

Conditional partial exchangeability: a probabilistic framework for multi-view clustering

Beatrice Franzolini*

Bocconi Institute for Data Science and Analytics, Bocconi University
and

Maria De Iorio[†] and Johan Eriksson

Yong Loo Lin School of Medicine, National University of Singapore,
Institute for Human Development and Potential, A*STAR

Abstract

Standard clustering techniques assume a common configuration for all features in a dataset. However, when dealing with multi-view or longitudinal data, the clusters' number, frequencies, and shapes may need to vary across features to accurately capture dependence structures and heterogeneity. In this setting, classical model-based clustering fails to account for within-subject dependence across domains. We introduce conditional partial exchangeability, a novel probabilistic paradigm for dependent random partitions of the same objects across distinct domains. Additionally, we study a wide class of Bayesian clustering models based on conditional partial exchangeability, which allows for flexible dependent clustering of individuals across features, capturing the specific contribution of each feature and the within-subject dependence, while ensuring computational feasibility.

Keywords: Bayesian nonparametrics, Dynamic clustering, Hierarchical processes, Partial exchangeability, Random partitions, Unsupervised learning

*Franzolini was partially supported by the National Recovery and Resilience Plan of Italy (PE1 FAIR—CUP B43C22000800006).

[†]This work was supported by the Singapore Ministry of Education Academic Research Fund Tier 2 under Grant MOE-T2EP40121-0004.

1 Clustering multi-view information

Clustering is arguably the most famous unsupervised learning technique. It involves grouping observations into clusters based on their *similarities*. Standard clustering techniques assume a common clustering configuration of subjects across all features observed in a sample. However, given the complexity and dimension of modern datasets, a unique clustering arrangement for all the features is often inadequate to describe the structure and the heterogeneity in the population under study. For instance, in longitudinal data analysis, the underlying clustering structure of individuals is likely to change over time; in multi-view data (see, for instance, [Yang & Wang 2018](#)) multivariate information collected across distinct domains may require clusters’ shapes and definitions to change from feature to feature. In this setting, a unique clustering configuration based on all the observed features not only may be hard to detect (often leading to clusters of small size to accommodate heterogeneity in multi-dimensional spaces, cf., [Chandra et al. 2023](#)), but will also mainly capture global patterns, down-weighting the idiosyncratic contribution of each feature. Moreover, it depends on features’ dimensions, favoring higher dimensional features as more important in explaining the heterogeneity across subjects. See Figure 1 for a toy example illustrating the latter problem. In this work, we focus on clustering problems where multi-view or longitudinal information is available for the same subjects, and we allow the underlying clustering structure to change across features/time.

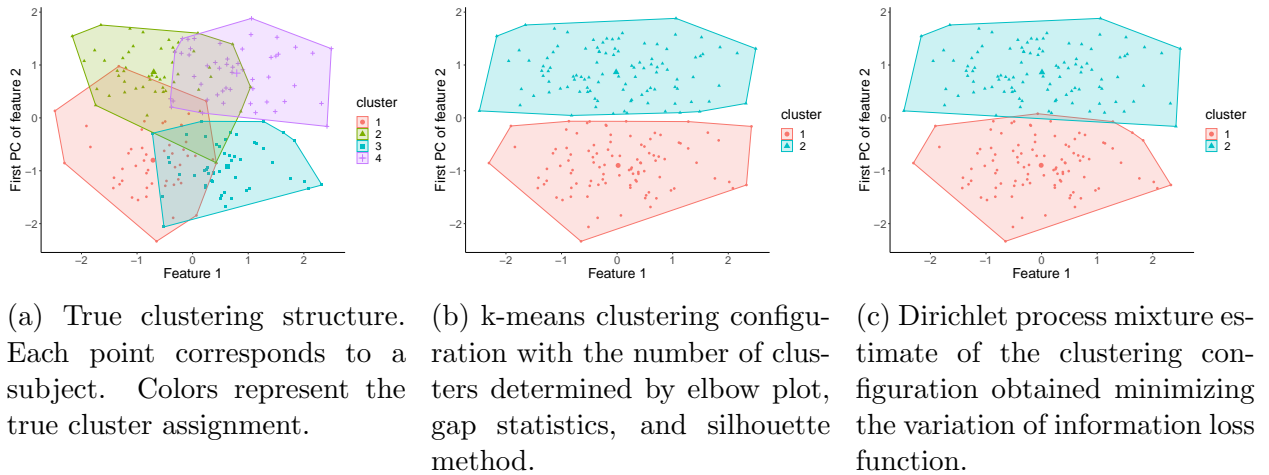


Figure 1: Toy example. Data was simulated for two features and 200 subjects, with the true clustering configuration displayed in panel (a). The first feature (x-axis) is sampled from a univariate Normal with unitary variance and mean equal either to 1 or -1, depending on the true cluster assignment. The second feature (y-axis) is sampled from a three-variate Normal with identity covariance matrix and mean equal to either (1, 1, 1) or (-1, -1, -1) depending on the true cluster assignment; it is represented via its first principal component on the y-axis. The Dirichlet process estimate in panel c) is obtained with a Multivariate Normal kernel with identity covariance matrix, standard Multivariate Normal base measure, and concentration parameter equal to 0.1. Both the clustering configurations obtained with k-means (panel b) and the Dirichlet process mixtures (panel c) are heavily informed by the second feature and ignore the information contained in the first feature.

The two main approaches for clustering are model-based and algorithmic methods. Model-based methods rely on distributional assumptions about the underlying data-generating mechanism of the observations in each cluster, leading to a mixture model. Unlike algorithmic techniques, model-based methods explicitly define the shape of a cluster in terms of

probability distribution functions. Most popular model-based approaches include Bayesian infinite mixture models (Ferguson 1983, Lo 1984, Barrios et al. 2013) and Bayesian mixtures with a random number of components (Nobile 1994, De Blasi et al. 2015, Miller & Harrison 2018, Argiento & De Iorio 2022). They allow for data-driven automatic selection of the number of clusters for which no finite upper bound has to be fixed. Most traditional clustering approaches (both model-based and algorithmic) are designed for single-view data and aim at detecting a unique clustering configuration of individuals in a sample. In recent years, a wealth of proposals for algorithms to integrate multi-view information has appeared in the machine learning literature (see, Yang & Wang 2018, Chen et al. 2022, for comprehensive reviews of the topic). Nonetheless, such methods, while recognizing the multi-view nature of the data, provide again a single clustering configuration common to all the features, which may still fail to highlight the complementary information of each feature (Yao et al. 2019). An interesting exception is provided by the algorithm proposed by Yao et al. (2019). In the Bayesian clustering literature, the focus is often placed on *multi-sample data*, rather than *multi-view data*, in the sense that there is an initial natural grouping of the subjects (for example, based on treatment groups in a clinical trial, or some level of a particular covariate) which is treated as deterministic and there is no overlap of subjects across groups. Then, clustering is performed within each group, with clusters possibly shared among groups. These models are obtained by inducing dependence between the group-specific random probability measures in the underlying mixture model and they have been the object of extensive research in recent decades (see, MacEachern 2000, Müller et al. 2004, Teh et al. 2006, Caron et al. 2007, Dunson & Park 2008, Ren et al. 2008, Dunson 2010, Rodríguez et al. 2010, Taddy 2010, Rodríguez & Dunson 2011, Lijoi et al. 2014, Foti & Williamson 2015, Caron et al. 2017, Griffin & Leisen 2017, DeYoreo & Kottas 2018, Argiento et al. 2020, Bassetti et al. 2020, Ascolani et al. 2021, Beraha et al. 2021, Denti et al. 2021, Zhou et al. 2021, Quintana et al. 2022, Lijoi et al. 2023). Models built with this strategy may be effectively employed for clustering multi-sample data, i.e., when different clustering configurations refer to disjoint sets of subjects. However, we note that these are not suitable for multi-view data. As we show in this work, when they are applied to cluster multi-view or longitudinal data, such methods focus on marginal inference based on each feature and fail to capture the true nature of the multivariate dependence. In particular, in Section 2, we show how this is a consequence of the fact that they disregard subjects’ identities, i.e., that subjects are indeed the same for all the observed features or times.

The Bayesian literature on clustering methods for longitudinal or multi-view information is rather limited. In this context, the core challenge is to define a probabilistic model able to account for subjects’ identities across multiple features. Bayesian clustering approaches that allow to both preserve subjects’ identity and provide multiple clustering configurations, appear limited to the following specific models: the hybrid Dirichlet process (Petrone et al. 2009), the enriched Dirichlet process (Wade et al. 2011), the separately exchangeable random partition models in Lee et al. (2013) and Rebaudo et al. (2021) and the temporal random partition model of Page et al. (2022). Even though these models are quite different in nature, they all belong to the novel probabilistic framework we introduce here, which also enables gaining new insights about these existing approaches.

The main contribution of this work is the introduction of *conditional partial exchangeability* (CPE), a modeling principle for multi-view and dynamic probabilistic clustering. CPE is a condition imposed on the conditional law of the observable in multi-view data, inducing dependence across distinct clustering configurations of the same subjects while preserving their identities. An additional contribution is developing a specific class of mixture models that satisfy CPE, which we refer to as *telescopic clustering models*. The introduction of these models further highlights the utility of CPE. We show that these models are analytically and

computationally tractable and establish Kolmogorov consistency of the distribution of the observable. Finally, we investigate two models within this class in more detail: one with an infinite number of components and another with a random number of components. We provide algorithms for posterior estimation and showcase the performance of the proposed framework on real and simulated data.

2 Conditional partial exchangeability

Let (X_{1i}, X_{2i}) , be features on the i -th observational unit, with $i = 1, \dots, n$. For simplicity of explanation, we partition the feature vector into two sub-components and discuss how to extend to a number L of components in Section 3.3: $X_{1i} \in \mathbb{X}_1 \subset \mathbb{R}^d$ is the observation recorded at layer 1, which can represent, for example, either a vector of *primary* features or observations corresponding to the initial time point $t = 1$, and $X_{2i} \in \mathbb{X}_2 \subset \mathbb{R}^p$ is the observation recorded at layer 2, which can refer to either a vector of *secondary* features or observations corresponding to a subsequent time point $t = 2$. Importantly, the support spaces \mathbb{X}_1 and \mathbb{X}_2 are not assumed to coincide. In particular, the dimensions d and p may be different. Our goal is to estimate two clustering configurations ρ_1 and ρ_2 , which correspond to the first and second layers, respectively, allowing for dependence between the two clustering configurations and employing a learning mechanism that takes into account subjects' identity thus capturing multivariate dependence. The partition ρ_j , $j = 1, 2$, can be represented by the vector $\mathbf{c}_j = (c_{j1}, \dots, c_{jn})$ of subject-specific allocation variables, whose elements take value in the set $[n] := \{1, 2, \dots, n\}$ and are such that $c_{ji} = c_{jl}$ if and only if subjects i and l belong to the same cluster at layer j . As typically done, we assume exchangeability for the joint observations $(X_{1i}, X_{2i})_{i \geq 1}$, i.e., $\mathbb{P}[(X_{1i}, X_{2i})_{i=1}^n \in A] = \mathbb{P}[(X_{1\sigma(i)}, X_{2\sigma(i)})_{i=1}^n \in A]$ for any σ permutation of n elements, $n \geq 1$, and measurable set $A \subseteq (\mathbb{X}_1 \times \mathbb{X}_2)^n$. A review of key preliminaries on exchangeable partitions and partial exchangeability is provided in Section S1 of the Supplement. For the sake of clarity, assume that the clustering configurations fully capture the dependence structure between first and second layers features, i.e., $(X_{11}, \dots, X_{1n}) \perp (X_{21}, \dots, X_{2n}) \mid \rho_1, \rho_2$. The latter is a common assumption in clustering models for multivariate data (see, e.g., Rogers et al. 2008, Kumar et al. 2011, Lock & Dunson 2013, Gao et al. 2020, Franzolini, Cremaschi, van den Boom & De Iorio 2023) as it often avoids identifiability issues. Nonetheless, it can be relaxed.

Exchangeability of the bivariate observations does not imply *conditional exchangeability* of one layer given the other, which is instead undesirable in a multi-view setting. To better understand this point, let us see what happens if we do assume exchangeability for observations in the second layer conditionally on the first-layer partition, ρ_1 . Formally, $(X_{2i})_{i \geq 1}$ is conditionally exchangeable with respect to ρ_1 , if, for any realization of ρ_1 , for any $n \geq 1$ and for any permutation σ ,

$$p(X_{21}, \dots, X_{2n} \mid \rho_{1n}) = p(X_{2\sigma(1)}, \dots, X_{2\sigma(n)} \mid \rho_{1n}) \quad (1)$$

where p denotes a joint density function (to be understood as a mass function in the discrete case). This implies, for instance, that the joint distribution of a pair of second-layer observations is invariant to their clustering allocations at layer 1, in the sense that, for any set of three subjects i, j and k , $p(X_{2i}, X_{2j} \mid c_{1i} = c_{1j} \neq c_{1k}) = p(X_{2i}, X_{2k} \mid c_{1i} = c_{1j} \neq c_{1k})$. Thus, under conditional exchangeability, knowing that subjects i and j belong to the same cluster at layer 1 does not provide any information at layer 2 specific to those same two subjects, i.e., the model does not preserve subjects' identities. This unusual behavior of the learning mechanism under *conditional exchangeability* arises because this assumption restricts the transfer of subject-level information (e.g., which subjects belong to the same cluster in ρ_1)

between layers. Instead, only population-level information, such as the number of clusters and their frequencies in ρ_1 , is propagated to the next layer. (cf., [Page et al. 2022](#)). However, this assumption is also at the core of the majority of dependent Bayesian clustering methods (see, [Quintana et al. 2022](#), for a recent review), which are thus appropriate for multi-sample data but cannot be effectively applied in longitudinal and multi-view settings. An alternative to *conditional exchangeability* (of the second layer given ρ_1) is offered in the Bayesian nonparametric literature by the Enriched Dirichlet process ([Wade et al. 2011](#)), which in contrast induces *conditional independence*, i.e., given ρ_1 , observations at the second layer are assumed exchangeable if they belong to the same first-layer cluster and independent otherwise. However, enriched constructions do not allow to define a prior for (ρ_1, ρ_2) with full support because they force second-layer clusters to be nested within first-layer clusters: if two items are assigned to distinct clusters at layer 1, they cannot be assigned to the same cluster at layer 2.

To define a flexible and general learning mechanism for Bayesian clustering of multi-view or longitudinal data, clusters defined by ρ_1 should be treated neither as almost irrelevant as under *conditional exchangeability* nor as too informative as under the *conditional independence* of enriched constructions. Ideally, an appropriate learning mechanism would a priori favor at layer 2 a clustering configuration similar to layer 1, but not necessarily nested. A general framework to achieve this goal is currently missing and it is the main contribution of this work. To this end, we introduce *conditional partial exchangeability* (CPE), formalized by the following definition and discussed thereafter. Given a partition ρ_1 and the corresponding subject-specific allocation variables \mathbf{c}_1 , let us denote with $\mathcal{P}(n; \rho_1)$ the set of permutations of $[n]$ that preserve ρ_1 , i.e. $\sigma \in \mathcal{P}(n; \rho_1)$ if and only if σ is a permutation of n elements such that $c_{1\sigma(i)} = c_{1i}$, for any $i \in [n]$.

Definition 1 (Conditional partial exchangeability). *Given a (marginally) exchangeable sequence $(X_{2i})_{i \geq 1}$ and a collection of coherent¹ random partitions $(\rho_{1n})_{n \geq 1}$, where ρ_{1n} is a partition of $[n]$, $(X_{2i})_{i \geq 1}$ is said to be conditionally partially exchangeable (CPE) with respect to $(\rho_{1n})_{n \geq 1}$ if and only if, for any $n \geq 1$, for any realization of ρ_{1n} , and for any $(i_1, \dots, i_\ell), (j_1, \dots, j_\ell) \subset [n]$, the following two conditions are satisfied*

$$\text{c-i)} \quad p(X_{21}, \dots, X_{2n} \mid \rho_{1n}) = p(X_{2\sigma(1)}, \dots, X_{2\sigma(n)} \mid \rho_{1n}), \quad \text{for any } \sigma \in \mathcal{P}(n; \rho_{1n});$$

$$\text{c-ii)} \quad p(X_{2i_1}, \dots, X_{2i_\ell} \mid c_{1i_1} = \dots = c_{1i_\ell}) = p(X_{2j_1}, \dots, X_{2j_\ell} \mid c_{1j_1} = \dots = c_{1j_\ell}).$$

To fully understand Definition 1, we need first to consider the fundamental differences between CPE (introduced here) and partial exchangeability (as presented by [de Finetti 1938](#), and quickly reviewed in Section S1 of the Supplement). Importantly, partial exchangeability is a condition on the marginal distributions of a sequence of observations, thereby defining a class of sequences. In contrast, CPE specifies a condition on the dependence between a sequence of exchangeable observations and a random partition, thereby defining a class of dependence relationships. Moreover, considering the conditions required for the conditional law of the observations, CPE differs from partial exchangeability not only in its conditional nature, imposed by condition (c-i), but also in its requirement for marginal invariance, imposed by condition (c-ii). Condition (c-ii) entirely lacks an analog in the definition of partial exchangeability. For a more intuitive understanding of CPE: assume the observations are partitioned according to a clustering configuration at layer 1. Conditioning on this layer-1 partition, the observations exhibit partial exchangeability at layer 2, i.e., in layer 2, exchangeability holds among observations that are co-clustered in layer 1, but

¹the collection of partitions $(\rho_{1n})_{n \geq 1}$ is said to be coherent if for any n , ρ_{1n} can be obtained from ρ_{1n+1} by removing the object $n+1$.

not necessarily across different layer-1 clusters (c-i). Moreover, CPE requires that groups of observations corresponding to different layer-1 clusters preserve the same within-cluster marginal distribution at the subsequent layer (c-ii), e.g., $p(X_{21} \mid c_{11}) = p(X_{22} \mid c_{12})$ and $p(X_{21}, X_{22} \mid c_{11} = c_{12}) = p(X_{23}, X_{24} \mid c_{13} = c_{14})$.

For the sake of notation, in the following, we omit the subscript n when denoting the partition. The next theorem shows that CPE allows the preservation of subjects' identities when moving from one layer to another.

Theorem 1 (Subjects' identity across layers). *If $(X_{2i})_{i \geq 1}$ is conditionally partially exchangeable with respect to ρ_1 , then, for any measurable A*

$$s-i) \mathbb{P}((X_{2i}, X_{2j}) \in A^2 \mid c_{1i} = c_{1j} \neq c_{1k}) \geq \mathbb{P}((X_{2i}, X_{2k}) \in A^2 \mid c_{1i} = c_{1j} \neq c_{1k});$$

$$s-ii) \text{ in general, } p(X_{2i}, X_{2j} \mid c_{1i} \neq c_{1j}) \neq p(X_{2i} \mid c_{1i} \neq c_{1j})p(X_{2j} \mid c_{1i} \neq c_{1j}),$$

where a strict inequality in s-i) is achievable as long as $(X_{2i})_{i \geq 1}$ is not conditionally exchangeable with respect to ρ_1 .

All proofs can be found in the Appendix. To preserve subjects' identity, we will always require that the inequality s-i) in Theorem 1 is strict for any measurable A : $0 < \mathbb{P}(X_{2i} \in A) < 1$. This ensures the non-degeneracy of *conditional exchangeability* in (1) for which the left and right terms in s-i) would be equal. Moreover, we will also require the inequality in s-ii) to hold, and, thus, the non-degeneracy of *conditional independence* of enriched constructions. Nonetheless, CPE recovers *conditional exchangeability* (as in the models in Quintana et al. 2022) and *conditional independence* (as in enriched constructions) as limiting cases of minimal and maximal preservation of subjects' identities, corresponding to $\mathbb{P}((X_{2i}, X_{2j}) \in A^2 \mid c_{1i} = c_{1j} \neq c_{1k}) = \mathbb{P}((X_{2i}, X_{2k}) \in A^2 \mid c_{1i} = c_{1j} \neq c_{1k})$ and $p(X_{2i}, X_{2j} \mid c_{1i} \neq c_{1j}) = p(X_{2i} \mid c_{1i} \neq c_{1j})p(X_{2j} \mid c_{1i} \neq c_{1j})$, respectively.

The following propositions illustrate how existing Bayesian clustering approaches do satisfy or not CPE.

Proposition 1 (Temporal random partition model). *If $(X_{1i}, \dots, X_{Ti})_{i=1}^n$ follows the temporal random partition model (t-RPM) of Page et al. (2022) - Section 2, then $(X_{ti})_{i \geq 1}$ is conditionally partially exchangeable with respect to ρ_{t-1} , but not conditionally exchangeable.*

Proposition 2 (Separately exchangeable NDP-CAM). *If $(X_{1i}, \dots, X_{ji})_{i \geq 1}$ follows the separately exchangeable random partition model of Rebaudo et al. (2021) - Section 3, then, for any j and j' , $(X_{ji})_{i \geq 1}$ is conditionally partially exchangeable with respect to ρ'_j but not conditionally exchangeable.*

Proposition 3 (Dependent Dirichlet processes). *If $(X_{1i}, X_{2i})_{i \geq 1}$ follows a mixture model with mixing probabilities provided by dependent processes of the type described in MacEachern (2000) and Quintana et al. (2022) = Section 2, then, conditionally on ρ_1 , $(X_{2i})_{i \geq 1}$ is conditionally exchangeable.*

Finally, for the remainder of this work, we demonstrate that the strength of CPE extends beyond providing a condition to preserve subjects' identities in multi-view probabilistic clustering. Instead, it can also serve as a *constructive* definition that, due to its conditional formulation, facilitates the development and analysis of various clustering processes while ensuring analytical and posterior computational tractability.

3 The class of telescopic clustering models

3.1 A general telescopic clustering model

First-layer observations $(X_{1i})_{i=1}^n$ are assumed to be distributed according to a mixture model (Ferguson 1983, Lo 1984):

$$X_{1i} \mid \tilde{p}_1 \stackrel{iid}{\sim} \int_{\Theta_1} k_1(X_{1i}, \theta) \tilde{p}_1(d\theta), \quad \text{for } i = 1, \dots, n, \quad (2)$$

where $k_1(\cdot, \cdot)$ is a kernel defined on (\mathbb{X}_1, Θ_1) , \tilde{p}_1 is an almost-surely discrete random probability, i.e., $\tilde{p}_1 \stackrel{a.s.}{=} \sum_{m=1}^M w_m \delta_{\theta_m^*}$, with $M \in \mathbb{N} \cup \{+\infty\}$ and $(w_m, \theta_m^*)_{m=1}^M$ random variables such that $\sum_{m=1}^M w_m \stackrel{a.s.}{=} 1$. In the following, for notational convenience, the set $[M] := \{1, \dots, M\}$ denotes the set of the first M natural numbers, when M is finite, and the set of the natural numbers \mathbb{N} , when $M = \infty$. Model (2) can be rewritten in terms of the allocation vector $\mathbf{c}_1 = (c_{11}, \dots, c_{1n})$, $i = 1, \dots, n$, defined in Section 2: $X_{1i} \mid c_{1i} = m, \boldsymbol{\theta}^* \stackrel{ind}{\sim} k_1(X_{1i}; \theta_m^*)$. In the following, we assume that the subject-specific allocation variables \mathbf{c}_1 and the independent and identically distributed cluster-specific parameters $\boldsymbol{\theta}^* = (\theta_m^*)_{m=1}^M$ are apriori independent so that the corresponding mixing random probability \tilde{p}_1 belongs to the class of species sampling processes (Pitman 1996). To satisfy CPE, the second-layer conditional model is defined as

$$X_{2i} \mid c_{1i} = m, (\tilde{p}_{21}, \dots, \tilde{p}_{2M}) \stackrel{ind}{\sim} \int_{\Theta_2} k_2(X_{2i}, \theta) \tilde{p}_{2m}(d\theta), \quad \text{for } i = 1, \dots, n, \quad (3)$$

where k_2 is a kernel defined on (\mathbb{X}_2, Θ_2) , $M \in \mathbb{N} \cup \{+\infty\}$ is the number of mixture components at the first layer, $(\tilde{p}_{21}, \dots, \tilde{p}_{2M})$ is a vector of (possibly dependent) almost-surely discrete and exchangeable random probability measures. Thus, when $M = \infty$, $(\tilde{p}_{21}, \dots, \tilde{p}_{2M})$ is a countably infinite number of probability measures indexed by \mathbb{N} . Bringing everything together, we arrive at the following definition for this class of models.

Definition 2 (Telescopic clustering model). *A random matrix $(X_{1i}, X_{2i})_{i \geq 1}$ taking values in $(\mathbb{X}_1 \times \mathbb{X}_2)^\infty$ is said to follow a telescopic clustering model (with two layers) if it admits the following representation:*

$$X_{1i} \mid \tilde{p}_1 \stackrel{iid}{\sim} \int_{\Theta_1} k_1(X_{1i}, \theta) \tilde{p}_1(d\theta), \quad \text{for } i = 1, 2, \dots$$

$$X_{2i} \mid c_{1i} = m, (\tilde{p}_{21}, \dots, \tilde{p}_{2M}) \stackrel{ind}{\sim} \int_{\Theta_2} k_2(X_{2i}, \theta) \tilde{p}_{2m}(d\theta), \quad \text{for } i = 1, 2, \dots$$

with $\tilde{p}_1 \sim P_1$ and $(\tilde{p}_{21}, \dots, \tilde{p}_{2M}) \sim P_2$, and where k_1 and k_2 are kernels defined on (\mathbb{X}_1, Θ_1) and (\mathbb{X}_2, Θ_2) , respectively; $\mathbf{c}_1 = (c_{11}, \dots, c_{1n})$ is a configuration of the allocation variables corresponding to the random partition ρ_1 induced by the marginal mixture model of $(X_{1i})_{i=1}^n$; $M \in \mathbb{N} \cup \{+\infty\}$ is the number of mixture components in the marginal model of $(X_{1i})_{i=1}^n$; the prior P_1 is such that \tilde{p}_1 is an almost-surely discrete random probability measure; the prior P_2 is such that $(\tilde{p}_{21}, \dots, \tilde{p}_{2M})$ are almost-sure discrete (possibly dependent) exchangeable random probability measures.

A specific model is then obtained when the prior distributions P_1 and P_2 for \tilde{p}_1 and $(\tilde{p}_{21}, \dots, \tilde{p}_{2M})$, respectively, are chosen.

The next theorem provides the hierarchical representation of the joint model for the data matrix.

Theorem 2 (Telescopic clustering - joint representation). *If $(X_{1i}, X_{2i})_{i \geq 1}$ follows a telescopic clustering model with two layers, as in Definition 2, then, for $i = 1, 2, \dots$, there exist θ_i, ξ_i , and \tilde{p} , such that*

$$(X_{1i}, X_{2i}) \mid (\theta_i, \xi_i) \stackrel{\text{ind}}{\sim} k_1(X_{1i}, \theta_i) k_2(X_{2i}, \xi_i), \quad (\theta_i, \xi_i) \mid \tilde{p} \stackrel{\text{iid}}{\sim} \tilde{p} \stackrel{\text{a.s.}}{=} \sum_{m=1}^M \sum_{s=1}^S w_m q_{ms} \delta_{(\theta_m^*, \xi_s^*)}.$$

From the conditional construction of telescopic clustering models, it is trivial to prove that they satisfy CPE. In Theorem 2 we prove that the joint observations $(X_{1i}, X_{2i})_{i \geq 1}$ are exchangeable across i and that their law is *Kolmogorov-consistent* in n . The latter condition is sometimes referred to as *marginal invariance* (Dahl et al. 2017) or *projectivity* (Betancourt et al. 2022) and, roughly speaking, implies that the model is suitable for drawing inferences on an infinite population since the distribution employed to model a finite sample n is extendable. Here, *Kolmogorov consistency* follows directly from the conditional-i.i.d. sampling of (θ_i, ξ_i) in Theorem 2 and de Finetti’s theorem (De Finetti 1937) for infinite exchangeable sequences of random variables.

Finally, if a global clustering structure (i.e., based on all layers) is of interest, the telescopic model still provides appropriate inference. Indeed, in telescopic clustering, global clusters are defined as the common refinement of the partitions at different layers, i.e., two subjects belong to the same global cluster if they belong to the same cluster at all layers. Still, the main goals of telescopic clustering models is different: (1) provide also, possibly different, clustering configuration at each layer, (2) allow global clusters to share all or a subset of latent parameters at any layer (cfr., Petrone et al. 2009), (3) allow more flexible transfer of information across features, which translates into better inferential performance (see Section 7.1), (4) allow investigating dependence between features in terms of dissimilarities between clustering configurations at different layers. The latter point is more extensively described in the next section.

3.2 Measures of telescopic dependence

The class of models described above allows for a bi-variate clustering configuration of the same observational units taking into account within-subject dependence. In this section, four dependence measures between clustering configurations (at the different layers) are presented. The measure of telescopic dependence and the telescopic adjusted Rand index are novel measures of dependence that capture specific properties of telescopic clustering models, while the remaining two are widely used measures: the expected Rand index and the expected Binder loss.

In telescopic clustering models, the probability of any two subjects being clustered together at layer 2 depends on whether they were clustered together at layer 1, with a higher probability if they were already clustered together than if they were not. This result follows directly from the fact that, in partially exchangeable mixture models with equal marginal distributions, the probability of ties within a group is always at least as high as the probability of ties across groups (for details, see, Ascolani et al. 2024, Franzolini 2022, Franzolini, Lijoi, Prünster & Rebaudo 2023). In light of this, we define a conditional measure of similarity between ρ_1 and ρ_2 as a normalized difference between conditional probabilities of ties.

Definition 3 (Measure of telescopic dependence). *Given two random partitions ρ_1 and ρ_2 of the same subjects,*

$$\tau = \frac{\mathbb{P}[c_{2i} = c_{2j} \mid c_{1i} = c_{1j}] - \mathbb{P}[c_{2i} = c_{2j} \mid c_{1i} \neq c_{1j}]}{\mathbb{P}[c_{2i} = c_{2j} \mid c_{1i} = c_{1j}]}$$

is called *measure of telescopic dependence* between ρ_1 and ρ_2 .

By definition, $\tau \in [0, 1]$ and $\tau = 1$ if and only if $\mathbb{P}[c_{2i} = c_{2j} \mid c_{1i} \neq c_{1j}] = 0$, while $\tau = 0$ if and only if $\mathbb{P}[c_{2i} = c_{2j} \mid c_{1i} = c_{1j}] = \mathbb{P}[c_{2i} = c_{2j} \mid c_{1i} \neq c_{1j}]$. It is immediate to show that when ρ_1 and ρ_2 are independent, then $\tau = 0$. On the other hand, under the enriched Dirichlet process $\tau = 1$, indicating maximum telescopic dependence, while in our framework $\tau \in [0, 1]$. This is due to the fact that in telescopic clustering $\mathbb{P}[c_{2i} = c_{2j} \mid c_{1i} \neq c_{1j}]$ can be positive, while in the enriched Dirichlet process the same probability is equal to zero for any value of the hyperparameters, resulting in a smaller support for the joint prior of the partitions. The measure τ of telescopic dependence is an asymmetric measure, which is computed conditionally on the allocation at layer 1.

Denote with $\Pi(n)$ the space of partitions of n elements and with $p(\rho_1, \rho_2)$ the joint probability law of the two clustering configurations induced by a telescopic clustering model, which we name *telescopic exchangeable partition probability function* (t-EPPF). In general, the t-EPPF has full support on the space of bi-variate clustering configurations $\Pi(n)^2$, while still encoding dependence between clustering configurations. In the following, we consider the expected Rand index (ER) and the expected Binder loss (EB) between ρ_1 and ρ_2 , defined respectively as

$$ER = \binom{n}{2}^{-1} \int_{\Pi(n)^2} [a(\rho_1, \rho_2) + b(\rho_1, \rho_2)] d p(\rho_1, \rho_2)$$

$$EB = \int_{\Pi(n)^2} [c(\rho_1, \rho_2) + d(\rho_1, \rho_2)] d p(\rho_1, \rho_2)$$

where a , b , c , and d are functions of the partitions: a returns the number of pairs of observations clustered together both at layer 1 and 2, b the number of pairs clustered together neither at layer 1 nor 2, c the number of pairs clustered together at layer 1, but not at layer 2, and d the number of pairs clustered together at layer 2 but not at layer 1.

Proposition 4 (Dependence measures as functions of the number of clusters). *In a telescopic clustering model, a priori*

$$\tau = \frac{\mathbb{P}(K_{22} = 1 \mid K_{12} = 1) - \mathbb{P}(K_{22} = 1 \mid K_{12} = 2)}{\mathbb{P}(K_{22} = 1 \mid K_{12} = 1)}$$

$$ER = \mathbb{P}(K_{12} = K_{22}), \quad EB = \binom{n}{2} \mathbb{P}(K_{12} \neq K_{22})$$

where $K_{\ell n}$ denotes the number of clusters at layer ℓ in a sample of n subjects.

As noted by [Hubert & Arabie \(1985\)](#), when the Rand index is used to compare random partitions, its expected value is not 0 in case of independence of the partitions. In a telescopic clustering, when ρ_1 and ρ_2 are independent, the expected value of the rand index is $ER^\perp = \sum_{\kappa=1}^2 \mathbb{P}(K_{12} = \kappa) \mathbb{P}(K_{22} = \kappa)$ where \perp denotes independence (see [Proposition 4](#) above). Thus, ER^\perp is typically positive. In the same spirit as that of the adjusted Rand index ([Hubert & Arabie 1985](#)), we define a *telescopic adjusted rand-index* that allows us to correct for the randomness of the partitions.

Definition 4 (Telescopic adjusted Rand index). *The telescopic adjusted Rand index between ρ_1 and ρ_2 is defined as*

$$TARI = \frac{[a(\rho_1, \rho_2) + b(\rho_1, \rho_2)] - ER^\perp}{1 - ER^\perp}$$

It is trivial to prove that, in the case of independence, the a priori expected value of the *TARI* equals 0.

3.3 Extension to L layers using polytrees

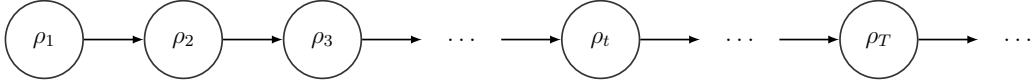


Figure 2: Layer dependence for longitudinal data.

The class of telescopic models as presented in the previous sections defines a prior distribution for the joint law of two partitions, ρ_1 and ρ_2 , through the product $p(\rho_1)p(\rho_2 | \rho_1)$ where $p(\rho_1)$ and $p(\rho_2 | \rho_1)$ are used to denote the marginal law of the partition ρ_1 and the conditional law of the partition ρ_2 , respectively.

The main advantage and novelty of this class of models lie in how the dependence between the two partitions is defined through the CPE, which ultimately specifies a one-way relationship from ρ_1 to ρ_2 , denoted in the following as $\rho_1 \rightarrow \rho_2$.

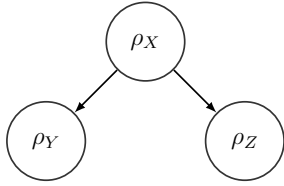


Figure 3: Triangular dependence for three layers.

A straightforward way to extend the modeling strategy to any number of layers is by combining multiple pairwise relationships in a polytree. For instance, in the context of longitudinal data, where different measurements are collected at different time points a Markovian structure across different layers can be imposed. The resulting telescopic clustering model is then obtained assuming CPE between Π_t and Π_{t+1} for any $t \in \mathbb{N}$, i.e., $p(\rho_t, t = 1, 2, \dots) = p(\rho_1) \prod_{t=2}^{\infty} p(\rho_t | \rho_{t-1})$. See Figure 2.

A second extension that we consider in this work involves combining the dependence across three sets of features through the triangular graph represented in Figure 3. In this setting, given the clustering configuration of X , which is the response variable of main interest, the goal is to also infer additional clustering configurations for two other sets of variables: Y and Z . Then, the t-EPPF of the model is given by $p(\rho_X, \rho_Y, \rho_Z) = p(\rho_X)p(\rho_Y | \rho_X)p(\rho_Z | \rho_X)$. The polytrees strategy is based on a partial ordering of the different layers, due to the fact that each node in the graph can have at most one parent node and the multivariate dependence across layers is obtained by combining pairwise dependence only. Nonetheless, in the structure in Figure 3, the CPE induces a mutual (undirected) dependence between ρ_Y and ρ_Z in the sense that, when $p(\rho_Y | \rho_X) = p(\rho_Z | \rho_X)$, the conditional law of ρ_Z given ρ_Y is the same as the conditional law of ρ_Y given ρ_Z .

4 A telescopic model with infinite number of labels

Hierarchical constructions for dependent processes, initially introduced in Teh et al. (2006), offer a powerful framework for modeling dependence across random distributions. In Teh et al. (2006), the construction is based on the Dirichlet process and it was further extended to encompass more general processes in Camerlenghi et al. (2019), Argiento et al. (2020), and Bassetti et al. (2020). We employ this construction to build up the telescopic mixtures with hierarchical Dirichlet processes (t-HDP), where we set as prior for the first-layer random probability \tilde{p}_1 an HDP, which defines the law of a single process (for details and generalization of this prior, see Camerlenghi et al. 2018) such that

$$\tilde{p}_1 | \gamma, \tilde{p}_0 \sim DP(\gamma, \tilde{p}_0), \quad \tilde{p}_0 | \gamma_0 \sim DP(\gamma_0, P_\theta), \quad (4)$$

while the second-layer conditional law is

$$\begin{aligned} X_{2i} \mid \mathbf{c}_1, (\tilde{p}_{21}, \tilde{p}_{22} \dots) &\stackrel{ind}{\sim} \int f(X_{2i}, \theta) \tilde{p}_{2c_i}(\mathrm{d}\theta) \\ \tilde{p}_{2m} \mid \alpha, \tilde{q}_0 &\stackrel{iid}{\sim} DP(\alpha, \tilde{q}_0), \quad \tilde{q}_0 \mid \alpha_0 \sim DP(\alpha_0, P_\xi), \end{aligned} \quad (5)$$

where $DP(\alpha, P)$ denotes a Dirichlet process with concentration parameter α and base distribution P . Consider a specific partition ρ_1 into K_{1n} sets of numerosities $n_1, \dots, n_{K_{1n}}$ for the first-layer partition. Then, we have (see, [Camerlenghi et al. 2018](#))

$$\mathbb{P}[\rho_1 = \rho] = \frac{\gamma_0^{K_{1n}}}{(\gamma)^{(n)}} \sum_{\ell} \frac{\gamma^{|\ell|}}{(\gamma_0)^{(|\ell|)}} \prod_{m=1}^{K_{1n}} (\ell_m - 1)! |s(n_m, \ell_m)| \quad (6)$$

where $|s(n, k)|$ denotes the signless Stirling number of the first kind and the sum in (6) runs over all vectors $(\ell_1, \dots, \ell_{K_{1n}})$ such that $\ell_m \in [n_m]$ and $(\gamma)^{(n)} = \Gamma(\gamma + n)/\Gamma(\gamma)$, where $\Gamma(x)$ denote the Gamma function in x . The conditional law of the partition at layer 2, given ρ_1 , is

$$\mathbb{P}[\rho_2 = \rho \mid \rho_1] = \frac{\alpha_0^{K_{2n}}}{\prod_{m=1}^{K_{1n}} (\alpha)^{(n_m)}} \sum_{\mathbf{t}} \frac{\alpha^{|\mathbf{t}|}}{(\alpha_0)^{(|\mathbf{t}|)}} \prod_{s=1}^{K_{2n}} (t_{\cdot s} - 1)! \prod_{m=1}^{K_{1n}} |s(n_{ms}, t_{ms})| \quad (7)$$

where the sum runs over all matrices $K_{1n} \times K_{2n}$, whose generic element t_{ms} belong to $[n_{ms}]$ provided that $n_{ms} \geq 1$, and is equal to 1 when $n_{ms} = 0$. Moreover, $t_{\cdot s} = \sum_m^{K_{1n}} t_{ms}$. See [Camerlenghi et al. \(2019\)](#).

Theorem 3 (t-EPPF in the t-HDP model). *Given a telescopic mixture model with hierarchical Dirichlet processes and two layers, $p(\rho_1, \rho_2)$ is given by*

$$\frac{\gamma_0^{K_{1n}} \alpha_0^{K_{2n}}}{(\gamma)^{(n)} \prod_{m=1}^{K_{1n}} (\alpha)^{(n_m)}} \sum_{\ell, \mathbf{t}} \frac{\gamma^{|\ell|} \alpha^{|\mathbf{t}|}}{(\gamma_0)^{(|\ell|)} (\alpha_0)^{(|\mathbf{t}|)}} \left(\prod_{m=1}^{K_{1n}} (\ell_m - 1)! |s(n_m, \ell_j)| \right) \prod_{s=1}^{K_{2n}} (t_{\cdot s} - 1)! \prod_{m=1}^{K_{1n}} |s(n_{ms}, t_{ms})|$$

Starting from the expression of the t-EPPF, it is straightforward to compute the indexes of dependence introduced in Section 3.2.

Corollary 1 (Measures of dependence in the t-HDP model). *In a t-HDP, the measure τ of telescopic dependence is $\tau = \frac{\alpha_0}{\alpha_0 + \alpha + 1}$ and the expected Rand index is $ER = \frac{(1 + \gamma_0 + \gamma)(1 + \alpha_0 + \alpha) + \gamma_0 \alpha_0 \gamma \alpha}{(\gamma_0 + 1)(\gamma + 1)(\alpha_0 + 1)(\alpha + 1)}$.*

Thus, τ tends to 0 as α tends to ∞ and to 1 as α_0 tends to ∞ . Finally, at layer 1, alternative priors might be more suitable depending on the application, such a non-hierarchical prior such as the classical Dirichlet Process ([Ferguson 1973](#)), or the Pitman-Yor process (e.g., [Pitman & Yor 1997](#)). However, incorporating a hierarchical structure in the subsequent layers is essential for achieving conditional partial exchangeability.

5 A telescopic model with random number of labels

The t-HDP model introduced in the previous section assumes that the number of sub-populations (or components) in the mixtures equals infinity, which is a classical modelling assumption in Bayesian nonparametric mixtures models. Nonetheless, an alternative successful strategy consists in assuming that the number M of sub-populations is almost-surely

finite and placing a prior over M . The second telescopic model introduced here lies within this framework. The prior for the first-layer random probability \tilde{p}_1 is defined by

$$\tilde{p}_1 = \sum_{m=1}^M w_m \delta_{\theta_m^*} \quad (8)$$

$$\mathbf{w} = (w_1, \dots, w_M) \mid M \sim P_w, \quad \theta_m^* \mid M \stackrel{iid}{\sim} P_\theta, \quad \text{for } m = 1, \dots, M, \quad M \sim P_M,$$

where \mathbf{w} and $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_M^*)$ are independent and P_M has support on the set of the natural numbers \mathbb{N} . The resulting marginal model for the first layer is a finite mixture with a random number of components (Nobile 1994, Miller & Harrison 2018, Argiento & De Iorio 2022). Depending on the choice of P_w different finite-dimensional prior processes can be employed as priors for the finite mixture construction. In the following, we focus on the Dirichlet distribution as prior for the weights, as it is the most popular in applications, i.e., $\mathbf{w} = (w_1, \dots, w_M) \mid M, \gamma \sim \text{Dirichlet}_M(\gamma, \dots, \gamma)$. Then, the conditional law of the second layer is defined employing a novel construction for the mixing random probability measures, whose formal construction is detailed in the following definition.

Definition 5 (Unique-atom process). *A vector of random probabilities $(\tilde{p}_1, \dots, \tilde{p}_K)$ is a unique-atom process if they admit the following almost-sure discrete representation: $\tilde{p}_m \stackrel{a.s.}{=} (1 - Z) \delta_{\xi_m^*} + Z \tilde{p}_0$ for $m = 1, \dots, K$, with $Z \sim \text{Bernoulli}(\omega)$, where \tilde{p}_0 is an almost-surely discrete random probability, $\xi_m^* \stackrel{iid}{\sim} P_\xi$, for $m = 1, \dots, K$, and \tilde{p}_0 , $(\xi_m^*)_{m=1}^K$, and Z are pairwise independent.*

In the following, we make use of unique-atom processes where the common \tilde{p}_0 in the previous definition is a random probability with a random (almost-surely finite) number of support points and Dirichlet weights, i.e., $\tilde{p}_0 \stackrel{a.s.}{=} \sum_{s=1}^S q_s \delta_{\xi_{0s}^*}$ with $S \sim P_S$, weights q_s distributed accordingly to a symmetric Dirichlet distribution and $\xi_{0s}^* \stackrel{iid}{\sim} P_\xi$. The rationale behind the construction in Definition 5 is the following: when the random variable $Z = 0$, the clustering structure is kept constant from one layer to the next, while when $Z = 1$, the clustering structure is estimated independently from the clustering arrangement at the previous layer. Employing unique-atom processes to build up CPE needed for telescopic clustering, we get the following second-layer specification

$$\begin{aligned} X_{2i} \mid \mathbf{c}_1, \mathbf{q}, \boldsymbol{\xi}, S, Z &\stackrel{iid}{\sim} (1 - Z) k_2(X_{2i}; \xi_{c_{1i}}^*) + Z \sum_{s=1}^S q_s k_2(X_{2i}; \xi_{0s}^*) \\ \mathbf{q} = (q_1, \dots, q_S) \mid S, \alpha &\sim \text{Dirichlet}_S(\alpha, \dots, \alpha) \\ \xi_{0s}^* \mid S &\stackrel{iid}{\sim} P_\xi, \quad \xi_m^* \mid K_{1n} \stackrel{iid}{\sim} P_\xi, \quad S \sim P_M, \quad Z \sim \text{Bernoulli}(\omega). \end{aligned} \quad (9)$$

The joint law of the two partitions is provided by the next theorem.

Theorem 4 (t-EPPF in the telescopic unique atom process). *Given a telescopic mixture with unique atom processes, the t-EPPF is*

$$\begin{aligned} p(\rho_1, \rho_2) = & (1 - \omega) V(n, K_{1n}) \prod_{m=1}^{K_{1n}} \frac{\Gamma(\gamma + n_m)}{\Gamma(\gamma)} \mathbb{1}(\rho_1 = \rho_2) \\ & + \omega V(n, K_{2n}) \prod_{s=1}^{K_{2n}} \frac{\Gamma(\alpha + \sum_{m=1}^{K_{1n}} n_{ms})}{\Gamma(\alpha)} V(n, K_{1n}) \prod_{m=1}^{K_{1n}} \frac{\Gamma(\gamma + n_m)}{\Gamma(\gamma)} \end{aligned}$$

where $V(n, K) = \sum_{M=1}^{+\infty} \frac{M_{(K)}}{(\gamma K)^{(n)}} p_M(M)$.

Corollary 2 (Measures of dependence in the telescopic unique atom process). *In a telescopic mixture with unique atom processes, $\tau = \frac{1-\omega}{1+\omega(\mathbb{E}[S]/\alpha-1)}$ and*

$$ER = \frac{\mathbb{E}[M]}{\gamma} \left(1 - \omega + \omega \frac{\mathbb{E}[S]}{\alpha} \right) + \frac{\mathbb{E}[M(M-1)]\gamma^2}{4\gamma^2 + 2\gamma} \left(1 - \omega + \omega \frac{\mathbb{E}[S(S-1)]}{4\alpha^2 + 2\alpha} \right).$$

Thus, τ tends to 1 as ω tends to 0 and to 0 as ω tends to 1.

6 Algorithms for posterior inference

Similarly to existing Bayesian mixture models, also in telescopic clustering models, posterior inference can be performed through either conditional or marginal Markov chain Monte Carlo (MCMC) algorithms. The conditional algorithms make use of representation theorems and also provide posterior samples of the underlying random probability measures (see, for instance, [Ishwaran & James 2001](#), [Walker 2007](#)). Nonetheless, when the number of components is infinite, conditional algorithms typically require to rely on a truncated approximation of the underlying random probability measure. In contrast, the marginal algorithms are derived through marginalization of the random probability (see, for instance, [Neal 2000](#)).

In the case of telescopic clustering models, marginal algorithms require evaluating the conditional law of the partition at the child nodes when sampling the cluster allocation at any given parent layer. However, evaluating this conditional law is typically computationally intensive, and introducing latent random variables to reduce the cost is not always straightforward. For example, in t-HDP models, the standard data augmentation provided by the Chinese restaurant franchise process ([Teh et al. 2006](#)) simplifies the conditional law of the partition to be evaluated but significantly slows down the mixing to unfeasible levels (for details, see Sections S3 and S4 of the Supplement). Therefore, enabling inference using marginal algorithms requires a tailored variable augmentation scheme for each telescopic clustering model.

On the other hand, the conditional sampling scheme for the t-HDP model exhibits good mixing and significantly lower computational time per iteration, making posterior inference feasible and, importantly, easily adaptable to different poly-tree structures and prior choices (for details, see Sections S2, S3, and S6 of the Supplement).

For these reasons, the results presented in the following sections are obtained via a truncated blocked Gibbs sampler. This algorithm is a conditional one, easier to generalize within the class of telescopic clustering models, provided that the full conditionals of the weights of the random probability measures are available. Unlike the marginal sampling scheme, it does not require model-specific data augmentation techniques, making it the preferred choice for this work. However, it relies on a truncated version of the random probability measures when the number of components is infinite, as in the t-HDP, and thus incurs a truncation error cost. Thus, this approach incurs a truncation error cost. A promising direction for future research is the adaptation of such schemes using slice sampling techniques ([Walker 2007](#), [Kalli et al. 2011](#)), as has been recently applied to the classical HDP in ([Amini et al. 2019](#)) and [Das et al. \(2024\)](#). It is important to notice that the availability of conditional sampling schemes depends on the existence of (conditional) representation theorems and underlying random probabilities, which, thus, for telescopic clustering, are not only an analytical and probabilistic result but a fundamental computational tool. A detailed derivation of the sampling schemes, computational cost, and mixing performance for telescopic models are in Sections S3, S4, and S7 of the Supplement.

Layer	Rand Index				# Mistakes			
	k-means	t-HDP	LSBP	E-DP	k-means	t-HDP	LSBP	E-DP
n.1	0.98	0.98	0.98	0.50	2	2	2	100
n.2	0.98	1.00	0.98	0.90	2	0	2	10
n.3	0.92	0.98	0.92	1.00	8	2	8	0
n.4	0.98	1.00	0.98	0.92	2	0	2	17
n.5	0.92	0.97	0.91	0.89	8	3	9	21
n.6	0.97	0.98	0.97	0.86	3	2	3	31
n.7	0.94	0.99	0.92	0.83	6	1	8	40
n.8	0.95	1.00	0.95	0.79	5	0	5	44
n.9	0.93	1.00	0.93	0.79	7	0	7	47
n.10	0.91	0.99	0.89	0.75	9	1	11	54
average	0.95	0.99	0.83	0.82	5.2	1.1	5.7	36.4

Table 1: Scenario 1, Rand indexes between the estimated and true clustering configurations and numbers of items allocated to the wrong cluster.

7 Numerical studies

7.1 Simulation study

Here we report results for a few simulations, additional simulation studies with different numbers of layers and misspecification can be found in Section S5 of the Supplement, together with additional results regarding the simulations described here below.

In the first simulation scenario (Scenario 1), we generate data on $n = 200$ items and $T = 10$ layers. At each layer, marginally we assume two clusters simulated from two univariate Normal distributions with unitary variance and centered in 0 and 4 respectively. From one layer to the next, 10 items (5% of the total) are selected at random and moved to the other cluster.

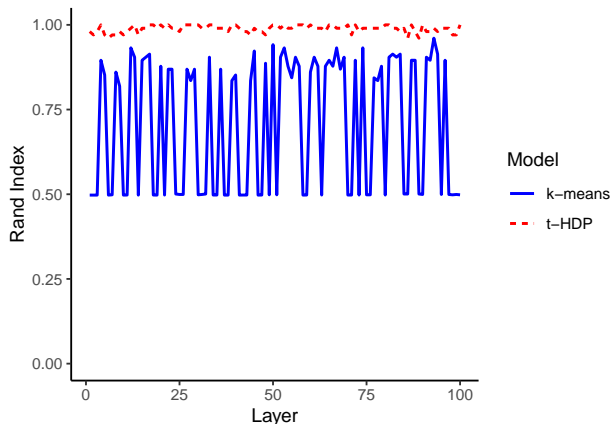


Figure 4: Scenario 2. Rand indexes between the truth and the estimated configuration.

estimate for the clustering configuration the one that minimizes the variation of information loss (Meilă 2007). Table 1 summarises the results. The t-HDP model outperforms the competitors both consistently at each layer and overall. In Scenario 2, data for $T = 100$ layers are simulated. At each layer, there are two clusters with 100 observations each. At layer 1, data are sampled from $X_{1i} | c_{1i} \stackrel{\text{ind}}{\sim} \mathcal{N}(0, 1)\mathbb{1}(c_{1i} = 1) + \mathcal{N}(3, 1)\mathbb{1}(c_{1i} = 2)$. Then, from layer ℓ to layer $\ell + 1$, 2% of the observations are selected at random and moved to the other cluster. Figures 4 and 5 summarize the results of the t-HDP model and independent k-means clustering, where again the t-HDP outperforms k-means. Posterior estimates are

We compare four methods: (i) k-means fitted independently at each layer, where the number of clusters is determined by the gap statistics (Tibshirani et al. 2001); (ii) the t-HDP’s estimate; (iii) the estimate obtained with a logit stick-breaking process (LSBP) (Ren et al. 2011); and (iv) the estimate from an Enriched Dirichlet process (E-DP) (Wade et al. 2011). For the LSBP, the layer’s number is used as a covariate for both the weights and the atoms (for more details and algorithms, see, Rigon & Durante 2021). For models (ii)-(iv), we use a Gaussian kernel for the nonparametric mixture with a Normal-InverseGamma for the mean and the variance as base measure. We report as a point

obtained by minimizing the variation of information loss (Meilă 2007) and by employing the gap statistics (Tibshirani et al. 2001).

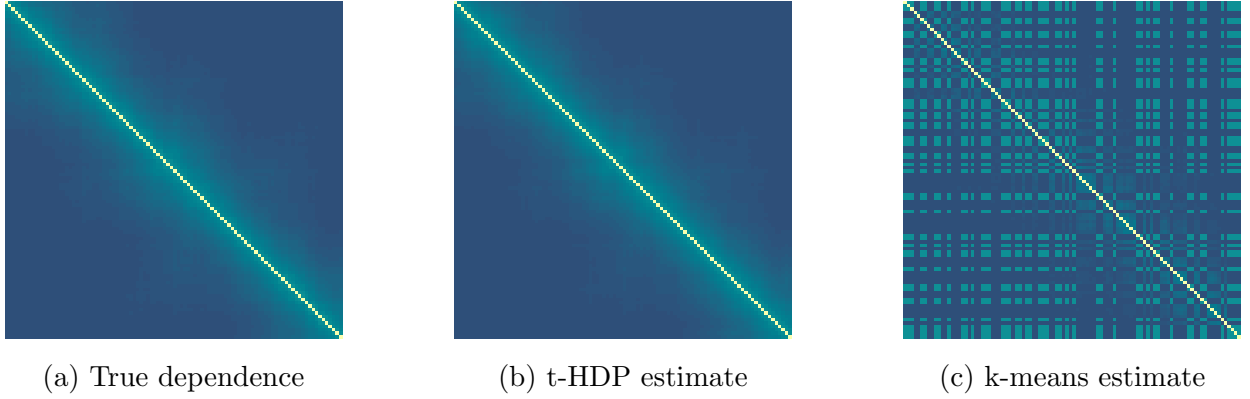


Figure 5: Simulation study: results for Scenario 2. Pairwise Rand indexes between any couple of layers for (a) the true clustering configurations; (b) the t-HDP model; (c) k-means.

7.2 An application to childhood obesity

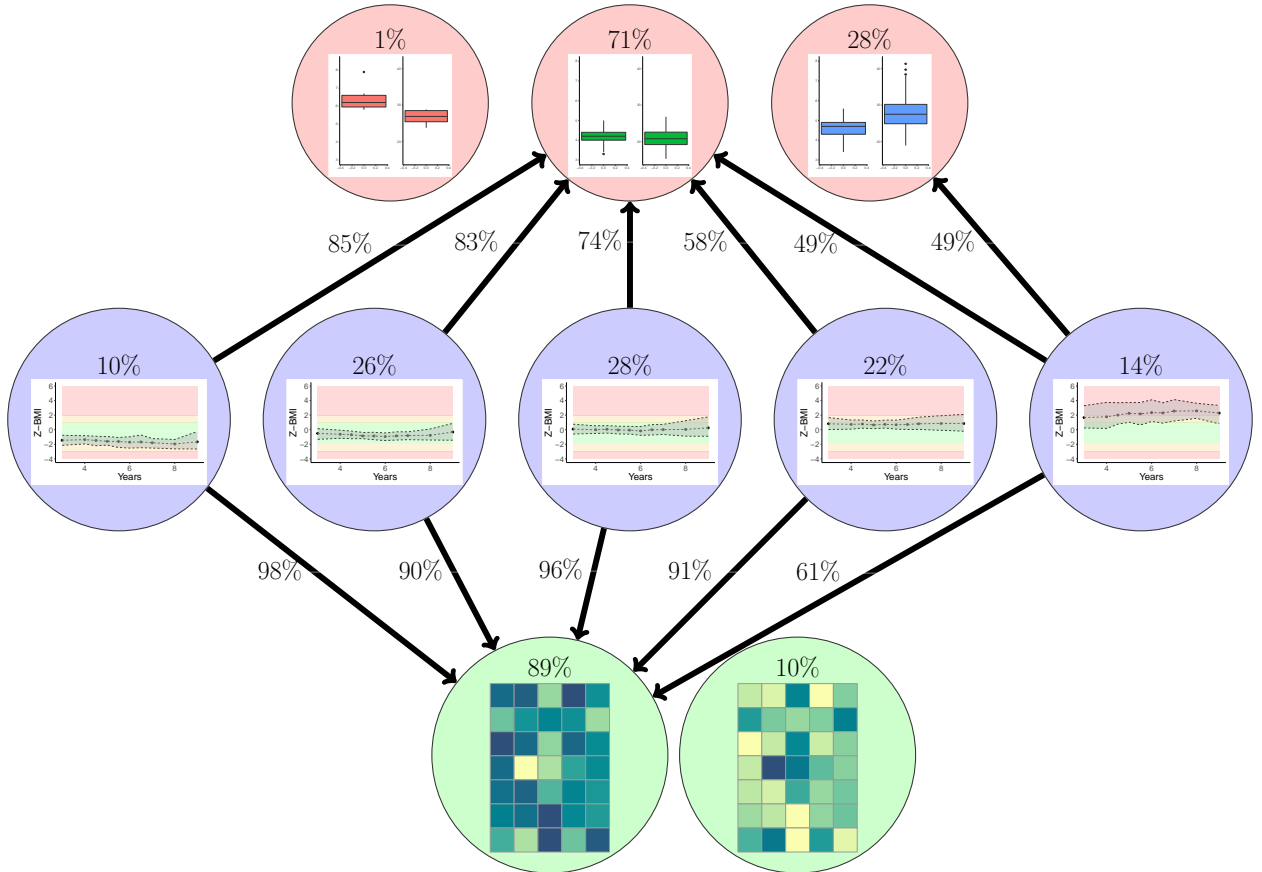


Figure 6: Estimated clustering configuration for the GUSTO cohort data. Nodes in the graph represent different clusters and colors different layers. The percentage within the nodes denotes the amount of children assigned to that cluster. Edges are drawn from each growth-trajectory cluster towards the mother cluster and metabolites cluster to which the majority of children in that particular cluster are assigned.

In this section, we investigate childhood obesity patterns and their relationship with

metabolic pathways, as well as traditional clinical markers for mothers employing the t-HDP. Data are available on a sample of $n = 553$ children from the Growing Up in Singapore Towards healthy Outcomes (GUSTO) cohort study, based in Singapore (Soh et al. 2014).

The first layer of information consists of z-BMI trajectories, including ten unequally spaced measurements per child observed from ages 3 to 9. The second layer contains information on the mother’s pre-pregnancy BMI (a known risk factor for childhood obesity) and the fasting oral glucose tolerance test (ogtt) result conducted at week 26 of pregnancy.

In this third layer, we include concentration data of 35 metabolites measured in the children using NMR spectroscopy. Before applying the t-HDP model, we compute principal components of the metabolite data in the third layer, selecting the first six components based on the scree plot and the elbow method, which collectively explain 66% of the variability. By clustering on the principal components, we focus on global patterns of the 35 metabolites, reducing noise and dimensionality, thus obtaining more robust and interpretable clusters. Data from the same cohort have been also analyzed by Cremaschi et al. (2024) with the goal of identifying metabolic pathways related to childhood obesity.

We fit the t-HDP model presented in Section 4 with multivariate independent Gaussian kernels and Normal-Inverse-Chi-Squared base measures for the vectors of means and variances. We specify a Gamma(1, 1) prior on all the concentration parameters. The total number of features is 18, divided into three layers of dimension 10, 2, and 6, respectively. The primary information is the growth trajectory of the child and conditionally on the clustering configuration of the trajectories, we define the model for the mother-layer and the metabolite-layer. We perform 100 000 iterations of the partially collapsed conditional block Gibbs sampler described in Section S4.2 of the Supplement, discard the first half as burn-in, and apply a thinning of 5 so that the final posterior sample is 10,000 draws. The estimated clustering configurations are summarized in Table S7.1 in the Supplement and shown in Figure 6. A detailed account of the results is provided in Section S7 of the Supplement.

The analysis identifies five distinct clusters that represent five different trajectories of z-BMI. The trajectories exhibit relatively stable patterns across the various time points considered but largely vary across clusters in terms of average z-BMI. More precisely, approximately 10% of children show consistently low z-BMI values (*underweight cluster*), around 14% of children fell into the cluster characterized by overweight/obesity status (*obesity cluster*), while 26%, 28%, and 22% of children are associated to normal-weight trajectories which are, respectively, below average, equal to average and above average, indicating a healthier weight status as compared to the *underweight cluster* and the *obesity cluster*. At layer 2, mothers’ clinical profiles are split into three distinct clusters. The first cluster contains a few outliers with exceptionally high glucose levels compared to the average in the sample. The remaining two clusters divide the mothers into two distinct groups. The first group, comprising 71% of mothers, exhibits *below-average* levels of glucose and BMI. In contrast, the second group, consisting of 28% of mothers, is characterized by *above-average* levels of both glucose and BMI. The percentage of children associated with the *below-average cluster* of mothers steadily decreases across the z-BMI clusters as the z-BMI trajectory increases. This finding suggests a positive relationship between the z-BMI trajectories of the child and the clinical markers of the mothers. Specifically, the majority of mothers in the *above-average cluster* have children with an overweight growth trajectory. This association is confirmed in the medical literature (see, for instance, Dalrymple et al. 2019, Josefson et al. 2020, Landon et al. 2020, Meek 2023, Ormindean et al. 2024). At the parallel layer 3, we estimate two distinct clusters characterized by different concentration profiles. The first cluster encompasses approximately 89% of the children and the second cluster consists of 10% of the children. Furthermore, the results indicate a relationship between obesity and metabolite concentrations. Specifically, conditioning on any of the *normal-weight clusters* or on the

underweight cluster at layer 1, leads to a very similar distribution of the children across the two metabolite clusters. On the contrary, conditioning to the *obesity cluster* at layer 1, a drastic variation in the distribution of children across the metabolite clusters is observed. These results emphasize the role of metabolite profiles in obesity development, as it is also well documented in the medical literature (see, for instance, [Perng et al. 2014](#), [Hellmuth et al. 2019](#), [Perng et al. 2020](#), [Handakas et al. 2022](#), [Schipper et al. 2024](#)). The observed associations between obesity trajectories and metabolite clusters provide further evidence of the complex interplay between metabolic factors and weight status. For a detailed account of the metabolite layer results, see Table S7.2 in the Supplement.

8 Conclusions and future directions

Standard clustering techniques often struggle when applied to datasets collected under a repeated measures design, such as multi-view or longitudinal data. These scenarios require potentially different clustering configurations for each view while still preserving subjects' identities across them. In particular, classical model-based clustering techniques fail to effectively address this issue, as they either impose a single clustering configuration across all views or disregard subjects' identities across views, thereby failing to capture both the multi-view nature of the problem and the repeated measures design underlying the data.

To overcome this challenge, we introduce conditional partial exchangeability (CPE), an invariance requirement for the conditional law of the observables in one view, given the clustering configuration of the same units with respect to another view. When satisfied by a probabilistic clustering model, CPE induces dependencies across views while ensuring that subjects' identities are preserved, as formally established in Theorem 1.

Furthermore, we introduce, characterize, and apply telescopic clustering models, a novel class of Bayesian mixture models. This class of models highlights that the utility of CPE extends beyond maintaining subject identities in multi-view probabilistic clustering. Rather, its conditional formulation provides a constructive definition that facilitates the development and analysis of diverse clustering processes while ensuring both analytical and posterior computational tractability. We motivate our approach theoretically and conduct extensive comparisons with a range of existing methods, consistently demonstrating that our approach outperforms all competitors.

Finally, our framework paves the way for exciting and insightful advancements in the study and development of dependent random partition models.

From a theoretical and probabilistic point of view, we have demonstrated that CPE preserves subjects' identities (as formally established in Theorem 1) and shown that some existing Bayesian models preserving subject identities indeed satisfy CPE. However, an interesting open question remains: whether all dependent partition models that preserve subject identities must necessarily satisfy CPE. Establishing this result would allow us to conclude that CPE is not only a sufficient but also a necessary condition for incorporating repeated measures designs into partition models.

From a statistical and modeling perspective, further exploration of the t-HDP model (and the telescopic clustering class in general) in its Markovian dependence formulation would be valuable, particularly in identifying conditions for the stationarity of the partition chain's law. Similarly, exploring the properties and applications of a more general polytree-dependent structure, particularly examining the marginal distribution of partitions at the leaves where mutual dependencies can arise, would be highly valuable.

A - Appendix

A1 Proof of Theorem 1

Proof of Theorem 1. By condition c-i) in the Definition 1 of Section 2 and de Finetti's representation theorem for partial exchangeability (de Finetti 1938) there exist $\tilde{p}_1, \dots, \tilde{p}_{K_1}$ random probability measures with distribution Q on \mathcal{P}^{K_1} , such that, conditionally on ρ_1 , for m and m' in $[K_1]$, with $m \neq m'$, we have

$$\mathbb{P}((X_{2i}, X_{2j}) \in A^2 \mid c_{1i} = c_{1j} = m, c_{1k} = m') = \int_{\mathcal{P}^{K_1}} \tilde{p}_m^2(A) dQ(\tilde{p}_1, \dots, \tilde{p}_{K_1})$$

and

$$\mathbb{P}((X_{2i}, X_{2k}) \in A^2 \mid c_{1i} = c_{1j} = m, c_{1k} = m') = \int_{\mathcal{P}^{K_1}} \tilde{p}_m(A) \tilde{p}_{m'}(A) dQ(\tilde{p}_1, \dots, \tilde{p}_{K_1})$$

where, by condition c-ii) in Definition 1, for any $A \in \mathbb{X}_2$,

$$\mathbb{E}[\tilde{p}_m^2(A)] \geq \mathbb{E}[\tilde{p}_m(A) \tilde{p}_{m'}(A)].$$

Moreover note that, in general, being $\tilde{p}_1, \dots, \tilde{p}_{K_1}$ dependent,

$$\mathbb{P}((X_{2i}, X_{2j}) \in A \times B \mid c_{1i} = m, c_{1j} = m') = \mathbb{E}[\tilde{p}_m(A) \tilde{p}_{m'}(B)] \neq \mathbb{E}[\tilde{p}_m(A)] \mathbb{E}[\tilde{p}_{m'}(B)].$$

□

A2 Proof of Proposition 1

Before proving Proposition 1, we first introduce the following Lemma.

Lemma 1. *Given a (non-random) partition ρ of n elements, a vector $\gamma = (\gamma_1, \dots, \gamma_n)$ with binary entries, and a permutation $\sigma : [n] \rightarrow [n]$ of n elements, let*

- $\sigma(\rho)$ be the partition obtained swapping the elements in the sets of ρ accordingly to σ ,
- $\mathcal{R}(\gamma) = \{i : \gamma_i = 1\}$ and $\sigma(\gamma) = (\gamma_{\sigma(1)}, \dots, \gamma_{\sigma(n)})$
- $\rho^{\mathcal{R}(\gamma)}$ be the “reduced partition” obtained removing from the sets in ρ all elements that are not in $\mathcal{R}(\gamma)$ and then removing empty sets.

then

1. $\rho^{\mathcal{R}(\gamma)} = \rho^{\mathcal{R}(\sigma(\gamma))}$
2. $\sigma^{-1} \left(\sigma \left(\rho^{\mathcal{R}(\gamma)} \right) \right) = \rho^{\mathcal{R}(\gamma)}$
3. $\rho^{\mathcal{R}(\gamma)} = \sigma \left(\rho^{\mathcal{R}(\gamma)} \right)$ for any $\gamma \in \{0, 1\}^n$ iff $\sigma \in \mathcal{P}(n; \rho)$

where σ^{-1} denotes the inverse of σ , i.e., $\sigma^{-1}(i) = j$, for j such that $\sigma(j) = i$ and $\mathcal{P}(n; \rho)$ denotes the space of permutations of n elements that preserve ρ , cf. Definition 1 in Section 2.2.

Proof of Lemma 1. The first statement follows trivially by definition of $\rho^{\mathcal{R}(\gamma)}$. The second statement follows by the definition of σ^{-1} inverse of σ . The last statement follows by considering $\gamma = (1, \dots, 1)$ and the definition of $\mathcal{P}(n; \rho)$. \square

Proof of Proposition 1. Denoting with X_{ti} a response measured on the i th unit at time t , for $i = 1, \dots, n$ and $t = 1, \dots, T$, the t-RPM mixture model of [Page et al. \(2022\)](#) is defined as

$$\begin{aligned} X_{ti} \mid \boldsymbol{\theta}_t^*, \mathbf{c}_t &\stackrel{iid}{\sim} k(X_{ti}, \theta_{tc_{ti}}^*) && \text{for } i = 1, \dots, n \text{ and } t = 1, \dots, T \\ \theta_{tj}^* \mid \mu_t &\stackrel{ind}{\sim} P_{\mu_t} && \text{for } j = 1, \dots, K_t \text{ and } t = 1, \dots, T \\ \{\mathbf{c}_t, \dots, \mathbf{c}_T\} \mid \boldsymbol{\alpha} &\sim \text{tRPM}(\boldsymbol{\alpha}, n) \end{aligned}$$

where $\boldsymbol{\theta}_t^* = (\theta_{t1}^*, \dots, \theta_{tK_t}^*)$, K_t is the number of clusters at time t , k denotes a kernel, P_{μ_t} is an absolutely continuous distribution, $\mathbf{c}_t = (c_{t1}, \dots, c_{tn})$ is the vector of allocation variables encoding the clustering configuration at time t , and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_T) \in [0, 1]^T$. For the formal and detailed definition of

$$\{\mathbf{c}_t, \dots, \mathbf{c}_T\} \mid \boldsymbol{\alpha} \sim \text{tRPM}(\boldsymbol{\alpha}, n)$$

we refer to the paper of [Page et al. \(2022\)](#), even though in the following we describe the core of the construction.

Denoting with ρ_{t-1} the partition encoded by \mathbf{c}_{t-1} , to prove CPE, we need to prove that

$$p(X_{t1}, \dots, X_{tn} \mid \rho_{t-1}) = p(X_{t\sigma(1)}, \dots, X_{t\sigma(n)} \mid \rho_{t-1})$$

for any $\sigma \in \mathcal{P}(n; \rho_{t-1})$, where, we recall, that $\mathcal{P}(n; \rho_{t-1})$ denotes the space of permutations of n elements that preserve ρ_{t-1} , see Section 2.

Given a partition ρ , we denote with $\sigma(\rho)$ the partition obtained by swapping the elements in the sets of ρ accordingly to the permutation σ . In the t-RPM mixture, the conditional law of $(X_{ti})_{i=1}^n$ conditionally on the partition at the previous time point ρ_{t-1} , is defined such that

$$p(X_{t1}, \dots, X_{tn} \mid \rho_{t-1}) = \sum_{\lambda} p(X_{t1}, \dots, X_{tn} \mid \rho_t = \lambda) \mathbb{P}(\rho_t = \lambda \mid \rho_{t-1})$$

where, the sum runs over all partitions λ of n elements. Each summand in the sum above is given by the product of two factors. For the first factor, we have trivially that:

$$p(X_{t1}, \dots, X_{tn} \mid \rho_t = \lambda) = p(X_{t\sigma(1)}, \dots, X_{t\sigma(n)} \mid \rho_t = \sigma(\lambda))$$

for any permutation σ of n elements. For what concerns the second factor, the conditional distribution $\mathbb{P}[\rho_t = \lambda \mid \rho_{t-1}]$ is defined by the introduction of the binary latent variables in $\gamma_t = (\gamma_{t1}, \dots, \gamma_{tn})$. The latent variables identify which subjects at time $t - 1$ will be considered for possible cluster reallocation at time t . Specifically, let γ_{it} be defined as

$$\gamma_{it} = \begin{cases} 1 & \text{if unit } i \text{ is not reallocated when moving from time } t - 1 \text{ to } t \\ 0 & \text{otherwise} \end{cases}$$

so that

$$\mathbb{P}[\rho_t = \lambda \mid \rho_{t-1}] = \sum_{\gamma_t} \mathbb{P}[\rho_t = \lambda \mid \gamma_t, \rho_{t-1}] p(\gamma_t)$$

where the sum runs over all binary vectors of length n and $p(\gamma_t) = \alpha_t^{\sum_{i=1}^n \gamma_{ti}}$. Each summand in the sum above is given by the product of two factors. The second factor $p(\gamma_t)$ is invariant with respect to any permutation σ of n elements.

Thus, denoting with $\sigma(\gamma_t)$ the vector $(\gamma_{\sigma(1)t}, \dots, \gamma_{\sigma(n)t})$, for any permutation σ of n elements, to prove that $(X_{ti})_{i \geq 1}$ is conditionally partially exchangeable with respect to ρ_{t-1} , we need to prove that

$$\mathbb{P}[\rho_t = \lambda \mid \gamma_t, \rho_{t-1}] = \mathbb{P}[\rho_t = \sigma(\lambda) \mid \sigma(\gamma_t), \rho_{t-1}]$$

for any $\sigma \in \mathcal{P}(n; \rho_{t-1})$.

In t-RPM, the left and right hand side of the equation above are respectively

$$\mathbb{P}[\rho_t = \lambda \mid \gamma_t, \rho_{t-1}] = \frac{\mathbb{P}[\rho_t = \lambda] \mathbb{I}(\lambda \in P(\gamma_t, \rho_{t-1}))}{\sum_{\lambda'} \mathbb{P}[\rho_t = \lambda'] \mathbb{I}(\lambda' \in P(\gamma_t, \rho_{t-1}))}$$

and

$$\mathbb{P}[\rho_t = \sigma(\lambda) \mid \sigma(\gamma_t), \rho_{t-1}] = \frac{\mathbb{P}[\rho_t = \sigma(\lambda)] \mathbb{I}(\sigma(\lambda) \in P(\sigma(\gamma_t), \rho_{t-1}))}{\sum_{\lambda'} \mathbb{P}[\rho_t = \sigma(\lambda')] \mathbb{I}(\sigma(\lambda') \in P(\sigma(\gamma_t), \rho_{t-1}))}$$

where, the sums at the denominators runs over all partitions λ' of n elements, \mathbb{I} is the indicator function, and $P(\gamma_t, \rho_{t-1})$ denotes the collection of partitions at time t that are compatible with ρ_{t-1} based on γ_t . This collection is the one denoted by P_{C_t} in the paper of [Page et al. \(2022\)](#).

By marginal exchangeability of ρ_t , we have that for any σ

$$\mathbb{P}[\rho_t = \lambda] = \mathbb{P}[\rho_t = \sigma(\lambda)]$$

Consider now the indication functions $\mathbb{I}(\lambda \in P(\gamma_t, \rho_{t-1}))$ and let $\mathcal{R}_t = \{i : \gamma_{it} = 1\}$ be the sets of indices of those subjects which will not be considered for reallocation time t . [Page et al. \(2022\)](#) show that

$$\mathbb{I}(\lambda \in P(\gamma_t, \rho_{t-1})) = \begin{cases} 1 & \lambda^{\mathcal{R}_t} = \rho_{t-1}^{\mathcal{R}_t} \\ 0 & \text{otherwise} \end{cases}$$

and, thus

$$\mathbb{I}(\sigma(\lambda) \in P(\sigma(\gamma_t), \rho_{t-1})) = \begin{cases} 1 & \sigma(\lambda)^{\sigma(\mathcal{R}_t)} = \rho_{t-1}^{\sigma(\mathcal{R}_t)} \\ 0 & \text{otherwise} \end{cases}$$

where $\rho^{\mathcal{R}_t}$ is the *reduced* partition obtained removing from the sets in ρ all elements that are not in the set \mathcal{R}_t .

By Lemma 1, we have

$$\sigma(\lambda)^{\sigma(\mathcal{R}_t)} = \rho_{t-1}^{\sigma(\mathcal{R}_t)} \quad \text{iff} \quad \sigma(\lambda)^{\mathcal{R}_t} = \rho_{t-1}^{\mathcal{R}_t} \quad \text{iff} \quad \lambda^{\mathcal{R}_t} = \sigma^{-1}(\rho_{t-1}^{\mathcal{R}_t})$$

Therefore, by Lemma 1, $\mathbb{I}(\lambda \in P(\gamma_t, \rho_{t-1})) = \mathbb{I}(\sigma(\lambda) \in P(\sigma(\gamma_t), \rho_{t-1}))$ for any possible realization of γ_t if and only if

$$\rho_{t-1}^{\mathcal{R}_t} = \sigma^{-1}(\rho_{t-1}^{\mathcal{R}_t}) \quad \text{iff} \quad \sigma(\rho_{t-1}^{\mathcal{R}_t}) = \rho_{t-1}^{\mathcal{R}_t} \quad \text{iff} \quad \sigma \in \mathcal{P}(n; \rho)$$

which proves that t-RPM mixtures are conditionally partially exchangeable.

To prove that t-RPM mixture are not conditionally exchangeable, consider the counterexample with $n = 3$, $\rho_{t-1} = \{\{1, 2\}, \{3\}\}$, $\rho_t = \{\{1\}, \{2, 3\}\}$, and $\sigma = (1, 3)$. In such a case, the permutation σ does not preserve ρ_{t-1} and, as a result, the law of ρ_t conditionally of $\rho_{t-1} = \{\{1, 2\}, \{3\}\}$ differs from the law of ρ_t conditionally of $\rho_{t-1} = \{\{1\}, \{2, 3\}\}$. As a result, the conditional law of the sequence of observations at time t is not invariant to any permutation, as prescribed by conditional exchangeability. \square

A3 Proof of Proposition 2

Before proving Proposition 2, we first introduce the following Lemma.

Lemma 2. *Given a partitions ρ of n elements and a permutation σ ,*

$$\sigma \in \mathcal{P}(n; \rho) \quad \text{iff} \quad \sigma^{-1} \in \mathcal{P}(n; \rho)$$

where σ^{-1} denotes the inverse of σ , i.e., $\sigma^{-1}(i) = j$, for j such that $\sigma(j) = i$ and $\mathcal{P}(n; \rho)$ denotes the space of permutations of n elements that preserve ρ , cf. Section 2.

Proof of Proposition 2. If $(X_{1i}, \dots, X_{ji})_{i \geq 1}$ follows the separate exchangeable random partition mixture of [Rebaudo et al. \(2021\)](#), then

$$\begin{aligned} X_{ji} \mid S_j = k, M_{ik} = \ell &\stackrel{\text{ind}}{\sim} k(X_{ji}, \theta_\ell^*) \quad \text{for } i = 1, 2, \dots \text{ and } j = 1, \dots, J \\ \mathbb{P}(M_{ik} = \ell \mid w_{k\ell}) &= w_{k\ell} \quad \mathbf{w}_k = (w_{k1}, w_{k2}, \dots) \stackrel{\text{iid}}{\sim} \text{GEM}(\alpha) \\ \mathbb{P}(S_j = k \mid \pi_k) &= \pi_k \quad \boldsymbol{\pi} = (\pi_1, \pi_2, \dots) \sim \text{GEM}(\beta) \\ \theta_\ell^* &\stackrel{\text{iid}}{\sim} G_0 \end{aligned}$$

where $\text{GEM}(\alpha)$ denote a stick-breaking prior for a sequence of weights ([Sethuraman 1994](#)) and G_0 is an absolutely continuous distribution. The partition ρ_j corresponding to the j th layer $(X_{ji})_{i \geq 1}$ is encoded by $(M_{iS_j})_{i \geq 1}$ and for any $n \geq 1$, $j, j' \in [J]$ and any realization ρ of the partition $\rho_{j'}$, we have

$$p(X_{j1}, \dots, X_{jn} \mid \rho_{j'} = \rho) = \mathbb{P}[S_j = S_{j'}] p(X_{j1}, \dots, X_{jn} \mid \rho_j = \rho) + \mathbb{P}[S_j \neq S_{j'}] p(X_{j1}, \dots, X_{jn})$$

and, similarly,

$$\begin{aligned} p(X_{j\sigma(1)}, \dots, X_{j\sigma(n)} \mid \rho_{j'} = \rho) &= \mathbb{P}[S_j = S_{j'}] p(X_{j\sigma(1)}, \dots, X_{j\sigma(n)} \mid \rho_j = \rho) \\ &\quad + \mathbb{P}[S_j \neq S_{j'}] p(X_{j\sigma(1)}, \dots, X_{j\sigma(n)}). \end{aligned}$$

Thus,

$$\begin{aligned} D &:= p(X_{j1}, \dots, X_{jn} \mid \rho_{j'} = \rho) - p(X_{j\sigma(1)}, \dots, X_{j\sigma(n)} \mid \rho_{j'} = \rho) \\ &= \mathbb{P}[S_j = S_{j'}] (p(X_{j1}, \dots, X_{jn} \mid \rho_j = \rho) - p(X_{j\sigma(1)}, \dots, X_{j\sigma(n)} \mid \rho_j = \rho)) \\ &= \mathbb{P}[S_j = S_{j'}] (p(X_{j1}, \dots, X_{jn} \mid \rho_j = \rho) - p(X_{j1}, \dots, X_{jn} \mid \rho_j = \sigma^{-1}(\rho))) \end{aligned}$$

By Lemma 2, for any $\sigma \in \mathcal{P}(n; \rho)$, we have $D = 0$, where, we recall, that $\mathcal{P}(n; \rho)$ denotes the space of permutations of n elements that preserve ρ , see Definition 1 in Section 2.

To prove that the separate exchangeable random partition mixture is not conditionally exchangeable, consider the counterexample with $n = 3$, $\rho_{j'} = \{\{1, 2\}, \{3\}\}$, $\sigma = (1, 3)$ and $(X_{j1}, X_{j2}, X_{j3}) \in (d(\bar{\theta}_\ell^* - \epsilon), d(\bar{\theta}_\ell^* + \epsilon), d(\bar{\theta}_\ell^* + 2\epsilon))$, where $\bar{\theta}_\ell^* = \mathbb{E}[\theta_\ell^*]$, $\epsilon > 0$ and $dy = [y, y + \nu)$, with ν arbitrarily small. In such a case, the permutation σ does not preserve $\rho_{j'}$ and, as a result, the law of ρ_j conditionally of $\rho_{j'} = \{\{1, 2\}, \{3\}\}$ differs from the law of ρ_j conditionally of $\rho_{j'} = \{\{1\}, \{2, 3\}\}$. The conditional law of the sequence of observations corresponding to ρ_j is not invariant to any permutation, as prescribed by conditional exchangeability. \square

A4 Proof of Proposition 3

Proof of Proposition 3. Denoting with X_{ix} the response measured on the i th unit corresponding to covariate's value $x \in \mathcal{X}$ and following a mixture model with mixing probability provided by the dependent processes in [MacEachern \(2000\)](#), then

$$\begin{aligned} X_{ix} \mid \boldsymbol{\theta}_x^* &\stackrel{iid}{\sim} k(X_{ix}, \theta_{xi}) \quad \text{for } i = 1, \dots, n \text{ and for any } x \\ \theta_{xi} &\stackrel{ind}{\sim} G_x \\ \{G_x : x \in \mathcal{X}\} &\sim DDP \end{aligned}$$

For a formal and detailed definition of $\{G_x : x \in \mathcal{X}\} \sim DDP$ we refer to the recent review paper of [Quintana et al. \(2022\)](#).

Denoting with ρ_x the partition induced by G_x , for any σ permutation of n elements, we have

$$\begin{aligned} p(X_{x'1}, \dots, X_{x'n} \mid \rho_x) &= \int p(X_{x'1}, \dots, X_{x'n} \mid G_{x'}, \rho_x) d p(G_{x'} \mid \rho_x) \\ &= \int p(X_{x'1}, \dots, X_{x'n} \mid G_{x'}) d p(G_{x'} \mid \rho_x) = \int p(X_{x'\sigma(1)}, \dots, X_{x'\sigma(n)} \mid G_{x'}) d p(G_{x'} \mid \rho_x) \\ &= p(X_{x'\sigma(1)}, \dots, X_{x'\sigma(n)} \mid \rho_x). \end{aligned}$$

□

A5 Proof of Theorem 2

Proof of Theorem 2. Note that, for any $n \geq 1$, the second layer observations, admit the following almost sure representation in terms of a latent collection of probability measures $(\tilde{q}_1, \dots, \tilde{q}_n)$ such as

$$X_{2i} \mid \tilde{q}_i \stackrel{ind}{\sim} \int k_2(X_{2i}; \xi) \tilde{q}_i(d\xi), \quad \text{where } \tilde{q}_i \mid \mathbf{w}, \tilde{p}_{21}, \dots, \tilde{p}_{2M} \stackrel{iid}{\sim} \sum_{m=1}^M w_m \delta_{\tilde{p}_{2m}}$$

and \mathbf{w} is the sequence of weights in the almost-sure representation of \tilde{p}_1 . Moreover, conditioning both layers to the allocations variables \mathbf{c}_1 and the unique values $\boldsymbol{\theta}^*$ corresponding to the first layer, we get

$$(X_{1i}, X_{2i}) \mid c_{1i} = m, \theta_m^*, \tilde{p}_{21}, \dots, \tilde{p}_{2M} \stackrel{ind}{\sim} k_1(X_{1i}; \theta_m^*) \left(\sum_{s=1}^S q_{ms} k_2(X_{2i}; \xi_s^*) \right).$$

□

A6 Proof of Proposition 4

Proof of Proposition 4. Note that, for any $i \neq j$, by exchangeability of the rows in the data matrix, we have

$$\mathbb{P}(c_{\ell i} = c_{\ell j}) = \mathbb{P}(c_{\ell 1} = c_{\ell 2}) \quad \text{and} \quad \mathbb{P}(c_{\ell i} = c_{\ell j}, c_{\ell' i} = c_{\ell' j}) = \mathbb{P}(c_{\ell 1} = c_{\ell 2}, c_{\ell' 1} = c_{\ell' 2})$$

Thus

$$\tau = \frac{\mathbb{P}[c_{21} = c_{22} \mid c_{11} = c_{12}] - \mathbb{P}[c_{21} = c_{22} \mid c_{11} \neq c_{12}]}{\mathbb{P}[c_{21} = c_{22} \mid c_{11} = c_{12}]}$$

where the event $c_{\ell 1} = c_{\ell 2}$ coincides with the event $K_{\ell 2} = 1$ and $c_{\ell 1} \neq c_{\ell 2}$ with the event $K_{\ell 2} = 2$, where $K_{\ell n}$ denote the number of cluster at layer ℓ in a sample of n subjects. Similarly,

$$\begin{aligned} ER &= \binom{n}{2}^{-1} \mathbb{E} \left[\sum_{i=1}^n \sum_{j=i+1}^n \mathbb{1}(c_{1i} = c_{1j}) \mathbb{1}(c_{2i} = c_{2j}) + \sum_{i=1}^n \sum_{j=i+1}^n \mathbb{1}(c_{1i} \neq c_{1j}) \mathbb{1}(c_{2i} \neq c_{2j}) \right] \\ &= \mathbb{P}(c_{1i} = c_{1j}, c_{2i} = c_{2j}) + \mathbb{P}(c_{1i} \neq c_{1j}, c_{2i} \neq c_{2j}) \\ &= \mathbb{P}(c_{11} = c_2, s_1 = s_2) + \mathbb{P}(c_{11} \neq c_2, s_1 \neq s_2) \\ &= \mathbb{P}(K_{12} = 1, K_{22} = 1) + \mathbb{P}(K_{12} = 2, K_{22} = 2). \end{aligned}$$

□

A7 Proof of Theorem 3 and Corollary 1

Proof of Theorem 3 follows directly by combining equations (6) and (7) in Section 4. Corollary 1 follows directly from Theorem 3 and Proposition 4.

A8 Proof of Theorem 4 and Corollary 2

Proof of Theorem 4. The marginal EPPF of the partition at layer 1 is a well-known result (see, e.g., [Green & Richardson 2001](#), [McCullagh & Yang 2008](#), [Miller & Harrison 2018](#), [Argiento & De Iorio 2022](#)). Considering a specific partition ρ_1 into K_{1n} sets of the n observations, under eq. (8) in Section 5, we have that

$$p(\rho_1) = V(n, K_{1n}) \prod_{m=1}^{K_{1n}} \frac{\Gamma(\gamma + n_m)}{\Gamma(\gamma)},$$

where n_m is the frequency of the m th cluster in order of appearance, i.e.,

$$n_m = \sum_{i=1}^n \mathbb{1}_m(c_{1i}^*) \text{ with } \sum_{m=1}^{K_{1n}} n_m = n \quad \text{and} \quad V(n, K_{1n}) = \sum_{M=1}^{+\infty} \frac{M_{(K_{1n})}}{(\gamma K_{1n})^{(n)}} p_M(M)$$

where $x^{(k)} = \Gamma(x+k)/\Gamma(x) = x(x+1)\dots(x+k-1)$ and $x_{(k)} = \Gamma(x+1)/\Gamma(x-k+1) = x(x-1)\dots(x-k+1)$, where $\Gamma(x)$ denote the Gamma function in x and $x_{(0)} = 1$ and $x_{(0)} = 1$ by convention. While from equation (9), we have that

$$p(\rho_2 \mid \rho_1) = (1 - \omega) \mathbb{1}(\rho_1 = \rho_2) + \omega V(n, K_{2n}) \prod_{s=1}^{K_{2n}} \frac{\Gamma(\alpha + \sum_{m=1}^{K_{1n}} n_{ms})}{\Gamma(\alpha)}$$

where n_{ms} is the number of observations in the first-layer cluster m and second-layer cluster s , when the clusters are in order of appearance.

Proof of Theorem 4 follows directly by combining the two partition functions above. Corollary 2 follows directly from Theorem 4 and Proposition 4. □

References

Amini, A. A., Paez, M., Lin, L. & Razaee, Z. S. (2019), ‘Exact slice sampler for hierarchical Dirichlet processes’, *arXiv preprint arXiv:1903.08829*.

- Argiento, R., Cremaschi, A. & Vannucci, M. (2020), ‘Hierarchical normalized completely random measures to cluster grouped data’, *Journal of the American Statistical Association* **115**(529), 318–333.
- Argiento, R. & De Iorio, M. (2022), ‘Is infinity that far? a Bayesian nonparametric perspective of finite mixture models’, *The Annals of Statistics* **50**(5), 2641–2663.
- Ascolani, F., Franzolini, B., Lijoi, A. & Prünster, I. (2024), ‘Nonparametric priors with full-range borrowing of information’, *Biometrika* **111**(3), 945–969.
- Ascolani, F., Lijoi, A. & Ruggiero, M. (2021), ‘Predictive inference with Fleming–Viot-driven dependent Dirichlet processes’, *Bayesian Analysis* **16**(2), 371–395.
- Barrios, E., Lijoi, A., Nieto-Barajas, L. E. & Prünster, I. (2013), ‘Modeling with normalized random measure mixture models’, *Statistical Science* **28**(3), 313 – 334.
- Bassetti, F., Casarin, R. & Rossini, L. (2020), ‘Hierarchical species sampling models’, *Bayesian Analysis* **15**(3), 809–838.
- Beraha, M., Guglielmi, A. & Quintana, F. A. (2021), ‘The semi-hierarchical Dirichlet process and its application to clustering homogeneous distributions’, *Bayesian Analysis* **16**(4), 1187–1219.
- Betancourt, B., Zanella, G. & Steorts, R. C. (2022), ‘Random partition models for micro-clustering tasks’, *Journal of the American Statistical Association* **117**(539), 1215–1227.
- Camerlenghi, F., Lijoi, A., Orbanz, P. & Prünster, I. (2019), ‘Distribution theory for hierarchical processes’, *The Annals of Statistics* **47**(1), 67–92.
- Camerlenghi, F., Lijoi, A. & Prünster, I. (2018), ‘Bayesian nonparametric inference beyond the Gibbs-type framework’, *Scandinavian Journal of Statistics* **45**(4), 1062–1091.
- Caron, F., Davy, M. & Doucet, A. (2007), Generalized polya urn for time-varying Dirichlet process mixtures, in ‘Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence’, pp. 33–40.
- Caron, F., Neiswanger, W., Wood, F., Doucet, A. & Davy, M. (2017), ‘Generalized Pólya urn for time-varying Pitman-Yor processes’, *Journal of Machine Learning Research* **18**, 1–32.
- Chandra, N. K., Canale, A. & Dunson, D. B. (2023), ‘Escaping the curse of dimensionality in Bayesian model based clustering’, *Journal of Machine Learning Research* **24**, 1–42.
- Chen, M.-S., Lin, J.-Q., Li, X.-L., Liu, B.-Y., Wang, C.-D., Huang, D. & Lai, J.-H. (2022), ‘Representation learning in multi-view clustering: A literature review’, *Data Science and Engineering* **7**(3), 225–241.
- Cremaschi, A., De Iorio, M., Kothandaraman, N., Yap, F., Tint, M. T. & Eriksson, J. (2024), ‘Joint modeling of association networks and longitudinal biomarkers: An application to child obesity’, *Statistics in Medicine* **43**(6), 1135–1152.
- Dahl, D. B., Day, R. & Tsai, J. W. (2017), ‘Random partition distribution indexed by pairwise information’, *Journal of the American Statistical Association* **112**(518), 721–732.

- Dalrymple, K. V., Thompson, J. M., Begum, S., Godfrey, K. M., Poston, L., Seed, P. T., McCowan, L. M., Wall, C., Shelling, A., North, R. et al. (2019), ‘Relationships of maternal body mass index and plasma biomarkers with childhood body mass index and adiposity at 6 years: The children of scope study’, *Pediatric obesity* **14**(10), e12537.
- Das, S., Niu, Y., Ni, Y., Mallick, B. K. & Pati, D. (2024), ‘Blocked Gibbs sampler for hierarchical Dirichlet processes’, *Journal of Computational and Graphical Statistics (in press)*.
- De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prunster, I. & Ruggiero, M. (2015), ‘Are Gibbs-type priors the most natural generalization of the Dirichlet process?’, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**(2), 212–229.
- De Finetti, B. (1937), La prévision: ses lois logiques, ses sources subjectives, in ‘Annales de l’institut Henri Poincaré’, Vol. 7, pp. 1–68.
- de Finetti, B. (1938), ‘Sur la condition d’équivalence partielle’, *Actualités Scientifiques et Industrielles* **739**, 5–18.
- Denti, F., Camerlenghi, F., Guindani, M. & Mira, A. (2021), ‘A common atoms model for the Bayesian nonparametric analysis of nested data’, *Journal of the American Statistical Association* **118**(541), 1–12.
- DeYoreo, M. & Kottas, A. (2018), ‘Modeling for dynamic ordinal regression relationships: An application to estimating maturity of rockfish in California’, *Journal of the American Statistical Association* **113**(521), 68–80.
- Dunson, D. B. (2010), Nonparametric Bayes applications to biostatistics, in ‘Bayesian Nonparametrics’, Cambridge Univ. Press, pp. 223–273.
- Dunson, D. B. & Park, J.-H. (2008), ‘Kernel stick-breaking processes’, *Biometrika* **95**(2), 307–323.
- Ferguson, T. S. (1973), ‘A Bayesian analysis of some nonparametric problems’, *The Annals of Statistics* **1**(2), 209–230.
- Ferguson, T. S. (1983), Bayesian density estimation by mixtures of normal distributions, in ‘Recent advances in statistics’, Elsevier, pp. 287–302.
- Foti, N. J. & Williamson, S. A. (2015), ‘A survey of non-exchangeable priors for Bayesian nonparametric models’, *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(2), 359–371.
- Franzolini, B. (2022), *On Dependent Processes in Bayesian Nonparametrics: Theory, Methods, and Applications*, Bocconi University.
- Franzolini, B., Cremaschi, A., van den Boom, W. & De Iorio, M. (2023), ‘Bayesian clustering of multiple zero-inflated outcomes’, *Philosophical Transactions of the Royal Society A* **81**(2247), 20220145.
- Franzolini, B., Lijoi, A., Prünster, I. & Rebaudo, G. (2023), ‘Multivariate species sampling processes’, *Working Paper*.
- Gao, L. L., Bien, J. & Witten, D. (2020), ‘Are clusterings of multiple data views independent?’, *Biostatistics* **21**(4), 692–708.

- Green, P. J. & Richardson, S. (2001), ‘Modelling heterogeneity with and without the Dirichlet process’, *Scandinavian Journal of Statistics* **28**(2), 355–375.
- Griffin, J. E. & Leisen, F. (2017), ‘Compound random measures and their use in Bayesian non-parametrics’, *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **79**(2), 525–545.
- Handakas, E., Lau, C. H., Alfano, R., Chatzi, V. L., Plusquin, M., Vineis, P. & Robinson, O. (2022), ‘A systematic review of metabolomic studies of childhood obesity: State of the evidence for metabolic determinants and consequences’, *Obesity Reviews* **23**, e13384.
- Hellmuth, C., Kirchberg, F. F., Brandt, S., Moß, A., Walter, V., Rothenbacher, D., Brenner, H., Grote, V., Gruszfeld, D., Socha, P. et al. (2019), ‘An individual participant data meta-analysis on metabolomics profiles for obesity and insulin resistance in european children’, *Scientific reports* **9**(1), 5053.
- Hubert, L. & Arabie, P. (1985), ‘Comparing partitions’, *Journal of classification* **2**(1), 193–218.
- Ishwaran, H. & James, L. F. (2001), ‘Gibbs sampling methods for stick-breaking priors’, *Journal of the American statistical Association* **96**(453), 161–173.
- Josefson, J. L., Catalano, P. M., Lowe, W. L., Scholtens, D. M., Kuang, A., Dyer, A. R., Lowe, L. P. & Metzger, B. E. (2020), ‘The joint associations of maternal bmi and glycemia with childhood adiposity’, *The Journal of Clinical Endocrinology & Metabolism* **105**(7), 2177–2188.
- Kalli, M., Griffin, J. E. & Walker, S. G. (2011), ‘Slice sampling mixture models’, *Statistics and computing* **21**, 93–105.
- Kumar, A., Rai, P. & Daume, H. (2011), ‘Co-regularized multi-view spectral clustering’, *Advances in neural information processing systems* **24**.
- Landon, M. B., Mele, L., Varner, M. W., Casey, B. M., Reddy, U. M., Wapner, R. J., Rouse, D. J., Tita, A. T., Thorp, J. M., Chien, E. K. et al. (2020), ‘The relationship of maternal glycemia to childhood obesity and metabolic dysfunction’, *The Journal of Maternal-Fetal & Neonatal Medicine* **33**(1), 33–41.
- Lee, J., Müller, P., Zhu, Y. & Ji, Y. (2013), ‘A nonparametric Bayesian model for local clustering with application to proteomics’, *Journal of the American Statistical Association* **108**(503), 775–788.
- Lijoi, A., Nipoti, B. & Prünster, I. (2014), ‘Bayesian inference with dependent normalized completely random measures’, *Bernoulli* **20**(3), 1260–1291.
- Lijoi, A., Prünster, I. & Rebaudo, G. (2023), ‘Flexible clustering via hidden hierarchical Dirichlet priors’, *Scandinavian Journal of Statistics* **50**(1), 213–234.
- Lo, A. Y. (1984), ‘On a class of Bayesian nonparametric estimates: I. Density estimates’, *The Annals of Statistics* **12**(1), 351–357.
- Lock, E. F. & Dunson, D. B. (2013), ‘Bayesian consensus clustering’, *Bioinformatics* **29**(20), 2610–2616.

- MacEachern, S. N. (2000), Dependent Dirichlet processes, Technical report, The Ohio State Univ.
- McCullagh, P. & Yang, J. (2008), ‘How many clusters?’, *Bayesian Analysis* **3**(1), 101–120.
- Meek, C. L. (2023), ‘An unwelcome inheritance: childhood obesity after diabetes in pregnancy’, *Diabetologia* **66**(11), 1961–1970.
- Meilă, M. (2007), ‘Comparing clusterings—an information based distance’, *Journal of Multivariate Analysis* **98**(5), 873–895.
- Miller, J. W. & Harrison, M. T. (2018), ‘Mixture models with a prior on the number of components’, *Journal of the American Statistical Association* **113**(521), 340–356.
- Müller, P., Quintana, F. & Rosner, G. (2004), ‘A method for combining inference across related nonparametric Bayesian models’, *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **66**(3), 735–749.
- Neal, R. M. (2000), ‘Markov chain sampling methods for Dirichlet process mixture models’, *Journal of computational and graphical statistics* **9**(2), 249–265.
- Nobile, A. (1994), *Bayesian analysis of finite mixture distributions*, Carnegie Mellon University.
- Ormindean, C. M., Ciortea, R., Bucuri, C. E., Măluțan, A. M., Iuhas, C. I., Porumb, C. G., Ormindean, V., Roman, M. P., Nati, I. D., Suci, V. et al. (2024), ‘Obesity, a single pathology influencing both mother and child—a retrospective analysis in hospital settings’, *Journal of Personalized Medicine* **14**(7), 683.
- Page, G. L., Quintana, F. A. & Dahl, D. B. (2022), ‘Dependent modeling of temporal sequences of random partitions’, *Journal of Computational and Graphical Statistics* **31**(2), 614–627.
- Perng, W., Gillman, M. W., Fleisch, A. F., Michalek, R. D., Watkins, S. M., Isganaitis, E., Patti, M.-E. & Oken, E. (2014), ‘Metabolomic profiles and childhood obesity’, *Obesity* **22**(12), 2570–2578.
- Perng, W., Rifas-Shiman, S. L., Sordillo, J., Hivert, M.-F. & Oken, E. (2020), ‘Metabolomic profiles of overweight/obesity phenotypes during adolescence: a cross-sectional study in project viva’, *Obesity* **28**(2), 379–387.
- Petrone, S., Guindani, M. & Gelfand, A. E. (2009), ‘Hybrid Dirichlet mixture models for functional data’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(4), 755–782.
- Pitman, J. (1996), ‘Some developments of the Blackwell-MacQueen urn scheme’, *Lecture Notes-Monograph Series* pp. 245–267.
- Pitman, J. & Yor, M. (1997), ‘The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator’, *The Annals of Probability* **25**(2), 855–900.
- Quintana, F. A., Müller, P., Jara, A. & MacEachern, S. N. (2022), ‘The dependent Dirichlet process and related models’, *Statistical Science* **37**(1), 24–41.

- Rebaudo, G., Lin, Q. & Mueller, P. (2021), ‘Separate exchangeability as modeling principle in Bayesian nonparametrics’, *arXiv preprint arXiv:2112.07755*.
- Ren, L., Du, L., Carin, L. & Dunson, D. B. (2011), ‘Logistic stick-breaking process.’, *Journal of Machine Learning Research* **12**, 203–239.
- Ren, L., Dunson, D. B. & Carin, L. (2008), The dynamic hierarchical Dirichlet process, in ‘Proceedings of the 25th international conference on machine learning’, pp. 824–831.
- Rigon, T. & Durante, D. (2021), ‘Tractable Bayesian density regression via logit stick-breaking priors’, *Journal of Statistical Planning and Inference* **211**, 131–142.
- Rodriguez, A. & Dunson, D. B. (2011), ‘Nonparametric Bayesian models through probit stick-breaking processes’, *Bayesian analysis* **6**(1), 45–178.
- Rodríguez, A., Dunson, D. B. & Gelfand, A. E. (2010), ‘Latent stick-breaking processes’, *Journal of the American Statistical Association* **105**(490), 647–659.
- Rogers, S., Girolami, M., Kolch, W., Waters, K. M., Liu, T., Thrall, B. & Wiley, H. S. (2008), ‘Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models’, *Bioinformatics* **24**(24), 2894–2900.
- Schipper, M. C., Blaauwendraad, S. M., Koletzko, B., Oei, E. H., Jaddoe, V. W. & Gaillard, R. (2024), ‘Associations of childhood bmi, general and visceral fat mass with metabolite profiles at school-age’, *International Journal of Obesity* pp. 1–11.
- Sethuraman, J. (1994), ‘A constructive definition of Dirichlet priors’, *Statistica Sinica* **4**(2), 639–650.
- Soh, S.-E., Tint, M. T., Gluckman, P. D., Godfrey, K. M., Rifkin-Graboi, A., Chan, Y. H., Stümel, W., Holbrook, J. D., Kwek, K., Chong, Y.-S. et al. (2014), ‘Cohort profile: Growing Up in Singapore Towards healthy Outcomes (GUSTO) birth cohort study’, *International journal of epidemiology* **43**(5), 1401–1409.
- Taddy, M. A. (2010), ‘Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking intensity of violent crime’, *Journal of the American Statistical Association* **105**(492), 1403–1417.
- Teh, Y., Jordan, M., Beal, M. & Blei, D. (2006), ‘Hierarchical Dirichlet processes’, *Journal of the American Statistical Association* **101**(476), 1566–1581.
- Tibshirani, R., Walther, G. & Hastie, T. (2001), ‘Estimating the number of clusters in a data set via the gap statistic’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(2), 411–423.
- Wade, S., Mongelluzzo, S. & Petrone, S. (2011), ‘An enriched conjugate prior for Bayesian nonparametric inference’, *Bayesian Analysis* **6**(3), 359–385.
- Walker, S. G. (2007), ‘Sampling the Dirichlet mixture model with slices’, *Communications in Statistics—Simulation and Computation* **36**(1), 45–54.
- Yang, Y. & Wang, H. (2018), ‘Multi-view clustering: A survey’, *Big Data Mining and Analytics* **1**(2), 83–107.

- Yao, S., Yu, G., Wang, J., Domeniconi, C. & Zhang, X. (2019), Multi-view multiple clustering, *in* ‘Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence’, International Joint Conferences on Artificial Intelligence Organization, pp. 4121–4127.
- Zhou, D., Gao, Y. & Paninski, L. (2021), Disentangled sticky hierarchical Dirichlet process hidden Markov model, *in* ‘Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020 Proceedings, Part I’, pp. 612–627.