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

Beatrice Franzolini[1], Maria De Iorio[2], and Johan Eriksson[3]

[1,2,3]Singapore Institute for Clinical Sciences (SICS), Agency for Science, Technology and Research (A*STAR), Singapore, Republic of Singapore
[2]Department of Statistical Science, University College London, London, UK
[2,3]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Republic of Singapore

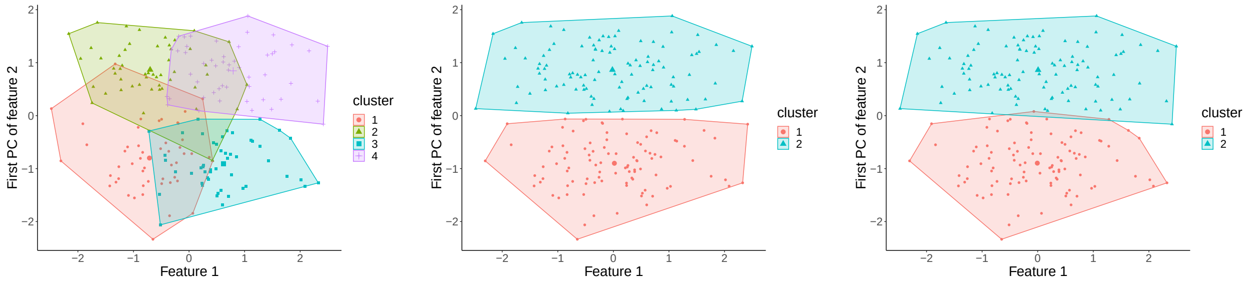
Abstract

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

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. Moreover, more and more frequently, datasets present multivariate information collected across distinct domains, with possibly different features support spaces. Data of this type are typically referred to as *multi-source* or *multi-view* data (see, for instance, Yang and Wang, 2018). Examples include samples of webpages whose characteristics are described by text, images, and videos, samples of physical objects represented by texture, shape and colour, or, as in the application we consider here, a cohort study of children for whom we consider growth trajectories, metabolites' concentration and clinical information about their mother. This type of data requires a more flexible approach to clustering, where clusters' shapes and definitions might change from feature to feature. A unique clustering configuration based on all the observed features



(a) True clustering structure. Observations in each cluster are simulated from a Multivariate Normal with identity variance and covariance matrix.

(b) k-means clustering configuration with the number of clusters determined by elbow plot, gap statistics (Tibshirani et al., 2001), and silhouette method (Kaufman and Rousseeuw, 2009).

(c) Dirichlet process mixture estimate of the clustering configuration obtained minimising the variation of information loss function (Meilă, 2007).

Figure 1: Toy example. Data simulated on two variables for 200 subjects, with true clustering structure displayed in panel (a). The first feature is one-dimensional, while the second feature is three-dimensional. 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 appear to ignore the information contained in the first feature, even though marginally the first feature has the same distribution of each dimension of the second feature.

not only may be hard to detect and interpret (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 shared by the different features, down-weighting the idiosyncratic contribution of each feature. Moreover, when standard clustering techniques are applied to multiple features, the common clustering solution also depends on each feature's dimension, favouring higher dimensional features as more important in explaining the heterogeneity across subjects, which is typically not desirable behaviour. See Figure 1 for a toy example illustrating this problem. A similar issue has also been noted in the literature on product partition models with covariates (Page and Quintana, 2018), when the covariates space is large compared to the response and inference on clustering is dominated by the covariates. 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, introducing dependence across the different clustering configurations.

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 the popular mixture model. The components in a mixture model can be thought of as potential sub-populations. The clusters in a finite sample are instead the allocated components, i.e., components to which at least one observation from the sample has been assigned. On the other hand, algorithmic methods allocate items to clusters without using distributional assumptions, but relying on optimization techniques to find the configuration which groups together *similar* observations, typically maximizing some distance between clusters. In both contexts, the (explicit or implicit) definition of clusters becomes crucial, often depending on the specific application at hand and the goal of the analysis, rather than on the dataset per se (see e.g., Hennig, 2015). Unlike algorithmic techniques, model-based methods define the shape of a cluster in terms of probability distribution functions and, as a consequence, enable us to conduct probabilistic assessments, providing quantification of uncertainty and a natural framework for predictions. Most popular model-based approaches include Bayesian infinite mixture models (Ferguson, 1983; Lo, 1984; Barrios, Lijoi, Nieto-Barajas, and Prünster, Barrios et al.) and Bayesian mixtures with a random number of components (Nobile, 1994; Richardson and Green, 1997; Miller and Harrison, 2018; Argiento and 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. Moreover, Bayesian clustering methods based on mixtures are used not only to detect well-separated groups of observations, but also for dimensionality reduction (Blei et al., 2003; Petrone et al., 2009), outlier-detection (Shotwell and Slate, 2011; Ngan et al., 2015; Franzolini et al., 2023), testing for distributional homogeneity (Rodríguez et al., 2008; Camerlenghi et al., 2019; Denti et al., 2021; Beraha et al., 2021; Balocchi et al., 2021; Lijoi et al., 2023), and data pre-processing (Zhang et al., 2006). For a recent review of Bayesian cluster analysis and its differences from algorithmic approaches see Wade (2023).

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 and Wang, 2018; Chen et al., 2022, for comprehensive reviews of the topic). Nonetheless, such methods while recognising 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. Then, clustering is performed within each group with clusters possibly shared among groups. Note that there is no overlap of subjects across groups. These models are obtained by inducing dependence between the group-specific random probability measures in the underlying mixture model (see, for instance, MacEachern, 1999, 2000; Müller et al., 2004; Teh et al., 2006; Caron et al., 2007; Dunson and Park, 2008; Ren et al., 2008; Dunson, 2010; Rodríguez et al., 2010; Taddy, 2010; Rodriguez and Dunson, 2011; Lijoi et al., 2014; Foti and Williamson, 2015; Caron et al., 2017; Griffin and Leisen, 2017; DeYoreo and Kottas, 2018; De Iorio et al., 2019; 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 with different clustering configurations for the same subjects, such methods focus on marginal inference based on each feature and fail to capture the true nature of the multivariate dependence (cf., Page et al., 2022). In particular, in Section 2.2, we show how this is a consequence of the fact that they do not incorporate any individual-specific effect and ultimately ignore that subjects are indeed the same for all the observed features.

The Bayesian literature on clustering methods for multi-view information is rather limited. In this context, the core challenge is to define a probabilistic model able to account for within-subject dependence across multiple features potentially taking values in different support spaces. By within-subject dependence, we refer to the relationship between observations corresponding to different features, times, or locations, but associated with the same subject. This is a typical setup arising in many applications, e.g., longitudinal data, repeated measures or panel data (see, for instance, Davis, 2002). In standard parametric regression models, within-subject dependence is usually captured through individual-specific random effects, which typically are real-valued parameters. This simple setup is not applicable when either features take values on different support spaces or the goal is multivariate clustering of subjects. Bayesian clustering approaches that allow to both deal with within-subject dependence and provide multiple clustering configurations, appear limited to the following: 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 Lin et al. (2021) and the temporal random partition model of Page et al. (2022). Even though these models are quite different in nature, we show that they all belong to the general probabilistic framework we develop here, which serves also to provide novel insights about these existing models. For instance, our results prove that the model by Page et al. (2022) admits a conditional representation in terms of mixture models with almost-surely discrete mixing measures, conditional on which the data can be seen as a random sample. Moreover, in their original formulation, these models cannot be applied to datasets where features take values into different support spaces.

The main contribution of this work is to introduce the concept of *conditional partial exchangeability* (CPE) as a modelling principle for multi-view clustering. CPE is a probabilistic framework able to induce dependence between clustering configurations of the same subjects but based on different features, still capturing the specific contribution of each feature and the within-subject dependence. Moreover, we present and study a general class of Bayesian models based on CPE, designed to accomplish multi-view clustering through a tailored learning mechanism. We refer to this class as *telescopic clustering models*. We show that telescopic clustering models admit a representation in terms of mixture models with random mixing measures, are analytically and computationally tractable and establish the Kolomogorov consistency of the predictive distribution. In particular, we investigate two cases: (i) telescopic mixture models with hierarchical Dirichlet processes as mixing measures (Teh et al., 2006); (ii) telescopic mixtures based on a novel construction of finite mixtures with a random number of components.

The paper is structured as follows. Section 2 introduces conditional partial exchangeability. Section 3 derives the class of telescopic clustering models, their properties, and the measures of dependence between partitions, provides details for two specific models within the class of telescopic clustering models, and discusses the algorithms to derive posterior inference. Section 4 demonstrates the approach through numerical simulations. Section 5 presents a real data application. Section 6 concludes the paper with a discussion.

2 Conditional partial exchangeability

2.1 Data structure and clustering problem

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 3.4. We assume that

- $X_{1i} \in \mathbb{X}_1 \subset \mathbb{R}^d$ 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 $t = 1$,
- $X_{2i} \in \mathbb{X}_2 \subset \mathbb{R}^p$ 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 $t = 2$.

The support spaces $\mathbb{X}_1$ and $\mathbb{X}_2$ are endowed with the corresponding Borel σ-algebras and are not assumed to coincide. In particular, the dimensions d and p may be different, although the observational units are either partially or completely overlapping at different layers, i.e. there is a subset of subjects for which all the features have been observed. In the following, we always refer to the same set of observational units $i = 1, \dots, n$ at each layer while allowing for missing data in the case of partially overlapping samples across layers.

We assume row-exchangeability for $(X_{1i}, X_{2i})_{i\geq 1}$. Formally, $(X_{1i}, X_{2i})_{i\geq 1}$ is said *row exchangeable* if and only if

$$\mathbb{P}\left[(X_{1i}, X_{2i})_{i=1}^n \in A\right] = \mathbb{P}\left[(X_{1\sigma(i)}, X_{2\sigma(i)})_{i=1}^n \in A\right]$$

for any $\sigma \in \mathcal{P}(n)$, $n \geq 1$, and measurable set $A \subseteq (\mathbb{X}_1 \times \mathbb{X}_2)^n$, where $\mathcal{P}(n)$ denotes the set of permutations of n elements. Row-exchangeability reflects the common assumption that the order in which the subjects have been observed does not provide any additional information about the overall population or for the prediction of new subjects. For a comprehensive overview of exchangeability and its extensions, we refer the reader to Aldous et al. (1985) and Kallenberg (1989, 2005).

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. The partition ρ_j, $j = 1, 2$, can be represented by the vector $\boldsymbol{c}_j = (c_{j1}, \ldots, c_{jn})$ of subject-specific allocation variables, whose elements take value in the set $[n] := \{1, 2, \ldots, n\}$ and are such that $c_{ji} = c_{jl}$ if and only if subjects i and l belong to the same cluster at layer j.

2.2 Conditional partial exchangeability

Assume that the clustering configurations fully capture the dependence structure between first and second layers features, i.e.,

$$(X_{11}, \ldots, X_{1n}) \perp (X_{21}, \ldots, X_{2n}) \mid \rho_1, \rho_2 \tag{1}$$

This is a common assumption in clustering models for multivariate responses (see, e.g., Rogers et al., 2008; Kumar et al., 2011; Lock and Dunson, 2013; Gao et al., 2020; Franzolini et al., 2023) as it often avoids identifiability issues. Nonetheless, we discuss strategies to relax this assumption in Section 3.4. The row-exchangeability of the observations implies that each layer is *marginally exchangeable*:

$$\mathbb{P}\left[(X_{j1}, \ldots, X_{jn}) \in A\right] = \mathbb{P}\left[(X_{j\sigma(1)}, \ldots, _{j\sigma(n)}) \in A\right] \qquad j = 1, 2$$

for any permutation $\sigma \in \mathcal{P}(n)$, $n \geq 1$, and any measurable set $A \subseteq \mathbb{X}_j^n$. Marginal exchangeability of the two layers clearly does not necessarily imply *conditional exchangeability* of one layer given the other. To better understand this point, assume exchangeability for observations in the second layer conditionally on the first-layer partition, ρ_1. This implies, for instance, that the joint distribution of a pair of second-layer observations is invariant with respect to their clustering allocation at layer 1, in the sense that, for any set of three subjects i, j and k, $\mathbb{P}(X_{2i}, X_{2j} \mid c_{1i} = c_{1j}) = \mathbb{P}(X_{2i}, X_{2k} \mid c_{1i} = c_{1j})$, and, in particular,

$$\mathbb{P}(c_{2i} = c_{2j} \mid c_{1i} = c_{1j}, c_{1i} \neq c_{1k}) = \mathbb{P}(c_{2i} = c_{2j} \mid c_{1i} \neq c_{1j}, c_{1i} = c_{1k}) \tag{2}$$

where, on both sides, the probability refers to the event of i and j belonging to the same cluster at layer 2, but, on the left side, it is conditional on i and j belonging to the same cluster at layer 1 and on the right side on i and j belonging to two distinct clusters at layer 1. This odd behaviour of the learning mechanism is due to the fact that *conditional exchangeability* prevents subject-level information (such as which subjects belong to the same cluster) to be carried from one layer to the next, and allows only population-level information (such as knowledge about the number of clusters or the clusters' frequencies) to be transferred at the next layer. Any model based on this probabilistic assumption ultimately induces a learning mechanism that ignores that observations at different layers refer to the same individuals, similar to what happens in a regression model on longitudinal data without introducing subject-specific random effects. However, this assumption is also at the core of many dependent Bayesian clustering methods (see, for instance, MacEachern, 1999, 2000; Müller et al., 2004; Teh et al., 2006; Caron et al., 2007; Dunson and Park, 2008; Ren et al., 2008; Dunson, 2010; Rodríguez et al., 2010; Taddy, 2010; Rodriguez and Dunson, 2011; Lijoi et al., 2014; Foti and Williamson, 2015; Caron et al., 2017; Griffin and Leisen, 2017; DeYoreo and Kottas, 2018; De Iorio et al., 2019; 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).

An alternative to *conditional exchangeability* (of the second layer given ρ_1) is offered by the Enriched Dirichlet process (Wade et al., 2011), where given ρ_1, observations at second layers are assumed exchangeable if they belong to the same first-layer cluster and independent otherwise. However, while this strategy is more coherent with the multi-view data structure since it accounts for within-subject dependence, it implies $\mathbb{P}(X_{2i}, X_{2j} \mid c_{1i} \neq c_{1j}) = \mathbb{P}(X_{2i} \mid c_{1i} \neq c_{1j})\mathbb{P}(X_{2j} \mid c_{1i} \neq c_{1j})$ and, most importantly,

$$\mathbb{P}(c_{2i} = c_{2j} \mid c_{1i} \neq c_{1j}) = 0 \tag{3}$$

The condition defined by (3) is again a strong assumption, especially for clustering purposes. It forces second-layer clusters to be nested within first-layer clusters in the following sense: if two items are assigned to distinct clusters at layer 1, they cannot be assigned to the same cluster at layer 2, regardless of the data-generating mechanism. This property also weakens the borrowing of information within same-layer observations and forces the number of second-layer clusters to be at least equal to the number of first-layer clusters.

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 in (2) nor as too informative as in (3). Conditionally on ρ_1, it is natural to describe second-layer data as multi-sample/grouped data, since a clustering configuration of the same observational units is already provided by ρ_1 and such information is relevant. Ideally, an appropriate learning mechanism would a-priori favour at layer 2 a clustering configuration similar to layer 1, but not necessarily identical or nested. To this end, (2) should be replaced by

$$\mathbb{P}(c_{2i} = c_{2j} \mid c_{1i} = c_{1j}, c_{1i} \neq c_{1k}) \geq \mathbb{P}(c_{2i} = c_{2j} \mid c_{1i} \neq c_{1j}, c_{1i} = c_{1k}) \tag{4}$$

while still assigning a positive prior probability to any clustering structure of second-layer observations, and thus, (3) should be replaced by

$$\mathbb{P}(c_{2i} = c_{2j} \mid c_{1i} \neq c_{1j}) \geq 0 \tag{5}$$

To build such a general framework, we impose that second-layer observations are partially exchangeable (de Finetti, 1938) conditionally on ρ_1. We introduce the following definition of CPE, as a modelling principle for dependent partitions of the same items.

Definition 1 *Given a (marginally) exchangeable sequence $(X_{2i})_{i\geq 1}$ and a collection of coherent*[1] *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*

$$\mathbb{P}\left[(X_{21}, \ldots, X_{2n}) \in A \mid \rho_{1n}\right] = \mathbb{P}\left[(X_{2\sigma(1)}, \ldots, X_{2\sigma(n)}) \in A \mid \rho_{1n}\right] \qquad \text{for any } \sigma \in \mathcal{P}(n; \rho_{1n})$$

for any measurable set A, $n \geq 1$, where $\mathcal{P}(n; \rho_{1n})$ denotes the space of permutations of n elements that preserve ρ_{1n}, i.e. $\sigma \in \mathcal{P}(n; \rho_{1n})$ if and only if σ is a permutation of n elements such that $c_{1\sigma(i)} = c_{1i}$, for any $i \in [n]$.

For the sake of notation, in the following, we omit the subscript n when denoting the partition. The proof that CPE implies (4) is a direct consequence of the results in Franzolini (2022) and Franzolini et al. (2023). Note the CPE still allows us to obtain (2) and (3) as limiting cases within the class of sequences defined by Definition 1. Nonetheless, in order to account for within-subject dependence, we will always require that $(X_{2i})_{i\geq 1}$ is not *conditionally exchangeable* and, thus, that the inequality in (4) is strict. As a consequence, the learning mechanism accounts for within-subject dependence.

The power of CPE is not limited to the definition of an appropriate probability invariance structure for multi-view clustering. It is a *constructive* definition that, thanks to its conditional formulation, facilitates the development and study of many clustering processes while guaranteeing posterior computational tractability. In the following sections, we introduce *telescopic clustering models*, a general class of clustering models, based on CPE and Bayesian a.s. discrete random measures. The following propositions illustrate how existing Bayesian clustering approaches do satisfy or not the CPE assumption.

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

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

Proof. *See Appendix A.1*

Proposition 2 *If $(X_{1i},\dots,X_{Ji})_{i\geq 1}$ follows the separate exchangeable random partition model of Lin et al. (2021), then, for any j and j', conditionally on ρ'_j, $(X_{ji})_{i\geq 1}$ is conditionally partially exchangeable and not conditionally exchangeable.*

Proof. *See Appendix A.2*

Proposition 3 *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 (1999, 2000) and Quintana et al. (2022), then, conditionally on ρ_1, $(X_{2i})_{i\geq 1}$ is conditionally exchangeable.*

Proof. *See Appendix A.3*

3 Telescopic clustering

3.1 The class of telescopic clustering models

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

$$X_{1i}\mid \tilde p_1 \overset{iid}{\sim} \int_{\Theta_1} k_1(X_{1i},\theta)\,\tilde p_1(\mathrm{d}\theta) \qquad \text{for } i=1,\dots,n \tag{6}$$

where $k_1(\cdot,\cdot)$ is a kernel defined on $(\mathbb{X}_1,\Theta_1)$, $\tilde p_1$ is an almost-surely discrete random probability, i.e., $\tilde p_1 \overset{a.s.}{=} \sum_{m=1}^M w_m\delta_{\theta^\star_m}$, with $M\in\mathbb{N}\cup\{+\infty\}$ and $(w_m,\theta^\star_m)_{m=1}^M$ random variables such that $\sum_{m=1}^M w_m \overset{a.s.}{=} 1$. 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 $M=\infty$. As prior distribution for $\tilde p_1$, many proposals are available in the Bayesian nonparametric literature (see, among others, Ferguson, 1973; Pitman and Yor, 1997; Regazzini et al., 2003; Lijoi et al., 2005a,b; Gil-Leyva and Mena, 2021; Argiento and De Iorio, 2022). In Sections 3.5 and 3.6, we consider two specific priors for the first layer: the hierarchical Dirichlet process, obtained when the hierarchical construction of Teh et al. (2006) is employed to define the law of a single random probability measure, as in Camerlenghi et al. (2018), and finite mixtures with a random number of components, (Nobile, 1994; Miller and Harrison, 2018; Argiento and De Iorio, 2022). However, we note that our construction is general.

The model in (6) admits an equivalent representation in terms of latent parameters due to the almost-sure discreteness of the random measure $\tilde p_1$. Such property induces the clustering of the observations. Model (6) can be rewritten in terms of the allocation vector $\boldsymbol{c}_1=(c_{11},\dots,c_{1n})$, $i=1,\dots,n$, defined in Section 2.2:

$$X_{1i}\mid c_{1i}=m,\boldsymbol{\theta}^\star \overset{ind}{\sim} k_1(X_{1i};\theta^\star_m) \tag{7}$$

The ties in the vector of latent parameters $\boldsymbol{c}_1$ determine the partition ρ_1 of the observations into different clusters, such that if $c_{1i}=c_{1j}$ individuals i and j belong to the same cluster at layer 1.

In the following, we assume that the subject-specific allocation variables $\boldsymbol{c}_1$ and the cluster-specific parameters $\boldsymbol{\theta}^\star=(\theta^\star_m)_{m=1}^M$ are a-priori independent so that the corresponding mixing random probability $\tilde p_1$ belongs to the class of species sampling processes (Pitman, 1996). However, our construction can be extended to the case in which allocation variables and cluster parameters are a-priori dependent at the cost of more involved computations. The labels $\boldsymbol{c}_1 = (c_{11}, \dots, c_{1n})$ are affected by the label switching problem (see, for instance, Stephens, 2000; Mena and Walker, 2015; McLachlan et al., 2019). To overcome this issue, we re-label the elements of the vector $\boldsymbol{c}_1$ in order of appearance of the observations and obtain $\boldsymbol{c}_1^\star = (c_{11}^\star, \dots, c_{1n}^\star)$. This means that $c_{11}^\star = 1$, i.e. the first observation X_{11} always belongs to the *first* cluster. Then either $c_{12}^\star = c_{11}^\star = 1$, if the second observation X_{12} is clustered together with X_{11}, or $c_{12}^\star = 2$, otherwise. Similarly, for subsequent observations (see, for instance, Ghosal and Van der Vaart, 2017, ch. 14). For each partition ρ_1 of the n units into K_{1n} sets, there exist $\binom{n}{K_{1n}}$ $K_{1n}!$ vectors $\boldsymbol{c}_1$ that encode the same partition ρ_1, due to the label switching problem and the fact that elements in $\boldsymbol{c}_1$ assume values in $[n]$, while $K_{1n} \leq n$. Note that, thanks to row-exchangeability, we can focus on an arbitrary order of the observations without affecting the joint law of the sample and, thus, posterior inference on the clustering configuration.

To satisfy CPE, the second-layer conditional model is defined as

$$X_{2i} \mid c_{1i} = m, (\tilde{p}_{21}, \dots, \tilde{p}_{2M}) \overset{ind}{\sim} \int_{\Theta_2} k_2(X_{2i}, \theta)\, \tilde{p}_{2m}(\mathrm{d}\theta) \qquad \text{for } i = 1, \dots, n \tag{8}$$

where k_2 is a kernel defined on $(\mathbb{X}_2, \Theta_2)$, $M \in \mathbb{N} \cup \{+\infty\}$ 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 $\mathbb{N}$. Note that (8) guarantees CPE of $(X_{2i})_{i\geq 1}$ with respect to $(\rho_{1n})_{n\geq 1}$ as a consequence of de Finetti's representation theorem of partial exchangeability (de Finetti, 1938). Consider the vector $\left(\tilde{p}_{21}^\star, \dots, \tilde{p}_{2K_{1n}}^\star\right)$ obtained from reordering and then selecting the first K_{1n} entries of $(\tilde{p}_{21}, \dots, \tilde{p}_{2M})$, where the reordering is accordingly to the order of appearance of first-layer clusters and K_{1n} is the number of clusters at the first layer. We have

$$X_{2i} \mid c_{1i}^\star = m, \left(\tilde{p}_{21}^\star, \dots, \tilde{p}_{2K_{1n}}^\star\right) \overset{ind}{\sim} \int_{\Theta_2} k_2(X_{2i}, \theta)\, \tilde{p}_{2m}^\star(\mathrm{d}\theta) \qquad \text{for } i = 1, \dots, n \tag{9}$$

The advantage of representation (9) compared to (8) is that, for any $n < \infty$, then $K_{1n} < \infty$ a.s., a fundamental property for devising sampling schemes. Differently, the sampling mechanism in (8) does not require to re-order of the probability measures based on the order of appearance. However, while (8) implies (9), the vice-versa is in general not true unless we assume that the whole (possibly infinite) vector $(\tilde{p}_{21}, \dots, \tilde{p}_{2M})$ is exchangeable. If the application under study requires more layers, as detailed in Section 3.4, then we need to assume re-ordering for all layers except for the final one.

We now can define the class of telescopic clustering of models (with two layers).

Definition 2 *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:*

$$X_{1i} \mid \tilde{p}_1 \overset{iid}{\sim} \int_{\Theta_1} k_1(X_{1i}, \theta)\, \tilde{p}_1(\mathrm{d}\theta) \qquad \textit{for } i = 1, 2, \dots$$

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

with

$$\tilde{p}_1 \sim P_1 \qquad \textit{and} \qquad (\tilde{p}_{21}, \dots, \tilde{p}_{2M}) \sim P_2$$

where

- *k_1 and k_2 are kernels defined on $(\mathbb{X}_1, \Theta_1)$ and $(\mathbb{X}_2, \Theta_2)$, respectively;*
- *$\boldsymbol{c}_1 = (c_{11}, \dots, c_{1n})$ is any possible configuration of the unordered 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 core of the learning mechanism lies in the choice of P_2. In the following, we explore in detail two prior specifications for P_2: the well-known hierarchical Dirichlet process of Teh et al. (2006) and a novel multivariate mixture with a random number of components. In Section S1 of the Supplement, we prove that, under Definition 2 and for any finite n, the rows of the original data matrix $(X_{1i}, X_{2i})_{i=1}^{n}$ are (finite) exchangeable. Nonetheless, in the next sections, we obtain infinite exchangeability of the rows as a by-product of a joint representation theorem.

3.2 Row-exchangeability and joint representation theorem

We now consider the joint distribution of both layers $(X_{1i}, X_{2i})_{i\geq 1}$. The next theorem provides a hierarchical representation of the joint model, which turns out to be again a mixture model with an almost surely discrete mixing random measure.

Theorem 1 *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*

$$(X_{1i}, X_{2i}) \mid (\theta_i, \xi_i) \overset{ind}{\sim} k_1(X_{1i}, \theta_i) k_2(X_{2i}, \xi_i)$$
$$(\theta_i, \xi_i) \mid \tilde{p} \overset{iid}{\sim} \tilde{p}$$

where $\tilde{p} \overset{a.s.}{=} \sum_{m=1}^{M} \sum_{s=1}^{S} w_m q_{ms} \delta_{(\theta_m^\star, \xi_s^\star)}$.

Proof. *See Appendix A.4*

Thanks to the Bayesian hierarchical representation in Theorem 1, our framework guarantees row-exchangeability of $(X_{1i}, X_{2i})_{i\geq 1}$. Moreover, Theorem 1 implies *Kolmogorov's consistency* in n, sometimes also referred to as *marginal invariance* (Dahl et al., 2017) or *projectivity* (Betancourt et al., 2022). This is a well-known advantage of Bayesian mixture models for clustering, where the induced clustering structure can be leveraged for predicting future data points. Projectivity of new layers is also possible in this framework and is discussed later in Section 3.4. More importantly, note that if a global clustering structure for the rows of the data matrix is of interest, the telescopic model still provides appropriate inference, similar to what can be obtained by clustering techniques (to which we refer as *constant-clustering* models) that consider all the features jointly to obtain a unique clustering configuration of the subjects. 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 goal of telescopic clustering models is different and, when compared to *constant-clustering* models, they present many advantages: (1) provide also marginal, 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 4), (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.

Finally, the following theorem state how telescopic clustering models in general do not imply column or conditional exchangeability.

Theorem 2 *If $(X_{1i}, X_{2i})_{i\geq 1}$ follows a telescopic clustering model with two layers, as defined by Definition 2, then, for every fixed $n \geq 1$,* *(i)* $(X_{1i}, X_{2i})_{i=1}^n$ *is in general not column exchangeable, i.e., Definition 2 does not imply that* $(X_{1i}, X_{2i})_{i=1}^n \overset{d}{=} (X_{2i}, X_{1i})_{i=1}^n$*, where* $\overset{d}{=}$ *denotes equality in distribution.*

(ii) $(X_{1i}, X_{2i})_{i=1}^n$ *is in general not conditionally exchangeable, i.e., Definition 2 does not imply that* $(X_{1i}, X_{2i})_{i=1}^n \overset{d}{=} \left(X_{1\sigma(i)}, X_{2\sigma'(i)}\right)_{i=1}^n$*, for any permutations* σ *and* σ' *of* n *elements, with* $\sigma \neq \sigma'$*, where* $\overset{d}{=}$ *denotes equality in distribution.*

Proof. *See Appendix A.5*

3.3 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 between clustering configurations that capture specific properties of telescopic clustering models, while the remaining two are based on widely used measures: the expected Rand index and the expected Binder loss.

Recall that c_{1i} and c_{2i} are the allocation variables for subject i at layer 1 and 2, respectively. We recall that thanks to CPE and by construction, equation (4) holds and thus subjects clustered together are layer 1 are more likely to be clustered together at layer 2 compared to subjects that do not belong to the same first-layer cluster. 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 *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*.*

where $\tau \in [0,1]$ and $\tau = 1$ iff $\mathbb{P}[c_{2i} = c_{2j} \mid c_{1i} \neq c_{1j}] = 0$, while $\tau = 0$ iff $\mathbb{P}[c_{2i} = c_{2j} \mid c_{1i} = c_{1j}] = \mathbb{P}[c_{2i} = c_{2j} \mid c_{1i} \neq c_{1j}]$. 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 smaller support for the joint prior of the clustering configurations. Notice that the measure τ of telescopic dependence is an asymmetric measure, which is computed conditionally on the allocation at layer 1.

To introduce, additional measures of dependence, we denote with $\Pi(n)$ the space of partitions of n elements and with t-EPPF(ρ_1, ρ_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). The t-EPPF can have 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 between ρ_1 and ρ_2, defined as

$$ER = \binom{n}{2}^{-1} \int\limits_{\Pi(n)^2} [a(\rho_1, \rho_2) + b(\rho_1, \rho_2)] \, \mathrm{d}\, \text{t-EPPF}(\rho_1, \rho_2)$$

and the expected Binder loss between ρ_1 and ρ_2, defined as

$$EB = \int\limits_{\Pi(n)^2} [c(\rho_1, \rho_2) + d(\rho_1, \rho_2)] \, \mathrm{d}\, \text{t-EPPF}(\rho_1, \rho_2)$$

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.

The next proposition provides the value of τ, ER, and EB, in any telescopic clustering model as a function of the distribution of the number of clusters.

Proposition 4 *In a telescopic clustering model, the a priori measure of telescopic dependence is*

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

The a priori expected Rand index equals

$$ER = \mathbb{P}(K_{12} = K_{22})$$

The a priori expected Binder's loss is

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

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

Proof. *See Appendix A.6*

As noted by Hubert and 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 below). Thus, $ER^{\perp}$ is typically positive. In the same spirit as that of the adjusted Rand index (Hubert and Arabie, 1985), we define a *telescopic adjusted rand-index* that allows us to correct for the randomness of the partitions. Note that, even if the well-known adjusted Rand index of Hubert and Arabie (1985) is used as a general measure of dependence in different contexts, it is actually based on the assumption that the two partitions are generated by a generalized hypergeometric distribution, conditionally on having fixed the number of clusters in each partition and the number of subjects in each cluster. Thus, since this is not the case in telescopic clustering, we avoid using the adjusted Rand index of Hubert and Arabie (1985) for measuring prior and posterior dependence. Instead, we define an analogue index, which reflects more closely the properties of the telescopic clustering.

Definition 4 *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 trivially to prove that, in the case of independence, the a priori expected value of $TARI$ equals 0.

The four indexes introduced in this section are here considered a priori and in this sense can be used for prior elicitation and hyperparameter tuning. The same indexes a posteriori are analogously defined employing the posterior t-EPPF, and can be numerically approximated by post-processing the outputs from Markov chain Monte Carlo sampling algorithms (which are described in Section 3.7 and Sections S2 and S3 of the Supplement).

3.4 Extension to L layers using polytrees

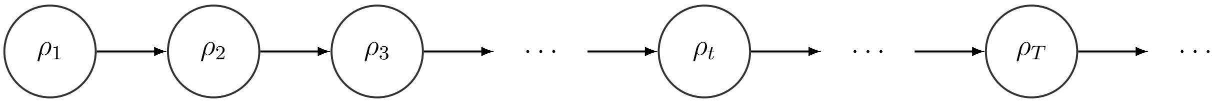


Figure 2: Graphical representation of layer dependence for longitudinal data.

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

$$\text{t-EPPF}(\rho_1, \rho_2) = \text{EPPF}(\rho_1)\, \text{p-EPPF}(\rho_2 \mid \rho_1)$$

where EPPF and p-EPPF are used to denote the marginal law of the partition ρ_1 and the conditional law of the partition ρ_2, respectively. We recall that ρ_1 is the random partition of an exchangeable sequence of observations and that ρ_2 is the random partition of a (conditionally) partially exchangeable sequence of observations. Such laws are typically referred to as *exchangeable partition functions* (EPPF) (Pitman, 1996) and partially exchangeable partition functions (p-EPPF) in the Bayesian nonparametric literature (cf., for instance, Lijoi et al., 2014).

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 $\rho_1 \to \rho_2$. A straightforward way to extend the modelling 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 straightforward extension can be obtained assuming a Markovian structure across different layers, obtaining a polytree. 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.,

$$\text{t-EPPF}(\rho_t,\, t = 1, 2, \ldots) = \text{EPPF}(\rho_1) \prod_{t=2}^{\infty} \text{p-EPPF}(\rho_t \mid \rho_{t-1}).$$

The resulting directed graph is displayed in Figure 2. Note that in this extension, the Markovian structure guarantees Kolmogorov consistency of the predictive distribution also to any new number of layers. This extension is explored in Section 4.2 on simulated data.

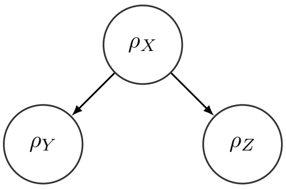


Figure 3: Graphical representation of multi-layer dependence for metabolomic study on BMI.

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

$$\text{t-EPPF}(\rho_X, \rho_Y, \rho_Z) = \text{EPPF}(\rho_X)\text{p-EPPF}(\rho_Y \mid \rho_X)\text{p-EPPF}(\rho_Z \mid \rho_X).$$

This modelling strategy is applied to data from the GUSTO cohort in Section 5.

Note that 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. The t-RPM model of Page et al. (2022) can be obtained with this strategy, but, for instance, the separate exchangeable models in Lin et al. (2021) with more than two layers cannot.

Finally, in the above discussion, we assume independence of the observations at different layers conditionally on the partitions, i.e., under the assumption in (1). Nonetheless, it is possible to induce further dependence, by defining a joint law for the cluster-specific unique values at different layers. For instance, for longitudinal data, it is straightforward to adopt a Markovian structure across layers through the specification of the base measures $P_0^{(\ell)} = \mathbb{E}[\tilde{p}_m^{(\ell)}]$, while for multivariate data with same support space, a hierarchical specification for the base measures can be adopted with little additional analytical and computational effort. In the next two sections, we introduce two special classes of telescopic models.

3.5 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 modelling 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) and Argiento et al. (2020). The primary purpose of hierarchical dependent processes is to enable density estimation and clustering for multi-sample data, allowing for information sharing and borrowing across different samples. The main idea behind the hierarchical dependent process is that the different processes are conditionally independent given a common base measure, which in itself is a random process. The randomness and the almost-sure discreteness of the common base measure induce dependence across both the weights and the atoms of the dependent processes. Moreover, the hierarchical construction offers computational feasibility, thanks, for instance, to marginal representations such as the Chinese franchise process.

In telescopic mixtures with hierarchical Dirichlet processes (t-HDP), 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

$$\begin{aligned} \tilde{p}_1 \mid \gamma, \tilde{p}_0 &\sim DP(\gamma, \tilde{p}_0) \\ \tilde{p}_0 \mid \gamma_0 &\sim DP(\gamma_0, P_\theta) \end{aligned} \tag{10}$$

while the second-layer conditional law is

$$\begin{aligned} X_{2i} \mid \boldsymbol{c}_1, (\tilde{p}_{21}, \tilde{p}_{22} \ldots,) &\overset{ind}{\sim} \int f(X_{2i}, \theta) \tilde{p}_{2c_i}(\mathrm{d}\theta) && \text{for } i = 1, \ldots, n \\ \tilde{p}_{2m} \mid \alpha, \tilde{q}_0 &\overset{iid}{\sim} DP(\alpha, \tilde{q}_0) && \text{for } m = 1, 2, \ldots \\ \tilde{q}_0 \mid \alpha_0 &\sim DP(\alpha_0, P_\xi) \end{aligned} \tag{11}$$

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 $n_1, \ldots, n_{K_{1n}}$ for the first-layer partition. Then, we have (see, Camerlenghi et al., 2018)

$$\mathbb{P}[\rho_1 = \rho] = \frac{\gamma_0^{K_{1n}}}{(\gamma)^{(n)}} \sum_{\boldsymbol{\ell}} \frac{\gamma^{|\boldsymbol{\ell}|}}{(\gamma_0)^{(|\boldsymbol{\ell}|)}} \prod_{m=1}^{K_{1n}} (\ell_m - 1)! |s(n_m, \ell_m)| \tag{12}$$

where $|s(n, k)|$ denotes the signless Stirling number of the first kind and the sum in (12) runs over all vectors $l = (l_1, \ldots, l_{K_{1n}})$ such that $l_m \in [n_m]$ and $(\gamma)^{(n)} = \Gamma(\gamma + n)/\Gamma(\gamma)$, where $\Gamma(x)$ denote the Gamma function in x. The p-EPPF describing the conditional law of the partition at layer 2, given ρ_1, is

$$\mathbb{P}[\rho_2 = \rho \mid \rho_1] = \frac{\alpha_0^{K_{2n}}}{\prod_{m=1}^{K_{1n}} (\alpha)^{(n_m)}} \sum_{\boldsymbol{t}} \frac{\alpha^{|\boldsymbol{t}|}}{(\alpha_0)^{(|\boldsymbol{t}|)}} \prod_{s=1}^{K_{2n}} (t_{\cdot s} - 1)! \prod_{m=1}^{K_{1n}} |s(n_{ms}, t_{ms})| \tag{13}$$

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 *Given a telescopic mixture model with hierarchical Dirichlet processes and two layers, the t-EPPF*(ρ_1, ρ_2) *is given by*

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

Proof. *Trivial by combining* (12) *and* (13).

Starting from the expression of the t-EPPF, it is straightforward to compute the indexes of dependence introduced in Section 3.3.

Corollary 1 *In a t-HDP, the measure* τ *of telescopic dependence is*

$$\tau = \frac{\alpha_0}{\alpha_0 + \alpha + 1}$$

which tends to 0 *as* α *tends to* ∞ *and to* 1 *as* α_0 *tends to* ∞.

The expected Rand index ER is

$$ER = \frac{(1 + \gamma_0 + \gamma)(1 + \alpha_0 + \alpha) + \gamma_0\, \alpha_0\, \gamma\, \alpha}{(\gamma_0 + 1)(\gamma + 1)(\alpha_0 + 1)(\alpha + 1)}$$

Proof. *Trivial by combining Proposition 4 and Theorem 3.*

3.6 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 unknown number M of sub-populations is finite but random. The second telescopic model introduced here lies within this framework. The prior for the first-layer random probability $\tilde{p}_1$ is defined by

$$\begin{aligned} \tilde{p}_1 &= \sum_{m=1}^{M} w_m \delta_{\theta_m^\star} \\ \boldsymbol{w} = (w_1, \ldots, w_M) \mid M &\sim P_w \\ \theta_m^\star \mid M &\overset{iid}{\sim} P_\theta \qquad \text{for } m = 1, \ldots, M \\ M &\sim P_M \end{aligned} \tag{14}$$

where $\boldsymbol{w}$ and $\boldsymbol{\theta}^\star = (\theta_1^\star, \ldots, \theta_M^\star)$ 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 and Harrison, 2018; Argiento and 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.

$$\boldsymbol{w} = (w_1, \ldots, w_M) \mid M, \gamma \sim \text{Dirichlet}_M (\gamma, \ldots, \gamma) \tag{15}$$

However, some alternatives that may be used to have an a.s. finite number of components are the Pitman-Yor multinomial process (Lijoi et al., 2020), normalized infinitely divisible multinomial processes (Lijoi et al., 2023), and the large class of normalized independent finite point processes (Argiento and De Iorio, 2022). Then, the conditional law of the second layer is defined employing a novel construction for the mixing random probability measures in (8). The vector of dependent random probabilities $(\tilde{p}_{21}, \ldots, \tilde{p}_{2K_{1n}})$, has probability ω, that all coordinates of the vector equal to the same almost-surely discrete probability $\tilde{p}_0$, while, with probability $1-\omega$, each coordinate of the vector $\tilde{p}_{2m}$ equals a Dirac measure in $\xi_m^\star$, with $\xi_m^\star \neq \xi_{m'}^\star$ a.s., for $m, m' = 1, \ldots, K_{1n}$. The formal construction is detailed in the following definition.

Definition 5 *A vector of random probability $(\tilde{p}_1, \ldots, \tilde{p}_K)$ is a unique-atom process if they admit the following almost-sure discrete representation:*

$$\begin{aligned} \tilde{p}_m &\overset{a.s.}{=} (1-Z)\,\delta_{\xi^\star_m} + Z\,\tilde{p}_0 \qquad \text{for} \quad m = 1, \ldots, K \\ Z &\sim \textit{Bernoulli}(\omega) \end{aligned}$$

where

1. *$\tilde{p}_0$ is an almost-surely discrete random probability,*
2. *$\xi^\star_m \overset{iid}{\sim} P_\xi$, for $m = 1, \ldots, K$,*
3. *$\tilde{p}_0$, $(\xi^\star_m)_{m=1}^K$, 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.,

$$\tilde{p}_0 \overset{a.s.}{=} \sum_{s=1}^{S} q_s \delta_{\xi^\star_{0s}}$$

with $S \sim P_S$, weights q_s distributed accordingly to a symmetric Dirichlet distribution and $\xi^\star_{0s} \overset{iid}{\sim} P_\xi$.

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. Therefore, the resulting telescopic model is a bivariate mixture of a *constant-clustering* model and an *independent-clustering* model. Employing unique-atom processes to build up the CPE needed for telescopic clustering, we get the following second-layer specification:

$$\begin{aligned} X_{2i} \mid \boldsymbol{c}_1, \boldsymbol{q}, \boldsymbol{\xi}, S, Z &\overset{ind}{\sim} (1-Z)\delta_{\xi^\star_{c_{1i}}} + Z \sum_{s=1}^{S} q_s k_2(X_{2i}; \xi^\star_{0s}) \qquad \text{for } i = 1, \ldots, n \\ \boldsymbol{q} = (q_1, \ldots, q_S) \mid S, \alpha &\sim \text{Dirichlet}_S(\alpha, \ldots, \alpha) \\ \xi^\star_{0s} \mid S &\overset{iid}{\sim} P_\xi \qquad \text{for } s = 1, \ldots, S \\ \xi^\star_m \mid K_{1n} &\overset{iid}{\sim} P_\xi \qquad \text{for } m = 1, \ldots, K_{1n} \\ S &\sim P_S \\ Z &\sim \text{Bernoulli}(\omega) \end{aligned} \tag{16}$$

As it is well known, the random probability $\tilde{p}_1$ in (14) is a species sampling process (Pitman, 1996) and, when used as mixing distribution as in (6), induces, a latent random partition of the observations. The marginal EPPF of the partition at layer 1 is a well-known result (see, e.g., Green and Richardson, 2001; McCullagh and Yang, 2008; Miller and Harrison, 2018; Argiento and De Iorio, 2022). Considering a specific partition ρ_1 into K_{1n} sets of the n observations, under (15) we have that

$$\mathbb{P}(\rho_1) = V(n, K_{1n}) \prod_{m=1}^{K_{1n}} \frac{\Gamma(\gamma + n_m)}{\Gamma(\gamma)} \tag{17}$$

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

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

where $x^{(k)} = \Gamma(x+k)/\Gamma(x) = x(x+1)\ldots(x+k-1)$ and $x_{(k)} = \Gamma(x+1)/\Gamma(x-k+1) = x(x-1)\ldots(x-k+1)$, where $\Gamma(x)$ denote the Gamma function in x and $x_{(0)} = 1$ and $x_{(0)} = 1$ by convention.

The law of second-layer partition ρ_2 conditionally on ρ_1 and the joint law of the two partitions are derived in the next theorem.

Theorem 4 *Given a telescopic mixture with unique atom processes, the probability of the second layer partition ρ_2 conditionally on the first layer partition ρ_1 is*

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

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.

The joint law of the two partitions is

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

Proof. *Trivial by combining (16) and (17).*

From the t-EPPF, it is possible to compute the indexes of dependence introduced in Section 3.3.

Corollary 2 *In a telescopic mixture with unique atom processes, the measure τ of telescopic dependence is*

$$\tau = \frac{1-\omega}{1 + \omega(\mathbb{E}[S]/\alpha - 1)}$$

which tends to 1 as ω tends to 0 and to 0 as ω tends to 1.

The expected Rand index ER is

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

Proof. *Trivial by Theorem 4.*

3.7 Algorithms for posterior inference

The posterior inference is performed through Markov chain Monte Carlo (MCMC) algorithms. In Sections S2 and S3 of the Supplement, we describe both conditional and marginal sampling schemes. The conditional algorithms make use of representation theorems and also provide posterior samples of the underlying random probability measures. The marginal algorithms are derived from the predictive distribution of the observations, obtained through marginalization of the random probability. In classical Bayesian clustering models, marginal sampling schemes tend to exhibit better mixing compared to conditional ones, but they come with a higher computational cost per iteration. This cost can increase significantly with the number of observations, and, in particular, with the number of clusters in the partitions visited by the MCMC chain.

Moreover, in the case of telescopic clustering models, marginal algorithms require evaluating the conditional pEPPF at the child nodes when sampling the cluster allocation at any given layer. The evaluation of the conditional pEPPF is typically computationally intensive. There are certain cases where the computational cost can be reduced by introducing latent random variables, but this is not always applicable. For example, in t-HDP models, the standard data augmentation provided by the Chinese franchise restaurant process simplifies the conditional pEPPF, but significantly slows down the mixing to unfeasible levels, as thoroughly detailed in Sections S2 and S3 of the

Supplement. On the other hand, the conditional sampling scheme for the t-HDP model shows a good mixing as well as a much lower computational time per iteration, making posterior inference feasible and more accurate. 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.

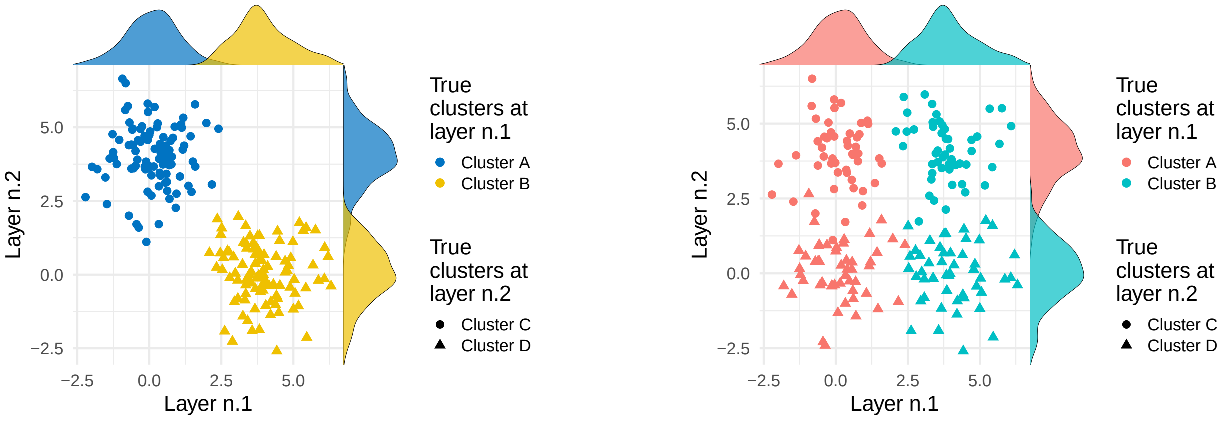
A detailed derivation of the sampling schemes for the general class of telescopic models and their tailored versions employed for inference in the remaining sections are in Sections S2 and S3 of the Supplement. Computational cost and mixing performance results are in Section S6 of the Supplement.

4 Numerical simulations

In this section, we present a simulation study to highlight the main learning properties of telescopic clustering. We consider four scenarios with different numbers of layers, varying from 2 to 100. The first three scenarios present well-separated clusters resulting in strong layer-marginal signals from the data concerning the partition. This is in order to highlight the learning mechanism induced by CPE. We consider more complex simulation setups, with also misspecification and up to 100 layers in Section 4.2 and Section S4 of the Supplement.

4.1 Simulation studies with two layers

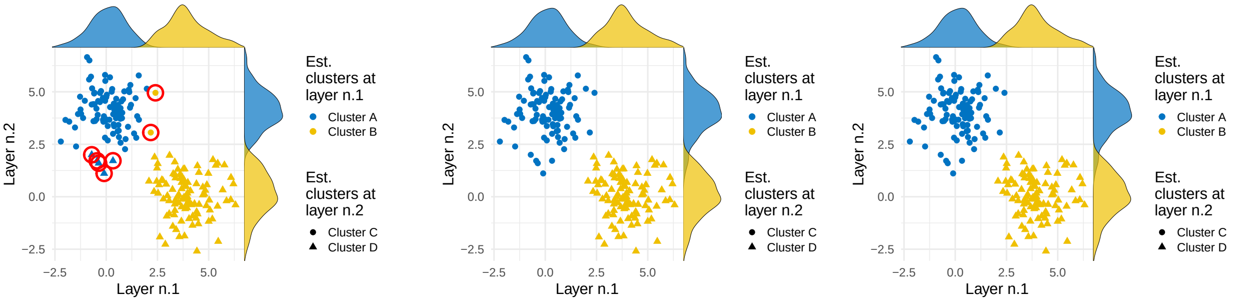
In the first two simulation scenarios, data for $n = 200$ observational units and $L = 2$ layers are generated. The first scenario (Scenario 1) is obtained keeping the clustering structure constant across the two layers. In particular, the first cluster is composed by 100 observations such that $(X_{1i}, X_{2i}) \overset{iid}{\sim} \mathcal{N}(0,1) \times \mathcal{N}(4,1)$, while the remaining 100 observations form a second cluster and are sampled according to $(X_{1i}, X_{2i}) \overset{iid}{\sim} \mathcal{N}(4,1) \times \mathcal{N}(0,1)$. Figure 4a shows the simulated data and highlights how the two clusters are well-separated both at layer 1 (on the x-axis) and at layer 2 (on the y-axis). The second scenario (Scenario 2) is obtained by imposing two highly different clustering



(a) Scenario 1: RI = 1, ARI = 1

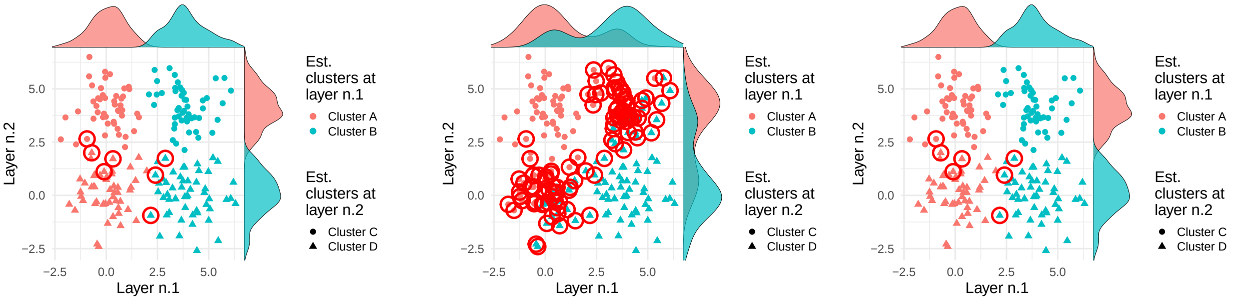
(b) Scenario 2: RI = 0.55, ARI = 0.01

Figure 4: Simulation study: simulated data and *true* cluster allocation for Scenarios 1 and 2. Each point corresponds to an item, colours denote clusters at layer 1 and shapes are clusters at layer 2. Under scenario 1, the clustering structure is the same at both layers, and the adjusted Rand index between the two partitions is equal to 1. Under scenario 2, the clustering structure drastically changes between the two layers and the adjusted Rand index (ARI) between the two partitions is approximately 0.



(a) Independent clustering (b) Constant clustering (c) Telescopic HDP

Figure 5: Simulation study: results for Scenario 1. Red circles denote observations assigned to the wrong cluster at least for one layer.



(a) Independent clustering (b) Constant clustering (c) Telescopic HDP

Figure 6: Simulation study: results for Scenario 2. Red circles denote observations assigned to the wrong cluster at least for one layer.

structures at the two layers while keeping the number of clusters and the clusters' frequencies fixed across layers. This is achieved by reassigning half of the observations in each cluster to the other cluster while moving from one layer to the next. Denoting with c_{1i} and c_{2i} the allocation variables at layer 1 and 2 respectively, the data generating process is: $(X_{1i}, X_{2i}) \mid c_{1i} = m, c_{2i} = s \overset{iid}{\sim} \mathcal{N}(\theta_m, 1) \times \mathcal{N}(\xi_s, 1)$, where the locations at layer 1 are $\boldsymbol{\theta} = (\theta_1, \theta_2) = (0, 4)$ and the locations at layer 2 are $\boldsymbol{\xi} = (\xi_1, \xi_2) = (4, 0)$. In this scenario, the true clustering structure coincides with the expected value of a random assignment procedure, where, moving from layer 1 to layer 2, each observation is reassigned to the other cluster with probability 1/2. Figure 4b shows the simulated data. Note that the layer-marginal distributions are the same in both simulation studies, what truly differentiates the two scenarios is how single items are reallocated moving from layer 1 to layer 2.

For both simulated datasets, as baseline comparisons, we estimate the clustering configuration independently at each layer as well as a constant clustering model, which assumes the same configuration at both layers. We compare such approaches with the results from the t-HDP model, presented in Section 3.5, based on CPE. The first two models are mixtures of the hierarchical Dirichlet process as described in Camerlenghi et al. (2018). What differentiates the three approaches is the type of dependence between layer-specific partitions, from independence to complete dependence. More details on the model specification and the hyperparameter setting are presented in Section S4 of the Supplement.

Figure 5 and 6 show the point estimates of the clustering allocations obtained minimizing the variation of information loss (Meilă, 2007). Obviously, the constant clustering approach performs extremely well under scenario 1, since the prior distribution is degenerate on the truth of a unique clustering configuration (cf. Figure 5b). The same model performs badly in the second simulation scenario since the true clustering configuration does not belong to the support of the prior (cf.

Figure 6b). On the other hand, the independent model performs worse than the constant model in simulation scenario 1, since it does not allow for any borrowing of information and modelling of within-subject dependence (cf. Figure 5a). In simulation scenario 2, the independent model has an advantage with respect to both the constant clustering and CPE, because under the truth there is no within-subject dependence and borrowing of information between clustering configurations is undesirable. Nonetheless, in this second scenario, the independent approach lead to seven allocation errors (four at layer 1 and three at layer 2), which can be explained by the fact that they correspond to observations that are more likely to be generated under the other mixture component (cf. Figure 6a).

The results of the telescopic clustering model coincide with the best performance in both scenarios. In fact, the model achieves the same results as the constant model when the clustering configuration is indeed constant (Scenario 1) and the same results as the independent model when the clustering configurations are the expectation of a random assignment (Scenario 2). The telescopic clustering appears able to detect the dependence structure between layers and accurately recover the clustering configuration of the observations.

4.2 Simulation study with ten and one hundred layers

In the third simulation scenario (Scenario 3), 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 centred in 0 and 4 respectively (analogously to the simulation studies in the previous section). However, from one layer to the next, 10 items (5% of the total) are selected at random and moved to the other cluster, so that the adjusted Rand index from one layer t to the next $t + 1$ equals 0.809. Simulated data are shown in Figure 7.

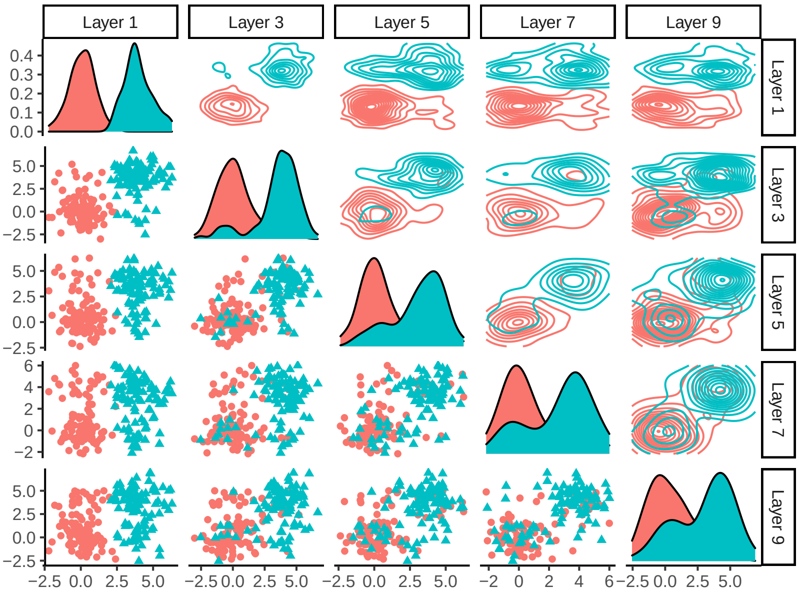


Figure 7: Simulation study: simulated data for Scenario 3. Plots refer to the observations in layers 1, 3, 5, 7, and 9. Colours and shapes denote the true clustering at layer 1. The diagonal plots show the marginal distribution at each layer, colour coded according to the clustering allocation at layer 1. Upper off-diagonal plots display the joint distribution of two pairs of layers, colour coded according to the clustering allocation at layer 1. Lower off-diagonal plots show the scatter plot of the data at the corresponding layers, colour coded according to the clustering allocation at layer 1. The adjusted Rand index between two consecutive configurations is 0.809.

	Rand Index				# Mistakes			
Layer	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: Simulation study: results for Scenario 3, Rand indexes between the estimated and true clustering configurations and numbers of items allocated to the wrong cluster. The two measures are reported for each layer and the last row contains the averages across layers.

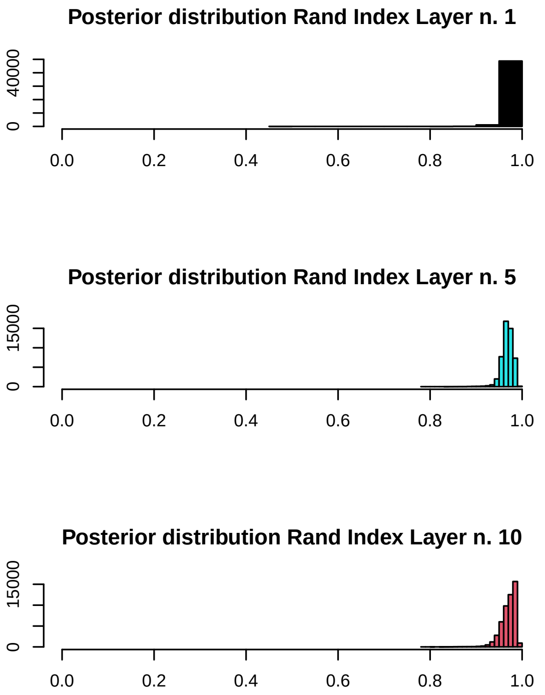


Figure 8: Simulation study: results for Scenario 3. Posterior distributions of Rand indexes between the posterior configurations and the truth for t-HDP model for layers 1, 5, 10.

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 of the random probabilities in the mixture model (for more details and algorithms, see, Rigon and Durante, 2021), leading to the inclusion of a linear trend that induces dependence at the level of the random probability measures. 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 estimate for the clustering configuration the one that minimises the variation of information loss (Meilă, 2007).

Table 1 summarises the results. The t-HDP model identifies the true clustering configuration at all layers with at most three out of 200 wrongly allocated subjects and a rand index between the truth and the estimate always higher than 0.97, the average rand index equals 0.99 and the average number of wrongly allocated subject per layer is 1.1 out 200. It outperforms the competitors both consistently at each layer and overall. Independent k-means and the LSBP perform well, even if they do not include within-subject dependence. This is to be expected in this scenario since the true clusters are well-separated (cf., with results of Scenario 4 below, where the k-means solution is often unable to identify the true number of clusters, even though the cluster have still Gaussian shapes). Nevertheless, both the k-means solution and the LSBP estimates are always dominated by the t-HDP estimates. Finally, the enriched Dirichlet process is the worst-performing model, as a direct consequence of the degeneracy issue of the model discussed in Section 2. Recall that under the enriched Dirichlet process, once two items are assigned to two different clusters at layer t, they cannot be assigned to the same cluster at any subsequent layer s, for $s > t$.

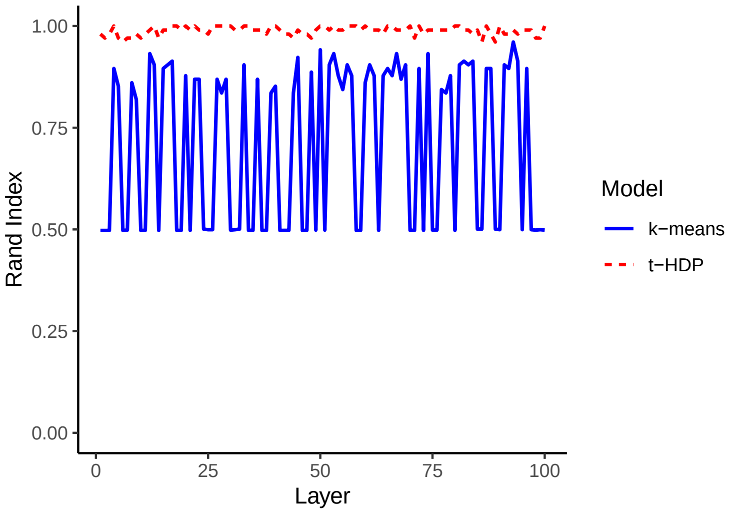


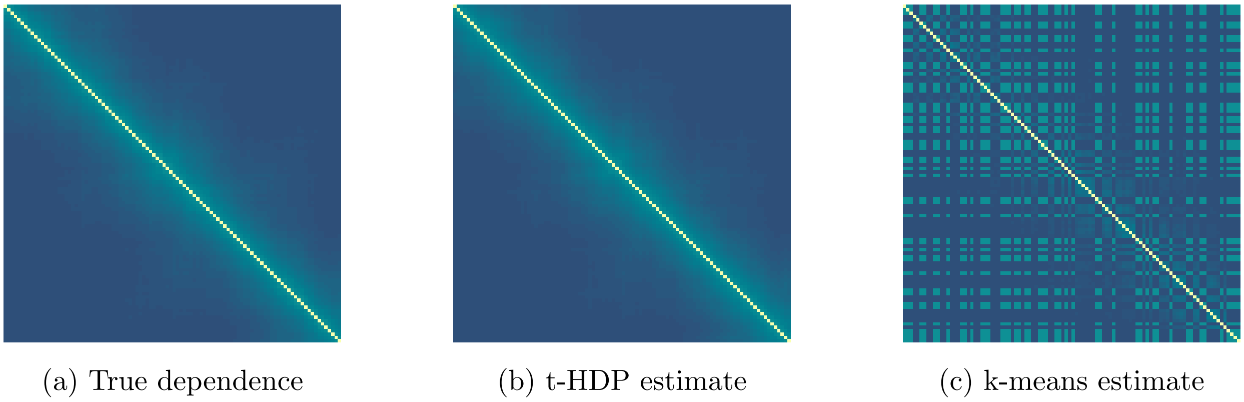
Figure 9: Simulation study: results for Scenario 4. Rand indexes between the truth and the estimated configuration at each layer estimates are obtained with the t-HDP model (dashed line) and independent k-means (solid line).

Figure 8 shows the posterior distribution of the Rand index between the true clustering configuration and the configurations visited by the posterior algorithm of the t-HDP model for layers 1, 5, and 10 after burn-in. The posterior is concentrated around 1, which corresponds to the truth, exhibiting small uncertainty around the estimated clustering configuration.

Finally, in Scenario 4, 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

$$X_{1i} \mid c_{1i} \overset{ind}{\sim} \mathcal{N}(0,1)\mathbb{1}(c_{1i} = 1) \\ + \mathcal{N}(3,1)\mathbb{1}(c_{1i} = 2)$$

Then, from layer ℓ to layer $\ell+1$, 2% of the observations are selected at random and moved to the other cluster. Figures 9 and 10 summarises the results of the t-HDP model and independent k-means clustering. Posterior estimates of the clustering configuration for the t-HDP model are obtained by minimising the variation of information loss (Meilă, 2007), while for k-means we employ the gap statistics (Tibshirani et al., 2001).


(a) True dependence

(b) t-HDP estimate

(c) k-means estimate

Figure 10: Simulation study: results for Scenario 4. Pairwise Rand indexes between any couple of layers for (a) the true clustering configurations; (b) the clustering configurations estimated by the t-HDP model; (c) the clustering configurations obtained with the k-means method.

5 An application to childhood obesity

In this section, we investigate childhood obesity patterns and their relationship with metabolic pathways, as well as traditional clinical markers for mothers. The prevalence of obesity in children and adolescents has escalated in recent years, reaching pandemic proportions worldwide.

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). Measuring children's growth trajectories is less trivial than in adults mainly because children are growing individuals and changes in the body mass index (BMI) over time are expected, even within a healthy state. A common measure for children's growth is to calculate the BMI, as for adults, and then compare it to the median values estimated for the same age class. The resulting measure is called the z-BMI score and is used in our analysis. The detailed procedure to compute the z-BMI can be found in WHO (2007). 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. The ogtt is a diagnostic procedure used to assess an individual's ability to regulate blood sugar levels. This test is aimed to evaluate how effectively the body processes respond to glucose, providing valuable information about insulin sensitivity, glucose metabolism, and the presence of conditions such as diabetes or impaired glucose tolerance. Finally, we introduce a third parallel layer at the same level as the mothers' information leading to the polytree structure in Figure 11. In this third layer, we include concentration data of 35 metabolites measured in the children using NMR spectroscopy. Metabolites are small molecular weight molecules that play a crucial role in the biochemical reactions occurring in an organism and are associated to numerous health conditions. Metabolomics provides an effective approach for detecting underlying biological mechanisms, uncovering genetic and environmental interactions, identifying therapeutic targets, and monitoring disease progression (see, for instance, Ellul et al., 2019). 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 analysed by Cremaschi et al. (2021) with the goal of identifying metabolic pathways related to childhood obesity. Their main inferential objective is the identification of a common clustering configuration for all the variates (growth curves and metabolic pathways), while here we focus on the dependence of the clustering structures across distinct sources of information.

We fit the t-HDP model presented in Section 3.5 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

Mother	**Metabolites**	**Growth trajectory**					**Total**
		Underweight	*Normal low*	*Normal*	*Normal high*	*Obesity*	
Low	*Conf. 1*	45	107	112	68	26	64.74%
	Conf. 2	1	11	5	3	12	5.79%
High	*Conf. 1*	8	20	39	42	21	23.51%
	Conf. 2	0	2	1	7	17	4.88%
Outliers	*Conf. 1*	0	2	0	2	1	0.90%
	Conf. 2	0	0	0	0	1	0.18%
Total		9.76%	25.68%	28.39%	22.06%	14.10%	

Table 2: Three-way cross-table of the estimated clustering configurations. Values within the table are absolute frequencies; the last row indicates the percentages of children in different growth trajectory clusters; the last column contains percentages of children assigned to different combinations of mother and metabolites clusters.

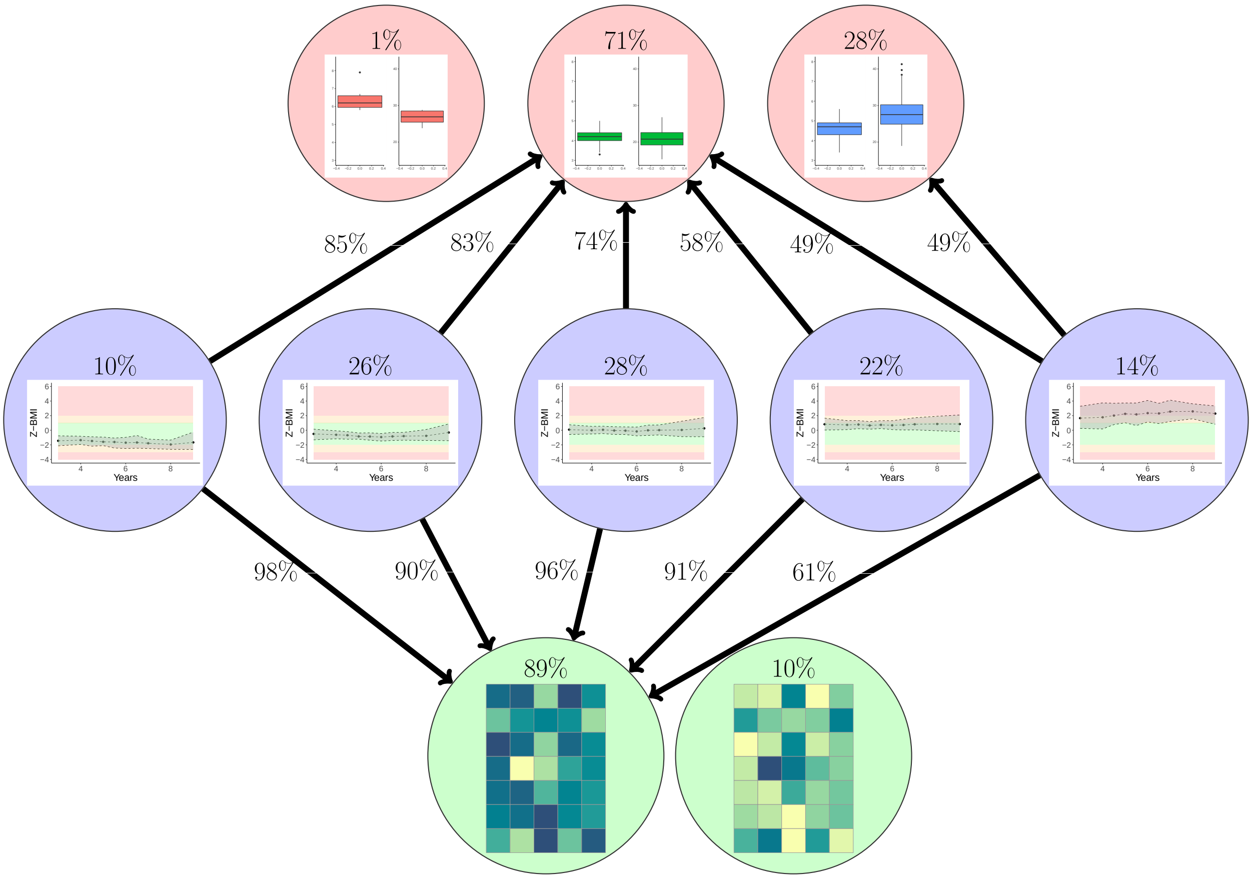


Figure 11: Estimated clustering configuration for the GUSTO cohort data. Nodes in the graph represent different clusters: the three nodes at the top refer to mothers' information, the five nodes in the centre correspond to growth trajectories, and the two nodes at the bottom are the two metabolites clusters. 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. On each edge, we report the conditional percentage of children assigned to the mothers-cluster or metabolites-cluster, conditional to belonging to a given growth-trajectory cluster.

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 conditional algorithm described in Section S3.1 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 summarised in Table 2 and shown in Figure 11. A detailed account of the results is provided in Section S5 of the Supplement. The estimates are obtained by minimizing the Binder loss function with equal costs for layers 2 and 3 and the variation of information loss (Meilă, 2007) for layer 1. The choice of the loss function used in each layer is driven by the interpretability of the results. This is due to the fact that the Binder loss function often tends to identify highly unbalanced and difficult-to-interpret clusters. On the other hand, the variation of information loss implicitly applies a stronger penalization to unbalanced clusters, leading to more balanced clusters' frequencies but often concentrating on too few clusters for interpretability purposes. These are well-known features of these loss functions, see for instance Dahl et al. (2022).

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. 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 characterised 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.

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. These findings highlight the heterogeneity in metabolite levels among the studied population. 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. 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 S5.1 in the Supplement. Understanding these relationships can shed light on the underlying childhood mechanisms involved in childhood obesity and potentially guide the development of targeted interventions aimed at addressing metabolic dysregulation and promoting healthier weight outcomes.

6 Conclusions and future directions

Standard clustering techniques often fall short when applied to datasets containing multi-view or longitudinal data, where the characteristics of clusters can vary across features, time or space. This setting requires flexible modelling of within-subject dependence across multiple features, even if they have different support spaces. Classical model-based clustering techniques are unable to address this issue effectively. To overcome this challenge, we introduce a novel class of Bayesian clustering models: the telescopic clustering models. The key idea behind this approach is the concept of conditional partial exchangeability, a probabilistic paradigm that effectively allows for multi-view clustering taking into account within-subject dependence and encompasses as special cases well-known construction in the Bayesian nonparametric literature. Our approach opens a promising direction for the development and exploration of dependent random partitions of the same subjects.

Furthermore, we demonstrate the proposed strategy on both simulated data and a real dataset analysis. The simulation study highlights that our approach consistently outperforms alternative methods, in terms of recovering the true cluster configuration across features. Additionally, we apply our methodology to investigate the relationships between childhood obesity and metabolite concentrations, while incorporating information about their mothers. This analysis not only demonstrates the effectiveness of our framework in handling complex, multi-dimensional data and performing data integration but also leads to interesting results in terms of biological mechanisms.

There are several directions for future work. Firstly, we have shown that constructions developed under CPE allow us to account for within-subject dependence, in the sense of (4), when they do not degenerate on *conditional exchangebility.* Nonetheless, from a probabilistic and statistical standpoint, it is interesting to determine whether there exist other invariance conditions that achieve the same inferential goal as CPE. Secondly, it is important to determine the extent of the class of telescopic clustering models within the CPE framework. This might lead to different constructions satisfying CPE. Moreover, while the telescopic class is built assuming a directed relationship $\rho_1 \to \rho_2$, it is worth investigating which symmetric relations between clustering configurations are included within this framework. Understanding the full scope and flexibility of this class would provide valuable insights into the range of applications and potential extensions of telescopic clustering. Furthermore, it is essential to explore other potential extensions of telescopic clustering models to more than two layers beside the polytrees. Lastly, the model described in Section 3.6 paves the way for further research on change-point detection in dynamic temporal clustering.

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

Appendix

A.1 Proof of Proposition 1

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

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

- *$\sigma(\rho)$ be the partition obtained swapping the elements in the sets of ρ accordingly to σ,*
- *$\mathcal{R}(\boldsymbol{\gamma}) = \{i : \gamma_i = 1\}$ and $\sigma(\boldsymbol{\gamma}) = (\gamma_{\sigma(1)}, \dots, \gamma_{\sigma(n)})$*
- *$\rho^{\mathcal{R}(\boldsymbol{\gamma})}$ be the "reduced partition" obtained removing from the sets in ρ all elements that are not in $\mathcal{R}(\boldsymbol{\gamma})$ 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* $\boldsymbol{\gamma} \in \{0,1\}^n$ *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, cfr. Definition 1 in Section 2.2.

***Proof.* (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 $\mathcal{P}(n;\rho)$.*

***Proof.* (Proposition 1)** *Denoting with X_{ti} a response measured on the ith 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*

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

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

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

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 $\boldsymbol{c}_{t-1}$, to prove CPE, we need to prove that

$$\mathbb{P}[(X_{t1}, \ldots, X_{tn}) \mid \rho_{t-1}] = \mathbb{P}[(X_{t\sigma(1)}, \ldots, X_{t\sigma(n)}) \mid \rho_{t-1}]$$

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 Definition 1 in Section 2.2.

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

$$\mathbb{P}[(X_{t1}, \ldots, X_{tn}) \mid \rho_{t-1}] = \sum_{\lambda} \left(\mathbb{P}[(X_{t1}, \ldots, X_{tn}) \mid \rho_t = \lambda] \;\; \mathbb{P}[\rho_t = \lambda \mid \rho_{t-1}]\right)$$

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:

$$\mathbb{P}[(X_{t1}, \ldots, X_{tn}) \mid \rho_t = \lambda] = \mathbb{P}[(X_{t\sigma(1)}, \ldots, X_{t\sigma(n)}) \mid \rho_t = \sigma(\lambda)]$$

for any permutation σ of n elements. For what concern 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 $\gamma_t = (\gamma_{1t}, \ldots, \gamma_{nt})$. The latent variables identify which subjects at time $t-1$ will be considered for possible cluster re-allocation at time t. Specifically, let γ_{it} be defined as

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

so that

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

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

Thus, denoting with $\sigma(\gamma_t)$ the vector $(\gamma_{\sigma(1)t}, \ldots, \gamma_{\sigma(n)t})$, for any permutation σ of n elements, to prove that $(X_{t\,i})_{i\geq 1}$ 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

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

and, thus

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

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

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

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

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

which prove that t-RPM mixtures are conditional 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 $\sigma = (1, 3)$.

A.2 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,*

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

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, cfr. Definition 1 in Section 2.2.

***Proof.* (Proposition 2)** *If $(X_{1i}, \ldots, X_{Ji})_{i\geq 1}$ follows the separate exchangeable random partition mixture of Lin et al. (2021), then*

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

where GEM(α) 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 jth 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*

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

and, similarly,

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

Thus

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

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 2.2.*

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* $(X_{j1}, X_{j2}, X_{j3}) \in (d(\bar{\theta}^\star_\ell - \epsilon), d(\bar{\theta}^\star_\ell + \epsilon), d(\bar{\theta}^\star_\ell + 2\epsilon))$, *where* $\bar{\theta}^\star_\ell = \mathbb{E}[\theta^\star_\ell]$, $\epsilon > 0$ *and* $dy = [y, y+\nu)$, *with* ν *arbitrarily small.*

A.3 Proof of Proposition 3

***Proof.* (Proposition 3)** *Denoting with* X_{ix} *the response measured on the ith 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*

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

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*

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

A.4 Proof of Theorem 1

***Proof.* (Theorem 1)** *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*

$$X_{2i} \mid \tilde{q}_i \overset{ind}{\sim} \int k_2(X_{2i};\xi)\tilde{q}_i(\mathrm{d}\xi) \qquad \tilde{q}_i \mid \boldsymbol{w}, \tilde{p}_{21},\dots,\tilde{p}_{2M} \overset{iid}{\sim} \sum_{m=1}^{M} w_m \delta_{\tilde{p}_{2m}}$$

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

$$(X_{1i}, X_{2i}) \mid c_{1i} = m, \theta^\star_m, \tilde{p}_{2M} \overset{ind}{\sim} k_1(X_{1i}; \theta^\star_m) \left(\sum_{s=1}^{S} q_{ms} k_2(X_{2i}; \xi^\star_s) \right)$$

A.5 Proof of Theorem 2

Proof. **(Theorem 2)** Proof of point(i)*: To prove point (ii) is sufficient to prove the existence of a telescopic clustering model with two layers that does not induce column-exchangeability. When $d \neq p$, the marginal laws of X_{11} and X_{21} are different and, thus, column-exchangeability does not hold true.* Proof of point(ii)*: Consider a telescopic model where the base measures at layer 1 and 2 are P_0 and Q_0, respectively. Consider the measurable sets $A_1, A_2, A_3 \subseteq \mathbb{X}_1$ and define the marginal likelihoods of each of the five possible cluster configurations for $n = 3$ observational units, i.e.*

$$\begin{aligned}
l_1(A_1, A_2, A_3; k_1, P_0) &:= \int k_1(A_1, \theta)\, k_1(A_2, \theta)\, k_1(A_3, \theta) \mathrm{d}P_0(\theta) && \textit{for } c_{11} = c_{12} = c_{13} \\
l_2(A_1, A_2, A_3; k_1, P_0) &:= \int k_1(A_1, \theta)\, k_1(A_2, \theta) \mathrm{d}P_0(\theta) \int k_1(A_3, \theta) \mathrm{d}P_0(\theta) && \textit{for } c_{11} = c_{12} \neq c_{13} \\
l_3(A_1, A_2, A_3; k_1, P_0) &:= l_2(A_1, A_3, A_2; k_1, P_0) && \textit{for } c_{11} = c_{13} \neq c_{12} \\
l_4(A_1, A_2, A_3; k_1, P_0) &:= l_3(A_2, A_1, A_3; k_1, P_0) && \textit{for } c_{12} = c_{13} \neq c_{11} \\
l_5(A_1, A_2, A_3; k_1, P_0) &:= \int k_1(A_1, \theta) \mathrm{d}P_0(\theta) \int k_1(A_2, \theta) \mathrm{d}P_0(\theta) \int k_1(A_3, \theta) \mathrm{d}P_0(\theta) && \textit{otherwise}
\end{aligned}$$

For measurable sets $A_1, A_2, A_3 \subset \mathbb{X}_1$ and $B_1, B_2, B_3 \subset \mathbb{X}_2$,

$$\begin{aligned}
&\mathbb{P}((X_{11}, X_{12}, X_{13}) \in A_1 \times A_2 \times A_3,\, (X_{21}, X_{22}, X_{23}) \in B_1 \times B_2 \times B_3) = \\
&\mathbb{P}(c_{11} = c_{12} = c_{13}) l_1(A_1, A_2, A_3; k_1, P_0) f_1(B_1, B_2, B_3) + \mathbb{P}(c_{11} = c_{13} \neq c_{12}) \sum_{i=2}^{4} l_i(A_1, A_2, A_3; k_1, P_0) f_i(B_1, B_2, B_3) \\
&+ \mathbb{P}(K_{13} = 3) l_5(A_1, A_2, A_3; k_1, P_0) f_5(B_1, B_2, B_3)
\end{aligned}$$

where f_i for $i = 1, 2, \ldots, 5$ are the prior predictive distributions $\mathbb{P}((X_{21}, X_{22}, X_{23}) \in B_1 \times B_2 \times B_3 \mid \rho_1)$ at the second layer conditional on each of the five first-layer clustering configurations. We have

$$\begin{aligned}
f_1(B_1, B_2, B_3) =& f_1(B_2, B_1, B_3) \\
f_2(B_1, B_2, B_3) =& f_2(B_2, B_1, B_3) \\
f_3(B_1, B_2, B_3) =& \mathbb{P}(c_{21} = c_{22} = c_{23} \mid c_{11} = c_{13} \neq c_{12}) l_1(B_1, B_2, B_3; k_2, Q_0) \\
&+ \mathbb{P}(c_{21} = c_{22} \neq c_{23} \mid c_{11} = c_{13} \neq c_{12}) l_2(B_1, B_2, B_3; k_2, Q_0) \\
&+ \mathbb{P}(c_{21} = c_{23} \neq c_{22} \mid c_{11} = c_{13} \neq c_{12}) l_3(B_1, B_2, B_3; k_2, Q_0) \\
&+ \mathbb{P}(c_{22} = c_{23} \neq c_{21} \mid c_{11} = c_{13} \neq c_{12}) l_4(B_1, B_2, B_3; k_2, Q_0) \\
&+ \mathbb{P}(K_{23} = 3 \mid c_{11} = c_{13} \neq c_{12}) l_5(B_1, B_2, B_3; k_2, Q_0) \\
f_4(B_1, B_2, B_3) =& \mathbb{P}(c_{21} = c_{22} = c_{23} \mid c_{12} = c_{13} \neq c_{11}) l_1(B_1, B_2, B_3; k_2, Q_0) \\
&+ \mathbb{P}(c_{21} = c_{22} \neq c_{23} \mid c_{12} = c_{13} \neq c_{11}) l_2(B_1, B_2, B_3; k_2, Q_0) \\
&+ \mathbb{P}(c_{21} = c_{23} \neq c_{22} \mid c_{12} = c_{13} \neq c_{11}) l_3(B_1, B_2, B_3; k_2, Q_0) \\
&+ \mathbb{P}(c_{22} = c_{23} \neq c_{21} \mid c_{12} = c_{13} \neq c_{11}) l_4(B_1, B_2, B_3; k_2, Q_0) \\
&+ \mathbb{P}(K_{23} = 3 \mid c_{12} = c_{13} \neq c_{11}) l_5(B_1, B_2, B_3; k_2, Q_0) \\
f_5(B_1, B_2, B_3) =& f_5(B_2, B_1, B_3)
\end{aligned}$$

Swapping now B_1 and B_2, we can compute the change in the joint probability as

$$\begin{aligned}
&\mathbb{P}((X_{11}, X_{12}, X_{13}) \in A_1 \times A_2 \times A_3,\, (X_{21}, X_{22}, X_{23}) \in B_1 \times B_2 \times B_3) \\
&- \mathbb{P}((X_{11}, X_{12}, X_{13}) \in A_1 \times A_2 \times A_3,\, (X_{21}, X_{22}, X_{23}) \in B_2 \times B_1 \times B_3)
\end{aligned} \tag{18}$$

which under partial exchangeability should equal 0 for any measurable sets $A_1, A_2, A_3, B_1, B_2, B_3$.However, the difference in (18) *equals*

$$[l_3(A_1, A_2, A_3; k_1, P_0) - l_4(A_1, A_2, A_3; k_1, P_0)] \times [l_3(B_1, B_2, B_3; k_2, Q_0) - l_4(B_1, B_2, B_3; k_2, Q_0)] \times$$
$$[\mathbb{P}(c_{21} = c_{23} \neq c_{22}, c_{11} = c_{13} \neq c_{12}) - \mathbb{P}(c_{22} = c_{23} \neq c_{21}, c_{11} = c_{13} \neq c_{12})]$$

which for $A_1 \neq A_2$ and $B_1 \neq B_2$ is in general different than zero.

A.6 Proof of Proposition 4

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

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

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,

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

References

Aldous, D. J., I. A. Ibragimov, J. Jacod, and D. J. Aldous (1985). *Exchangeability and related topics*. Springer.

Argiento, R., A. Cremaschi, and M. Vannucci (2020). Hierarchical normalized completely random measures to cluster grouped data. *Journal of the American Statistical Association 115*(529), 318–333.

Argiento, R. and M. De Iorio (2022). Is infinity that far? a Bayesian nonparametric perspective of finite mixture models. *The Annals of Statistics 50*(5), 2641–2663.

Ascolani, F., A. Lijoi, and M. Ruggiero (2021). Predictive inference with Fleming–Viot-driven dependent Dirichlet processes. *Bayesian Analysis 16*(2), 371–395.

Balocchi, C., E. I. George, and S. T. Jensen (2021). Clustering areal units at multiple levels of resolution to model crime incidence in Philadelphia. *Preprint at arXiv: 2112.02059*.

Barrios, E., A. Lijoi, L. E. Nieto-Barajas, and I. Prünster. Modeling with normalized random measure mixture models. *Statistical Science 28*(3), 313 – 334.

Bassetti, F., R. Casarin, and L. Rossini (2020). Hierarchical species sampling models. *Bayesian Analysis 15*(3), 809–838.

Beraha, M., A. Guglielmi, and F. A. Quintana (2021). The semi-hierarchical Dirichlet process and its application to clustering homogeneous distributions. *Bayesian Analysis 16*(4), 1187–1219.

Betancourt, B., G. Zanella, and R. C. Steorts (2022). Random partition models for microclustering tasks. *Journal of the American Statistical Association 117*(539), 1215–1227.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Camerlenghi, F., D. B. Dunson, A. Lijoi, I. Prünster, and A. Rodríguez (2019). Latent nested nonparametric priors (with discussion). *Bayesian Analysis 14*(4), 1303–1356.

Camerlenghi, F., A. Lijoi, P. Orbanz, and I. Prünster (2019). Distribution theory for hierarchical processes. *The Annals of Statistics 47*(1), 67–92.

Camerlenghi, F., A. Lijoi, and I. Prünster (2018). Bayesian nonparametric inference beyond the Gibbs-type framework. *Scandinavian Journal of Statistics 45*(4), 1062–1091.

Caron, F., M. Davy, and A. Doucet (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., W. Neiswanger, F. Wood, A. Doucet, and M. Davy (2017). Generalized Pólya urn for time-varying Pitman-Yor processes. *Journal of Machine Learning Research 18*, 1–32.

Chandra, N. K., A. Canale, and D. B. Dunson (2023). Escaping the curse of dimensionality in Bayesian model based clustering. *Journal of Machine Learning Research 24*, 1–42.

Chen, M.-S., J.-Q. Lin, X.-L. Li, B.-Y. Liu, C.-D. Wang, D. Huang, and J.-H. Lai (2022). Representation learning in multi-view clustering: A literature review. *Data Science and Engineering 7*(3), 225–241.

Cremaschi, A., M. De Iorio, N. Kothandaraman, F. Yap, M. T. Tint, and J. Eriksson (2021). Joint modelling of association networks and longitudinal biomarkers: an application to child obesity. *arXiv preprint arXiv:2111.06212*.

Dahl, D. B., R. Day, and J. W. Tsai (2017). Random partition distribution indexed by pairwise information. *Journal of the American Statistical Association 112*(518), 721–732.

Dahl, D. B., D. J. Johnson, and P. Müller (2022). Search algorithms and loss functions for Bayesian clustering. *Journal of Computational and Graphical Statistics 31*(4), 1189–1201.

Davis, C. S. (2002). Statistical methods for the analysis of repeated measurements. Technical report, Springer.

de Finetti, B. (1938). Sur la condition d'equivalence partielle. *Actualitès Scientifiques et Industrielles 739*, 5–18.

De Iorio, M., S. Favaro, A. Guglielmi, and L. Ye (2019). Bayesian nonparametric temporal dynamic clustering via autoregressive Dirichlet priors. *arXiv preprint arXiv:1910.10443*.

Denti, F., F. Camerlenghi, M. Guindani, and A. Mira (2021). A common atoms model for the Bayesian nonparametric analysis of nested data. *Journal of the American Statistical Association* (541), 1–12.

DeYoreo, M. and A. Kottas (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*, pp. 223–273. Cambridge Univ. Press.

Dunson, D. B. and J.-H. Park (2008). Kernel stick-breaking processes. *Biometrika 95*(2), 307–323.

Ellul, S., M. Wake, S. A. Clifford, K. Lange, P. Würtz, M. Juonala, T. Dwyer, J. B. Carlin, D. P. Burgner, and R. Saffery (2019). Metabolomics: population epidemiology and concordance in Australian children aged 11–12 years and their parents. *BMJ open 9*(Suppl 3).

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*, pp. 287–302. Elsevier.

Foti, N. J. and S. A. Williamson (2015). A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE Trans. Pattern Anal. Mach. Intell. 37*, 359–371.

Franzolini, B. (2022). *On Dependent Processes in Bayesian Nonparametrics: Theory, Methods, and Applications.* Bocconi University.

Franzolini, B., A. Cremaschi, W. v. d. Boom, and M. De Iorio (2023). Bayesian clustering of multiple zero-inflated outcomes. *Philosophical Transactions of the Royal Society A 81*(2247), 20220145.

Franzolini, B., A. Lijoi, and I. Prünster (2023). Model selection for maternal hypertensive disorders with symmetric hierarchical Dirichlet processes. *Annals of Applied Statistics 17*(1), 313–332.

Franzolini, B., A. Lijoi, I. Prünster, and G. Rebaudo (2023). Multivariate species sampling models. *Working Paper*.

Gao, L. L., J. Bien, and D. Witten (2020). Are clusterings of multiple data views independent? *Biostatistics 21*(4), 692–708.

Ghosal, S. and A. Van der Vaart (2017). *Fundamentals of nonparametric Bayesian inference*, Volume 44. Cambridge University Press.

Gil-Leyva, M. F. and R. H. Mena (2021). Stick-breaking processes with exchangeable length variables. *Journal of the American Statistical Association 18*(541), 1–14.

Green, P. J. and S. Richardson (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics 28*(2), 355–375.

Griffin, J. E. and F. Leisen (2017). Compound random measures and their use in Bayesian nonparametrics. *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 79*(2), 525–545.

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters 64*, 53–62.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of classification 2*(1), 193–218.

Kallenberg, O. (1989). On the representation theorem for exchangeable arrays. *Journal of Multivariate Analysis 30*(1), 137–154.

Kallenberg, O. (2005). *Probabilistic symmetries and invariance principles*, Volume 9. Springer.

Kaufman, L. and P. J. Rousseeuw (2009). *Finding groups in data: an introduction to cluster analysis.* John Wiley & Sons.

Kumar, A., P. Rai, and H. Daume (2011). Co-regularized multi-view spectral clustering. *Advances in neural information processing systems 24*.

Lee, J., P. Müller, Y. Zhu, and Y. Ji (2013). A nonparametric Bayesian model for local clustering with application to proteomics. *Journal of the American Statistical Association 108*(503), 775–788.

Lijoi, A., R. H. Mena, and I. Prünster (2005a). Bayesian nonparametric analysis for a generalized Dirichlet process prior. *Statistical Inference for Stochastic Processes 8*(3), 283–309.

Lijoi, A., R. H. Mena, and I. Prünster (2005b). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association 100*(472), 1278–1291.

Lijoi, A., B. Nipoti, and I. Prünster (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli 20*(3), 1260–1291.

Lijoi, A., I. Prünster, and G. Rebaudo (2023). Flexible clustering via hidden hierarchical Dirichlet priors. *Scandinavian Journal of Statistics 50*(1), 213–234.

Lijoi, A., I. Prünster, and T. Rigon (2020). The Pitman–Yor multinomial process for mixture modelling. *Biometrika 107*(4), 891–906.

Lijoi, A., I. Prünster, and T. Rigon (2023). Finite-dimensional discrete random structures and Bayesian clustering. *Journal of the American Statistical Association*, forthcoming.

Lin, Q., G. Rebaudo, and P. Mueller (2021). Separate exchangeability as modeling principle in Bayesian nonparametrics. *arXiv preprint arXiv:2112.07755*.

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. and D. B. Dunson (2013). Bayesian consensus clustering. *Bioinformatics 29*(20), 2610–2616.

MacEachern, S. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*.

MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical report, The Ohio State Univ.

McCullagh, P. and J. Yang (2008). How many clusters? *Bayesian Analysis 3*(1), 101–120.

McLachlan, G. J., S. X. Lee, and S. I. Rathnayake (2019). Finite mixture models. *Annual Review of Statistics and Its Application 6*, 355–378.

Meilă, M. (2007). Comparing clusterings–an information based distance. *Journal of Multivariate Analysis 98*(5), 873–895.

Mena, R. H. and S. G. Walker (2015). On the Bayesian mixture model and identifiability. *Journal of Computational and Graphical Statistics 24*(4), 1155–1169.

Miller, J. W. and M. T. Harrison (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association 113*(521), 340–356.

Müller, P., F. Quintana, and G. Rosner (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.

Ngan, H. Y. T., N. H. C. Yung, and A. G. O. Yeh (2015). Outlier detection in traffic data based on the Dirichlet process mixture model. *IET Intell. Transp. Syst. 9*, 773–781.

Nobile, A. (1994). *Bayesian analysis of finite mixture distributions.* Carnegie Mellon University.

Page, G. L. and F. A. Quintana (2018). Calibrating covariate informed product partition models. *Statistics and Computing 28*, 1009–1031.

Page, G. L., F. A. Quintana, and D. B. Dahl (2022). Dependent modeling of temporal sequences of random partitions. *Journal of Computational and Graphical Statistics 31*(2), 614–627.

Petrone, S., M. Guindani, and A. E. Gelfand (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*, 245–267.

Pitman, J. and M. Yor (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability 25*(2), 855–900.

Quintana, F. A., P. Müller, A. Jara, and S. N. MacEachern (2022). The dependent Dirichlet process and related models. *Statistical Science 37*(1), 24–41.

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

Ren, L., L. Du, L. Carin, and D. B. Dunson (2011). Logistic stick-breaking process. *Journal of Machine Learning Research 12*(1).

Ren, L., D. B. Dunson, and L. Carin (2008). The dynamic hierarchical Dirichlet process. In *Proceedings of the 25th international conference on machine learning*, pp. 824–831.

Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology) 59*(4), 731–792.

Rigon, T. and D. Durante (2021). Tractable Bayesian density regression via logit stick-breaking priors. *Journal of Statistical Planning and Inference 211*, 131–142.

Rodriguez, A. and D. B. Dunson (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian analysis 6*(1), 45–178.

Rodríguez, A., D. B. Dunson, and A. E. Gelfand (2008). The nested Dirichlet process (with discussion). *Journal of the American Statistical Association 103*(483), 1131–1154.

Rodríguez, A., D. B. Dunson, and A. E. Gelfand (2010). Latent stick-breaking processes. *Journal of the American Statistical Association 105*(490), 647–659.

Rogers, S., M. Girolami, W. Kolch, K. M. Waters, T. Liu, B. Thrall, and H. S. Wiley (2008). Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics 24*(24), 2894–2900.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*(2), 639–650.

Shotwell, M. S. and E. H. Slate (2011). Bayesian outlier detection with Dirichlet process mixtures. *Bayesian Analysis 6*(4), 665–690.

Soh, S.-E., M. T. Tint, P. D. Gluckman, K. M. Godfrey, A. Rifkin-Graboi, Y. H. Chan, W. Stünkel, J. D. Holbrook, K. Kwek, Y.-S. Chong, et al. (2014). Cohort profile: Growing Up in Singapore Towards healthy Outcomes (GUSTO) birth cohort study. *International journal of epidemiology 43*(5), 1401–1409.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(4), 795–809.

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

Tibshirani, R., G. Walther, and T. Hastie (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.

Van Dyk, D. A. and T. Park (2008). Partially collapsed gibbs samplers: Theory and methods. *Journal of the American Statistical Association 103*(482), 790–796.

Wade, S. (2023). Bayesian cluster analysis. *Philosophical Transactions of the Royal Society A 381*(2247), 20220149.

Wade, S. and Z. Ghahramani (2018). Bayesian cluster analysis: point estimation and credible balls (with discussion). *Bayesian Analysis 13*(2), 559–626.

Wade, S., S. Mongelluzzo, and S. Petrone (2011). An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Analysis 6*(3), 359–385.

WHO (2007). Computation of centiles and z-scores for height-for-age, weight-for-age and BMI-for-age. *Geneva: World Health Organization.*

Yang, Y. and H. Wang (2018). Multi-view clustering: A survey. *Big Data Mining and Analytics 1*(2), 83–107.

Yao, S., G. Yu, J. Wang, C. Domeniconi, and X. Zhang (2019). Multi-view multiple clustering. *arXiv preprint arXiv:1905.05053*.

Zhang, C., Y. Qin, X. Zhu, J. Zhang, and S. Zhang (2006). Clustering-based missing value imputation for data preprocessing. In *IEEE Int. Conf. Industr. Inform.*, pp. 1081–1086.

Zhou, D., Y. Gao, and L. Paninski (2021). Disentangled sticky hierarchical Dirichlet process hidden Markov model. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part I*, pp. 612–627. Springer.

Supplement to *"Conditional partial exchangeability: a probabilistic framework for multi-view clustering"*

Beatrice Franzolini, Maria De Iorio, and Johan Eriksson

S1 Finite row-exchangeability of telescopic clustering

Assume that $(X_{1i}, X_{2i})_{i=1}^n$ is a finite sample distributed accordingly to a telescopic clustering model. By marginal exchangeability of $(X_{1i})_{i=1}^n$, we have that, for any measurable $A \subset \mathbb{X}$ and any $\sigma \in \mathcal{P}(n)$, with $\mathcal{P}(n)$ set of permutations of n elements,

$$\mathbb{P}\left[(X_{1i})_{i=1}^n \in A\right] = \mathbb{P}\left[(X^{(1)}_{\sigma(i)})_{i=1}^n \in A\right] \tag{19}$$

Moreover, since the marginal model at layer 1 admits the equivalent representation

$$\begin{aligned} (X_{1i}, \theta_i) \mid \tilde{p}_1 &\overset{ind}{\sim} k_1(X_{1i}, \theta_i) \times \tilde{p}_1(\mathrm{d}\theta_i) \qquad \text{for } i = 1, \dots, n \\ \tilde{p}_1 &\sim P_1 \end{aligned}$$

we have that $(X_{1i}, \theta_i)_{i=1}^n$ is exchangeable, i.e.,

$$\mathbb{P}\left[(X_{1i}, \theta_i)_{i=1}^n \in A \times B\right] = \mathbb{P}\left[(X^{(1)}_{\sigma(i)}, \theta_{\sigma(i)})_{i=1}^n \in A \times B\right] \tag{20}$$

for any measurable $B \subset \Theta$, and, therefore, by (19) and (20)

$$\mathbb{P}\left[(\theta_i)_{i=1}^n \in B \mid (X_{1i})_{i=1}^n \in A\right] = \mathbb{P}\left[(\theta_{\sigma(i)})_{i=1}^n \in B \mid (X^{(1)}_{\sigma(i)})_{i=1}^n \in A\right]$$

Moreover, we note that the partition ρ_1 is a deterministic function of the latent parameters $(\theta_1, \dots, \theta_n)$ thus its posterior law has to preserve the same invariance of the posterior of the latent parameters, i.e., for any $\sigma \in \mathcal{P}(n)$,

$$\mathbb{P}\left[\rho_1 = \rho_1 \mid (X_{1i})_{i=1}^n \in A\right] = \mathbb{P}\left[\rho_1 = \sigma(\rho_1) \mid (X^{(1)}_{\sigma(i)})_{i=1}^n \in A\right] \tag{21}$$

where $\sigma(\rho_1)$ is the partition obtained applying the permutation σ to the elements in the clusters identified by ρ_1.

Consider now the second layer and a measurable rectangle $C = \bigotimes_{i=1}^n C_i$, note that

$$\mathbb{P}\left[(X_{2i})_{i=1}^n \in C \mid (X_{1i})_{i=1}^n \in A)\right]$$

equals

$$\sum_{\rho_1 \in \Pi(n)} \left\{\mathbb{P}\left[(X_{2i})_{i=1}^n \in C \mid \rho_1 = \rho_1\right] \mathbb{P}\left[\rho_1 = \rho_1 \mid (X_{1i})_{i=1}^n \in A\right]\right\} \tag{22}$$

where $\Pi(n)$ is the set of partitions of n elements and $\mathbb{P}[(X_{2i})_{i=1}^n \in C \mid \rho_1 = \rho_1]$ is

$$\begin{aligned} &\int_{\mathcal{P}^M_{\mathbb{X}_2}} \prod_{m \in \boldsymbol{m}} \prod_{i: c_{1i} = m} \int_{\Theta_2} \int_{C_i} k_2(x, \theta) \mathrm{d}x \, \tilde{p}_{2m}(\mathrm{d}\theta) P_2(\mathrm{d}\tilde{p}_{21} \dots \mathrm{d}p_{2M}) \\ &= \int_{\mathcal{P}^M_{\mathbb{X}_2}} \prod_{m \in \boldsymbol{m}} \prod_{i: c_{\sigma(i)} = m} \int_{\Theta_2} \int_{C_{\sigma(i)}} k_2(x, \theta) \mathrm{d}x \, \tilde{p}_{2m}(\mathrm{d}\theta) P_2(\mathrm{d}\tilde{p}_{21} \dots \mathrm{d}p_{2M}) \\ &= \mathbb{P}((X_{2\sigma(i)})_{i=1}^n \in C \mid \rho_1 = \sigma(\rho_1)) \end{aligned} \tag{23}$$

where $\mathcal{P}_{\mathbb{X}_2}$ is the space of all probability measures on $\mathbb{X}_2$ and the mixing or de Finetti measure P_2 is a probability measure on $\mathcal{P}^M_{\mathbb{X}_2}$.

The extension of the result in (23) to any measurable set C can be obtained thanks to Dynkin's π-λ theorem, recalling that rectangles are a generating π-system of the Borel product σ-algebra and that the set of measurable C for which (23) holds true is easily proved to be a λ-system.

Putting together (22) with (21) and (23), for any measurable A and C, we get

$$\mathbb{P}\left[(X_{2i})_{i=1}^{n} =\in C \mid\ (X_{1i})_{i=1}^{n} \in A\right] = \mathbb{P}\left[(X_{2\sigma(i)})_{i=1}^{n} \in C \mid (X_{\sigma(i)}^{(1)})_{i=1}^{n} \in A\right]$$

Finally, point (i) is proved considering the joint prior predictive distribution of the whole matrix $(X_{1i}, X_{2i})_{i=1}^{n}$ obtained as

$$\mathbb{P}\left[(X_{1i})_{i=1}^{n} \in A\right] \mathbb{P}\left[(X_{2i})_{i=1}^{n} \in C \mid\ (X_{1i})_{i=1}^{n} \in A\right]$$

S2 Sampling schemes for generic telescopic clustering models

For simplicity of exposition, the algorithms for the general class of telescopic clustering models are here presented referring to the Markovian graphical structure in Figure 2, whose special cases include telescopic clustering with two layers. Algorithms for different graph structures can be obtained analogously. As an example of this, see the sampling strategy derived for the t-HDP in Section S3.1 which is suitable for any polytree structure of dependence across layers.

S2.1 Marginal MCMC

In this section, both the underlying random probabilities and the cluster-specific parameters are marginalized out. The sampling of the partitions is then performed based on the exchangeable partition probability function (EPPF) of the first layer and the conditional partial exchangeable partition probability functions (c-pEPPF) of the subsequent layers. The algorithms' output is a posterior sample from the telescopic clustering configuration only. The marginal MCMC's core structure is in Algorthm 1.

Algorithm 1 Markov chain Monte Carlo - Marginal algorithm

```
    **Input**: Data matrix $(X_{ti}, t = 1, \dots, T)_{i=1}^{n}$
    **Output**: posterior distribution of $(\rho_t, t = 1, \dots, T)$

  Sample $\rho_1$ from its full conditional proportional to
      $$\mathbb{P}(\rho_1)\mathbb{P}(\boldsymbol{X}_1 \mid \rho_1)\mathbb{P}(\rho_2 \mid \rho_1)$$
  **for** $t$ in 2:$(T-1)$ **do**
      Sample $\rho_t$ from its full conditional proportional to
          $$\mathbb{P}(\rho_t \mid \rho_{t-1})\mathbb{P}(\boldsymbol{X}_t \mid \rho_t)\mathbb{P}(\rho_{t+1} \mid \rho_t)$$
  Sample $\rho_T$ from its full conditional proportional to
      $$\mathbb{P}(\rho_T \mid \rho_{T-1})\mathbb{P}(\boldsymbol{X}_T \mid \rho_T)$$
```

Algorthm 1 requires to sample from the full conditional of the partition ρ_t, for $t = 1, \dots, T$. To derive the full conditional, we recall that $\boldsymbol{X}_t$ are the observations at layer t, c_{ti} is the unordered allocation variable for the ith subject referring to the partition at layer t. We denote with $\boldsymbol{c}_t$ the collection of all allocation variables identifying the partition ρ_t, i.e., $\boldsymbol{c}_t = (c_{ti} : i = 1, \ldots, n)$ and with $\boldsymbol{c}_t^{-i}$ the vector where the ith entry as been removed, i.e., $\boldsymbol{c}_t^{-i} = (c_{tj} : j \in [n] \setminus \{i\})$. First of all we note that the full conditional of the partition at layer t is

$$\begin{aligned}\mathbb{P}(\rho_t \mid \rho_1, \ldots, \rho_{t-1}, \rho_{t+1}, \ldots, \rho_T, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_T) &\propto \mathbb{P}(\rho_1) \prod_{s=2}^{T} \mathbb{P}(\rho_s \mid \rho_{s-1}) \prod_{s=1}^{T} \mathbb{P}(\boldsymbol{X}_s \mid \rho_s) \\ &\propto \mathbb{P}(\rho_{t-1} \mid \rho_t) \mathbb{P}(\rho_t \mid \rho_{t-1}) \mathbb{P}(\boldsymbol{X}_t \mid \rho_t)\end{aligned}$$

Sampling ρ_t from its full conditional is typically unfeasible since it requires evaluating the c-pEPPF, i.e., $\mathbb{P}(\rho_t \mid \rho_{t-1})$, for all possible realizations of ρ_t. This problem is not specific of telescopic clustering models. EPPFs and similar probability mass functions describing the law of partitions have always a large support that increases with n accordingly to the Bell number of n, and thus a posteriori is typically unfeasible to sample directly from them. The sampling of the partition in probabilistic clustering models is usually done by sampling each subject-specific allocation variable c_{ti} at a time, conditional on all the others. Following this strategy for telescopic clustering, we have:

$$\begin{aligned}\mathbb{P}(c_{ti} = m \mid \boldsymbol{X}_t, \boldsymbol{c}_t^{-i}, \rho_{t-1}, \rho_{t+1}) &\propto \mathbb{P}(c_{ti} = m, X_{it}, \rho_{t+1}, \mid \boldsymbol{X}_t^{-i}, \boldsymbol{c}_t^{-i}, \rho_{t-1}) \\ &= \mathbb{P}(c_{ti} = m, X_{it} \mid \boldsymbol{X}_t^{-i}, \boldsymbol{c}_t^{-i}, \rho_{t-1}) \mathbb{P}(\rho_{t+1} \mid c_{ti} = m, \boldsymbol{X}_t, \boldsymbol{c}_t^{-i}, \rho_{t-1}) \\ &= \mathbb{P}(c_{ti} = m, X_{it} \mid \boldsymbol{X}_t^{-i}, \boldsymbol{c}_t^{-i}, \rho_{t-1}) \mathbb{P}(\rho_{t+1} \mid c_{ti} = m, \boldsymbol{c}_t^{-i}) \\ &= \frac{\mathbb{P}(c_{ti} = m, \boldsymbol{c}_t^{-i}, \boldsymbol{X}_t \mid \rho_{t-1})}{\mathbb{P}(\boldsymbol{c}_t^{-i}, \boldsymbol{X}_t^{-i} \mid \rho_{t-1})} \mathbb{P}(\rho_{t+1} \mid c_{ti} = m, \boldsymbol{c}_t^{-i}) \\ &= \frac{\mathbb{P}(c_{ti} = m, \boldsymbol{c}_t^{-i} \mid \rho_{t-1})}{\mathbb{P}(\boldsymbol{c}_t^{-i} \mid \rho_{t-1})} \frac{\mathbb{P}(\boldsymbol{X}_t \mid c_{ti} = m, \boldsymbol{c}_t^{-i})}{\mathbb{P}(\boldsymbol{X}_t^{-i} \mid \boldsymbol{c}_t^{-i})} \mathbb{P}(\rho_{t+1} \mid c_{ti} = m, \boldsymbol{c}_t^{-i})\end{aligned}$$

This means that c_{ti} should be sampled accordingly to

$$\begin{aligned}&p(c_{ti} = m \mid \boldsymbol{c}_t^{-i}, \boldsymbol{c}_{t-1}, \boldsymbol{c}_{t+1}, X^{(t)}) = \\ &\qquad \text{Past}_{imt}(\boldsymbol{c}_t^{-i}, \boldsymbol{c}_{t-1}) \times \text{Fut}_{imt}(\boldsymbol{c}_t^{-i}, \boldsymbol{c}_{t+1}) \times \text{Lik}_{imt}(\boldsymbol{c}_t^{-i}, \boldsymbol{X}_t)\end{aligned}$$

where

$$\text{Past}_{imt}(\boldsymbol{c}_t^{-i}, \boldsymbol{c}_{t-1}) = \begin{cases} \frac{\mathbb{P}(c_{ti}=m, \boldsymbol{c}_t^{-i})}{\mathbb{P}(\boldsymbol{c}_t^{-i})} & \text{for } t = 1 \\ \frac{\mathbb{P}(c_{ti}=m, \boldsymbol{c}_t^{-i} \mid \rho_{t-1})}{\mathbb{P}(\boldsymbol{c}_t^{-i} \mid \rho_{t-1})} & \text{for } t = 2, \ldots, T \end{cases}$$

$$\text{Fut}_{imt} = \begin{cases} \mathbb{P}(\rho_{t+1} \mid c_{ti} = m, \boldsymbol{c}_t^{-i}) & \text{for t=1,\ldots,T-1} \\ 1 & \text{for T=1} \end{cases}$$

and

$$\text{Lik}_{imt} = \begin{cases} \dfrac{\int k_t(x_i^{(t)}, \theta) \prod_{\substack{j: c_{jt} = m \\ j \neq i}} k_t(x_j^{(t)}, \theta) \mathrm{d}P_\theta(\theta)}{\int \prod_{\substack{j: c_{jt} = m \\ j \neq i}} k_t(x_j^{(t)}, \theta) \mathrm{d}P_\theta(\theta)} & \text{if } m \in \boldsymbol{c}_t^{-i} \\ \int k_t(x_i^{(t)}, \theta) \mathrm{d}P_\theta(\theta) & \text{otherwise} \end{cases}$$

Thus, the complexity and the mixing performance of this strategy largely depend on two aspects. The first is how fast the cluster-specific marginal likelihood

$$\int_{\Theta_t} \prod_{j: c_{tj} = m} k_t(x_{tj}, \theta) \mathrm{d}P_\theta(\theta)$$

can be computed. In this regard, the best scenario is when the kernel and the base measure are conjugate so that typically a closed-form expression for the marginal likelihood is available. The second important aspect is how fast the ratio Past_{imt} and the factor Fut_{imt} can be computed. These both depend on the specific model chosen and may require the use of auxiliary random variables to be computed. For instance, when we devise a marginal algorithm for the t-HDP, Past_{imt} can be simplified by introducing the auxiliary variables referring to the labels of the tables in the restaurant franchise metaphor (see, Teh et al., 2006, for more details). Nonetheless, computing Fut_{imt} still requires evaluating the c-pEPPF $\mathbb{P}(\rho_{t+1} \mid \rho_t)$ for a high number of realizations of ρ_t at each iteration. Moreover, the introduction of latent variables to simplify this computation may not be a viable strategy. For instance, with the t-HDP, introducing the table labels of the subsequent layer slow the mixing of the chain of parents nodes to unfeasible levels. Whenever a specific telescopic clustering model is affected by these problems there exist two possible solutions: the first is to employ a conditional algorithm, and the second is to derive a block marginal Gibbs sampler. They are described in the next two sections.

S2.2 Conditional MCMC sampler

Conditional algorithms are a convenient strategy when the full posterior of the random probability is easier to sample compared to the evaluation of the partition's probability mass function. In fact, conditionally on the random probabilities, the full conditional of the allocation variable c_{ti} largely simplifies since it does not depend on observations other than X_{ti}, for t varying.

To derive the conditional sampler for a generic telescopic clustering model, denote with

- $\pi(m, k, t)$ the weight associated to the kth component of $\tilde{p}_m^{(t)}$
- $\theta^\star(m, k, t)$ the atom associated to the kth component of $\tilde{p}_m^{(t)}$

Algorithm 2 Conditional sampler

```
    Input: Data matrix $(X_{ti}, t = 1, \dots, T)_{i=1}^n$
    Output: posterior distribution of $\rho_1$ and $\rho_2$

for $i$ in 1:$n$ do
    Sample $(c_{ti})_t$ from
        $$p[(c_{ti})_{t=1}^T = (c_t)_{t=1}^T] \propto \prod_{t=1}^T [\pi(c_{t-1}, c_t, t)\kappa_t(X_{ti}; \theta^\star(c_{t-1}, c_t, t))]$$
Sample $\pi(m, k, t)$ and $\theta^\star(m, k, t)$ (full conditional does not depends on $(\boldsymbol{X}_s)_{s \neq t}$)
```

S2.3 Block Marginal Gibbs sampling for two layers

When the number of layers is small, e.g., $T = 2$, a marginal sampling scheme can be devised accordingly to Algorithm 3 can be employed. Contrary to Algorithm 1, each allocation variable is sampled by integrating out the allocation variables of descendant/future layers of the same subject, resulting in a block structure where each subject is allocated to all layers conditional on the other subjects' allocation. However, since the allocation variable at descendent layers is integrated out, each layer is sampled from a distribution that depends also on observations at subsequent layers. Algorithm 3 should, in general, provide a better mixing per iteration compared to Algorithm 2 thanks to the fact that descendant layers and all random probabilities are integrated out, however, such marginalization is only feasible for a limited number of layers since it increases the computational time per iteration proportionally to the number of descendant layers.

Algorithm 3 Block Marginal Gibbs sampling

```
    Input: Data matrix $(X_{1i}, X_{2i})_{i=1}^{n}$
    Output: smoothing posterior distribution of $\rho_1$ and $\rho_2$

for $i$ in 1:$n$ do
    Sample $c_{1i}$ from $p(c_{1i} \mid \boldsymbol{c}_1^{-i}, \boldsymbol{c}_2^{-i}, \boldsymbol{X}_1, \boldsymbol{X}_2)$, where
        $$p(c_{1i} = m \mid \boldsymbol{c}_1^{-i}, \boldsymbol{c}_2^{-i}, \boldsymbol{X}_1, \boldsymbol{X}_2) \propto \begin{cases} C_m \frac{p(\rho_1^{-i} \cap \{c_{1i}=m\})}{p(\pi_1^{-i})} \frac{\int k_1(X_{1i},\theta) \prod_{j:c_{1j}=m} k_1(X_{1j},\theta) \mathrm{d}P_\theta(\theta)}{\int \prod_{j:c_{1j}=m} k_1(X_{1j},\theta) \mathrm{d}P_\theta(\theta)} & \text{if } m \in \boldsymbol{c}_1^{-i} \\ C_m \frac{p(\rho_1^{-i} \cap \{c_{1i}=m\})}{p(\pi_1^{-i})} \int k_1(X_{1i}, \theta) \mathrm{d}P_\theta(\theta) & \text{otherwise} \end{cases}$$
     where $C_m$ is the marginal likelihood of the second layer, i.e., for $m \in \boldsymbol{c}_1^{-i}$,
        $$C_m = \sum_s \frac{p(\rho_2^{-i} \cap \{c_{2i} = s\}) \mid \rho_1)}{p(\rho_2^{-i} \mid \rho_1)} \frac{\int k_2(X_{2i}, \theta) \prod_{j:c_{2j}=s} k_2(X_{2j}, \xi) \mathrm{d}P_\xi(\xi)}{\int \prod_{j:c_{2j}=s} k_2(X_{2j}, \xi) \mathrm{d}P_\xi(\xi)}$$
    Sample $c_{2i}$ from $p(c_{2i} \mid \boldsymbol{c}_1, \boldsymbol{c}_2^{-i}, \boldsymbol{X}_2)$, where
        $$p(c_{2i} = s \mid \boldsymbol{m}, \boldsymbol{s}^{-i}, \boldsymbol{X}_2) \propto \begin{cases} \frac{p(\rho_2^{-1} \cap \{c_{2i}=s\} | \rho_1)}{p(\pi_2^{-i} | \rho_1)} \frac{\int k_2(X_{2i},\theta) \prod_{j:c_{2j}=s} k_2(X_{2j},\xi) \mathrm{d}P_\xi(\xi)}{\int \prod_{j:c_{2j}=s} k_2(X_{2j},\xi) \mathrm{d}P_\xi(\xi)} & \text{if } s \in \boldsymbol{c}_2^{-i} \\ \frac{p(\rho_2^{-1} \cap \{c_{2i}=s\} | \rho_1)}{p(\pi_2^{-i} | \rho_1)} \int k_2(X_{2i}, \xi) \mathrm{d}P_\xi(\xi) & \text{otherwise} \end{cases}$$
  where: $\prod_{s \in \emptyset} := 1$.
```

S3 Sampling schemes for t-HDP

As already noticed in Section S2, the marginal sampling scheme as devised for a general telescopic sampler is not a viable alternative for the t-HDP. In particular, adopting the general marginal sampler, require to evaluate Fut_{imt} which is computationally non-feasible, and cannot be solved with the introduction of the typical latent variables employed with the HDP, because will results in a slow mixing, which decreases drastically for layers with a high number of descendants.

Thus in the following, we provide a faster conditional sampler, obtained by combining block and partially collapsed Gibbs sampling steps, that can be employed for any reasonable number of layers, we tested the performance to up to 100 layers, and a Block Marginal Gibbs sampling that can be employed when the number of layers is small.

S3.1 Partially collapsed conditional block Gibbs sampler

Denote with

- $\pi_0(k, \ell)$ the weight associated to the kth component of $\tilde{q}_0^{(\ell)}$

- $\theta_0^\star(k,\ell)$ the atom associated to the kth component of $\tilde{q}_0^{(\ell)}$
- $\pi(m,k,\ell)$ the weight associated to the kth component of $\tilde{p}_m^{(\ell)}$
- $c(\ell,i)$ label of the table at layer ℓ of the ith customer
- $k(\ell,c)$ label of the dish served at layer ℓ at table c
- $m(\ell,i)$ label of the dish eaten at layer ℓ by the ith customer (thus: $m(0,i)=1$ for all i and $m(\ell,i)=k(\ell,c(\ell,i))$)

The truncated stick breaking version of the t-HDP can be written as follows

$$\begin{aligned}
\pi_0(\cdot,\ell) = [\pi_0(1,\ell),\dots,\pi_0(H_0,\ell)] &\overset{iid}{\sim} \mathrm{TSB}(\alpha_0,H_0) && \text{for } \ell=1,\dots,L\\
\pi(m,\cdot,\ell) = [\pi(m,1,\ell),\dots,\pi_0(m,H,\ell)] &\overset{iid}{\sim} \mathrm{TSB}(\alpha,H) && \text{for } m=1,\dots,H_0\\
&&& \text{and } \ell=1,\dots,L\\
\theta_0^\star(\cdot,\ell)[\theta_0^\star(1,\ell),\dots,\theta_0^\star(H_0,\ell)] &\overset{iid}{\sim} \bigtimes_{h=1}^{H_0} P_0 && \text{for } \ell=1,\dots,L\\
k(\ell,\cdot) = [k(\ell,1),\dots,k(\ell,H_0\times H)] \mid \pi_0(\cdot,\ell) &\overset{ind}{\sim} \bigtimes_{c=1}^{H_0\times H}\left(\sum_{h}^{H_0}\pi_0(h,\ell)\delta_h\right) && \text{for } \ell=1,\dots,L\\
c(\ell,\cdot) = [c(\ell,1),\dots,c(\ell,n)] \mid \pi(\cdot,\cdot,\ell), m(\mathrm{par}(\ell),\cdot)&\\
&\overset{ind}{\sim} \bigtimes_{i=1}^{n}\left(\sum_{h=1}^{H}\pi(m(\mathrm{par}(\ell),i),h,\ell)\delta_{[(m(\mathrm{par}(\ell),i)-1)H+h]}\right) && \text{for } \ell=1,\dots,L\\
m(\ell,i) \mid k(\ell,\cdot),c(\ell,\cdot) &\overset{ind}{\sim} \delta_{k(\ell,c(\ell,i))} && \text{for } i=1,\dots,n\\
&&& \text{and } \ell=1,\dots,L\\
X_{\ell i} \mid \theta_0^\star, m(\ell,i) &\overset{ind}{\sim} \kappa_\ell(\cdot,\theta_0^\star(m(\ell,i),\ell)) && \text{for } i=1,\dots,n\\
&&& \text{and } \ell=1,\dots,L
\end{aligned}$$

Denote also with

- C_ℓ the set of unique values in $c(\ell,\cdot)$ (actually occupied tables)
- $n(\ell,c)$ number of customer at layer ℓ sat at table c
- $q(\ell,h)$ number of tables at layer ℓ serving dish h
- $\bar{n}(\ell,h_1,h_2)=n(\ell,(h_1-1)\times H+h_2)$

Figure S3.1 shows the corresponding graphical model when the number of layers equals three and the dependence across layers is triangular as in Figure 3. Algorithm 4 contains the pseudo-code of the conditional algorithm to estimate the t-HDP model for any number of layers and any polytree structure. The algorithm is derived based on the truncated stick-breaking version of the t-HDP, described here above, and it is obtained by combining block and partially collapsed Gibbs sampling steps. In particular, $c(\ell,i)$ and $m(\ell,i)$ are sampled as a block from which $\{c(\ell,i), \text{with } \ell\in\mathrm{child}(\ell)\}$ are marginalized out. This drastically improves the mixing of the chain compared to a classical Gibbs sampler and leads to the correct stationary distribution for the chain (cfr., Van Dyk and Park, 2008).

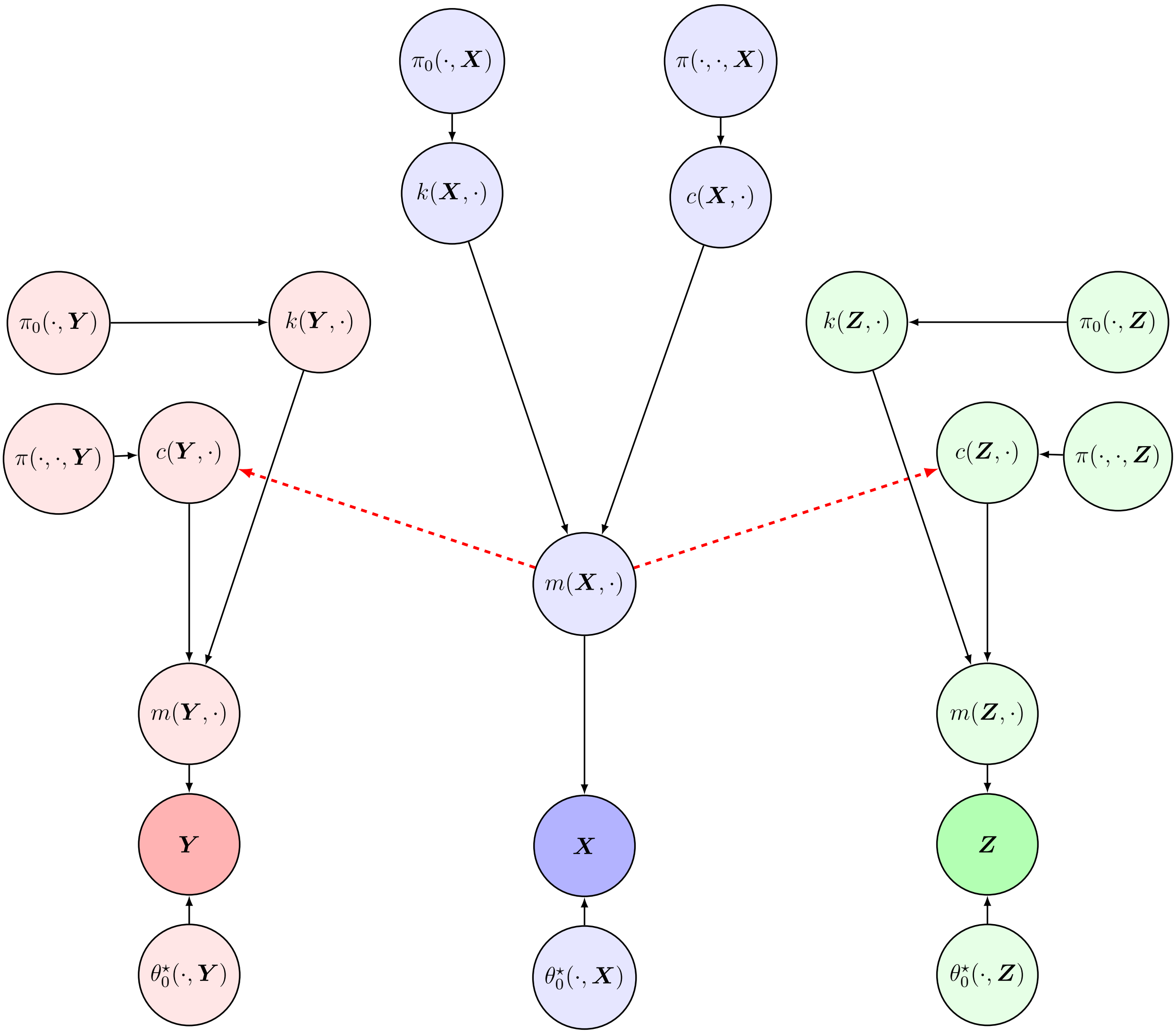

$\pi_0(\cdot, \boldsymbol{X})$
$\pi(\cdot, \cdot, \boldsymbol{X})$
$k(\boldsymbol{X}, \cdot)$
$c(\boldsymbol{X}, \cdot)$
$\pi_0(\cdot, \boldsymbol{Y})$
$k(\boldsymbol{Y}, \cdot)$
$k(\boldsymbol{Z}, \cdot)$
$\pi_0(\cdot, \boldsymbol{Z})$
$\pi(\cdot, \cdot, \boldsymbol{Y})$
$c(\boldsymbol{Y}, \cdot)$
$c(\boldsymbol{Z}, \cdot)$
$\pi(\cdot, \cdot, \boldsymbol{Z})$
$m(\boldsymbol{X}, \cdot)$
$m(\boldsymbol{Y}, \cdot)$
$m(\boldsymbol{Z}, \cdot)$
$\boldsymbol{Y}$
$\boldsymbol{X}$
$\boldsymbol{Z}$
$\theta_0^\star(\cdot, \boldsymbol{Y})$
$\theta_0^\star(\cdot, \boldsymbol{X})$
$\theta_0^\star(\cdot, \boldsymbol{Z})$

Figure S3.1: Graphical model corresponding to a t-HDP with the truncated stick-breaking representation with a triangular layer dependence.

Algorithm 4 Conditional sampler - t-HDP

```
**Input**: Data matrix $(X_{1i}, X_{2i})_{i=1}^{n}$
**Output**: smoothing posterior distribution of $\rho_1$ and $\rho_2$

for $\ell$ in 1:$L$ do
    for $h_1$ in 1:$H_0$ do
        Sample $b_0(h_1,\ell)$ from Beta $\left(1+q(\ell,h_1), \alpha_0+\sum_{s=h_1+1}^{H_0} q(\ell,s)\right)^{\diamond}$
        $\pi_0(h_1,\ell) \leftarrow b_0(h_1,\ell)\prod_{s=1}^{h_1-1} b_0(s,\ell)^{\diamond}$
        Sample $\theta_0^{\star}(h_1,\ell)$ from $p(\theta) \propto \prod_{i:m(\ell,i)=h_1} \kappa_\ell(X_{\ell i},\theta)P_0(\mathrm{d}\theta)^{\diamond}$
        for $h_2$ in 1:$H$ do
            Sample $b(h_1,h_2,\ell)$ from Beta $\left(1+\bar{n}(\ell,h_1,h_2), \alpha_0+\sum_{s=h_2+1}^{H} \bar{n}(\ell,h_1,s)\right)^{\diamond}$
            $\pi(h_1,h_2,\ell) \leftarrow b(h_1,h_2,\ell)\prod_{s=1}^{h_2-1} b(h_1,s,\ell)^{\diamond}$

for $\ell$ in 1:$L$ do
    for $i$ in 1:$n$ do
        $m \leftarrow m(\ell-1,i)$
        $f \leftarrow m(\ell+1,i)$
        Sample $c(\ell,i)$ from $p(c)$ with $c \in \{(m-1)H+1,\dots,mH\}$, where

            $p(c) \propto \pi(m,c,\ell) \times \kappa_\ell(X_{\ell i}\,;\theta_0^{\star}(k(\ell,c),\ell)) \times \prod_{\ell^{\star}\in\mathrm{child}(\ell)} \left(\sum_{d\in\mathcal{M}_{\ell cf}} \pi(k(\ell,c),d,\ell^{\star})\right)$
                    where $\mathcal{M}_{\ell cf} = \{d : k(\ell^{\star},[k(\ell,c)-1]H+d) = f\}$

    for $c$ in $1:(H\times H_0)$ do
        Sample $k(\ell,c)$ from $p(k)$, with, for $k \in \{1,\dots H_0\}$,

            $p(k) \propto \pi_0(k,\ell)\prod_{i:c(\ell,i)=c} \kappa_\ell(X_{\ell i}\,;\theta_0^{\star}(k,\ell))^{\diamond}$

    for $i$ in 1:$n$ do
        $m(\ell,i) \leftarrow k(\ell,c(\ell,i))$
    for $h$ in 1:$H_0$ do
        $q(\ell,h) \leftarrow \sum_{c=1}^{H\times H_0} \mathbb{1}(k(\ell,c)=h)\,\mathbb{1}(c\in\mathcal{C}_\ell)$
    for $c$ in 1:$H\times H_0$ do
        $n(\ell,c) \leftarrow \sum_{i=1}^{n} \mathbb{1}(c(\ell,i)=c)$
```

$\diamond$ we use the conventions: $\sum_{s=H_0+1}^{H_0} q(\ell,s) := 0$, $\prod_{s=1}^{0} b_0(s,\ell) := 1$, $\prod_{i\in\emptyset} x_i = 1$

S3.2 Block Marginal Gibbs sampling for two layers

Referring to the Chinese restaurant metaphor used to describe the predictive law of the hierarchical Dirichlet process as in Teh et al. (2006), denote with c_{ti}, the label of the table at which the ith client is sat at layer t and with c_{ti} the dish eaten by the ith client at layer t. We recall that all clients that sat at the same table eat the same dish and that the same dish can be served at more than one table. According to the metaphor, c_{ti} encodes the clustering structure of interest, while c_{ti} are auxiliary latent parameters that are used to simplify the full conditional distribution from which the cluster configuration has to be sampled in a Gibbs sampler.

Denote with

- $\mathcal{C}_1$ the set of tables' labels at layer 1
- $\mathcal{M}_1$ the set of dishes' labels at layer 1
- $\mathcal{C}_2$ the set of tables' labels at layer 2
- $\mathcal{M}_2$ the set of dishes' labels at layer 2
- $\mathcal{C}_{2|m}$ the set of tables' labels at layer 2 restricted to those clients that at layer 1 were eating dish m
- n_{1c} number of customer at layer 1 sat at table c
- $n_{2,c|m}$ number of customer sat at table c at layer 2 and eating dish m at layer 1
- q_{1m} number of tables at layer 1 serving dish m
- q_{2m} number of tables at layer 2 serving dish m
- $d_\ell(c)$ a function returning the label of the dish served at table c of layer ℓ

At layer 1, to sample c_{1i} from $p(c_{1i} \mid \boldsymbol{c}_1^{-i}, \boldsymbol{X}_1, \boldsymbol{X}_2)$, we first sample the table allocation variable c_{1i} from

$$p(c_{1i} = c \,|\boldsymbol{c}_1^{-i}, \boldsymbol{c}_1^{-i}, \boldsymbol{X}_1, \boldsymbol{X}_2)$$

$$\propto \begin{cases} C_{d_1(c)}\, n_{1c}^{-i} \dfrac{\int k_1(X_{1i},\theta) \prod\limits_{j:c_{1j}=c} k_1(x_j^{(1)},\theta)\mathrm{d}P_\theta(\theta)}{\int \prod\limits_{j:c_{1j}=c} k_1(x_j^{(1)},\theta)\mathrm{d}P_\theta(\theta)} & \text{if } c \in \mathcal{C}_1^{-i} \\ \alpha\Bigg(\sum\limits_{m\,\in\,\mathcal{M}_1^{-i}} C_m \dfrac{q_{1m}^{-i}}{q_1^{*-i}+\alpha_0} \dfrac{\int k_1(X_{1i},\theta) \prod\limits_{j:c_{1j}=m} k_1(x_j^{(1)},\theta)\mathrm{d}P_\theta(\theta)}{\int \prod\limits_{j:c_{1j}=m} k_1(x_j^{(1)},\theta)\mathrm{d}P_\theta(\theta)} + \\ \quad C_0 \dfrac{\alpha_0}{q_1^{*-i}+\alpha_0} \int k_1(X_{1i},\theta)\mathrm{d}P_\theta(\theta) \Bigg) & \text{otherwise} \end{cases}$$

where

1.

$$C_m = \sum_{s\,\in\,\mathcal{C}_{2|m}^{-i}} \frac{n_{2,s|m}^{-i}}{n_{1m}^{\star-i}+\alpha} \frac{\int k_2(X_{2i},\theta) \prod\limits_{j:c_{2j}=s} k_2(X_{2j},\xi)\mathrm{d}P_\xi(\xi)}{\int \prod\limits_{j:c_{2j}=s} k_2(X_{2j},\xi)\mathrm{d}P_\xi(\xi)} +$$

$$+ \frac{\alpha}{(n_{1m}^{\star-i}+\alpha)} \sum_{s\,\in\,\mathcal{M}_2^{-i}} \frac{q_{2s}^{-i}}{(q_2^{\star-i}+\alpha_0)} \frac{\int k_2(X_{2i},\theta) \prod\limits_{j:c_{2j}=s} k_2(X_{2j},\xi)\mathrm{d}P_\xi(\xi)}{\int \prod\limits_{j:c_{2j}=s} k_2(X_{2j},\xi)\mathrm{d}P_\xi(\xi)} +$$

$$+ \frac{\alpha}{(n_{1m}^{\star-i}+\alpha)} \frac{\alpha_0}{(q_2^{\star-i}+\alpha_0)} \int k_2(X_{2i},\theta)\mathrm{d}P_\xi(\xi)$$

with

$$n_{1m}^{\star-i} = \sum_{s\,\in\,\mathcal{C}_{2|m}^{-i}} n_{2,s|m}^{-i} \qquad q_2^{\star-i} = \sum_{s\,\in\,\mathcal{M}_2^{-i}} q_{2s}$$

note that $n_{1m}^{\star-i}$ is the number of subjects assigned to dish m at layer 1 (excluding subject i.

2.

$$C_0 = \sum_{s\,\in\,\mathcal{M}_2^{-i}} \frac{q_{2s}^{-i}}{(q_2^{\star-i}+\alpha_0)} \frac{\int k_2(X_{2i},\theta) \prod\limits_{j:c_{2j}=s} k_2(X_{2j},\xi)\mathrm{d}P_\xi(\xi)}{\int \prod\limits_{j:c_{2j}=s} k_2(X_{2j},\xi)\mathrm{d}P_\xi(\xi)} + \\ + \frac{\alpha_0}{(q_2^{\star-i}+\alpha_0)} \int k_2(X_{2i},\theta)\mathrm{d}P_\xi(\xi)$$

Then the dish allocation variable $c_{1,}$ is sampled from $p(c_{1i} \mid \boldsymbol{m}^{-i}, \boldsymbol{c}, X^{(1)}, \boldsymbol{X}_2)$. Notice that the full conditional is degenerate if at the previous step the customer has sat at an already occupied table, contrary, if $c_{1i} \notin \mathcal{C}_1^{-i}$,

$$p(c_{1i} = m \,|\boldsymbol{m}^{-i}, \boldsymbol{c}, X^{(1)}, \boldsymbol{X}_2) \\ \propto \begin{cases} C_m \, q_{1m}^{-i} \dfrac{\int k_1(X_{1i},\theta) \prod\limits_{j:c_{1j}=m} k_1(x_j^{(1)},\theta)\mathrm{d}P_\theta(\theta)}{\int \prod\limits_{j:c_{1j}=m} k_1(x_j^{(1)},\theta)\mathrm{d}P_\theta(\theta)} & \text{if } m \in \boldsymbol{c}_1^{-i} \\ C_0\, \alpha_0 \int k_1(X_{1i},\theta)\mathrm{d}P_\theta(\theta) & \text{otherwise} \end{cases}$$

The second layer is sampled following a classical marginal MCMC for the HDP (see Teh et al., 2006), which can be obtained from the two full conditionals above setting $C_m = 1$ for all m and $C_0 = 1$.

S4 Simulation studies

	Num. of	Num. of	Num. of	Num. of		
Scenario	items	layers	var. per layer	clusters	Adj. RI	Mispecified
n.1	200	2	1	2	1.000	No
n.2	200	2	1	2	0.010	No
n.3	200	10	1	2	0.809	No
n.4	200	100	1	2	0.921	No
n.5	200	2	2	3	0.914	Yes

Table S4.1: Simulation scenarios summaries: number of layers, layers' dimension (i.e., number of variables per each layer), number of clusters at each layer, adjusted Rand index between partitions at consecutive layers, whether the t-HDP estimated over the simulated data has a mispecified kernel or not.

Scenario n.1 and n.2

Simulating scenario description: see Section 4.1.

Model: t-HDP model with univariate Normal kernel with mean μ and variance equal to 1. Prior distribution for the mean is Normal centered in 0 and variance equal to 0.1. Concentration parameters are fixed to 0.1.

Algorithm: 100 000 iterations of the block partially collapsed conditional sampler in Algorithm 4 are performed and the first half is disregarded as burn-in. The chain is initialized to the k-means solutions computed independently for each layer.

Results: see Section 4.1.

Scenario n.3

Simulating scenario description: see Section 4.2.

Model: t-HDP model with univariate Normal kernel with mean μ and variance σ^2. Prior distribution for the mean is Normal centered in 0 and variance equal to $\sigma^2/0.1$. Concentration parameters are fixed to 0.1. The prior for the precision $1/\sigma^2$ is a Gamma distribution with shape and rate parameters equal to 0.1.

Algorithm: 100 000 iterations of the block partially collapsed conditional sampler in Algorithm 4 are performed and the first half is disregarded as burn-in. The chain is initialized to the k-means solutions computed independently for each layer.

Results: see Section 4.2.

Scenario n.4

Simulating scenario description: Data for 100 layers are simulated. At each layer there are two clusters and data are univariate. In particular, at layer 1 half of the dataset forms the first cluster, i.e., $c_{1i} = 1$ for $i = 1, \dots, 50$, and the other half the second cluster, i.e. $c_{1i} = 2$ for $i = 51, \dots, 100$. At layer 1, values are sampled from

$$X_{1i} \mid c_{1i} \overset{ind}{\sim} \mathcal{N}(0,1)\mathbb{1}(c_{1i} = 1) + \mathcal{N}(3,1)\mathbb{1}(c_{1i} = 2)$$

Then, from layer ℓ to layer $\ell + 1$, 2% of the observations are selected at random and moved to the cluster they were not assigned to.

Model: t-HDP model with univariate Normal kernel with mean μ and variance σ^2. The prior distribution for the mean is Normal centred in 0 and variance equal to $\sigma^2/0.1$. The prior for the precision $1/\sigma^2$ is a Gamma distribution with shape and rate parameters equal to 0.1. The concentration parameters of the t-HDP have prior Gamma with rate and shape parameters equal to 3.

Algorithm: 70 000 iterations of the block partially collapsed conditional sampler in Algorithm 4 are performed and the first 20 000 are disregarded as burn-in. The chain is initialized to the k-means solutions computed independently for each layer.

Results: see Section 4.2.

Scenario n.5

Simulating scenario description: Data for two layers are simulated. At each layer, there are three clusters and data are bi-variate. In particular, at layer 1 approximately one-third of the dataset forms the first cluster, i.e., $c_{1i} = 1$ for $i = 1, \dots, 66$, approximately one-third forms the second cluster, i.e. $c_{1i} = 2$ for $i = 67, \dots, 132$ and the remaining observations form a third cluster. At layer 1, bivariate values are sampled from bivariate student t distributions

$$X_{1i} \mid c_{1i} \overset{ind}{\sim} \mathcal{T}_2(\boldsymbol{\mu_1}, 1, \Sigma_1)\mathbb{1}(c_{1i} = 1) + \mathcal{T}_2(\boldsymbol{\mu_2}, 1, \Sigma_2)\mathbb{1}(c_{1i} = 2) + \mathcal{T}_2(\boldsymbol{\mu_3}, 1, \Sigma_3)\mathbb{1}(c_{1i} = 3)$$

where $\mathcal{T}_2(\boldsymbol{\mu}, \nu, \Sigma)$ denotes a bivariate t-Student distribution with ν degrees of freedom, centered in $\boldsymbol{\mu}$ and with scale matrix given by Σ.

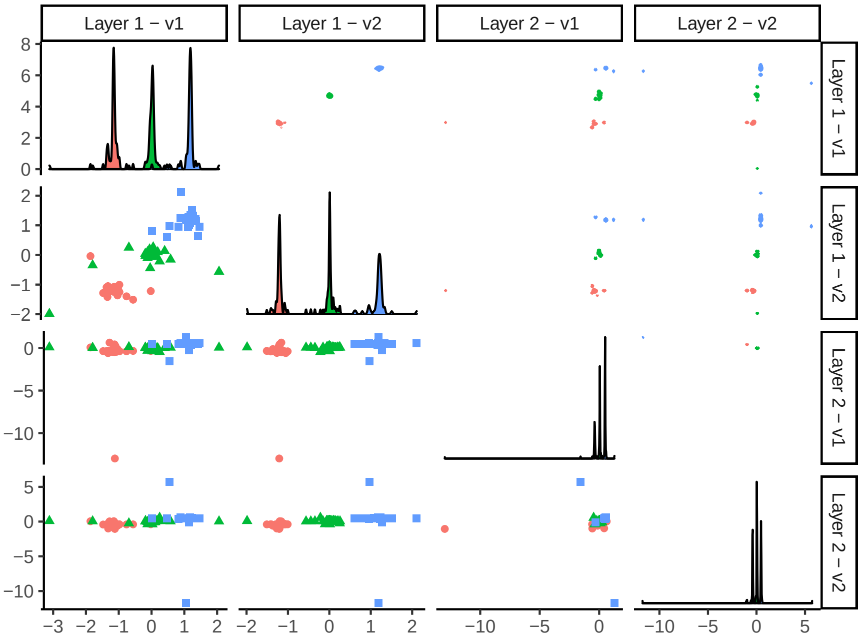


Figure S4.1: Simulation study: simulated data for Scenario n.5. Colours and shapes denote the true clustering at layer 1. The diagonal plots show the marginal distribution of each variable at each layer, colour coded according to the clustering allocation at layer 1. Upper and lower off-diagonal plots display the joint distribution of two pairs of variables, colour coded according to the clustering allocation at layer 1.

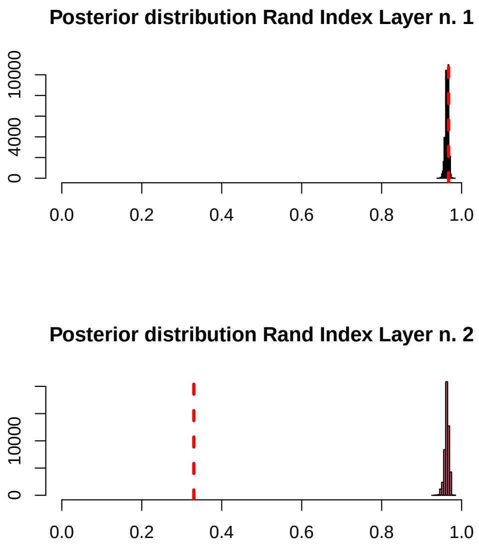


Figure S4.2: Simulation study: posterior distributions of Rand indexes between the posterior configurations and the truth for t-HDP model in Scenario n.5. Red dashed vertical lines denote the Rand indexes corresponding to the k-means' solution.

Then, from layer 1 to layer 2, 5% of the observations in the first two clusters are selected at random and moved to the cluster they were not assigned to, while the third cluster is kept constant. Bivariate values for the second layer are sampled from

$$\begin{aligned} X_{2i} \mid c_{2i} \overset{ind}{\sim} & \mathcal{T}_2(\boldsymbol{\mu_1}, 1, \Sigma_1)\mathbb{1}(c_{2i}=1) \\ & + \mathcal{T}_2(\boldsymbol{\mu_2}, 1, \Sigma_2)\mathbb{1}(c_{2i}=2) \\ & + \mathcal{T}_2(\boldsymbol{\mu_3}, 1, \Sigma_3)\mathbb{1}(c_{2i}=3) \end{aligned}$$

The true clusters' means are $\boldsymbol{\mu_1} = (0,0)^T$, $\boldsymbol{\mu_2} = (4,4)^T$, and $\boldsymbol{\mu_3} = (8,8)^T$.

Model: t-HDP model with univariate Normal kernel with mean $\boldsymbol{\mu}$ and diagonal variance and covariance matrix Σ^2. The prior distribution for the mean and variance and covariance matrix is a Normal-Inverse-Chi-Squared-distribution, in particular, for $j = 1, 2$, μ_j are a priori independent and Normal distributed with mean 0 and variance $\sigma_j^2/0.1$, while σ_j^2 are independently distributed accordingly to an inverse Chi-Squared with 1 degrees of freedom. The concentration parameters of the t-HDP have prior Gamma with rate and shape parameters equal to 3.

Algorithm: 100 000 iterations of the block partially collapsed conditional sampler in Algorithm 4 are performed and the first half is disregarded as burn-in. The chain is initialized to the k-means solutions computed independently for each layer for 10 clusters.

Results: The Rand index between the true configuration and the point estimates derived minimizing the variation of information loss function (Meilă, 2007; Wade and Ghahramani, 2018) are 0.97 and 0.96 for layer 1 and layer 2 respectively. The same values obtained with two independent k-means algorithms where the number of cluster is chosen based on the gap statistics (Tibshirani et al., 2001), are respectively 0.97 and 0.33. Figure S4.2 shows the distribution of the Rand index between the true clustering configuration and the configurations visited by the posterior algorithm of the t-HDP model after burnin.

S5 Application to metabolic concentrations in obese children: additional details and results

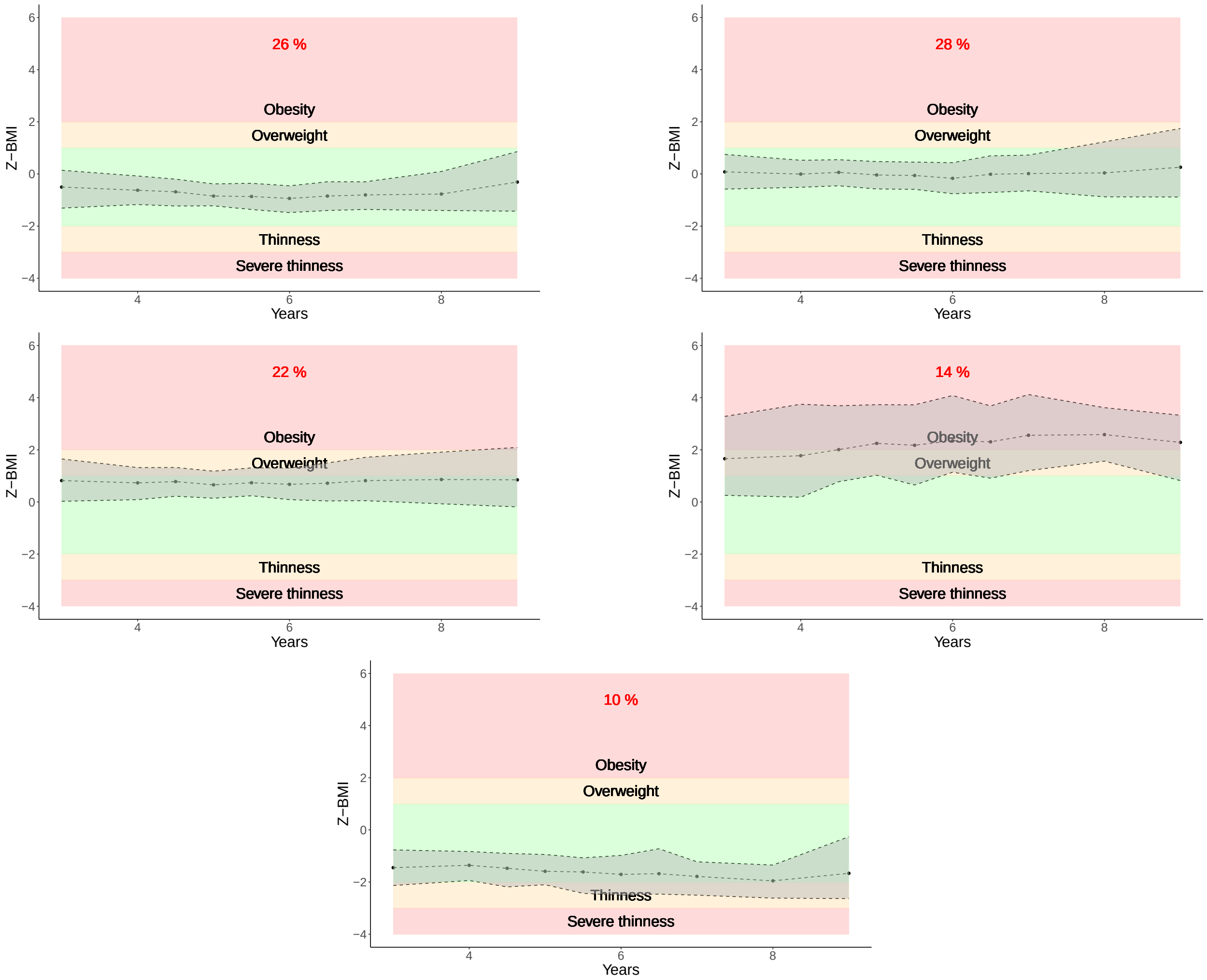


Figure S5.1: Average growth trajectories in the five estimated clusters at the z-BMI layer, shaded area include 95% of the observations assigned to the cluster, bands in the background corresponds to WHO classification of growth trajectories into Obesity, Overweight, Normal, Thinness, and Severe thinness. Percentages correspond to the proportions of children assigned to each of the five clusters.

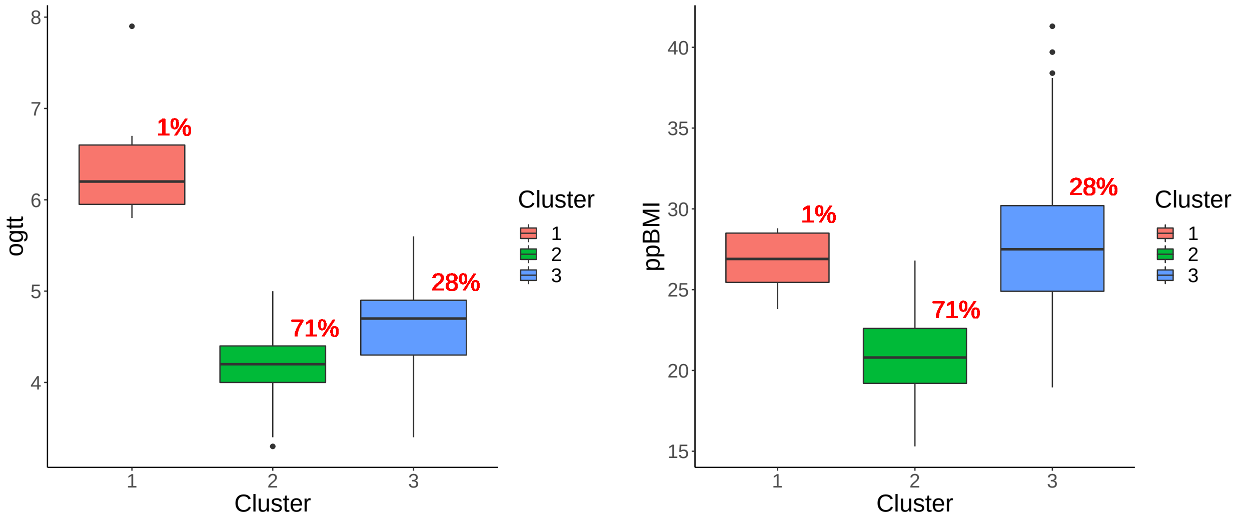


Figure S5.2: Boxplots by cluster assignment of the variables OGTT and PPBMI corresponding to the mother layer. Percentages correspond to the proportions of mothers assigned to each of the three clusters.

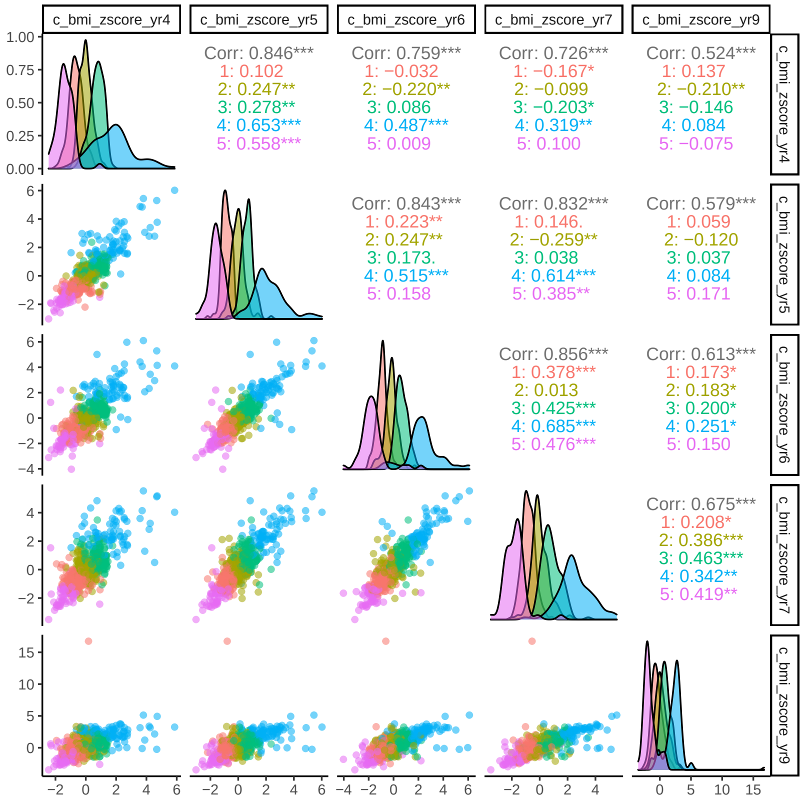


Figure S5.3: Scatter plots, density estimates and correlation values of z-BMI scores by cluster assignment at years 4, 5, 6, 7, and 9. Colours denote the cluster assignment of the children at the growth trajectory layer. The diagonal plots show the marginal distribution of the z-BMIs at each time point, colour coded according to the clustering allocation. Upper off-diagonal plots display the correlation between any two pairs of time points, overall and by cluster. Lower off-diagonal plots show the scatter plot of the data, colour coded according to the clustering allocation.

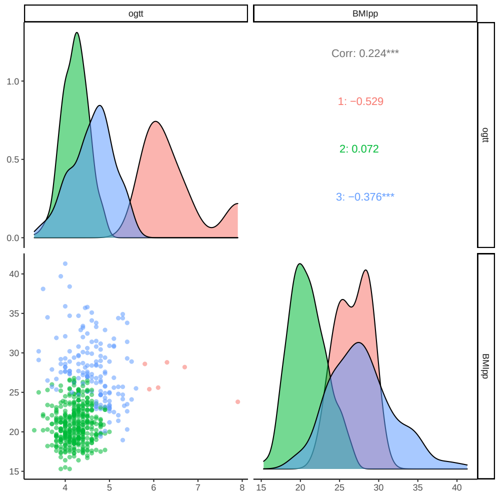


Figure S5.4: Scatter plots, density estimates and correlation values of the OGTT and PPBMI variables. Colours denote the cluster assignment at the mother layer. The diagonal plots show the marginal distribution of OGTT and PPBMI. Upper off-diagonal plots display the correlation between the two variables overall and by cluster. Lower off-diagonal plots show the scatter plot of the data.

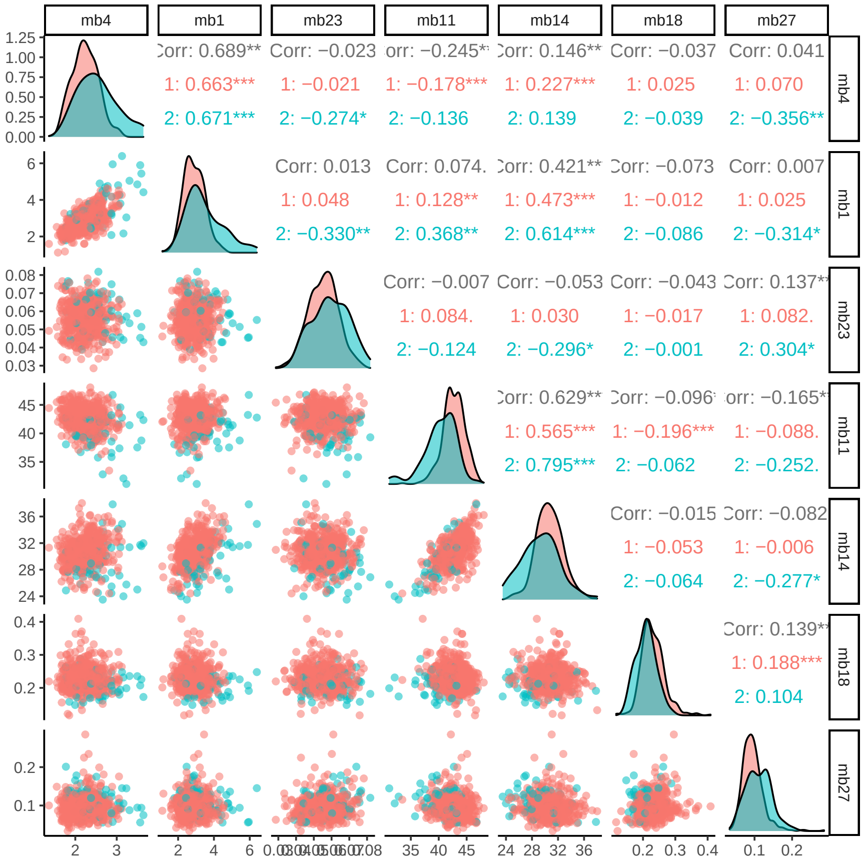


Figure S5.5: Scatter plots, density estimates and correlation values of seven randomly selected metabolites. Colours denote the cluster assignment at the metabolites layer. The diagonal plots show the marginal distributions. Upper off-diagonal plots display the correlation overall and by cluster. Lower off-diagonal plots show the scatter plot of the data.

	Average		IQR		Kruskal-Wallis
Metabolite	cluster 1	cluster 2	cluster 1	cluster 2	p-value
Clinical LDL Cholesterol	2.8918	3.4800	0.8383	1.3361	0.0000
HDL Cholesterol	1.6027	1.5545	0.3234	0.4035	0.0366
Triglycerides	0.7663	1.2641	0.3232	0.7728	0.0000
Phosphoglycerides	2.2372	2.5376	0.4324	0.7174	0.0000
Cholines Phosphoglycerides	2.5614	2.8604	0.4464	0.6645	0.0000
Sphingomyelins	0.5001	0.5468	0.0936	0.1424	0.0014
APO A1	1.4819	1.4974	0.2747	0.3638	0.7219
APO B	0.8092	0.9997	0.2181	0.3844	0.0000
Omega 3	0.4182	0.4702	0.1377	0.1862	0.0077
Omega 6	4.2964	4.7016	0.5689	0.8934	0.0000
Poly-Unsaturated FA (PUFA)	42.7171	40.2715	2.5448	3.4999	0.0000
Mono-Unsaturated FA (MUFA)	23.2455	24.8581	1.8430	2.8237	0.0000
Saturated FA (SFA)	34.0375	34.8704	1.0974	1.5096	0.0000
Linoleic acid	30.7019	29.3047	2.7123	4.0315	0.0001
Docosahexaenoic acid (DHA)	2.1145	1.9018	0.5387	0.4947	0.0005
Alanine	0.3079	0.3502	0.0901	0.1114	0.0000
Glutamine	0.5775	0.5375	0.1343	0.1423	0.0026
Glycine	0.2310	0.2091	0.0436	0.0412	0.0000
Histidine	0.0877	0.0878	0.0128	0.0136	0.8621
Isoleucine	0.0512	0.0650	0.0126	0.0126	0.0000
Leucine	0.1020	0.1230	0.0203	0.0254	0.0000
Valine	0.2311	0.2722	0.0422	0.0395	0.0000
Phenylalanine	0.0553	0.0594	0.0114	0.0140	0.0032
Tyrosine	0.0686	0.0802	0.0139	0.0224	0.0000
Glucose	4.8412	4.9235	0.5471	0.4521	0.0219
Lactate	2.0602	2.5197	0.7823	0.8316	0.0000
Pyruvate	0.0952	0.1108	0.0314	0.0461	0.0001
Citrate	0.1059	0.1034	0.0176	0.0198	0.1641
beta-Hydroxybutyric acid	0.1237	0.2036	0.1257	0.1227	0.4402
Acetate	0.0357	0.0294	0.0156	0.0101	0.0000
Acetoacetate	0.0434	0.0709	0.0406	0.0468	0.5898
Acetone	0.0183	0.0258	0.009	0.0091	0.2576
Creatinine	45.3778	47.9555	9.1402	12.7015	0.0318
Albumin	42.5924	43.4704	3.7106	4.6384	0.3151
Glycoprotein acetyls	0.8306	0.9525	0.1344	0.2099	0.0000

Table S5.1: Summary of metabolite clusters: average concentration of each metabolite by cluster, interquartile range (IQR) of the metabolites concentration distribution by cluster and p-value of the Kruskal-Wallis test for difference in distribution between the two clusters.

S6 Computational cost and mixing performance of posterior algorithms

Simulation study n.1: n=200, P=2, L=2				
	ESS / N	ESS / N	time for 1	time for 1000
	layer n.1	layer n.2	iteration	effective draws
	0.03830	0.02410	0.0132 sec	9.12 min

Simulation study n.2: n=200, P=2, L=2				
	ESS / N	ESS / N	time for 1	time for 1000
	layer n.1	layer n.2	iteration	effective draws
	0.05762	0.04580	0.0134 sec	4.88 min

Simulation study n.3: n=200, P=10, L=10				
ESS / N	ESS / N	ESS / N	time for 1	time for 1000
layer n.1	layer n.5	layer n.10	iteration	effective draws
0.09904	0.05823	0.03056	0.072 sec	39.26 min

Simulation study n.4: n=200, P=100, L=100				
ESS / N	ESS / N	ESS / N	time for 1	time for 1000
layer n.1	layer n.50	layer n.100	iteration	effective draws
0.0254	0.02035	0.0872	1.028 sec	841 min

Table S6.1: Effective sample size per iteration (ESS/N) after burn-in for the Rand index between chain and truth, time in seconds per iteration, and time in minutes for 1000 effective draws. The latter is computed as the maximum of the value $(time) \times 1000/(ESS/N)$ across layers. n denotes the sample size, P is the total number of considered variables, and L is the total number of layers to which the variables are assigned. Algorithms are coded in R and run on Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz CPU.

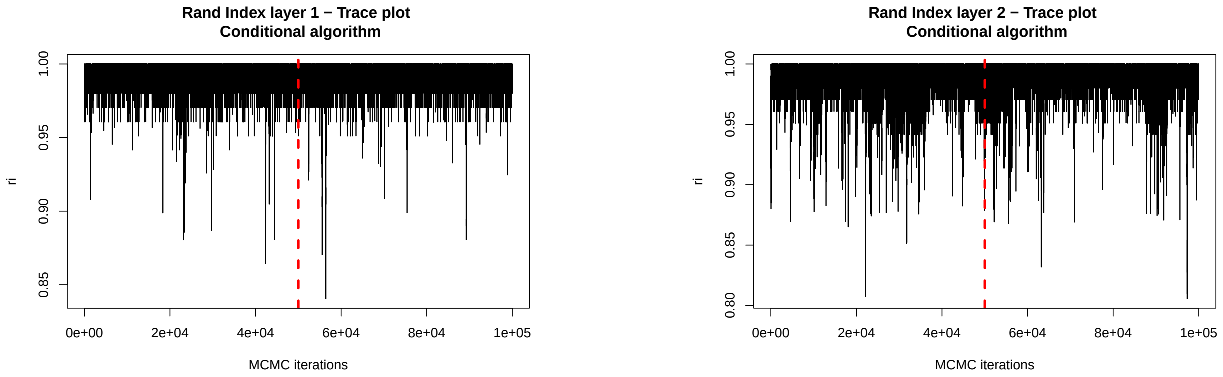


Figure S6.1: Simulation study: Scenario n.1. Trace plots of the Rand index between the chain configuration and the true configuration. Vertical dashed lines locate the burn-in period.

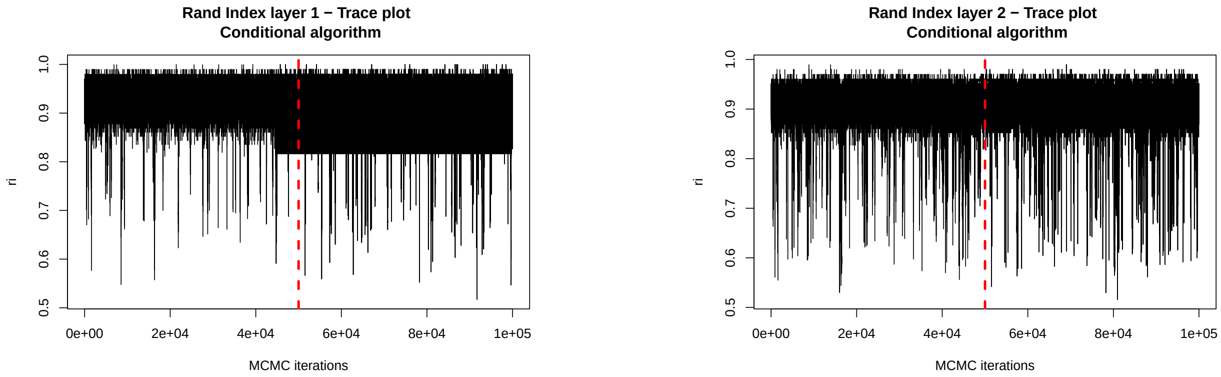


Figure S6.2: Simulation study: Scenario n.2. Trace plots of the Rand index between the chain configuration and the true configuration. Vertical dashed lines locate the burn-in period.