Bayesian Analysis (2024) **TBA**, Number TBA, pp. 1–29

Hierarchical Mixture of Finite Mixtures (with Discussion)

Alessandro Colombi[*], Raffaele Argiento[†], Federico Camerlenghi[‡], and Lucia Paci[§]

Abstract. Statistical modelling in the presence of data organized in groups is a crucial task in Bayesian statistics. The present paper conceives a mixture model based on a novel family of Bayesian priors designed for multilevel data and obtained by normalizing a finite point process. In particular, the work extends the popular Mixture of Finite Mixtures model to the hierarchical framework to capture heterogeneity within and between groups. A full distribution theory for this new family and the induced clustering is developed, including the marginal, posterior, and predictive distributions. Efficient marginal and conditional Gibbs samplers are designed to provide posterior inference. The proposed mixture model outperforms the Hierarchical Dirichlet Process, the foremost tool for handling multilevel data, in terms of analytical feasibility, clustering discovery, and computational time. The motivating application comes from the analysis of shot put data, which contains performance measurements of athletes across different seasons. In this setting, the proposed model is exploited to induce clustering of the observations across seasons and athletes. By linking clusters across seasons, similarities and differences in athletes' performances are identified.

Keywords: model-based clustering, multilevel data, partial exchangeability, sports analytics, vector of finite Dirichlet processes.

MSC2020 subject classifications: Primary 62F15, 62H30, 62G05.

1 Introduction

Statistical modelling of population heterogeneity is a recurrent challenge in real-world applications. In this study, our focus shifts towards hierarchical data scenarios, where observations emerge from distinct groups (or levels) and our objective is to model heterogeneity both within and between these groups. Specifically, heterogeneity within groups is handled via mixture modelling to get group-specific clustering of observations, as well as density estimation. Concurrently, between-group heterogeneity can be addressed through two extreme modelling choices: (i) pooling all observations and (ii) conducting independent analyses for each of the d groups. However, both alternatives pose limitations. In the first case, differences across groups are not accounted for, while in the second one, groups are not linked, preventing sharing of statistical strength. In

[*] Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy, a.colombi10@campus.unimib.it

[†] Department of Economics, Università degli studi di Bergamo, Italy, raffaele.argiento@unibg.it

[‡] Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy, federico.camerlenghi@unimib.it

[§] Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy, lucia.paci@unicatt.it

https://doi.org/10.1214/24-BA1501

this work, a tunable balance between the two alternatives is proposed to model heterogeneity across groups.

In Bayesian formalism, the sharing of information is naturally achieved through hierarchical modelling; parameters are shared among groups, and the randomness of the parameters induces dependencies among the groups. In particular, in model-based clustering, such sharing leads to group-specific clusters that are common among the groups, i.e., clusters that have the same interpretation across groups allow to define a global clustering. In general, the number of clusters within each group, as well as the number of global clusters, are unknown and need to be inferred from the data. In the hierarchical landscape, several Bayesian nonparametric approaches have been proposed, the pioneering work being the Hierarchical Dirichlet Process (HDP) mixture model (Teh et al., 2006). The HDP has been recently extended to encompass normalized random measures (Camerlenghi et al., 2019; Argiento et al., 2020) and species sampling models (Bassetti et al., 2020). Despite their mathematical elegance, these methods suffer from considerable computational complexity and costs, posing a barrier for practitioners.

This work targets these problems and proposes a finite mixture model for each group, where the random number of mixture components, as well as the mixture parameters, are shared among groups. Rather, the mixture weights are assumed to be group-specific in order to accommodate differences between groups. The proposed approach bridges the gap between infinite and finite mixtures in the hierarchical setting by introducing a novel family of Bayesian priors, named Vector of Finite Dirichlet Processes (Vec-FDP), to capture heterogeneity within and between groups. In particular, we use the Vec-FDP prior to building a new class of hierarchical mixture models, named Hierarchical Mixture of Finite Mixtures (HMFM), that encompasses the popular Mixture of Finite Mixtures (MFM) model (Miller and Harrison, 2018; Frühwirth-Schnatter et al., 2021) as a special case when the number of groups $d = 1$. We point out that the name Hierarchical Mixture of Finite Mixtures was previously used by Miller (2014), whose model falls into a special case of the hierarchical species sampling models discussed by Bassetti et al. (2020). However, note that both Miller (2014) and Bassetti et al. (2020) use *hierarchical* to refer to the hierarchy among the random probability measures, while we refer to the hierarchy of the data.

Furthermore, following the same approach of Argiento and De Iorio (2022), we introduce a more general construction for the mixing weights. This involves the normalization of positive unnormalized weights distributed according to any probability distribution over the positive real numbers. By doing so, we define a broader family of Bayesian priors, referred to as Vector of Normalized Independent Finite Point Processes (Vec-NIFPP), which encompasses the Vec-FDP as a special case when the unnormalized weights are Gamma distributed. Although we borrow Bayesian nonparametric tools to derive the distributional results and the clustering properties, the model leverages on a finite, yet random, number of mixture components, enhancing its accessibility to a broader audience, e.g., those interested in model-based clustering and mixture of experts modelling (Jacobs et al., 1991).

Posterior inference poses computational challenges in the context of Bayesian nonparametrics, particularly when dealing with hierarchical data. Rather, leveraging the posterior characterization of the Vec-FDP, we design an efficient Markov chain Monte Carlo (MCMC) sampling strategy, which significantly improves the existing methods based on infinite dimensional processes, such as the HDP. Notably, given a fixed number of clusters and groups, the HDP's computational time scales quadratically with the volume of data. In contrast, our approach achieves linear scaling.

The reason behind such improvement is evident from the restaurant franchise-like representation of the Vec-FDP prior. In contrast to the HDP's complex distinction between tables and menus, our representation is straightforward due to the absence of stacked layers of infinite dimensional objects in the model structure. This simplification enhances model interpretability, which is preserved when moving from a single group to multiple groups, unlike the HDP. The restaurant franchise metaphor also illuminates the flexibility of the sharing of information mechanism induced by the proposed method. In this regard, a comprehensive simulation study compares the proposed HMFM model with both the HDP mixture model and the MFM model assumed independently for each group. Experiments shed light on the advantages of employing joint modelling rather than independent analyses and demonstrate how the excessive sharing of information induced by the HDP may lead to misleading conclusions. Therefore, the HMFM strikes a balanced compromise between the HDP and the independent analyses, offering a tunable approach that combines the strengths of both methods.

The motivating example for this work comes from sports analytics. See Page et al. (2013) for Bayesian nonparametric methods in this domain. Our methodology finds application in the analysis of data from shot put, a track and field discipline that involves propelling a heavy spherical ball, or *shot*, over the greatest distance attainable. The dataset comprises measurements, specifically throw lengths or marks, recorded during professional shot put competitions from 1996 to 2016, for a total of $35,637$ marks on 403 athletes. The data are organized by aligning marks by season to ensure equitable comparison among athletes. Athletes have varying durations, with a maximum of 15 seasons, which correspond to the d different groups. We employ the proposed HMFM model to cluster athletes' performances within each season while preserving the interpretability of clusters across different seasons. This allows us to enrich the understanding of the evolution in athletes' performance trends. A remarkable finding is that the estimated clusters are gender-free, thanks to the inclusion of an additional season-specific regression parameter. Notably, the model identifies a special cluster consisting of six exceptional women's performances achieved by Olympic or world champions; no men have ever been able to outperform their competitors in such a neat way.

Summing up, our methodological contribution is to propose a new family of dependent Bayesian nonparametric priors designed for analyzing hierarchical data in a mixture setting. The proposed model allows to identify random clusters of observations within and between groups. We deeply investigate the theoretical properties of the HMFM, including the distribution of the random partition, the predictive distribution of the process, and its correlation structure. Additionally, we fully characterize its posterior distribution. The theoretical properties of the more general Vec-NIFPP model have also been thoroughly investigated. On the computational side, two efficient MCMC procedures based on a conditional and a marginal sampler provide inference on the number of clusters, and the clustering structure as well as group-dependent density estimation. Both algorithms show improved computational efficiency with respect to common strategies for posterior sampling in hierarchical mixture models. The flexibility of the proposed model is illustrated through an extensive simulation study and a real-world application in sports analytics.

The rest of the paper is organized as follows. Section 2 introduces the HMFM model and defines the local and global clustering. Section 3 provides the main distributional results, such as the marginal, posterior, and predictive distributions of the Vec-FDP. In Section 4 we present the hyperpriors choice and the computational strategies. Section 5 showcases a simulation study where we compare the HMFM with the HDP and the MFM, the latter fitted independently within each group as well as on the pooled dataset. The analysis of shot put data is illustrated in Section 6. A discussion in Section 7 concludes the paper and Supplementary materials (Colombi et al., 2024) complement it.

2 Hierarchical mixture of finite mixtures

Given d groups (or levels), let $\boldsymbol{y}_j = \left(y_{j,1}, \ldots, y_{j,n_j}\right)$ denote the data collected over n_j individuals in group $j = 1, \ldots, d$, where $y_{j,i} \in \mathbb{Y}$ and $\mathbb{Y}$ is the sampling space. We assume that the data in each group j come from a finite mixture of M components, that is

$$y_{j,1}, \ldots, y_{j,n_j} \mid P_j \overset{\text{iid}}{\sim} \int_{\Theta} f(\,\cdot \mid \tau) P_j(\mathrm{d}\tau) \quad \text{for each } j = 1, \ldots, d, \tag{1}$$

where $\{f(\,\cdot \mid \tau),\ \tau \in \Theta\}$ is a parametric family of densities over $\mathbb{Y}$ and $(P_1, \ldots, P_d)$ is a vector of random probability measures over the parameter space $\Theta \subset \mathbb{R}^s$. We focus on random probability measures having almost surely discrete realizations. More specifically, we define a vector of finite dependent random probability measures $(P_1, \ldots, P_d)$ as follows,

$$P_j(\cdot) = \sum_{m=1}^{M} \frac{S_{j,m}}{T_j} \delta_{\tau_m}(\cdot), \tag{2}$$

where $S_{j,m}$ are the unnormalized weights, δ_{τ_m} stands for the delta-Dirac mass at τ_m, and $T_j = \sum_{m=1}^{M} S_{j,m}$ is referred to as the total mass.

As a prior for M, we place a 1-shifted Poisson distribution with parameter Λ, denoted by $\mathrm{Pois}_1(\Lambda)$, so that we are sure that there always exists at least one mixture component. Then, conditionally to M, $(\tau_1, \ldots, \tau_M)$ are common random atoms across the d random probability measures, which are assumed independent and identically distributed (i.i.d.) with common distribution P_0, that is a diffuse probability measure on Θ. Moreover, given M, the unnormalized weights $S_{j,m}$ are conditionally independent both within and between the groups. In particular, we assume the components of $\boldsymbol{S}_j = (S_{j,1}, \ldots, S_{j,M})$ to be i.i.d. from Gamma $(\gamma_j, 1)$, independently with respect to $j = 1, \ldots, d$. Throughout this work, we always refer to a shape-rate parametrization of the gamma distribution.

The induced prior on the normalized mixture weights $(\pi_{j,1},\dots,\pi_{j,M})$ is

$$(\pi_{j,1},\dots,\pi_{j,M}) = \left(\frac{S_{j,1}}{T_j},\dots,\frac{S_{j,M}}{T_j}\right) \sim \mathrm{Dir}_M\left(\gamma_j,\dots,\gamma_j\right), \quad \text{for each } j=1,\dots,d,$$

where $\mathrm{Dir}_M\left(\gamma_j,\dots,\gamma_j\right)$ denotes the M-dimensional symmetric Dirichlet distribution with parameter γ_j. The mixing measure P_j is obtained by normalization:

$$P_j(\cdot) = \frac{\mu_j(\cdot)}{\mu_j(\Theta)}, \tag{3}$$

where $\mu_j(\cdot) = \sum_{m=1}^{M} S_{j,m}\delta_{\tau_m}(\cdot)$. The seminal contribution of Regazzini et al. (2003) has spurred the construction of random probability measures via the normalization approach, which turns out to be a convenient framework to face posterior inference; see, e.g., Lijoi et al. (2014) Camerlenghi et al. (2019), Argiento et al. (2020) and Argiento and De Iorio (2022) for allied contributions. We point out that, marginally, each component P_j is a finite Dirichlet process as the one described in Argiento and De Iorio (2022). These discrete random measures are also known as Gibbs-type priors with a negative parameter (Gnedin and Pitman, 2006; De Blasi et al., 2015).

Our model construction generalizes the work of Argiento and De Iorio (2022) by allowing for the sharing of information across groups thanks to shared atoms and a shared number of components. Besides, the proposed model retains mathematical tractability thanks to the normalization approach and the conditional independence of the unnormalized weights. The prior specification for the mixture parameters $(M,\boldsymbol{S},\boldsymbol{\tau})$, where $\boldsymbol{S} = (S_1,\dots,S_d)$ and $\boldsymbol{\tau} = (\tau_1,\dots,\tau_M)$, is equivalent to specify the joint law of the vector $(P_1,\dots,P_d)$, called *Vector of Finite Dirichlet Process* (Vec-FDP) and denoted by

$$(P_1,\dots,P_d) \sim \text{Vec-FDP}\left(\Lambda,\boldsymbol{\gamma},P_0\right), \tag{4}$$

where $\boldsymbol{\gamma} = (\gamma_1,\dots,\gamma_d)$. Summing up, the model can be formulated in the following hierarchical form,

$$\begin{aligned}
y_{j,1},\dots,y_{j,n_j} \mid \boldsymbol{S}_j,\boldsymbol{\tau},M &\overset{\text{iid}}{\sim} \sum_{m=1}^{M} w_{j,m} f\left(\cdot \mid \tau_m\right) \\
\tau_1,\dots,\tau_M \mid M &\overset{\text{iid}}{\sim} P_0(\cdot) \\
w_{j,1},\dots,w_{j,M} \mid M,\gamma_j &\overset{\text{iid}}{\sim} \mathrm{Dir}_M\left(\gamma_j,\dots,\gamma_j\right), \quad \text{for } j=1,\dots,d \\
M \mid \Lambda &\sim \mathrm{Pois}_1(\Lambda).
\end{aligned} \tag{5}$$

We notice that, when $d=1$, the model in (5) coincides with the MFM model (Miller and Harrison, 2018; Frühwirth-Schnatter et al., 2021). It follows that the proposed model is an extension of the MFM model to hierarchical data and so we refer to it as the *Hierarchical Mixture of Finite Mixtures* (HMFM) model.

Remark The HMFM model can be framed into a more general and flexible class of Bayesian nonparametric models. In particular, the choice for the prior distribution of the number of components M and the unnormalized weights $S_{j,m}$, can be generalized with respect to the choices in (5), while keeping the mathematical tractability. The theoretical properties of this broader family, named *Vector of Normalized Independent Finite Point Processes*, and a detailed construction using point processes are presented in Section S1 and Section S2 of the Supplementary materials, respectively.

2.1 Clustering

It is worth noticing that the mixture model in (1) can be equivalently written as

$$y_{j,i} \mid \theta_{j,i} \overset{\text{ind}}{\sim} f(\,\cdot \mid \theta_{j,i}), \qquad \theta_{j,i} \mid P_j \overset{\text{iid}}{\sim} P_j$$

with $i = 1, \ldots, n_j$ and $j = 1, \ldots, d$. Under this formulation, we get rid of the integral in (1) by introducing a latent variable $\theta_{j,i}$ for each observation $y_{j,i}$. Conditionally on $(P_1, \ldots, P_d)$, the latent variables $\theta_{j,i}$'s are i.i.d. within the same group and independent across groups. In other words, by virtue of the de Finetti representation theorem, $\boldsymbol{\theta} := (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$, where $\boldsymbol{\theta}_j = (\theta_{j,1}, \ldots, \theta_{j,n_j})$, is a sample from a partially exchangeable array of latent variables. This is tantamount to saying that the distribution of the $\theta_{j,i}$'s is invariant under a specific class of permutations; see (Kallenberg, 2005) and references therein. The distributional properties of $\boldsymbol{\theta}$ play a pivotal role in mixture models both in defining the clustering and devising efficient computational procedures. More precisely, since the P_j's in (4) have discrete realizations, almost surely, the latent variables feature ties within and across groups; thus they naturally induce both a local and a global clustering of the observations.

Local clustering For each group $j = 1, \ldots, d$, P_j is almost surely discrete, then ties are expected with positive probability among $\theta_{j,1}, \ldots, \theta_{j,n_j}$. Let $K_j := K_{j,(n_j)}$ be the random number of distinct values in this sample. Such group-specific distinct values are collected in the set $\mathcal{T}_j = \left\{\theta^*_{j,1}, \ldots, \theta^*_{j,K_j}\right\}$. Furthermore, let $\widetilde{\rho}_j = \left\{C_{j,1}, \ldots, C_{j,K_j}\right\}$ be the random partition of $\{1, \ldots, n_j\}$ induced by $\mathcal{T}_j$ through the following rule:

$$\theta_{j,i} \in C_{j,k} \iff \exists k \in \{1, \ldots, K_j\} \text{ such that } \theta_{j,i} = \theta^*_{j,k},$$

for each $i = 1, \ldots, n_j$. Note that the random partition $\widetilde{\rho}_j$ is exchangeable due to the exchangeability of $\theta_{j,1}, \ldots, \theta_{j,n_j}$ and it is called *local clustering* of group j (or group-specific clustering).

Global clustering Since the P_j's share the same support, we also expect ties between groups, i.e., $\mathbb{P}\left(\theta^*_{j,k} = \theta^*_{j',k'}\right) > 0$, with $j \neq j'$. We define $\mathcal{T} = \{\theta^{**}_1, \ldots, \theta^{**}_K\}$ the set of unique values among the $\theta^*_{j,k}$, $j = 1, \ldots, d$ and $k = 1, \ldots, K_j$; we also observe that $\mathcal{T} = \bigcup_{j=1}^d \mathcal{T}_j$. The corresponding random partition $\rho = \{C_1, \ldots, C_K\}$ is induced by $\mathcal{T}$ as follows

$$\theta_{j,i} \in C_k \iff \exists k \in \{1, \ldots, K\} \text{ such that } \theta_{j,i} = \theta^{**}_k,$$

for each $j = 1, \dots, d$ and for each $i = 1, \dots, n_j$. The random partition ρ is called *global clustering* and the number of global clusters has been denoted by $K := K_{(n_1,\dots,n_d)}$. Since values are expected to be shared also across groups, then $K \leq \sum_{j=1}^{d} K_j$.

To shed light on the relationship between local and global clustering, we introduce $\rho_j = \{C_{j,1}, \dots, C_{j,K}\}$ so that $C_k = \bigcup_{j=1}^{d} C_{j,k}$. Note that $C_{j,k}$ can be empty for some $k = 1, \dots, K$. We define $n_{j,k} = |C_{j,k}|$ the number of observations of the j-th group in the k-th cluster. Then, we let $\boldsymbol{n}_j = (n_{j,1}, \dots, n_{j,K})$ and the following hold true

$$\sum_{j=1}^{d} n_{j,k} > 0, \text{ and } \sum_{k=1}^{K} n_{j,k} = n_j, \tag{6}$$

for any $k = 1, \dots, K$ and for any $j = 1, \dots, d$, respectively.

The random partition induced by the whole sample $\boldsymbol{\theta}$ of size $n := n_1 + \cdots + n_d$ may be described through a probabilistic object called *partially Exchangeable Partition Probability Function* (pEPPF). The pEPPF, denoted here as $\Pi_K^{(n)}(\boldsymbol{n}_1, \dots, \boldsymbol{n}_d)$ is the probability distribution of both the local and global clustering, where $\boldsymbol{n}_1, \dots, \boldsymbol{n}_d$ satisfy the constraints given in (6). The pEPPF is formally defined as follows:

$$\Pi_K^{(n)}(\boldsymbol{n}_1, \dots, \boldsymbol{n}_d) := \mathbb{E}\left[\int_{\Theta^K} \prod_{k=1}^{K} P_j^{n_{j,k}}(\mathrm{d}\theta_k^{**})\right].$$

We refer to Camerlenghi et al. (2019) for additional details.

Observe that, conditionally to M and $\boldsymbol{\tau}$, the unique values $\theta_1^{**}, \dots, \theta_K^{**}$ are such that the following properties hold: (i) $K \leq M$; (ii) there exists $m \in \{1, \dots, M\}$ such that $\theta_k^{**} = \tau_m$, for each $k = 1, \dots, K$. Property (i) implies that a distinction between allocated mixture components (clusters) and non-allocated components is required (Nobile, 2004; Argiento and De Iorio, 2022). From property (ii), it follows that there are exactly K allocated components collected in the following set:

$$\mathcal{M}^{(a)} = \{m \in \{1, \dots, M\} : \exists k \in \{1, \dots, K\} \text{ such that } \theta_k^{**} = \tau_m\}.$$

3 Properties of the HMFM model

3.1 Distributional results

In this section, we derive all the theoretical properties for the latent variables $\theta_{j,i}$ modeled as follows

$$\theta_{j,i} \mid P_j \overset{\text{iid}}{\sim} P_j, \quad (P_1, \dots, P_d) \sim \text{Vec-FDP}(\Lambda, \boldsymbol{\gamma}, P_0), \tag{7}$$

where $j = 1, \dots, d$ and $i = 1, \dots, n_j$. In addition, we assume conditional independence across groups, i.e., $\boldsymbol{\theta}_j, \boldsymbol{\theta}_l \mid P_j, P_l$ are independent for $j \neq l$. The distributional results derived for the model in (7) are the theoretical guidance to understand the clustering mechanism, to elicit the prior properly, and they play a pivotal role in devising efficient marginal and conditional algorithms to perform posterior inference. The proofs of the theoretical properties are obtained as a special case of the more general theorems presented in Section S3 of the Supplementary materials.

Before moving forward to the main results, we remind that for a Gamma$(\gamma_j, 1)$ distributed random variable, its Laplace transform $\psi_j(u_j)$ and the corresponding derivative $\kappa(u_j, n_{j,k})$ equal to

$$\psi_j\left(u_j\right) = \frac{1}{\left(1+u_j\right)^{\gamma_j}}, \quad \kappa_j\left(u_j, n_{j,k}\right) = \frac{1}{\left(1+u_j\right)^{n_{j,k}+\gamma_j}} \frac{\Gamma\left(n_{j,k}+\gamma_j\right)}{\Gamma\left(\gamma_j\right)}. \tag{8}$$

The almost sure discreteness of the P_j, coupled with their common supports, entails that the hierarchical sample $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ is equivalently characterized by $(\boldsymbol{\theta}^{**}, \rho)$, previously defined in Section 2.1. The following theorem specifies the distribution of the pEPPF for the HMFM model.

Theorem 3.1 (pEPPF)**.** *The probability to observe a sample $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ of size n from (7) exhibiting K distinct values $(\theta_1^{**}, \ldots, \theta_K^{**})$ with respective counts $\boldsymbol{n}_1, \ldots, \boldsymbol{n}_d$ is given by the following pEPPF*

$$\Pi_K^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_d) = \mathrm{V}(K; \boldsymbol{\gamma}, \Lambda) \prod_{j=1}^{d} \prod_{k=1}^{K} \frac{\Gamma\left(n_{j,k}+\gamma_j\right)}{\Gamma\left(\gamma_j\right)}, \tag{9}$$

where $\mathrm{V}(K; \boldsymbol{\gamma}, \Lambda)$ *equals*

$$\mathrm{V}(K; \boldsymbol{\gamma}, \Lambda) = \int_{(\mathbb{R}^+)^d} \Psi(K, \boldsymbol{u}) \prod_{j=1}^{d} \frac{u_j^{n_j-1}}{\Gamma\left(n_j\right)} \frac{1}{\left(1+u_j\right)^{n_j+K\gamma_j}} \mathrm{d}u_1 \ldots \mathrm{d}u_d,$$

and, setting $\boldsymbol{\psi}(\boldsymbol{u}) = \prod_{j=1}^{d} \psi_j\left(u_j\right)$*,* $\Psi(K, \boldsymbol{u})$ *is defined as follows*

$$\Psi(K, \boldsymbol{u}) \ = \ \Lambda^{K-1}\left(K + \Lambda \boldsymbol{\psi}(\boldsymbol{u})\right) e^{-\Lambda(1-\boldsymbol{\psi}(\boldsymbol{u}))}. \tag{10}$$

See Section S3.1 for an extended discussion and derivation of $\Psi(K, \boldsymbol{u})$.

The predictive distributions, i.e., the distribution of θ_{j,n_j+1} given $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ for each possible group j, easily follow from Theorem 3.1. These unveil important intuitions about the clustering mechanism, and they are the building block for a marginal posterior sampler; we defer their detailed presentation to Section 3.2. Recall that in Section 2.1 we defined the number of global clusters, denoted as $K_{(n_1,\ldots,n_d)}$. This is a random quantity and in the case of $d = 2$ groups whose cardinalities are n_1 and n_2, respectively, we can compute an explicit expression for the prior distribution of $K_{(n_1,n_2)}$.

Theorem 3.2 (Prior distribution of global number of clusters)**.** *Consider $d = 2$ groups of observations with size n_1 and n_2, respectively. Then, the prior distribution for the* *global number of clusters $K_{(n_1,n_2)}$ is*

$$\begin{aligned}\mathbb{P}&\left(K_{(n_1,n_2)}=K\right)\\&=V(K;\boldsymbol{\gamma},\Lambda)\sum_{r_1=0}^{K}\sum_{r_2=0}^{K-r_1}\binom{K-r_1}{r_2}\frac{(K-r_2)!}{r_1!}\prod_{j=1}^{2}\left|C(n_j,K-r_j;-\gamma_j)\right|,\end{aligned}\tag{11}$$

where for any non-negative integers $n\geq 0$ and $0\leq K\leq n$, $C(n,K;-\gamma_j)$ denotes the central generalized factorial coefficients. See Charalambides (2002).

The following theorems provides functionals of $(P_1,\dots,P_d)$, which are useful for prior elicitation and to shed light on the dependence structure introduced by our prior.

Theorem 3.3 (Mixed moments)**.** *Let $(P_1,\dots,P_d)\sim$ Vec-FDP $(\Lambda,\boldsymbol{\gamma},P_0)$ be a vector of normalized random probability measures defined through normalization as in (3). Then the following hold:*

(i) for any measurable sets A,B and for any $j,l\in\{1,\dots,d\}$,

$$\mathbb{E}\left[P_j(A)P_l(B)\right]=\mathbb{P}\left(K_{(1,1)}=1\right)\left(P_0(A\cap B)-P_0(A)P_0(B)\right);\tag{12}$$

(ii) for any measurable set A and for any $j,l\in\{1,\dots,d\}$,

$$\mathbb{E}\left[P_j(A)^{n_j}P_l(A)^{n_l}\right]=\mathbb{E}\left[P_0(A)^{K_{(n_j,n_l)}}\right]=\sum_{k=1}^{n_j+n_l}P_0(A)^k\mathbb{P}\left(K_{(n_j,n_l)}=k\right),\tag{13}$$

where $K_{(n_j,n_l)}$ is the global number of clusters across two groups with size n_j and n_l and it can be evaluated via (11).

Furthermore, both (12) and (13) can be extended to the case of more than two groups. As a byproduct of Theorem 3.3 we obtain a closed form expression for pairwise correlation between the components of $(P_1,\dots,P_d)$ evaluated on specific sets. Let A be a measurable set, then, for any $j,l\in\{1,\dots,d\}$:

$$\operatorname{corr}\left(P_j(A),P_l(A)\right)\;=\;\frac{1-e^{-\Lambda}}{\Lambda\left(\gamma_j+1\right)\left(\gamma_l+1\right)I\left(\gamma_j,\Lambda\right)I\left(\gamma_l,\Lambda\right)},\tag{14}$$

where $I\left(\gamma_j,\Lambda\right)=\int_0^1\left(1+\Lambda x\right)e^{-\Lambda(1-x)}(1-x^{1/\gamma_j})\mathrm{d}x$. The expression in (14) does not depend on the choice of the set A. Thus, it may be considered an overall measure of dependence between the two random probability measures. The limits of (14) when both γ_j and γ_l goes to 0 and $+\infty$ equal to

$$\lim_{\gamma_j,\gamma_l\to 0}\operatorname{corr}\left(P_j(A),P_l(A)\right)\;=\;\frac{1-e^{-\Lambda}}{\Lambda},\qquad\lim_{\gamma_j,\gamma_l\to\infty}\operatorname{corr}\left(P_j(A),P_l(A)\right)\;=\;1.\tag{15}$$

These limits are interesting because we see that, given Λ, decreasing γ_j and γ_l, the correlation does not go to 0 but reaches a lower bound that depends on Λ, which, in turn, goes to 0 if Λ increases. On the other hand, increasing values of γ_j and γ_l lead correlation equal to 1, regardless of the value of Λ. See the left panel in Figure S1. We refer to Section S3.6 of the Supplementary materials for the proofs of (14) and (15) and a generalization of (14).

We now aim at giving a posterior characterization for a vector $(P_1, \ldots, P_d)$ distributed as in (4). Since $(P_1, \ldots, P_d)$ is obtained via normalization of $(\mu_1, \ldots, \mu_d)$, it is sufficient to provide a posterior characterization for the latter vector. In order to do this, we follow the same approach of James et al. (2009), Camerlenghi et al. (2019) and Argiento and De Iorio (2022). Thus, we introduce a vector of auxiliary variables $\boldsymbol{U}_n = (U_1, \ldots, U_d)$ such that $U_j \mid T_j \overset{\text{ind}}{\sim} \text{Gamma}\,(n_j, T_j)$, where $T_j = \mu_j(\mathbb{X})$. This is possible since the marginal distribution of $\boldsymbol{U}_n$ does exist, see Section S3.8. Hence, conditionally to $\boldsymbol{U}_n$ and to $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$, $(\mu_1, \ldots, \mu_d)$ is a superposition of two independent processes, one driving the non-allocated components and the other one driving the allocated components.

Theorem 3.4 (Posterior representation)**.** *Let $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ be a sample from the statistical model in (7). Then, the posterior distribution of $(\mu_1, \ldots, \mu_d)$ is characterized as the superposition of two independent processes on $(\mathbb{R}^+)^d \times \Theta$:*

$$(\mu_1, \ldots, \mu_d) \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d, \boldsymbol{U}_n \overset{\text{d}}{=} \left(\mu_1^{(a)}, \ldots, \mu_d^{(a)}\right) + \left(\mu_1^{(na)}, \ldots, \mu_d^{(na)}\right), \text{ where:}$$

(i) the process of allocated components $\left(\mu_1^{(a)}, \ldots, \mu_d^{(a)}\right)$ equals

$$\mu_j^{(a)} = \sum_{k=1}^{K} S_{j,k}^{(a)} \delta_{\theta_k^{**}}, \text{ as } j = 1, \ldots, d,$$

where the random variables $S_{j,k}^{(a)} \mid \boldsymbol{U}_n \overset{ind}{\sim} \text{Gamma}(n_{j,k} + \gamma_j, u_j + 1)$, for each $j \in \{1, \ldots, d\}$ and $k \in \{1, \ldots, K\}$;

(ii) the process of non-allocated components $\left(\mu_1^{(na)}, \ldots, \mu_d^{(na)}\right)$ equals

$$\mu_j^{(na)} = \sum_{m^*=0}^{M^*} S_{j,k}^{(na)} \delta_{\tau_{m^*}}, \text{ as } j = 1, \ldots, d.$$

In particular, M^* is a random variable distributed as a mixture of Poisson distributions, namely

$$M^* \sim (1 - w_K(\boldsymbol{u}))\, \text{Pois}_1 \left(\Lambda \prod_{j=1}^{d} \psi_j\,(u_j)\right) + w_K(\boldsymbol{u})\, \text{Pois}\left(\Lambda \prod_{j=1}^{d} \psi_j\,(u_j)\right),$$

where we have set $w_K(\boldsymbol{u}) := K/(\Lambda \prod_{j=1}^{d} \psi_j\,(u_j) + K)$, and the random variables $S_{j,1}^{(na)}, \ldots, S_{j,M^*}^{(na)} \mid \boldsymbol{U}_n, M^* \overset{iid}{\sim} \text{Gamma}(\gamma_j, u_j + 1)$, for $j \in \{1, \ldots, d\}$.

Note that for the process of non-allocated component $\mathbb{P}\,(M^* = 0) > 0$ even if $\mathbb{P}\,(M = 0) = 0$. Hence, it is possible to have zero non-allocated components.

3.2 Predictive distribution and franchise metaphor

Further intuitions on the cluster mechanism described in Section 2.1 are available when considering the predictive distributions. Consider a realization $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d)$ with K distinct values $(\theta_1^{**}, \dots, \theta_K^{**})$ and a partition $\rho = \{C_1, \dots, C_K\}$ with counts $(\boldsymbol{n}_1, \dots, \boldsymbol{n}_d)$ satisfying the constraints in (6). Following the approach of James et al. (2009), Favaro and Teh (2013) and Argiento and De Iorio (2022) we work conditionally to $\boldsymbol{U}_n = \boldsymbol{u}$. Then, for each group, say j, we have

$$\begin{aligned}\mathbb{P}\left(\theta_{j,n_j+1} \in \cdot \mid \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d, \boldsymbol{u}, \boldsymbol{\gamma}, \Lambda\right)& \\ \propto \sum_{k=1}^{K} (n_{j,k} + \gamma_j)\, \delta_{\theta_k^{**}}(\cdot) + \boldsymbol{\psi}(\boldsymbol{u})\gamma_j \Lambda \frac{K + 1 + \Lambda \boldsymbol{\psi}(\boldsymbol{u})}{K + \Lambda \boldsymbol{\psi}(\boldsymbol{u})} P_0(\cdot).&\end{aligned} \tag{16}$$

Such a predictive distribution can be interpreted in terms of a restaurant franchise metaphor. Consider a franchise of d Chinese restaurants each with possibly infinitely many tables. Here $\theta_{j,i}$ represents the dish served to customer i in restaurant j, and each θ_k^{**} represents a dish. All customers sitting at the same table must eat the same dish. The same dish can not be served at different tables in the same restaurant, but it can be served across different restaurants. According to the predictive law, the first customer entering the first restaurant sits at the first table eating dish $\theta_{1,1} = \theta_1^{**}$, which is drawn from P_0. At the same time, an empty table serving dish θ_1^{**} must be prepared in all the other restaurants: this step corresponds to the first cluster allocation, i.e., C_1. Then, the second customer of the first restaurant arrives and can either: (i) sit at the same table as the first customer, with probability proportional to $1 + \gamma_1$ or (ii) sit at a new table with probability proportional to $\boldsymbol{\psi}(\boldsymbol{u})\gamma_1\Lambda\frac{2+\Lambda\boldsymbol{\psi}(\boldsymbol{u})}{1+\Lambda\boldsymbol{\psi}(\boldsymbol{u})}$. In the latter case, the customer chooses a new dish θ_2^{**}, drawn from P_0, and the number of clusters K is increased by 1; moreover, an empty table serving dish θ_2^{**} must be prepared in all the other restaurants of the franchise. Then, the process evolves according to (16). Figure S2 displays a graphical representation of the process.

Interestingly, our model is more parsimonious than the HDP by Teh et al. (2006) in sharing information across restaurants. While the HDP relies on the popularity of a dish throughout the entire franchise to influence a new customer's choice, in our model, such probability depends on the sample only through the dish's popularity within the specific restaurant the customer enters. This distinctive feature proves to be appealing as it mitigates the excessive borrowing of information across groups that is induced by hierarchical processes. Subsequent numerical experiments highlight the advantage of our model by showing that the HDP can lead to misleading results in posterior inference. An additional advantage of our model with respect to the HDP is that different tables within the same restaurant cannot serve the same dish. This simplifies the local clustering structure, and it improves the computational efficiency in posterior sampling, as discussed in Section 4.2. We conclude by reminding that there are situations where the stronger capability of sharing information of the HDP can be a preferable feature. Experiment 1 in Section 5.1 is one such case, although the HMFM still remains a competitive alternative.

4 Fitting details

4.1 Hyperpriors

We consider the following hyperpriors for the process hyperparameters $(\Lambda, \boldsymbol{\gamma})$,

$$\pi(\Lambda, \boldsymbol{\gamma}) = \pi(\boldsymbol{\gamma} \mid \Lambda)\pi(\Lambda) = \prod_{j=1}^{d} \text{Gamma}(a_\gamma, \Lambda b_\gamma) \times \text{Gamma}(a_\Lambda, b_\Lambda). \tag{17}$$

The prior distribution in (17) extends, to our setting, the prior choice introduced by Frühwirth-Schnatter and Malsiner-Walli (2019) to encourage sparsity in the mixture, whose advantages have been studied both theoretically (Rousseau and Mengersen, 2011; Van Havre et al., 2015) and empirically (Malsiner-Walli et al., 2016, 2017). Furthermore, this prior formulation assumes the γ_js to be conditionally independent given Λ, so tuning the sharing of information between groups, see also Figure S1. In particular, note that $\Lambda \mid \boldsymbol{\gamma}$ is still Gamma distributed, i.e., $\Lambda \mid \boldsymbol{\gamma} \sim \text{Gamma}\left(a_\Lambda + d a_\gamma, b_\Lambda + b_\gamma \sum_{j=1}^{d} \gamma_j\right)$, which yields tractable posterior inference. We provide practical guidelines for setting the values of hyperparameters $(a_\gamma, b_\gamma, a_\Lambda, b_\Lambda)$ by exploiting the equivalent sample principle (Diaconis and Ylvisaker, 1979). To this end, we design a suitable reparametrization of the prior based on three quantities, Λ_0, V_Λ and γ_0. The first two quantities represent the prior expected value and variance of Λ, respectively, while γ_0 represent a common prior guess on γ_j. Thus, the new specification relies on hyperparameters that are easy to interpret and allows to elicit prior knowledge when available. Further details are presented in Section S5.1.

4.2 Computational methods

As customary in hierarchical modelling, we introduce latent allocation vectors $\boldsymbol{c}_j = \left(c_{j,1}, \dots, c_{j,n_j}\right)$ whose element $c_{j,i} \in \{1, \dots, M\}$ denotes to which component observation $y_{j,i}$ is assigned, for each $j = 1, \dots, d$. Setting $\theta_{j,i} = \tau_{c_{j,i}}$, we are able to link the mixture parameters and the observation-specific parameters. We suggest two MCMC strategies to carry out posterior inference for mixture modelling. The first one is a conditional algorithm that provides full Bayesian inference on both the mixing parameters $(P_1, \dots, P_d)$ and the clustering structure ρ. Namely, we draw a sample of the vector of random probability measures $(P_1, \dots, P_d)$ from its posterior distribution given in Theorem 3.4 by sampling from the joint posterior distribution of $(\boldsymbol{S}_1, \dots, \boldsymbol{S}_d, \boldsymbol{\tau}, \boldsymbol{c}_1, \dots, \boldsymbol{c}_d, M)$. Note that the global number of clusters K is automatically deduced from the cluster allocation vectors $(\boldsymbol{c}_1, \dots, \boldsymbol{c}_d)$. To do so, we resort to auxiliary variables $\boldsymbol{U}_n$ and the hyperparameters $(\Lambda, \boldsymbol{\gamma})$. For the sake of brevity, we denote $\boldsymbol{S} = (\boldsymbol{S}_1, \dots, \boldsymbol{S}_d)$ and $\boldsymbol{c} = (\boldsymbol{c}_1, \dots, \boldsymbol{c}_d)$. We adopt a blocked Gibbs sampling strategy. In particular, let $\Delta = (\boldsymbol{S}, \boldsymbol{\tau}, \boldsymbol{c}, M, \boldsymbol{U}, \boldsymbol{\gamma}, \Lambda)$ be a vector collecting all the parameters and let $\boldsymbol{y}$ be the collection of all variables $y_{j,i}$, for each group j and for each individual i. We partition Δ in two blocks $\Delta_1 = (\boldsymbol{S}, \boldsymbol{\tau}, \boldsymbol{c}, \Lambda)$ and $\Delta_2 = (\boldsymbol{U}, M, \boldsymbol{\gamma})$. The sampling scheme proceeds by iterating two steps: (i) sampling Δ_1 conditionally to Δ_2 and $\boldsymbol{y}$; (ii) sampling Δ_2 conditionally to Δ_1 and $\boldsymbol{y}$. We refer to Section S5.2 for a step-by-step description of the algorithm, where all full conditional distributions are detailed together with their computational times.

The second algorithm is a marginal sampler that simplifies the computation by integrating the mixture parameters while providing inference on the sole clustering structure. The algorithm is derived from the predictive distributions detailed in Section 3.2, and the full conditional distribution of $\boldsymbol{U}_n$ given in Theorem S1.3. A detailed description of the marginal algorithm is given in Section S5.3, along with the derivation of its computational time.

Notably, the HMFM achieves linear scaling, taking $O(n(M+K))$ and $O(n(K+1))$ time for one iteration of the conditional and marginal sampler, respectively. In contrast, the HDP faces scalability issues that may eventually limit its practical feasibility. A naive implementation of HDP, based on the traditional Chinese restaurant franchise process, takes $O(n^2)$ time. To overcome this computational burden, Teh et al. (2006) also propose a direct assignment scheme whose computational bottleneck is the computation of the unsigned Stirling numbers $s(n_{j,k}, m)$ for each $j = 1, \dots, d$, $k = 1, \dots, K$ and for all positive integers $m \leq n_{j,k}$, where $n_{j,k}$ is the number of observations in group j assigned to cluster k. The computational time to compute $s(n_{j,k}, m)$ is $O((n_{j,k})^2)$. By doing so, the quadratic cost of the algorithm is deferred to the calculation of the Stirling numbers which, however, can be precomputed and saved. Once they are available, the cost per iteration is also linear for the HDP. Thus, while the precomputation of Stirling numbers makes HDP competitive with HMFM for moderate values of n the linear complexity of our proposed method makes it more scalable and appealing for large datasets.

The R code implementing both MCMC algorithms is available at `https://github.com/alessandrocolombi/HMFM`, along with the simulation study.

5 Simulation study

We carried out an extensive simulation study comparing the HMFM in (5) with: (i) the HDP (Teh et al., 2006) mixture model; (ii) the MFM model assumed independently for each group (MFM-indep); (iii) the MFM model assumed for the pooled data, i.e., ignoring the groups (MFM-pooled). The MFM model is fitted using the algorithm by Argiento and De Iorio (2022). Regarding the HMFM, we investigate the performance of both the conditional (HMFM-cond) and the marginal (HMFM-marg) sampler. To assess the performance of recovering the true clustering, we compute the Co-Clustering Error (CCE; Dahl 2006; Bassetti et al. 2020) and the Adjusted Rand Index (ARI; Hubert and Arabie 1985). Additional details are given in Section S6, together with a comparison in terms of density estimation.

In summary, the simulation study consists of three experiments: the first one considers $d = 2$ groups that share a component and clearly shows the advantage of the joint modelling approach against both independent group-specific analyses and pooled analyses. The second experiment is an illustrative example with again $d = 2$ groups, but without any components shared across the groups. The study highlights the limitations of the HDP in situations where borrowing information from other groups can lead to misleading conclusions. Rather, this issue is mitigated by the HMFM that borrows less information across the groups relative to the HDP. The third experiment generates data from $d = 15$ groups and further evidences how the HMFM outperforms the HDP; the results of this experiment are deferred to Section S6 of the Supplementary materials.

5.1 Experiment 1

This experiment considers $d = 2$ groups, both having two local clusters ($K_1 = 2$, $K_2 = 2$) one of which is shared; hence, the global number of clusters is $K = 3$. The mixing probabilities are set so that the shared component has a lower value in the second group. In particular, 50 independent datasets are generated from $y_{1,i} \overset{\text{iid}}{\sim} 0.5\text{N}(-3, 0.1) + 0.5\text{N}(0, 0.5)$ and $y_{2,i} \overset{\text{iid}}{\sim} 0.2\text{N}(0, 0.5) + 0.8\text{N}(1.75, 1.5)$, for $i = 1, \ldots, 300$. Note that the second group is defined so that the two components strongly overlap. As a result, the shared component is completely masked in this group. Nevertheless, the masked component can be spotted by exploiting the sharing of information with the other group. Figure 1 shows the empirical distribution of a simulated dataset, as well as the pooled data, and the underlying densities.

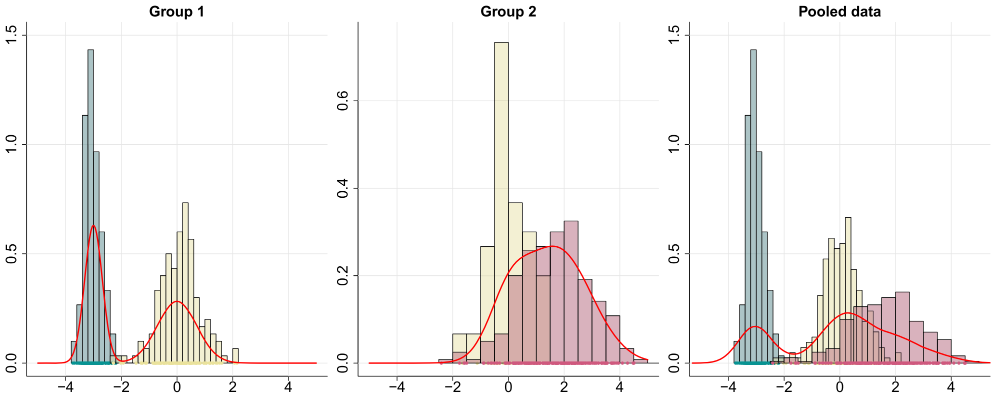

Figure 1: Empirical distributions of a dataset simulated under Experiment 1. Dots represent the observations while lines represent the underlying densities. Colours relate to the mixing components.

For each simulated dataset, we fit the HDP, the independent group-specific MFM, and the proposed HMFM by setting $\Lambda_0 = 5$, $\text{V}_\Lambda = 5$, $\gamma_0 = 0.5$, following the guidelines in Section S5.1. The HDP and MFMs are fitted employing default hyperpriors (see Section S6). A clear advantage of joint modelling is the possibility of deriving model-based clustering also across different groups, which is not possible when we run independent analyses.

Table 1 presents the results of model comparison based on the mean and standard deviation (in brackets) of the ARI over the simulated datasets. The results show that

	Group 1	Group 2	% 1 cluster	Global
HMFM-cond	0.991 (0.012)	0.133 (0.185)	62%	0.720 (0.024)
HMFM-marg	0.991 (0.011)	0.130 (0.181)	62%	0.719 (0.024)
HDP	0.993 (0.010)	0.097 (0.169)	74%	0.721 (0.018)
MFM-indep	0.993 (0.011)	0.027 (0.100)	90%	–
MFM-pooled	0.971 (0.033)	0.100 (0.126)	28%	0.540 (0.071)

Table 1: The first two columns report the mean and the standard deviation (in brackets) of the ARI for the two groups over the 50 simulated datasets under Experiment 1. The third column shows the percentage of times that each method gathers all observations of the second group in a single cluster. The final column reports mean and standard deviation of the ARI relative to the final partition of the data.

all methods, with negligible differences, are able to perfectly recover the true clustering in the first group, where the two components are well separated. Instead, the advantage of the sharing of information allowed by hierarchical modeling is evident in the second group, where components overlap. Indeed, the MFM-indep fails to identify the presence of two different clusters in the second group, gathering all observations together and obtaining an ARI value close to zero. On the other hand, the HDP and the HMFM are able to borrow information from the first group to recognize the presence of two clusters in the second group, leading to higher values of ARI, with HMFM (both marginal and conditional samplers) outperforming the HDP. The MFM-pooled method represents an intermediate situation between independent analyses and hierarchical approaches. By pooling all the data, it retains information from the first group, enabling it to identify the presence of two clusters even in the second group with performance comparable to the HMFM and the HDP. To clarify, we compute the percentage of simulated datasets for which each method obtains a partition of the second group consisting only of a single cluster. Note that lower values are better as we know that the true number of clusters is two. The results are reported in the third column of Table 1 and show how the MFM-indep consistently fails to identify at least two clusters for the second group, showing the limitations of the independent analyses against the hierarchical approaches. This limitation is even more pronounced with respect to the MFM-pooled, which represents the extreme case of information sharing. By pooling all data together, the presence of two clusters in the second group becomes evident.

However, we point out that the competitiveness of the MFM-pooled compared to the HMFM and the HDP holds only when looking at the group-specific metrics. Rather, when we evaluate the global ARI in the final column of Table 1, i.e., the ARI relative to the final partition of all the data, it becomes clear that the pooled solution is still sub-optimal compared to the joint modeling; recall that the ARI for the global partition is not defined for MFM-indep.

To conclude the clustering comparison, Figure 2 provides a picture of the distribution of the CCE within each group over the simulated datasets. Unlike the ARI, which considers only the final clustering estimate, the CCE accounts for all MCMC iterations, thus better reflecting the posterior variance. However, since it relies on the group-specific posterior similarity matrix, it is not well-defined for MFM-pooled. Therefore, the latter

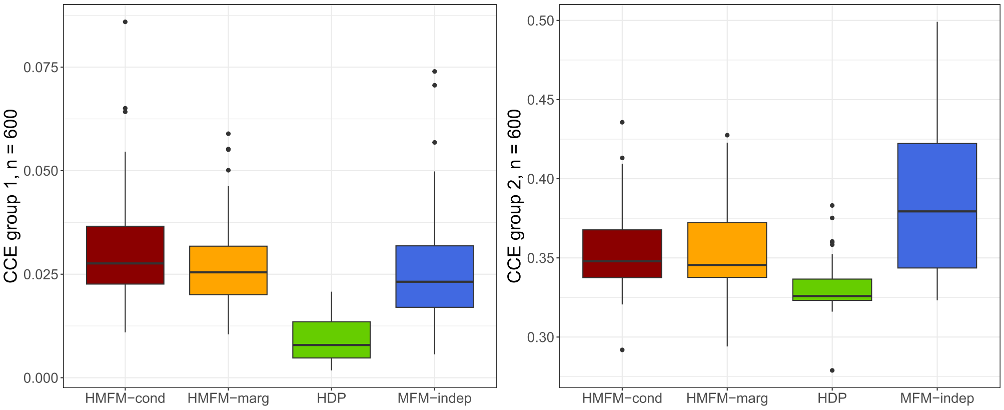

Figure 2: Co-Clustering Error in the first group (left) and second group (right). Boxplots are obtained over 50 datasets simulated under Experiment 1.

case is discarded from the results. In the first group, differences between the HMFM (conditional and marginal) and the MFM-indep are negligible. On the other hand, the MFM-indep commits a higher error in the second group, with also higher variability among different datasets, which confirms the findings reported in Table 1 about the sub-optimality of using independent analyses relative to joint modeling of the groups. Finally, we point out that the HDP performs better than the HMFM in the first and second groups. Indeed, in this example, the high sharing of information of the HDP enhances the inference as the two clusters strongly overlap, i.e., sharing as much information as possible from the two groups becomes beneficial. Hence, Figure 2 indicates that the HDP is preferable overall, considering all the MCMC iterations. However, such improvement becomes negligible, if not completely reversed, when considering only the final estimate obtained by minimizing the variation of information criterion, see Table 1.

5.2 Experiment 2

This experiment considers $d = 2$ groups coming from two components, not shared across the groups, namely $K_1 = 2$, $K_2 = 2$, so that $K = 4$. In particular, 50 datasets are generated from the model $y_{1,i} \overset{\text{iid}}{\sim} 0.5\text{N}(-3,1) + 0.5\text{N}(1,1)$ and $y_{2,i} \overset{\text{iid}}{\sim} 0.5\text{N}(-4,1) + 0.5\text{N}(0,1)$, each for $i = 1,\ldots,n_j$. We repeat the experiment for an increasing number of observations, $n = 50, 100, 200$, which we equally divided into the two groups. See Figure 3 for an example of the empirical distribution of a simulated dataset with $n = 100$ and the underlying densities. This experiment is designed to assess whether the borrowing of information may lead to misleading results in situations where groups do not share any features. Hence, for this scenario, we confine the comparison in terms of global clustering among the HMFM, the HDP, and the MFM-pooled. Here, the HMFM is fitted by setting $\Lambda_0 = 10$, $\text{V}_\Lambda = 2$, $\gamma_0 = 0.01$ while the HDP and MFM are fitted employing default hyperpriors.

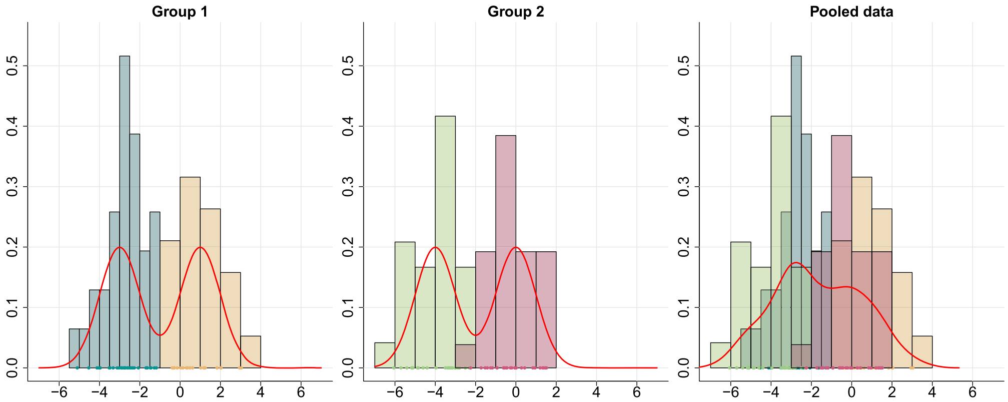

Figure 3: Empirical distributions of a dataset simulated under Experiment 2. Dots represent the observations, while lines represent the underlying densities. Colors relate to the mixing components.

	$n = 50$	$n = 100$	$n = 200$
HMFM-cond	0.60 (0.35)	0.74 (0.24)	0.90 (0.10)
HMFM-marg	0.59 (0.33)	0.74 (0.24)	0.89 (0.08)
HDP	0.30 (0.27)	0.48 (0.28)	0.73 (0.18)
MFM-pooled	0.36 (0.13)	0.40 (0.72)	0.41 (0.04)

Table 2: Mean and standard deviation (in brackets) of the ARI for the two groups over the 50 simulated datasets under Experiment 2.

Table 2 presents the results of model comparison based on the mean and standard deviation (in brackets) of the ARI over the simulated datasets. Although the scenario is simple, the table shows that the HDP struggles to find the underlying global clustering, and it is outperformed by the HMFM, both marginal and conditional. For what concerns MFM-pooled, Figure 3 clearly shows that discarding the group-membership information makes the clustering task much harder as all clusters strongly overlap. This explains the poor performances of the MFM-pooled reported in Table 2, which do not significantly improve increasing the sample size. The same conclusions can be drawn from the CCE relative to the global partition (Figure 4) which takes into account all the posterior pairwise probabilities of observations to be clustered together. Clearly, the HMFM outperforms both HDP and the MFM-pooled in terms of clustering estimation.

Finally, Figure 5 shows the frequencies of the estimated number of clusters over the 50 different datasets. The HMFM always prefers a number of clusters greater than two and, as the sample size increases, it selects four clusters, which is the true value. In contrast, the HDP tends to identify only two clusters when n is small, discarding this preference when more data are added. This example showcases that the poorer clustering performance of the HDP is due to the oversharing of information which can compromise the recovery of the true underlying partition. Indeed, in the extreme

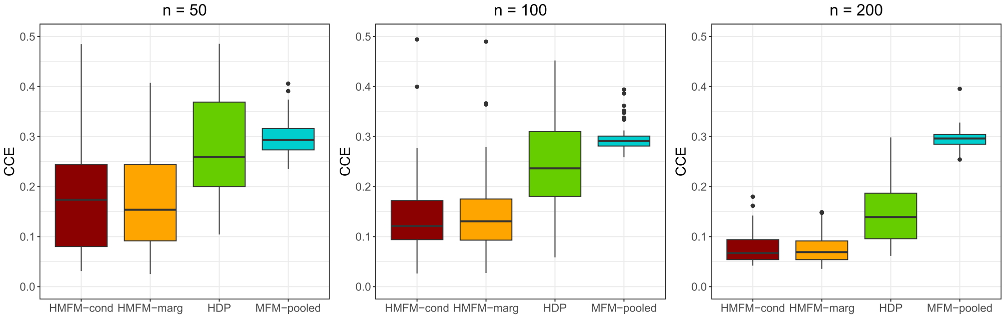

Figure 4: Co-Clustering Error for different sample sizes. Boxplots are obtained over 50 datasets simulated under Experiment 2.

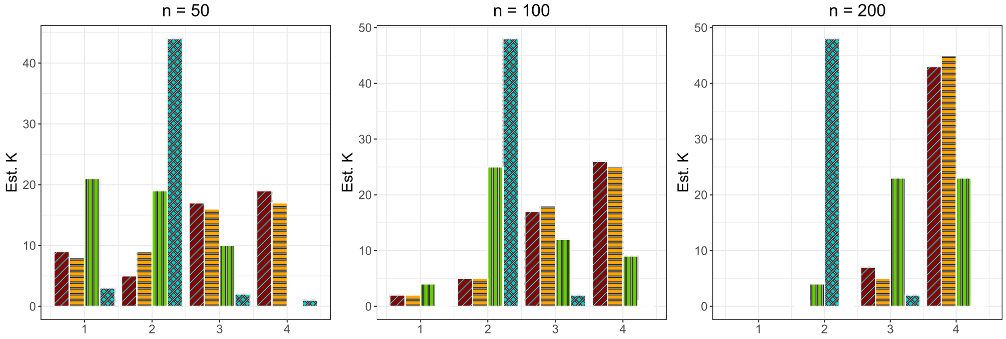

Figure 5: Estimated number of global clusters. Frequencies are obtained over 50 datasets simulated under Experiment 2. Red and orange bars represent the HMFM, conditional and marginal algorithms, respectively. Green bars are for the HDP, while the light blue bars are for the MFM-pooled.

situation (i.e., ignoring the groups), the MFM-pooled remains fixed at the two-cluster solution, even when the sample size increases.

To better clarify the clustering mechanism of the different models, we resort to the restaurant franchise metaphor discussed in Section 3.2. According to the HDP, when a new customer arrives, the probability of consuming a dish not yet served in that specific restaurant, but available in other franchise restaurants, hinges on the cumulative number of clients eating that dish across all restaurants. In contrast, the HMFM uses information only about whether a dish is being consumed or not in other restaurants. This experiment unveils a potential issue in the HDP's clustering mechanism; the concentration of all clients within the first restaurant (group) eating the same dish increases the probability of that dish being offered in the second restaurant (group), even though no components are shared between them. Rather, this confusion is circumvented by the HMFM's clustering mechanism. We notice that this phenomenon is directly tied to the choice of the prior, and its impact diminishes as the sample size increases, as shown in Table 2. Nevertheless, the HMFM model still outperforms the competing approaches even when $n = 200$.

6 Analysis of shot put data

Shot put is a track and field event in which athletes throw a heavy spherical ball, known as the shot, as far as possible. Our dataset comprises measurements, specifically the throw lengths or marks, recorded during professional shot put competitions from 1996 to 2016, for a total of $35,637$ measurements on 403 athletes. Each athlete's record includes the mark achieved, competition details, and personal information, namely, age, gender, and whether the event took place indoors or outdoors. The analyzed data are publicly available (`www.tilastopaja.eu`). Our objective is to model the seasonal performance for each shot putter, interpreted as the mean and variance of his/her seasonal marks. In particular, the season number assigned to each observation corresponds to the number of seasons the athlete has participated in, excluding seasons where he/she did not compete. For example, season 1 represents the athlete's first active season. This grouping of observations into seasons reflects the athletes' years of experience. Figure S9 in the Supplementary materials visually illustrates the performance evolution throughout the career of four randomly selected shot putters from the dataset. Each athlete has different participation in competitions, and the length and trajectory of their performance careers vary. While performance is expected to vary over the athlete's career, the figure evidences that the performance remains relatively consistent within each season. We characterized the seasonal performances as arising primarily from random fluctuations around a mean value. The values of this mean and the associated variability are unknown and are inferred from the data. Although it is a simplified representation, this captures the essential characteristics of athletes' careers.

In a previous study, Dolmeta et al. (2023) employed a generalized autoregressive conditional heteroskedasticity (GARCH) model to account for the volatility clustering of athletes' results over time. Rather, in this work, we frame the data into a hierarchical structure where each season represents a different group. Hence, we assume the HMFM model described in Section 2 for analyzing the athletes' performance; the proposed model allows us to capture the variability among different seasons and clustering the performances both within the seasons and across them.

Let n_j be the number of athletes competing in season j, with $j = 1, \ldots, d$. The longest career consists of 15 seasons, which is then the total number of groups $d = 15$. Each active athlete i in season j, with $i = 1, \ldots, n_j$, takes part in $N_{j,i}$ events. At each event, indexed by $h = 1, \ldots, N_{j,i}$, the athlete's mark $y_{j,i,h}$ is measured. Moreover, r event-specific covariates are available, $\boldsymbol{x}_{j,i,h} \in \mathbb{R}^r$, and collected in the design matrices $X_{j,i} \in \mathbb{R}^{N_{j,i} \times r}$.

Assuming that observations are noisy measurements of an underlying athlete-specific function, the model we employ for these data is $y_{j,i,h} = \mu_{j,i} + X_{j,i}\boldsymbol{\beta}_j + \varepsilon_{j,i,h}$, with $\varepsilon_{j,i,h} \overset{\text{iid}}{\sim} \mathrm{N}\left(0, \sigma^2_{j,i}\right)$, where $\mu_{j,i}$ is a season-specific random intercept, $\boldsymbol{\beta}_j$ is a r-dimensional vector of regression parameters, shared among all the athletes in season j, and $\sigma^2_{j,i}$ denotes the error variance. Therefore, within each season j, the athlete's observations $\boldsymbol{y}_{j,i} = (y_{j,i,1}, \ldots, y_{j,i,N_{j,i}})$ are distributed as

$$\boldsymbol{y}_{j,i} \mid \mu_{j,i}, \sigma^2_{j,i}, \boldsymbol{\beta}_j, X_{i,j} \overset{\text{ind}}{\sim} \mathrm{N}_{N_{j,i}} \left(\mu_{j,i} \mathbf{1}_{N_{j,i}} + X_{j,i} \boldsymbol{\beta}_j, \sigma^2_{j,i} \mathbf{I}_{N_{j,i}} \right), \tag{18}$$

where $\mathrm{N}_{N_{j,i}}$ denotes the $N_{j,i}$-dimensional normal distribution, $\mathbf{1}_{N_{j,i}}$ is a vector of length $N_{j,i}$ with all entries equal to 1 and $\mathbf{I}_{N_{j,i}}$ is the identity matrix of size $N_{j,i}$. To ensure identifiability, observations $\boldsymbol{y}_{j,i}$ have been centered within each season, i.e., $\sum_{i=1}^{n_j} \sum_{h=1}^{N_{j,i}} y_{j,i,h} = 0$ for each j.

Letting $\theta_{j,i} = \left(\mu_{j,i}, \sigma^2_{j,i} \right)$, we place a Vec-FDP prior for $\theta_{j,i}$ so that a clustering of athletes' performances both within and across different seasons is achieved. We assume a multivariate normal prior distribution for the regression coefficients, whose prior mean is denoted by $\boldsymbol{\beta}_0$ and covariance matrix Σ_0. We define $\boldsymbol{y}$ as the collection of all observations across seasons j and athletes i. Based on evidence from a previous analysis (Dolmeta et al., 2023) and for ease of interpretation, we use only gender as a covariate in our analysis. In particular, we use male athletes as reference baseline and set $\boldsymbol{\beta}_0 = -2\mathbf{1}_d$ and $\Sigma_0 = \mathbf{I}_d$ expecting male athletes to throw longer than females.

We set the base probability measure P_0 for $\theta = \left(\mu, \sigma^2 \right)$ to be a Normal-Inverse Gamma, exploiting the conjugacy with the likelihood in (18). In particular, the Normal-Inverse Gamma distribution is parametrized as is Hoff (2009), i.e.,

$$\left(\mu, \sigma^2 \right) \sim \operatorname{InvGamma}(\mu_0, k_0, \nu_0, \sigma_0^2) = \mathrm{N}\left(\mu_0, \frac{\sigma^2}{k_0} \right) \times \operatorname{Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2} \sigma_0^2 \right).$$

Following the approach of Richardson and Green (1997) and Lijoi et al. (2007a), we set $\mu_0 = 0$ and $k_0 = \frac{1}{\operatorname{range}(\boldsymbol{y})^2}$. Then, we set $\nu_0 = 4$ and $\sigma_0^2 = 10$ to have a vague InvGamma with infinite variance. For the process hyperparameters Λ and $\boldsymbol{\gamma}$, we set the hyperprior in (17). To achieve sparsity in the mixture, we follow the approach in Section S5.1, and set $\Lambda_0 = 25$, $\mathrm{V}_\Lambda = 3$ and $\gamma_0 = 1/\sum_{j=1}^d n_j = 0.00027$, leading to $a_\gamma = 13.89$, $b_\gamma = 2007.78$, $a_\Lambda = 208.33$, $b_\Lambda = 8.33$. The complete formulation of the hierarchical model can be found in Section S7. The burn-in period has been set equal to $50,000$, then $200,000$ additional iterations were run with a thinning of 10. The initial partition has been set using the k-means on the pooled dataset with 20 centrers. Posterior analysis is not sensible for such a choice.

Figure 6 shows the posterior 95% credible intervals of the regression coefficients. The posterior distribution is concentrated on negative values, meaning that the athletes' marks are, on average, higher for males than females. Also, the effect of gender on athletes' performance is significantly different across seasons, e.g., it is more evident in the first years and it reduces over career years, with the exception of the final season.

The final clustering has been obtained through minimization of the variation of information (Wade and Ghahramani, 2018; Dahl et al., 2022) loss function, and it consists of 13 clusters, with the posterior mode being equal to 15. Among these, we identify 11 main clusters since two of them capture noisy observations with high variance. The estimated clusters have been relabeled according to decreasing means.

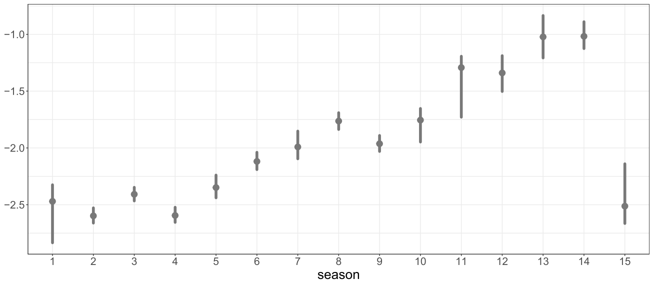

Figure 6: 95% posterior credible intervals of season-specific coefficients $\boldsymbol{\beta}$'s.

A remarkable finding is that the cluster interpretation does not depend on gender, whose effect has been filtered out by the season-specific parameter β. In other words, our clustering does not trivially distinguish between males and females, but it models the athletes' performance regardless of their gender. This claim is supported by the fact that, when ordering the clusters according to their means, both males' and females' average performance are ordered too, with the exception of one cluster which is made of female athletes only. Moreover, Figure 7 reports the athletes' marks, coloured according to their cluster membership, for male and female players, respectively. The two plots are similar, highlighting that the cluster interpretation is gender-free.

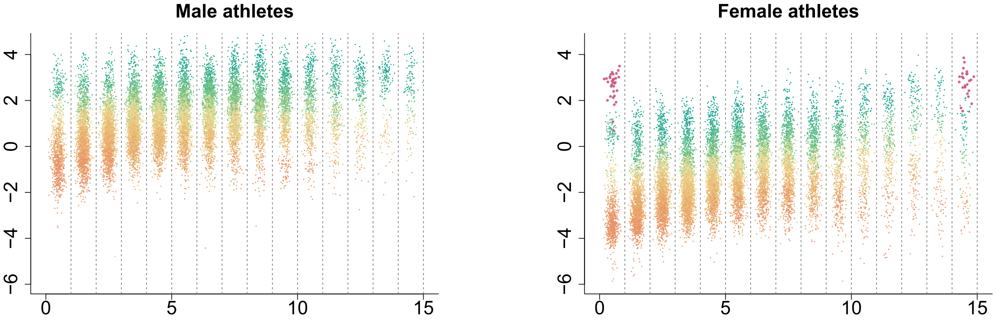

Figure 7: Shot put marks for male athletes (left panel) and female athletes (right panel). Vertical dotted lines delimit seasons. Dots are coloured according to their cluster membership.

Nevertheless, we are able to identify the presence of a particular cluster, whose points are highlighted in the right panel of Figure 7, including three exceptional women performances, which are much above the average mean throw for female athletes. No man belongs to such a cluster, meaning that no one has ever been able to outperform competitors in such a neat way. In particular, in this cluster, we find Astrid Kumbernuss, who is a three-time World champion and one-time Olympic champion; Valerie Adams, who, during her outstanding career, won two Olympic Games and four World Championships and Nadzeya Ostapchuk, who won a bronze medal at the Olympic Games.

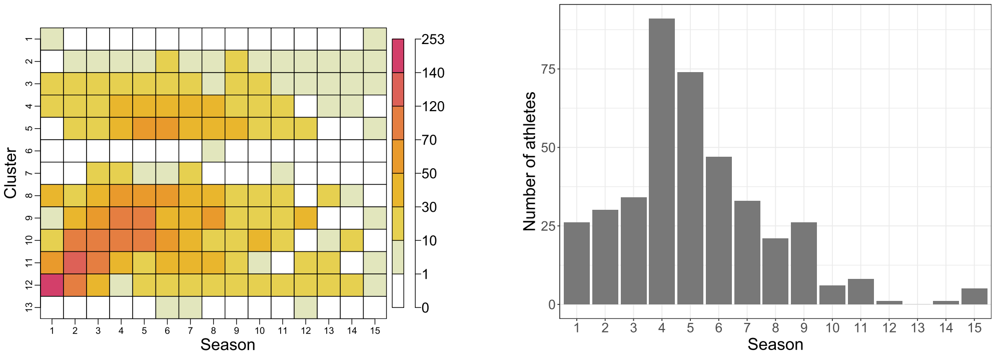

Figure 8: Left panel: local cluster sizes. Right panel: absolute frequencies of the seasons in which each athlete reaches their peak cluster for the first time, i.e., the one with the highest average.

According to the model (18), clusters are interpreted according to their corresponding mean and variances, but many insights can be gained from the clustering results. Notably, our analysis focuses on global considerations about the average evolution of the athletes' careers, as well as on the individual-specific sequence of clusters that characterizes every athlete. We refer to Section S7 for tables reporting all season-specific cluster sizes and cluster summaries and to the left panel of Figure 8 for a representation of the local cluster sizes. Finally, see Section S7.1 for a comparison with a naive modeling, which consists of fitting a mixture model to the pooled data.

From the global interpretation of the clustering results, we observe that in the first season of their careers, most of the athletes are grouped into the lowest-ranked cluster. This finding aligns with the expectation that rookies tend to exhibit similar performances. Then, the remaining athletes are divided among other low-ranked clusters, with the exception of some athletes who already belong to high-ranked clusters. The intermediate-level clusters are notably empty, highlighting a highly polarized situation. An interesting feature of our model is the ability to evaluate how the larger cluster changes across seasons. Notice in the left panel of Figure 8 the darkest squares progressively shift from the lowest-ranked cluster toward the intermediate-ranked clusters up to season 7. This trend likely reflects both the athletes' increasing experience and ongoing physical development during the early stages of their careers. However, from season 7 onwards, this progression is eventually tempered by external factors not captured by the model, such as training consistency and injuries. Furthermore, the right panel of Figure 8 shows the absolute frequencies of the seasons in which each athlete first reached the highest-ranked cluster of their career. This depicts a skewed distribution, with a mode at the season 4, and subsequent seasons being more frequent compared to the first three seasons. The frequencies decrease significantly from season 10 onwards.

The analysis of cluster summaries provides additional insights. From Table S2, we note that male athletes in the highest-ranked clusters have higher average ages than female athletes. This indicates that women tend to reach their peak performance at a younger age than men. A possible explanation for this disparity is that female bodies develop earlier than male bodies; this is supported by the fact that women begin their careers at an average age of 18.2, while men start at an average age of 19.4. Cluster 1 stands out from this trend as it comprises female athletes with a mean age of 28, which is considerably older than the average peak age. Lastly, it is worth noting that top-level clusters, specifically clusters 2 and 3, have a higher proportion of female athletes than male athletes. This is despite the overall number of observations for women being smaller than for men. This suggests that male competitions may be more balanced, making it more challenging for athletes to distinguish themselves from the average level.

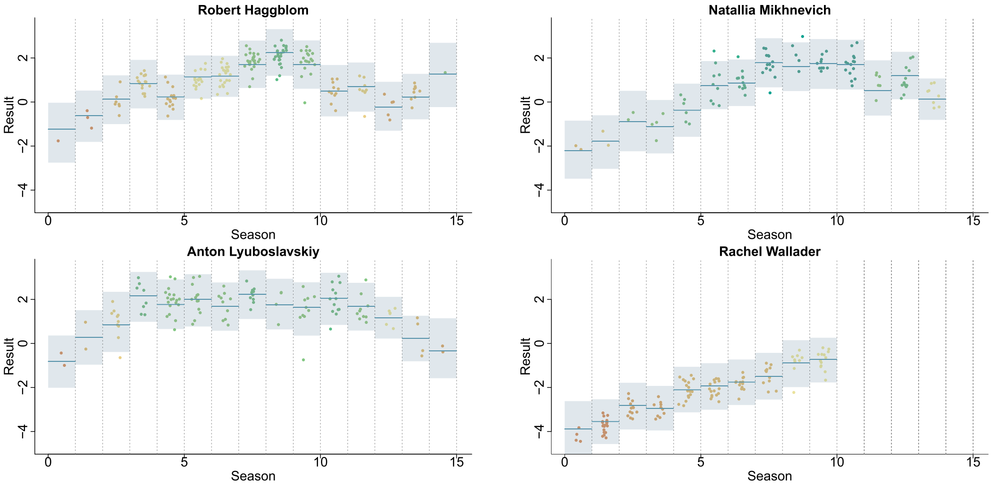

Figure 9: Shot put measurement for four randomly selected athletes. Points are coloured according to the cluster membership of the corresponding performance. Solid means represent the estimated cluster means. Shaded areas represent the 95% credible bands.

A key feature of our analysis is the possibility of studying the evolution of season-specific cluster membership of each player. This is exemplified in Figure 9, which showcases the trajectories of four players (two men and two women). In the plot, marks representing each season are colour-coded according to their cluster memberships, while solid lines represent the estimated seasonal mean performances, and the shaded areas represent 95% credible bands. One notable pattern emerges from the trajectory of Robert Häggblom, where a drop occurred immediately after the peak of his career, where he even participated in the 2008 Olympic games. A back injury conditioned the final part of its career. This sudden change would have been challenging to capture by a time-smoothing model. In contrast, Rachel Wallader demonstrates significant improvements throughout her career, starting from cluster 12, which we recall being the predominant cluster among rookies, and eventually reaching the higher-performing cluster before retiring, also managing to win the British title. Anton Lyuboslavskiy and Natallia Mikhnevich share remarkable career paths, showcasing exceptional performance not only for a single season but for extended periods of time, both around 9 years.

Indeed, both are Olympic level players. Unlike Rachel Wallader, Anton Lyuboslavskiy continued to compete beyond his prime, maintaining high levels of performance but eventually transitioning to intermediate cluster levels. Lastly, we highlight that Robert Häggblom and Natallia Mikhnevich achieved comparable marks, but the performances of the second athlete, a woman, are assigned to higher-ranked clusters. This demonstrates our model's ability to recognize top players, regardless of their gender. Indeed, Natallia Mikhnevich's career has been richer in success, as she won both gold and silver at the European Championships. Finally, we point out that we identify athlete-specific sequences of clusters, which could themselves be clustered to capture similarities and differences in the development trajectories of the athletes. A model-based solution to this task would necessitate moving beyond the framework of partial exchangeability across the seasons and adopting temporal modeling; see Page et al. (2022) for a possible alternative.

7 Discussion

We have introduced an innovative Bayesian nonparametric model for the analysis of grouped data, leveraging the normalization of finite dependent point processes as its foundation. Furthermore, we provided a comprehensive Bayesian analysis of this novel model class, delving into the examination of the pEPPF, posterior distributions and predictive distributions. A special emphasis has been placed on vectors of finite Dirichlet processes, which stand out as a noteworthy example in this context. Besides, we have also defined the HMFM as a natural extension of the work by Miller and Harrison (2018). Based on our theory, marginal and conditional algorithms have been developed. One significant benefit of HMFM is its ability to effectively capture dependence across groups. Moreover, we have empirically shown that HMFM better calibrates the borrowing of information across groups than a traditional HDP and also performs better in terms of computational time. In other words, the HMFM is a harmonious balance between HDP and independent analyses. Finally, the analysis of the shot put data illuminated the flexibility of the model employed to infer athlete career trajectories and group them into clusters with a meaningful interpretation.

We now pinpoint several open problems related to our work, which are left for future research. First, note that our construction allows a local and a global clustering. However, in numerous applications, one is also interested in clustering the different groups of observations. Nested structures, introduced by Rodríguez et al. (2008), have gained popularity as valuable Bayesian tools for accomplishing this task. Recent developments also include Denti et al. (2023) and D'Angelo et al. (2023). We intend to explore the use of NIFPP to simplify the complexity of nested models, and to alleviate the computational burden associated with traditional nested structures.

Other future directions of research aim to enhance between-group dependence based on our approach. An intriguing extension we plan to investigate is to adopt a construction akin to compound random measures (Griffin and Leisen, 2017). Our idea is to replace $S_{j,m}$, the unnormalized weight referring to group j and atom m, with a product between a shared component across groups, i.e., depending only on m, and an idiosyncratic component, which depends on both group j and the specific atom m. This modification would break the conditional independence, given M, of the unnormalized weights. However, it could be advantageous in situations where additional information sharing is desired, as for the overlapping community in modular graphs, see (Todeschini et al., 2020).

Throughout the paper, we assumed that the atoms $\tau_1, \dots, \tau_M$ are i.i.d. according to a diffuse base measure P_0. However, recent research lines have explored the use of repulsive point processes as priors for mixture parameters (Petralia et al., 2012; Beraha et al., 2021) and applied them to model-based clustering for high-dimensional data (Ghilotti et al., 2024). Furthermore, the new general theory presented by Beraha et al. (2023) unifies various types of dependence on location, including independence, repulsiveness, and attractiveness. Palm calculus, which was a fundamental tool in our analysis, provides a mathematical framework for analysing these models. As a consequence, it is a promising avenue to extend our model to incorporate more sophisticated forms of dependence across the atoms $\tau_1, \dots, \tau_M$.

The use of vectors of NIFPP is not limited to the mixture framework. Indeed they can be helpful to face extrapolation problems when multiple populations of species are available. To have a glimpse of this, consider two populations of animals composed of different species with unknown proportions. Given samples from the first and the second populations, extrapolation problems refer to out-of-sample prediction. For instance, a typical question is: how many new and distinct species are shared across two additional samples from the populations? The seminal work of Lijoi et al. (2007b) faced extrapolation problems in the simplified framework of a single population, but no results are available in the presence of multiple groups. We think that the use of vectors of NIFPP can help to face the multiple-sample setting, still unexplored in the Bayesian framework.

Acknowledgments

We would like to thank the Editor, the Associate Editor, and the Referees for their insightful and constructive comments. We would like to thank Mario Beraha (Department of Mathematics, Politecnico di Milano) and Silvia Montagna (Department of Economics, University of Modena and Regio Emilia) for the helpful discussions.

Funding

The first and the third authors gratefully acknowledge support from the Italian Ministry of Education, University and Research (MUR), “Dipartimenti di Eccellenza” grant 2023-2027. The research of the second and the third authors was partially supported by MUR-PRIN grant 2022 CLTYP4, funded by the European Union – Next Generation EU. The research of the fourth author was partially supported by MUR-PRIN grant 2022 SMNNKY, funded by the European Union – Next Generation EU.

Supplementary Material

Supplementary Material for: Hierarchical Mixture of Finite Mixtures (DOI: 10.1214/24-BA1501SUPP; .pdf). The Supplementary materials includes: details on the general Vec-

NIFPP model; details on Palm's calculus, the mathematical tool we used for technical proofs; the proofs of the theoretical results, including additional details on the mixed moments and the predictive distribution; a detailed description of the MCMC algorithms for the HMFM; additional results on the simulation study and on shot put data analysis.

References

Argiento, R., Cremaschi, A., and Vannucci, M. (2020). "Hierarchical normalized completely random measures to cluster grouped data." *Journal of the American Statistical Association*, 115(529): 318–333. MR4078466. doi: https://doi.org/10.1080/01621459.2019.1594833. 2, 5

Argiento, R. and De Iorio, M. (2022). "Is infinity that far? A Bayesian nonparametric perspective of finite mixture models." *The Annals of Statistics*, 50(5): 2641–2663. MR4505373. doi: https://doi.org/10.1214/22-aos2201. 2, 5, 7, 10, 11, 13

Bassetti, F., Casarin, R., and Rossini, L. (2020). "Hierarchical species sampling models." *Bayesian Analysis*, 15(3): 809–838. MR4132651. doi: https://doi.org/10.1214/19-BA1168. 2, 13

Beraha, M., Argiento, R., Camerlenghi, F., and Guglielmi, A. (2023). "Normalized random meaures with interacting atoms for Bayesian nonparametric mixtures." arXiv:2302.09034. 25

Beraha, M., Argiento, R., Møller, J., and Guglielmi, A. (2021). "MCMC computations for Bayesian mixture models using repulsive point processes." *Journal of Computational and Graphical Statistics*, 31: 1–37. MR4425075. doi: https://doi.org/10.1080/10618600.2021.2000424. 25

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2019). "Distribution theory for hierarchical processes." *The Annals of Statistics*, 47(1): 67–92. MR3909927. doi: https://doi.org/10.1214/17-AOS1678. 2, 5, 7, 10

Charalambides, C. A. (2002). *Enumerative combinatorics*. CRC Press. MR1937238. 9

Colombi, A. Argiento, R. Camerlenghi, F., and Paci, L. (2024). "Supplementary Materials for: Hierarchical Mixture of Finite Mixtures"." *Bayesian Analysis*. doi: https://doi.org/10.1214/24-BA1501SUPP. 4

Dahl, D. B. (2006). "Model-Based clustering for expression data via a Dirichlet process mixture model." In Do, K.-A., Müller, P., and Vannucci, M. (eds.), *Bayesian Inference for Gene Expression and Proteomics*, Cambridge University Press, 201–218. MR2706330. 13

Dahl, D. B., Johnson, D. J., and Müller, P. (2022). "Search algorithms and loss functions for Bayesian clustering." *Journal of Computational and Graphical Statistics*, 31: 1189–1201. MR4513380. doi: https://doi.org/10.1080/10618600.2022.2069779. 20

D'Angelo, L., Canale, A., Yu, Z., and Guindani, M. (2023). "Bayesian nonparametric analysis for the detection of spikes in noisy calcium imaging data." *Biometrics*, 79: 1370–1382. MR4606359. doi: https://doi.org/10.1111/biom.13626. 24

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2): 212–229. 5

Denti, F., Camerlenghi, F., Guindani, M., and Mira, A. (2023). "A common atoms model for the Bayesian nonparametric analysis of nested data." *Journal of the American Statistical Assocciation*, 118(541): 405–416. MR4571130. doi: https://doi.org/10.1080/01621459.2021.1933499. 24

Diaconis, P. and Ylvisaker, D. (1979). "Conjugate priors for exponential families." *The Annals of Statistics*, 7(2): 269–281. MR0520238. 12

Dolmeta, P., Argiento, R., and Montagna, S. (2023). "Bayesian GARCH modeling of functional sports data." *Statistical Methods & Applications*, 32: 401–423. MR4606279. doi: https://doi.org/10.1007/s10260-022-00656-z. 19, 20

Favaro, S. and Teh, Y. W. (2013). "MCMC for normalized random measure mixture models." *Statistical Science*, 28(3): 335–359. MR3135536. doi: https://doi.org/10.1214/13-STS422. 11

Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2019). "From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering." *Advances in data analysis and classification*, 13: 33–64. MR3935190. doi: https://doi.org/10.1007/s11634-018-0329-y. 12

Frühwirth-Schnatter, S., Malsiner-Walli, G., and Grün, B. (2021). "Generalized mixtures of finite mixtures and telescoping sampling." *Bayesian Analysis*, 16(4): 1279–1307. MR4381135. doi: https://doi.org/10.1214/21-BA1294. 2, 5

Ghilotti, L., Beraha, M., and Guglielmi, A. (2024). "Bayesian clustering of high-dimensional data via latent repulsive mixtures." *Biometrika*, asae059. doi: https://doi.org/10.1093/biomet/asae059. 25

Gnedin, A. and Pitman, J. (2006). "Exchangeable Gibbs partitions and Stirling triangles." *Journal of Mathematical sciences*, 138: 5674–5685. MR2160320. doi: https://doi.org/10.1007/s10958-006-0335-z. 5

Griffin, J. E. and Leisen, F. (2017). "Compound random measures and their use in Bayesian non-parametrics." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(2): 525–545. MR3611758. doi: https://doi.org/10.1111/rssb.12176. 24

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer. MR2648134. doi: https://doi.org/10.1007/978-0-387-92407-6. 20

Hubert, L. and Arabie, P. (1985). "Comparing partitions." *Journal of Classification*, 2: 193–218. 13

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). "Adaptive mixtures of local experts." *Neural Computation*, 3(1): 79–87. 2

James, L. F., Lijoi, A., and Prünster, I. (2009). "Posterior analysis for normalized random measures with independent increments." *Scandinavian Journal of Statistics*, 36(1): 76–97. MR2508332. doi: https://doi.org/10.1111/j.1467-9469.2008.00609.x. 10, 11

Kallenberg, O. (2005). *Probabilistic symmetries and invariance principles.* Probability and its Applications. Springer, New York. MR2161313. 6

Lijoi, A., Mena, R. H., and Prünster, I. (2007a). "Controlling the reinforcement in Bayesian non-parametric mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4): 715–740. MR2370077. doi: https://doi.org/10.1111/j.1467-9868.2007.00609.x. 20

Lijoi, A., Mena, R. H., and Prünster, I. (2007b). "Bayesian nonparametric estimation of the probability of discovering new species." *Biometrika*, 94(4): 769–786. MR2416792. doi: https://doi.org/10.1093/biomet/asm061. 25

Lijoi, A., Nipoti, B., and Prünster, I. (2014). "Bayesian inference with dependent normalized completely random measures." *Bernoulli*, 20(3): 1260–1291. MR3217444. doi: https://doi.org/10.3150/13-BEJ521. 5

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2017). "Identifying mixtures of mixtures using Bayesian estimation." *Journal of Computational and Graphical Statistics*, 26(2): 285–295. MR3640186. doi: https://doi.org/10.1080/10618600.2016.1200472. 12

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). "Model-based clustering based on sparse finite Gaussian mixtures." *Statistics and computing*, 26(1-2): 303–324. MR3439375. doi: https://doi.org/10.1007/s11222-014-9500-2. 12

Miller, J. W. (2014). "Nonparametric and variable-dimension Bayesian mixture models: Analysis, comparison, and new methods." Ph.D. thesis, Brown University. 2

Miller, J. W. and Harrison, M. T. (2018). "Mixture models with a prior on the number of components." *Journal of the American Statistical Association*, 113(521): 340–356. MR3803469. doi: https://doi.org/10.1080/01621459.2016.1255636. 2, 5, 24

Nobile, A. (2004). "On the posterior distribution of the number of components in a finite mixture." *The Annals of Statistics*, 32(5): 2044–2073. MR2102502. doi: https://doi.org/10.1214/009053604000000788. 7

Page, G., Barney, B., and McGuire, A. (2013). "Effect of position, usage rate, and per game minutes played on NBA player production curves." *Journal of Quantitative Analysis in Sports*, 9: 337–345. 3

Page, G. L., Quintana, F. A., and Dahl, D. B. (2022). "Dependent modeling of temporal sequences of random partitions." *Journal of Computational and Graphical Statistics*, 31(2): 614–627. MR4425090. doi: https://doi.org/10.1080/10618600.2021.1987255. 24

Petralia, F., Rao, V., and Dunson, D. (2012). "Repulsive Mixtures." In Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc. 25

Regazzini, E., Lijoi, A., and Prünster, I. (2003). "Distributional results for means of normalized random measures with independent increments." *The Annals of Statistics*, 31(2): 560–585. MR1983542. doi: https://doi.org/10.1214/aos/1051027881. 5

Richardson, S. and Green, P. J. (1997). "On Bayesian analysis of mixtures with an unknown number of components (with discussion)." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4): 731–792. MR1483213. doi: https://doi.org/10.1111/1467-9868.00095. 20

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The nested Dirichlet process." *Journal of the American Statistical Association*, 103(483): 1131–1144. MR2528831. doi: https://doi.org/10.1198/016214508000000553. 24

Rousseau, J. and Mengersen, K. (2011). "Asymptotic behaviour of the posterior distribution in overfitted mixture models." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 73(5): 689–710. MR2867454. doi: https://doi.org/10.1111/j.1467-9868.2011.00781.x. 12

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). "Hierarchical Dirichlet processes." *Journal of the American Statistical Association*, 101(476): 1566–1581. MR2279480. doi: https://doi.org/10.1198/016214506000000302. 2, 11, 13

Todeschini, A., Miscouridou, X., and Caron, F. (2020). "Exchangeable random measures for sparse and modular graphs with overlapping communities." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(2): 487–520. MR4084173. doi: https://doi.org/10.1111/rssb.12363. 25

Van Havre, Z., White, N., Rousseau, J., and Mengersen, K. (2015). "Overfitting Bayesian mixture models with an unknown number of components." *PloS one*, 10(7): e0131739. 12

Wade, S. and Ghahramani, Z. (2018). "Bayesian cluster analysis: Point estimation and credible balls (with discussion)." *Bayesian Analysis*, 13(2): 559–626. MR3807860. doi: https://doi.org/10.1214/17-BA1073. 20