Hierarchical Normalized Completely Random Measures to Cluster Grouped Data

Raffaele Argiento, Andrea Cremaschi and Marina Vannucci

No. 557
November 2018

© 2018 by Raffaele Argiento, Andrea Cremaschi and Marina Vannucci. Any opinions expressed here are those of the authors and not those of the Collegio Carlo Alberto.
Hierarchical Normalized Completely Random Measures to Cluster Grouped Data

Raffaele Argiento
ESOMAS Department, University of Torino and Collegio Carlo Alberto, Torino, Italy
and
Andrea Cremaschi
Department of Cancer Immunology, Institute of Cancer Research, Oslo University Hospital, Oslo, Norway
Oslo Centre for Biostatistics and Epidemiology, University of Oslo, Oslo, Norway
and
Marina Vannucci
Department of Statistics, Rice University, Houston, TX, USA

Abstract

In this paper we propose a Bayesian nonparametric model for clustering grouped data. We adopt a hierarchical approach: at the highest level, each group of data is modelled according to a mixture, where the mixing distributions are conditionally independent normalized completely random measures (NormCRMs) centered on the same base measure, which is itself a NormCRM. The discreteness of the shared base measure implies that the processes at the data level share the same atoms. This desired feature allows to cluster together observations of different groups. We obtain a representation of the hierarchical clustering model by marginalizing with respect to the infinite dimensional NormCRMs. We investigate the properties of the clustering structure induced by the proposed model and provide theoretical results concerning the distribution of the number of clusters, within and between groups. Furthermore, we offer an interpretation in terms of generalized Chinese restaurant franchise process, which allows for posterior inference under both conjugate and non-conjugate models. We develop algorithms for fully Bayesian inference and assess performances by means of a simulation study and a real-data illustration. Supplementary Materials for this work is available online.

Keywords: Bayesian Nonparametrics; Clustering; Mixture Models; Hierarchical Models.
1 Introduction

In statistical modeling, dependency among observations can be captured in a number of different ways, for example through the inclusion of additional components (covariates) that link data in different groups. A specific type of dependency among observations is the membership to a specific group or category, where data share similar characteristics. This relates to the concept of partial exchangeability, where classical exchangeability does not hold for the whole dataset, but it does within each group. Let \( \theta = (\theta_1, \ldots, \theta_d) \) indicate a multidimensional vector of random variables divided into \( d \) groups, each of size \( n_j \), for \( j = 1, \ldots, d \). Partial exchangeability coincides with assigning a probability distribution \( P_j \) to each group, such that \( (\theta_{j1}, \ldots, \theta_{jn_j}) \mid P_j \stackrel{\text{iid}}{\sim} P_j \), for each \( j = 1, \ldots, d \), under a suitable prior (de Finetti measure) for the vector of random probabilities \( (P_1, \ldots, P_d) \), see Kallenberg (2005) for an excellent overview on the topic. From an inferential point of view, the specification of the joint distribution of \( (P_1, \ldots, P_n) \) is crucial as it defines the dependence structure among the random probability measures and, consequently, the sharing of information. In the Bayesian framework, it is common to impose the mild condition of exchangeability, i.e., \( (P_1, \ldots, P_d) \mid P \stackrel{\text{iid}}{\sim} P \), for a suitable probability distribution \( P \). In Bayesian nonparametrics, such hierarchical structure has been used to introduce the celebrated hierarchical Dirichlet process (Teh et al. 2005, 2006), with successful applications in genetics, image segmentation and topic modeling, to mention a few (Blei 2012; Teh and Jordan 2010). More recently, hierarchical processes have been investigated from an analytical perspective by Camerlenghi et al. (2017, 2018), while Bassetti et al. (2018) have focused on hierarchical species sampling models. These authors have shown that extensions to normalized completely random measures encompassing the Dirichlet process allow for richer predictive structures.

Undoubtedly, some of the most popular models in the Bayesian nonparametric framework are mixture models, (see, for example, Ferguson 1983; Lo 1984). In this setting, conditionally upon a set of latent variables, the observations are assumed independent from a family of parametric densities, while the latent parameters are distributed according to an almost surely discrete random probability measure (for further details, see Ishwaran and James 2001; Lijoi et al. 2007). These models owe their popularity to ease of interpretation, computational availability, and elegant mathematical properties. Any mixture model with an almost surely discrete mixing measure leads to ties in \( \theta \) with positive probability. This induces a random partition of the subject labels via the values of the parameters \( \theta \), meaning that two subjects share the same cluster if and only if the corresponding latent variables take on the same value. We refer to this as the natural clustering. Pitman (1996, 2003) showed that assigning the law of the discrete mixing measure is equivalent to assigning the law of the parameter that identifies the natural clustering. The prior on this partition is then obtained by marginalizing with respect to the infinite-dimensional parameter, and it is expressed via the so-called exchangeable partition probability function.

In this paper, we aim at obtaining a similar result in the context of hierarchical normalized completely random measures. We define a hierarchical normalized completely random measure mixture model by assuming that, conditionally upon \( \theta = (\theta_1, \ldots, \theta_d) \), the data are independent from some parametric family of distributions, and the prior on \( \theta \) is the hierarchical process discussed above. Marginalizing with respect to \( (P_1, \ldots, P_d) \) and \( P \), we write
our hierarchical model in terms of the cluster parameters and their prior distributions (i.e.,
\(d + 1\) distinct exchangeable partition probability functions). As a result, we obtain a two-
layered hierarchical clustering structure: a clustering within each of the groups (that we will
refer to as \(l\)-clustering), and a natural clustering across the whole multidimensional array \(\theta\).
We study such clustering structure by considering a nonparametric mixture model in which
the completely random measure has a discrete centering measure, and provide theoretical
results concerning the distribution of the number of clusters, within and between groups.
Furthermore, we offer an interpretation in terms of generalized Chinese restaurant franchise
process, enabling posterior inference for both conjugate and non-conjugate models.

With respect to the recent contributions on Bayesian nonparametric hierarchical pro-
cesses of Camerlenghi et al. (2018) and Bassetti et al. (2018), which investigate more the-
etorical aspects of these models, in this paper we provide a detailed study of the mixture
model in terms of the two-layered hierarchical clustering structure it induces, as particularly
useful in the context of Bayesian nonparametric model-based clustering. Original contribu-
tions of the paper include: a characterization of the mixture model in terms of the clustering
structure; interpretation of the clustering model through the metaphor of the generalized
Chinese restaurant franchise; an MCMC algorithm to compute the posterior of the cluster
structure, which makes use of data augmentation techniques; expressions of moments of the
Bayesian nonparametric ingredients; applications to simulated and benchmark data sets, to
illustrate the effect of critical hyperparameters on the clustering structure. Additionally, we
use original techniques in the proofs of some of the theoretical results (see Proposition 2 in
Section 2.2). Finally, we also notice that the two-layered hierarchical clustering induced by
our model can be interpreted as a “cluster of clusters”, or “mixture of mixtures”, as intro-
duced in Argiento et al. (2014) and Malsiner-Walli et al. (2017). These authors, however,
address a different problem from the one we consider here.

The rest of the paper is organized as follows: in Section 2 we first introduce completely
random measures and normalized completely random measures for which the centering dis-
tribution is discrete. We also discuss the properties that characterize the clustering induced
by a normalized completely random measure, such as the expression of mixed moments and
the law of the number of clusters. Next, we extend the proposed characterization to hierar-
chical normalized completely random measures for grouped data and employ them as mixing
distributions in mixture models for clustering. In Section 3, we provide insights on the pos-
terior sampling process for the proposed mixture models, including prediction. Section 4
offers some simulation results, as well as an application to a benchmark dataset. Section
5 concludes the paper. Proofs of the theoretical results, algorithmic details and additional
results are reported in the Supplementary Materials.

2 Methodology

2.1 Normalized completely random measures with discrete centering

Let \(\Theta\) be a complete and separable metric space, endowed with the corresponding Borel
\(\sigma\)-algebra \(\mathcal{B}\). A completely random measure (CRM) on \(\Theta\) is a random measure \(\mu_1\) taking
values on the space of boundedly finite measures on \((\Theta, \mathcal{B})\) and such that, for any collection of disjoint sets \(\{B_1, \ldots, B_n\} \in \mathcal{B}\), the random variables \(\mu_1(B_1), \ldots, \mu_1(B_n)\) are independent (see Kingman 1993). In this paper, we focus on the subclass of CRMs that can be written as \(\mu_1(\cdot) = \sum_{l \geq 1} J_l \delta_{\tau_l}(\cdot)\), describing an almost surely discrete random probability measure with random masses \(\mathcal{J} = \{J_l\} = \{J_l, l \geq 1\}\) independent from the random locations \(\mathcal{T} = \{\tau_l\}\).

The law of this subclass of CRMs, called homogeneous CRMs, is characterized by a Lévy intensity measure \(\nu\) that factorizes into \(\nu(ds, d\tau) = \alpha(s)P(d\tau)ds\), where \(\alpha\) is the density of a nonnegative measure, absolutely continuous with respect to the Lebesgue measure on \(\mathbb{R}^+\), and \(P\) is a probability measure over \((\Theta, \mathcal{B})\). Hence, the random locations are independent and identically distributed according to the base distribution \(P\), while the random masses \(\mathcal{J}\) are distributed according to a Poisson random measure with intensity \(\alpha\). In what follows, we will refer to \(\alpha\) only as the Lévy intensity.

A homogeneous normalized completely random measure (NormCRM) \(P_1\) on \(\Theta\) is a random probability measure having the following representation:

\[
P_1(\cdot) = \frac{\mu_1(\cdot)}{\mu_1(\Theta)} = \sum_{l \geq 1} \frac{J_l}{T_1} \delta_{\tau_l}(\cdot) = \sum_{l \geq 1} w_l \delta_{\tau_l}(\cdot),
\]

where \(T_1 = \mu_1(\Theta) = \sum_{l \geq 1} J_l\), and hence \(\sum_{l \geq 1} w_l = 1\). We point out that the law of the infinite sequence \(\{w_l\}\) depends only on the Lévy intensity \(\alpha\). We indicate with \(P_1 \sim \text{NormCRM}(\alpha, P)\) a NormCRM with Lévy intensity \(\alpha\) and centering measure \(P\). The acronym NRMI is also used in the literature, in reference to the original definition of NormCRMs on the real line as normalized random measures with independent increments (Regazzini et al. 2003). To ensure that the normalization in (1) is well-defined, the random variable \(T_1\) has to be positive and almost surely finite. This is guaranteed by imposing the regularity conditions

\[
\int_{\mathbb{R}^+} \alpha(s)ds = +\infty, \quad \text{and} \quad \int_{\mathbb{R}^+} (1 - e^{-s})\alpha(s)ds < +\infty.
\]

The class of NormCRMs encompasses the well-known Dirichlet process \(\text{Dir}(\kappa, P)\), obtained by normalization of a gamma process, with Lévy intensity \(\alpha(s) = \kappa s^{-1}e^{-s}\mathbb{I}_{(0, +\infty)}(s)\), for \(\kappa > 0\). It also includes the normalized generalized gamma process \(\text{NGG}(\kappa, \sigma, P)\) of Lijoi et al. (2007), obtained when \(\alpha(s) = \frac{\kappa}{\Gamma(1 - \sigma)}s^{-1-\sigma}e^{-s}\mathbb{I}_{(0, +\infty)}(s)\), for \(0 \leq \sigma < 1\), and the normalized Bessel process \(\text{NormBessel}(\kappa, \omega, P)\) of Argiento et al. (2016), when \(\alpha(s) = \frac{s}{\omega}e^{-\omega s}I_0(s)\mathbb{I}_{(0, +\infty)}(s)\), for \(\omega \geq 1\) and \(I_0(s)\) the modified Bessel function of the first kind. In the expressions of the Lévy intensities above \(\mathbb{I}_A(s)\) is the indicator function of the set \(A\), i.e., \(\mathbb{I}_A(s) = 1\) if \(s \in A\) and \(\mathbb{I}_A = 0\) otherwise.

A sample from \(P_1\) is an exchangeable sequence such that \((\tilde{\theta}_1, \ldots, \tilde{\theta}_n)|\mathcal{P}_1 \overset{iid}{\sim} P_1\). In this paper we adopt a slightly different representation of a sample from \(P_1\), which will be useful to characterize the clustering induced by \(P_1\) when the centering measure \(P\) is discrete. Let \(P_1\) be defined as in (1) and let \(P_1^*\) be a random probability measure on the positive integers
\[ P^*_1(\cdot) = \sum_{l \geq 1} w_l \delta_l(\cdot). \]

**Lemma 1.** Let \((\tilde{\theta}_1, \ldots, \tilde{\theta}_n)\) be a sample from \(P_1\) defined as in (1) and let \((l_1, \ldots, l_n)\) be a sample from \(P^*_1\) defined as in (3). Define \(\theta_1 = \tau_1, \ldots, \theta_n = \tau_n\) with \(\{\tau_i\} \overset{iid}{\sim} P\) and \(P\) the centering measure of \(P_1\). Then

\[ (\theta_1, \ldots, \theta_n) \overset{\mathcal{L}}{=} (\tilde{\theta}_1, \ldots, \tilde{\theta}_n). \]

**Proof:** See Supplementary Materials. We therefore refer to a sample from \(P_1\) as a sequence \((\theta_1, \ldots, \theta_n)\) obtained under (3) following the construction in Lemma 1. Since \(P_1\) is discrete, a sample \((\tilde{\theta}_1, \ldots, \tilde{\theta}_n)\) from (1) induces a random partition \(\tilde{\rho} = \{\tilde{C}_1, \ldots, \tilde{C}_{\hat{K}_n}\}\) on the set of indices \(\{1, \ldots, n\}\). We refer to this as the **natural clustering**, with \(\tilde{C}_j = \{i : \tilde{\theta}_i = \tilde{\theta}_j^*\}\), for \(j = 1, \ldots, \hat{K}_n\), and \((\tilde{\theta}_1^*, \ldots, \tilde{\theta}_n^*)\) the set of unique values derived from the sequence \((\tilde{\theta}_1, \ldots, \tilde{\theta}_n)\). When the centering distribution \(P\) is diffuse, it is well known (see Pitman 1996; Ishwaran and James 2003) that the joint marginal distribution of a sample \((\tilde{\theta}_1, \ldots, \tilde{\theta}_n)\) can be uniquely characterized by the law of the natural clustering \((\tilde{\rho}, \tilde{\theta}_1^*, \ldots, \tilde{\theta}_n^*)\) as

\[ \mathcal{L}(\tilde{\rho}, d\tilde{\theta}_1^*, \ldots, d\tilde{\theta}_n^*) = \mathcal{L}(\tilde{\rho}) \mathcal{L}(d\tilde{\theta}_1^*; \tilde{K}_n) = \pi(\tilde{\rho}) \prod_{l=1}^{\hat{K}_n} P(d\tilde{\theta}_l^*), \]

with \(\pi(\tilde{\rho})\) the probability law on the set of the partitions of \(\{1, \ldots, n\}\), which is called the exchangeable partition probability function, or eppf. Since the eppf depends only on the Lévy intensity \(\alpha\) of the NormCRM, we write \(\pi(\tilde{\rho}) = \text{eppf}(\tilde{e}_1, \ldots, \tilde{e}_{\hat{K}_n}; \alpha)\), where eppf is a unique symmetric function depending only on \(\tilde{e}_j = \text{Card}({\tilde{C}_j})\), the cardinalities of the sets \(\tilde{C}_j\), for \(j = 1, \ldots, \hat{K}_n\). A formula for the eppf of a generic NormCRM can be obtained as (see formulas (36)-(37) in Pitman (2003))

\[ \text{eppf}(\tilde{e}_1, \ldots, \tilde{e}_{\hat{K}_n}; \alpha) = \int_0^{+\infty} \frac{u^{n-1}}{\Gamma(n)} e^{-\phi(u)} \prod_{l=1}^{\hat{K}_n} c_{\tilde{e}_l}(u) du, \]

where \(\phi(u)\) and the functions \(c_m(u)\), for \(m = 1, 2, \ldots, \) are defined as

\[ \phi(u) = \int_0^{+\infty} (1 - e^{-us})\alpha(s)ds, \quad c_m(u) = (-1)^{m-1} \phi_m(u) = \int_0^{+\infty} s^me^{-us}\alpha(s)ds. \]

Here, \(\phi(u)\) is the **Laplace exponent** of the unnormalized CRM \(\mu_1(\cdot)\), and \(\phi_m(u) = \frac{d^m}{du^m} \phi(u)\). Decomposition (4) sheds light on the law of the clustering structure induced by a NormCRM when the centering measure is diffuse. It can be decomposed into two factors: the law of the partition \(\tilde{\rho}\), that depends only on the intensity parameter \(\alpha\), and the law of the cluster-specific parameters \((\tilde{\theta}_1^*, \ldots, \tilde{\theta}_{\hat{K}_n}^*)\), that conditionally upon the number of unique values \(\hat{K}_n\)
is the $\bar{K}_h$-product of the centering measure $P$.

We want to show that equation (4) can still be valid in the case of a discrete centering measure $P$, even though with a slightly different interpretation. With this aim, consider a sample $(l_1, \ldots, l_n)$ from $P^*_1$ as in (3), and the vector $(\theta_1, \ldots, \theta_n)$ as in Lemma 1. Also, let $l^* = (l^*_1, \ldots, l^*_K)$ be the vector of unique values among $(l_1, \ldots, l_n)$ and $\rho$ the induced clustering, i.e., $\rho = \{C_1, \ldots, C_K\}$, where $i \in C_h$ iff $l_i = l^*_h$, with $i = 1, \ldots, n$ and $h = 1, \ldots, K$.

The law of $\rho$ can be characterized in terms of generalized Chinese restaurant process (see Pitman 2006) as the eppf induced by the Lévy intensity $\alpha$. To prove this we first observe that, if $(l_1, \ldots, l_n) \mid P^*_1 \overset{iid}{\sim} P^*_1$, then $L(l_1, \ldots, l_n) = E(w_{l_1} \ldots w_n)$. Then, using the equivalent representation of $(l_1, \ldots, l_n)$ in terms of $(l^*_1, \ldots, l^*_K)$ and $\rho = \{C_1, \ldots, C_K\}$, with $e_j = \text{Card}(C_j)$, for $j = 1, \ldots, K$, a change of variable leads to $L(l^*_1, \ldots, l^*_K, \rho) = E(w_{l^*_1} \ldots w_{l^*_K})$. Hence, by formula (4) in Pitman (2003), the law of $\rho$ is $L(\rho) = \sum_{l^*_1, \ldots, l^*_K} E(w_{l^*_1} \ldots w_{l^*_K}) = \text{ep}(e_1, \ldots, e_K; \alpha)$, where $l^*_1, \ldots, l^*_K$ ranges over all permutations of $K$ positive integers.

We are now ready to show that, even if we do not assume $P$ to be diffuse, the law of the sample $(\theta_1, \ldots, \theta_n)$ has a unique representation as in (4), provided that $\rho$ is the partition induced by $(l_1, \ldots, l_n)$ and that $(\rho^*_1, \ldots, \rho^*_K)$ is an i.i.d. sample from $P$. To this end we first give the following:

**Definition 1.** An $l$-clustering representation of $(\theta_1, \ldots, \theta_n)$ is a vector $(\rho^*_1, \ldots, \rho^*_K)$ s.t.:

1. $\rho = \{C_1, \ldots, C_K\}$ is the clustering induced by the $l^*$ on the data indices (i.e., $i \in C_h$ iff $l_i = l^*_h$ for $i = 1, \ldots, n$ and $h = 1, \ldots, K$),
2. $\rho^*_h$ is the value shared by all the $\theta$’s in group $C_h$, for $h = 1, \ldots, K$.

We point out that, in an $l$-clustering representation, $(\rho^*_1, \ldots, \rho^*_K)$ is not the vector of unique values among the $\theta$’s, and so $K$ is not the random variable representing the number of different values among the $\theta$’s, as it is usually denoted in the Bayesian nonparametric literature. Due to the discreteness of the centering measure $P$, we could have coincidence also among the $\rho^*$’s. Moreover, from $(\rho, \rho^*_1, \ldots, \rho^*_K)$, we can recover $(\theta_1, \ldots, \theta_n)$. While from $(\theta_1, \ldots, \theta_n)$ we cannot recover the $l$-clustering unless the knowledge of $(l_1, \ldots, l_n)$ is provided.

As a simple example to better understand this point, consider a sample of dimension $n = 8$

![Figure 1: Illustration of an l-clustering based on a sample of dimension n = 8 from a NormCRM whose centering measure is a discrete distribution on the colored lines.](image)

...from a NormCRM whose centering measure is a discrete distribution on the colored lines. In this sample we have $\theta = \{\text{continuous green, dashed orange, continuous green, dotted blue, continuous green, dashed orange, dashed orange, dashed orange}\}$, obtained under the hypothesis of Lemma 1, with $(l_1, \ldots, l_n) = (1, 2, 1, 3, 1, 4, 2, 4)$, $\tau_1 = \text{continuous green, } \tau_2 = \text{dashed}$. 


orange, \( \tau_3 = \text{dotted blue}, \tau_4 = \text{dashed orange} \). The corresponding \( l \)-clustering is represented in Figure 1, and it is formed by \( \rho = \{C_1 = \{1,3,5\}, C_2 = \{2,7\}, C_3 = \{4\}, C_4 = \{6,8\}\} \) and \((\theta_1, \ldots, \theta_{K_n}^\ast) = (\text{continuous green, dashed orange, dotted blue, dashed orange})\). It is clear that we can recover the sample from the \( l \)-clustering partition by letting \( \theta_i = \theta_i^\ast, \ i = 1, \ldots, n \). The opposite, however, is not possible: without the knowledge of \((l_1, \ldots, l_n)\) the \( l \)-clustering cannot be recovered just by looking at \((\theta_1, \ldots, \theta_n)\). We also notice that the partition \( \rho \) of the \( l \)-clustering is not the one induced by the unique values among the \( \theta \)'s. In fact, \( \theta_2^\ast = \theta_4^\ast = \text{dashed orange} \) with both clusters \( C_2 \) and \( C_4 \) containing orange dashed lines, while the natural clustering is \( \tilde{\rho} = \{C_1 = \{1,3,5\}, C_2 = \{2,7,6,8\}, C_3 = \{4\}\} \). These considerations yield to the following:

**Proposition 1.** The marginal law of a sample \((\theta_1, \ldots, \theta_n)\) from a NormCRM \( P_1 \) defined as in (1), with a general centering measure \( P \), has a unique characterization in terms of \( \rho \) and \((\theta_1^\ast, \ldots, \theta_{K_n}^\ast)\) of an \( l \)-clustering representation. In particular,

\[
\mathcal{L}(\rho, d\theta_1^\ast, \ldots, d\theta_{K_n}^\ast) = \pi(\rho) \prod_{l=1}^{K_n} P(d\theta_l^\ast) = \text{eppf}(e_1, \ldots, e_{K_n}; \alpha) \prod_{l=1}^{K_n} P(d\theta_l^\ast).
\]

**Proof:** See Supplementary Materials. This result implies that \( \rho \) is chosen according to the eppf induced by \( \alpha \) and that \((\theta_1^\ast, \ldots, \theta_{K_n}^\ast)\) is an i.i.d. sample from \( P \).

### 2.2 Hierarchical NormCRMs

In many applications, multiple sources of information can be observed, hence generating sequences of data points that are related. In particular, data sampled under the same experimental condition can be considered exchangeable, introducing group-specific parameters. In our setting, this translates into including an additional hierarchical level in the model, yielding \((\theta_{1j}, \ldots, \theta_{nj_j})|P_j \text{ iid } P_j\), for group \( j = 1, \ldots, d \). Here, we assume that \((P_1, \ldots, P_d)\) are also exchangeable, that infinite experimental conditions are possible, and that \((P_1, \ldots, P_d)|P \text{ iid } \text{NormCRM}(\alpha, P)\), where \( P \sim \text{NormCRM}(\alpha_0, P_0) \), with \( P_0 \) a diffuse measure on \( \Theta \) and \( \alpha_0 \) and \( \alpha \) Lévy intensities satisfying the regularity conditions reported in (2). We define a hierarchical NormCRM as the joint law of \((P_1, \ldots, P_d)\). A sample from a hierarchical NormCRM is a multidimensional array \( \theta = (\theta_1, \ldots, \theta_d) \), with \( \theta_j = (\theta_{1j}, \ldots, \theta_{nj_j}) \), \( j = 1 \ldots, d \), with elements in \( \Theta \), such that

\[
(\theta_{1j}, \ldots, \theta_{nj_j})|P_j \text{ iid } P_j, \ j = 1, \ldots, d,
\]

\[
(P_1, \ldots, P_d)|P \sim \text{NormCRM}(\alpha; P),
\]

\[
P \sim \text{NormCRM}(\alpha_0; P_0).
\]

The class of hierarchical NormCRMs reduces to the celebrated hierarchical Dirichlet process of Teh et al. (2006) when \( \alpha_0 \) and \( \alpha \) are the Lévy intensities of a gamma process. Theoretical aspects of hierarchical NormCRMs have been thoroughly investigated by Camerlenghi et al. (2018) and further extended to the class of species sampling models by Bassetti et al. (2018). With respect to these contributions, in this paper we focus more on investigating aspects of
the model in terms of the two-layered hierarchical clustering structure it induces. We also derive a central limit theorem for the number of clusters induced by \( P_1, \ldots, P_d \) at the second level of hierarchy in formula (8) where, unlike other asymptotic results on NormCRMs, we let the number of observations \( n_j \) in each group be bounded and the number of groups \( d \) go to infinity. Furthermore, we provide expressions for the mixed moments of a hierarchical NormCRM. Finally, we use different techniques in the proofs of some of the theoretical results presented below (see Proposition 2 and Supplementary Materials).

Using Proposition 1, we integrate out the infinite-dimensional parameters \((P_1, \ldots, P_d)\) and \( P \). Firstly, given \( P \), we observe that for each \( j = 1, \ldots, d \), \( \theta_j = (\theta_{j1}, \ldots, \theta_{jn_j}) \) is a sample from \( P_j \), that is a NormCRM with discrete centering measure \( P \) as centering measure. Consequently, \( \theta_j \) can be obtained from its \( l \)-clustering representation \( \rho_j = (C_{j1}, \ldots, C_{jk_j}) \), \( \theta^*_j = (\theta^*_{j1}, \ldots, \theta^*_{jK_j}) \). We refer to \( C_{jh} \), for \( h = 1, \ldots, K_j \) as \( l \)-clusters hereafter. Hence, we can marginalize model (8) with respect to \( P_j \) as:

\[
\rho_j|\alpha \sim \text{eppf}(e_j; \alpha), \quad (\theta^*_{j1}, \ldots, \theta^*_{jK_j})|K_j, P \overset{iid}{\sim} P, \quad P \sim \text{NormCRM}(\alpha_0; P_0),
\]

where \( e_j = (e_{j1}, \ldots, e_{jK_j}) \) is the vector of \( l \)-cluster sizes in the \( j \)-th group, for \( j = 1, \ldots, d \). Let \((K_1, \ldots, K_d) = (k_1, \ldots, k_d)\) be the number of \( l \)-clusters in each group of data, and let \( T = \sum_{j=1}^{d} K_j \) be the total number of \( l \)-clusters across all groups. We define the index transformation \( t : \{(j, h) : j = 1, \ldots, d; h = 1, \ldots, k_j\} \to \{1, \ldots, T\} \) as follows:

\[
t(j, h) = \sum_{s=0}^{j-1} k_s + h, \quad h = 1, \ldots, k_j, \quad j = 1, \ldots, d,
\]

where \( k_0 = 0 \). Consider now the multidimensional array \( \theta = (\theta_1, \ldots, \theta_d) \), where each element \( \theta_j = (\theta_{j1}, \ldots, \theta_{jn_j}) \) is a sample of size \( n_j \) from \( P_j \), for \( j = 1, \ldots, d \). Furthermore, consider \( \theta^* = (\theta^*_1, \ldots, \theta^*_d) \), the multidimensional array where each element \( \theta^*_j \) is the vector of shared value of the \( l \)-clustering representation of \( \theta_j \) for each group of data, as described in definition (1). The action of the function \( t \) in (10) on the indices of \( \theta^* \) results in a transformation of the multidimensional array into a vector of length \( T \) where the rows of the array are sequentially aligned. Since the information carried by this vector remains unchanged, we will refer to it by using the same notation, \( \theta^* \). Conditionally on \( T \), the vector \( \theta^* \) is a sample of size \( T \) from \( P \), that is \((\theta^*_1, \ldots, \theta^*_T)|T, P \overset{iid}{\sim} P \). Since \( P \) is a NormCRM with diffuse centering measure \( P_0 \), the \( l \)-clustering representation of \( \theta^* \) coincides with the natural clustering induced by the NormCRM. In particular, \( \psi = (\psi_1, \ldots, \psi_M) \) will denote the vector of unique values in \( \theta^* \), and \( \eta = (D_1, \ldots, D_M) \) the clustering induced by \( \psi \) on the index set \( \{1, \ldots, T\} \), that is \( t \in D_m \) iff \( \theta^*_t = \psi_m \), for \( t = 1, \ldots, T \) and \( m = 1, \ldots, M \). We denote by \( d = (d_1, \ldots, d_M) \) the size of the clusters in \( \eta \). Interestingly, the vector \( \psi \) is also the vector of unique values among the multidimensional array \( \theta \).

**Proposition 2.** The marginal law of the multidimensional array \( \theta = (\theta_1, \ldots, \theta_d) \) from a
hierarchical NormCRM defined as in (8), with $P_0$ diffuse, can be characterized in terms of $(\rho, \eta, \psi)$, with $\rho = (\rho_1, \ldots, \rho_d)$, as:

$$
\mathcal{L}(\rho, \eta, d\psi) = \mathcal{L}(\eta|\rho) \prod_{j=1}^{d} \mathcal{L}(\rho_j) \prod_{m=1}^{M} P_0(d\psi_m) = \text{eppf}(d; \alpha_0) \prod_{j=1}^{d} \text{eppf}(e_j; \alpha) \prod_{m=1}^{M} P_0(d\psi_m). \tag{11}
$$

We note that the so called pEPPF, introduced by Camerlenghi et al. (2018), can be obtained by (11) marginalizing with respect to $\eta$ and $\psi$. Moreover, according to (11), the multidimensional array $\theta$ can be drawn as follows: first, choose the partitions $\rho_j$ of the indices $\{1, \ldots, n_j\}$, for $j = 1, \ldots, d$, independently and according to the Chinese restaurant process governed by $\alpha$. After identifying the number of elements $K_j = k_j$ of $\rho_j$ in each group, compute $T = \sum_{j=1}^{d} k_j$ and draw a partition $\eta$ of the indices $\{1, \ldots, T\}$ according to a Chinese restaurant process governed by $\alpha_0$. Sample $\psi = (\psi_1, \ldots, \psi_M)$, with $M = \#\eta$, iid from $P_0$, and build $\theta^* = (\theta_1^*, \ldots, \theta_T^*)$, where $\theta_t^* = \psi_m$ iff $t \in D_m$, for $t = 1, \ldots, T$ and $m = 1, \ldots, M$. Invert the transformation in (10) by finding $(j, h)$ such that $\theta_{jh}^* = \theta_{t=\psi_m}^*$. Finally, for each $i = 1, \ldots, n_j$ and $j = 1, \ldots, d$, set $\theta_{ji} = \theta_{jh}^*$ iff $i \in C_{jh}$, for $h = 1, \ldots, k_j$.

Let $\theta$ be a multidimensional array sampled according to a hierarchical NormCRM as in formula (8). We call natural clustering the partition of the data corresponding to indices $I = \{I_1, \ldots, I_M\}$ such that $(j, i) \in I_m$ iff $\theta_{ji} = \psi_m$, where $\psi = (\psi_1, \ldots, \psi_M)$ is the vector of unique values among $\theta$. On the other hand, if we have a sample $(\rho, \eta, \psi)$ from (11), a multidimensional array $\theta$ can be recovered whose natural clustering is obtained by letting $I_m = I_m^{(\rho, \eta)} = \{(j, i) : j \in \{1, \ldots, d\}, i \in C_{jh}, t(j, h) \in D_m\}$, and $\theta_{ji} = \psi_m$, for $(j, i) \in I_m$ and $m = 1, \ldots, M$. Hence, formula (11) characterizes the natural clustering induced by a sample $\theta$ form a hierarchical NormCRM. The partition identified by $(\rho, \eta)$ can be expressed in terms of the generalized Chinese restaurant franchise process as follows: $d$ restaurants with $n_j$ customers each, for $j = 1, \ldots, d$, share the same menu of dishes. The customers entering the $j$-th restaurant are allocated to the tables according to $\text{eppf}(e_j; \alpha)$, independently from the other restaurants in the franchise, generating the partition $\rho_j = (C_{j1}, \ldots, C_{jK_j})$, for $j = 1, \ldots, d$. Conditionally on $T = \sum_{j=1}^{d} K_j$, the tables of the franchise are grouped according to the law described by $\text{eppf}(d; \alpha_0)$, thus obtaining a partition of tables in $M$ clusters (i.e., a clustering of clusters). In addition, conditionally on $M$, all tables in a same cluster share the same dish, for a totality of $M$ different dishes served in the franchise, and indicated by the vector $\psi = (\psi_1, \ldots, \psi_M)$, sampled i.i.d. from $P_0$. In the Chinese restaurant franchise metaphor, the natural clustering is formed of clusters of customers that share the same dish across the franchise, and not only in the same restaurant.

### 2.2.1 Number of Clusters of a Hierarchical NormCRM

When studying a random partition induced by a hierarchical NormCRM, the quantities of interest are the random variables identifying the number of clusters. Here we have $K_j$, the number of $l$-clusters in the $j$-th group of data, and $M$, the number of natural clusters (or equivalently the number of elements in the partition $\eta$). Moreover, we are interested in
\[ T = \sum_{j=1}^{d} K_j. \] As already observed, the \( l \)-clustering partitions in each group are independent and sampled according to the eppf induced by the Lévy intensity \( \alpha \). It is well known that the distribution of \( K_j \) (see Pitman (2006)) can be recovered as

\[
P(K_j = k) = \frac{1}{k!} \sum_{e_1 + \cdots + e_k = n_j} \binom{n_j}{e_1, \ldots, e_k} \text{eppf}(e_1, \ldots, e_k; \alpha), \quad k = 1, \ldots, n_j,
\]

where the last sum is over all the compositions of \( n_j \) into \( k \) parts, i.e., all positive integers such that \( e_1 + \cdots + e_k = n_j \). In the same way, given \( K_1, \ldots, K_d \), the distribution of \( M \), the number of clusters of a partition sampled according to the eppf induced by \( \alpha_0 \) on a sample of size \( T \), can be written as

\[
P(M = m) = \sum_{t=d}^{n} \mathbb{P}(M = m | T = t) \mathbb{P}(T = t), \quad m = 1, \ldots, n, \quad n = \sum_{j=1}^{d} n_j, \tag{13}
\]

\[
P(M = m | T = t) = \frac{1}{m!} \sum_{d_1 + \cdots + d_m = t} \binom{t}{d_1, \ldots, d_m} \text{eppf}(d_1, \ldots, d_m; \alpha_0), \quad m = 1, \ldots, t.
\]

A derivation of this formulas can also be found in Camerlenghi et al. (2018). In the sensitivity analysis, contained in the Supplementary Materials, we present a description of the a priori behavior of \( M \). A numerical evaluation of expression (13) can be quite burdensome, since it involves the computation of the distribution of \( T = \sum_{j=1}^{d} K_j \), that is a convolution of \( d \) random variables with probability mass functions defined in (12). To simplify the computation, we observe that \( T \) is a sum of \( d \) independent random variables, so that the Central Limit Theorem can be used, to approximate the distribution of \( T \) when \( d \) is large. In particular, we adopt the Berry-Esseen Theorem (see, for instance, Durrett 1991) to quantify the error of the approximation. Let \( \mu_j = \mathbb{E}(K_j) \), \( \sigma_j = \text{Var}(K_j) \), and \( \gamma_j = \mathbb{E}|K_j - \mu_j|^3 \). Let

\[
\mu = \sum_{j=1}^{d} \mu_j, \quad \sigma^2 = \sum_{j=1}^{d} \sigma_j^2, \quad \gamma = \sum_{j=1}^{d} \gamma_j,
\]

and let \( \Phi(x; \mu, \sigma^2) \) be the cdf of a Normal distribution with mean \( \mu \) and variance \( \sigma^2 \). Then:

\[
\sup_{x \in \mathbb{R}} |F_{T/d}(x) - \Phi(x; \mu, \sigma^2)| \leq c \frac{\gamma}{\sigma^3 \sqrt{d}}, \tag{14}
\]

where \( F_{T/d} \) is the cdf of the random variable \( T/d \), and \( c \) is an absolute constant. The upper bound on the smallest possible value of \( c \) has decreased from Esseen’s original estimate of 7.59 to its current value of 0.4785 provided by Tyrin (2010). We observe that, since \( K_j < n_j \) for each \( j \), then \( \gamma_j \leq (n_j - 1)^3 \). Furthermore, under the assumption that the number of observations in each group is bounded, i.e. \( 2 \leq n_j \leq H < \infty \) for each \( j \) and for a positive constant \( H \), we have that \( \sigma_j^2 > 0 \). Under the latter two hypotheses: \( c \frac{\gamma}{\sigma^3 \sqrt{d}} \leq c \frac{dH}{d^{1/2} \sigma_{\min}} = \frac{c}{\sqrt{d}} \frac{H}{\sigma_{\min}} \). Therefore, from (14), we obtain both the Central Limit Theorem for \( d \to \infty \) with the usual rate of convergence, and an easy bound for the approximation error. We observe also that if the number of observations in each group is constant, i.e. \( n_j = n \), and both previous hypotheses hold true, then the bound is equal to \( c \gamma_1 / (\sqrt{d} \sigma_1^3) \).
2.2.2 Moments of a Hierarchical NormCRM

Other quantities of interest are the moments of a hierarchical NormCRM. Here, we firstly derive two formulas for a NormCRM with generic centering measure, clarifying how the moments of a NormCRM are related to the distribution of the number of clusters studied in the previous section. We also generalize these formulas to the hierarchical NormCRM case. Let $P_1$ be a NormCRM($\alpha, P$), and let $A$ be a measurable subset of $\Theta$. We can characterize the moments of the random variable $P_1(A)$ in terms of the distribution of the number of clusters $K_n$, for each $n > 1$, as

$$E(P_1(A)^n) = E(P(A)^{K_n}).$$

(15)

This formula can be easily extended to compute the mixed moments. In particular, if $A$ and $B$ are two disjoint measurable subsets of $\Theta$, and $n_1, n_2$ are positive integers, then

$$E(P_1(A)^{n_1}P_1(B)^{n_2}) = \sum_{k_1=1}^{n_1} \sum_{k_2=1}^{n_2} P(A)^{k_1}P(B)^{k_2}g(k_1, k_2),$$

(16)

where $(k_1, k_2)$ is a composition of $K_n = k$, and $n = n_1 + n_2$. Moreover, $g$ is defined as

$$g(k_1, k_2) = \frac{1}{k_1!k_2!} \sum_{\epsilon_{1,1}+\ldots+\epsilon_{1,k_1}=n_1} \sum_{\epsilon_{2,1}+\ldots+\epsilon_{2,k_2}=n_2} \binom{n_1}{\epsilon_{1,1}, \ldots, \epsilon_{1,k_1}} \times \binom{n_2}{\epsilon_{2,1}, \ldots, \epsilon_{2,k_2}} \text{eppf}(\epsilon_{1,1}, \ldots, \epsilon_{1,k_1}, \epsilon_{2,1}, \ldots, \epsilon_{2,k_2}).$$

(17)

We refer to the Supplementary Material for the proofs of (15) and (16). Using (17), we can recover some well-known results for the moments of a NormCRM.

Consider now $P_1, P_2|P \overset{\text{iid}}{\sim} \text{NormCRM}(\alpha, P)$, and $P \sim \text{NormCRM}(\alpha_0, P_0)$, and let $A, B \subset \Theta$ measurable, we have the following:

$$E(P_1(A)^{n_1}) = E(P_0(A)^M)$$

$$\text{Cov}(P_1(A), P_2(B)) = \eta_0 (P_0(A \cap B) - P_0(A)P_0(B))$$

$$\text{Cov}(P_1(A), P_1(B)) = (\eta_0 + \eta_1 - \eta_0\eta_1) (P_0(A \cap B) - P_0(A)P_0(B)),$$

where $M$ is the number of natural clusters of a sample of size $n_1$ from a NormCRM($\alpha, P$) with just one group, $\eta_0 = \text{eppf}(2; \alpha_0)$, and $\eta_1 = \text{eppf}(2; \alpha)$. Therefore, the expression of the covariance of the measure across groups of observations is governed only by $\eta_0$, the probability of ties in a sample from $P$, that is the probability a priori that two tables share the same dish in the Chinese franchise metaphor. The covariance takes into account only the linear dependence of two random variables, and thus indices involving higher moments are needed in order to consider different forms of dependences. In the Supplementary Materials, we extend formula (16) to the hierarchical case, and use it to derive an expression for the coskewness, $\text{CoSk}(P_1(A), P_2(A))$, as a measure of departure from linearity.
2.3 A Nonparametric Mixture Model

A NormCRM is almost surely discrete. Thus, it can be conveniently used as a mixing distribution in a mixture model to induce a clustering of the observations (Argiento et al. 2014; Favaro and Teh 2013). Similarly, a hierarchical NormCRM can be used as a mixing distribution in a hierarchical mixture model. Let $(Y_{11}, \ldots, Y_{n1}, \ldots, Y_{d1}, \ldots, Y_{dn_d})$ be a set of observations split into $d$ groups, each containing $n_j$ observations, for $j = 1, \ldots, d$, with $Y_j = (Y_{j1}, \ldots, Y_{jn_j})$ the vector of observations in the $j$-th group. We write a hierarchical NormCRM mixture model as follows:

$$Y_j | \theta_j \overset{\text{ind}}{\sim} \prod_{i=1}^{n_j} f(y_{ji} | \theta_{ji}), \quad \text{for } j = 1, \ldots, d, \quad (18)$$

$$\theta_j = (\theta_{j1}, \ldots, \theta_{jn_j}) | P_j \overset{\text{iid}}{\sim} P_j, \quad \text{for } j = 1, \ldots, d,$$

$$P_1, \ldots, P_d | P \overset{\text{iid}}{\sim} \text{NormCRM}(\alpha, P),$$

$$P \sim \text{NormCRM}(\alpha_0, P_0).$$

The law of the multidimensional array $\theta = (\theta_1, \ldots, \theta_d)$ is assigned using the $l$-clustering representation, as discussed in Section 2.1. Using Proposition 2, the infinite-dimensional parameters $(P_1, \ldots, P_d)$ and $P$ can be marginalized from (18), yielding

$$(Y_1, \ldots, Y_d) | \rho, \eta, \psi \sim \prod_{m=1}^{M} \prod_{(j,i) \in \mathcal{I}_m^{(\rho, \eta)}} f(y_{ji} | \psi_m)$$

$$\rho_j | \alpha \overset{\text{ind}}{\sim} \text{eppf}(e_{j1}, \ldots, e_{jK_j}; \alpha),$$

$$\eta | T, \alpha_0 \sim \text{eppf}(d_1, \ldots, d_M; \alpha_0),$$

$$\psi_1, \ldots, \psi_M | M \overset{\text{iid}}{\sim} P_0,$$

where the link between $\eta$ and $\rho$ is described by the transformation $t$ defined in (10). Here, $\mathcal{I}_m^{(\rho, \eta)} = \{(j, i) : j \in \{1, \ldots, d\}, i \in C_{jh}, t(j, h) \in D_m\}$ coincides with the natural clustering induced by $\theta$, i.e., the set of customers in the whole franchise eating the same dish $m \in \{1, \ldots, M\}$, according to the Chinese restaurant franchise metaphor. Model (19) is fully specified by choosing $\alpha, \alpha_0$ and $P_0$. The choice of $P_0$ can be any convenient diffuse prior that would be used in a parametric setting where the sampling model is $f(\cdot | \psi)$ (e.g., a conjugate prior). As for the choice of the Lévy intensities $\alpha$ and $\alpha_0$, satisfying (2), we suggest to depart from the Dirichlet process in cases where the aim of the statistical analysis is clustering, see Section 4 for more details. Model (19) induces a standard framework for clustering: data are considered i.i.d. within cluster $\mathcal{I}_m$, for $m = 1, \ldots, M$, while there is independence between clusters. Moreover, the prior on the random partition $\mathcal{I} := \{\mathcal{I}_1, \ldots, \mathcal{I}_M\}$ (i.e. the natural clustering) is made explicit by introducing $\rho$ and $\eta$, and letting $\mathcal{I}_m = \mathcal{I}_m^{(\rho, \eta)}$, for each $m$. 
3 Posterior Inference

In this section, we illustrate the sampling procedure for posterior inference from model (19). With this aim, we introduce a set of auxiliary variables that greatly simplifies the calculation of the predictive probabilities for the hierarchical NormCRM. We then use the Chinese restaurant franchise metaphor to describe the sampling process.

3.1 Data Augmentation

Proposition 2 is the main ingredient to fully characterize the predictive structure of a hierarchical NormCRM, which is obtained by first marginalizing (11) with respect to $\psi$, and then computing the ratio of the marginals evaluated at different clustering configurations. However, the eppf’s in formula (11) involve the computation of integrals that depend on the two Lévy intensities $\alpha$ and $\alpha_0$, see (5). In order to avoid the computation of such integrals, we resort to a standard approach for NormCRMs (see for instance Lijoi and Prünster 2010). Specifically, we introduce a vector of auxiliary variables $U = (U_1, \ldots, U_d, U_0)$, and consider only the integrand terms in formula (5), re-writing (11) as

$$L(\rho, \eta, du) = \left( \prod_{j=1}^{d} \text{eppf}(e_j; \alpha, u_j) du_j \right) \text{eppf}(d; \alpha_0, u_0) du_0$$

where $\text{eppf}(\cdot; \alpha, u_j)$, for $j = 1, \ldots, d$, and $\text{eppf}(\cdot; \alpha_0, u_0)$ are the integrand in (5) after disintegration. We mention here that, conditionally upon the total mass $T_j = \mu_j(\Theta)$ of the $j$-th unnormalized CRM, $U_j$ is gamma-distributed with shape $n_j$ and scale $T_j$, while, conditionally upon the total number of groups of the $l$-clustering across sources $T$ and the total mass $T_0 = \mu_0(\Theta)$, the variable $U_0$ is gamma-distributed with shape $T$ and scale $T_0$.

We now describe the predictive distribution of a sample from a hierarchical NormCRM, conditionally to the auxiliary variables $U$, when a new observation from one of the $d$ groups of data is added to the dataset. For sake of clarity, we will use the Chinese restaurant franchise metaphor introduced in Section 2.2, shedding light on how a new observation modifies the clustering induced by a hierarchical NormCRM. After all the customers have taken their seats in the $d$ restaurants, thus generating the table allocations $\rho$, the new $(n_j + 1)$-th customer in the $j$-th restaurant will choose an existing table with probability $P_{j_{th}}^{(to)}$, or a new one with probability $P_{j_{th}}^{(tn)}$, given by:

$$P_{j_{th}}^{(to)} := \mathbb{P}((n_j + 1) \in C_{j_{th}} | U_j = u_j) \propto \frac{\text{eppf}(e_{j1}, \ldots, e_{j_{th}} + 1, \ldots, e_{jk}; \alpha, u_j)}{\text{eppf}(e_{j1}, \ldots, e_{jk}; \alpha, u_j)} = \frac{1}{A_j} \frac{c_{e_{j_{th}}+1}(u_j)}{c_{e_{j_{th}}}(u_j)},$$

$$P_{j_{th}}^{(tn)} := \mathbb{P}((n_j + 1) \in C_{j_{th}+1} | U_j = u_j) \propto \frac{\text{eppf}(e_{j1}, \ldots, e_{jk}, 1; \alpha, u_j)}{\text{eppf}(e_{j1}, \ldots, e_{jk}; \alpha, u_j)} = \frac{c_1(u_j)}{A_j},$$

(21)
Choose table

Choose dish

\[ \begin{align*}
\text{old } P_{jh}^{*} & \quad \rightarrow \quad \text{old } P_{jh}^{*} \tag{new} \\
\text{new } P_{j}^{(n)} & \quad \rightarrow \quad P_{j}^{(n)} \tag{old} \\
\text{new } P_{j}^{(m)} & \quad \rightarrow \quad P_{j}^{(m)} \tag{new} \\
\end{align*} \]

\[ \begin{align*}
P^{(dn)} & \quad \rightarrow \quad P^{(dn)} \\
P^{(do)} & \quad \rightarrow \quad P^{(do)} \\
\end{align*} \]

Same dish

\[ \begin{align*}
\text{old } P_{jh} & \quad \rightarrow \quad \text{old } P_{jh} \tag{1} \\
\text{new } P_{j}^{(n)} & \quad \rightarrow \quad P_{j}^{(n)} \tag{new} \\
\end{align*} \]

\[ \begin{align*}
\theta_{j(n_j + 1)} & = \theta^*_{m} \quad m = 1, \ldots, M \\
\theta_{j(n_j + 1)} & = \psi_{M+1} \quad \text{new} \\
\theta_{j(n_j + 1)} & = \psi_{m} \quad m = 1, \ldots, M \\
\theta_{j(n_j + 1)} & = \theta^*_{t(j,h)} \quad h = 1, \ldots, k_j \\
\end{align*} \]

Figure 2: Probability tree of the Chinese restaurant franchise process for a hierarchical NormCRM

where \( A_j = \sum_{h=1}^{k_j} \frac{c_{j,h+1}(u_j)}{c_{j,h}(u_j)} + c_1(u_j), \) for \( h = 1, \ldots, k_j \) and \( j = 1, \ldots, d. \) If an existing table is chosen, the partition \( \eta \) is not modified. However, if the customer chooses to sit at a new \((T+1)\)-th table, the allocation structure \( \eta \) is also modified, according to the dish served in the newly-created table. This will be assigned a dish already served elsewhere with probability \( P^{(do)}_m, \) for \( m = 1, \ldots, M, \) or a new one with probability \( P^{(dn)} \), given by:

\[ P^{(do)}_m := \Pr((T + 1) \in D_m|U_0 = u_0) \propto \frac{\text{eppf}(d_1, \ldots, d_m + 1, \ldots, d_M; \alpha_0, u_0)}{\text{eppf}(d_1, \ldots, d_M; \alpha_0, u_0)} = \frac{1}{A_0} \frac{c_{dm+1}(u_0)}{c_{dm}(u_0)}, \tag{22} \]

\[ P^{(dn)} := \Pr((T + 1) \in D_{M+1}|U_0 = u_0) \propto \frac{\text{eppf}(d_1, \ldots, d_m, 1; \alpha_0, u_0)}{\text{eppf}(d_1, \ldots, d_M; \alpha_0 u_0)} = \frac{c_1(u_0)}{A_0}, \]

where \( A_0 = \sum_{m=1}^{M} \frac{c_{dm+1}(u_0)}{c_{dm}(u_0)} + c_1(u_0), \) for \( m = 1, \ldots, M. \) If a new dish is selected to be served at the \((T+1)\)-th table, its label \( \psi_{M+1} \) will be drawn from \( P_0. \) The Chinese restaurant franchise, jointly with the dish choices, is outlined in Figure 2, while the derivations of formulas (21) and (22) are given in the Supplementary Materials.

3.2 MCMC Algorithm

In this section, we concisely illustrate the steps of the MCMC algorithm required for posterior sampling under model (19). The core idea of this algorithm is to extend the one of Teh et al. (2005) for the hierarchical Dirichlet process to the more general class of hierarchical NormCRMs, hence reproducing an extended version of the generalized Chinese restaurant franchise inspired by the work of James et al. (2009) and Favaro and Teh (2013).
Conditionally to the vector of auxiliary variables $U$, the joint distribution of our model is

$$\mathcal{L}(Y_1, \ldots, Y_d | \rho, \eta, \psi, U) \mathcal{L}(\rho_1, \ldots, \rho_d | U_1, \ldots, U_d) \mathcal{L}(\eta | U_0) \prod_{m=1}^{M} P_0(d\psi_m)$$

$$= \prod_{m=1}^{M} \prod_{(j,i) \in \mathcal{I}_m(\rho, \alpha)} f(y_{ji} | \psi_m) \prod_{j=1}^{d} \text{eppf}(e_{j1}, \ldots, e_{jk_j}; \alpha, u_j) \text{eppf}(d_1, \ldots, d_M; \alpha_0, u_0) \prod_{m=1}^{M} P_0(d\psi_m).$$

The state space of the Gibbs sampler is given by $(\rho, \eta, \psi, U)$.

- **Updates of $U$ and $\psi$:** By using the expression of eppf$(\cdot; \alpha, u_j)$, for $j = 1, \ldots, d$, and eppf$(\cdot; \alpha_0, u_0)$, as well as the centering measure $P_0$, it is straightforward to see how the updates of $U$ and $\psi$ can be achieved by using standard techniques, such as Metropolis-Hastings. We give details of these sampling steps for the case of a hierarchical NGG (HNGG) mixture model (Lijoi et al. 2007) in the Supplementary Materials.

- **Updates of $(\rho, \eta)$:** The clustering parameters are updated by first marginalizing with respect to the vector of unique values $\psi$, and then updating $(\rho, \eta)$ conditionally on $U$, resorting to the generalized Chinese restaurant franchise process. Let the superscript $(-ji)$ denote the random variables modified after the removal of the $i$-th observation of the $j$-th group, for $i = 1, \ldots, n_j$ and $j = 1, \ldots, d$. Conditionally to $Y$ and $(\rho^{-ji}, \eta^{-ji})$, the probability of assigning the $i$-th customer to the $h$-th table of the $j$-th restaurant, where the $m$-th dish is served (i.e., $t(j, h) \in D_m^{-ji}$), is

$$\mathbb{P}(i \in C_{jh}^{-ji}, t(j, h) \in D_m^{-ji} | Y, \rho^{-ji}, \eta^{-ji}, U_j = u_j) \propto \mathcal{M}(y_{ji} | y_{m\rho^{-mji}, \eta^{-mji}}) \mathbb{P}(i \in C_{jh}^{-ji}, t(j, h) \in D_m^{-ji} | \rho^{-ji}, \eta^{-ji}, U_j = u_j). \quad (23)$$

In the latter Formula (23), when $h = 1, \ldots, k_j^{-ji}$, then $m$ is the value such that $t(j, h) \in D_m^{-ji}$, i.e. $C_{jh}^{-ji}$ is an already occupied table where the $m$-th dish is served. Moreover, when $h = (k_j^{-ji} + 1)$, then $m = 1, \ldots, M^{-ji} + 1$, i.e. when a new table is allocated one of the already served dishes or a new one can be chosen. Here, $\mathcal{M}(y_{ji} | y_{m\rho^{-mji}, \eta^{-mji}})$ is the predictive density of a parametric Bayesian model where the sampling model is $f(y | \theta)$, the prior is $P_0$, and the observations are all the data with index in $\mathcal{I}_m$, with proviso that $\mathcal{I}_{M^{-ji}+1}$ is the empty set. Finally, probability $\mathbb{P}(i \in C_{jh}^{-ji}, t(j, h) \in D_m^{-ji} | \rho^{-ji}, \eta^{-ji}, U_j = u_j)$ is reported at the leaves of the tree in Figure 2 (see also formula (25) below). Thus, (23) is proportional to the prior probability that the $i$-th customer of the $j$-th restaurant will seat at the $h$-th table, updated by considering the information yielded by the customers of the franchise eating the same dish. This dish is served at table $h$ and at all the other tables such that $t(j, h) \in D_m^{-ji}$, for $j = 1, \ldots, d$. This step of the algorithm clarifies how the sharing of information between individuals and across groups takes place in model (19).

The updating process continues by re-allocating $C_{jh}$ to a cluster of tables. To this end, we have to assign $t = t(j, h)$ to a new/old cluster of tables $D_m$. More formally, let the superscript $(-t)$ indicate the variables after the removal of all the observations in $C_{jh}$

15
such that \( t = t(j, h) \). Conditionally on \( Y \) and \((\rho^{-t}, \eta^{-t})\), the probability of assigning the \( t \)-th table to the \( m \)-th cluster is

\[
P(t \in D^{-t}_m | Y, \rho^{-t}, \eta^{-t}, U_0 = u_0) \propto \mathcal{M}(y_t | Y_{\mathcal{I}_m^{(\rho^{-t}, \eta^{-t})}}) \P(t \in D^{-t}_m | \rho^{-t}, \eta^{-t}, U_0 = u_0),
\]

for \( m = 1, \ldots, M^{-t} + 1 \), where \((M^{-t} + 1)\) indicates the new cluster of tables. Here, \( Y_t := \{Y_{ij} : i \in C_{jh}, t = t(j, h)\} \) is the set of all the observations (customers) in the \( l \)-cluster \( C_{jh} \) (i.e., the \( t \)-th table), and \( \mathcal{M}(y_t | y_{\mathcal{I}_m}) \) is the joint predictive density of \( e_{jh} \) observations from a parametric Bayesian model where the sampling model is \( f(y | \theta) \), the prior is \( P_0 \), and the observations are all the data with index in \( \mathcal{I}_m \). Probability \( \P(t \in D^{-t}_m | \rho^{-t}, \eta^{-t}, U_0 = u_0) \) has been introduced in (22), and it is the predictive probability prescribed by the generalized Chinese restaurant process of the table indices. Formula (24) is proportional to the prior probability that the \( t \)-th table of the \( j \)-th restaurant will share the \( m \)-th dish, updated by considering the information yielded by the customers of the franchise eating the same dish. This step of the algorithm clarifies how the sharing of information within and between groups takes place in model (19).

The computation of the conditional probabilities \( \mathcal{M} \) in (23) and (24) is available in closed analytical form only when \( P_0 \) and \( f(\cdot | \theta) \) in (19) are conjugate. For the non-conjugate case, an extension of the proposed algorithm, that uses Algorithm 8 of Neal (2000) and the algorithm of Favaro and Teh (2013), is described in the Supplementary Materials.

3.3 Prediction

Consider a configuration \((\rho, \eta, \psi)\), and hence \( \theta \), of \( n \) observations in \( d \) groups, as well as a new dish label indicated as \( \psi_{M+1} \). We want to make a prediction within the \( j \)-th group of data, according to the law of \( \theta_{j(n_j + 1)} \). Combining equations (21) and (22) as outlined in Figure 2, we express this law, jointly with the \( j \)-th \( l \)-clustering configuration, as follows:

\[
P(\theta_{j(n_j + 1)} = \psi_m, (n_j + 1) \in C_{jh} | \rho, \eta, \psi) = \begin{cases} 
P^{(to)}_{jh} & h = 1, \ldots, k_j, \quad m = t(j, h) \\
P^{(tn)}_{j} P^{(do)}_m & h = k_j + 1, \quad m = 1, \ldots, M \\
P^{(tn)}_{j} P^{(dn)} & h = k_j + 1, \quad m = M + 1 
\end{cases}
\]

(25)

where \( \psi_{M+1} \sim P_0 \). Interestingly, since we have assumed that the groups are exchangeable and infinitely many are allowed, as it is usual in hierarchical models, we can make a prediction on the arrival of a new observation in a new group as:

\[
P(\theta_{(d+1)1} = \psi_m | \rho, \eta, \psi) = \begin{cases} 
P^{(do)}_m & h = 1 \text{ and } m = 1, \ldots, M, \\
P^{(dn)} & h = 1 \text{ and } m = M + 1.
\end{cases}
\]

(26)

We now derive the law of a new observation given the multidimensional array of data \( Y = (Y_1, \ldots, Y_d) \) for the mixture in model (19). It is easy to see that, conditionally on \((\rho, \eta, \psi)\), this quantity changes depending on the group. In particular, the predictive densities in an
existing or in a new group are:

\[
p(y_{j(n_j+1)}|Y, \rho, \eta, \psi) = \sum_{h=1}^{k_j} P^{(to)}_{jh} f(y_{j(n_j+1)}|\theta_{jh}) + P^{(tn)}_{j} \sum_{m=1}^{M} P^{(do)}_{m} f(y_{j(n_j+1)}|\psi_{m}) + P^{(tn)}_{j} \int f(y_{j(n_j+1)}|\psi)P_{0}(d\psi),
\]

\[
p(y_{(d+1)}|Y, \rho, \eta, \psi)) = \sum_{m=1}^{M} P^{(do)}_{m} f(y_{(d+1)}|\psi_{m}) + P^{(dn)} \int f(y_{(d+1)}|\psi)P_{0}(d\psi), \tag{27}
\]

where the marginal distribution \( M(y) = \int f(y|\psi)P_{0}(d\psi) \) is often not available in practice, and can be approximated via Monte Carlo integration. Finally, the unconditional predictive distribution is computed by averaging (27) with respect to the posterior sample of \((\rho, \eta, \psi)\), drawn using the Gibbs sampler described in Section 3.2.

As a final remark, in the non-hierarchical case, where the \( d \) processes in model (18) are independent (i.e., \( P \) degenerates to the diffuse centering measure \( P_{0} \)), \( P^{(do)}_{m} = 0 \), for each \( m = 1, \ldots, M \), and \( P^{(dn)} = 1 \). Hence, within each group, the predictive structure of a standard Chinese Restaurant process is recovered, and the predictive distribution in a new group coincides with \( M(y) \).

4 Applications

In this section, we assess the performance of the proposed mixture model on both simulated and benchmark datasets. We focus on the normalized generalized gamma process (NGG) (Lijoi et al. 2007), yielding a hierarchical NGG (HNGG) mixture model. This is obtained from model (19) when the Lévy intensities are

\[
\alpha(s) = \frac{\kappa}{\Gamma(1-\sigma)}s^{-1-\sigma}e^{-s}I_{(0, +\infty)}(s),
\]

\[
\alpha_{0}(s) = \frac{\kappa_{0}}{\Gamma(1-\sigma_{0})}s^{-1-\sigma_{0}}e^{-s}I_{(0, +\infty)}(s),
\]

so that the hyperparameters of the HNGG process are \((\kappa, \kappa_{0}, \sigma, \sigma_{0})\), with \( \kappa, \kappa_{0} > 0 \) and \( \sigma, \sigma_{0} \in (0, 1) \). We choose \( f(\cdot|\theta) \) as a Gaussian kernel with \( \theta = (\mu, \tau^{2}) \) representing the mean and variance parameters, and a centering measure \( P_{0} \) as a conjugate normal-inverse-gamma with parameters \((m_{0}, k_{0}, a_{\tau^{2}}, b_{\tau^{2}})\). Under this model, the marginal distributions are known, allowing us to adopt the marginal algorithm introduced in Section 3.2. Non-conjugate independent normal and inverse-gamma priors on the mean and variance parameters can also be used (see Supplementary Materials). The NGG adds flexibility to random partition models, mitigating the “rich-gets-richer” effect of the commonly used Dirichlet process via the introduction of an additional parameter. We refer readers to the Supplementary Materials for the expressions of equations (21) and (22) in this setting.

4.1 Simulation Study \((d = 2)\)

Our goal is to assess the performance of the nonparametric HNGG process in recovering the original clustering of the observations, as well as to evaluate its goodness of fit. Here, we first study performances for different values of the hyperparameters \( \kappa, \kappa_{0}, \sigma, \sigma_{0} \) of the HNGG. This allows us to obtain different models, including the Dirichlet counterpart, i.e., the Hierarchical
Dirichlet Process (HDP) model obtained for $\sigma = \sigma_0 = 0$, and an independent model where $(P_1, \ldots, P_d) \overset{iid}{\sim} \text{NGG}(\kappa_0, \sigma_0, P_0)$. We also consider the case of random hyperparameters.

We simulated the data from two distinct groups of 100 observations, each sampled from a two-component Gaussian mixture, with one component shared between the two groups. Membership to the Gaussian components identifies a clustering structure across groups into three clusters, which we will refer to as the true partition of the data. For $i = 1, \ldots, 100$:

$$
y_{1i} \overset{iid}{\sim} 0.2N(y_{1i} | -3, 0.1) + 0.8N(y_{1i} | 0, 0.5); \quad y_{2i} \overset{iid}{\sim} 0.1N(y_{2i} | 0, 0.5) + 0.9N(y_{2i} | 1, 1.5).
$$

We observe how the component shared by the two groups has a higher weight in the first mixture and we argue that, unlike our hierarchical model that shares information across groups, the independent model will not be able to adequately recover this component.

We fitted a HNGG model with two groups, i.e. $j = 1, 2$ and assessed performance of the proposed model using the log-pseudo marginal likelihood (LPML) and the Rand index (RI). The first measures the goodness-of-fit of the model, while the second measures the goodness-of-clustering. For the RI, we first estimated the natural clustering as the partition minimizing the Binder’s loss function (see Lau and Green 2007), and then computed the Rand index against the true partition of the data in the second group, since this presents the most interesting features from a clustering perspective. Throughout the simulation study, we fixed the hyperparameters of $P_0$ equal to $(m_0, k_0, a, b) = (0.25, 0.62, 2.07, 0.66)$, a specification that corresponds to $\mathbb{E}[\mu] = \hat{y}_n$, $\mathbb{E}[\tau^2] = \hat{s}_n^2/3$, $\text{Var}(\mu) = 1$, $\text{Var}(\tau^2) = 5$, where $\hat{y}_n$ and $\hat{s}_n^2$ represent the sample mean and variance of the whole dataset, respectively.

The choice of the hyperparameters of the centering measure $P_0$ is crucial, as it influences both the fitting and the clustering estimation. Results obtained from a sensitivity analysis, reported in the Supplementary Materials, show that there is a trade-off between goodness-of-clustering and goodness-of-fit, when varying the values of the variances of the parameters a priori. In particular, while too large variances yield poor clustering performance in terms of Rand index (RI), they can help improve the model fitting in term of log-pseudo marginal likelihood (LPML). This behaviour is observed also in the sensitivity analysis concerning the hyperparameters $(\kappa, \kappa_0, \sigma, \sigma_0)$ of the HNGG process.

We first show results obtained with several combinations of values for the parameters $(\kappa, \kappa_0, \sigma, \sigma_0)$. This strategy, in particular, allows comparison with various HDP and independent models. Values of LPML and RI are reported in Tables 1 and 2, respectively, for some of the parameter settings. The estimated number of natural clusters in the second group (i.e., the number of different dishes served in the second restaurant) is reported in brackets in Table 2. Results on additional settings are reported in the Supplementary Materials. In general, the best LPML values are observed when $\sigma$ or $\sigma_0$ are different from zero, meaning that the HNGG outperforms the HDP in terms of goodness-of-fit. On the other hand, higher values of LPML, obtained for large values of the parameters $\sigma$ or $\sigma_0$, correspond also to higher numbers of clusters in each group, as it can be observed in Table 2. Furthermore, we observe that the RI alone is not indicative of good recovery of the true partition, since its maximum value is reached when all the observations in the second group are clustered together. A better clustering of the data in the second group is instead provided by two estimated clusters and a relatively high RI, as we can observe in some of the cases where...
\[ \sigma = \sigma_0 \approx 0 \quad \text{and} \quad (\sigma, \sigma_0) = (0.3, 0.1) \quad \text{and} \quad (\sigma, \sigma_0) = (0.1, 0.3) \]

<table>
<thead>
<tr>
<th>( \kappa )</th>
<th>( \kappa_0 = 0.1 )</th>
<th>( \kappa_0 = 1 )</th>
<th>( \kappa_0 = 10 )</th>
<th>( \kappa_0 = 0.1 )</th>
<th>( \kappa_0 = 1 )</th>
<th>( \kappa_0 = 10 )</th>
<th>( \kappa_0 = 0.1 )</th>
<th>( \kappa_0 = 1 )</th>
<th>( \kappa_0 = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \kappa = 0.1 )</td>
<td>-0.2743</td>
<td>-0.2864</td>
<td>-0.2739</td>
<td>-0.2835</td>
<td>-0.2751</td>
<td>-0.2754</td>
<td>-0.2758</td>
<td>-0.2786</td>
<td>-0.2770</td>
</tr>
<tr>
<td>( \kappa = 1 )</td>
<td>-0.2826</td>
<td>-0.2812</td>
<td><strong>-0.2705</strong></td>
<td>-0.2753</td>
<td>-0.2708</td>
<td>-0.2727</td>
<td>-0.2760</td>
<td>-0.2774</td>
<td><strong>-0.2696</strong></td>
</tr>
<tr>
<td>( \kappa = 10 )</td>
<td>-0.2775</td>
<td>-0.2740</td>
<td>-0.2734</td>
<td>-0.2723</td>
<td><strong>-0.2698</strong></td>
<td>-0.2725</td>
<td>-0.2706</td>
<td>0.2712</td>
<td>0.2752</td>
</tr>
</tbody>
</table>

Table 1: Simulated Data \((d = 2)\) – LPML(10^3) values for different combinations of the parameters \((\kappa, \kappa_0, \sigma, \sigma_0)\).

\[ \sigma = \sigma_0 \approx 0 \quad \text{and} \quad (\sigma, \sigma_0) = (0.3, 0.1) \quad \text{and} \quad (\sigma, \sigma_0) = (0.1, 0.3) \]

<table>
<thead>
<tr>
<th>( \kappa )</th>
<th>( \kappa_0 = 0.1 )</th>
<th>( \kappa_0 = 1 )</th>
<th>( \kappa_0 = 10 )</th>
<th>( \kappa_0 = 0.1 )</th>
<th>( \kappa_0 = 1 )</th>
<th>( \kappa_0 = 10 )</th>
<th>( \kappa_0 = 0.1 )</th>
<th>( \kappa_0 = 1 )</th>
<th>( \kappa_0 = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \kappa = 0.1 )</td>
<td>0.5840</td>
<td>0.5079</td>
<td><strong>0.7568</strong></td>
<td>0.5194</td>
<td>0.4958</td>
<td>0.4885</td>
<td>0.4952</td>
<td>0.5048</td>
<td>0.4958</td>
</tr>
<tr>
<td>( \kappa = 1 )</td>
<td>0.5291</td>
<td>0.5079</td>
<td>0.4317</td>
<td><strong>0.5404</strong></td>
<td>0.4380</td>
<td>0.4083</td>
<td>0.4590</td>
<td><strong>0.5222</strong></td>
<td>0.4115</td>
</tr>
<tr>
<td>( \kappa = 10 )</td>
<td>0.4968</td>
<td>0.3844</td>
<td>0.3319</td>
<td>0.5048</td>
<td>0.3519</td>
<td>0.2749</td>
<td>0.3325</td>
<td>0.3513</td>
<td>0.3123</td>
</tr>
</tbody>
</table>

Table 2: Simulated Data \((d = 2)\) – Rand Index (RI) and estimated number of natural clusters in the second group for different combinations of the parameters \((\kappa, \kappa_0, \sigma, \sigma_0)\).

\( \sigma \) and \( \sigma_0 \) differ from 0. Results on additional quantities of interest a priori, such as the expectation and variance of \( M \), and the covariance and coskewness of \( P_1(A) \) and \( P_2(A) \), with \( A \) a neighbourhood of zero, are reported in the Supplementary Materials. These results show how larger values of the hyperparameters \((\kappa, \kappa_0, \sigma, \sigma_0)\) influence the distribution of the number of clusters, as they induce an increase in the expectation and variance of \( M \). On the other hand, increasing the values of \((\kappa_0, \sigma_0)\) also increases the dependency between measures, while an opposite trend is observed when increasing \((\kappa, \sigma)\). As expected, the highest values of RI are obtained when the prior expected value of \( M \) is close to the truth. On the other hand, optimal LPML values are obtained for high dependency a priori between \( P_1(A) \) and \( P_2(A) \).

Next, we assess the performance of an independent NGG model obtained as \((P_1, \ldots, P_d) \overset{iid}{\sim} \text{NGG}(\kappa_0, \sigma_0, P_0)\) with \( \sigma_0 = 0.1 \) and \( \kappa_0 = 0.1 \). Posterior density estimates are reported in Figure 3(a), with the histograms of the data coloured according to the true partition. The histogram of the data associated with the component shared between the two groups is depicted in purple. The density estimation in the second group is characterized by one mode rather than two, resulting from the absence of sharing of information between groups, see also Figure 3(d). Furthermore, the predictive distribution of a new group coincides with the marginal distribution \( M(y) \), as depicted in Figure 3(b). We contrast these results with the estimated densities obtained by fitting a HNGG model with \( \kappa = \kappa_0 = 0.1 \) and \( \sigma, \sigma_0 \overset{\text{Beta}(2,18)}{\sim} (\mathbb{E}[\sigma] = \mathbb{E}[\sigma_0] = 0.1) \). Posterior density estimates are reported in Figure 4(a). We notice how the predictive density in the second group is now able to estimate the component with fewer observations, taking advantage of the sharing of information. In Figure 4(b), the estimate of the predictive density in a new group is plotted over the histogram of the whole dataset, regardless of the group information. The shared component in the second group is no longer visible, as expected. However, this is clearly recovered in the infer-
Figure 3: Simulated Data \((d = 2)\) – Summary plots for the independent NGG case with \(\sigma_0 = 0.1\) and \(\kappa_0 = 0.1\). (a): Posterior density estimates and histograms of the data, colored according to the true partition. (b): Predictive density for a new group. Posterior distribution of the number of clusters for all observations, \(M\) (c), and in each group (d).

ence, as shown by the posterior distribution of the number of clusters for all observations in Figure 4(c), and at group level in Figure 4(d). Figures 4(e)-4(f) depict posterior histograms and kernel density estimates of \(\sigma\) and \(\sigma_0\), showing a clear departure from the HDP case.

We performed a comparison between the proposed approach and two simpler models: the Bayesian parametric model of Hoff (2009), and a frequentist mixed-effects model of Pinheiro and Bates (2000). Density estimation under the frequentist approach was obtained via a parametric bootstrap technique. We refer readers to the Supplementary Materials for additional details on these comparisons. Figure 5(a) reports the density estimation results under the two parametric models, clearly showing how both models fail to recover the bi-modality of the densities in the groups.

In order to study the behavior of our proposed model for larger numbers of groups, we performed an additional simulation study with \(d = 100\). For this simulation, we show results in terms of receiver operating characteristic (ROC) curves, computed averaging over 25 replicated datasets, for the hierarchical and independent NGG models, respectively, see Figure 5(b). The ROC curves are computed by considering as true positive the event that two elements are correctly clustered together, and as false positive the event that they are
Figure 4: Simulated Data ($d = 2$) – Summary plots for the HNGG model with $\sigma, \sigma_0 \sim \text{Beta}(2, 18)$ and $\kappa = \kappa_0 = 0.1$. (a): Posterior density estimates and histograms of the data, colored according to the true partition. (b): Predictive density for a new group. Posterior distribution of the number of clusters for all observations, $M$ (c), and in each group (d). (e,f): Posterior distributions for the parameters $\sigma$ and $\sigma_0$. 
erroneously clustered together. Our results clearly show that the HNGG model outperforms its independent counterpart in terms of accuracy of the clustering. Additional details of these comparisons can be found in the Supplementary Materials.

4.2 Application to the school data (Hoff, 2009)

In this section, we show an application of the proposed HNGG model to the school dataset used in the popular textbook by Hoff (2009), where Bayesian hierarchical models are presented. The data are part of the 2002 Educational Longitudinal Study (ELS), a survey of students from a large sample of schools across the United States. The observations represent the math scores of 10th grade children from \( d = 100 \) American high-schools. Here, we report the results obtained by fitting the HNGG model (19) with a non-conjugate prior, such that: 

\[
P_0(\mu, \tau^2) = p(\mu)p(\tau^2) = N(\mu|50, 25) \text{inv-gamma}(\tau^2|0.5, 50),
\]

where the hyperparameters were fixed as in Hoff (2009), chap. 8. In order to allow for more robustness in the inference process, we imposed prior distributions on \( \kappa, \kappa_0 \sim \text{gamma}(1, 1) \) and \( \sigma, \sigma_0 \sim \text{Beta}(2, 18) \).

Figure 6(a) shows the data organized by school. The order of the schools is given by increasing sample mean in each group, and the color of each data point refers to its natural cluster assignment, obtained by minimizing the Binder’s loss function, which identified 5 clusters. Three major clusters can be observed, corresponding to students with low (squares), medium (dots), or high (diamonds and triangles) math scores, respectively. However, these clusters also characterize different school compositions: on one hand, low-sized schools are composed of only one type of students, while on the other hand, when the number of students increases, we observe more heterogeneity in the school composition. We argue that this could be explained by additional latent variables representing some socio-economical information.

To explore the clustering structure at group level, in Figure 6(c) we plot the posterior mean of the number of elements in \( \rho_j \), i.e. \( l \)-clusters, for \( j = 1, \ldots, 100 \). We observe some het-
erogeneity, with some schools having just one $l$-cluster of students, and others with up to three different $l$-clusters. We then selected the 3 schools with the highest posterior expected numbers of $l$-clusters (schools 98, 1, 12) and the 3 with the lowest (schools 67, 51, 72), and estimated the corresponding predictive densities, see Figure 6(d). The composition of the selected schools is shown by plotting the observations underneath the predictive densities, colored according to the natural clustering estimated via the Binder’s loss. The intensity of the grey scale increases with the posterior expected number of $l$-clusters. Schools 67 and 51 have students with higher math scores, while school 72 is characterized by lower math scores. The other three selected schools present a more heterogeneous composition. This confirms our interpretation of the results in Figure 6(d). In Figure 6(b), the predictive density in a new group is depicted, with the histogram of the whole dataset obtained without considering the group information. The predictive density does not appear to be multimodal, showing how the proposed mixture model preserves the shrinkage effect typical of the Bayesian hierarchical models while the underlying clustering allows for a more detailed interpretation of the information in the data.

Finally, following the suggestion of one of the reviewers, we performed a comparison of our results with a simple parametric hierarchical model fitted in (Hoff 2009, Chapter 8). In Figure 6(e) the predictive densities under the parametric model are reported for a selection of schools. Comparing these densities with the corresponding ones in panel (d), it is clear how the parametric model does not capture the skewness and the heavy tails of the data, as it does not allow for heterogeneity within groups. Additional details on this comparison can be found in the Supplementary Materials.
Figure 6: School data by Hoff (2009) – (a): Data sorted by increasing sample mean in each school (vertical lines). Colors and marker shapes identify the estimated clustering. (b): Predictive density of the math score for a student in a new school. (c): Posterior mean of the number of elements in $\rho_j$, i.e. $l$-clustering, for $j = 1, \ldots, 100$. (d): Predictive densities of the math score for a new student in selected schools under the HNGG model. The points reported on the bottom lines are the observations in the groups colored according to the estimated partition. The gray scale reflects the posterior expected number of $l$-clusters in the schools. (e): Predictive densities for the same schools are in panel (d), estimated under the Bayesian parametric model (Hoff 2009). The gray scale of the points reflects the posterior expected number of $l$-clusters in the schools under the HNGG model.
5 Conclusion

In this paper, we have fully investigated the clustering induced by a NormCRM mixture model. This model is suitable in cases of data belonging to specific groups or categories in which similar characteristics are shared. At group level, each NormCRM is centered on the same base measure, which is a NormCRM itself. The discreteness of the shared base measure implies that the processes at data level share the same atoms. This desired feature allows to cluster together observations of different groups. By integrating out the nonparametric components of our prior (i.e. $P_1, \ldots, P_d, P$), we have obtained a representation of our model through formula (19) that sheds light on the hierarchical clustering induced by the mixture. At the first level of the hierarchy, data are clustered within each of the groups ($l$-clustering). These partitions are i.i.d. with law that is identified by the eppf induced by the NormCRM($\alpha, P$), that is the law of the mixing measure at the same level of the hierarchy. These $l$-clusters can in turn be aggregated into $M$ clusters according to the partition induced by the eppf at the lowest level of the hierarchy, corresponding to NormCRM($\alpha_0, P_0$). This clustering structure reveals the sharing of information among the groups of observations in the mixture model. Furthermore, we have offered an interpretation of this hierarchical clustering in terms of the generalized Chinese restaurant franchise process, which has allowed us to perform posterior inference in the presence of both conjugate and non-conjugate models. We have provided theoretical results concerning the a priori distribution of the number of clusters, within or between groups, and a general formula to compute moments and mixed moments of general order. To evaluate the model performance and the elicitation of the hyperparameters, we have conducted a simulation study and an analysis on a benchmark dataset. Results have shed insights on the sharing of information among clusters and groups of data, showing how our model is able to identify components of the mixture that are less represented in a group of data.

The proposed characterization of the mixture model in terms of the clustering structure it induces has the potential to be generalized. For example, an interesting future avenue is to investigate extensions to situations where covariates are available. The use of covariate-dependent priors in Bayesian nonparametrics is relatively new. In his seminal work, MacEachern (1999) considered a semiparametric mixture model framework, manipulating the infinite-dimensional mixing measure in order to incorporate covariate information. A different approach, based on product partition models, was proposed by Müller and Quintana (2010), where covariates directly modify the eppf to drive prior knowledge on the random partition. An interesting review is provided by Barcella et al. (2017) who also show how a product partition model with covariates reduces to a mixture model on the joint distribution of the response and the covariates.

6 Acknowledgements

Raffaele Argiento gratefully acknowledges Collegio Carlo Alberto for partially funding this work. The authors also thank the Norway Centre for Molecular Medicine (NCMM) IT facility for the computational support.

SUPPLEMENTARY MATERIAL
Title: Supplementary Materials  The file:  
Argiento_Cremaschi_Vannucci_HNCRM_Supplementary_Materials.pdf 
reports additional details on the material presented in the main paper. This includes 
proofs of the theoretical results presented in the paper and details on how to com-
pute covariance and coskewness of a hierarchical NormCRM, details on the MCMC 
algorithm and additional results from the simulation studies.

Title: Code  The Matlab code implementing the algorithm described in the paper (both 
conjugate and non), is publicly available on GitHub https://github.com/AndCre87/HNCRM

References

Argiento, R., Bianchini, I., Guglielmi, A., et al. (2016). Posterior sampling from ε-
approximation of normalized completely random measure mixtures. *Electronic Journal 

ter analysis using species sampling Gaussian mixture models. *Journal of Computational 

techniques for covariate dependent Dirichlet process mixture models. *Canadian Journal 


Camerlenghi, F., Lijoi, A., and Prünster, I. (2017). Bayesian prediction with multiple-

Brooks/Cole.


Hierarchical Normalized Completely Random Measures to Cluster Grouped Data: Supplementary Materials

Raffaele Argiento
ESOMAS Department, University of Torino and Collegio Carlo Alberto, Torino, Italy
and
Andrea Cremaschi
Centre for Molecular Medicine Norway, Nordic EMBL Partnership, University of Oslo, Oslo, Norway
Oslo Centre for Biostatistics and Epidemiology, University of Oslo, Oslo, Norway
and
Marina Vannucci
Department of Statistics, Rice University, Houston, TX, USA
In Section 1 we report details of the proofs of the theoretical results presented in the paper, while in Section 2 we describe the MCMC algorithm for posterior inference and its theoretical justification. Section 3 contains additional numerical results from the simulation study with two groups \((d = 2)\) and the computation of some quantities of interest a priori, such as the moments of the distribution of the number of clusters \(M\) and the mixed moments for two random measures \(P_1\) and \(P_2\). Furthermore, Section 3 illustrates the results of two extensive sensitivity analysis performed on the hyperparameters of the HNGG process and of the centering measure \(P_0\), respectively. To further explore the efficacy of the proposed model, comparisons with parametric Bayesian and frequentist approaches are reported in Section 4. Finally, results for an additional simulation study characterized by \(d = 100\) groups is reported in Section 5.

1 Proofs

1.1 Proof of Lemma 1

Proof. Let \(W := \{w_l : \sum_{l \geq 1} w_l = 1\}\) be an infinite sequence of random weights, obtained by normalization of a CRM as in formula (1) in the paper. Additionally, let \(T_1 := \{\tau_l : l \geq 1\}\) be an infinite sequence of i.i.d. random locations from \(P\). Firstly we observe that:

\[
\mathcal{L}(d\tilde{\theta}_1, \ldots, d\tilde{\theta}_n) = \int P_1(d\tilde{\theta}_1) \cdots P_1(d\tilde{\theta}_n) \mathcal{L}(dP_1)
= \int_{\Delta^\infty \times \Theta^\infty} \sum_{l_1 \geq 1} w_{l_1} \delta_{\tau_{l_1}}(d\tilde{\theta}_1) \cdots \sum_{l_n \geq 1} w_{l_n} \delta_{\tau_{l_n}}(d\tilde{\theta}_n) \mathcal{L}(dW) \mathcal{L}(dT_1),
\]
where $\mathcal{P}$ is the space of all the probability measures on $\Theta$, and $\Delta^\infty$ is the infinite dimensional simplex. On the other hand, we have that:

$$\mathcal{L}(d\theta_1, \ldots, d\theta_n) = \int_{\Theta^\infty} \sum_{l_1 \geq 1} \cdots \sum_{l_n \geq 1} \mathcal{L}(d\theta_1, \ldots, d\theta_n | l_1, \ldots, l_n, T_1) \mathcal{L}(l_1, \ldots, l_n) \mathcal{L}(dT_1)$$

$$= \int_{\Theta^\infty} \sum_{l_1 \geq 1} \cdots \sum_{l_n \geq 1} \delta_{\tau_1}(d\theta_1) \cdots \delta_{\tau_n}(d\theta_n) \int_{\mathcal{P}^\ast} \mathcal{L}(l_1, \ldots, l_n | P_1^\ast) \mathcal{L}(dP_1^\ast) \mathcal{L}(dT_1)$$

$$= \int_{\Theta^\infty} \sum_{l_1 \geq 1} \cdots \sum_{l_n \geq 1} \delta_{\tau_1}(d\theta_1) \cdots \delta_{\tau_n}(d\theta_n) \int_{\mathcal{P}^\ast} P_1^\ast(l_1) \cdots P_1^\ast(l_n) \mathcal{L}(dP_1^\ast) \mathcal{L}(dT_1)$$

$$= \int_{\Theta^\infty} \sum_{l_1 \geq 1} \cdots \sum_{l_n \geq 1} \delta_{\tau_1}(d\theta_1) \cdots \delta_{\tau_n}(d\theta_n) \int_{\Delta^\infty} \sum_{h_1 \geq 1} \sum_{h_n \geq 1} w_{h_1} \delta_{h_1}(l_1) \cdots w_{h_n} \delta_{h_n}(l_n) \mathcal{L}(dW) \mathcal{L}(dT_1)$$

$$= \int_{\Theta^\infty} \sum_{l_1 \geq 1} \cdots \sum_{l_n \geq 1} \delta_{\tau_1}(d\theta_1) \cdots \delta_{\tau_n}(d\theta_n) \int_{\Delta^\infty} w_1 \cdots w_n \mathcal{L}(dW) \mathcal{L}(dT_1)$$

$$= \int_{\Delta^\infty \times \Theta^\infty} \sum_{l_1 \geq 1} \cdots \sum_{l_n \geq 1} w_1 \delta_{\tau_1}(d\theta_1) \cdots \sum_{l_n \geq 1} w_n \delta_{\tau_n}(d\theta_n) \mathcal{L}(dW) \mathcal{L}(dT_1).$$

This ends the proof, since the two marginal laws have the same integral representation. \hfill \Box

### 1.2 Proof of Proposition 1

**Proof.** We first observe that a sample $(\theta_1, \ldots, \theta_n)$ can be reconstructed via the partition $\rho$ and the vector $(\theta_1^*, \ldots, \theta_n^*)$ of an $l$-clustering representation, since it is sufficient to let $\theta_{i} = \theta_{i}^*$ if $i \in C_i$ for $i = 1, \ldots, n$ and $l = 1, \ldots, K_n$. From Lemma 1, we have

$$\mathcal{L}(d\theta_1, \ldots, d\theta_n) = \int_{\Theta^\infty} \sum_{l_1 \geq 1} \cdots \sum_{l_n \geq 1} \delta_{\tau_1}(d\theta_1) \cdots \delta_{\tau_n}(d\theta_n) \int_{\Delta^\infty} w_1 \cdots w_n \mathcal{L}(dW) \mathcal{L}(dT)$$

$$= \int_{\Theta^\infty} \sum_{l_1 \geq 1} \cdots \sum_{l_n \geq 1} \delta_{\tau_1}(d\theta_1) \cdots \delta_{\tau_n}(d\theta_n) \mathcal{L}(l_1, \ldots, l_n) \mathcal{L}(dT),$$

and

$$\mathcal{L}(d\theta_1, \ldots, d\theta_n, l_1, \ldots, l_n) = \mathcal{L}(l_1, \ldots, l_n) \int_{\Theta^\infty} \delta_{\tau_1}(d\theta_1) \cdots \delta_{\tau_n}(d\theta_n) \mathcal{L}(dT).$$

Hence, using the $l$-clustering representation and Lemma 1:

$$\mathcal{L}(d\theta_1^*, \ldots, d\theta_k^*, l_1^*, \ldots, l_k^*, \rho) = \mathcal{L}(l_1^*, \ldots, l_k^*, \rho) \int_{\Theta^\ast} \delta_{\tau_1^*}(d\theta_1^*) \cdots \delta_{\tau_k^*}(d\theta_k^*) \mathcal{L}(dT)$$

$$= \mathcal{L}(l_1^*, \ldots, l_k^*, \rho) P(\theta_1^*) \cdots P(\theta_k^*),$$

3
where the latter holds because $\mathcal{L}(dT) = P^\infty(d\{\tau_h, h = 1, 2, \ldots\})$. Finally, marginalizing over $(l_1^*, \ldots, l_k^*)$:

$$
\mathcal{L}(d\theta_1^*, \ldots, d\theta_k^*, \rho) = \sum_{l_1^*, \ldots, l_k^*} \mathcal{L}(l_1^*, \ldots, l_k^*, \rho) P(\theta_1^*) \ldots P(\theta_k^*)
$$

$$
= \mathcal{L}(\rho) P(\theta_1^*) \ldots P(\theta_k^*)
$$

$$
= \text{eppf}(e_1, \ldots, e_{k_n}; \alpha) \prod_{h=1}^{k_n} P(d\theta_h^*).
$$

\[ \square \]

1.3 Proof of Proposition 2

Proof. Given model (8) in the paper, we can exploit the hierarchy used to model the multidimensional array $\theta = (d\theta_1, \ldots, d\theta_d)$ to write the joint law as

$$
\mathcal{L}(d\theta) = \int_{\mathcal{P}_1} \ldots \int_{\mathcal{P}_d} \mathcal{L}(d\theta|P_1, \ldots, P_d) \mathcal{L}(dP_1, \ldots, dP_d) = \int_{\mathcal{P}} \prod_{j=1}^d \mathcal{L}(d\theta_j|P_j) \mathcal{L}(dP_j|P) \mathcal{L}(dP),
$$

where $(\mathcal{P}_1, \ldots, \mathcal{P}_d)$, and $\mathcal{P}$ equivalently represent the space of all probability measures on $\Theta$ (the index is used for notational convenience). As for Proposition 1, we can write each $j$-th inner integral as

$$
\int_{\mathcal{P}_j} \prod_{j=1}^d \mathcal{L}(d\theta_j^*, \ldots, d\theta_j^{n_j}, \rho_j) \mathcal{L}(dP_j) = \int_{\mathcal{P}_j} \mathcal{L}(\rho_j) \mathcal{L}(d\theta_j^*, \ldots, d\theta_j^{n_j}|K_j) \mathcal{L}(dP).
$$

Then, using the transformation $t = t(j, h)$ defined in formula (10) in the paper, we can use the law of the vector of length $T$ originated by the $d$ different $l$-clusterings, which in turn is an i.i.d. sample from $P$, to write

$$
\left( \prod_{j=1}^d \mathcal{L}(\rho_j) \right) \int_{\mathcal{P}} \mathcal{L}(d\theta_1^*, \ldots, d\theta_T^*) | P \mathcal{L}(dP) = \left( \prod_{j=1}^d \mathcal{L}(\rho_j) \right) \mathcal{L}(\eta|T) \prod_{m=1}^M P_0(d\psi_m),
$$

where $\eta$ is the partition of the vector $(\theta_1^*, \ldots, d\theta_T^*)$ into $M$ groups of size $d = (d_1, \ldots, d_M)$ characterised by the vector of unique values $(\psi_1, \ldots, \psi_M)$. We can now apply Proposition
1 to each random partition \((\rho_1, \ldots, \rho_d)\) and \(\eta\), yielding

\[
\mathcal{L}(d\theta) = \left( \prod_{j=1}^{d} \text{eppf}(e_j; \alpha) \right) \text{eppf}(d; \alpha_0) \prod_{m=1}^{M} P_0(d\psi_m).
\]

\[\square\]

### 1.4 Moments of a Hierarchical NormCRM

**Proof of equation** (15):

**Proof.** For each measurable set \(A \subset \Theta\), the Laplace transform of the unnormalized completely random measure is

\[
L_A(u) = \mathbb{E} \left( e^{-u\mu_1(A)} \right) = \exp \left\{ - \int_{R^+ \times A} \kappa \left( 1 - e^{-us} \right) \nu(ds, d\theta) \right\} = \exp \left\{ - \kappa P(A) \int_{R^+} (1 - e^{us})\alpha(s)ds \right\},
\]

that is, \(L_A(u) = e^{-P(A)\phi(u)}\), with \(\phi(u)\) as defined in Section 2.1 in the main paper. Moreover, we recall that the \(n\)th derivative of the Laplace transform is given by

\[
\mathbb{E} \left( \mu_1(A)^n e^{-u\mu_1(A)} \right) = (-1)^n \frac{d}{du^n} L_A(u) = (-1)^n \frac{d}{du^n} e^{-P(A)\phi(u)}.
\]

So applying the Faà di Bruno formula (see for instance Johnson 2002) we obtain

\[
\mathbb{E} \left( \mu_1(A)^n e^{-u\mu_1(A)} \right) = n! \sum_{k=1}^{n} \frac{1}{k!} e^{-P(A)\phi(u)} \sum_{\epsilon_1 + \cdots + \epsilon_k = n} P(A)^k \prod_{j=1}^{k} \frac{c_{\epsilon_j}(u)}{\epsilon_j!},
\]

where the function \(c_m(u)\) has also been defined in Section 2.1. If \(A \subset \Theta\), let \(\bar{A} = \Theta \setminus A\) be its complementary, we can compute:

\[
\mathbb{E} \left( P_1(A)^n \right) = \mathbb{E} \left( \frac{\mu_1(A)^n}{T^n} \right) = \mathbb{E} \left( \int_0^\infty \frac{u^{n-1}}{\Gamma(n)} e^{-uT} \mu_1(A)^n du \right)
\]

\[
= \mathbb{E} \left( \int_0^\infty \frac{u^{n-1}}{\Gamma(n)} e^{-u\mu_1(\Theta)} \mu_1(A)^n du \right)
\]

\[
= \int_0^\infty \frac{u^{n-1}}{\Gamma(n)} \mathbb{E} \left( e^{-u\mu_1(\bar{A})} \right) \mathbb{E} \left( e^{-u\mu_1(A)} \mu_1(A)^n \right) du
\]

\[
= \int_0^\infty \frac{u^{n-1}}{\Gamma(n)} L_{\bar{A}}(u)(-1)^n \frac{d}{du^n} L_A(u) du.
\]
By formulas (1) and (3) we obtain
\[ E(P_1(A)^n) = \sum_{k=1}^{n} P(A)^k \frac{1}{k!} \sum_{e_1 + \ldots + e_k = n} \left( \begin{array}{c} n \\ e_1, \ldots, e_k \end{array} \right) \int_0^\infty \frac{u^{n-1}}{\Gamma(n)} e^{-\phi(u)} \prod_{j=1}^{k} c_{e_j}(u) \, du \]
\[ = \sum_{k=1}^{n} P(A)^k \frac{1}{k!} \sum_{e_1 + \ldots + e_k = n} \left( \begin{array}{c} n \\ e_1, \ldots, e_k \end{array} \right) \text{eppf}(e_1, \ldots, e_k) \]
\[ = \sum_{k=1}^{n} P(A)^k \mathbb{P}(K_n = k) \]
\[ = E(P(A)^{K_n}). \]

**Proof of equation (16):**

**Proof.** The result follow by mimic the proof of Equation (15), just observing that since \( A \) and \( B \) are disjoint then \( \mu_1(A) \) and \( \mu_1(B) \) are independent random variables.

Consider now \( P_1 \sim \text{NormCRM}(P, \alpha) \) and let \( n = 2 \) so that \( K_2 \), the number of cluster in a sample of size \( n = 2 \), takes values in \( \{1, 2\} \). We denote with \( \eta_1 \) the probability that \( K_2 = 1 \). It easy to realize that \( \eta_1 \) is the probability that two observations from a NormCRM with Levy intensity \( \alpha \) will share the same cluster, i.e. \( \eta_1 := \text{eppf}(2; \alpha) \). Therefore, the second moment of a NormCRM is \( E(P_1(A)^2) = P(A)\eta_1 + P(A)^2(1 - \eta_1) \), and consequently

\[ \text{Var}(P_1(A)) = \eta_1 P(A)(1 - P(A)) \quad (4) \]

Let \( A \) and \( B \) be two disjoint measurable subsets of \( \Theta \) so that, for \( n_1 = n_2 = 1 \),

\[ E(P_1(A)P_1(B)) = P(A)P(B)\text{eppf}(1, 1; \alpha) = P(A)P(B)(1 - \eta_1). \]

The general case when \( A \) and \( B \) are not disjoint follows easily as

\[ E(P_1(A)P_1(B)) = E((P_1(A \cap B))^2) + E(P_1(A \setminus B)P_1(A \cap B)) + E(P_1(B \setminus A)P_1(A \cap B)) + E(P_1(A \setminus B)P_1(B \setminus A)). \]

Applying the result for the disjoint sets given above we find

\[ E(P_1(A)P_1(B)) = \eta_1 P(A \cap B) + (1 - \eta_1)P(A)P(B), \]
and consequently
\[ \text{Cov}(P_1(A), P_1(B)) = \eta_1 (P(A \cap B) - P(A)P(B)), \]
for any \( A, B \in \mathcal{B}(\Theta) \).

We are now ready to compute the moments of a hierarchical NormCRM. Consider first \( P_1|P \sim \text{NormCRM}(\alpha, P) \), and \( P \sim \text{NormCRM}(\alpha_0, P_0) \), and let \( A \) and \( B \) be a measurable sets. Indicate with \( \eta_0 = \text{eppf}(2; \alpha_0) \) and \( \eta_1 = \text{eppf}(2; \alpha) \), we have
\[
\text{Var}(P_1(A)) = \mathbb{E}(P_1(A)^2) - \mathbb{E}^2(P_1(A)) = \mathbb{E}(P(A)^k) - P_0(A)^2 \\
= \mathbb{E}(P(A))\mathbb{P}(K_2 = 1) + \mathbb{E}(P(A)^2)\mathbb{P}(K_2 = 2) - P_0(A)^2 \\
= P_0(A)\eta_1 + \mathbb{E}(P^2(A))(1 - \eta_1) - P_0(A)^2 \\
= P_0(A)\eta_1 + (P_0(A)\eta_0 + P_0(A)^2(1 - \eta_0)) - P_0(A) \\
= P_0(A)(1 - P_0(A))(\eta_0 + \eta_1 - \eta_0\eta_1),
\]
while for the covariance of a measure in two groups \( A \) and \( B \), we have
\[
\text{Cov}(P_1(A), P_1(B)) = \mathbb{E}(P_1(A)P_1(B)) - \mathbb{E}(P_1(A))\mathbb{E}(P_1(B)) = \mathbb{E}(\mathbb{E}(P_1(A)P_1(B)|P)) - P_0(A)P_0(B) \\
= \mathbb{E}(\eta_1 P(A \cap B) + (1 - \eta_1)P(A)P(B)) - P_0(A)P_0(B) \\
= \eta_1 P_0(A \cap B) + (1 - \eta_1)\mathbb{E}(P(A)P(B)) - P_0(A)P_0(B) \tag{5} \\
= \eta_1 P_0(A \cap B) + (1 - \eta_1) [\eta_0 P_0(A \cap B) + (1 - \eta_0)P_0(A)P_0(B)] - P_0(A)P_0(B) \\
= (\eta_0 + \eta_1 - \eta_0\eta_1) [P_0(A \cap B) - P_0(A)P_0(B)],
\]
and for the covariance of two measures in two groups we obtain
\[
\text{Cov}(P_1(A), P_2(B)) = \mathbb{E}(P_1(A)P_2(B)) - P_0(A)P_0(B) \\
= \mathbb{E}(\mathbb{E}(P_1(A)P_2(B)|P)) - P_0(A)P_0(B) \\
= \mathbb{E}(\mathbb{E}(P_1(A)|P)\mathbb{E}(P_2(B)|P)) - P_0(A)P_0(B) \tag{6} \\
= \mathbb{E}(P(A)P(B)) - P_0(A)P_0(B) \\
= \eta_0 P_0(A \cap B) + (1 - \eta_0)P_0(A)P_0(B) - P_0(A)P_0(B) \\
= \eta_0 (P_0(A \cap B) - P_0(A)P_0(B)).
\]
Hence, the covariance of the measure across groups of observations is governed only by \( \alpha_0 \).

**Mixed moments of a Hierarchical NormCRM**

Let \( P_1, P_2|P \overset{\text{iid}}{\sim} \text{NormCRM}(\alpha, P) \) and \( P \sim \text{NormCRM}(\alpha_0, P_0) \) be a sample of size \( n = n_1 + n_2 \) from a hierarchical NormCRM. When considering \( A \) and \( B \in \mathcal{B}(\Theta) \) such that
\[ A \cap B = \emptyset \text{ we have} \]
\[
\mathbb{E} [P_1(A)^{n_1} P_2(B)^{n_2}] = \mathbb{E} [\mathbb{E} [P_1(A)^{n_1} P_2(B)^{n_2} | P]] = \mathbb{E} [\mathbb{E} [P_1(A)^{n_1} | P] \mathbb{E} [P_2(B)^{n_2} | P]] \\
= \mathbb{E}_P [\mathbb{E}_{K_1} [P(A)^{k_1}] \mathbb{E}_{K_2} [P(B)^{k_2}]] = \mathbb{E}_P [\mathbb{E}_{K_1, K_2} [P(A)^{k_1} P(B)^{k_2}]] \\
= \mathbb{E}_{K_1, K_2} [\mathbb{E}_P [P(A)^{k_1} P(B)^{k_2}]] = \mathbb{E}_{K_1, K_2} \left( \sum_{t_1=1}^{k_1} \sum_{t_2=1}^{k_2} P_0(A)^{t_1} P_0(B)^{t_2} g_0(k_1, k_2) \right),
\]
where we used the fact that \( K_1 \perp K_2 \) and swapped the expectations in the last step. The function \( g_0(k_1, k_2) \) is defined in equation (17) in the main paper. Considering now the case when \( A = B, \) and \( T = K_1 + K_2, \) we obtain:
\[
\mathbb{E} [P_1(A)^{n_1} P_2(A)^{n_2}] = \mathbb{E}_T [\mathbb{E}_P [P(A)^T]] = \\
\sum_{t=d}^{n_1+n_2} \mathbb{E}_P [P(A)^t] \mathbb{P}(T = t) = \sum_{t=d}^{n_1+n_2} \mathbb{E}_M [P(A)^M | T = t] \mathbb{P}(T = t) = \mathbb{E} [P_0(A)^M].
\]

1.5 Coskewness

Let \( X \) and \( Y \) be two random variable with finite third moments and such that \( \mathbb{E}(X) = \mu_X, \) \( \mathbb{E}(Y) = \mu_Y, \) and \( \text{Var}(X) = \sigma_X^2, \) \( \text{Var}(Y) = \sigma_Y^2. \) Furthermore, let the skewness of \( X \) and \( Y \) be defined as \( \text{Sk}(X) = \frac{\mathbb{E}((X-\mu_X)^3)}{\sigma_X^3} \) and \( \text{Sk}(Y) = \frac{\mathbb{E}((X-\mu_Y)^3)}{\sigma_Y^3}, \) respectively. The coskewness of \( X \) over \( Y \) is defined as:
\[
\text{CoSk}(X, Y) = \frac{\mathbb{E}((X-\mu_X)^2(\sigma_Y - \mu_Y))}{\sigma_X^2 \sigma_Y} = \frac{\mathbb{E}(X^2 Y) - 2 \mu_X \mathbb{E}(XY) - \mathbb{E}(X^2) \mu_Y + 2 \mu_X^2 \mu_Y}{\sigma_X^2 \sigma_Y} \\
= \frac{\text{Cov}(X^2, Y) - 2 \mu_X \text{Cov}(X, Y)}{\sigma_X^2 \sigma_Y}. \tag{7}
\]

Clearly, CoSk is not symmetric and CoSk(Y, X) is defined analogously. Furthermore,
\[
\sigma_X \text{CoSk}(X, Y) + \sigma_Y \text{CoSk}(Y, X) = \frac{\sigma_X^3 \text{Sk}(X + Y) - \sigma_X^3 \text{Sk}(X) - \sigma_Y^3 \text{Sk}(Y)}{3 \sigma_X \sigma_Y}.
\]

We note that the coskewness can be different from zero even when \( X \) and \( Y \) are symmetric, since it is governed by the random variable \( X + Y. \) This mixed third moment is a measure of the joint asymmetry of \( X \) and \( Y \) (see for instance Friend and Westerfield 1980; Fang and Lai 1997). It is mainly used in the field of Econometrics, with the aim of studying the risk associated with financial portfolios. In our setting, if we let \( X = P_1(A) \) and \( Y = P_2(A), \) so
that $\mu_X = \mu_Y = P_0(A) = P_0$, then $\text{Cov}(P_1(A), P_2(A)) = \eta_0 P_0(1 - P_0)$, from equation (6) in the paper, and
\[
\text{CoSk}(P_1(A), P_2(A)) = \frac{\mathbb{E} \left( P^{M_{2,1}}_0 \right) - 2 \eta_0 P^2_0(1 - P_0) - P_0 \mathbb{E} \left( P^{M_2}_0 \right)}{(\mathbb{E}(P^{M_2}_0) - \mathbb{E}(P_0)^2)^{3/2}},
\]
where $M_{2,1}$ is the number of clusters in the natural clustering rule of a hierarchical Norm-CRM with two groups, with 2 individuals in the first group and 1 individual in the second group, and $M_2$ is the number of clusters in the natural clustering rule of a NormCRM (i.e., with one group) with 2 individuals. Since $\text{Var}(P_1(A)) = \text{Var}(P_2(A)) = \mathbb{E}(P^{M_2}_0) - \mathbb{E}(P_0)^2 = P_0(A)(1 - P_0(A))(\eta_0 + \eta_1 - \eta_0 \eta_1)$, we can re-write
\[
\text{CoSk}(P_1(A), P_2(A)) = \frac{\mathbb{E} \left( P^{M_{2,1}}_0 \right) - 2 \eta_0 P^2_0(1 - P_0) - P_0 \mathbb{E} \left( P^{M_2}_0 \right)}{(P_0(1 - P_0)(\eta_0 + \eta_1 - \eta_0 \eta_1))^{3/2}}.
\] (8)

2 MCMC Algorithm

In Section 3.2 of the main text, we introduced the general MCMC sampling scheme targeting the posterior distribution of $(U, \psi, \rho, \eta)$ under model (19). Here we provide more details on the algorithm, with a detailed formulation for the case of the HNGG process. As mentioned in the paper, conditionally on the vector of auxiliary variables $U$, the law of our joint model is
\[
\mathcal{L}(Y_1, \ldots, Y_d|\rho, \eta, \psi, U) \mathcal{L}(\rho_1, \ldots, \rho_d|U_1, \ldots, U_d) \mathcal{L}(\eta|U_0) \prod_{m=1}^{M} P_0(d\psi_m)
\]
\[
= \prod_{m=1}^{M} \prod_{(j,i) \in \mathcal{I}_m^{(\rho, \eta)}} f(y_{ji}|\psi_m) \prod_{j=1}^{d} \text{eppf}(e_{j1}, \ldots, e_{jk_j}; \kappa, \sigma, u_j) \text{eppf}(d_1, \ldots, d_M; \kappa_0, \sigma_0, u_0) \prod_{m=1}^{M} P_0(d\psi_m),
\]
where $\mathcal{I}_m^{(\rho, \eta)} = \{(j, i) : j \in \{1, \ldots, d\}, i \in C_{jh}, t(j, h) \in D_m \}$, for $m = 1, \ldots, M$. As reported in the paper, $e_j = (e_{j1}, \ldots, e_{jk_j})$ are the cluster sizes of partition $\rho_j$, for $j = 1, \ldots, d$, while $d = (d_1, \ldots, d_M)$ the ones of partition $\eta$.

We show in this section how the prior allocation probabilities of a generalise Chinese restaurant franchise process are derived (formulas (21) and (22) in the paper), and give their expression for the HNGG case. Furthermore, we show the sampling steps required for the update of $(U, \psi, \rho, \eta)$, which are the main parameters of interest in this study. We describe how the posterior sampling of the random partitions $(\rho, \eta)$ can be performed in both conjugate and non-conjugate models. Finally, we give details on the posterior update of the hyperparameters of the HNGG($\kappa, \sigma, \kappa_0, \sigma_0$) process.
2.1 Derivation of formulas (21) and (22)

Without loss of generality, consider the $j$-th random partition $\rho_j$ introduced in model (19) of the paper. From Pitman (1996) and Ishwaran and James (2003) it is known that the distribution of $\rho_j = \{C_{j1}, \ldots, C_{jk_j}\}$ of the indices $\{1, \ldots, n_j\}$ for the case of a NGG($\kappa, \sigma$), is the following exchangeable partition probability function (eppf):

$$
\pi(\rho_j) = \text{eppf}(e_{j1}, \ldots, e_{j k_j}; \kappa, \sigma) = \int_0^{+\infty} \frac{u_j^{n_j-1}}{\Gamma(n_j)} e^{-\phi(u_j)} \prod_{l=1}^{k_j} c_{e_{jl}}(u_j) du_j, \quad (9)
$$

where the Laplace exponent $\phi(u_j)$ and the cumulants $c_m(u_j)$ are defined as:

$$
\phi(u_j) = \frac{\kappa}{\sigma} ((u_j + 1)^\sigma - 1), \quad (10)
$$

$$
c_m(u_j) = \frac{\kappa}{(u_j + 1)^{m-\sigma}} \frac{\Gamma(m-\sigma)}{\Gamma(1-\sigma)}. \quad (11)
$$

Let now $\Gamma_j \sim \text{gamma}(n_j, 1)$ which is independent from $T_j$, the total mass of the NGG measure as described in Section 2.1 of the paper, and set $U_j := \Gamma_j / T_j$. For the NGG process case, for any $n_j \geq 1$, the marginal density function of $U_j$ is given by

$$
f_{U_j}(u_j; n_j, \kappa, \sigma) = \frac{u_j^{n_j-1}}{\Gamma(n_j)} \frac{d^n_j}{d u_j^n} e^{-\phi(u_j)}. \quad (12)
$$

By disintegration of formula (9) (see also James et al. 2009) we can see how the conditional distribution of $U_j$, given the random partition $\rho_j$, is

$$
f_{U_j|\rho_j}(u_j) \propto \frac{u_j^{n_j-1}}{\Gamma(n_j)} \prod_{l=1}^{k_j} c_{e_{jl}}(u_j).
$$

We denote with $\text{eppf}(e_{j1}, \ldots, e_{j k_j}; \kappa, \sigma)$ the integrand of (9):

$$
\text{eppf}(e_{j1}, \ldots, e_{j k_j}; \kappa, \sigma) = \int_0^{+\infty} \text{eppf}(e_{j1}, \ldots, e_{j k_j}; \kappa, \sigma, u_j) du_j. \quad (12)
$$

Then,

$$
f_{U_j|\rho_j}(u_j) = \frac{u_j^{n_j-1}}{\Gamma(n_j)} \prod_{l=1}^{k_j} c_{e_{jl}}(u_j) \int_0^{+\infty} \frac{u_j^{n_j-1}}{\Gamma(n_j)} \prod_{l=1}^{k_j} c_{e_{jl}}(u_j) du_j = \frac{\text{eppf}(e_{j1}, \ldots, e_{j k_j}; \kappa, \sigma, u_j)}{\text{eppf}(e_{j1}, \ldots, e_{j k_j}; \kappa, \sigma)}, u_j \geq 0.
$$
Consider now a new variable indexed as \((n_j + 1)\). The predictive distribution of the cluster allocation of the new variable can be written as an integral with respect to \(f_{U_j|\rho_j}(u_j)du_j\). 
To this aim, let \(l_{new} \in \{1, \ldots, k_j + 1\}\) be the cluster label of the \((n_j + 1)\)-th observation, and denote with \(\rho_{j}^{new} = \{C_{j1}^{new}, \ldots, C_{jk_j}^{new}\}\) the resulting new partition. Thus, \(\rho_{j}^{new} = (\rho_j \setminus \tilde{C}_{jl_{new}}) \cup (\tilde{C}_{jl_{new}} \cup \{p + 1\})\), where \(\tilde{C}_{jk_j + 1} = \emptyset\), and \(\tilde{C}_{jl} = C_{jl}\), for \(l = 1, \ldots, k_j\). Thus, \(k_{j}^{new}\) can be \(k_j\) or \(k_j + 1\). Moreover:

\[
\pi(\rho_{j}^{new}) = \operatorname{eppf}(e_{j1}^{new}, \ldots, e_{jk_j}^{new}; \kappa, \sigma) = \int_{0}^{+\infty} \frac{u_{j}^{n_j}}{\Gamma(n_j + 1)} e^{-\phi(u_j)} \prod_{l=1}^{k_{j}^{new}} \varphi_{jl}^{new}(u_j)du_j,
\]

\[
= \int_{0}^{+\infty} \operatorname{eppf}(e_{j1}^{new}, \ldots, e_{jk_j}^{new}; \kappa, \sigma) \operatorname{eppf}(e_{j1}, \ldots, e_{jk_j}; \kappa, \sigma, u_j)du_j,
\]

\[
= \int_{0}^{+\infty} \frac{\operatorname{eppf}(e_{j1}^{new}, \ldots, e_{jk_j}^{new}; \kappa, \sigma)}{\operatorname{eppf}(e_{j1}, \ldots, e_{jk_j}; \kappa, \sigma)} \operatorname{eppf}(e_{j1}, \ldots, e_{jk_j}; \kappa, \sigma, u_j)\operatorname{eppf}(e_{j1}, \ldots, e_{jk_j}; \kappa, \sigma, u_j)du_j,
\]

We introduce an abuse of notation by indicating with \((n_j + 1) \in C_{jl_{new}}\), for \(l_{new} = 1, \ldots, k_j\), the event that the new variable is allocated to the \(l_{new}\)-th cluster in the \(j\)-th group, implying that \(\rho_{j}^{new}\) is obtained by letting \(k_{j}^{new} = k_j\), \(C_{jl_{new}}^{new} = C_{jl_{new}} \cup \{n_j + 1\}\), and by leaving the other clusters unchanged. Similarly, with \((n_j + 1) \in C_{jk_j}^{new}\), we indicate the event that the new variable is assigned to a new cluster, such that \(l_{new} = k_j + 1, k_{j}^{new} = k_j + 1\) and \(C_{jl_{new}}^{new} = \{n_j + 1\}\). For \(l_{new} = 1, \ldots, k_j + 1\), we have the predictive probabilities

\[
\mathbb{P}((n_j + 1) \in C_{jl_{new}}|\rho_j) = \frac{\operatorname{eppf}(e_{j1}^{new}, \ldots, e_{jk_j}^{new}; \kappa, \sigma)}{\operatorname{eppf}(e_{j1}, \ldots, e_{jk_j}; \kappa, \sigma)}
\]

\[
= \int_{0}^{+\infty} \frac{\operatorname{eppf}(e_{j1}^{new}, \ldots, e_{jk_j}^{new}; \kappa, \sigma)}{\operatorname{eppf}(e_{j1}, \ldots, e_{jk_j}; \kappa, \sigma, u_j)} \operatorname{eppf}(e_{j1}, \ldots, e_{jk_j}; \kappa, \sigma, u_j)du_j,
\]

\[
= \int_{0}^{+\infty} \operatorname{eppf}(e_{j1}^{new}, \ldots, e_{jk_j}^{new}; \kappa, \sigma) \operatorname{eppf}(e_{j1}, \ldots, e_{jk_j}; \kappa, \sigma, u_j)\operatorname{eppf}(e_{j1}, \ldots, e_{jk_j}; \kappa, \sigma, u_j)du_j,
\]

finally we have that

\[
\mathbb{P}((n_j + 1) \in C_{jl_{new}}|\rho_j, U_j = u_j) \propto \frac{\operatorname{eppf}(e_{j1}^{new}, \ldots, e_{jk_j}^{new}; \kappa, \sigma, u_j)}{\operatorname{eppf}(e_{j1}, \ldots, e_{jk_j}; \kappa, \sigma, u_j)}.
\]

from which the expression the formulas (21) is easily recovered.

An analogous argument holds for the case of the partition \(\eta_j\), whose marginal law is:

\[
\pi(\eta) = \operatorname{eppf}(d_1, \ldots, d_M; \kappa_0, \sigma_0) = \int_{0}^{+\infty} \frac{u_0^{T-1}}{\Gamma(T)} e^{-\phi(u_0)} \prod_{m=1}^{M} c_{dm}(u_0)du_0,
\]
where \( \mathbf{d} = (d_1, \ldots, d_M) \) is the vector of cluster sizes, \( M \) is the number of clusters and \( \{1, \ldots, T\} \) the set of indices on which the partition is induced. Proceeding as for the case of \( \rho_j \), we can show how formulas (22) in the paper are derived.

### 2.2 Updates of \( U \) and \( \psi \):

From the expression of \( \text{eppf}(\cdot; u_0, \alpha_0) \) and \( \text{eppf}(\cdot; u_j, \alpha) \), as well as the centering measure \( P_0 \), we obtain the following full-conditionals

\[
p(U_j | \rho_j, \alpha) \propto u_j^{n_j - 1} e^{-\phi(u_j)} \prod_{h=1}^{k_j} c_{e_{jh}}(u_j), \quad j = 1, \ldots, d
\]

\[
p(U_0 | \eta, \alpha_0, T) \propto u_0^{T-1} e^{-\phi(u_0)} \prod_{m=1}^{M} c_{d_m}(u_0), \quad (13)
\]

\[
p(\psi_m | Y, \rho, \eta) \propto \prod_{(j,i) \in \mathcal{I}_{[0]}^{(\rho, \eta)}} f(y_{ji} | \psi_m) P_0(d\psi_m), \quad m = 1, \ldots, M,
\]

with the quantities \( \phi(u) \) and \( c_l(u) \) defined as in (10) in the main text. These quantities are often known up to a normalizing constant, making necessary a series of Metropolis-Hastings steps. Here, we use an adaptive MCMC scheme (see Andrieu and Thoms 2008; Griffin and Stephens 2013, for more details). In the case of HNGG(\( \kappa, \sigma, \kappa_0, \sigma_0 \)):

\[
p(U_j | \rho_j, \alpha) \propto u_j^{n_j - 1} e^{-\frac{\mathbb{H}}{\mathbb{E}}((u_j+1)^{\gamma}-1)}(u_j+1)^{k_j \sigma - n_j}, \quad j = 1, \ldots, d
\]

\[
p(U_0 | \eta, \alpha_0, T) \propto u_0^{T-1} e^{-\frac{\mathbb{H}}{\mathbb{E}}((u_0+1)^{\gamma}-1)}(u_0+1)^{M \sigma_0 - T}.
\]

Further sharing of information among the hierarchies of the model can be achieved by imposing a prior distribution over the parameters \( \kappa, \sigma, \kappa_0, \sigma_0 \).

### 2.3 Updates of \( (\rho, \eta) \) (conjugate models):

Consider now removing the \( i \)-th customer from the \( j \)-th restaurant, and compute the corresponding allocation probabilities, indicating with the superscript \( -ji \) the variables that are affected in the removal process. Conditionally on \( Y \), \( (\rho^{-ji}, \eta^{-ji}) \) and the latent variable \( U_j \), the probability of assigning the \( i \)-th customer to the \( h \)-th table of the \( j \)-th restaurant, where the \( m \)-th dish is served (i.e., \( t(j, h) \in D_{m}^{-ji} \)), is

\[
P(i \in C_{jhi}^{-ji}, t(j, h) \in D_{m}^{-ji} | Y, \rho^{-ji}, \eta^{-ji}, U_j = u_j) dy_{ij}
\]

\[
\propto P(Y_{ji} = dy_{ij} | Y^{-ji}, i \in C_{jhi}^{-ji}, \rho^{-ji}, \eta^{-ji}) P(i \in C_{jhi}^{-ji}, t(j, h) \in D_{m}^{-ji} | \rho^{-ji}, \eta^{-ji}, U_j = u_j),
\]

12
for $h = 1, \ldots, k^{-ji}$, $m$ is the value such that $t(j, h) \in D^{-ji}_m$, and for $h = (k^{-ji} + 1)$, $m = 1, \ldots, M^{-ji} + 1$. By recalling the definition of the set of indices $\mathcal{I}_m^{(\rho, \eta)}$ of the natural clustering rule used in the paper (19), we recognize the first term as

$$
\mathbb{P}(Y_{ji} = dy_{ij}|Y^{-ji}, i \in C_{jh}^{-ji}, \rho^{-ji}, \eta^{-ji}) \propto 
\int \Theta \left( \sum_{ij} y_{ij} \cup \mathbf{y}_{\mathcal{I}_m^{(\rho^{-ji}, \eta^{-ji})}} | \psi \right) P_0(d\psi) \int \Theta \left( \sum_{ij} y_{ij} \cup \mathbf{y}_{\mathcal{I}_m^{(\rho^{-ji}, \eta^{-ji})}} | \psi \right) P_0(d\psi)
$$

where $\mathcal{M}(y_{ji}|\mathbf{y}_{\mathcal{I}_m})$ is the predictive density of a parametric Bayesian model where the sampling model is $f(y|\theta)$, the prior is $P_0$, and the observations are all the data with index in cluster $\mathcal{I}_m$, with proviso that $\mathcal{I}_{M^{-ji}+1}$ is the empty set. Hence:

$$
\mathbb{P}(i \in C_{jh}^{-ji}, t(j, h) \in D^{-ji}_m|Y, \rho^{-ji}, \eta^{-ji}, U_j = u_j) 
\propto \mathcal{M}(y_{ji}|\mathbf{y}_{\mathcal{I}_m^{(\rho^{-ji}, \eta^{-ji})}}) \mathbb{P}(i \in C_{jh}^{-ji}, t(j, h) \in D^{-ji}_m|\rho^{-ji}, \eta^{-ji}, U_j = u_j). \quad (14)
$$

In formula (14), when $h = 1, \ldots, k^{-ji}$, then $m = t(j, h)$, i.e. $C_{jh}^{-ji}$ is an already occupied table where the $m$-th dish is served. Moreover, when $h = (k^{-ji} + 1)$, then $m = 1, \ldots, M^{-ji} + 1$, i.e. when a new table is occupied one of the already served dishes or a new one can be chosen. Finally, probability $\mathbb{P}(i \in C_{jh}^{-ji}, t(j, h) \in D^{-ji}_m|\rho^{-ji}, \eta^{-ji}, U_j = u_j)$ is reported at the leaves of the tree in Figure 2.

The updating proceeds by re-allocating $C_{jh}$ to a cluster of tables. To this end, consider temporarily removing the table $t = t(j, h)$ from its cluster of tables, and hence all the indices in $C_{jh}$, and compute the probability of assigning it to an existing/new cluster of tables. We indicate with the superscript $-t$ the variables affected in the removal process. More formally, conditionally to $Y$, $(\rho^{-t}, \eta^{-t})$, and the latent variable $U_0$, the probability of assigning the $t$-th table to the $m$-th cluster is

$$
\mathbb{P}(t \in D^{-t}_m|Y, \rho^{-t}, \eta^{-t}, U_0 = u_0)dy_t 
\propto \mathbb{P}(Y_t = dy_t|Y^{-t}, t \in D^{-t}_m, \rho^{-t}, \eta^{-t})\mathbb{P}(t \in D^{-t}_m|\rho^{-t}, \eta^{-t}, U_0 = u_0),
$$

for $m = 1, \ldots, M^{-t} + 1$, where $(M^{-t} + 1)$ indicates the new cluster of tables. Here, $Y_t := \{Y_{ij}: i \in C_{jh}, t = t(j, h)\}$ is the set of all the observations (customers) in cluster $C_{jh}$ (seated at table $t$), and $\mathcal{M}(y_t|\mathbf{y}_{\mathcal{I}_m})$ is the joint predictive density of $e_{jh}$ observations from a parametric Bayesian model where the sampling model is $f(y|\theta)$, the prior is $P_0$, and the observations are all the data with index in cluster $\mathcal{I}_m$. Probability $\mathbb{P}(t \in D^{-t}_m|\rho^{-t}, \eta^{-t}, U_0 = u_0)$ has been introduced in (22) in the main text, and it is the predictive probability
prescribed by the generalised Chinese restaurant process of the table indices. On the other hand

\[
\mathbb{P}(Y_t = d_{yt}, t \in D_m^{-t}, \rho^{-jt}, \eta^{-jt}) = \\
\frac{\int_{\Theta} f(y_t \cup Y_{I_m^{\rho^{-jt}, \eta^{-jt}}}|\psi)P_0(d\psi)}{\int_{\Theta} f(Y_{I_m^{\rho^{-jt}, \eta^{-jt}}}|\psi)P_0(d\psi)} dy_t = \mathcal{M}\left(y_t \bigg| Y_{I_m^{\rho^{-jt}, \eta^{-jt}}}, \psi\right) dy_t,
\]

and, similarly to the updating step of \(\rho\), we obtain

\[
\mathbb{P}(t \in D_m^{-t}|Y, \rho^{-t}, \eta^{-t}) \propto \mathcal{M}\left(y_t \bigg| Y_{I_m^{\rho^{-t}, \eta^{-t}}}, \psi\right) \mathbb{P}(t \in D_m^{-t}|\rho^{-t}, \eta^{-t}, U_0 = u_0),
\]

for \(m = 1, \ldots, M^{-t} + 1\), where \((M^{-t} + 1)\) indicates the new cluster of tables.

This step of the algorithm requires knowledge of the conditional distributions \(\mathcal{M}\), which is often not available, for instance when the centering distribution \(P_0\) is not conjugate with respect to the sampling model \(f(\cdot|\theta)\). For such cases, we extend our algorithm in the next part of the section, following the idea of the popular Algorithm 8 of Neal (2000), jointly with the one in the re-use algorithm of Favaro and Teh (2013).

### 2.4 Updates of (\(\rho, \eta\)) (non-conjugate models):

Consider augmenting the state space with \(N_{aux}\) new random variables associated with a set of \(N_{aux}\) empty clusters, s.t. \((\psi^{e}_1, \ldots, \psi^{e}_{N_{aux}}) \sim \mathcal{P}_0\). The role of these auxiliary variables (and clusters) is to approximate the probability of sampling a new label in a Monte Carlo fashion, hence avoiding the computation of the integrals involved in the definition of \(\mathcal{M}\). However, a drawback of such algorithm is the waste of i.i.d. samples from \(P_0\), since the \(N_{aux}\) auxiliary variables have to be renovated at every iteration of the Gibbs sampler, in order for the method to be invariant. To overcome this issue, Favaro and Teh (2013) proposed a re-use algorithm that bypasses the waste of i.i.d. samples.

Referring to the notation used earlier in Section 2.3, consider equation (14):

\[
\mathbb{P}(i \in C^{-ji}_j|Y, \rho^{-ji}, \eta^{-ji}, U_j = u_j) \propto \mathcal{M}\left(y_{ji} \bigg| Y_{I_m^{\rho^{-ji}, \eta^{-ji}}}, \psi\right) \mathbb{P}(i \in C^{-ji}_j|\rho^{-ji}, \eta^{-ji}, U_j = u_j),
\]

which is used in the reallocation process of the \(ji\)-th customer to the \(h\)-th table in the \(j\)-th restaurant. In order to avoid the computation of the predictive density \(\mathcal{M}\) when the model is non-conjugate, we can augment the state space by avoiding the marginalisation of the unique values \(\psi\), as it was instead done in the previous section. In this way, for
$h = 1, \ldots, k_j$, we can simply disintegrate the integral $\mathcal{M}$ by conditioning with respect to $\psi_m = \theta^*_j$, obtaining:

$$
\mathbb{P}(i \in C_{jh}^j | Y, \rho^{-ji}, \eta^{-ji}, U_j = u_j, \psi_m = \theta^*_j) \propto f(y_{ji} | \psi_m) \mathbb{P}(i \in C_{jh}^j | \rho^{-ji}, \eta^{-ji}, U_j = u_j).
$$

For a new cluster $h = k_j + 1$ such that $\theta^*_{jk_j+1} = \psi_{M+1}$, i.e. a new dish from the menu is served, we observe that:

$$
\mathcal{M}(y_{ji}) = \int f(y_{jn+1} | \psi_{M+1}) P_0(d \psi_{M+1}) = \mathbb{E}\left( \frac{1}{N_{aux}} \sum_{c=1}^{N_{aux}} f(y_{jn+1} | \psi^e_c) \right)
$$

where $(\psi^e_1, \ldots, \psi^e_{N_{aux}}) \overset{iid}{\sim} P_0$ are therefore used to disintegrate this integral. Now, by augmenting the state space with $\psi_m$, for $m = 1, \ldots, M$, and $(\psi^e_1, \ldots, \psi^e_{N_{aux}})$ we obtain the following full conditionals:

$$
\begin{align*}
\mathbb{P}(i \in C_{jh}^j | Y, \rho^{-ji}, \eta^{-ji}, \psi^{-ji}, U, \psi^e) \\
\propto \left\{ \begin{array}{ll}
P_{j}^{(to)} f(y_{ji} | \psi_m) & h = 1, \ldots, k_j^{-ji}, \ m = t(j, h) \\
P_{j}^{(tn)} f(y_{ji} | \psi_m) & h = k_j^{-ji} + 1, \ m = 1, \ldots, M^{-ji} \\
P_{j}^{(tn)} P_{m}^{(dn)} f(y_{ji} | \psi^e_m) / N_{aux} & h = k_j^{-ji} + 1, \ m = M^{-ji} + 1, \ldots, M^{-ji} + N_{aux}
\end{array} \right.
\end{align*}
$$

using the a priori allocation probabilities of Figure 2 in the main text. Similarly, for the update of $\eta$,

$$
\begin{align*}
\mathbb{P}(t \in D_{-t}^{-m} | Y, \rho^{-t}, \eta^{-t}, \psi, U_0, \psi^e) \\
\propto \left\{ \begin{array}{ll}
P_{m}^{(do)} f(y_t | \psi_m) & m = 1, \ldots, M^{-t} \\
P_{m}^{(dn)} f(y_t | \psi^e_m) / N_{aux} & m = M^{-t} + 1, \ldots, M^{-t} + N_{aux}
\end{array} \right.
\end{align*}
$$

To complete our MCMC for non-conjugate models, we also need to produce the full-conditional distributions of the variables $\psi = (\psi_1, \ldots, \psi_M)$ and $(\psi^e_1, \ldots, \psi^e_{N_{aux}})$. While the auxiliary variables are i.i.d from $P_0$ by construction, the full-conditionals for $\psi$ can be derived from formula (13).

### 2.5 Updates of $(\kappa, \sigma, \kappa_0, \sigma_0)$

It is easy to see that the full-conditional distributions for the hyperparameters $\kappa, \sigma, \kappa_0, \sigma_0$ can be derived from the expressions of $\text{eppf}(e_j; \kappa, \sigma, u_j)$ for $j = 1, \ldots, d$, and $\text{eppf}(d; \kappa_0, \sigma_0, u_0)$. 

15
The mass parameters $\kappa$ and $\kappa_0$ are typically assigned a gamma prior $\text{gamma}(a_\kappa, b_\kappa)$ and $\text{gamma}(a_{\kappa_0}, b_{\kappa_0})$, yielding conjugate full-conditionals:

$$p(\kappa | \sigma, \mathbf{u}, \mathbf{e}_1, \ldots, \mathbf{e}_d) = \text{gamma}(a_\kappa + T, b_\kappa + \sum_{j=1}^{d} \left((u_j + 1)^{\sigma} - 1\right)/\sigma),$$

$$p(\kappa_0 | \sigma_0, u_0, \mathbf{d}) = \text{gamma}(a_{\kappa_0} + M, b_{\kappa_0} + \left((u_0 + 1)^{\sigma_0} - 1\right)/\sigma_0).$$

As far as $\sigma$ and $\sigma_0$ are concerned, we assign to $\sigma$ and $\sigma_0$ prior distributions $\text{Beta}(a_\sigma, b_\sigma)$ and $\text{Beta}(a_{\sigma_0}, b_{\sigma_0})$, respectively. The full-conditionals result proportional to:

$$p(\sigma | \kappa, \mathbf{u}, \mathbf{e}_1, \ldots, \mathbf{e}_d) \propto \sigma^{a_\sigma-1}(1 - \sigma)^{b_\sigma-1} \prod_{j=1}^{d} e^{-\frac{u_j}{\sigma}((u_j+1)^{\sigma} - 1)} \prod_{l=1}^{k_j} \Gamma(e_{jl} - \sigma),$$

$$p(\sigma_0 | \kappa_0, u_0, \mathbf{d}) \propto \sigma_0^{a_{\sigma_0}-1}(1 - \sigma_0)^{b_{\sigma_0}-1} e^{-\frac{u_0}{\sigma_0}((u_0+1)^{\sigma_0} - 1)} \prod_{m=1}^{M} \Gamma(d_m - \sigma_0).$$

We resort to an adaptive Metropolis-Hastings algorithm for updating these parameters.

### 3 Simulation Study: $d = 2$

In this Section, we provide additional insights on the simulation study results presented in Section 4.1 of the main paper. Referring to the notation used in equations (21), (22), the prior assignment probabilities used for the MCMC updates in the specifica case of HNGG process are:

$$\mathbb{P}(i \in C^{-ji}_{jh} | \mathbf{p}^{-ji}, \mathbf{\eta}^{-ji}) = \begin{cases} P^{(to)}_{jh} = e^{-\frac{ji}{A_j}} \frac{\kappa_0}{A_j}, & h = 1, \ldots, k^{-ji}, \\ P^{(tn)}_{j} = \frac{\kappa_0(u_j + 1)^{\sigma_0}}{A_j}, & h = k^{-ji} + 1, \end{cases}$$

$$\mathbb{P}(t \in D^{-t}_m | \mathbf{p}^{-t}, \mathbf{\eta}^{-t}) = \begin{cases} P^{(do)}_{m} = \frac{d^{-t}_{m} - \sigma_0}{A_0}, & h = k^{-t}_m + 1, \ m = 1, \ldots, M^{-t}, \\ P^{(dn)}_{m} = \frac{\kappa_0(u_0 + 1)^{\sigma_0}}{A_0}, & h = k^{-t}_m + 1, \ m = M^{-t} + 1, \end{cases}$$

where $A_j = (n_j - k^{-ji}_j \sigma) + \kappa_0(u_j + 1)^{\sigma}$ and $A_0 = (T - M^{-t}\sigma_0) + \kappa_0(u_0 + 1)^{\sigma_0}$, and where $e^{-ji}$ and $d^{-t}$ represent the cardinality of the table and dish clusters in the configurations $(\mathbf{p}^{-ji}, \mathbf{\eta}^{-ji})$ and $(\mathbf{p}^{-t}, \mathbf{\eta}^{-t})$, respectively.
3.1 Moments of the number of clusters \( M \) a priori

In addition to the results of Section 2.2.1 (main text), we report first and second order moments of the a priori law of the number of clusters \( M \), for all the combinations of the hyperparameters \((\kappa, \kappa_0, \sigma, \sigma_0)\) shown so far (including the additional ones presented above). The prior distribution of \( M \) can be approximated via Monte Carlo averaging, hence providing the estimates of its mean and variance (see Table 1). When comparing these results with the LPML estimates in Table 1 in the manuscript and supplementary Table 5, we can observe that the best fit is achieved for higher number of clusters a priori. However, an opposite behaviour is observed when it comes to clustering, as it appears from comparing supplementary Table 1 with Table 2 and supplementary Table 6, where lower expected numbers of clusters a priori are associated with higher RI index values.

3.2 Mixed moments for the a priori random measures

It is of interest to study the distribution a priori of the random measures involved in model (19), based on the setting of the simulated data of Section 4.1 of the manuscript. In particular, under the specification thereby introduced, we have that the mean and variance parameters \((\mu_{ji}, \tau_{ji}^2)\) are assigned a HNGG process prior, where the indices vary in each group of observations \(j = 1, \ldots, d\), and for each data point in the \(j\)-th group \(i = 1, \ldots, n_j\). Given the measurable sets \(A\) and \(B\), we are interested in computing the mixed moments of the random measures \(P_1(A)\) and \(P_2(B)\) by using formulas (5) and (6) for the covariance terms, and equation (8) for the coskewness. We restrict the computation to the one-dimensional parameter \(\mu\), and use its marginal centering measure

\[
P_0(d\mu) = t_{2a_2} \left( d\mu \bigg| \frac{b_{r_2} k_0 + 1}{a_{r_2} k_0} \right),
\]

which is a location-scale \(t\)-Student distribution. We set the hyperparameters \((m_0, k_0, a_{r_2}, b_{r_2})\) to the values \((0.25, 0.62, 2.07, 0.66)\) used the simulations. We specify the sets \(A\) and \(B\) to one standard deviation from the expected values in each group. Namely, \(A = (\mu_1 - \sqrt{\sigma_1^2}, \mu_1 + \sqrt{\sigma_1^2})\) and \(B = (\mu_2 - \sqrt{\sigma_2^2}, \mu_2 + \sqrt{\sigma_2^2})\), where \((\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = (-0.6, 0.42, 0.9, 1.4)\). Also, \(A \cap B = (\mu_2 - \sqrt{\sigma_2^2}, \mu_1 + \sqrt{\sigma_1^2})\), focusing the attention on a small neighbourhood of zero. The values of the marginal probabilities are \(P_0(A) = 0.3309\), \(P_0(B) = 0.649\), and \(P_0(A \cap B) = 0.123\). Results on the covariances are reported in Tables 2 and 3. We point out that, according to formula (6), the values of \(\text{Cov}(P_1(A), P_2(B))\) only depend on the NormCRM governed by \(\alpha_0\), and hence by NGG(\(\kappa_0, \sigma_0\)). The values of the coskewness for the set \(A \cap B\) are reported in Table 4, following formula (8).
<table>
<thead>
<tr>
<th>( \kappa = \kappa_0 = 1 )</th>
<th>( \sigma_0 \approx 0 )</th>
<th>( \sigma_0 = 0.1 )</th>
<th>( \sigma_0 = 0.3 )</th>
<th>( \kappa = \kappa_0 = 1 )</th>
<th>( \sigma_0 \approx 0 )</th>
<th>( \sigma_0 = 0.1 )</th>
<th>( \sigma_0 = 0.3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma \approx 0 )</td>
<td>2.9973</td>
<td>3.2327</td>
<td>3.7975</td>
<td>( \sigma \approx 0 )</td>
<td>1.5009</td>
<td>1.7088</td>
<td>2.2384</td>
</tr>
<tr>
<td>( \sigma = 0.1 )</td>
<td>3.3840</td>
<td>3.6932</td>
<td>4.4619</td>
<td>( \sigma = 0.1 )</td>
<td>1.8077</td>
<td>2.0941</td>
<td>2.8664</td>
</tr>
<tr>
<td>( \sigma = 0.3 )</td>
<td>4.4251</td>
<td>4.9727</td>
<td>6.4492</td>
<td>( \sigma = 0.3 )</td>
<td>2.8531</td>
<td>3.5049</td>
<td>5.6153</td>
</tr>
<tr>
<td>( \kappa = \kappa_0 = 0.01 )</td>
<td>( \sigma_0 \approx 0 )</td>
<td>( \sigma_0 = 0.1 )</td>
<td>( \sigma_0 = 0.3 )</td>
<td>( \kappa = \kappa_0 = 0.01 )</td>
<td>( \sigma_0 \approx 0 )</td>
<td>( \sigma_0 = 0.1 )</td>
<td>( \sigma_0 = 0.3 )</td>
</tr>
<tr>
<td>( \sigma \approx 0 )</td>
<td>1.0219</td>
<td>1.0312</td>
<td>1.0517</td>
<td>( \sigma \approx 0 )</td>
<td>0.0216</td>
<td>0.0310</td>
<td>0.0518</td>
</tr>
<tr>
<td>( \sigma = 0.1 )</td>
<td>1.1223</td>
<td>1.1795</td>
<td>1.3139</td>
<td>( \sigma = 0.1 )</td>
<td>0.1120</td>
<td>0.1807</td>
<td>0.3614</td>
</tr>
<tr>
<td>( \sigma = 0.3 )</td>
<td>1.3513</td>
<td>1.5540</td>
<td>2.1035</td>
<td>( \sigma = 0.3 )</td>
<td>0.2686</td>
<td>0.5721</td>
<td>1.6977</td>
</tr>
<tr>
<td>( \sigma = \sigma_0 = 0.1 )</td>
<td>( \kappa_0 = 0.1 )</td>
<td>( \kappa_0 = 1 )</td>
<td>( \kappa_0 = 10 )</td>
<td>( \sigma = \sigma_0 = 0.1 )</td>
<td>( \kappa_0 = 0.1 )</td>
<td>( \kappa_0 = 1 )</td>
<td>( \kappa_0 = 10 )</td>
</tr>
<tr>
<td>( \kappa = 0.1 )</td>
<td>1.3886</td>
<td>1.6881</td>
<td>2.1107</td>
<td>( \kappa = 0.1 )</td>
<td>0.3683</td>
<td>0.7138</td>
<td>1.293</td>
</tr>
<tr>
<td>( \kappa = 1 )</td>
<td>2.3474</td>
<td>3.6932</td>
<td>5.7638</td>
<td>( \kappa = 1 )</td>
<td>0.9846</td>
<td>2.0941</td>
<td>4.152</td>
</tr>
<tr>
<td>( \kappa = 10 )</td>
<td>3.9451</td>
<td>9.1034</td>
<td>21.1094</td>
<td>( \kappa = 10 )</td>
<td>2.1367</td>
<td>5.0186</td>
<td>12.604</td>
</tr>
<tr>
<td>( \sigma = \sigma_0 \approx 0 )</td>
<td>( \kappa_0 = 0.1 )</td>
<td>( \kappa_0 = 1 )</td>
<td>( \kappa_0 = 10 )</td>
<td>( \sigma = \sigma_0 \approx 0 )</td>
<td>( \kappa_0 = 0.1 )</td>
<td>( \kappa_0 = 1 )</td>
<td>( \kappa_0 = 10 )</td>
</tr>
<tr>
<td>( \kappa = 0.1 )</td>
<td>1.1519</td>
<td>1.3031</td>
<td>1.4851</td>
<td>( \kappa = 0.1 )</td>
<td>0.1412</td>
<td>0.2884</td>
<td>0.4697</td>
</tr>
<tr>
<td>( \kappa = 1 )</td>
<td>1.8449</td>
<td>2.9973</td>
<td>4.5808</td>
<td>( \kappa = 1 )</td>
<td>0.5633</td>
<td>1.5009</td>
<td>2.9527</td>
</tr>
<tr>
<td>( \kappa = 10 )</td>
<td>2.8197</td>
<td>7.4472</td>
<td>18.3780</td>
<td>( \kappa = 10 )</td>
<td>0.9162</td>
<td>3.8042</td>
<td>10.4049</td>
</tr>
<tr>
<td>( (\sigma, \sigma_0) = (0.3,0.1) )</td>
<td>( \kappa_0 = 0.1 )</td>
<td>( \kappa_0 = 1 )</td>
<td>( \kappa_0 = 10 )</td>
<td>( (\sigma, \sigma_0) = (0.3,0.1) )</td>
<td>( \kappa_0 = 0.1 )</td>
<td>( \kappa_0 = 1 )</td>
<td>( \kappa_0 = 10 )</td>
</tr>
<tr>
<td>( \kappa = 0.1 )</td>
<td>1.6141</td>
<td>3.3529</td>
<td>7.6986</td>
<td>( \kappa = 0.1 )</td>
<td>0.6352</td>
<td>2.0010</td>
<td>8.0010</td>
</tr>
<tr>
<td>( \kappa = 1 )</td>
<td>1.8473</td>
<td>4.4619</td>
<td>13.1155</td>
<td>( \kappa = 1 )</td>
<td>0.9229</td>
<td>2.8664</td>
<td>8.6887</td>
</tr>
<tr>
<td>( \kappa = 10 )</td>
<td>2.1938</td>
<td>6.1790</td>
<td>23.8944</td>
<td>( \kappa = 10 )</td>
<td>1.4190</td>
<td>4.6033</td>
<td>15.0348</td>
</tr>
<tr>
<td>( (\sigma, \sigma_0) = (0.1,0.3) )</td>
<td>( \kappa_0 = 0.1 )</td>
<td>( \kappa_0 = 1 )</td>
<td>( \kappa_0 = 10 )</td>
<td>( (\sigma, \sigma_0) = (0.1,0.3) )</td>
<td>( \kappa_0 = 0.1 )</td>
<td>( \kappa_0 = 1 )</td>
<td>( \kappa_0 = 10 )</td>
</tr>
<tr>
<td>( \kappa = 0.1 )</td>
<td>1.8905</td>
<td>2.775</td>
<td>4.0879</td>
<td>( \kappa = 0.1 )</td>
<td>0.9002</td>
<td>1.3453</td>
<td>2.3386</td>
</tr>
<tr>
<td>( \kappa = 1 )</td>
<td>2.8108</td>
<td>4.9727</td>
<td>9.9275</td>
<td>( \kappa = 1 )</td>
<td>2.4988</td>
<td>3.5049</td>
<td>5.9328</td>
</tr>
<tr>
<td>( \kappa = 10 )</td>
<td>4.5242</td>
<td>9.3185</td>
<td>25.8077</td>
<td>( \kappa = 10 )</td>
<td>7.5565</td>
<td>10.1261</td>
<td>17.3192</td>
</tr>
</tbody>
</table>

Table 1: \( \mathbb{E}(M) \) and Var(\( M \)) for different combinations of \((\kappa, \kappa_0, \sigma, \sigma_0)\) in the HNGG model.
Table 2: Cov($P_1(A), P_1(B)$) for different combinations of ($\kappa, \kappa_0, \sigma, \sigma_0$) in the HNGG model.

<table>
<thead>
<tr>
<th>$\kappa = \kappa_0 = 1$</th>
<th>$\sigma_0 \approx 0$</th>
<th>$\sigma_0 = 0.1$</th>
<th>$\sigma_0 = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma \approx 0$</td>
<td>-0.0688</td>
<td>-0.066</td>
<td>-0.0608</td>
</tr>
<tr>
<td>$\sigma = 0.1$</td>
<td>-0.066</td>
<td>-0.0629</td>
<td>-0.057</td>
</tr>
<tr>
<td>$\sigma = 0.3$</td>
<td>-0.0608</td>
<td>-0.057</td>
<td>-0.05</td>
</tr>
<tr>
<td>$\kappa = \kappa_0 = 0.01$</td>
<td>$\sigma_0 \approx 0$</td>
<td>$\sigma_0 = 0.1$</td>
<td>$\sigma_0 = 0.3$</td>
</tr>
<tr>
<td>$\sigma \approx 0$</td>
<td>-0.0917</td>
<td>-0.0916</td>
<td>-0.0915</td>
</tr>
<tr>
<td>$\sigma = 0.1$</td>
<td>-0.0916</td>
<td>-0.0906</td>
<td>-0.0886</td>
</tr>
<tr>
<td>$\sigma = 0.3$</td>
<td>-0.0915</td>
<td>-0.0886</td>
<td>-0.0829</td>
</tr>
<tr>
<td>$\sigma = \sigma_0 = 0.1$</td>
<td>$\kappa_0 = 0.1$</td>
<td>$\kappa_0 = 1$</td>
<td>$\kappa_0 = 10$</td>
</tr>
<tr>
<td>$\kappa = 0.1$</td>
<td>-0.0885</td>
<td>-0.082</td>
<td>-0.0758</td>
</tr>
<tr>
<td>$\kappa = 1$</td>
<td>-0.082</td>
<td>-0.0629</td>
<td>-0.0445</td>
</tr>
<tr>
<td>$\kappa = 10$</td>
<td>-0.0758</td>
<td>-0.0445</td>
<td>-0.0143</td>
</tr>
<tr>
<td>$\sigma = \sigma_0 \approx 0$</td>
<td>$\kappa_0 = 0.1$</td>
<td>$\kappa_0 = 1$</td>
<td>$\kappa_0 = 10$</td>
</tr>
<tr>
<td>$\kappa = 0.1$</td>
<td>-0.091</td>
<td>-0.0876</td>
<td>-0.0842</td>
</tr>
<tr>
<td>$\kappa = 1$</td>
<td>-0.0876</td>
<td>-0.0688</td>
<td>-0.05</td>
</tr>
<tr>
<td>$\kappa = 10$</td>
<td>-0.0842</td>
<td>-0.05</td>
<td>-0.0159</td>
</tr>
<tr>
<td>($\sigma, \sigma_0$) = (0.3, 0.1)</td>
<td>$\kappa_0 = 0.1$</td>
<td>$\kappa_0 = 1$</td>
<td>$\kappa_0 = 10$</td>
</tr>
<tr>
<td>$\kappa = 0.1$</td>
<td>-0.0851</td>
<td>-0.0721</td>
<td>-0.0595</td>
</tr>
<tr>
<td>$\kappa = 1$</td>
<td>-0.0801</td>
<td>-0.057</td>
<td>-0.0349</td>
</tr>
<tr>
<td>$\kappa = 10$</td>
<td>-0.0755</td>
<td>-0.0435</td>
<td>-0.0127</td>
</tr>
<tr>
<td>($\sigma, \sigma_0$) = (0.1, 0.3)</td>
<td>$\kappa_0 = 0.1$</td>
<td>$\kappa_0 = 1$</td>
<td>$\kappa_0 = 10$</td>
</tr>
<tr>
<td>$\kappa = 0.1$</td>
<td>-0.0851</td>
<td>-0.0801</td>
<td>-0.0755</td>
</tr>
<tr>
<td>$\kappa = 1$</td>
<td>-0.0721</td>
<td>-0.057</td>
<td>-0.0435</td>
</tr>
<tr>
<td>$\kappa = 10$</td>
<td>-0.0595</td>
<td>-0.0349</td>
<td>-0.0127</td>
</tr>
</tbody>
</table>

Table 3: Cov($P_1(A), P_2(B)$) for different combinations of ($\kappa_0, \sigma_0$) in the HNGG model.

<table>
<thead>
<tr>
<th>$\kappa_0 = 1$</th>
<th>$\sigma_0 \approx 0$</th>
<th>$\sigma_0 = 0.1$</th>
<th>$\sigma_0 = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa_0 = 0.01$</td>
<td>-0.0459</td>
<td>-0.0403</td>
<td>-0.0299</td>
</tr>
<tr>
<td>$\sigma_0 \approx 0$</td>
<td>-0.0908</td>
<td>-0.0817</td>
<td>-0.0633</td>
</tr>
<tr>
<td>$\sigma_0 = 0.1$</td>
<td>-0.0744</td>
<td>-0.0403</td>
<td>-0.0074</td>
</tr>
<tr>
<td>$\sigma_0 = 0.3$</td>
<td>-0.0566</td>
<td>-0.0299</td>
<td>-0.0057</td>
</tr>
<tr>
<td>$\kappa = \kappa_0 = 1$</td>
<td>$\sigma_0 \approx 0$</td>
<td>$\sigma_0 = 0.1$</td>
<td>$\sigma_0 = 0.3$</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----------------</td>
<td>----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>$\sigma \approx 0$</td>
<td>1.4729</td>
<td>1.3347</td>
<td>1.0551</td>
</tr>
<tr>
<td>$\sigma = 0.1$</td>
<td>1.5293</td>
<td>1.3942</td>
<td>1.1171</td>
</tr>
<tr>
<td>$\sigma = 0.3$</td>
<td>1.6491</td>
<td>1.5239</td>
<td>1.2593</td>
</tr>
<tr>
<td>$\kappa = \kappa_0 = 0.01$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma \approx 0$</td>
<td>2.2736</td>
<td>2.0460</td>
<td>1.5897</td>
</tr>
<tr>
<td>$\sigma = 0.1$</td>
<td>2.2758</td>
<td>2.0687</td>
<td>1.6405</td>
</tr>
<tr>
<td>$\sigma = 0.3$</td>
<td>2.2803</td>
<td>2.1157</td>
<td>1.7541</td>
</tr>
<tr>
<td>$\sigma = \sigma_0 = 0.1$</td>
<td>$\kappa_0 = 0.1$</td>
<td>$\kappa_0 = 1$</td>
<td>$\kappa_0 = 10$</td>
</tr>
<tr>
<td>$\kappa = 0.1$</td>
<td>1.9287</td>
<td>1.1061</td>
<td>0.20855</td>
</tr>
<tr>
<td>$\kappa = 1$</td>
<td>2.0773</td>
<td>1.3942</td>
<td>0.2909</td>
</tr>
<tr>
<td>$\kappa = 10$</td>
<td>2.2473</td>
<td>1.9345</td>
<td>0.6810</td>
</tr>
<tr>
<td>$\sigma = \sigma_0 \approx 0$</td>
<td>$\kappa_0 = 0.1$</td>
<td>$\kappa_0 = 1$</td>
<td>$\kappa_0 = 10$</td>
</tr>
<tr>
<td>$\kappa = 0.1$</td>
<td>2.1043</td>
<td>1.1937</td>
<td>0.2196</td>
</tr>
<tr>
<td>$\kappa = 1$</td>
<td>2.1849</td>
<td>1.4729</td>
<td>0.3023</td>
</tr>
<tr>
<td>$\kappa = 10$</td>
<td>2.2728</td>
<td>1.9862</td>
<td>0.6999</td>
</tr>
<tr>
<td>$(\sigma, \sigma_0) = (0.3, 0.1)$</td>
<td>$\kappa_0 = 0.1$</td>
<td>$\kappa_0 = 1$</td>
<td>$\kappa_0 = 10$</td>
</tr>
<tr>
<td>$\kappa = 0.1$</td>
<td>2.0029</td>
<td>1.2356</td>
<td>0.2416</td>
</tr>
<tr>
<td>$\kappa = 1$</td>
<td>2.1279</td>
<td>1.5239</td>
<td>0.3422</td>
</tr>
<tr>
<td>$\kappa = 10$</td>
<td>2.2572</td>
<td>1.9776</td>
<td>0.7559</td>
</tr>
<tr>
<td>$(\sigma, \sigma_0) = (0.1, 0.3)$</td>
<td>$\kappa_0 = 0.1$</td>
<td>$\kappa_0 = 1$</td>
<td>$\kappa_0 = 10$</td>
</tr>
<tr>
<td>$\kappa = 0.1$</td>
<td>1.5234</td>
<td>0.8349</td>
<td>0.1599</td>
</tr>
<tr>
<td>$\kappa = 1$</td>
<td>1.7960</td>
<td>1.1171</td>
<td>0.2253</td>
</tr>
<tr>
<td>$\kappa = 10$</td>
<td>2.1891</td>
<td>1.7893</td>
<td>0.5718</td>
</tr>
</tbody>
</table>

Table 4: CoSk($P_1(A \cap B), P_2(A \cap B)$) for different combinations of $(\kappa, \kappa_0, \sigma, \sigma_0)$ in the HNGG model.
Analogously to what presented in the main manuscript, we show posterior estimates for the LPML (Table 5) and RI in the second group (Table 6), obtained when fitting the HNGG model with some additional combinations of the hyperparameters \((\kappa, \kappa_0, \sigma, \sigma_0)\). These simulations show how the inference is influenced by the choice of such pivotal quantities. In particular, we can observe how an increased variability in the a priori number of clusters (higher values of the hyperparameters \((\kappa, \kappa_0, \sigma, \sigma_0)\)) is associated with a better fit in terms of LPML values. However, these combinations are not associated with good recovery of the partition of the data in the second group, as it was observed in the examples reported in the paper.

### 3.4 Sensitivity analysis for the hyperparameters of \(P_0\)

In Bayesian nonparametrics, it is well known how the choice of the centering measure \(P_0\) is crucial for posterior inference. In order to explore the influence of this parameter on the analysis carried out in the paper, we set up an extensive sensitivity analysis for the choice of the hyperparameters of \(P_0\). In the simulation study of Section 4.1 of the paper, \(P_0\) is a
normal-inverse-gamma distribution for the mean and variance parameters $(\mu, \tau^2)$:

$$P_0(d\mu, d\tau^2) = N(d\mu|m_0, \tau^2/k_0)\text{inv-gamma}(d\tau^2|a_{\tau^2}, b_{\tau^2}),$$

characterised by the set of hyperparameters $(m_0, k_0, a_{\tau^2}, b_{\tau^2})$. In order to fix these quantities on a varied range of values, we impose the values of the a priori moments of $\mu$ and $\tau^2$. Let $\hat{y}_n$ and $\hat{s}_n^2$ be the sample mean and variance of the whole dataset, respectively. Introduce the quantity $d_n$, representing an initial guess on the value of $\mathbb{E}(M)$, i.e. the mean number of clusters in the dataset. Consider the following system of equations, to be solved with respect to $(m_0, k_0, a_{\tau^2}, b_{\tau^2})$:

$$\mathbb{E}[\mu] = m_0 = \hat{y}_n,$$
$$\mathbb{E}[\tau^2] = \frac{b_{\tau^2}}{a_{\tau^2} - 1} = \frac{\hat{s}_n^2}{d_n},$$
$$\text{Var}(\mu) = \mathbb{E}[\tau^2]/k_0,$$
$$\text{Var}(\tau^2) = \mathbb{E}^2[\tau^2] \frac{b_{\tau^2}}{a_{\tau^2} - 2},$$

where the variances of $\mu$ and $\tau^2$ are assigned a value in $\{0.1, 1, 5\}$, while $d_n$ is ranged in $\{1, 2, 3, 5, 10\}$. Finally, $\kappa, \kappa_0, \sigma, \sigma_0$ are fixed to 0.1.

The performance of the proposed model in each setting is evaluated via the LPML and RI in the second group, as shown in the paper. From the estimates of the LPML (Table 7), it is clear how higher a priori variances for both $\mu$ and $\tau^2$ yield better fitting of the data, regardless of the value of $d_n$. This behaviour can be explained by considering that the hyperparameters of the HNGG process are fixed to a low value, yielding a small a priori number of clusters, hence requiring higher variability in the values of the parameters $\mu$ and $\tau^2$ in order to correctly recover the distribution of the data. Scenarios with high LPML values are also characterised by low RI estimates (Table 7), indicating poor recovery of the cluster assignment of the data. The correct number of clusters (2) is identified for most combinations of the hyperparameters $(m_0, k_0, a_{\tau^2}, b_{\tau^2})$, with the exception of the ones characterised by $d_n = 1$, where most of the times the data are clustered all together. Notice that this configuration is also the one maximizing the RI estimates. This behaviour is due to the larger a priori mean and variances of $\mu$ and $\tau^2$, making it difficult for the model to identify multiple clusters in the second group. On the other hand, the effect of smaller a priori variances on the estimated number of clusters, e.g. when $d_n = 10$, is mitigated by the low values of the hyperparameters $(\kappa, \kappa_0, \sigma, \sigma_0)$. As a matter of fact, the maximum estimated number of clusters among the simulations is 3.
Table 7: Simulated Data \((d = 2)\) – LPML(10^3) and Rand Index (RI) for different combinations of \((m_0, k_0, a_{\tau^2}, b_{\tau^2})\).
4 Comparison with existing methods

4.1 Bayesian hierarchical parametric model

We compare the results obtained with the proposed model versus the hierarchical parametric model of (Hoff 2009):

\[ y_j^1, \ldots, y_j^n \mid \mu_j, \tau_j^2 \overset{\text{iid}}{\sim} N(\mu_j, \tau_j^2), \quad j = 1, \ldots, d \quad (16) \]

\[ \mu_1, \ldots, \mu_d \mid \vartheta, \xi^2 \overset{\text{iid}}{\sim} N(\vartheta, \xi^2), \]

\[ \tau_1^2, \ldots, \tau_d^2 \mid \nu_0, \tau_0^2 \overset{\text{iid}}{\sim} \text{inv-gamma} \left( \frac{\nu_0}{2}, \frac{\nu_0 \tau_0^2}{2} \right), \]

\[ \vartheta \sim N(\vartheta_0, \gamma_0^2), \]

\[ \xi^2 \sim \text{inv-gamma} \left( \frac{\eta_0}{2}, \frac{\eta_0 \xi_0^2}{2} \right), \]

\[ \tau_0^2 \sim \text{gamma}(a_{\tau_0}, b_{\tau_0}), \]

\[ \nu_0 \sim \text{Geom}(1 - e^{-\alpha}), \]

where Geom(p) is the Geometric distribution on \( \{1, 2, \ldots\} \) with mean \( 1/p \) and \( \alpha > 0 \).

In the following comparison, we will focus on the fitting properties of this parametric model, comparing them with the results obtained with our model, since a comparison of the clustering properties is not possible in the parametric setting.

4.1.1 Simulated data \((d = 2)\)

First, we need to specify the hyperparameters involved in model (16). This can be done by solving a system of equations based on the estimates of the first and second moments of \( \mu_j \) and \( \tau_j^2 \), for \( j = 1, \ldots, d \). In particular, we indicate by \( \bar{X} \) the sample mean of the sequence of random variables \( X_1, \ldots, X_d \) and fix its first and second moments. When \( X_j = \mu_j \), for \( j = 1, \ldots, d \), we recover the overall mean and variability. When \( X_j = \tau_j^2 \), for \( j = 1, \ldots, d \), we recover the information regarding the variability within or between the groups involved in the study. These quantities are of interest in standard statistical analysis when group labels are available for the data at hand. The hyperparameters are then fixed after solving
the following system of equations:

\[
\begin{align*}
E[\hat{\mu}] &= E[\mu_j] = E[E[\mu_j|\vartheta, \xi^2]] = \vartheta_0 = \hat{y}_n \quad \text{(overall mean)}, \\
\text{Var}(\hat{\mu}) &= \frac{1}{d} \left( \gamma_0^2 + \frac{\nu_0 \xi^2_0}{\nu_0} \right) = \hat{s}^2_n \quad \text{(overall variability)}, \\
E[\tau^2] &= E[\tau^2_j] = \frac{\nu_0 \tau^2_0}{\nu_0 - 1} = \frac{1}{d} \sum_{j=1}^{d} \hat{s}^2_j \quad \text{(within-group variability)}, \\
E[\xi^2] &= \frac{\nu_0 \xi^2_0}{\nu_0 - 1} = \frac{1}{d - 1} \sum_{j=1}^{d} (\hat{y}_j - \hat{y}_n)^2 \quad \text{(between-group variability)},
\end{align*}
\]

where \(\hat{y}_j\) and \(\hat{s}^2_j\) for \(j = 1, \ldots, d\) are group-specific sample mean and variances, respectively, while \(\hat{y}_n\) and \(\hat{s}^2_n\) are the sample mean and variance of the whole dataset. In order for the first moment of the inv-gamma distributions to be defined, we fixed \(\eta_0 = \nu_0 = 3\). This setting yields \((\vartheta_0, \gamma_0^2, \xi^2_0, \tau^2_0) = (0.25, 2.79, 0.30, 0.47)\). For posterior inference, we ran an MCMC chain of 50,000 iterations, saving the last 5,000 samples after a thinning of 2. Figure 1 shows the density estimation results within each group, overlapped with the histogram of the data (histograms within each group are scaled to the sampling probability used to simulate the data). The parametric model is not able to recover the heterogeneity in the data with only one component assigned to each group, and yields a poorer fitting (LPML(10^3) = -0.3202) when compared to Table 1. The posterior distribution of the hyperparameters \(\tau^2_0\) and \(\nu_0\) are shown in Figures 1, panels (c) and (d). These parameters yield insights on the sharing of information between the components of the mixture model. In particular, low values \(\nu_0\), as depicted in Figure 1(d), indicate that the variances are quite different and that sharing of information occurs among groups.

### 4.1.2 School data

We now fit model (16) to the School dataset of (Hoff 2009). Prior elicitation follows the textbook guidelines. We ran 50,000 iterations of the posterior sampling algorithm, and saved the last 5,000 samples for inference, after a thinning of 2. In an effort to reproduce the analysis proposed by Hoff (2009), as well as to show the advantages of the proposed model, we show in Figure 2 the predictive densities for some selected schools. In particular, we focus on schools number 46 and 82, characterised by exceptionally low sample means of the math scores. In addition to these two schools, that were originally reported in the textbook, we add the ones with the higher and lower posterior mean for the number of clusters, namely schools number 12 and 67, respectively. Overall, the parametric model yields lower performance in terms of fitting (LPML(10^3) = -7.3023 VS -7.2925 for
Figure 1: Simulated data ($d = 2$) – Posterior density estimation within each group. (a) Parametric model (Hoff); (b) Proposed model. Posterior distribution of $\tau_0^2$ (c) and $\nu_0$ (d).
the nonparametric case). Additionally, the parametric model is struggling to capture the heterogeneity in the school data, due to the enhanced shrinkage effect. In particular, we highlight the case of school 12 (Figure 2(c)), where heavy tails are detected, and of school 67 (Figure 2(d)), where negative skewness is present. It is evident how a single component used to model each group is not sufficient to provide flexible estimation of the density of the data in these groups.

4.2 Frequentist approach

The proposed model is hereby compared with a standard mixed-effects model representing the frequentist counterpart. Pinheiro and Bates (2000) deeply investigated mixed-effects models within the frequentist framework, and also provided an R package to perform analysis in this setting, namely the \texttt{nlme} package (Pinheiro et al. 2018). In particular, we consider the following model:

\begin{equation}
 y_j = X_j \beta + Z_j b_j + \epsilon_j, \quad j = 1, \ldots, d
\end{equation}

\[ \epsilon_{j1}, \ldots, \epsilon_{jn_j} \overset{iid}{\sim} N(0, \tau^2_j), \quad j = 1, \ldots, d, \]

\[ b_1, \ldots, b_d \overset{iid}{\sim} N(0, \Psi) \]

where \(X_j\) is an \(n_j \times p\) matrix of fixed covariates, \(\beta\) is a \(p \times 1\) parameter vector of fixed effects, \(Z_j\) is an \(n_j \times v\) matrix of covariates for the \(v \times 1\) vector of random effects \(b_j\), and \(\epsilon_j\) is an \(n_j \times 1\) vector of errors. By letting \(X_j = 1\) and \(Z_j = 1\), for \(j = 1, \ldots, d\), we obtain a frequentist model \(\beta + b_j\) has the same interpretation as \(\mu_j\) in model (16). We used the function \texttt{lme} to fit model (17) to the simulated data presented in Section 4.1 of the paper. For a visual comparison, the estimates of the distribution of a new observation in each of the two groups is produced via nonparametric bootstrap, and are reported in Figure 3.

5 Simulation Study \((d = 100)\)

We present in this section an additional simulation study, where the number of groups is relatively large \((d = 100)\), and the data in each group are characterised by different sample sizes. The group sizes \(n_j\), for \(j = 1, \ldots, d\), are simulated from a Poisson distribution with mean 20. Observations in each group are simulated from a mixture of Gaussian distributions with a different number of components. The components are selected uniformly at random from a set of 10 unique labels reported in Table 8, hence defining the values for the vectors of labels \(\theta^*_j\) and the number of clusters \(K_j\), for \(j = 1, \ldots, d\). The mixture components are shown in Figure 4(a). The weights of the components within each group are simulated
Figure 2: School data – Posterior density estimation within each of the selected group. (a) Parametric model (Hoff); (b) Proposed model; (c) School 12; (d) School 67.
Figure 3: Simulated data \((d = 2)\) – Bootstrap density estimation within each group and confidence intervals for the random effects.
from a Dirichlet distribution with parameters $\alpha_1 = \cdots = \alpha_{K_j} = 0.1$. Figure 4(b) shows a histogram of the resulting dataset, normalized within each group with respect to the weights of the mixture. It is clear how the process used to simulate the data yields strong heterogeneity, both within and between groups, and requires a flexible model that allows sharing of information for a suitable analysis, especially in the regions where different groups overlap. Figures 4(c,d) show the histogram of the sample means $\hat{y}_j$, for $j = 1, \ldots, d$, computed within each group, and the relationship between these quantities and the sizes of the groups. We can observe positive skewness in the distribution of the sample means, and a dependency on the group size.

The hyperparameters of the HNGG process are fixed such that where $\kappa = \kappa_0 = 0.1$, $\sigma, \sigma_0 \sim \text{Beta}(2,18)$, and the hyperparameters of $P_0$ are chosen such that $E[\mu] = \hat{y}_n$, $E[\tau^2] = \hat{s}_n^2/10$, $\text{Var}(\mu) = \text{Var}(\tau^2) = 1$, where $\hat{y}_n$ and $\hat{s}_n^2$ represent the sample mean and variance of the whole dataset, respectively.

Inference on the clustering structure of the data is summarized in Figure 5, where the data within each group are plotted according to their increasing sample mean $\hat{y}_j$, for $j = 1, \ldots, 100$ in panel (a). We can observe how the sharing of information allows one to cluster data points belonging to different groups, but sharing similar characteristics. This feature would be partially lost if we instead were to fit the whole dataset, regardless of the group assignment, yielding a predictive density as in Figure 5(d). However, strong heterogeneity is still clear in the aggregated histogram. The posterior distribution of the number of clusters is close to the truth ($10$), and it is reported in Figure 5(c). Finally, Figures 5(e,f) show how the posterior distributions of both $\sigma$ and $\sigma_0$ are concentrated on values away from zero, indicating the need to depart from the simple Dirichlet process case.

For the sake of completeness, a comparison of the performance of the proposed model with its independent counterpart (both NGG and Dirichlet) is presented. The inferential procedure has been applied to 25 datasets simulated using the same procedure introduced in this section. For each MCMC chain so produced, we estimated the posterior probability of inclusion $\hat{\pi}_{i_1i_2}$ between two elements $(i_1, i_2)$, by averaging over the clustering configuration of the whole dataset (i.e., by aligning the $d = 100$ groups and considering the cluster allocation labels). We then considered $(i_1, i_2)$ a “positive” if $\hat{\pi}_{i_1i_2} > \epsilon$, for $\epsilon$ ranging in $(0,1)$. This procedure allowed us to produce a receiver operating characteristic (ROC) curve comparing the true and false positive rates as average over the 25 replicates. The curves are depicted in Figure 6, where it is clear how the HNGG model outperforms its

<table>
<thead>
<tr>
<th>$\mu^*$</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>0</th>
<th>1</th>
<th>2.5</th>
<th>3</th>
<th>4.5</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau^2$</td>
<td>1</td>
<td>0.25</td>
<td>0.1</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>0.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Table 8: Simulated Data ($d = 100$) – unique labels used for simulation.
Figure 4: Simulated data ($d = 100$) – Densities used to simulate the data in the $d = 100$ simulation study (unnormalized) (a). Histograms of the data in each group (normalized within each group) (b). Histogram of $\hat{y}_j$, the sample means in each group (c). Sample sizes $n_j$ VS sample means $\hat{y}_j$ (d).
Figure 5: Simulated data ($d = 100$) – Scatter plot of the data organized by increasing sample mean. The colours refer to the cluster assignment in the Binder partition (a). Posterior means of the number of clusters in each group (b). Posterior distribution of the number of clusters $M$ (c). Predictive distribution in a new group and histogram of the data (d). Posterior distribution of $\sigma$ (e) and $\sigma_0$ (f).
Figure 6: Simulated data \((d = 100)\) – ROC curves obtained using the posterior inclusion probability of the \(d = 100\) groups of data. It is clear the improvement yielded by the possibility of sharing the information in the HNGG process.

independent counterparts in terms of accuracy of the clustering (the area under the ROC curve is evidently larger).
References


