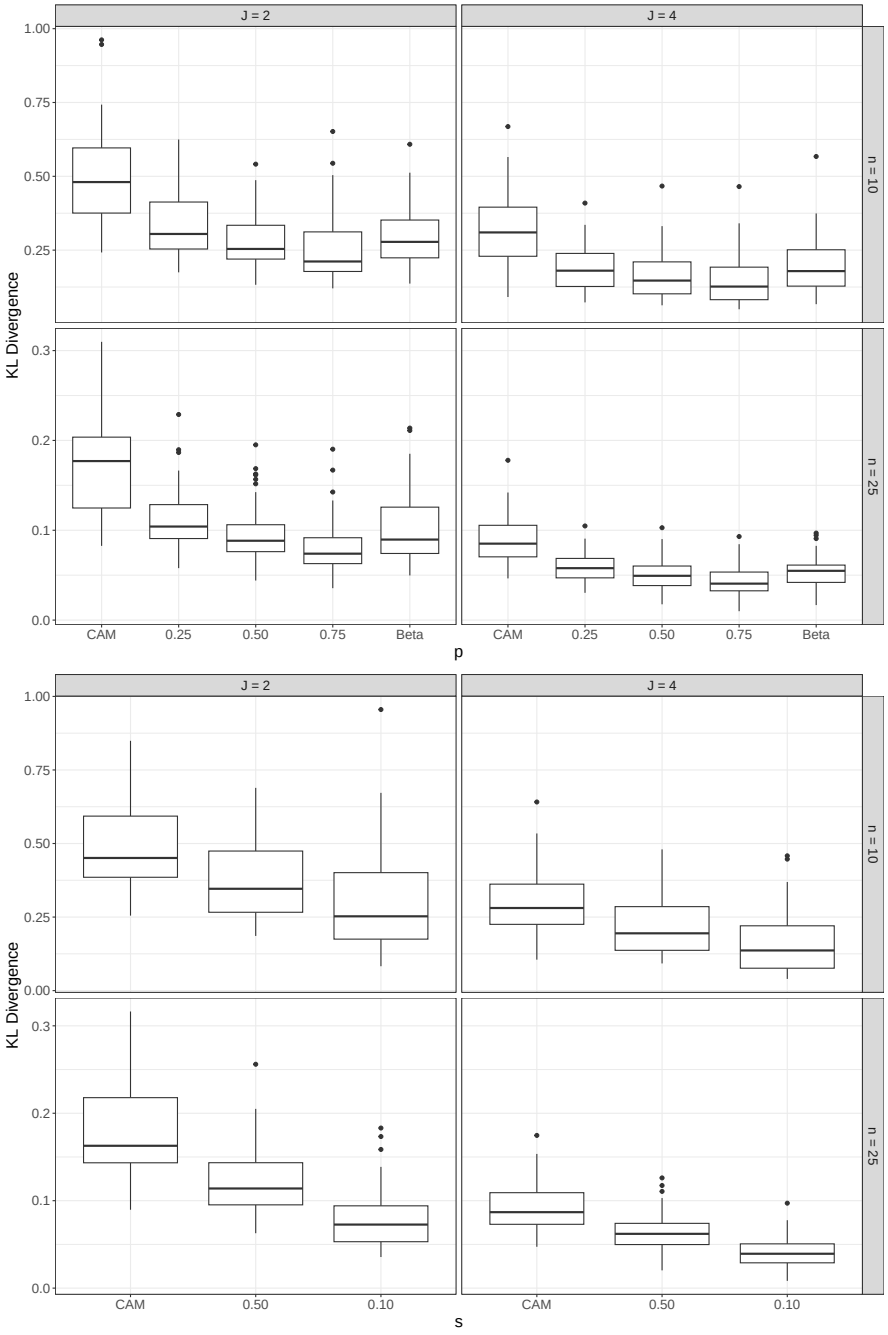

Figure S10: Distributions of the KL divergence over the 50 replications for a geCAM embedded with a SKBP$(1, 1, p)$ (top graphs) and the 2PBP(s) (bottom graphs), for different specifications of p and s, respectively.

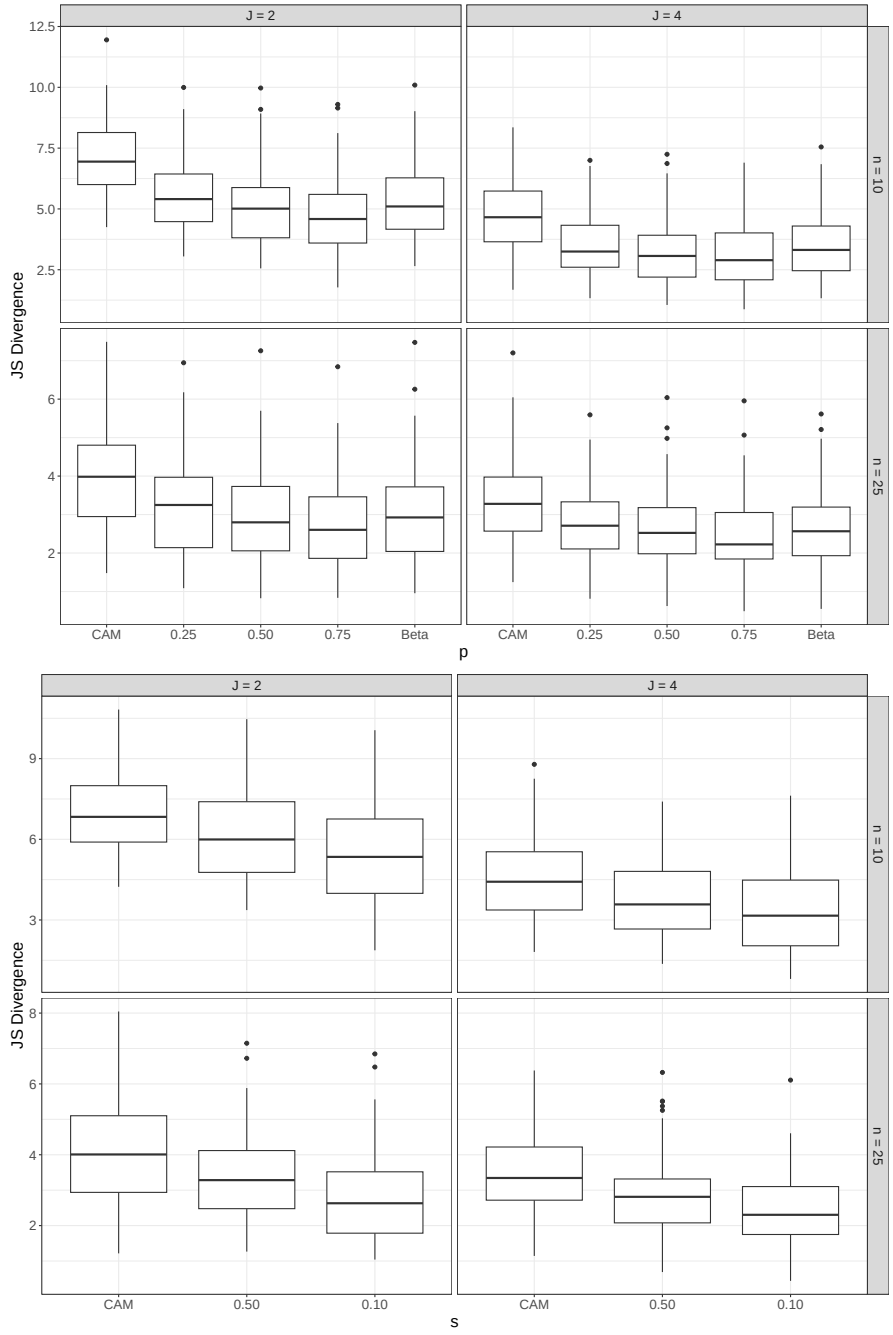

Figure S11: Distributions of the JS divergence over the 50 replications for a geCAM embedded with a SKBP$(1, 1, p)$ (top graphs) and the 2PBP(s) (bottom graphs), for different specifications of p and s, respectively.

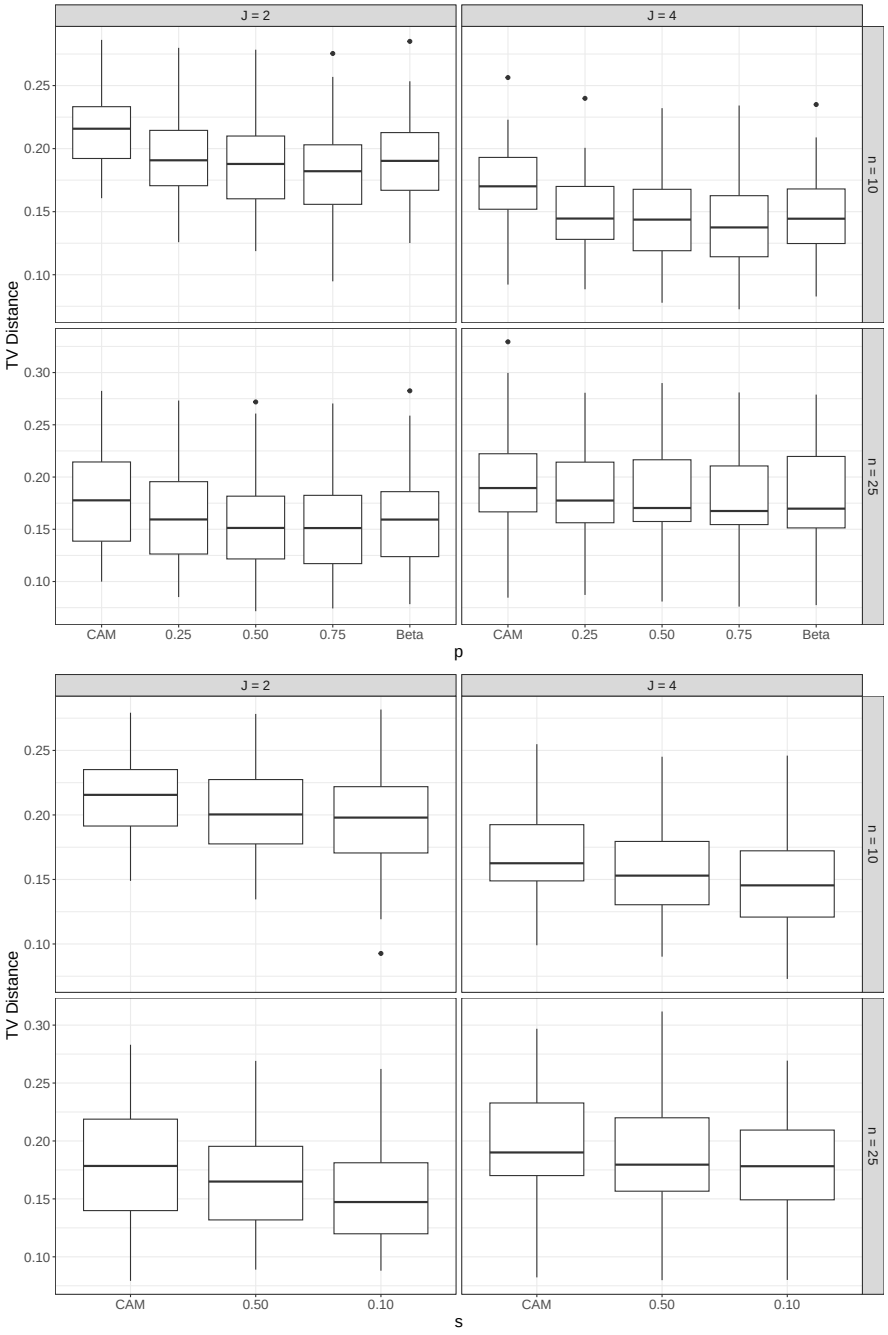

Figure S12: Distributions of the TV distance over the 50 replications for a geCAM embedded with a SKBP$(1, 1, p)$ (top graphs) and the 2PBP(s) (bottom graphs), for different specifications of p and s, respectively.

B.4. Application to CPP data

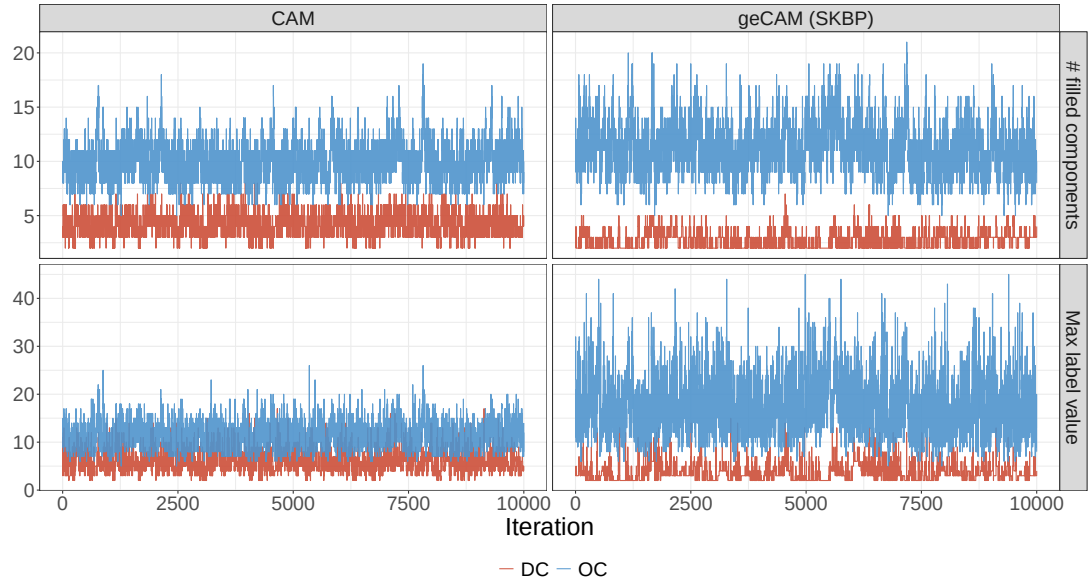

Figure S13: MCMC traceplots of the number of filled distributional and observational components (defined in the main text as K^* and L^*, top row) and values of the largest label used ($\bar{S}$ and $\bar{M}$, bottom row) obtained with the geCAM (SKBP) and CAM models when fitted to the CPP data. All traces are below the truncation values chosen for model fitting.