Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Hierarchical Normalized Completely Random Measures to Cluster Grouped Data

Raffaele Argiento, Andrea Cremaschi & Marina Vannucci

To cite this article: Raffaele Argiento, Andrea Cremaschi & Marina Vannucci (2019): Hierarchical Normalized Completely Random Measures to Cluster Grouped Data, Journal of the American Statistical Association, DOI: 10.1080/01621459.2019.1594833

To link to this article: https://doi.org/10.1080/01621459.2019.1594833

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 modeled 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 $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ 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 $(\theta_{j1}, \ldots, \theta_{jn_j}) \mid P_j \overset{\text{iid}}{\sim} P_j$, for each $j = 1, \ldots, d$, under a suitable prior (*de Finetti measure*) for the vector of random probabilities $(P_1, \ldots, P_d)$. Readers are referred to 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., $(P_1, \ldots, P_d) \mid P \overset{\text{iid}}{\sim} P$, 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 their ease of interpretation, computational availability, and elegant mathematical properties. Any mixture model with an almost surely discrete mixing measure leads to ties in θ with positive probability. This induces a random partition of the subject labels via the values of the parameters θ, 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 $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$, the data are independent from some parametric family of distributions, and the prior on $\boldsymbol{\theta}$ 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 the *l*-clustering), and a natural clustering across the whole multidimensional array $\boldsymbol{\theta}$. 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 the 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 processes of Camerlenghi et al. (2018) and Bassetti et al. (2018), which investigate more theoretical aspects, in this paper we provide a detailed study of the two-layered hierarchical clustering structure induced by these models. Original contributions of our 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. While in Section 2 below we acknowledge some overlap between our theoretical results and those of Bassetti et al. (2018), we point out that our work was developed independently of theirs. In addition, we use original techniques in the proofs of some of the 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 introduced in Argiento et al. (2014) and Malsiner-Walli et al. (2017). These authors, however, address a different problem from the one considered here.

The rest of the paper is organized as follows: completely random measures and normalized completely random measures with discrete centering distribution are introduced in Section 2. Characteristic properties of the clustering induced by a normalized completely random measure are also discussed, such as the expression of mixed moments and the law of the number of clusters. Next, the proposed characterization is extended to hierarchical normalized completely random measures for grouped data. In Section 3, insights on the posterior sampling process for the proposed mixture models are provided, including prediction. Simulation results as well as an application to a benchmark dataset are given in Section 4. Finally, Section 5 concludes the paper. Proofs of the theoretical results, algorithmic details and

additional results are reported in the Supplementary Materials available online.

2 Methodology

2.1 Normalized completely random measures with discrete centering

Let Θ be a complete and separable metric space, endowed with the corresponding Borel σ-algebra $\mathcal{B}$. A completely random measure (CRM) on Θ is a random measure μ_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 *Lèvy intensity measure* ν that factorizes into $\nu(ds, d\tau) = \alpha(s)P(d\tau)ds$, where α 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 α. In what follows, we will refer to α only as the Lèvy intensity.

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

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

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 Lèvy intensity α. We indicate with $P_1 \sim \text{NormCRM}(\alpha, P)$ a NormCRM with Lèvy intensity α 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*

$$\int_{\mathbb{R}^+} \alpha(s)ds = +\infty, \quad \text{and} \quad \int_{\mathbb{R}^+} (1-e^{-s})\alpha(s)ds < +\infty. \qquad (2)$$

The class of NormCRMs encompasses the well-known Dirichlet process Dir $^{(\kappa, P)}$, obtained by normalization of a gamma process, with Lèvy intensity $\alpha(s) = \kappa s^{-1}\mathrm{e}^{-s}\mathbb{I}_{(0,+\infty)}(s)$, for $\kappa > 0$. It also includes the normalized generalized gamma process NGG $^{(\kappa, \sigma, P)}$ of Lijoi et al. (2007), obtained when $\alpha(s) = \frac{\kappa}{\Gamma(1-\sigma)} s^{-1-\sigma}\mathrm{e}^{-s}\mathbb{I}_{(0,+\infty)}(s)$, for $0 \le \sigma < 1$, and the normalized Bessel process NormBessel $^{(\kappa, \omega, P)}$ of Argiento et al. (2016), when $\alpha(s) = \frac{\kappa}{s}\mathrm{e}^{-\omega s} I_0(s)\mathbb{I}_{(0,+\infty)}(s)$, for $\omega \ge 1$ and $I_0(s)$ the modified Bessel function of the first kind. In the expressions of the Levy 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 $(\tilde{\theta}_1, \ldots, \tilde{\theta}_n) \mid P_1 \overset{iid}{\sim} P_1$. 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 $P_1^\star$ be a random probability measure on the positive integers $\mathbb{N} = \{1, 2, \ldots\}$ whose weights coincide with the weight of P_1, that is,

$$P_1^\star(\cdot) = \sum_{l \ge 1} w_l \delta_l(\cdot). \qquad (3)$$

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* $P_1^\star$ *defined as in (3). Define* $\theta_1=\tau_{l_1},\ldots,\theta_n=\tau_{l_n}$ *with* $\{\tau_l\}\overset{\text{iid}}{\sim}P$ *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 Section 1.1 of the Supplementary Materials. A similar result is shown in Proposition 1 of Bassetti et al. (2018).

Here we 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 $\tilde{\rho}=\{\tilde{C}_1,\ldots,\tilde{C}_{\tilde{K}_n}\}$ 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 $j=1,\ldots,\tilde{K}_n$, and $(\tilde{\theta}^*_1,\ldots,\tilde{\theta}^*_{\tilde{K}_n})$ 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 $(\tilde{\rho},\tilde{\theta}^*_1,\ldots,\tilde{\theta}^*_{\tilde{K}_n})$ as

$$\mathcal{L}(\tilde{\rho},d\tilde{\theta}^*_1,\ldots,d\tilde{\theta}^*_{\tilde{K}_n})=\mathcal{L}(\tilde{\rho})\mathcal{L}(d\tilde{\theta}^*_1,\ldots,d\tilde{\theta}^*_{\tilde{K}_n}\mid\tilde{K}_n)=\pi(\tilde{\rho})\prod_{l=1}^{\tilde{K}_n}P(d\tilde{\theta}^*_l),\quad(4)$$

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 α of the NormCRM, we write $\pi(\tilde{\rho})=\text{eppf}(\tilde{e}_1,\ldots,\tilde{e}_{\tilde{K}_n};\alpha)$, 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 $j=1,\ldots,\tilde{K}_n$. A formula for the eppf of a generic NormCRM can be obtained as (see formulas (36)-(37) in Pitman (2003))

$$\text{eppf}(\tilde{e}_1,\ldots,\tilde{e}_{\tilde{K}_n};\alpha)=\int_0^{+\infty}\frac{u^{n-1}}{\Gamma(n)}\text{e}^{-\phi(u)}\prod_{l=1}^{\tilde{K}_n}c_{\tilde{e}_l}(u)du,\qquad(5)$$

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

$$\phi(u) = \int_0^{+\infty} (1 - \mathrm{e}^{-us})\alpha(s)ds, \quad c_m(u) = (-1)^{m-1}\phi_m(u) = \int_0^{+\infty} s^m \mathrm{e}^{-us}\alpha(s)ds. \qquad (6)$$

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 α, and the law of the cluster-specific parameters $(\tilde{\theta}_1^*, \dots, \tilde{\theta}_{\tilde{K}_n}^*)$, that conditionally upon the number of unique values $\tilde{K}_n$ is the $\tilde{K}_n$-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 slight different interpretation. With this aim, consider a sample $(l_1, \dots, l_n)$ from $P_1^\star$ as in (3), and the vector $(\theta_1, \dots, \theta_n)$ as in Lemma 1. Also, let $\boldsymbol{l}^* = (l_1^*, \dots, l_{K_n}^*)$ be the vector of unique values among $(l_1, \dots, l_n)$ and ρ the induced clustering, i.e., $\rho = \{C_1, \dots, C_{K_n}\}$, where $i \in C_h$ iff $l_i = l_h^*$, with $i = 1, \dots, n$ and $h = 1, \dots, K_n$. The law of ρ can be characterized in terms of generalized Chinese restaurant process (see Pitman 2006) as the eppf induced by the Lévy intensity α. To prove this we first observe that, if $(l_1, \dots, l_n) \mid P_1^\star \overset{\text{iid}}{\sim} P_1^\star$, then $\mathcal{L}(l_1, \dots, l_n) = \mathbb{E}(w_{l_1} \dots w_{l_n})$. Then, using the equivalent representation of $(l_1, \dots, l_n)$ in terms of $(l_1^*, \dots, l_{K_n}^*)$ and $\rho = \{C_1, \dots, C_{K_n}\}$, with $e_j = \mathrm{Card}(C_j)$, for $j = 1, \dots, K_n$, a change of variable leads to $\mathcal{L}(l_1^*, \dots, l_{K_n}^*, \rho) = \mathbb{E}(w_{l_1^*}^{e_1} \dots w_{l_{K_n}^*}^{e_{K_n}})$. Hence, by formula (4) in Pitman (2003), the law of ρ is $\mathcal{L}(\rho) = \sum_{l_1^*, \dots, l_{K_n}^*} \mathbb{E}(w_{l_1^*}^{e_1} \dots w_{l_{K_n}^*}^{e_{K_n}}) = \mathrm{eppf}(e_1, \dots, e_{K_n}; \alpha)$, where $l_1^*, \dots, l_{K_n}^*$ ranges over all permutations of K_n positive integers. We are now ready to show that, even if we do not assume P to be diffuse, the law of the sample $(\theta_1, \dots, \theta_n)$ has a unique representation as in (4), provided that ρ is the partition induced by

$(l_1,\ldots,l_n)$ and that $(\theta_1^*,\ldots,\theta_{K_n}^*)$ 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* $(\rho,\theta_1^*,\ldots,\theta_{K_n}^*)$ *s.t.:*

1. $\rho=\{C_1,\ldots,C_{K_n}\}$ *is the clustering induced by the* $\mathbf{l}^*$ *on the data indices (i.e.,* $i\in C_h$ *iff* $l_i=l_h^*$ *for* $i=1,\ldots,n$ *and* $h=1,\ldots,K_n$*),*
2. θ_h^* *is the value shared by all the θ's in group* C_h*, for* $h=1,\ldots,K_n$*.*

We point out that, in an *l*-clustering representation, $(\theta_1^*,\ldots,\theta_{K_n}^*)$ is not the vector of unique values among the θs, and so K_n is not the random variable representing the number of different values among the θs, 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 θ^*'s. Moreover, from $(\rho,\theta_1^*,\ldots,\theta_{K_n}^*)$, 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 from a NormCRM whose centering measure is a discrete distribution on the colored lines. In this sample we have $\theta=$ *(continuous green, dashed orange, continuous green, dotted blue, continuous green, dashed orange, dashed orange, dashed orange)*, obtained under the hypothesis of Lemma 1, with $(l_1,\ldots,l_n)=(1,2,1,3,1,4,2,4)$, τ_1= *continuous green*, τ_2= *dashed orange*, $\tau_3=$ *dotted blue*, τ_4= *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 $(\theta_1^*,\ldots,\theta_{K_n}^*)=$ *(continuous green, dashed orange, dotted blue,dashed orange)*. It is clear that we can recover the sample from the *l*-clustering partition by letting $\theta_i=\theta_{l_i}^*, i=1,\ldots,n$. 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 ρ of the *l*-clustering is not the one induced by the unique

values among the θs. In fact, $\theta_2^* = \theta_4^*$ = *dashed orange* with both clusters C_2 and C_4 containing orange dashed lines, while the natural clustering is $\tilde{\rho} = \{\tilde{C}_1 = \{1,3,5\}, \tilde{C}_2 = \{2,7,6,8\}, \tilde{C}_3 = \{4\}\}$.

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 ρ and* $(\theta_1^*, \ldots, \theta_{K_n}^*)$ *of an l-clustering representation. In particular,*

$$\mathcal{L}(\rho, d\theta_1^*, \ldots, d\theta_{K_n}^*) = \pi(\rho) \prod_{h=1}^{K_n} P(d\theta_h^*) = eppf(e_1, \ldots, e_{K_n}; \alpha) \prod_{h=1}^{K_n} P(d\theta_h^*). \quad (7)$$

Proof: See Section 1.2 of the Supplementary Materials. A different presentation of this result is given in Proposition 1 of Bassetti et al. (2018). The result implies that ρ is chosen according to the eppf induced by α and that $(\theta_1^*, \ldots, \theta_{K_n}^*)$ 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 $(\theta_{j1}, \ldots, \theta_{jn_j}) \mid P_j \overset{\text{iid}}{\sim} P_j$, 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 $(P_1, \ldots, P_d) \mid P \overset{\text{iid}}{\sim} \text{NormCRM}(\alpha, P)$, where $P \sim \text{NormCRM}(\alpha_0, P_0)$, with P_0 a diffuse measure on Θ and α_0 and α 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 $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$, with $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jn_j}),\ j = 1 \ldots, d$, with elements in Θ, such that

$$
\begin{aligned}
&(\theta_{j1}, \ldots, \theta_{jn_j}) \mid P_j \overset{\text{iid}}{\sim} P_j, \quad j = 1, \ldots, d, \\
&(P_1, \ldots, P_d) \mid P \overset{\text{iid}}{\sim} \text{NormCRM}(\alpha; P), \qquad (8) \\
&P \sim \text{NormCRM}(\alpha_0; P_0).
\end{aligned}
$$

The class of hierarchical NormCRMs reduces to the celebrated hierarchical Dirichlet process of Teh et al. (2006) when α_0 and α are the Lèvy 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). Below we 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.

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 $j = 1, \ldots, d, \boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jn_j})$ is a sample from P_j, that is a NormCRM with the discrete distribution P as centering measure. Consequently, $\boldsymbol{\theta}_j$ can be obtained from its *l*-clustering representation $\rho_j = (C_{j1}, \ldots, C_{jK_j}), \boldsymbol{\theta}_j^* = (\theta_{j1}^*, \ldots, \theta_{jK_j}^*)$. 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:

$$
\begin{aligned}
&\rho_j \mid \alpha \overset{\text{ind}}{\sim} \text{eppf}(\boldsymbol{e}_j; \alpha), \\
&(\theta_{j1}^*, \ldots, \theta_{jK_j}^*) \mid K_j, P \overset{\text{iid}}{\sim} P, \qquad (9) \\
&P \sim \text{NormCRM}(\alpha_0; P_0),
\end{aligned}
$$

where $\boldsymbol{e}_j = (e_{j1}, \ldots, e_{jK_j})$ 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:

$$t(j,h) = \sum_{s=0}^{j-1} k_s + h, \quad h = 1,\ldots,k_j, \quad j = 1,\ldots,d, \qquad (10)$$

where $k_0 = 0$. Consider now the multidimensional array $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_d)$, where each element $\boldsymbol{\theta}_j = (\theta_{j1},\ldots,\theta_{jn_j})$ is a sample of size n_j from P_j, for $j = 1,\ldots,d$. Furthermore, consider $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^*,\ldots,\boldsymbol{\theta}_d^*)$, the multidimensional array where each element $\boldsymbol{\theta}_j^*$ is the vector of shared value of the *l*-clustering representation of $\boldsymbol{\theta}_j$ for each group of data, as described in definition (1). The action of the function t in (10) on the indices of $\boldsymbol{\theta}^*$ 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, $\boldsymbol{\theta}^*$. Conditionally on T, the vector $\boldsymbol{\theta}^*$ is a sample of size T from P, that is $(\theta_1^*,\ldots,\theta_T^*) \mid T, P \overset{\text{iid}}{\sim} P$. Since P is a NormCRM with diffuse centering measure P_0, the *l*-clustering representation of $\boldsymbol{\theta}^*$ coincides with the natural clustering induced by the NormCRM. In particular, $\boldsymbol{\psi} = (\psi_1,\ldots,\psi_M)$ will denote the vector of unique values in $\boldsymbol{\theta}^*$, and $\eta = (D_1,\ldots,D_M)$ the clustering induced by $\boldsymbol{\psi}$ 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 $\boldsymbol{d} = (d_1,\ldots,d_M)$ the size of the clusters in η. Interestingly, the vector $\boldsymbol{\psi}$ is also the vector of unique values among the multidimensional array $\boldsymbol{\theta}$.

Proposition 2. *The marginal law of the multidimensional array* $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_d)$ *from a hierarchical NormCRM defined as in (8), can be characterized in terms of* $(\boldsymbol{\rho},\eta,\boldsymbol{\psi})$*, with* $\boldsymbol{\rho} = (\rho_1,\ldots,\rho_d)$*, as:*

$$\mathcal{L}(\boldsymbol{\rho},\eta,d\boldsymbol{\psi}) = \mathcal{L}(\eta \mid \boldsymbol{\rho}) \prod_{j=1}^{d} \mathcal{L}(\rho_j) \prod_{m=1}^{M} P_0(d\psi_m) = eppf(\boldsymbol{d};\alpha_0) \prod_{j=1}^{d} eppf(\boldsymbol{e}_j;\alpha) \prod_{m=1}^{M} P_0(d\psi_m). \qquad (11)$$

We note that Proposition 2 can be derived from Proposition 4 of Bassetti et al. (2018), as a special case of the larger class of species sampling models, and also that the so called pEPPF, introduced by Camerlenghi et al. (2018), can be obtained by (11) marginalizing with respect to η and $\boldsymbol{\psi}$. Moreover, according to (11), the multidimensional array $\boldsymbol{\theta}$ can be drawn as follows: first, choose the partitions ρ_j of the indices $\{1,\ldots,n_j\}$, for $j=1,\ldots,d$, independently and according to the Chinese restaurant process governed by α. After identifying the number of elements $K_j = k_j$ of ρ_j in each group, compute $T=\sum_{j=1}^{d} k_j$ and draw a partition η of the indices $\{1,\ldots,T\}$ according to a Chinese restaurant process governed by α_0. Sample $\boldsymbol{\psi}=(\psi_1,\ldots,\psi_M)$, with $M=\#\eta$, i.i.d. from P_0, and build $\boldsymbol{\theta}^*=(\theta_1^*,\ldots,\theta_T^*)$, 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 $\theta_{jh}^*=\theta_{t=t(j,h)}^*$. 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 $\boldsymbol{\theta}$ be a multidimensional array sampled according to a hierarchical NormCRM as in formula (8). The natural clustering is represented by the partition of the data corresponding to indices $\mathcal{I}=\{\mathcal{I}_1,\ldots,\mathcal{I}_M\}$ such that $(j,i)\in\mathcal{I}_m$ iff $\theta_{ji}=\psi_m$, with $\boldsymbol{\psi}=(\psi_1,\ldots,\psi_M)$ the vector of unique values among $\boldsymbol{\theta}$. On the other hand, if we have a sample $(\boldsymbol{\rho},\eta,\boldsymbol{\psi})$ from (11), a multidimensional array $\boldsymbol{\theta}$ can be recovered whose natural clustering is obtained by letting $\mathcal{I}_m=\mathcal{I}_m^{(\rho,\eta)}=\{(j,i): j\in\{1,\ldots,d\}, i\in C_{jh}, t(j,h)\in D_m\}$, and $\theta_{ji}=\psi_m$, for $(j,i)\in\mathcal{I}_m$ and $m=1,\ldots,M$. Hence, formula (11), under the assumption that P_0 is diffuse, characterizes the natural clustering induced by a sample $\boldsymbol{\theta}$ form a hierarchical NormCRM. The partition identified by $(\boldsymbol{\rho},\eta)$ 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 eppf$(\boldsymbol{e}_j;\alpha)$, 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 eppf$(\boldsymbol{d}; \alpha_0)$, 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 $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_M)$, 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 η). 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 α. It is well known that the distribution of K_j (see Pitman (2006)) can be recovered as

$$\mathbb{P}\left(K_j = k\right) = \frac{1}{k!} \sum_{e_{j1}+\cdots+e_{jk}=n_j} \binom{n_j}{e_{j1}, \ldots, e_{jk}} \text{eppf}\,(e_{j1}, \ldots, e_{jk}; \alpha), \quad k = 1, \ldots, n_j, \qquad (12)$$

where the last sum is over all the compositions of n_j into *k* parts, i.e., all positive integers such that $e_{j1} + \ldots + e_{jk} = n_j$. 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 α_0 on a sample of size *T*, can be written as

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

$$\mathbb{P}\left(M = m \,|\, T = t\right) = \frac{1}{m!} \sum_{d_1+\cdots+d_m=t} \binom{t}{d_1, \ldots, d_m} \text{eppf}\,(d_1, \ldots, d_m; \alpha_0), \quad m = 1, \ldots, t. \qquad (13)$$

A derivation of this formulas can also be found in Camerlenghi et al. (2018). In the sensitivity analysis, contained in Section 3.1 of 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 $\mu_j = \mathbb{E}(K_j), \sigma_j = \mathrm{Var}(K_j)$, and $\gamma_j = \mathbb{E}\,|K_j - \mu_j|^3$. Let $\mu = \sum_{j=1}^{d} \mu_j, \sigma^2 = \sum_{j=1}^{d} \sigma_j^2, \gamma = \sum_{j=1}^{d} \gamma_j$, and let $\Phi(x; \mu, \sigma^2)$ be the cdf of a Normal distribution with mean μ and variance σ^2. Then:

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

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 Tyurin (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: $c \frac{\gamma}{\sigma^3 \sqrt{d}} \leq c \frac{dH}{d^{3/2} \sigma_{min}^3} = \frac{c}{\sqrt{d}} \frac{H}{\sigma_{min}^3}$. 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(α, P), and let A be a measurable subset of Θ. 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

$$\mathbb{E}(P_1(A)^n) = \mathbb{E}(P(A)^{K_n}). \qquad (15)$$

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

$$\mathbb{E}(P_1(A)^{n_1} P_1(B)^{n2}) = \sum_{k_1=1}^{n_1} \sum_{k_2=1}^{n_2} P(A)^{k_1} P(B)^{k_2} g(k_1, k_2), \qquad (16)$$

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

$$g(k_1, k_2) = \frac{1}{k_1! k_2!} \sum_{e_{1,1}+\cdots+e_{1,k_1}=n_1} \sum_{e_{2,1}+\cdots+e_{2,k_2}=n_2} \binom{n_1}{e_{1,1}, \ldots, e_{1,k_1}} \times \binom{n_2}{e_{2,1}, \ldots, e_{2,k_2}} \text{eppf}(e_{1,1}, \ldots, e_{1,k_1}, e_{2,1}, \ldots, e_{2,k_2}). \qquad (17)$$

We refer to Section 1.4 of the Supplementary Materials for the proofs of (15) and (16). Using (17), we can recover some well-known results for the moments of a NormCRM.

Consider now $P_1, P_2 \mid P \overset{\text{iid}}{\sim}$ NormCRM(α, P), and $P \sim$ NormCRM(α_0, P_0), and let $A, B \subset \Theta$ measurable, we have the following:

$$\mathbb{E}(P_1(A)^{n_1}) = \mathbb{E}(P_0(A)^M)$$
$$\mathrm{Cov}(P_1(A), P_2(B)) = \eta_0 \left(P_0(A \cap B) - P_0(A)P_0(B)\right)$$
$$\mathrm{Cov}(P_1(A), P_1(B)) = (\eta_0 + \eta_1 - \eta_0\eta_1)\left(P_0(A \cap B) - P_0(A)P_0(B)\right),$$

where M is the number of *natural* clusters of a sample of size n_1 from a NormCRM(α, P) with just one group, $\eta_0 = \mathrm{eppf}(2; \alpha_0)$, and $\eta_1 = \mathrm{eppf}(2; \alpha)$. Therefore, the expression of the covariance of the measure across groups of observations is governed only by η_0, which is the probability of ties in a sample from P, i.e. the probability 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 Section 1.5 of the Supplementary Materials, we extend formula (16) to the hierarchical case, and use it to derive an expression for the coskewness, $\mathrm{CoSk}(P_1(A), P_2(A))$, 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 $(Y_{11}, \ldots, Y_{1n_1}, \ldots, Y_{d1}, \ldots, Y_{dn_d})$ be a set of observations split into d groups, each containing n_j observations, for $j = 1, \ldots, d$, with $\boldsymbol{Y}_j = (Y_{j1}, \ldots, Y_{jn_j})$ the vector of observations in the j-th group. We write a hierarchical NormCRM mixture model as follows:

$$\begin{aligned}
&\boldsymbol{Y}_j \mid \boldsymbol{\theta}_j \overset{\text{ind}}{\sim} \prod_{i=1}^{n_j} f(y_{ji} \mid \theta_{ji}), \qquad \text{for } j = 1, \ldots, d, \\
&\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jn_j}) \mid P_j \overset{\text{iid}}{\sim} P_j, \text{ for } j = 1, \ldots, d, \qquad (18) \\
&P_1, \ldots, P_d \mid P \overset{\text{iid}}{\sim} \mathrm{NormCRM}(\alpha, P), \\
&P \sim \mathrm{NormCRM}(\alpha_0, P_0).
\end{aligned}$$

The law of the multidimensional array $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ is assigned using the *I*-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

$$
\begin{aligned}
&(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_d) \mid \boldsymbol{\rho}, \eta, \boldsymbol{\psi} \sim \prod_{m=1}^{M} \prod_{(j,i) \in \mathcal{I}_m^{(\boldsymbol{\rho}, \eta)}} f(y_{ji} \mid \psi_m) \\
&\rho_j \mid \alpha \overset{\text{ind}}{\sim} \text{eppf}(e_{j1}, \ldots, e_{jK_j}; \alpha), \qquad (19) \\
&\eta \mid T, \alpha_0 \sim \text{eppf}(d_1, \ldots, d_M; \alpha_0), \\
&(\psi_1, \ldots, \psi_M) \mid M \overset{\text{iid}}{\sim} P_0,
\end{aligned}
$$

where the link between η and $\boldsymbol{\rho}$ is described by the transformation t defined in (10). Here, $\mathcal{I}_m^{(\boldsymbol{\rho}, \eta)} = \{(j,i) : j \in \{1, \ldots, d\}, i \in C_{jh}, t(j,h) \in D_m\}$ coincides with the natural clustering induced by $\boldsymbol{\theta}$, 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 α, α_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 $f(\cdot \mid \psi)$ (e.g., a conjugate prior). As for the choice of the Lèvy intensities α and α_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 $\boldsymbol{\rho}$ and η, and letting $\mathcal{I}_m = \mathcal{I}_m^{(\boldsymbol{\rho}, \eta)}$, for $m = 1, \ldots, M$.

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 $\boldsymbol{\psi}$, 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 α and α_0, see (5). In order to avoid the computation of such integrals, we resort to a standard approach for NormCRMs (see Lijoi and Prünster 2010). Specifically, we introduce a vector of auxiliary variables $\boldsymbol{U} = (U_1, \ldots, U_d, U_0)$, and consider only the integrand terms in formula (5), re-writing (11) as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\rho}, \eta, d\boldsymbol{u}) &= \left(\prod_{j=1}^{d} \operatorname{eppf}(\boldsymbol{e}_j; \alpha, u_j) du_j \right) \operatorname{eppf}(\boldsymbol{d}; \alpha_0, u_0) du_0 \\ &= \left(\prod_{j=1}^{d} \frac{u_j^{n_j - 1}}{\Gamma(n_j)} \mathrm{e}^{-\phi(u_j)} \prod_{h=1}^{K_j} c_{e_{jh}}(u_j) du_j \right) \frac{u_0^{T-1}}{\Gamma(T)} \mathrm{e}^{-\phi(u_0)} \prod_{m=1}^{M} c_{d_m}(u_0) du_0, \end{aligned} \tag{20}$$

where $\operatorname{eppf}(\cdot; \alpha, u_j)$, for $j = 1, \ldots, d$, and $\operatorname{eppf}(\cdot; \alpha_0, u_0)$ are the integrands in (5) after disintegration. We mention here that, conditionally upon the total mass $\mathcal{T}_j = \mu_j(\Theta)$ of the j-th unnormalized CRM, U_j is gamma-distributed with shape n_j and scale $\mathcal{T}_j$, while, conditionally upon the total number of groups of the I-clustering across sources T and the total mass $\mathcal{T}_0 = \mu_0(\Theta)$, the variable U_0 is gamma-distributed with shape T and scale $\mathcal{T}_0$.

We now describe the predictive distribution of a sample from a hierarchical NormCRM, conditionally to the auxiliary variables $\boldsymbol{U}$, 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 $\boldsymbol{\rho}$, the new (n_j+1)-th customer in the j-th restaurant will choose the h-th existing table with probability $P_{jh}^{(to)}$, or a new one with probability $P_j^{(tn)}$, given by:

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

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

(21)

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

$$P_m^{(do)} := \mathbb{P}((T+1) \in D_m \,|\, U_0 = u_0) \propto \frac{\text{eppf}\,(d_1,\ldots,d_m+1,\ldots,d_M;\alpha_0,u_0)}{\text{eppf}\,(d_1,\ldots,d_M;\alpha_0,u_0)} = \frac{1}{A_0}\frac{c_{d_m+1}(u_0)}{c_{d_m}(u_0)},$$

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

(22)

where $A_0 = \sum_{m=1}^{M} \frac{c_{d_m+1}(u_0)}{c_{d_m}(u_0)} + c_1(u_0)$, for $m = 1,\ldots,M$. If a new dish is selected to be served at the $(T+1)$-th table, its label ψ_{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 $\boldsymbol{U}$, the joint distribution of our model is

$$\mathcal{L}(\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_d \mid \boldsymbol{\rho},\eta,\boldsymbol{\psi},\boldsymbol{U})\mathcal{L}(\rho_1,\ldots,\rho_d \mid U_1,\ldots,U_d)\mathcal{L}(\eta \mid U_0)\prod_{m=1}^{M} P_0(d\psi_m)$$
$$=\prod_{m=1}^{M}\prod_{(j,i)\in\mathcal{I}_m^{(\boldsymbol{\rho},\eta)}} f(y_{ji}\mid\psi_m)\prod_{j=1}^{d}\mathrm{eppf}(e_{j1},\ldots,e_{jK_j};\alpha,u_j)\mathrm{eppf}(d_1,\ldots,d_M;\alpha_0,u_0)\prod_{m=1}^{M}P_0(d\psi_m).$$

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

- **Updates of $\boldsymbol{U}$ and $\boldsymbol{\psi}$:** By using the expression of $\mathrm{eppf}(\cdot;\alpha,u_j)$, for $j=1,\ldots,d$, and $\mathrm{eppf}(\cdot;\alpha_0,u_0)$, as well as the centering measure P_0, it is straightforward to see how the updates of $\boldsymbol{U}$ and $\boldsymbol{\psi}$ 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 Section 2.2 of the Supplementary Materials.
- **Updates of $(\boldsymbol{\rho},\eta)$:** The clustering parameters are updated by first marginalizing with respect to the vector of unique values $\boldsymbol{\psi}$, and then updating $(\boldsymbol{\rho},\eta)$ conditionally on $\boldsymbol{U}$, 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 $\boldsymbol{Y}$ and $(\boldsymbol{\rho}^{-ji},\eta^{-ji})$, 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

$$\begin{aligned}&\mathbb{P}(i\in C_{jh}^{-ji}, t(j,h)\in D_m^{-ji}\mid \boldsymbol{Y},\boldsymbol{\rho}^{-ji},\eta^{-ji},U_j=u_j)\\&\propto \mathcal{M}\left(y_{ji}\mid \boldsymbol{y}_{\mathcal{I}_m^{(\boldsymbol{\rho}^{-ji},\eta^{-ji})}}\right)\mathbb{P}(i\in C_{jh}^{-ji}, t(j,h)\in D_m^{-ji}\mid \boldsymbol{\rho}^{-ji},\eta^{-ji},U_j=u_j).\end{aligned} \tag{23}$$

In the latter, 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 the chosen dish can be either one of those already served or a new one. Here, $\mathcal{M}(y_{ji} \mid \boldsymbol{y}_{\mathcal{I}_m})$ is the predictive density of a parametric Bayesian model where the sampling model is $f(y \mid \theta)$, the prior is P_0, and the observations are all the data with index in $\mathcal{I}_m$, with $\mathcal{I}_{M^{-ji}+1}$ the empty set. Probability $\mathbb{P}(i \in C_{jh}^{-ji}, t(j,h) \in D_m^{-ji} \mid \boldsymbol{\rho}^{-ji}, \eta^{-ji}, U_j = u_j)$ is depicted 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 sit 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} such that $t = t(j,h)$. Conditionally on $\boldsymbol{Y}$ and $(\boldsymbol{\rho}^{-t}, \eta^{-t})$, the probability of assigning the t-th table to the m-th cluster is

$$\mathbb{P}(t \in D_m^{-t} \mid \boldsymbol{Y}, \boldsymbol{\rho}^{-t}, \eta^{-t}, U_0 = u_0) \propto \mathcal{M}\left(\boldsymbol{y}_t \mid \boldsymbol{Y}_{\mathcal{I}_m^{(\boldsymbol{\rho}^{-t}, \eta^{-t})}}\right) \mathbb{P}(t \in D_m^{-t} \mid \boldsymbol{\rho}^{-t}, \eta^{-t}, U_0 = u_0), \tag{24}$$

for $m = 1, \ldots, M^{-t} + 1$, where $(M^{-t} + 1)$ indicates the new cluster of tables. Here, $\boldsymbol{Y}_t := \{Y_{ij} : i \in C_{jh}, t = t(j,h)\}$ is the set of all the observations (customers) in the l-cluster C_{jh} (i.e., the t-th table), and $\mathcal{M}(\boldsymbol{y}_t \mid \boldsymbol{y}_{\mathcal{I}_m})$ 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 $\mathbb{P}(t \in D_m^{-t} \mid \boldsymbol{\rho}^{-t}, \eta^{-t}, U_0 = u_0)$ 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 Section 2.4 of the Supplementary Materials.

3.3 Prediction

Consider a configuration $(\boldsymbol{\rho}, \eta, \boldsymbol{\psi})$, and hence $\boldsymbol{\theta}$, of n observations in d groups, as well as a new dish label indicated as ψ_{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:

$$\mathbb{P}(\theta_{j(n_j+1)} = \psi_m, (n_j+1) \in C_{jh} \mid \boldsymbol{\rho}, \eta, \boldsymbol{\psi}) = \begin{cases} P_{jh}^{(to)} & h = 1, \ldots, K_j, \quad m = t(j,h) \\ P_j^{(tn)} P_m^{(do)} & h = K_j + 1, \quad m = 1, \ldots, M \\ P_j^{(tn)} P^{(dn)} & h = K_j + 1, \quad m = M+1 \end{cases} \tag{25}$$

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:

$$\mathbb{P}(\theta_{(d+1)1} = \psi_m \mid \boldsymbol{\rho}, \eta, \boldsymbol{\psi}) = \begin{cases} P_m^{(do)} & h = 1 \text{ and } m = 1, \ldots, M, \\ P^{(dn)} & h = 1 \text{ and } m = M+1. \end{cases} \qquad (26)$$

We now derive the law of a new observation given the multidimensional array of data $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_d)$ for the mixture in model (19). It is easy to see that, conditionally on $(\boldsymbol{\rho}, \eta, \boldsymbol{\psi})$, this quantity changes depending on the group. In particular, the predictive densities in an existing or in a new group are:

$$\begin{aligned} & p(y_{j(n_j+1)} \mid \boldsymbol{Y}, \boldsymbol{\rho}, \eta, \boldsymbol{\psi}) \\ & = \sum_{h=1}^{K_j} P_{jh}^{(to)} f(y_{j(n_j+1)} \mid \theta_{jh}) + P_j^{(tn)} \sum_{m=1}^{M} P_m^{(do)} f(y_{j(n_j+1)} \mid \psi_m) + P_j^{(tn)} P^{(dn)} \int_{\Theta} f(y_{j(n_j+1)} \mid \psi) P_0(d\psi), \\ & p(y_{(d+1)1} \mid \boldsymbol{Y}, \boldsymbol{\rho}, \eta, \boldsymbol{\psi})) = \sum_{m=1}^{M} P_m^{(do)} f(y_{(d+1)1} \mid \psi_m) + P^{(dn)} \int_{\Theta} f(y_{(d+1)1} \mid \psi) P_0(d\psi), \end{aligned} \qquad (27)$$

where the marginal distribution $\mathcal{M}(y) = \int_{\Theta} f(y \mid \psi) P_0(d\psi)$ 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 $(\boldsymbol{\rho}, \eta, \boldsymbol{\psi})$, 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), $P_m^{(do)} = 0$, 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 $\mathcal{M}(y)$.

4 Applications

In this section, we assess the performance of the proposed mixture model on both simulated and benchmark datasets. We focus on the hierarchical NGG (HNGG) mixture model, obtained from model (19) when the Lévy intensities are $\alpha(s) = \frac{\kappa}{\Gamma(1-\sigma)} s^{-1-\sigma} e^{-s} \mathbb{I}_{(0,+\infty)}(s)$, and $\alpha_0(s) = \frac{\kappa_0}{\Gamma(1-\sigma_0)} s^{-1-\sigma_0} e^{-s} \mathbb{I}_{(0,+\infty)}(s)$, 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 application in Section 4.2). 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 Section 3 of 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 $(P_1, \ldots, P_d) \overset{\text{iid}}{\sim} NGG(\kappa_0, \sigma_0, P_0)$. 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 $i = 1, \ldots, 100$:

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

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 the 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 (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 $(m_0, k_0, a_{\tau^2}, b_{\tau^2}) = (0.25, 0.62, 2.07, 0.66)$, a specification that corresponds to $\mathbb{E}[\mu] = \hat{y}_n, \mathbb{E}[\tau^2] = \hat{s}_n^2 / 3, \mathrm{Var}(\mu) = 1, \mathrm{Var}(\tau^2) = 5$, 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 Section 3.4 of 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 large variances yield poor clustering performance in terms of RI, they can help improve the model fitting in terms of 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 Section 3.3 of the Supplementary Materials. In general, the best LPML values are observed when σ or σ_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 σ or σ_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 σ and σ_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 Sections 3.1 and 3.2 of 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 (κ_0, σ_0) also increases the dependency between measures, while an opposite trend is observed when increasing (κ, σ). 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 $(P_1, \ldots, P_d) \overset{\text{iid}}{\sim} \text{NGG}(\kappa_0, \sigma_0, P_0)$ 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 $\mathcal{M}(y)$, 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 $\sigma, \sigma_0 \sim Beta(2,18)$ ($\mathbb{E}[\sigma] = \mathbb{E}[\sigma_0] = 0.1$). 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 inference, 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 the histograms of the posterior samples of σ and σ_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 Section 4 of 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 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 Section 5 of 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). 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:
$P_0(\mu, \tau^2) = p(\mu)\,p(\tau^2) = \mathrm{N}\left(\mu \mid 50, 25\right)\text{inv-gamma}(\tau^2 \mid 0.5, 50)$, where the hyperparameters are set as in Hoff (2009), chap. 8. In order to allow for more robustness in the inference process, we impose prior distributions $\kappa, \kappa_0 \sim \text{gamma}(1,1)$ 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 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 ρ_j, i.e. *l*-clusters, for $j = 1, \ldots, 100$. We observe some heterogeneity, 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 ones (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, specifically according to the natural clustering estimated via the Binder's loss. The intensity of the grey scale for the predictive densities increases with the posterior expected number of 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 as in Hoff (2009) (Chapter 8). In Figure 6(e) the predictive densities under the parametric model are reported for a selection of the 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 Section 4 of the Supplementary Materials.

5 Conclusion

In this paper, we have conducted a thorough investigation of the clustering induced by a NormCRM mixture model. This model is suitable for data belonging to groups or categories that share similar characteristics. 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 desirable feature allows to cluster together observations of different groups. By integrating out

the nonparametric components of our prior (i.e. $P_1, \cdots, P_d, P$), 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 identified by the eppf induced by the NormCRM (α, 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 (α_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 has the potential to be generalized. For example, an interesting future direction is to investigate extensions to situations where covariates are available, following either the approach of MacEachern (1999), via dependent nonparametric processes, or the product partition model approach of Müller and Quintana (2010).

Acknowledgements

Raffaele Argiento gratefully acknowledges *Collegio Carlo Alberto* for partially funding this work.

Andrea Cremaschi thanks the Norway Centre for Molecular Medicine (NCMM) IT facility for the computational support.

SUPPLEMENTARY MATERIAL

Title: Supplementary Materials The file:
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 compute 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 of Statistics*, 10(2):3516–3547.

Argiento, R., Cremaschi, A., and Guglielmi, A. (2014). A "density-based" algorithm for cluster analysis using species sampling Gaussian mixture models. *Journal of Computational and Graphical Statistics*, 23(4):1126–1142.

Bassetti, F., Casarin, R., and Rossini, L. (2018). Hierarchical species sampling models. *arXiv preprint arXiv:1803.05793*.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2018). Distribution theory for hierarchical processes. *The Annals of Statistics, to appear.*

Camerlenghi, F., Lijoi, A., and Prünster, I. (2017). Bayesian prediction with multiple-samples information. *Journal of Multivariate Analysis*, 156:18–28.

Durrett, R. (1991). *Probability: Theory and Examples*. Pacific Grove, CA: Wadsworth & Brooks/Cole.

Favaro, S. and Teh, Y. (2013). MCMC for normalized random measure mixture models. *Statistical Science*, 28(3):335–359.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics*, pages 287–302. Elsevier.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer Verlag.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Ishwaran, H. and James, L. F. (2003). Generalized weighted chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, 13(4):1211–1235.

James, L. F., Lijoi, A., and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36(1):76–97.

Kallenberg, O. (2005). *Probabilistic Symmetries and Invariance Principles*. Springer Science & Business Media.

Kingman, J. F. C. (1993). *Poisson Processes*, volume 3. Oxford university press.

Lau, J. W. and Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3):526–558.

Lijoi, A., Mena, R. H., and Prünster, I. (2007). Controlling the reinforcement in bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):715–740.

Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In Hjort, N., Holmes, C., Müller, P., and Walker, editors, *In Bayesian Nonparametrics*, pages 80–136. Cambridge University Press.

Lo, A. Y. (1984). On a class of bayesian nonparametric estimates: I. density estimates. *The annals of statistics*, 12(1):351–357.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pages 50–55.

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2017). Identifying mixtures of mixtures using bayesian estimation. *Journal of Computational and Graphical Statistics*, 26(2):285–295.

Müller, P. and Quintana, F. (2010). Random partition models with regression on covariates. *Journal of Statistical Planning and Inference*, 140(10):2801–2808.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Pinheiro, J. and Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. *Lecture Notes-Monograph Series*, pages 245–267.

Pitman, J. (2003). Poisson-Kingman partitions. In *Science and Statistics: a Festschrift for Terry Speed*, volume 40 of *IMS Lecture Notes-Monograph Series*, pages 1–34. Institute of Mathematical Statistics, Hayward (USA).

Pitman, J. (2006). *Combinatorial Stochastic Processes*. LNM n. 1875. Springer, New York.

Regazzini, E., Lijoi, A., Prünster, I., et al. (2003). Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31(2):560–585.

Teh, Y. W. and Jordan, M. I. (2010). Hierarchical bayesian nonparametric models with applications. volume 1, pages 158–207. Camb. Ser. Stat. Probab. Math.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Tyurin, I. S. (2010). An improvement of upper estimates of the constants in the lyapunov theorem. *Russian Mathematical Surveys*, 65(3):201–202.

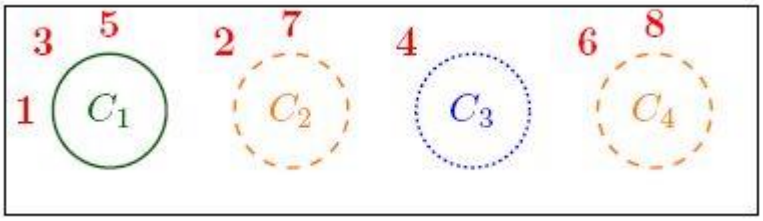

Fig. 1 Illustration of an *I*-clustering based on a sample of dimension $n = 8$ from a NormCRM whose centering measure is a discrete distribution on the colored lines.

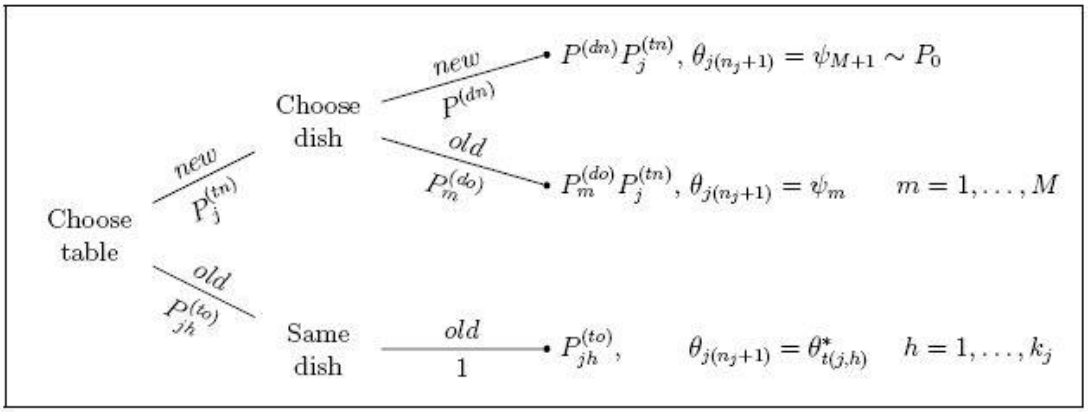

Fig. 2 Probability tree of the Chinese restaurant franchise process for a hierarchical NormCRM

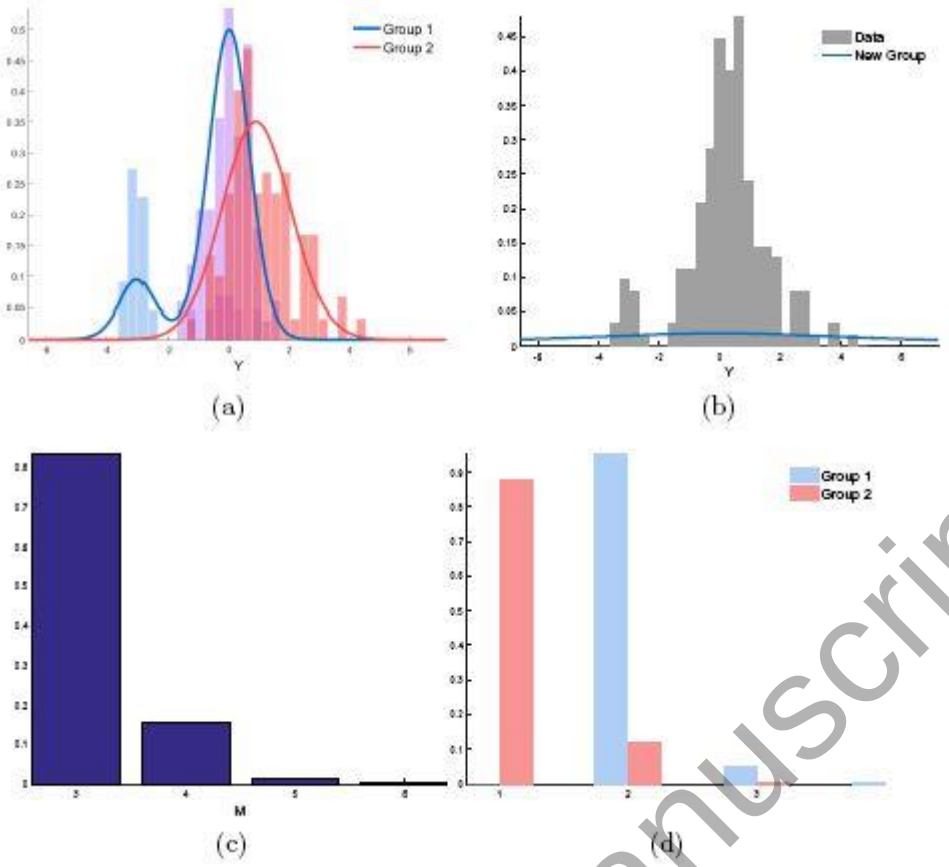

Fig. 3 Simulated Data ($d = 2$) – Summary plots for the independent NGG case with $\sigma_0 = 0.1$ and $\kappa_0 = 0.1$. **(a)**: Posterior density estimates and histograms of the data, colored according to the true partition. **(b)**: Predictive density for a new group. Posterior distribution of the number of clusters for all observations, M **(c)**, and in each group **(d)**.

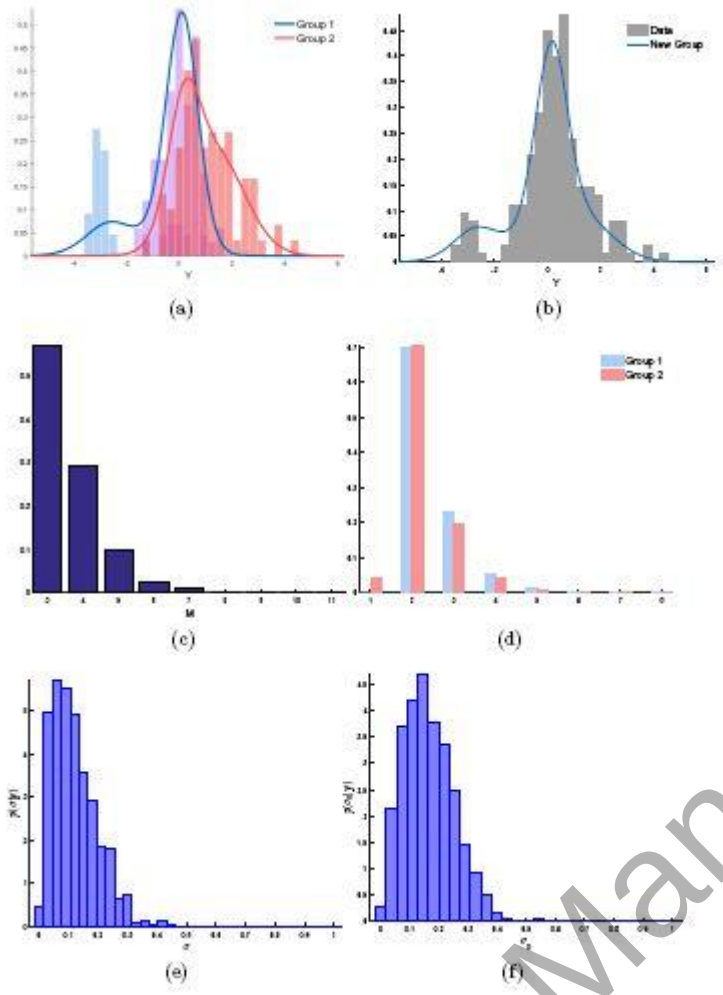

Fig. 4 Simulated Data ($d = 2$) – Summary plots for the HNGG model with $\sigma, \sigma_0 \sim \text{Beta}(2,18)$ and $\kappa = \kappa_0 = 0.1$. **(a)**: Posterior density estimates and histograms of the data, colored according to the true partition. **(b)**: Predictive density for a new group. Posterior distribution of the number of clusters for all observations, M **(c)**, and in each group **(d)**. **(e,f)**: Posterior distributions for the parameters σ and σ_0.

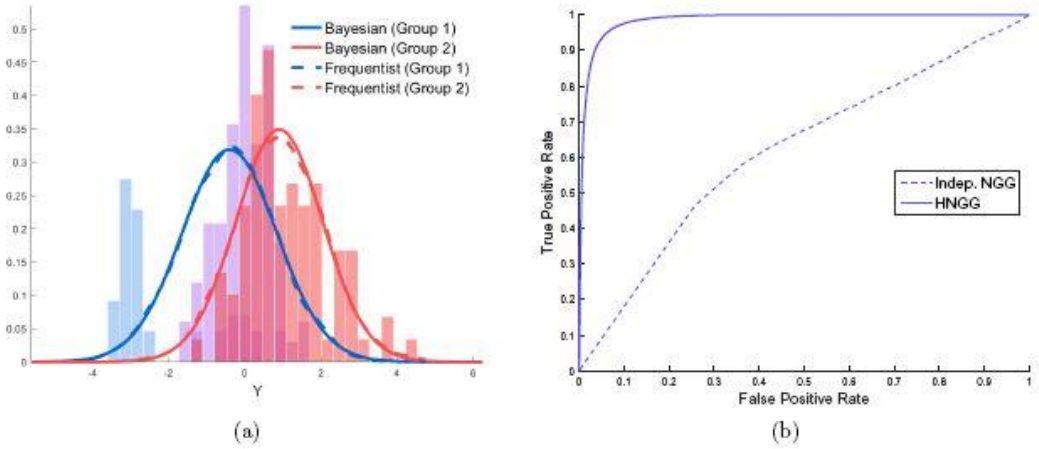

Fig. 5 Simulated Data – **(a)**: Density estimation for the simulation study ($d = 2$) under Bayesian and frequentist parametric models. **(b)**: ROC curves for the simulation study with $d = 100$ groups, averaged over 25 replicated datasets.

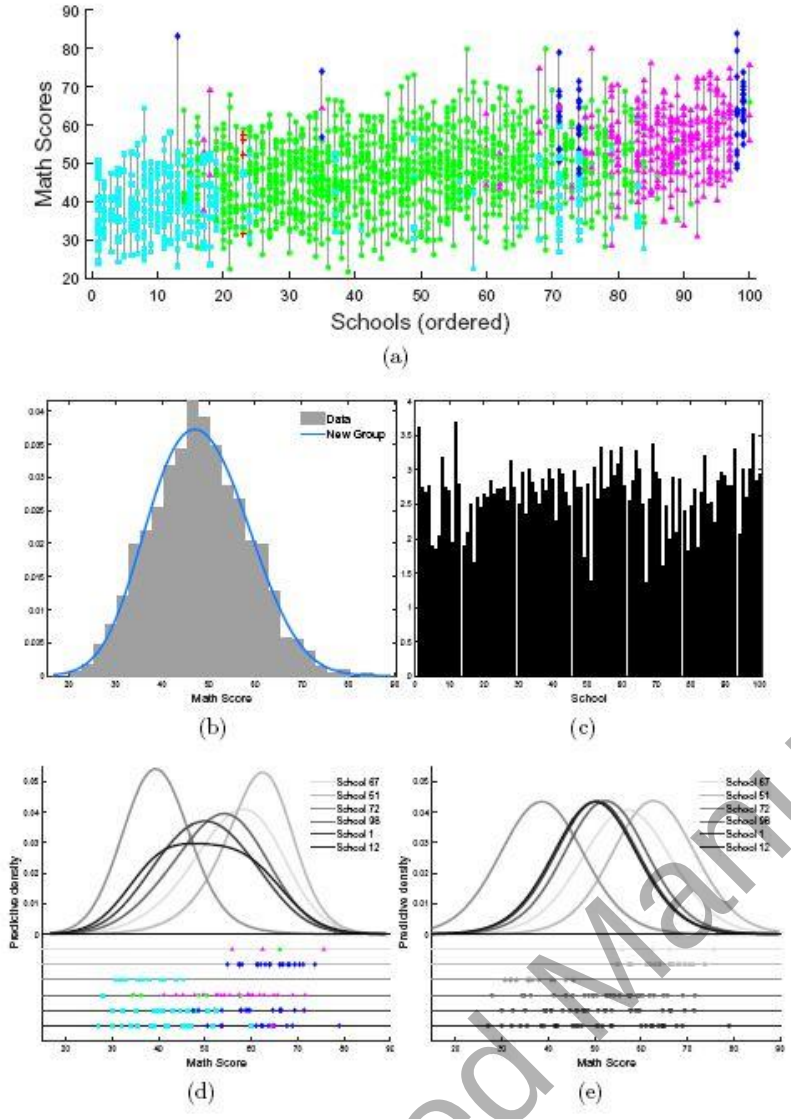

Fig. 6 School data by Hoff (2009) – **(a)**: Data sorted by increasing sample mean in each school (vertical lines). Colors and marker shapes identify the estimated clustering. **(b)**: Predictive density of the math score for a student in a new school. **(c)**: Posterior mean of the number of elements in ρ_j, i.e. *I*-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 *I*-clusters in the schools. **(e)**: Predictive densities for the same schools are in panel (d), estimated under the Bayesian parametric model of Hoff (2009). The gray scale of the points reflects the posterior expected number of *I*-clusters in the schools under the HNGG model.

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

	$\sigma = \sigma_0 \approx 0$			$(\sigma, \sigma_0) = (0.3, 0.1)$			$(\sigma, \sigma_0) = (0.1, 0.3)$		
	$\kappa_0 = 0.1$	$\kappa_0 = 1$	$\kappa_0 = 10$	$\kappa_0 = 0.1$	$\kappa_0 = 1$	$\kappa_0 = 10$	$\kappa_0 = 0.1$	$\kappa_0 = 1$	$\kappa_0 = 10$
$\kappa = 0.1$	-0.2743	-0.2864	-0.2739	-0.2835	-0.2751	-0.2754	-0.2758	-0.2786	-0.2770
$\kappa = 1$	-0.2826	-0.2812	**-0.2705**	-0.2753	-0.2708	-0.2727	-0.2760	-0.2774	**-0.2696**
$\kappa = 10$	-0.2775	-0.2740	-0.2734	-0.2723	**-0.2698**	-0.2725	-0.2706	0.2712	0.2752

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 $(\kappa, \kappa_0, \sigma, \sigma_0)$.

	$\sigma = \sigma_0 \approx 0$			$(\sigma, \sigma_0) = (0.3, 0.1)$			$(\sigma, \sigma_0) = (0.1, 0.3)$		
	$\kappa_0 = 0.1$	$\kappa_0 = 1$	$\kappa_0 = 10$	$\kappa_0 = 0.1$	$\kappa_0 = 1$	$\kappa_0 = 10$	$\kappa_0 = 0.1$	$\kappa_0 = 1$	$\kappa_0 = 10$
$\kappa = 0.1$	0.5840	0.5079	**0.7568**	0.5194	0.4958	0.4885	0.4952	0.5048	0.4958
	(2)	(2)	**(1)**	(2)	(2)	(6)	(2)	(2)	(2)
κ = 1	0.5291	0.5079	0.4317	**0.5404**	0.4380	0.4083	0.4590	**0.5222**	0.4115
	(2)	(2)	(4)	**(2)**	(7)	(14)	(8)	**(4)**	(9)
κ = 10	0.4968	0.3844	0.3319	0.5048	0.3519	0.2749	0.3325	0.3513	0.3123
	(2)	(8)	(14)	(2)	(12)	(19)	(15)	(15)	(23)