Journal of Machine Learning Research 25 (2024) 1-56 Submitted 8/23; Revised 9/24; Published 10/24

Graphical Dirichlet Process for Clustering Non-Exchangeable Grouped Data

Arhit Chakrabarti ARHIT.CHAKRABARTI@STAT.TAMU.EDU
Department of Statistics
Texas A&M University
College Station, TX 77843-3143, USA

Yang Ni YNI@STAT.TAMU.EDU
Department of Statistics
CPRIT Single Cell Data Science Core
Texas A&M University
College Station, TX 77843-3143, USA

Ellen Ruth A. Morris ELLENRUTH@TAMU.EDU
Department of Nutrition
Program in Integrative Nutrition & Complex Diseases
Current address: Texas A&M Veterinary Medical Diagnostic Laboratory
Texas A&M University
College Station, TX 77843-4471, USA

Michael L. Salinas MLSALINAS4@TAMU.EDU
Department of Nutrition
Program in Integrative Nutrition & Complex Diseases
CPRIT Single Cell Data Science Core
Texas A&M University
College Station, TX 77843-2253, USA

Robert S. Chapkin ROBERT.CHAPKIN@AG.TAMU.EDU
Department of Nutrition
Program in Integrative Nutrition & Complex Diseases
CPRIT Single Cell Data Science Core
Texas A&M University
College Station, TX 77843-2253, USA

Bani K. Mallick BMALLICK@STAT.TAMU.EDU
Department of Statistics
Texas A&M University
College Station, TX 77843-3143, USA

Editor: Mingyuan Zhou

Abstract

We consider the problem of clustering grouped data with possibly non-exchangeable groups whose dependencies can be characterized by a known directed acyclic graph. To allow the sharing of clusters among the non-exchangeable groups, we propose a Bayesian nonparametric approach, termed graphical Dirichlet process, that jointly models the dependent group-specific random measures by assuming each random measure to be

distributed as a Dirichlet process whose concentration parameter and base probability measure depend on those of its parent groups. The resulting joint stochastic process respects the Markov property of the directed acyclic graph that links the groups. We characterize the graphical Dirichlet process using a novel hypergraph representation as well as the stick-breaking representation, the restaurant-type representation, and the representation as a limit of a finite mixture model. We develop an efficient posterior inference algorithm and illustrate our model with simulations and a real grouped single-cell data set.

Keywords: Bayesian nonparametrics, clustering, directed acyclic graph, family-owned restaurant process, non-exchangeable groups

1. Introduction

This article considers clustering of grouped data where the groups are *non-exchangeable*. We are interested in settings where the data are *partially exchangeable* (de Finetti, 1938), which entails the exchangeability of the observations within each group but not across the groups, (see Kallenberg, 2005 for an extensive bibliography). We consider dependent group–specific random probability measures, thereby allowing the borrowing of information across non-exchangeable groups. We represent the dependencies among groups through a known *directed acyclic graph* (DAG) with nodes denoting groups and directed edges denoting the group dependencies. Such data are abundant in many areas such as genomics. For example, our motivating application is a single-cell RNA-sequencing (scRNA-seq) study that aimed to investigate intestinal stem cell differentiation processes in mice with colorectal cancer. The experiments started from a baseline group where the mice were genetically wild-type, fed with a normal diet, and treated with no cancer therapy (placebo). Then to understand the main effects of genotype, diet, and cancer therapy on colonic crypt and tumor niche cell composition, the experimenters introduced three new groups of mice, each differing from the baseline group by exactly one factor (Apc knock-out, a high-fat diet, or a new cancer treatment AdipoRon). To determine the two-way interaction effects, three additional groups of mice were studied, each of which differed from the baseline group by two factors (e.g., mice with Apc knock-out, a high-fat diet, and no cancer treatment). Lastly, for a three-way interaction, they introduced the eighth group of mice with Apc knock-out, a high-fat diet, and the new treatment AdipoRon. The progression of these experiments from baseline to the study of main effects, two-way interactions, and three-way interactions manifests the non-exchangeability of the experimental groups (e.g., the baseline group is expected to be more similar to the "main effect" groups than the "three-way interaction" group). With this grouped scRNA-seq data set, our goal is to cluster cells based on gene expression at the single-cell level within each experimental group while allowing information to be shared across these non-exchangeable groups with a novel DAG-based Bayesian nonparametric model.

Our proposed model extends beyond the specific motivating problem previously discussed. Grouped data can emerge across various disciplines, where the groups exhibit inherent non-exchangeability. Furthermore, the dependencies among the groups can be naturally represented through a known DAG. The first example is time-series data. One might be interested in clustering stocks based on daily prices for each year. Each calendar

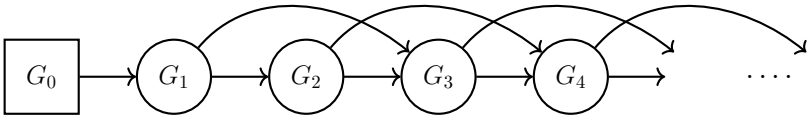


Figure 1: DAG for AR2.

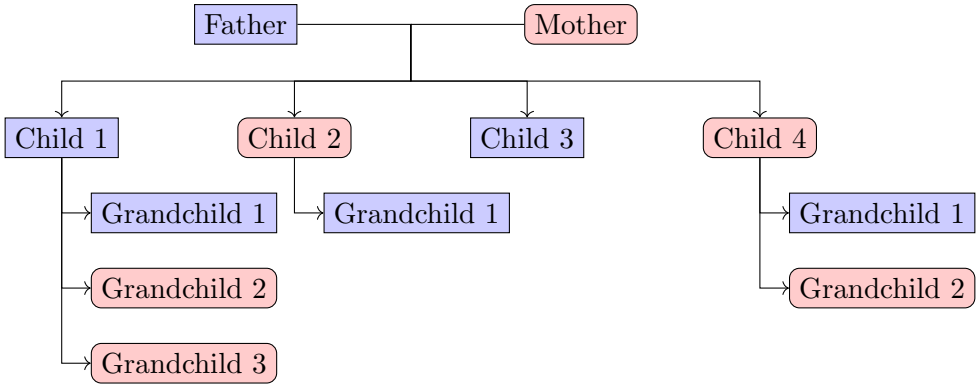


Figure 2: Family tree with three generations. Males are depicted with the color blue and females with red.

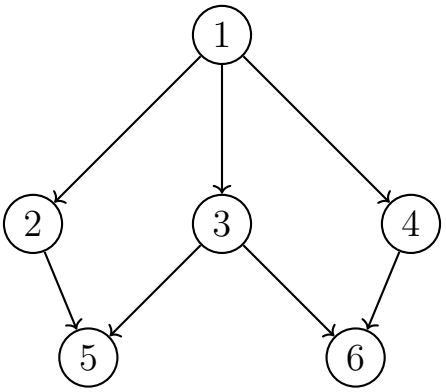
year is then a group. The groups naturally have time dependence (i.e., one does not expect the clustering of stocks to change dramatically in consecutive years), which may be represented by an autoregressive (AR) model. AR model is one type of DAG model; see Figure 1 for the underlying DAG of AR(2). The second example arises in family tree (Figure 2, another type of DAG). For example, it may be of interest to study the evolution of the gene expressions of a family of three generations to understand how the expression patterns change with the generations. Clustering the gene expression of the family members, where the dependencies between the members (each family member constitutes a group) are naturally explained by the underlying tree, may provide valuable information to the understanding of phenotypic features and/or disease progression (e.g., hemophilia, cancer, etc.).

The *Dirichlet process* (DP, Ferguson, 1973) and its variations (De Blasi et al., 2013; Barrios et al., 2013) have been the backbone of numerous model-based Bayesian nonparametric clustering methods (Hjort et al., 2010; Müller et al., 2015). The DP, $DP(\alpha_0, G_0)$, is a probability measure on probability measures, where $\alpha_0 > 0$ is the concentration parameter and G_0 is a base probability measure. There have been extensive studies on DP mixture models (Antoniak, 1974; Escobar and West, 1995; MacEachern and Müller, 1998), which enable clustering without having to fix the number of clusters *a priori*. When there are groups present in the data, naively, one could consider either a separate DP mixture model for each group on one extreme or a single DP mixture model ignoring the groups on the other extreme. However, it is often desirable to identify group-specific clusters while allowing the groups to be linked so that clusters are comparable across groups. Given the goal of clustering the observations within each

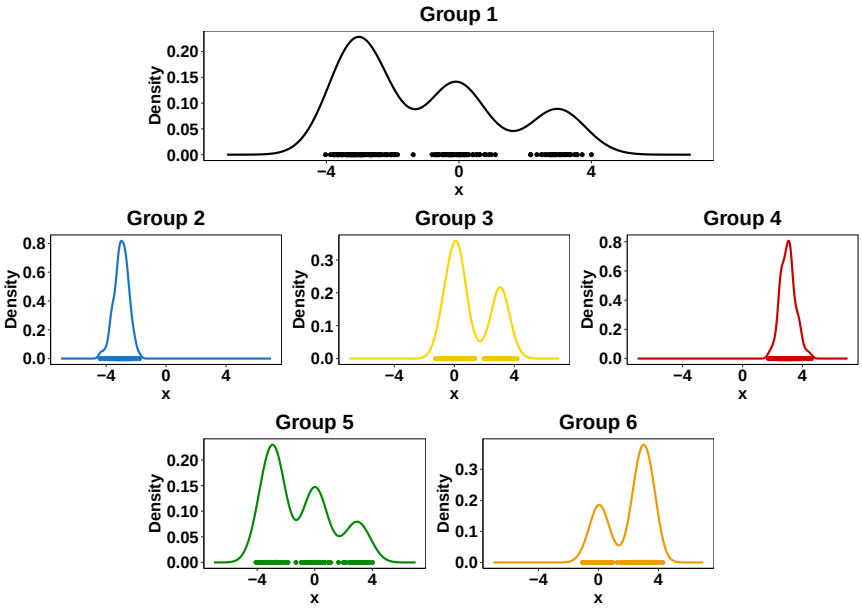
group, consider a set of random probability measures, G_j, one for each group j, where each G_j is distributed as $DP(\alpha_{0j}, G_{0j})$ with group-specific concentration parameter α_{0j} and base probability measure G_{0j}. Many methods have been proposed to link these group-specific DPs to induce dependencies through the parameter α_{0j} and/or G_{0j} (Cifarelli and Regazzini, 1978; Mallick and Walker, 1997; Kleinman and Ibrahim, 1998; Müller et al., 2004). Perhaps one of the most well-known methods is the hierarchical Dirichlet process (HDP, Teh et al., 2006), which falls in the general framework of dependent DP (MacEachern, 1999, 2000) and assumes each group-specific G_j is distributed as $DP(\alpha_0, G_0)$ where α_0 is the shared concentration parameter and G_0 is the shared base probability measure for all groups. They further assume that G_0 follows another DP, $G_0 \sim DP(\gamma, H)$. Since draws from a DP are discrete with probability one (Sethuraman, 1994), the base measure G_0 is almost surely discrete, which ensures that the group-specific probability measure G_j shares the same set of atoms. The corresponding HDP mixture model is thus capable of identifying group-specific clusters while borrowing strength across groups. By construction, HDP mixture model assumes that both the observations within each group and the groups are exchangeable. Recently, several authors have proposed hierarchies of discrete probability measures extending beyond the hierarchical Dirichlet process (Teh, 2006; Thibaux and Jordan, 2007; Zhou, 2016; Tillinghast et al., 2022). See Teh and Jordan, 2010; Foti and Williamson, 2013 for a summary of non-exchangeable priors for Bayesian nonparametric models. Furthermore, Camerlenghi et al., 2019b provides a distribution theory for the entire class of hierarchical processes. A similar approach with a different scope, the nested DP (Rodríguez et al., 2008), assumes G_j follows a DP-distributed random probability measure with another DP as the base measure, $G_j \sim Q$ and $Q \sim DP(\alpha_0, DP(\gamma, H))$. The nested structure allows for the clustering of groups but restricts the clusters of observations within each group to be either identical or completely unrelated across groups. Models based on the nested DP have been widely employed in various contexts (Rodriguez and Dunson, 2014; Graziani et al., 2015; Zuanetti et al., 2018). However, similarly to HDP, nested DP also assumes both the observations within each group and the groups to be exchangeable. Moreover, the nested DP is known to suffer from a degeneracy property (Camerlenghi et al., 2019a)–two distributions sharing even one atom in their support are automatically assigned to the same cluster. Several recent works (Beraha et al., 2021; Lijoi et al., 2022; Bi and Ji, 2023) have been proposed to take advantage of the cluster-sharing feature of the HDP and the group-clustering feature of the nested DP. In contrast to methods relying on the HDP or its variants, some other works rely on models with additive structure or common atoms (Camerlenghi et al., 2019a; Chandra et al., 2023; Denti et al., 2023; D'Angelo et al., 2023; D'Angelo and Denti, 2024). Dependent DP has also been extensively used to model random distributions with various other types of dependencies such as spatial and temporal dependencies (Iorio et al., 2004; De Iorio et al., 2009; Dunson and Herring, 2005; Gelfand et al., 2005; Griffin and Steel, 2006; Nieto-Barajas and Contreras-Cristán, 2014; Dahl et al., 2017); see Quintana et al., 2020 for a recent review of different dependent DPs.

In this paper, we are interested in modeling a set of group-specific random distributions of which the (conditional) dependencies can be characterized by a DAG whose nodes represent the groups. More precisely, we assume that the joint distribution of

the set of group-specific random distributions factorizes with respect to a DAG and, therefore, respects its Markov property (i.e., conditional independencies). We call such graph-dependent DP, the *graphical Dirichlet process* (GDP). Using GDP as a mixing distribution, the GDP mixture model gives rise to group-specific clusters, which depend directly on their Markov blanket. As an illustration, for a grouped data with six groups, whose dependencies are shown by the underlying DAG in Figure 3a, the corresponding group-specific distributions are shown in Figure 3b. Clearly, the distributions corresponding to the different groups are similar to their parents. In particular, the distribution of group 6 resembles its parents (group 3 and 4) and is different from the group 5 even though they share a parent (group 3). However, all groups share some or all of the components of the ancestor group (group 1), highlighting the sharing feature of the proposed GDP.



(a) The DAG denoting dependency between the groups.



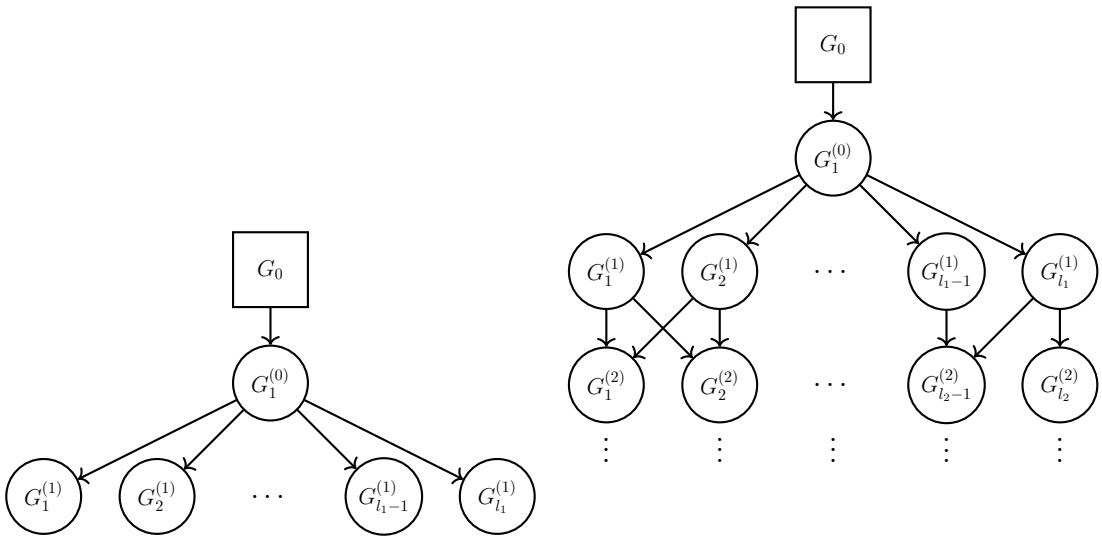
(b) The underlying distributions for the groups.

Figure 3: Sharing of features by GDP.

The known flexibility of DAG in representing conditional dependencies renders the generality of the proposed GDP for modeling dependent random distributions and group-specific clusters beyond exchangeable groups. The use of DAGs in Bayesian nonparametrics has been considered in recent literature. Dey et al., 2022 proposed a graphical Gaussian process to parsimoniously model multivariate spatial data by incorporating conditional independencies among variables encoded by a DAG. Gu and Dunson, 2023 proposes a pyramid-shaped deep latent variable model for categorical data using a DAG to represent the layer-wise latent conditional dependency structure. These works showcased the usefulness of DAGs through their factorization in Bayesian nonparametric models. We also exploit such factorization in this paper but our model is significantly different from theirs in both approaches and scopes. For example, their graphs link variables whereas ours link groups, and they focus on the modeling of multivariate spatial fields or generative models for categorical data whereas we focus on clustering non-exchangeable grouped data. The proposed GDP is a general model. The well-known HDP is a special case of GDP with a specific type of DAG–a fork, i.e., one parent node and many children nodes (detailed in Section 3.1); see Figure 4a. Several existing works on time-evolving topic models can also be reformulated using a DAG to capture the time-dependency structure (Srebro and Roweis, 2005; Ren et al., 2008; Zhang et al., 2010). Furthermore, the *tree-structured HDP* (Figure 2) considered by Alam et al., 2019 is a special case of our proposed GDP.

In this paper, we will characterize the proposed GDP by a novel *hypergraph* representation, which uses the fact that Dirichlet distribution/process is a normalized gamma distribution/process. We will also provide several other representations analogous to those for the HDP, i.e., a stick-breaking representation, a restaurant-type representation, and a representation as an infinite limit of a finite mixture model. We develop efficient posterior sampling based on the SALTSampler (Director et al., 2017) and a Blocked Gibbs sampler for DP/HDP (Ishwaran and James, 2001; Das et al., 2024). Simulations and the motivating grouped single-cell data are used to demonstrate our method. In summary, our main contribution is three-fold. We propose a general Bayesian nonparametric approach, GDP, to incorporate non-exchangeable group dependencies for clustering. Second, we provide several characterizations of GDP, each providing a different perspective. Furthermore, we develop a Metropolis-within-blocked-Gibbs sampler for posterior inference. Since HDP is a special case of GDP, this also contributes to a new sampler for HDP. The difficulty of sampling the global weights for HDP is mitigated by using the specialized proposal of SALTSampler (Director et al., 2017).

The remainder of the paper is organized as follows. Section 2 provides a brief overview of some preliminaries needed for the remainder of the paper. Section 3 introduces the proposed GDP and the corresponding nonparametric mixture model. We introduce the hyperpriors of our model and also present two lemmas, which are the backbone of our main result in Theorem 3. In Section 4, we present different representations of the proposed GDP. In Section 5, we provide simulations to illustrate our method. Section 6 presents a real data analysis using the proposed method on the motivating single-cell data. The paper concludes with a brief discussion in Section 7. The source codes used for the analysis, including those for simulations and real data, can be found in the repository `https://github.com/Arhit-Chakrabarti/GDPSamp`.



(a) HDP (b) GDP for DAG with a unique root node.

Figure 4: Schematic illustration of HDP and GDP. HDP is a special case of GDP when the DAG is a fork.

2. Preliminaries

2.1 Directed Acyclic Graph

We first provide a brief background on DAG. Let $D = (V, E)$ be a DAG consisting of a set of nodes $V = \{1, 2, \ldots, p\}$ and a set of directed edges $E \subset V \times V$ that does not contain any directed cycles. We denote a directed edge from the node i to node j by $j \leftarrow i$ and call i a *parent* of j. A node without parents is called a *root*. For a DAG, there exists at least one root. Let $\boldsymbol{Y} = \{Y_1, \ldots, Y_p\}$ be a set of random variables. Every node $j \in V$ represents a random variable Y_j; later in this paper, Y_j will be a random probability measure. In a DAG model, also known as a Bayesian network, the probability distribution $\mathcal{P}(\boldsymbol{Y})$ is assumed to factorize over D, $\mathcal{P}(\boldsymbol{Y}) = \prod_{j=1}^{p} \mathcal{P}(Y_j \mid Y_{pa(j)})$, where $pa(j) = \{k \in V | j \leftarrow k\}$ denotes the collection of parents of node j. This DAG factorization implies that the distribution $\mathcal{P}$ respects the conditional independence relationships encoded by the graph D via the notion of d-separation (Pearl, 2009); and vice versa. For instance, any node is conditional independent of its non-descendants given its parents, i.e., $Y_j \perp Y_{nd(j)} | Y_{pa(j)}$ for any $j \in V$ where $\perp$ denotes independence, $nd(j) = V \backslash de(j) \backslash \{j\}$ denotes the non-descendants of node j, and $de(j) = \{k \in V | k \leftarrow \cdots \leftarrow j\}$ denotes the descendants of node j. A *Markov blanket* of any node j from V is any subset V_1 of V such that $Y_j \perp Y_{V \backslash V_1} | Y_{V_1}$. In other words, V_1 contains all the information in V about the node j. DAG models are convenient tools to parsimoniously specify a multivariate distribution through its conditionals, which is especially useful in this paper for specifying a multivariate distribution of a set of random probability measures.

2.2 Infinite Mixture Model

Next, we present a brief overview of infinite mixture models for a single population, the DP mixture model, and for multiple exchangeable populations, the HDP mixture model.

2.2.1 Dirichlet Process Mixture Model

For a single population, let x_i denote the ith realization of a random variable X. We consider a mixture model,

$$\begin{aligned} \theta_i \mid G &\overset{iid}{\sim} G, \\ x_i \mid \theta_i &\overset{ind}{\sim} F(\theta_i), \end{aligned} \tag{1}$$

where $F(\theta_i)$ denotes the distribution of x_i parameterized by θ_i. The parameters θ_i's are conditionally independent given the prior distribution G. In a DP mixture model, G is assigned a DP prior, $G \sim DP(\alpha_0, G_0)$ with concentration α_0 and base probability measure G_0.

Sethuraman, 1994 presented the *stick-breaking representation* of the DP based on independent sequences of i.i.d. random variables $(\pi'_k)_{k=1}^{\infty}$ and $(\phi_k)_{k=1}^{\infty}$, which is given by,

$$\pi'_k \mid \alpha_0 \overset{iid}{\sim} Beta(1, \alpha_0), \qquad \phi_k \mid G_0 \overset{iid}{\sim} G_0, \tag{2}$$

$$\pi_k = \pi'_k \prod_{l=1}^{k-1}(1 - \pi'_l), \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}, \tag{3}$$

where δ_ϕ is a point mass at ϕ and ϕ_k's are called the *atoms* of G. The sequence of random weights $\boldsymbol{\pi} = (\pi_k)_{k=1}^{\infty}$ constructed from Eq. (2) and Eq. (3) satisfies $\sum_{k=1}^{\infty} \pi_k = 1$ with probability one. The random probability measure on the set of integers is denoted by $\boldsymbol{\pi} \sim \text{GEM}(\alpha_0)$ for convenience where GEM stands for Griffiths, Engen and McCloskey (Pitman, 2002). It is clear from Equation (1) and Equation (3) that θ_i takes the value ϕ_k with probability π_k. Let z_i be a categorical variable such that $z_i = k$ if $\theta_i = \phi_k$. An equivalent representation of a Dirichlet process mixture is given by,

$$\begin{aligned} \boldsymbol{\pi} \mid \alpha_0 &\sim \text{GEM}(\alpha_0), & z_i \mid \boldsymbol{\pi} &\overset{iid}{\sim} \boldsymbol{\pi}, \\ \phi_k \mid G_0 &\overset{iid}{\sim} G_0, & x_i \mid z_i, (\phi_k)_{k=1}^{\infty} &\overset{ind}{\sim} F(\phi_{z_i}). \end{aligned} \tag{4}$$

2.2.2 Hierarchical Dirichlet Process Mixture Model

Suppose observations are now organized into multiple exchangeable groups. Let x_{ji} denote the observation i from group j and θ_{ji} denote the parameter specifying the mixture component associated with the corresponding observation. Let $F(\theta_{ji})$ denote the distribution of x_{ji} given θ_{ji} and G_j denote a prior distribution for θ_{ji}. The group-specific mixture model is given by,

$$\begin{aligned} \theta_{ji} \mid G_j &\overset{ind}{\sim} G_j, \\ x_{ji} \mid \theta_{ji} &\overset{ind}{\sim} F(\theta_{ji}). \end{aligned} \tag{5}$$

As with the DP mixture model, when the random measures G_j's are assigned an HDP prior,

$$\begin{aligned} G_0 \mid \gamma, H &\sim DP(\gamma, H), \\ G_j \mid \alpha_0, G_0 &\sim DP(\alpha_0, G_0), \end{aligned} \tag{6}$$

the corresponding mixture model is referred to as the HDP mixture model. The global random probability measure G_0 is distributed as a DP with concentration parameter γ and base probability measure H. The group-specific random measures G_j's are conditionally independent given G_0 and hence are exchangeable (de Finetti, 1938). They are distributed as DP with the base measure G_0 and some concentration parameter α_0. The probability model (5) along with (6) completes the specification of an HDP mixture model. Because DP-distributed G_0 is almost surely discrete, the atoms of G_j's and hence the group-specific clusters are necessarily shared across groups.

3. Graphical Dirichlet Process

When groups are non-exchangeable (e.g., due to study design), the joint distribution of G_j's specified by (6) may not be appropriate. Our approach to the problem of sharing clusters among non-exchangeable groups is through specifying a general joint distribution of G_j's that respect the Markov property of a DAG D that links the groups. We assume that the underlying DAG D is known and we define the appropriate prior on the nodes of the DAG and refer to the resulting stochastic process on the graph as the graphical Dirichlet process (GDP). We show how this prior can be used in the non-exchangeable grouped mixture model setting.

3.1 The Proposed GDP

Let the nodes V of DAG $D = (V, E)$ now represent the group-specific random probability measures G_j's. The edges E represent the conditional dependence of G_j's. Then the joint distribution of the random probability measures follows the DAG factorization $\mathcal{P}(G_1, \ldots, G_p|D) = \prod_{j=1}^{p} \mathcal{P}(G_j \mid G_{pa(j)})$, where $G_{pa(j)}$ is the set of random probability measures indexed by the parents $pa(j)$ of node j. For convenience, we assume D has a unique root; see Figure 4b. This assumption does not diminish the generality of our approach as a DAG with multiple roots can always be converted, without losing any conditional dependencies, to a DAG with a unique root by simply augmenting the DAG with a hidden common parent of the roots; that hidden common parent becomes the unique root of the new DAG (Figure 5). The augmentation only changes the Markov blanket of the original root nodes. Specifically, the Markov blanket of any original root node is simply augmented with the hidden parent node. As the Markov blanket of any other node remains unchanged, the distributions of all other nodes remain the same, and hence this augmentation does not alter the conditional dependencies of the original DAG.

Let us introduce a few terms before describing the proposed GDP. We denote the root node, which may be hidden, as the *layer 0* of DAG D. The child nodes of the root node are termed as the *layer-1* nodes, and we assume that there are l_1 of them. Similarly, we assume that there are a total of l_2 child nodes from the layer-1 nodes, which we refer to as the *layer-2* nodes. We assume that there are K layers in the given DAG D and at any layer k, there are l_k nodes. The total number of non-root nodes is $\sum_{k=1}^{K} l_k = p$. We define the concentration parameters and random measures of node j in the layer k of DAG D as $\alpha_j^{(k)}$ and $G_j^{(k)}$, $\quad j = 1, \ldots, l_k$. We denote by $an^{(k,l)}(j)$ the collection of generation-l ancestors of node j in layer k of the DAG. For example, $an^{(k,1)}(j)$ denotes the parents (generation-1

ancestors) of the node j in layer k, and $an^{(k,2)}(j)$ denotes the collection of the parents of the nodes in $an^{(k,1)}(j)$ or in other words, $an^{(k,2)}(j)$ denotes the collection of "grand-parents" (generation-2 ancestors) of node j in layer k of the DAG.

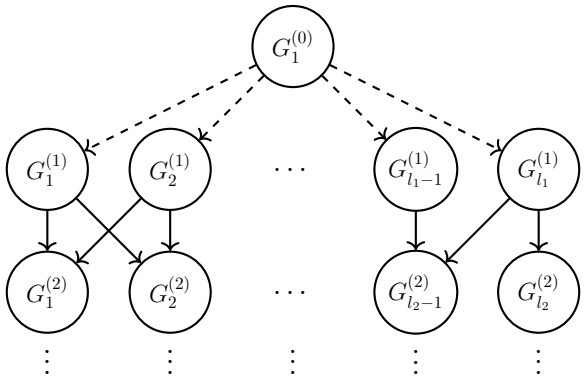


Figure 5: DAG augmented with a hidden root $G_1^{(0)}$, indicated by the dashed arrows. The original root nodes are $G_1^{(1)}, \ldots, G_{l_1}^{(1)}$.

We define GDP recursively from layer 0, the root node,

$$G_1^{(0)} \mid \alpha_1^{(0)}, G_0 \sim DP\left(\alpha_1^{(0)}, G_0\right), \tag{7}$$

where G_0 is a fixed base probability measure. Then the distribution of the random probability measure of node j in layer k of DAG D conditional on the concentration parameters and random probability measures of its parent nodes is given by,

$$G_j^{(k)} \mid \alpha_j^{(k)}, \{G_l^{(k-1)} : l \in an^{(k,1)}(j)\} \sim DP\left(\alpha_j^{(k)}, \sum_{l \in an^{(k,1)}(j)} \pi_{jl}^{(k)} G_l^{(k-1)}\right), \tag{8}$$

for $j = 1, 2, \ldots, l_k$. In other words, node j in layer k of the DAG is distributed according to a DP with its own concentration parameter $\alpha_j^{(k)}$ and its base distribution being a weighted average of the random probability measures of its parents in layer $k-1$ of the DAG, $\{G_l^{(k-1)} : l \in an^{(k,1)}(j)\}$, where the weights are given by $\{\pi_{jl}^{(k)} : l \in an^{(k,1)}(j)\}$, which have a unit sum $\sum_{l \in an^{(k,1)}(j)} \pi_{jl}^{(k)} = 1$. Moreover, from the Markov properties of DAG D, $G_{j_1}^{(k)}$ and $G_{j_2}^{(k)}$ are conditionally independent given their parents, $\{G_l^{(k-1)} : l \in an^{(k,1)}(j_1)\}$ and/or $\{G_l^{(k-1)} : l \in an^{(k,1)}(j_2)\}$, and $G_j^{(k)}$ is conditionally independent of all other random probability measures given its Markov blanket.

We remark that HDP is a special case of the proposed GDP with a specific DAG, fork-DAG (Figure 4a). Using the notations introduced, a fork-DAG is a DAG with a unique

root node and only one layer of l_1 child nodes. With this specific DAG, the GDP is given by

$$G_1^{(0)} \mid \alpha_1^{(0)}, G_0 \sim DP\left(\alpha_1^{(0)}, G_0\right),$$
$$G_j^{(1)} \mid \alpha_j^{(1)}, G_1^{(0)} \sim DP\left(\alpha_j^{(1)}, G_1^{(0)}\right), \quad j = 1, 2, \ldots, l_1,$$

which is clearly an HDP.

3.2 GDP Mixture Model

To cluster observations that are organized into possibly non-exchangeable groups, we use the proposed GDP in Section 3.1 as a mixing distribution of a mixture model. Letting j index the groups and i index the observations within each group, we assume that the observations $x_{j1}, x_{j2}, \ldots, x_{jn_j}$ are exchangeable within each group j but the groups may not be exchangeable. We assume that each observation within a group is drawn independently from the mixture model (5) and G_j's follow the GDP (7) and (8).

3.3 Hyperpriors

We assign a Dirichlet prior on the weights $\{\pi_{jl}^{(k)} : l \in an^{(k,1)}(j)\}$ in (8),

$$\{\pi_{jl}^{(k)} : l \in an^{(k,1)}(j)\} \sim Dir\left(\{\alpha_l^{(k-1)} : l \in an^{(k,1)}(j)\}\right), \tag{9}$$

where the parameters $\{\alpha_l^{(k-1)} : l \in an^{(k,1)}(j)\}$ correspond to the concentration parameters of the parents (generation-1 ancestors) of node j. Since the concentration parameter of a DP relates to its precision (inverse-variance), assuming a Dirichlet prior for the mixture weights of any node with Dirichlet parameters proportional to the precisions of the parent nodes is a natural choice. This gives more "weightage" to a parent node with a higher precision as opposed to a parent node with a lower precision.

The other distributional consideration that significantly simplifies the distribution of the random measure of any particular node is by considering a gamma-DAG distribution on the concentration parameters $\alpha_j^{(k)}$'s, which, like the distribution of $G_j^{(k)}$'s, also respects the same Markov property of DAG D. Specifically, we assume that

$$\begin{aligned}
&\alpha_1^{(0)} \mid \alpha_0 \sim Gamma(\alpha_0, 1), \\
&\alpha_j^{(k)} \mid \{\alpha_l^{(k-1)} : l \in an^{(k,1)}(j)\} \sim Gamma\left(\sum_{l \in an^{(k,1)}(j)} \alpha_l^{(k-1)}, 1\right), \quad j = 1, 2, \ldots, l_k.
\end{aligned} \tag{10}$$

In other words, the concentration parameter of the root node follows a gamma distribution with a fixed shape α_0 and a unit rate. The concentration parameter at any level of the DAG follows a conditionally gamma distribution with the shape parameter equal to the sum of the shape parameters of its parents. Such a choice of Gamma hyperprior on the concentration parameters of bottom level DPs of HDP have been considered in Williamson et al., 2013. We extend such a construction for the more general framework of our proposed

GDP. In the next section, we will see how our choice of hyperpriors and hyperparameters leads to several compact representations of the proposed GDP, which requires two lemmas. The first lemma is Lemma 3.1 from Sethuraman, 1994, which we state here.

Lemma 1 (Sethuraman, 1994) *Let* $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1k})$ *and* $\boldsymbol{\alpha}_2 = (\alpha_{21}, \alpha_{22}, \ldots, \alpha_{2k})$ *be* k*-dimensional vectors with* $\alpha_{ij} > 0 \ \forall \ j = 1, 2, \ldots, k, \ \ i = 1, 2$*. Let* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$ *be independent* k*-dimensional random vectors distributed as Dirichlet distribution with parameters* $\boldsymbol{\alpha}_1$ *and* $\boldsymbol{\alpha}_2$*, respectively. Let* $\alpha_{1\cdot} = \sum_{j=1}^{k} \alpha_{1j}$ *and* $\alpha_{2\cdot} = \sum_{j=1}^{k} \alpha_{2j}$*. Let* π *be independent of* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$ *and have a beta distribution Beta* $(\alpha_{1\cdot}, \alpha_{2\cdot})$*. Then the distribution of* $\pi \boldsymbol{X}_1 + (1 - \pi) \boldsymbol{X}_2$ *is the Dirichlet distribution with parameter* $\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2$*.*

The proof is provided in Section B of the Appendix for completeness. The next lemma is an immediate extension of Theorem 1 for more than two independent Dirichlet distributed random vectors. As the Dirichlet distribution is a multivariate analog of the beta distribution, by considering a Dirichlet distribution on the weights, we arrive at a similar result. This lemma is a finite-dimensional version of Theorem 1 of Williamson et al., 2013, which essentially states that a finite Dirichlet mixture of DPs is, in turn, a DP with its concentration parameter being the sum of the concentration parameters of the component DPs, and the base measure being a weighted mixture of the corresponding mixing base measures.

Lemma 2 *Let* $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L$ *be* k*-dimensional vectors where* $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{ik})$ *with* $\alpha_{ij} > 0 \ \forall \ j = 1, 2, \ldots, k, \ \ i = 1, 2, \ldots, L$*. Let* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_L$ *be independent* k*-dimensional random vectors distributed as Dirichlet distribution with parameters* $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L$*, respectively. Let* $\alpha_{i\cdot} = \sum_{j=1}^{k} \alpha_{ij}, \ \ i = 1, 2, \ldots, L$*. Let* $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_L)$ *be independent of* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_L$ *and have a Dirichlet distribution Dir* $(\alpha_{1\cdot}, \alpha_{2\cdot}, \ldots, \alpha_{L\cdot})$*. Then the distribution of* $\sum_{i=1}^{L} \pi_i \boldsymbol{X}_i$ *is the Dirichlet distribution with parameter* $\sum_{i=1}^{L} \boldsymbol{\alpha}_i$*.*

The proof is provided in Section B of the Appendix, which uses the fact that Dirichlet distribution is normalized gamma distribution. This lemma will be used to prove the hypergraph representation of GDP in the next section.

4. Representations of the Graphical Dirichlet Process

In this section, we characterize the proposed GDP through (i) the hypergraph representation, (ii) the stick-breaking representation, (iii) the restaurant-type process representation, and (iv) the limit of finite mixture representation.

4.1 The Hypergraph Representation

The GDP, along with the hyperpriors on the concentration parameters and mixture weights, can be represented hierarchically as,

$$
\begin{aligned}
&\alpha_1^{(0)} \mid \alpha_0 \sim Gamma\left(\alpha_0, 1\right), \\
&G_1^{(0)} \mid \alpha_1^{(0)}, G_0 \sim DP\left(\alpha_1^{(0)}, G_0\right), \\
&\alpha_j^{(k)} \mid \{\alpha_l^{(k-1)} : l \in an^{(k,1)}(j)\} \sim Gamma\left(\sum_{l \in an^{(k,1)}(j)} \alpha_l^{(k-1)}, 1\right), \\
&\{\pi_{jl}^{(k)} \,:\, l \in an^{(k,1)}(j)\} \mid \{\alpha_l^{(k-1)} : l \in an^{(k,1)}(j)\} \sim Dir\left(\{\alpha_l^{(k-1)} : l \in an^{(k,1)}(j)\}\right), \\
&G_j^{(k)} \mid \alpha_j^{(k)}, \{G_l^{(k-1)} : l \in an^{(k,1)}(j)\} \sim DP\left(\alpha_j^{(k)}, \sum_{l \in an^{(k,1)}(j)} \pi_{jl}^{(k)} G_l^{(k-1)}\right),
\end{aligned}
\tag{11}
$$

for $j = 1, 2, \ldots, l_k$ and $k = 1, \ldots, K$.

The hyperparameters of the GDP consist of the base probability measure G_0 and the concentration parameter α_0. The probability measure $G_1^{(0)}$ of the root node varies around the base measure G_0 with the amount of variability governed by $\alpha_1^{(0)}$, which in turn is governed by the hyperparameter α_0. We now present a novel hypergraph representation of GDP, which simplifies the graph-based distribution. The representation follows from the gamma-DAG distribution on the concentration parameters and standard properties of Dirichlet distribution.

Theorem 3 (Hypergraph Representation) *Consider a DAG D that has K layers and l_k distinct nodes in layer k for $k = 1, \ldots, K$. Under model* (11)*, the distribution of the random measure $G_j^{(k)}$ of node j in layer k of DAG D can be equivalently represented as,*

$$
\begin{aligned}
&G_j^{(k)} \mid \alpha_j^{(k)}, H_j^{(k,k)} \sim DP\left(\alpha_j^{(k)}, H_j^{(k,k)}\right), \\
&H_j^{(k,k)} \mid \{\alpha_l^{(k-1)} : l \in an^{(k,1)}(j)\}, H_j^{(k,k-1)} \sim DP\left(\sum_{l \in an^{(k,1)}(j)} \alpha_l^{(k-1)}, H_j^{(k,k-1)}\right), \\
&H_j^{(k,k-1)} \mid \{\alpha_l^{(k-2)} : l \in an^{(k,2)}(j)\}, H_j^{(k,k-2)} \sim DP\left(\sum_{l \in an^{(k,2)}(j)} \alpha_l^{(k-2)}, H_j^{(k,k-2)}\right), \\
&\vdots \\
&H_j^{(k,2)} \mid \{\alpha_l^{(1)} : l \in an^{(k,k-1)}(j)\}, G_1^{(0)} \sim DP\left(\sum_{l \in an^{(k,k-1)}(j)} \alpha_l^{(1)}, G_1^{(0)}\right).
\end{aligned}
$$

The proof is provided in the Appendix A. In words, Theorem 3 essentially states the following. The distribution of $G_j^{(k)}$ is a DP with a hidden base measure $H_j^{(k,k)}$ and the

concentration parameter $\alpha_j^{(k)}$. The hidden base measure $H_j^{(k,k)}$, in turn, is again a DP with base measure $H_j^{(k,k-1)}$ and concentration parameter being the sum of the concentration parameters of the generation-1 ancestors of $G_j^{(k)}$. Recursively, the hidden base measure $H_j^{(k,k-1)}$ is a DP with base measure $H_j^{(k,k-2)}$ and the concentration parameter being the sum of the concentration parameters of the generation-2 ancestors. This distributional pattern continues in a hierarchical fashion. Through $k-1$ hidden base measures, any node in layer k can be seen to depend on the root node $G_1^{(0)}$ through its ancestral relationships. We call the representation of GDP in Theorem 3 as the hypergraph representation because one can view $H_j^{(k,k-a)}$ for $a = 0, \ldots, k-2$ as a hypernode that contains all the sufficient information from generation-$(a+1)$ ancestors of $G_j^{(k)}$. We provide in Figure 6 an illustrative example of the hypergraph representation showing how the hypernodes contain all the ancestral information. From Figure 6a, we can see that the distribution of G_6 depends on the distribution of its parents, G_2 and G_3. We refer to H_2, consisting of $\{G_2, G_3\}$, as a hypernode. Hypernode H_2 contains all the information about the parents of G_6. Loosely speaking, the information of the root node G_1 (e.g., its atoms) is passed to G_6 through H_2. Similarly, H_3, being the hypernode of $\{G_3, G_4\}$, contains all the information about G_7 from its parent nodes allowing the flow of information from the root node (see Figure 6b). For node G_8, we have two levels of hypernodes–H_4 denotes the first layer and consists of the parents of G_8, and H^* denotes the second layer and consists of generation-2 ancestors of G_8. Thus, hypernodes H_4 and H^* carry all the information from the root node G_1 to G_8 as illustrated in Figure 6c.

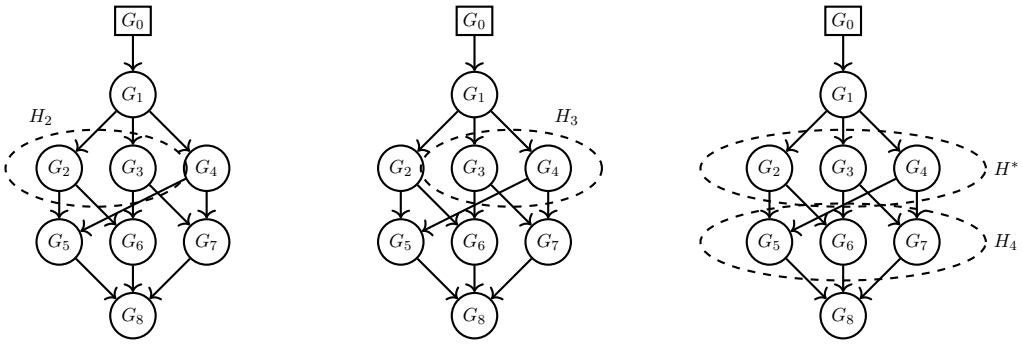


Figure 6: Illustration of hypernodes (represented by dashed ovals) of the DAG for our motivational problem. (a) Hypernode H_2 consists of the generation-1 ancestors (i.e., G_2 and G_3) of node G_6. (b) Hypernode H_3 consists of the generation-1 ancestors (i.e., G_3 and G_4) of node G_7. (c) Hypernode H_4 consists of the generation-1 ancestors (i.e., G_5, G_6, and G_7) of node G_8. Hypernode H^* consists of the generation-2 ancestors (i.e., G_2, G_3, and G_4) of node G_8.

We will exploit this representation to derive the stick-breaking representation and the limit of finite mixture representation of the proposed GDP in the next subsections.

4.2 The Stick-Breaking Representation

Given that the random measure $G_1^{(0)}$ of the root node is distributed as a DP, it can be expressed using a stick-breaking representation,

$$G_1^{(0)} = \sum_{l=1}^{\infty} \beta_{1l}^{(0)} \delta_{\phi_l}, \tag{12}$$

where $\phi_l \overset{iid}{\sim} G_0$ and $\boldsymbol{\beta}_1^{(0)} = \left(\beta_{1l}^{(0)}\right)_{l=1}^{\infty} \sim \text{GEM}\left(\alpha_1^{(0)}\right)$ are mutually independent. We interpret $\boldsymbol{\beta}_1^{(0)}$ as a probability measure on the positive integers. Since $G_1^{(0)}$ has support at the atoms $\boldsymbol{\phi} = (\phi_l)_{l=1}^{\infty}$, each $G_j^{(k)}$ necessarily has support at these atoms as well and hence can be expressed as,

$$G_j^{(k)} = \sum_{l=1}^{\infty} \beta_{jl}^{(k)} \delta_{\phi_l}. \tag{13}$$

As with Theorem 3, the stick-breaking weights depend hierarchically on a set of hidden weights. Letting $\boldsymbol{\beta}_j^{(k)} = \left(\beta_{jl}^{(k)}\right)_{l=1}^{\infty}$ be the stick-breaking weights for node j in layer k of DAG D and letting $\boldsymbol{\nu}_j^{(k,m)} = \left(\nu_{jl}^{(k,m)}\right)_{l=1}^{\infty}, m = 2, \ldots, k$ be their hidden weights, we have the following corollary.

Corollary 4 (Stick-Breaking Representation) *Consider a DAG D that has K layers and l_k distinct nodes in layer k for $k = 1, \ldots, K$. The stick-breaking weights $\boldsymbol{\beta}_j^{(k)}$ of node j at layer k of DAG D can be represented as*

$$\begin{aligned}
&\boldsymbol{\beta}_j^{(k)} \mid \alpha_j^{(k)}, \boldsymbol{\nu}_j^{(k,k)} \sim DP_{\mathbb{Z}^+}\left(\alpha_j^{(k)}, \boldsymbol{\nu}_j^{(k,k)}\right), \\
&\boldsymbol{\nu}_j^{(k,k)} \mid \{\alpha_l^{(k-1)} : l \in an^{(k,1)}(j)\}, \boldsymbol{\nu}_j^{(k,k-1)} \sim DP_{\mathbb{Z}^+}\left(\sum_{l \in an^{(k,1)}(j)} \alpha_l^{(k-1)}, \boldsymbol{\nu}_j^{(k,k-1)}\right), \\
&\boldsymbol{\nu}_j^{(k,k-1)} \mid \{\alpha_l^{(k-2)} : l \in an^{(k,2)}(j)\}, \boldsymbol{\nu}_j^{(k,k-2)} \sim DP_{\mathbb{Z}^+}\left(\sum_{l \in an^{(k,2)}(j)} \alpha_l^{(k-2)}, \boldsymbol{\nu}_j^{(k,k-2)}\right), \\
&\vdots \\
&\boldsymbol{\nu}_j^{(k,2)} \mid \{\alpha_l^{(1)} : l \in an^{(k,k-1)}(j)\}, \boldsymbol{\beta}_1^{(0)} \sim DP_{\mathbb{Z}^+}\left(\sum_{l \in an^{(k,k-1)}(j)} \alpha_l^{(1)}, \boldsymbol{\beta}_1^{(0)}\right),
\end{aligned}$$

where $DP_{\mathbb{Z}^+}(a, \boldsymbol{\eta})$ denotes the random probability measure on the positive integers distributed as a Dirichlet process with the concentration parameter $a > 0$ and base measure on the positive integers, $\boldsymbol{\eta}$.

The proof of this corollary directly follows from the hypergraph representation of Theorem 3 and is hence omitted. We call this representation *the stick-breaking representation* where $\boldsymbol{\nu}_j^{(k,k)}$ is interpreted as a hidden probability measure on the set of

positive integers corresponding to the first hidden layer. Each hidden layer of stick-breaking weights depend hierarchically on its previous hidden layer, denoted by $\boldsymbol{\nu}_j^{(k,k-1)}$, $\boldsymbol{\nu}_j^{(k,k-2)}$, and so on, and finally on the weights $\boldsymbol{\beta}_1^{(0)}$ of the root node.

4.3 The Family-Owned Restaurant Process Representation

DP and HDP have the well-known Chinese restaurant process and franchise representations. Here, we provide a culinary analog for the proposed GDP. We refer to this process as the *family-owned restaurant process* as it is customary to use familial relationships to describe the relationships between nodes in a DAG. The metaphor is as follows. An original restaurant is opened by the ancestor of a family (the root node), which serves some dishes from a global menu containing an infinite number of dishes. The descendants of the ancestor open their own respective restaurants, which serve some of the dishes already being served in the restaurants owned by their parents and possibly some new dishes from the global menu. At each table of the original restaurant, one dish is ordered from the menu by the first customer occupying the table, and the dish is shared by all the other customers who sit at that table. Any subsequent customer may either join an occupied table and share the dish being served at that table or open a new table with a new dish from the menu. In restaurants other than the original restaurant, however, the first customer might choose to select a dish being served at one of the tables of its parent restaurant or order a new dish from the menu. Since the hypergraph representation of GDP involves hypernodes with hidden probability measures, we introduce a notation for the number of tables serving a dish in any restaurant and demarcate them with the notation for the number of tables serving the dish in the hypernodes, which we refer to as hyper-restaurants.

As before, assume that there are K generations in the family and there are l_k different restaurants in generation k. The restaurants correspond to the nodes of DAG D. The customers coming in restaurant j of generation k correspond to parameters $\theta_{ji}^{(k)}$. Let $\phi_1, \phi_2, \ldots, \phi_L$ denote i.i.d. random variables distributed according to the base distribution G_0, which are dishes from the global menu. To maintain a count of customers and tables, we introduce two notations. We use the notation $n_{jt}^{(k)}$ to denote the number of customers at table t in the restaurant j of generation k and the notation $m_{jl}^{(k)}$ to denote the number of tables in the restaurant j of generation k that serve dish l. Marginal counts are represented by dots at the appropriate indices. For example, $m_{j\cdot}^{(k)}$ denotes the count of all the tables (regardless of what dishes being served) in the restaurant j of generation k. We introduce the notation $\psi_{jt}^{(k,k)}$ to denote the dish served at table t in restaurant j of generation k, chosen from the corresponding layer-1 hyper-restaurant ($H_j^{(k,k)}$).

We integrate out random measures $\left\{G_j^{(k)}, H_j^{(k,k)}, H_j^{(k,k-1)}, \ldots, G_1^{(0)}\right\}$ sequentially. First, we find the conditional distribution of $\theta_{ji}^{(k)}$ given $\theta_{j1}^{(k)}, \theta_{j2}^{(k)}, \ldots, \theta_{j,i-1}^{(k)}, \alpha_j^{(k)}$, and $H_j^{(k,k)}$ with $G_j^{(k)}$ integrated out,

$$\theta_{ji}^{(k)} \mid \theta_{j1}^{(k)}, \theta_{j2}^{(k)}, \ldots, \theta_{j,i-1}^{(k)}, \alpha_j^{(k)}, H_j^{(k,k)} \sim \sum_{t=1}^{m_{j\cdot}^{(k)}} \frac{n_{jt}^{(k)}}{i-1+\alpha_j^{(k)}} \delta_{\psi_{jt}^{(k,k)}} + \frac{\alpha_j^{(k)}}{i-1+\alpha_j^{(k)}} H_j^{(k,k)}, \quad (14)$$

We let $\psi_{jt}^{(k,k-1)}$ to denote the dish served at table t in the layer-1 hyper-restaurant corresponding to restaurant j of generation k, chosen from the dishes served in the layer-2 hyper-restaurants ($H_j^{(k,k-1)}$). Integrating out the hidden measure from the current layer $H_j^{(k,k)}$, the conditional distribution of $\psi_{jt}^{(k,k)}$ given $\psi_{j1}^{(k,k-1)}, \psi_{j2}^{(k,k-1)}, \ldots, \psi_{j1}^{(k,k)}, \ldots, \psi_{j,t-1}^{(k,k)}, \{\alpha_l^{(k-1)} : l \in an^{(k,1)}(j)\}$, and the hidden measure from the previous layer, $H_j^{(k,k-1)}$ is given by,

$$\begin{aligned}
&\psi_{jt}^{(k,k)} \mid \psi_{j1}^{(k,k-1)}, \psi_{j2}^{(k,k-1)}, \ldots, \psi_{j1}^{(k,k)}, \ldots, \psi_{j,t-1}^{(k,k)}, \{\alpha_l^{(k-1)} : l \in an^{(k,1)}(j)\}, H_j^{(k,k-1)} \\
&\sim \sum_{l=1}^{M_j^{(k,k-1)}} \frac{m_{jl}^{(k,k-1)}}{m_{j\cdot}^{(k,k-1)} + \sum_{l \in an^{(k,1)}(j)} \alpha_l^{(k-1)}} \delta_{\psi_{jl}^{(k,k-1)}} + \frac{\sum_{l \in an^{(k,1)}(j)} \alpha_l^{(k-1)}}{m_{j\cdot}^{(k,k-1)} + \sum_{l \in an^{(k,1)}(j)} \alpha_l^{(k-1)}} H_j^{(k,k-1)},
\end{aligned} \tag{15}$$

where the notation $m_{jl}^{(k,k-1)}$ denotes the number of tables in layer-1 hyper-restaurant, corresponding to restaurant j of generation k serving the dish l. We denote by $M_j^{(k,k-1)}$ the number of dishes served in the layer-1 hyper-restaurants and by $m_{j\cdot}^{(k,k-1)}$ the total number of tables in the layer-1 hyper-restaurant, corresponding to the restaurant j of generation k. Similarly, integrating out the measure $H_j^{(k,k-1)}$ and introducing the next layer of variables $\psi_{jt}^{(k,k-2)}$, the conditional distribution of $\psi_{jt}^{(k,k-1)}$ given $\psi_{j1}^{(k,k-2)}, \psi_{j2}^{(k,k-2)}, \ldots, \psi_{j1}^{(k,k-1)}, \ldots, \psi_{j,t-1}^{(k,k-1)}, \{\alpha_l^{(k-2)} : l \in an^{(k,2)}(j)\}$, and the hidden measure from the previous layer $H_j^{(k,k-2)}$ is given by,

$$\begin{aligned}
&\psi_{jt}^{(k,k-1)} \mid \psi_{j1}^{(k,k-2)}, \psi_{j2}^{(k,k-2)}, \ldots, \psi_{j1}^{(k,k-1)}, \ldots, \psi_{j,t-1}^{(k,k-1)}, \{\alpha_l^{(k-2)} : l \in an^{(k,2)}(j)\}, H_j^{(k,k-2)} \\
&\sim \sum_{l=1}^{M_j^{(k,k-2)}} \frac{m_{jl}^{(k,k-2)}}{m_{j\cdot}^{(k,k-2)} + \sum_{l \in an^{(k,2)}(j)} \alpha_l^{(k-2)}} \delta_{\psi_{jl}^{(k,k-2)}} + \frac{\sum_{l \in an^{(k,2)}(j)} \alpha_l^{(k-2)}}{m_{j\cdot}^{(k,k-2)} + \sum_{l \in an^{(k,2)}(j)} \alpha_l^{(k-2)}} H_j^{(k,k-2)}.
\end{aligned} \tag{16}$$

As in the stick-breaking representation, we can recursively integrate out hidden measures and eventually arrive at the conditional distribution of $\psi_{jt}^{(k,2)}$ given $\psi_{j1}^{(0)}, \psi_{j2}^{(0)}, \ldots, \psi_{j1}^{(k,2)}, \ldots, \psi_{j,t-1}^{(k,2)}, \{\alpha_l^{(1)} : l \in an^{(k,k-1)}(j)\}$, and the probability measure of the root node $G_1^{(0)}$,

$$\begin{aligned}
&\psi_{jt}^{(k,2)} \mid \psi_{j1}^{(0)}, \psi_{j2}^{(0)}, \ldots, \psi_{j1}^{(k,2)}, \ldots, \psi_{j,t-1}^{(k,2)}, \{\alpha_l^{(1)} : l \in an^{(k,k-1)}(j)\}, G_1^{(0)} \\
&\qquad \sim \sum_{l=1}^{M_j^{(k,1)}} \frac{m_{jl}^{(k,1)}}{m_{j\cdot}^{(k,1)} + \sum_{l \in an^{(k,k-1)}(j)} \alpha_l^{(1)}} \delta_{\psi_{jl}^{(0)}} + \frac{\sum_{l \in an^{(k,k-1)}(j)} \alpha_l^{(1)}}{m_{j\cdot}^{(k,1)} + \sum_{l \in an^{(k,k-1)}(j)} \alpha_l^{(1)}} G_1^{(0)},
\end{aligned} \tag{17}$$

and the conditional distribution of $\psi_{jt}^{(0)}$ given $\psi_{j1}^{(0)}, \ldots, \psi_{j,t-1}^{(0)}, \alpha_1^{(0)}$, and the base measure G_0,

$$\psi_{jt}^{(0)} \mid \psi_{j1}^{(0)}, \ldots, \psi_{j,t-1}^{(0)}, \alpha_1^{(0)}, G_0 \sim \sum_{l=1}^{L} \frac{m_l^{(0)}}{m_{\cdot}^{(0)} + \alpha_1^{(0)}} \delta_{\phi_l} + \frac{\alpha_1^{(0)}}{m_{\cdot}^{(0)} + \alpha_1^{(0)}} G_0, \tag{18}$$

where $m_l^{(0)}$ denotes the number of tables in the original restaurant serving dish l and $m_\cdot^{(0)}$ denotes the total number of tables in the original restaurant. Note that (18) corresponds to the case where the root node is hidden (the same as in HDP). When the root node is not hidden, a similar formula can be derived, which is omitted for simplicity.

4.4 The Infinite Limit of Finite Mixture Model

The GDP mixture model can be derived as the infinite limit of a finite mixture model. Let us denote the observations and the mixture component indicator from node j in layer k of DAG D by $x_{ji}^{(k)}$ and $z_{ji}^{(k)}$, respectively. Suppose $\boldsymbol{\beta}_1^{(0)}$ is the vector of mixing weights for the root node. Denoting by $\boldsymbol{\beta}_j^{(k)}$ the mixing weights of node j in layer k and by $\nu_j^{(k,m)}$ the corresponding mixing weights for the hidden layer m, with $m = 2, \ldots, k$, we consider a finite mixture version of the proposed GDP,

$$
\begin{aligned}
\boldsymbol{\beta}_1^{(0)} \mid \alpha_1^{(0)} &\sim Dir\left(\alpha_1^{(0)}/L, \ldots, \alpha_1^{(0)}/L\right), \\
\boldsymbol{\nu}_j^{(k,2)} \mid \{\alpha_l^{(1)} : l \in an^{(k,k-1)}(j)\}, \boldsymbol{\beta}_1^{(0)} &\sim Dir\left(\sum_{l \in an^{(k,k-1)}(j)} \alpha_l^{(1)} \left(\beta_{11}^{(0)}, \ldots, \beta_{1L}^{(0)}\right)\right), \\
&\vdots \\
\boldsymbol{\nu}_j^{(k,k)} \mid \{\alpha_l^{(k,k-1)} : l \in an^{(k,1)}(j)\}, \boldsymbol{\nu}_j^{(k,k-1)} &\sim Dir\left(\sum_{l \in an^{(k,1)}(j)} \alpha_l^{(k-1)} \left(\nu_{j1}^{(k,k-1)}, \ldots, \nu_{jL}^{(k,k-1)}\right)\right), \\
\boldsymbol{\beta}_j^{(k)} \mid \alpha_j^{(k)}, \boldsymbol{\nu}_j^{(k,k)} &\sim Dir\left(\alpha_j^{(k)} \left(\nu_{j1}^{(k,k)}, \ldots, \nu_{jL}^{(k,k)}\right)\right), \\
\phi_l \mid G_0 &\sim G_0, \\
z_{ji}^{(k)} \mid \boldsymbol{\beta}_j^{(k)} &\sim \boldsymbol{\beta}_j^{(k)}, \\
x_{ji}^{(k)} \mid z_{ji}^{(k)}, (\phi_l)_{l=1}^L &\sim F\left(\phi_{z_{ji}^{(k)}}\right).
\end{aligned} \tag{19}
$$

The distribution of this finite mixture model approaches the GDP mixture model as $L \to \infty$. Refer to Section C of the Appendix for the proof. Based on this finite mixture model approximation with a large enough truncation level L, we develop an efficient posterior inference procedure of our model using a Metropolis-within-blocked-Gibbs sampler with a specialized proposal (Director et al., 2017); see Section D of the Appendix for details.

5. Simulations

Our simulations are designed to mimic the motivating application where we have 8 experimental groups, whose relationships are represented by the DAG in Figure 7.

We generated data within each of the 8 groups from a four-component mixture of bivariate Gaussian distributions with different covariance matrices for each group. We drew the DP concentration parameters α_j's for the different groups from their prior distribution (10) respecting the DAG in Figure 7 with $\alpha_0 = 5$. The weights of the finite mixture model corresponding to the different groups were drawn using (19) and the same DAG. The true cluster indicators of each group were sampled from a multinomial distribution with probabilities equal to the mixture weights. Using these true cluster indices for each group, samples were drawn from the Gaussian distribution with the cluster-specific mean and group-specific covariance matrix, given in Tables 4 and 5, respectively, in Section E of the Appendix. Refer to the same section in the Appendix for more details on our

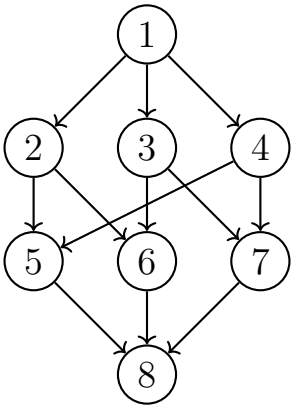


Figure 7: The DAG of experimental groups.

simulation strategy. In our Gibbs sampler, the truncation level of the finite mixture model was set to $L = 10$, and the base measure for GDP, G_0, was specified as the normal-inverse-Wishart distribution, $\mathcal{NIW}(\mathbf{0}, 0.01, \mathbb{I}_2, 2)$. Upon the completion of the Gibbs sampler, the clusters were estimated by using the least squares criterion (Dahl, 2006), and they were compared with the true cluster labels for evaluation. We considered various sample sizes in each group, which are summarized in Table 1. In all cases, we ran 15,000 iterations of our Gibbs sampler and after discarding the first 5,000 samples as burn-in, we retained every 10th iteration of posterior samples.

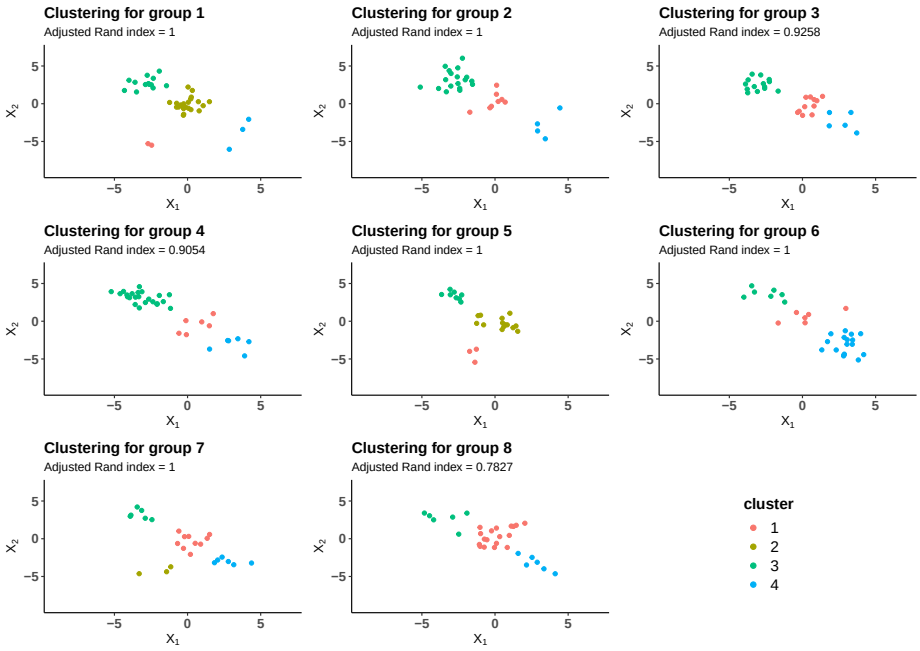
Sample sizes	Groups							
	1	2	3	4	5	6	7	8
small	40	30	30	35	25	30	25	30
moderate	80	70	70	75	83	88	92	88
large	150	160	180	170	155	175	185	145
unbalanced	350	30	40	45	25	25	35	35

Table 1: The sample sizes for the different groups that were used to simulate the data.

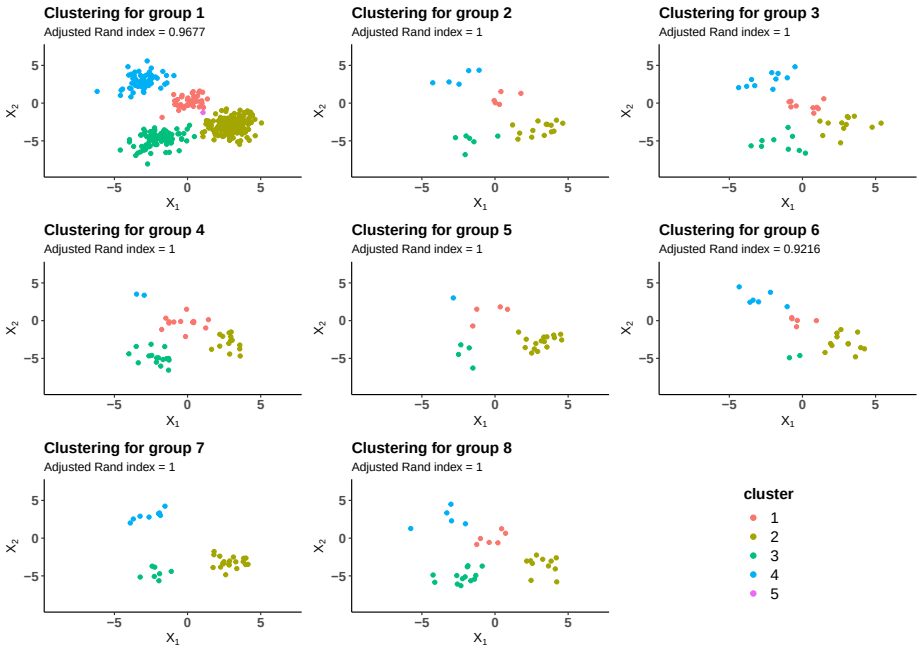
The clustering results of GDP for small and unbalanced sample sizes are visualized in Figure 8. The remaining clustering plots are shown in Appendix E. Across different sample sizes, the proposed GDP was able to identify the clusters within each group with very good accuracy and was able to link clusters across non-exchangeable groups.

We also looked at the clustering performance of GDP under a more difficult scenario. The simulation details and clustering results are shown in Appendix E. Since HDP is a special case of the proposed GDP, we compared the two methods for this difficult scenario. We also compared the clustering performance of GDP with k-means, a widely used non-Bayesian clustering technique. The number of clusters in k-means was taken to be the truncation level of our GDP. All simulations were replicated 50 times.

GDP significantly outperformed both HDP and k-means. For example, the boxplots of adjusted Rand indices (Hubert and Arabie, 1985; higher is better) for the different methods are shown in Figure 9. It is evident that the adjusted Rand indices of GDP were almost uniformly higher than those of HDP because HDP was not able to handle non-exchangeable groups. Similarly, the higher adjusted Rand indices of GDP indicated its superior clustering performance over the k-means algorithm. Moreover, k-means algorithm does not allow sharing of relevant clusters across the groups.



(a) Small sample size in each group



(b) Unbalanced sample sizes between groups

Figure 8: Clustering performance of GDP for different sample sizes. The colors indicate the estimated clusters by GDP. Adjusted Rand index is reported at the top of each panel.

We further explored additional simulations for grouped data characterized by dependencies that can be represented by a known DAG, beyond those that mimick our motivational application. Specifically, we investigated time-dependent grouped data, which can be represented using an autoregressive (AR) model. Such an AR model may be analyzed using the GDP. The proposed model, simulation details, and clustering results for various number of time points (groups) are shown in Appendix Section F. In summary, our model was able to identify the clusters within each group and link them across groups with good accuracy.

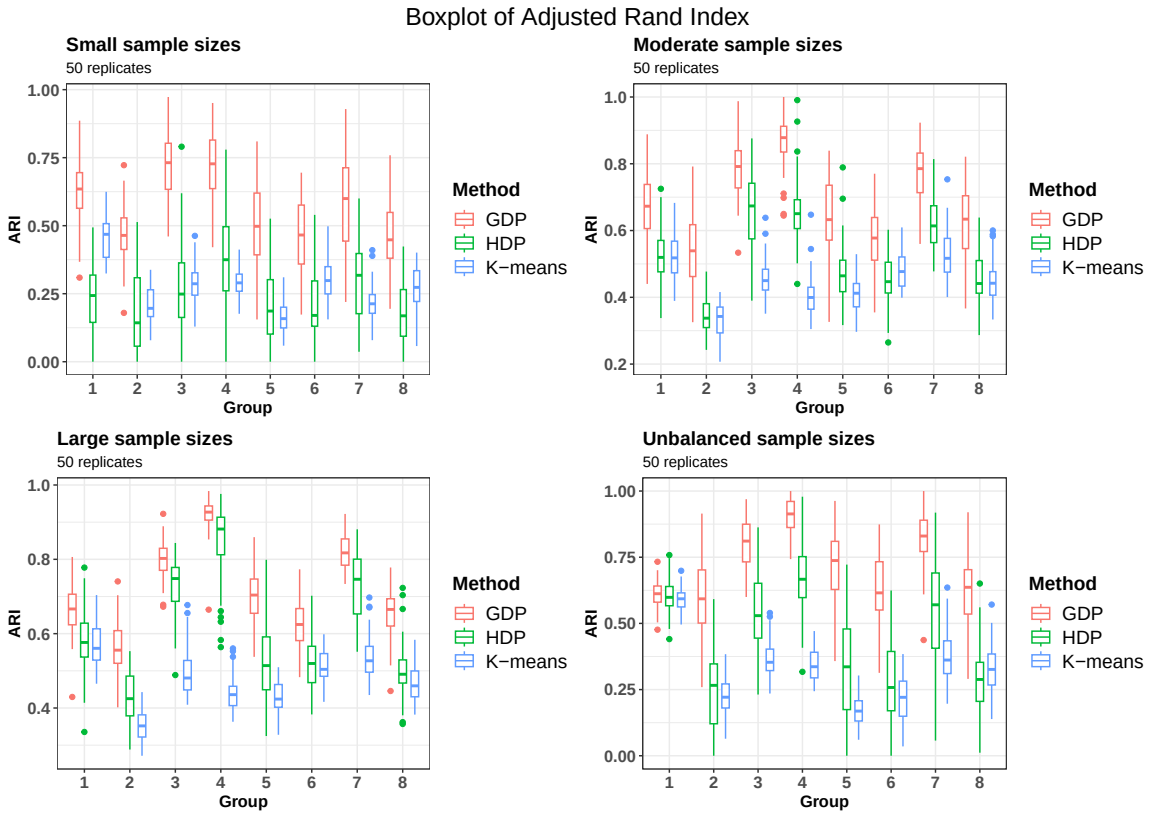


Figure 9: The boxplots of the adjusted Rand indices for GDP, HDP, and k-means for all sample sizes.

6. Real Data Analysis

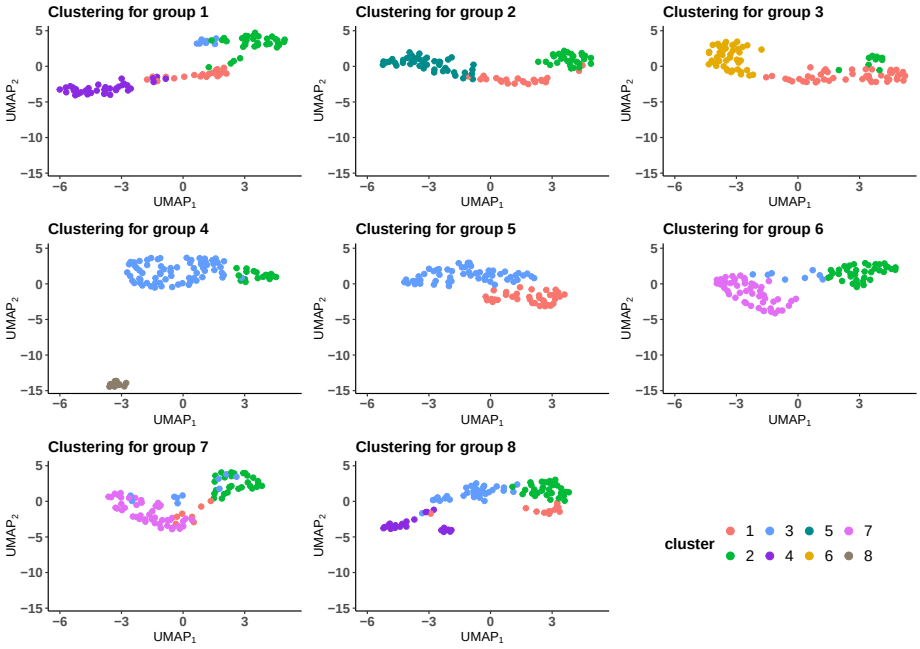
With the advancement of next-generation sequencing techniques in recent years, it is now possible to molecularly characterize individual cells, which may provide valuable insights into complex biological systems, ranging from cancer genomics to diverse microbial communities (Hwang et al., 2018). Colorectal cancer is the third most common type of cancer after breast and lung cancers. It is known that the mutation of tumor-suppressor gene Apc is an initial step in most colorectal tumors (Morin et al., 1997). In addition, numerous studies have been conducted to understand the effect of high-fat vs low-fat diet on gene expressions (Jump and Clarke, 1999; Bouchard-Mercier et al., 2013; Fan et al., 2020). We are motivated by a study that aimed to investigate how diet, genotype, and treatment with a new cancer prevention drug (AdipoRon) against placebo interacted to influence

the expression of genes in intestinal crypt and tumor cells. The experiments started from a baseline group where the mice were genetically wild-type, fed with a normal diet, and treated with placebo. Then to understand the main effects of genotype, diet, and cancer treatment on stem cell gene expression, the experimenters introduced three new groups of mice, each differing from the baseline group by exactly one factor (Apc knock-out, high-fat diet, or new cancer treatment AdipoRon). To determine the two-way interaction effects, three additional groups of mice were studied, each of which differed from the baseline group by two factors (e.g., mice with Apc knock-out, high-fat diet, and placebo). Lastly, for a three-way interaction, the experimenters introduced the eighth group of mice with Apc knock-out, a high-fat diet, and the new treatment AdipoRon. By design, these 8 experimental groups are non-exchangeable and their relationships can be delineated by the DAG in Figure 7. The goal of this analysis is to identify potential intestinal molecular subtypes within each experimental group while allowing information to be shared across these non-exchangeable groups with the proposed GDP model. For illustration, we randomly sampled 100 cells from each of the eight groups. The scRNA-seq data were pre-processed following standard procedure as outlined by Hao et al. (2021) using the R package `Seurat`. The data was log-normalized and scaled such that the mean expression across cells was 0 and the variance across cells was 1. As a common practice in single-cell data analyses, the uniform manifold approximation and projection (UMAP) (McInnes et al., 2018) was used to reduce the data to two dimensions. We considered the truncation level, $L = 30$, and the same base probability measure, G_0, as in the simulations. Furthermore, we have considered several choices of the truncation level of the GDP, which shows our method is relatively robust for $L \geq 30$; see Appendix G for details. We ran four parallel chains of the Gibbs sampler for $50,000$ iterations. To monitor the convergence of the sampler, we drew the traceplots of the log-likelihood for each of the four chains, after discarding the initial $35,000$ samples and thinning the samples by a factor of 15, which indicated no lack of convergence of our sampler. We pooled the Monte Carlo samples across different chains for posterior inference. We compared the clustering performance with that obtained from HDP on the same data.

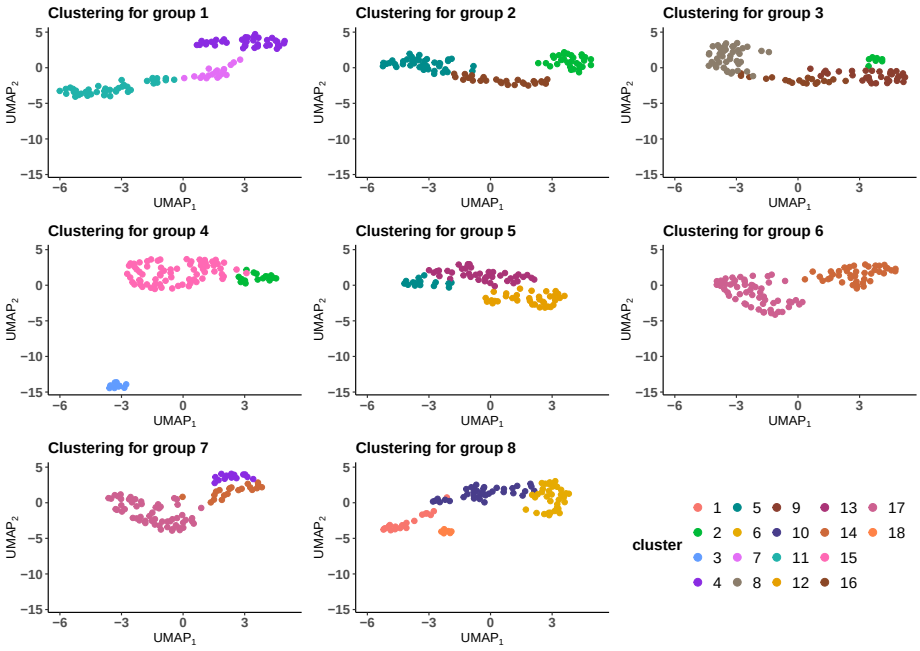
The estimated clusters from GDP and HDP are shown in Figures 10a and 10b, respectively. As shown in Table 3 in Section D of the Appendix, group 1 is the wild-type group receiving the placebo and a normal diet. Each of group 2, 3, and 4 are obtained from group 1 by changing the three factors one at a time, and hence shares some similar clusters with group 1. Group 4 is similar to the baseline group 1 but with the Apc gene knocked out. The corresponding clustering plot (Figure 10a) of GDP indicates that the Apc knock-out group seems to exhibit more heterogeneity of cells (suggesting possibly new cellular subtypes) as compared to the wild-type group. Group 5 is the Apc knock-out group receiving a high-fat diet and the placebo. The clustering plot shows some resemblance with its parent groups (groups 2 and 4) but with the absence of some parental clusters. Groups 6 and 7 show similar clustering patterns, indicating possibly similar impact of changing the corresponding factors from their parent groups. Groups 7 and 8 correspond to the Apc knock-out group receiving the new treatment and fed with a normal and high-fat diet, respectively. It can be seen that the high-fat diet group appears to have greater molecular heterogeneity than the normal diet group. The Figure 10b, on the other hand, clearly shows that HDP fails to capture meaningful clusters across the non-exchangeable groups, i.e., some points that seemingly belong to the same cluster are assigned different labels across groups. To quantify the difference between GDP and HDP, we computed several internal clustering validation measures; see Liu et al., 2010 for a review of several such measures. Table 2 compares the Calinski-Harabasz, Davies–Bouldin, and Silhouette Index between GDP and HDP. Clearly, all of them indicate the superior clustering performance of GDP over HDP.

7. Discussion

We have introduced the GDP as a graph-based stochastic process for modeling dependent random measures that are linked by a DAG. We have also introduced the corresponding infinite mixture



(a) Clustering plot for different groups by GDP.



(b) Clustering plot for different groups by HDP.

Figure 10: Clustering of the group-specific single-cell data whose dimensions are reduced to 2 by UMAP by (a) GDP and (b) HDP.

	Calinski-Harabasz Index	Davies–Bouldin Index	Silhouette Index
GDP	418.518	1.559	0.198
HDP	225.327	4.868	-0.044

Table 2: Different measures of internal clustering for GDP and HDP. Higher values of Calinski-Harabasz Index and Silhouette Index indicates better clustering. Lower values of Davies–Bouldin Index indicate well separated clusters.

model and presented how the GDP mixture model can be used for clustering grouped data with non-exchangeable groups. We provided different representations of the GDP including a novel hypergraph representation of the original process. The posterior inference was relatively straightforward. We illustrated our method using both simulations and an application to a real grouped scRNA-seq data set.

There are a few possible future directions for this work. First, it may be possible to replace the DAG in our GDP with an undirected or chain graph. The challenge is to define the joint distribution over a set of random measures given the graph where the convenient DAG factorization no longer applies. Second, it may also be possible to learn the DAG structure instead of assuming it is known, which may require independent realizations of the GDP. In theory, if there are replicates from the underlying joint distribution of random probability measures, it is possible to identify the underlying DAG up to its Markov equivalent class. In that case, we can either consider a uniform prior for DAG D, $p(D) \propto 1$ or a prior that penalizes the graph complexity, $p(D) \propto \theta^{|D|}$ where $\theta \in (0, 1)$ and $|D|$ is the number of edges in D. Then the posterior inference can be carried out by searching the DAG space via MCMC with edge addition, deletion, and reversal moves; see e.g., Section 2.4 of Choi et al. (2020). For multivariate data, a DAG may be uniquely identifiable under certain conditions such as non-Gaussianity (Shimizu et al., 2006). In the proposed construction of the GDP, we have assumed that the random probability measures corresponding to each node is non-Gaussian, i.e., a Dirichlet process. Therefore, it might be possible to extend the ideas of Shimizu et al., 2006 in our setup, replacing random variables corresponding to each node with random probability measures for the unique identifiability of the underlying DAG. Third, the proposed construction of the GDP can be extended for other processes including the Pitman-Yor process and more generally the completely random measures and normalized random measures. For the proposed *Graphical Dirichlet Process*, we have assumed that G_j is a Dirichlet Process with the base measure being a weighted mixture of the measures of the corresponding parent nodes ($\{G_l : l \in pa(j)\}$), i.e., $G_j \mid \alpha_j, \{G_l : l \in pa(j)\} \sim DP\left(\alpha_j, \sum_{l \in pa(j)} \pi_{jl} G_l\right)$. This can, by construction, be replaced by the Pitman-Yor process i.e., $G_j \mid \alpha_j, \sigma_j, \{G_l : l \in pa(j)\} \sim \mathcal{PY}\left(\alpha_j, \sigma_j, \sum_{l \in pa(j)} \pi_{jl} G_l\right)$, where $\mathcal{PY}(a, b, \pi)$ denotes a Pitman-Yor process with concentration parameter a, discount parameter $b > -a$, and base measure π, which leads to a *Graphical Pitman-Yor Process*. Alternatively, one may consider a *Graphical Gamma Process* by assuming $G_j \mid \alpha_j, \{G_l : l \in pa(j)\} \sim \Gamma P\left(\alpha_j, \sum_{l \in pa(j)} \pi_{jl} G_l\right)$, where $\Gamma P(a, \pi)$ is a Gamma process with concentration parameter a, and base measure π. One caveat though is that the proposed hypergraph representation of the proposed GDP (Theorem 3) relies on the fact that a Dirichlet mixture of DPs is a DP. Therefore, for other processes, to get an equivalent hypergraph representation the choice of the hyperpriors must be redesigned, which could be an interesting future direction..

Acknowledgments

Ni's research was partially supported by Cancer Prevention and Research Institute of Texas (CPRIT) RP230204, 1R01GM148974-01, and NSF DMS-2112943. Chapkin's research was partially supported by CPRIT RP230204, NIH RO1 CA244359, and the Allen Endowed Chair in Nutrition & Chronic Disease Prevention. Mallick's research was partially supported by NSF CCF-1934904. The authors would also like to thank the Editor, AE, and the three anonymous reviewers, whose feedback led to a substantial improvement of the paper.

Appendix A. Proof of the Hypergraph Representation

We prove Theorem 3 (the hypergraph representation of the proposed GDP) of the main manuscript in the case of our motivational problem where we have 8 groups. Note that the proof for any general DAG follows in a similar fashion by repeated application of the two lemmas in Section 3.3 of the main manuscript and properties of gamma and Dirichlet distributions, which, however, requires more involved bookkeeping of the corresponding random distributions and hence is omitted. Our proof also illustrates how the random distribution of any particular node of the DAG is related to the root node through a number of hidden random measures, which shows the clustering property of our model. In our motivating example, each group corresponds to a combination of treatment, diet, and genotype, as summarized in Table 3 in Section D of the Appendix. The underlying DAG for the problem is given in Figure 7 of the main manuscript where group 1 is the root node, groups 2-4 are the layer-1 nodes, groups 5-7 are the layer-2 nodes, and group 8 is the layer-3 node. For ease of notation, instead of using $G_1^{(0)}$ and $\alpha_1^{(0)}$ to denote the random measure and the concentration parameter of the root node, we use simply G_1 and α_1 instead; similarly for all the other nodes. Using these simplified notations, Figures 11a and 11b show the relationships among the group-specific random measures and concentration parameters according to Figure 7 of the main manuscript.

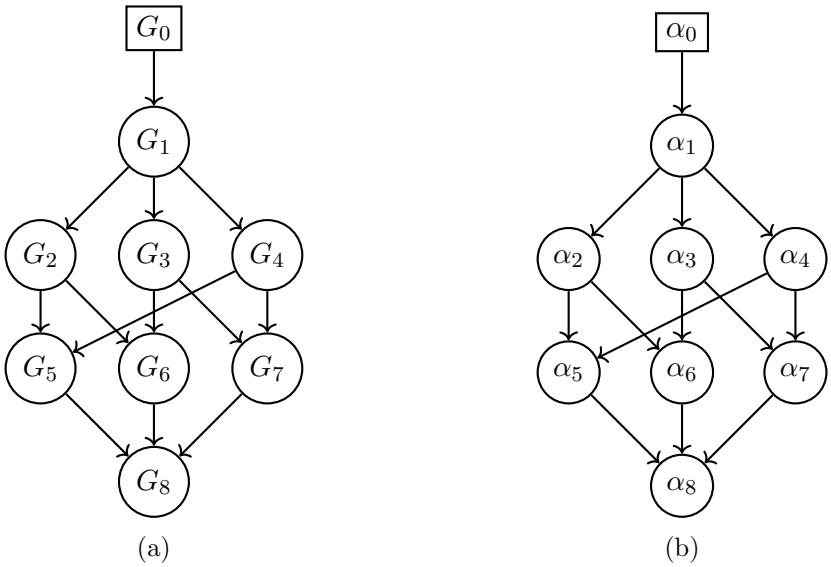


Figure 11: The DAG of the (a) random measures G_j's and (b) concentration parameters α_j's.

The proposed GDP mixture model for this problem is given hierarchically as,

$$\alpha_1 \sim Gamma(\alpha_0, 1), \qquad G_1 \sim DP(\alpha_1, G_0),$$

$$
\begin{aligned}
&\alpha_j \sim Gamma(\alpha_1, 1), \quad j = 2,3,4, && G_j \sim DP(\alpha_j, G_1), \quad j = 2,3,4,\\
&\alpha_5 \sim Gamma(\alpha_2 + \alpha_4, 1), && G_5 \sim DP(\alpha_5, \pi_1 G_2 + (1-\pi_1) G_4),\\
& && \pi_1 \sim Beta(\alpha_2, \alpha_4),\\
&\alpha_6 \sim Gamma(\alpha_2 + \alpha_3, 1), && G_6 \sim DP(\alpha_6, \pi_2 G_2 + (1-\pi_2) G_3),\\
& && \pi_2 \sim Beta(\alpha_2, \alpha_3),\\
&\alpha_7 \sim Gamma(\alpha_3 + \alpha_4, 1), && G_7 \sim DP(\alpha_7, \pi_3 G_3 + (1-\pi_3) G_4),\\
& && \pi_3 \sim Beta(\alpha_3, \alpha_4),\\
&\alpha_8 \sim Gamma(\alpha_5 + \alpha_6 + \alpha_7, 1), && G_8 \sim DP(\alpha_8, \gamma_1 G_5 + \gamma_2 G_6 + \gamma_3 G_7),\\
& && \boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3) \sim Dir(\alpha_5, \alpha_6, \alpha_7),\\
&\theta_{ji} \mid G_j \overset{ind}{\sim} G_j, & &\\
&x_{ji} \mid \theta_{ji} \overset{ind}{\sim} F(\theta_{ji}), && i = 1, \ldots, n_j, \quad j = 1, \ldots, 8.
\end{aligned} \tag{20}
$$

Now, from Theorem 3, we have the following hypergraph representation, which we are going to prove,

$$
\begin{aligned}
&\alpha_1 \sim Gamma(\alpha_0, 1), && G_1 \sim DP(\alpha_1, G_0),\\
&\alpha_j \sim Gamma(\alpha_1, 1), \quad j = 2,3,4, && G_j \sim DP(\alpha_j, G_1), \quad j = 2,3,4,\\
&\alpha_5 \sim Gamma(\alpha_2 + \alpha_4, 1), && G_5 \sim DP(\alpha_5, H_1)\\
& && H_1 \sim DP(\alpha_2 + \alpha_4, G_1),\\
&\alpha_6 \sim Gamma(\alpha_2 + \alpha_3, 1), && G_6 \sim DP(\alpha_6, H_2),\\
& && H_2 \sim DP(\alpha_2 + \alpha_3, G_1),\\
&\alpha_7 \sim Gamma(\alpha_3 + \alpha_4, 1), && G_7 \sim DP(\alpha_7, H_3),\\
& && H_3 \sim DP(\alpha_3 + \alpha_4, G_1),\\
&\alpha_8 \sim Gamma(\alpha_5 + \alpha_6 + \alpha_7, 1), && G_8 \sim DP(\alpha_8, H_4),\\
& && H_4 \sim DP(\alpha_5 + \alpha_6 + \alpha_7, H^*),\\
& && H^* \sim DP(2(\alpha_2 + \alpha_3 + \alpha_4), G_1),\\
&\theta_{ji} \mid G_j \overset{ind}{\sim} G_j, & &\\
&x_{ji} \mid \theta_{ji} \overset{ind}{\sim} F(\theta_{ji}), && i = 1, \ldots, n_j, \quad j = 1, \ldots, 8.
\end{aligned} \tag{21}
$$

Proof

Note that the random measures G_2, G_3, and G_4 are the layer-1 nodes. Their relationships to the root node G_1 are the same as those in an HDP. We shall consider the relationships of the random measures of the layer-2 and layer-3 nodes (i.e., G_5, G_6, G_7, and G_8) to the root node. Let $H_1 = \pi_1 G_2 + (1-\pi_1) G_4$ where $G_2 \sim DP(\alpha_2, G_1)$ and $G_4 \sim DP(\alpha_4, G_1)$ independently. Let $A_1, A_2, \ldots, A_r$ be a finite measurable partition of the sample space Θ. Then by the definition of DP, we have

$$
\begin{aligned}
&(G_2(A_1), G_2(A_2), \ldots, G_2(A_r)) \sim Dir\left(\alpha_2 G_1(A_1), \alpha_2 G_1(A_2), \ldots, \alpha_2 G_1(A_r)\right),\\
&(G_4(A_1), G_4(A_2), \ldots, G_4(A_r)) \sim Dir\left(\alpha_4 G_1(A_1), \alpha_4 G_1(A_2), \ldots, \alpha_4 G_1(A_r)\right),
\end{aligned}
$$

which are conditionally independent given α_2, α_4 and G_1. As $\pi_1 \sim Beta(\alpha_2, \alpha_4)$ independently of G_2 and G_4, using Theorem 1, we have that, given α_2, α_4 and G_1,

$$
\pi_1 \left(G_2(A_1), \ldots, G_2(A_r)\right) + (1-\pi) \left(G_4(A_1), \ldots, G_4(A_r)\right)
$$

$$\begin{aligned} & \sim Dir((\alpha_2+\alpha_4)(G_1(A_1),\ldots,G_1(A_r))) \\ \Rightarrow (H_1(A_1),\ldots,H_1(A_r)) \mid \alpha_2,\alpha_4,G_1 & \sim Dir((\alpha_2+\alpha_4)(G_1(A_1),\ldots,G_1(A_r))) \\ \Rightarrow H_1 \mid \alpha_2,\alpha_4,G_1 & \sim DP(\alpha_2+\alpha_4,G_1) \end{aligned}$$

Thus, we have

$$\begin{aligned} G_5 \mid H_1,\alpha_5 &\sim DP(\alpha_5,H_1), \\ H_1 \mid \alpha_2,\alpha_4,G_1 &\sim DP(\alpha_2+\alpha_4,G_1). \end{aligned} \tag{22}$$

Similarly, the other layer-2 measures G_6 and G_7 have the following representations:

$$\begin{aligned} G_6 \mid H_2,\alpha_6 &\sim DP(\alpha_6,H_2), \\ H_2 \mid \alpha_2,\alpha_3,G_1 &\sim DP(\alpha_2+\alpha_3,G_1), \end{aligned} \tag{23}$$

and,

$$\begin{aligned} G_7 \mid H_3,\alpha_7 &\sim DP(\alpha_7,H_3), \\ H_3 \mid \alpha_3,\alpha_4,G_1 &\sim DP(\alpha_3+\alpha_4,G_1), \end{aligned} \tag{24}$$

where $H_2=\pi_2G_2+(1-\pi_2)G_3$ and $H_3=\pi_3G_3+(1-\pi_3)G_4$.

Let $H_4=\gamma_1G_5+\gamma_2G_6+\gamma_3G_7$ and $\boldsymbol{\gamma}=(\gamma_1,\gamma_2,\gamma_3)\sim Dir(\alpha_5,\alpha_6,\alpha_7)$. Since G_5, G_6, and G_7 are conditionally independent given G_2, G_3, and G_4, they are also independent given H_1,H_2, and H_3. Therefore, we have,

$$\begin{aligned} G_5 \mid \alpha_5,H_1 &\sim DP(\alpha_5,H_1), \\ G_6 \mid \alpha_6,H_2, &\sim DP(\alpha_6,H_2), \\ G_7 \mid \alpha_7,H_3 &\sim DP(\alpha_7,H_3). \end{aligned}$$

For any finite measurable partition $A_1,A_2,\ldots,A_r$ of Θ, from Theorem 2, we have

$$\begin{aligned} & (H_4(A_1),\ldots,H_4(A_r)) \mid \alpha_5,\alpha_6,\alpha_7,H_1,H_2,H_3 \\ = \;& \gamma_1\left(G_5(A_1),\ldots,G_5(A_r)\right)+\gamma_2\left(G_6(A_1),\ldots,G_6(A_r)\right)+\gamma_3\left(G_7(A_1),\ldots,G_7(A_r)\right) \\ & \sim Dir\left(\left(\alpha_5H_1+\alpha_6H_2+\alpha_7H_3\right)(A_1),\ldots,\left(\alpha_5H_1+\alpha_6H_2+\alpha_7H_3\right)(A_r)\right) \\ & \equiv Dir\left(\alpha^*\left(\left(\frac{\alpha_5}{\alpha^*}H_1+\frac{\alpha_6}{\alpha^*}H_2+\frac{\alpha_7}{\alpha^*}H_3\right)(A_1),\ldots,\left(\frac{\alpha_5}{\alpha^*}H_1+\frac{\alpha_6}{\alpha^*}H_2+\frac{\alpha_7}{\alpha^*}H_3\right)(A_r)\right)\right) \\ & \equiv Dir\left(\alpha^*\left(H^*(A_1),\ldots,H^*(A_r)\right)\right) \\ & \Rightarrow H_4 \mid \alpha^*,H^* \sim DP(\alpha^*,H^*), \end{aligned} \tag{25}$$

where $\alpha^*=\alpha_5+\alpha_6+\alpha_7$ and $H^*=\frac{\alpha_5}{\alpha^*}H_1+\frac{\alpha_6}{\alpha^*}H_2+\frac{\alpha_7}{\alpha^*}H_3$. Note that α_5, α_6, and α_7 are independent gamma random variables conditionally on $\alpha_2,\alpha_3,\alpha_4$ with shape parameters $\alpha_2+\alpha_4,\alpha_2+\alpha_3$, and $\alpha_3+\alpha_4$, respectively. Thus,

$$\left(\frac{\alpha_5}{\alpha^*},\frac{\alpha_6}{\alpha^*},\frac{\alpha_7}{\alpha^*}\right) \mid \alpha_2,\alpha_3,\alpha_4 \sim Dir(\alpha_2+\alpha_4,\alpha_2+\alpha_3,\alpha_3+\alpha_4) \tag{26}$$

Thus, given $G_1,G_2,G_3,G_4,\alpha_2,\alpha_3,\alpha_4$, and from Eqs. (22–24), and Eq. (26), using Theorem 2 and using a similar measurable finite partition of Θ argument, we have,

$$H^* \mid \alpha_2,\alpha_3,\alpha_4,G_1 \sim DP(2(\alpha_2+\alpha_3+\alpha_4),G_1), \tag{27}$$

which completes the proof. ■

Appendix B. Proof of Lemma 1 and Lemma 2

B.1 Proof of Lemma 1

Lemma 1 (**Sethuraman, 1994**) *Let* $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1k})$ *and* $\boldsymbol{\alpha}_2 = (\alpha_{21}, \alpha_{22}, \ldots, \alpha_{2k})$ *be* k*-dimensional vectors with* $\alpha_{ij} > 0 \ \forall \ j = 1, 2, \ldots, k, \ i = 1, 2$. *Let* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$ *be independent* k*-dimensional random vectors distributed as Dirichlet distribution with parameters* $\boldsymbol{\alpha}_1$ *and* $\boldsymbol{\alpha}_2$*, respectively. Let* $\alpha_{1\cdot} = \sum_{j=1}^{k} \alpha_{1j}$ *and* $\alpha_{2\cdot} = \sum_{j=1}^{k} \alpha_{2j}$. *Let* π *be independent of* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$ *and have a beta distribution* $Beta\,(\alpha_{1\cdot}, \alpha_{2\cdot})$. *Then the distribution of* $\pi \boldsymbol{X}_1 + (1 - \pi)\, \boldsymbol{X}_2$ *is the Dirichlet distribution with parameter* $\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2$.

Proof Let $T_i \overset{ind}{\sim} Gamma\,(\alpha_{1i}, \lambda)\,, \ i = 1, 2, \ldots, k$ and $S_i \overset{ind}{\sim} Gamma\,(\alpha_{2i}, \lambda)\,, \ i = 1, 2, \ldots, k$ independently of T_i, where $\lambda > 0$. Let $T = \sum_{i=1}^{k} T_i$ and $S = \sum_{i=1}^{k} S_i$. We know from the reproductive property of independent gamma distributions that $T \sim Gamma\left(\sum_{i=1}^{k} \alpha_{1i}, \lambda\right) \equiv Gamma\,(\alpha_{1\cdot}, \lambda)$ and $S \sim Gamma\left(\sum_{i=1}^{k} \alpha_{2i}, \lambda\right) \equiv Gamma\,(\alpha_{2\cdot}, \lambda)$ independently of T. Define

$$\boldsymbol{X}_1 := \left(\frac{T_1}{T}, \frac{T_2}{T}, \ldots, \frac{T_k}{T}\right), \quad \boldsymbol{X}_2 := \left(\frac{S_1}{S}, \frac{S_2}{S}, \ldots, \frac{S_k}{S}\right), \quad \text{and} \quad \pi := \frac{T}{T+S}.$$

It is easy to see that $\boldsymbol{X}_1 \sim Dir(\alpha_{11}, \ldots, \alpha_{1k})$ is independent of $\boldsymbol{X}_2 \sim Dir(\alpha_{21}, \ldots, \alpha_{2k})$, and that $\pi \sim Beta\,(\alpha_{1\cdot}, \alpha_{2\cdot})$. We now need to show that π as defined above is indeed independent of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ as required by the lemma. For any fixed $\alpha_{11}, \ldots, \alpha_{1k}$, we have that $\sum_{i=1}^{k} T_i$ is a complete and sufficient statistic for λ. Because $\boldsymbol{X}_1 \sim Dir(\alpha_{11}, \ldots, \alpha_{1k})$ is ancillary for λ, by the Basu's theorem (Basu, 1955), we have that $\boldsymbol{X}_1$ is independent of $\sum_{i=1}^{k} T_i = T$. Furthermore, due to the independence of S_i and T_i, $i = 1, \ldots, k$, $\boldsymbol{X}_1$ is independent of S, and, therefore, $\boldsymbol{X}_1$ is independent of $\pi = \frac{T}{T+S}$. Similarly, $\boldsymbol{X}_2$ is also independent of π. Then,

$$\begin{aligned}
\pi \boldsymbol{X}_1 + (1 - \pi)\, \boldsymbol{X}_2 &= \frac{T}{T+S}\left(\frac{T_1}{T}, \frac{T_2}{T}, \ldots, \frac{T_k}{T}\right) + \frac{S}{T+S}\left(\frac{S_1}{S}, \frac{S_2}{S}, \ldots, \frac{S_k}{S}\right) \\
&= \left(\frac{T_1 + S_1}{T+S}, \frac{T_2 + S_2}{T+S}, \ldots, \frac{T_k + S_k}{T+S}\right) \sim Dir\,(\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2)\,,
\end{aligned}$$

because $T_i + S_i \overset{ind}{\sim} Gamma(\alpha_{1i} + \alpha_{2i}, \lambda) \ \ i = 1, 2, \ldots, k$ and $T + S \sim Gamma(\alpha_{1\cdot} + \alpha_{2\cdot}, \lambda)$. ■

B.2 Proof of Lemma 2

Lemma 2 *Let* $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L$ *be* k*-dimensional vectors where* $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{ik})$ *with* $\alpha_{ij} > 0 \ \forall \ j = 1, 2, \ldots, k, \ i = 1, 2, \ldots, L$. *Let* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_L$ *be independent* k*-dimensional random vectors distributed as Dirichlet distribution with parameters* $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L$*, respectively. Let* $\alpha_{i\cdot} = \sum_{j=1}^{k} \alpha_{ij}, \quad i = 1, 2, \ldots, L$. *Let* $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_L)$ *be independent of* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_L$ *and have a Dirichlet distribution* $Dir\,(\alpha_{1\cdot}, \alpha_{2\cdot}, \ldots, \alpha_{L\cdot})$. *Then the distribution of* $\sum_{i=1}^{L} \pi_i \boldsymbol{X}_i$ *is the Dirichlet distribution with parameter* $\sum_{i=1}^{L} \boldsymbol{\alpha}_i$.

Proof The proof is similar to that of Appendix B.1. By noting that

$$\boldsymbol{\pi} \sim Dir\,(\alpha_{1\cdot}, \alpha_{2\cdot}, \ldots, \alpha_{L\cdot}) \overset{d}{=} \left(\frac{\gamma_1}{\gamma}, \frac{\gamma_2}{\gamma}, \ldots, \frac{\gamma_L}{\gamma}\right),$$

where $\gamma_i \overset{ind}{\sim} Gamma(\alpha_{i\cdot}, \lambda), \ \ i = 1, 2, \ldots, L$ and $\gamma = \sum_{i=1}^{L} \gamma_i \sim Gamma\left(\sum_{i=1}^{L} \alpha_{i\cdot}, \lambda\right)$. The remaining proof follows from standard properties of Dirichlet distributions and mimics the proof of

Appendix B.1. ■

Appendix C. Proof of the Infinite Limit of Finite Mixture Model

The GDP mixture model can be derived as the infinite limit of a finite mixture model. Let us denote the observations and the mixture component indicator from node j in layer k of DAG D by $x_{ji}^{(k)}$ and $z_{ji}^{(k)}$, respectively. Let $\boldsymbol{\beta}_1^{(0)}$ be the vector of mixing weights for the root node. Denoting by $\boldsymbol{\beta}_j^{(k)}$ the mixing weights of node j in layer k and by $\nu_j^{(k,m)}$ the corresponding mixing weights for the hidden layer m, with $m = 2, \ldots, k$, we have

$$
\begin{aligned}
\boldsymbol{\beta}_1^{(0)} \mid \alpha_1^{(0)} &\sim Dir\left(\alpha_1^{(0)}/L, \ldots, \alpha_1^{(0)}/L\right), \\
\boldsymbol{\nu}_j^{(k,2)} \mid \{\alpha_l^{(1)} : l \in an^{(k,k-1)}(j)\}, \boldsymbol{\beta}_1^{(0)} &\sim Dir\left(\sum_{l \in an^{(k,k-1)}(j)} \alpha_l^{(1)} \left(\beta_{11}^{(0)}, \ldots, \beta_{1L}^{(0)}\right)\right), \\
&\vdots \\
\boldsymbol{\nu}_j^{(k,k)} \mid \{\alpha_l^{(k,k-1)} : l \in an^{(k,1)}(j)\}, \boldsymbol{\nu}_j^{(k,k-1)} &\sim Dir\left(\sum_{l \in an^{(k,1)}(j)} \alpha_l^{(k-1)} \left(\nu_{j1}^{(k,k-1)}, \ldots, \nu_{jL}^{(k,k-1)}\right)\right), \\
\boldsymbol{\beta}_j^{(k)} \mid \alpha_j^{(k)}, \boldsymbol{\nu}_j^{(k,k)} &\sim Dir\left(\alpha_j^{(k)} \left(\nu_{j1}^{(k,k)}, \ldots, \nu_{jL}^{(k,k)}\right)\right), \\
\phi_l \mid G_0 &\sim G_0, \\
z_{ji}^{(k)} \mid \boldsymbol{\beta}_j^{(k)} &\sim \boldsymbol{\beta}_j^{(k)}, \\
x_{ji}^{(k)} \mid z_{ji}^{(k)}, (\phi_l)_{l=1}^L &\sim F\left(\phi_{z_{ji}^{(k)}}\right).
\end{aligned}
\tag{28}
$$

Proof Consider the random probability measure

$$G_1^{(0)\,L} = \sum_{l=1}^{L} \beta_{1l}^{(0)} \delta_{\phi_l}.$$

Ishwaran and Zarepour, 2002 shows that for every measurable function g, integrable with respect to G_0, we have, given $\alpha_1^{(0)}$, as $L \to \infty$

$$\int g(\theta) dG_1^{(0),L}(\theta) \xrightarrow{\mathcal{D}} \int g(\theta) dG_1^{(0)}(\theta).$$

Further, consider

$$
\begin{aligned}
G_j^{(k)\,L} &= \sum_{l=1}^{L} \beta_{jl}^{(k)} \delta_{\phi_l}, \\
H_j^{(k,m)\,L} &= \sum_{l=1}^{L} \nu_{jl}^{(k,m)} \delta_{\phi_l}, \quad m = 2, \ldots, k.
\end{aligned}
$$

Let $(A_1, \ldots, A_r)$ be a measurable partition of the sample space Θ. Let $K_t = \{l = 1, \ldots, L : \phi_l \in A_t\}$, $t = 1, \ldots, r$, where $r \leq L$. Assuming that G_0 is non-atomic, the ϕ_l's are distinct with

probability one, implying that any partition of $\{1,\dots,L\}$ corresponds to some partition of Θ. Thus, as $\boldsymbol{\beta}_j^{(k)} \mid \alpha_j^{(k)}, \boldsymbol{\nu}_j^{(k,k)} \sim Dir\left(\alpha_j^{(k)}\left(\nu_{j1}^{(k,k)},\dots,\nu_{jL}^{(k,k)}\right)\right)$, from the properties of Dirichlet distribution, we have,

$$\begin{aligned}\left(G_j^{(k)\,L}(A_1),\dots,G_j^{(k)\,L}(A_r)\right) &= \left(\sum_{l\in K_1}\beta_{jl}^{(k)},\dots,\sum_{l\in K_r}\beta_{jl}^{(k)}\right)\\ &\sim Dir\left(\alpha_j^{(k)}\sum_{l\in K_1}\nu_{jl}^{(k,k)},\dots,\alpha_j^{(k)}\sum_{l\in K_r}\nu_{jl}^{(k,k)}\right).\end{aligned}$$

Thus,

$$G_j^{(k)\,L} \mid \alpha_j^{(k)}, H_j^{(k,k)\,L} \sim DP\left(\alpha_j^{(k)}, H_j^{(k,k)\,L}\right).$$

Similarly,

$$\begin{aligned}&H_j^{(k,k)\,L} \mid \{\alpha_l^{(k-1)} : l\in an^{(k,1)}(j)\}, H_j^{(k,k-1)\,L} \sim DP\left(\sum_{l\in an^{(k,1)}(j)}\alpha_l^{(k-1)}, H_j^{(k,k-1)\,L}\right),\\ &H_j^{(k,k-1)\,L} \mid \{\alpha_l^{(k-2)} : l\in an^{(k,2)}(j)\}, H_j^{(k,k-2)\,L} \sim DP\left(\sum_{l\in an^{(k,2)}(j)}\alpha_l^{(k-2)}, H_j^{(k,k-2)\,L}\right),\\ &\vdots\\ &H_j^{(k,2)\,L} \mid \{\alpha_l^{(1)} : l\in an^{(k,k-1)}(j)\}, G_1^{(0)\,L} \sim DP\left(\sum_{l\in an^{(k,k-1)}(j)}\alpha_l^{(1)}, G_1^{(0)\,L}\right).\end{aligned}$$

By letting $L\to\infty$, the marginal distribution that this finite mixture model induces on the observations, $\boldsymbol{x}_j^{(k)} = (x_{j1}^{(k)}, x_{j2}^{(k)},\dots)$, approaches the proposed GDP mixture model. ■

Appendix D. Finite Mixture Model Approximation and Posterior Inference

The posterior inference of the proposed GDP mixture model is carried out using a blocked Gibbs sampler. For concreteness, we will present the finite mixture model approximation of the GDP for our motivating example and posterior inference based on this approximation. In our motivating application, we have 8 experimental groups. Each group corresponds to a combination of treatment, diet, and genotype; see Table 3 where we use binary indicators to denote the genotype, the two levels of diet, and the two treatment regimes. The design of the experiments naturally introduces dependencies among the experimental groups, which are represented by the DAG in Figure 12, where group 1 is the root node, groups 2-4 are the layer-1 nodes, groups 5-7 are the layer-2 nodes, and group 8 is the layer-3 node. For ease of notation, instead of using $G_1^{(0)}$ and $\alpha_1^{(0)}$ to denote the random measure and the concentration parameter of the root node, we use simply G_1 and α_1 instead; similarly for all the other nodes.

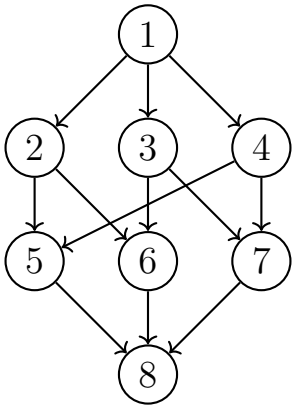


Figure 12: The DAG of experimental groups.

Group	Diet	Treatment	Genotype
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1
5	1	0	1
6	1	1	0
7	0	1	1
8	1	1	1

Table 3: Each experimental group corresponds to a combination of diet, treatment, and genotype. Diet = 1 corresponds to high-fat diet and 0 corresponds to normal diet, Treatment = 1 corresponds to AdipoRon and 0 corresponds to no therapy, Genotype = 1 corresponds to Apc knock-out and 0 corresponds to wild type.

Recall that from the main text, the finite truncation of the infinite mixture model representation is given by,

$$
\begin{aligned}
\boldsymbol{\beta}_1^{(0)} \mid \alpha_1^{(0)} &\sim Dir\left(\alpha_1^{(0)}/L, \ldots, \alpha_1^{(0)}/L\right), \\
\boldsymbol{\nu}_j^{(k,2)} \mid \{\alpha_l^{(1)} : l \in an^{(k,k-1)}(j)\}, \boldsymbol{\beta}_1^{(0)} &\sim Dir\left(\sum_{l \in an^{(k,k-1)}(j)} \alpha_l^{(1)} \left(\beta_{11}^{(0)}, \ldots, \beta_{1L}^{(0)}\right)\right), \\
&\vdots \\
\boldsymbol{\nu}_j^{(k,k)} \mid \{\alpha_l^{(k,k-1)} : l \in an^{(k,1)}(j)\}, \boldsymbol{\nu}_j^{(k-1)} &\sim Dir\left(\sum_{l \in an^{(k,1)}(j)} \alpha_l^{(k-1)} \left(\nu_{j1}^{(k,k-1)}, \ldots, \nu_{jL}^{(k,k-1)}\right)\right), \quad (29) \\
\boldsymbol{\beta}_j^{(k)} \mid \alpha_j^{(k)}, \boldsymbol{\nu}_j^{(k,k)} &\sim Dir\left(\alpha_j^{(k)} \left(\nu_{j1}^{(k,k)}, \ldots, \nu_{jL}^{(k,k)}\right)\right), \\
\phi_l \mid G_0 &\sim G_0, \\
z_{ji}^{(k)} \mid \boldsymbol{\beta}_j^{(k)} &\sim \boldsymbol{\beta}_j^{(k)}, \\
x_{ji}^{(k)} \mid z_{ji}^{(k)}, (\phi_l)_{l=1}^L &\sim F\left(\phi_{z_{ji}^{(k)}}\right).
\end{aligned}
$$

Using the simplified notations for the group-specific random measures and concentration parameter, from the finite truncation of the infinite mixture model representation from Equation (29), we have, for this motivating problem,

$$
\begin{aligned}
\alpha_1 \mid \alpha_0 &\sim Gamma(\alpha_0, 1), & \boldsymbol{\beta}_1 \mid \alpha_1 &\sim Dir(\alpha_1/L, \ldots, \alpha_1/L), \\
\alpha_2 \mid \alpha_1 &\sim Gamma(\alpha_1, 1), & \boldsymbol{\beta}_2 \mid \alpha_2, \boldsymbol{\beta}_1 &\sim Dir(\alpha_2 \boldsymbol{\beta}_1), \\
\alpha_3 \mid \alpha_1 &\sim Gamma(\alpha_1, 1), & \boldsymbol{\beta}_3 \mid \alpha_3, \boldsymbol{\beta}_1 &\sim Dir(\alpha_3 \boldsymbol{\beta}_1), \\
\alpha_4 \mid \alpha_1 &\sim Gamma(\alpha_1, 1), & \boldsymbol{\beta}_4 \mid \alpha_4, \boldsymbol{\beta}_1 &\sim Dir(\alpha_4 \boldsymbol{\beta}_1), \\
\alpha_5 \mid \alpha_2, \alpha_4 &\sim Gamma(\alpha_2 + \alpha_4, 1), & \boldsymbol{\beta}_5 \mid \alpha_5, \boldsymbol{\nu}_1 &\sim Dir(\alpha_5 \boldsymbol{\nu}_1), \\
& & \boldsymbol{\nu}_1 \mid \alpha_2, \alpha_4, \boldsymbol{\beta}_1 &\sim Dir\left((\alpha_2 + \alpha_4)\boldsymbol{\beta}_1\right), \\
\alpha_6 \mid \alpha_2, \alpha_3 &\sim Gamma(\alpha_2 + \alpha_3, 1), & \boldsymbol{\beta}_6 \mid \alpha_6, \boldsymbol{\nu}_2 &\sim Dir(\alpha_6 \boldsymbol{\nu}_2), \\
& & \boldsymbol{\nu}_2 \mid \alpha_2, \alpha_3, \boldsymbol{\beta}_1 &\sim Dir\left((\alpha_2 + \alpha_3)\boldsymbol{\beta}_1\right), \\
\alpha_7 \mid \alpha_3, \alpha_4 &\sim Gamma(\alpha_3 + \alpha_4, 1), & \boldsymbol{\beta}_7 \mid \alpha_7, \boldsymbol{\nu}_3 &\sim Dir(\alpha_7 \boldsymbol{\nu}_3), \\
& & \boldsymbol{\nu}_3 \mid \alpha_3, \alpha_4, \boldsymbol{\beta}_1 &\sim Dir\left((\alpha_3 + \alpha_4)\boldsymbol{\beta}_1\right), \\
\alpha_8 \mid \alpha_5, \alpha_6, \alpha_7 &\sim Gamma(\alpha_5 + \alpha_6 + \alpha_7, 1), & \boldsymbol{\beta}_8 \mid \boldsymbol{\nu}_4, \alpha_8, &\sim Dir(\alpha_8 \boldsymbol{\nu}_4), \\
& & \boldsymbol{\nu}_4 \mid \alpha_5, \alpha_6, \alpha_7, \boldsymbol{\eta} &\sim Dir\left((\alpha_5 + \alpha_6 + \alpha_7)\boldsymbol{\eta}\right), \\
& & \boldsymbol{\eta} \mid \alpha_2, \alpha_3, \alpha_4, \boldsymbol{\beta}_1 &\sim Dir\left(2(\alpha_2 + \alpha_3 + \alpha_4)\boldsymbol{\beta}_1\right), \\
z_{ji} \mid \boldsymbol{\beta}_j &\overset{ind}{\sim} Cat(1:L, \boldsymbol{\beta}_j), & & \\
x_{ji} \mid z_{ji}, (\phi_l)_{l=1}^L &\overset{ind}{\sim} F(\phi_{z_{ji}}), & & i = 1, \ldots, n_j, \;\; j = 1, \ldots, 8, \quad (30)
\end{aligned}
$$

where $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1L})$, $\boldsymbol{\nu}_1 = (\nu_{11}, \ldots, \nu_{1L})$, $\boldsymbol{\nu}_2 = (\nu_{21}, \ldots, \nu_{2L})$, $\boldsymbol{\nu}_3 = (\nu_{31}, \ldots, \nu_{3L})$, $\boldsymbol{\nu}_4 = (\nu_{41}, \ldots, \nu_{4L})$, and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_L)$. With the above distributional structure, Gibbs sampling is straightforward. We use $\pi(.)$ and $\pi(. \mid -)$ to denote the prior distribution and the conditional distribution, respectively, of the parameter specified in the argument. The full

conditional distribution for the atoms is given by,

$$\pi(\{\phi_l\}_{l=1}^L \mid -) \propto \prod_{l=1}^{L} \left[\left\{ \prod_{j=1}^{8} \prod_{i=1}^{n_j} F(x_{ji} \mid \phi_l)^{\mathbb{1}(z_{ji}=l)} \right\} \pi(\phi_l) \right]. \tag{31}$$

The full conditional distributions for the latent cluster labels are given by,

$$P(z_{ji} = l \mid -) \propto \beta_{jl} F(x_{ji} \mid \phi_l), \quad l = 1, \ldots, L, \quad i = 1, \ldots, n_j \quad j = 1, \ldots, 8. \tag{32}$$

The full conditional distribution for the stick-breaking weights is given by,

$$\pi(\boldsymbol{\beta}_1 \mid -) \propto \frac{\prod_{l=1}^{L} \beta_{1l}^{m_{1l}+\frac{\alpha_1}{L}} \left\{ \beta_{2l}^{\alpha_2} \beta_{3l}^{\alpha_3} \beta_{4l}^{\alpha_4} \nu_{1l}^{\alpha_2+\alpha_4} \nu_{2l}^{\alpha_2+\alpha_3} \nu_{3l}^{\alpha_3+\alpha_4} \eta_l^{2(\alpha_2+\alpha_3+\alpha_4)} \right\}^{\beta_{1l}}}{\prod_{l=1}^{L} \left\{ \Gamma\left((\alpha_2+\alpha_4)\beta_{1l}\right) \Gamma\left((\alpha_2+\alpha_3)\beta_{1l}\right) \Gamma\left((\alpha_3+\alpha_4)\beta_{1l}\right) \Gamma\left(2(\alpha_2+\alpha_3+\alpha_4)\beta_{1l}\right) \right\}} \times \frac{1}{\prod_{l=1}^{L} \left\{ \Gamma\left(\alpha_2\beta_{1l}\right) \Gamma\left(\alpha_3\beta_{1l}\right) \Gamma\left(\alpha_4\beta_{1l}\right) \right\}}, \tag{33}$$

where $m_{1l} = \sum_{i=1}^{n_1} \mathbb{1}(z_{1i} = l)$, $l = 1, \ldots, L$. The full conditionals for $\boldsymbol{\beta}_j$, $j = 2, \ldots, 8$, are in closed form,

$$\pi(\boldsymbol{\beta}_j \mid -) \sim Dir(\boldsymbol{m}_j + \alpha_j \boldsymbol{\beta}_1), \quad \text{where } \boldsymbol{m}_j = (m_{j1}, \ldots, m_{jL}) \text{ and } m_{jl} = \sum_{i=1}^{n_j} \mathbb{1}(z_{ji} = l),\ l = 1, \ldots, L. \tag{34}$$

By letting $\boldsymbol{B}(\boldsymbol{a})$ to denote the multivariate beta function, i.e., for a L-dimensional vector $\boldsymbol{a} = (a_1, \ldots, a_L)$ with $a_i > 0$, we have,

$$\boldsymbol{B}(\boldsymbol{a}) = \frac{\prod_{l=1}^{L} \Gamma(a_l)}{\Gamma(\sum_{l=1}^{L} a_l)},$$

where $\Gamma(\cdot)$ is the gamma function. Then the full-conditional distribution of the hidden weights are given by,

$$\pi(\boldsymbol{\nu}_1 \mid -) \propto \frac{1}{\boldsymbol{B}(\alpha_5 \boldsymbol{\nu}_1)} \prod_{l=1}^{L} \left\{ \beta_{5l}^{\alpha_5 \nu_{1l}} \nu_{1l}^{(\alpha_2+\alpha_4)\beta_{1l}-1} \right\}, \tag{35}$$

$$\pi(\boldsymbol{\nu}_2 \mid -) \propto \frac{1}{\boldsymbol{B}(\alpha_6 \boldsymbol{\nu}_2)} \prod_{l=1}^{L} \left\{ \beta_{6l}^{\alpha_6 \nu_{2l}} \nu_{2l}^{(\alpha_2+\alpha_3)\beta_{1l}-1} \right\}, \tag{36}$$

$$\pi(\boldsymbol{\nu}_3 \mid -) \propto \frac{1}{\boldsymbol{B}(\alpha_7 \boldsymbol{\nu}_3)} \prod_{l=1}^{L} \left\{ \beta_{7l}^{\alpha_7 \nu_{3l}} \nu_{3l}^{(\alpha_3+\alpha_4)\beta_{1l}-1} \right\}, \tag{37}$$

$$\pi(\boldsymbol{\nu}_4 \mid -) \propto \frac{1}{\boldsymbol{B}(\alpha_8 \boldsymbol{\nu}_4)} \prod_{l=1}^{L} \left\{ \beta_{4l}^{\alpha_8 \nu_{4l}} \nu_{4l}^{(\alpha_5+\alpha_6+\alpha_7)\eta_l-1} \right\}, \tag{38}$$

$$\pi(\boldsymbol{\eta} \mid -) \propto \frac{1}{\boldsymbol{B}((\alpha_5+\alpha_6+\alpha_7)\boldsymbol{\eta})} \prod_{l=1}^{L} \left\{ \eta_l^{2(\alpha_2+\alpha_3+\alpha_4)\beta_{1l}-1} \nu_{4l}^{(\alpha_5+\alpha_6+\alpha_7)\eta_l} \right\}. \tag{39}$$

The full conditionals for the concentration parameters are given by,

$$\pi(\alpha_1 \mid -) \propto \frac{e^{-\alpha_1} \alpha_1^{\alpha_0-1} \alpha_2^{\alpha_1} \alpha_3^{\alpha_1} \alpha_4^{\alpha_1}}{\{\Gamma(\alpha_1)\}^3 \boldsymbol{B}((\alpha_1/L, \ldots, \alpha_1/L)} \prod_{l=1}^{L} \beta_{1l}^{\frac{\alpha_1}{L}} \tag{40}$$

$$\pi(\alpha_2 \mid -) \propto \frac{e^{-\alpha_2}\alpha_2^{\alpha_1-1}\alpha_5^{\alpha_2}\alpha_6^{\alpha_2}\left[\prod_{l=1}^{L}\left\{\beta_{2l}^{\beta_{1l}}\nu_{1l}^{\beta_{1l}}\nu_{2l}^{\beta_{1l}}\eta_l^{2\beta_{1l}}\right\}^{\alpha_2}\right]\Gamma(\alpha_2)\Gamma(2(\alpha_2+\alpha_3+\alpha_4))}{\prod_{l=1}^{L}\{\Gamma(\alpha_2\beta_{1l})\Gamma((\alpha_2+\alpha_4)\beta_{1l})\Gamma((\alpha_2+\alpha_3)\beta_{1l})\Gamma(2(\alpha_2+\alpha_3+\alpha_4)\beta_{1l})\}}, \tag{41}$$

$$\pi(\alpha_3 \mid -) \propto \frac{e^{-\alpha_3}\alpha_3^{\alpha_1-1}\alpha_6^{\alpha_3}\alpha_7^{\alpha_3}\left[\prod_{l=1}^{L}\left\{\beta_{3l}^{\beta_{1l}}\nu_{2l}^{\beta_{1l}}\nu_{3l}^{\beta_{1l}}\eta_l^{2\beta_{1l}}\right\}^{\alpha_3}\right]\Gamma(\alpha_3)\Gamma(2(\alpha_2+\alpha_3+\alpha_4))}{\prod_{l=1}^{L}\{\Gamma(\alpha_3\beta_{1l})\Gamma((\alpha_2+\alpha_3)\beta_{1l})\Gamma((\alpha_3+\alpha_4)\beta_{1l})\Gamma(2(\alpha_2+\alpha_3+\alpha_4)\beta_{1l})\}}, \tag{42}$$

$$\pi(\alpha_4 \mid -) \propto \frac{e^{-\alpha_4}\alpha_4^{\alpha_1-1}\alpha_5^{\alpha_4}\alpha_7^{\alpha_4}\left[\prod_{l=1}^{L}\left\{\beta_{4l}^{\beta_{1l}}\nu_{1l}^{\beta_{1l}}\nu_{3l}^{\beta_{1l}}\eta_l^{2\beta_{1l}}\right\}^{\alpha_4}\right]\Gamma(\alpha_4)\Gamma(2(\alpha_2+\alpha_3+\alpha_4))}{\prod_{l=1}^{L}\{\Gamma(\alpha_4\beta_{1l})\Gamma((\alpha_2+\alpha_4)\beta_{1l})\Gamma((\alpha_3+\alpha_4)\beta_{1l})\Gamma(2(\alpha_2+\alpha_3+\alpha_4)\beta_{1l})\}}, \tag{43}$$

$$\pi(\alpha_5 \mid -) \propto \frac{e^{-\alpha_5}\alpha_5^{\alpha_2+\alpha_4-1}\alpha_8^{\alpha_5}\left[\prod_{l=1}^{L}\{\beta_{5l}^{\nu_{1l}}\nu_{4l}^{\eta_l}\}^{\alpha_5}\right]\Gamma(\alpha_5)}{\prod_{l=1}^{L}\{\Gamma(\alpha_5\nu_{1l})\Gamma((\alpha_5+\alpha_6+\alpha_7)\nu_l)\}}, \tag{44}$$

$$\pi(\alpha_6 \mid -) \propto \frac{e^{-\alpha_6}\alpha_6^{\alpha_2+\alpha_3-1}\alpha_8^{\alpha_6}\left[\prod_{l=1}^{L}\{\beta_{6l}^{\nu_{2l}}\nu_{4l}^{\eta_l}\}^{\alpha_6}\right]\Gamma(\alpha_6)}{\prod_{l=1}^{L}\{\Gamma(\alpha_6\nu_{2l})\Gamma((\alpha_5+\alpha_6+\alpha_7)\nu_l)\}}, \tag{45}$$

$$\pi(\alpha_7 \mid -) \propto \frac{e^{-\alpha_7}\alpha_7^{\alpha_3+\alpha_4-1}\alpha_8^{\alpha_7}\left[\prod_{l=1}^{L}\{\beta_{7l}^{\nu_{3l}}\nu_{4l}^{\eta_l}\}^{\alpha_7}\right]\Gamma(\alpha_7)}{\prod_{l=1}^{L}\{\Gamma(\alpha_7\nu_{3l})\Gamma((\alpha_5+\alpha_6+\alpha_7)\nu_l)\}}, \tag{46}$$

$$\pi(\alpha_8 \mid -) \propto \frac{e^{-\alpha_8}\alpha_8^{\alpha_5+\alpha_6+\alpha_7-1}\left[\prod_{l=1}^{L}\beta_{8l}^{\alpha_8\nu_{4l}}\right]\Gamma(\alpha_8)}{\prod_{l=1}^{L}\Gamma(\alpha_8\nu_{4l})}. \tag{47}$$

Note that the full conditionals of α_j, $j = 1, \ldots, 8$, $\boldsymbol{\beta}_1, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\nu}_3, \boldsymbol{\nu}_4$, and $\boldsymbol{\eta}$ are not standard distributions that have direct samplers. We adopt a Metropolis-within-Gibbs strategy to sample from their corresponding full conditional distributions. Since α_j's are real-valued, sampling using a Metropolis step is straightforward. However, the main bottleneck in sampling are the weights $\boldsymbol{\beta}_1, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\nu}_3, \boldsymbol{\nu}_4$ and $\boldsymbol{\eta}$, which have a complex structure on the simplex. To mitigate this problem, we use the SALTSampler (Director et al., 2017) for which the implementation is publicly available as an R package.

Appendix E. Simulation details

Our simulations are designed to mimick the motivating application where we have 8 experimental groups. See Table 3 for our experimental design represented in terms of binary indicators denoting the levels of diet, treatment, and genotype. The corresponding DAG is given in Figure 12.

For our simulation study, we generated data within each of the 8 groups from a four-component mixture of bivariate Gaussian distributions with different covariance matrices for each group. Taking $\alpha_0 = 5$, we drew the concentration parameters for the different groups α_j's, the mixture model weights, $\boldsymbol{\beta}_j$'s, $\boldsymbol{\nu}_j$'s, and $\boldsymbol{\eta}_j$, and the true cluster indicators z_{ji}'s for each of the different groups using (30). Given the cluster indicators, the data were generated from the Gaussian distribution with the true cluster-specific means ϕ_l's given in Table 4 and the group-specific covariance matrices given in Table 5. Note that within each group, the same covariance matrix was used for all clusters.

In our Gibbs sampler, the truncation level of the finite mixture model was set to $L = 10$, and the base measure for GDP, G_0, was specified as the normal-inverse-Wishart distribution, $\mathcal{NIW}(\mathbf{0}, 0.01, \mathbb{I}_2, 2)$. Upon the completion of the Gibbs sampler, the clusters were estimated by using the least squares criterion (Dahl, 2006), and they were compared with the true cluster labels for evaluation. We considered a variety of sample sizes as well as a case with very imbalanced design, which are summarized in Table 6. In all cases, we ran 15,000 iterations of our Gibbs sampler and after discarding the first 5,000 samples as burn-in, we retained every 10th iteration of posterior samples.

Cluster	Mean
1	(-2, -5)
2	(0, 0)
3	(-3, 3)
4	(3, -3)

Table 4: True cluster-specific means.

Group	Covariance
1	$\begin{bmatrix} 0.8 & 0.3 \\ 0.3 & 0.8 \end{bmatrix}$
2	$\begin{bmatrix} 0.85 & 0.25 \\ 0.25 & 0.85 \end{bmatrix}$
3	$\begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}$
4	$\begin{bmatrix} 0.8 & -0.1 \\ -0.1 & 0.8 \end{bmatrix}$
5	$\begin{bmatrix} 0.8 & -0.2 \\ -0.2 & 0.9 \end{bmatrix}$
6	$\begin{bmatrix} 0.8 & 0 \\ 0 & 0.8 \end{bmatrix}$
7	$\begin{bmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{bmatrix}$
8	$\begin{bmatrix} 1.1 & 0.1 \\ 0.1 & 1.1 \end{bmatrix}$

Table 5: True covariance matrices for different groups.

Group	Sample sizes			
	small	moderate	large	unbalanced
1	40	80	150	350
2	30	70	160	30
3	30	70	180	40
4	35	75	170	45
5	25	83	155	25
6	30	88	175	25
7	25	92	185	35
8	30	88	145	35

Table 6: The sample sizes for the different groups that were used to simulate the data.

We presented the results of clustering for small sample sizes and unbalanced sample sizes in the main manuscript. Figure 13 shows the results of clustering for moderate and large sample sizes in each group.

We further considered the case, wherein the simulation scenario was difficult. We generated data within each of the 8 groups from a ten-component mixture of bivariate Gaussian distributions with different covariance matrices for each group. The choice of mixture model weights for the first fours groups are summarized in Table 7.

Group	Mixture weights $\boldsymbol{\beta}_j$
1	$(0.100, 0.100, 0.100, 0.100, 0.100, 0.100, 0.100, 0.100, 0.100, 0.100)^\top$
2	$(0.167, 0.167, 0.167, 0.167, 0.167, 0.056, 0.056, 0.056, 0.000, 0.000)^\top$
3	$(0.095, 0.095, 0.095, 0.000, 0.000, 0.143, 0.143, 0.143, 0.143, 0.143)^\top$
4	$(0.030, 0.030, 0.030, 0.182, 0.182, 0.182, 0.182, 0.182, 0.000, 0.000)^\top$

Table 7: True group-specific mixture model weights.

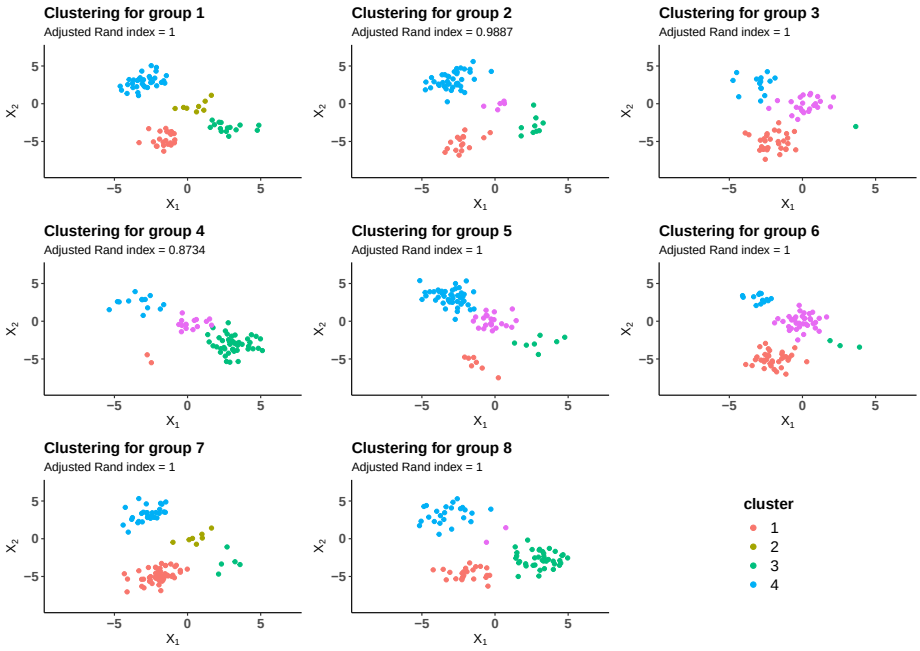
The mixture weights for all other groups were taken to be the mean of the mixture weights of their parent, e.g., the mixture weight for group 5 was the mean of the mixture weights of groups 2 and 4. The true cluster indicators z_{ji}'s for each of the different groups were drawn using (30) and the true mixture weights. Given the cluster indicators, the data were generated from the Gaussian distribution with the true cluster-specific means ϕ_l's given in Table 8 and the group-specific covariance matrices given in Table 5.

Cluster	Mean
1	(-2.5, 0)
2	(0, 0)
3	(2.5, 0)
4	(2.5, -2.5)
5	(-3, -3)
6	(2, 2)
7	(-2, 5)
8	(5, 8)
9	(-5, -8)
10	(8, -8)

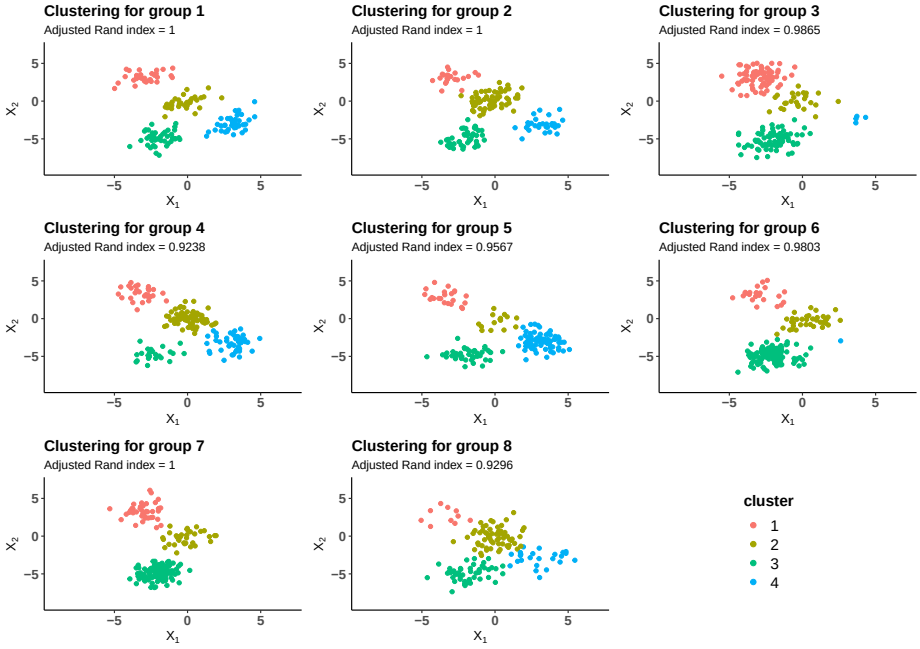
Table 8: True cluster-specific means.

In our Gibbs sampler, the truncation level of the finite mixture model was set to $L = 20$, the hyperparameter α_0 was taken to be 1, and the base measure for GDP, G_0, was specified as the normal-inverse-Wishart distribution, $\mathcal{NIW}(\mathbf{0}, 0.01, \mathbb{I}_2, 2)$. Upon the completion of the Gibbs sampler, the clusters were estimated by using the least squares criterion (Dahl, 2006), and they were compared with the true cluster labels for evaluation. We again considered a variety of sample sizes as summarized in Table 6. In all cases, we ran 25,000 iterations of our Gibbs sampler and after discarding the first 15,000 samples as burn-in, considered thinning of the samples by a factor 10. The clustering results are shown in Figure 14. Clearly, GDP was able to identify the overlapping clusters within each group and link them across groups for all simulation scenarios with reasonable accuracy as measured by the adjusted Rand indices (Hubert and Arabie, 1985) for each group (shown in the plots).

In the main manuscript, we reported the boxplot of Adjusted Rand indices for 50 replicates. Further investigation regarding the choice of α_0 revealed no significant impact in clustering

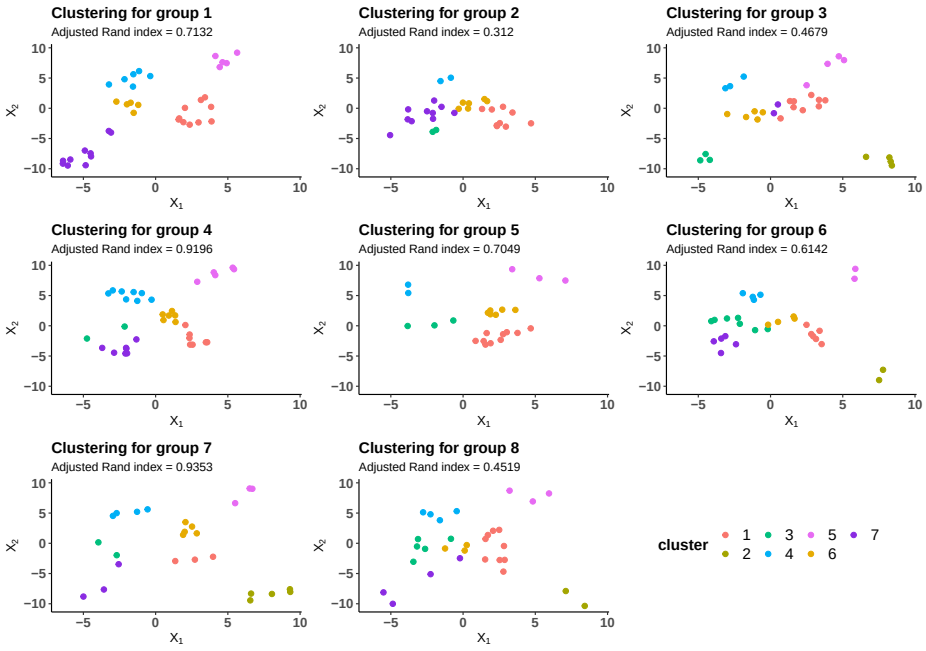


(a) Moderate sample size in each group

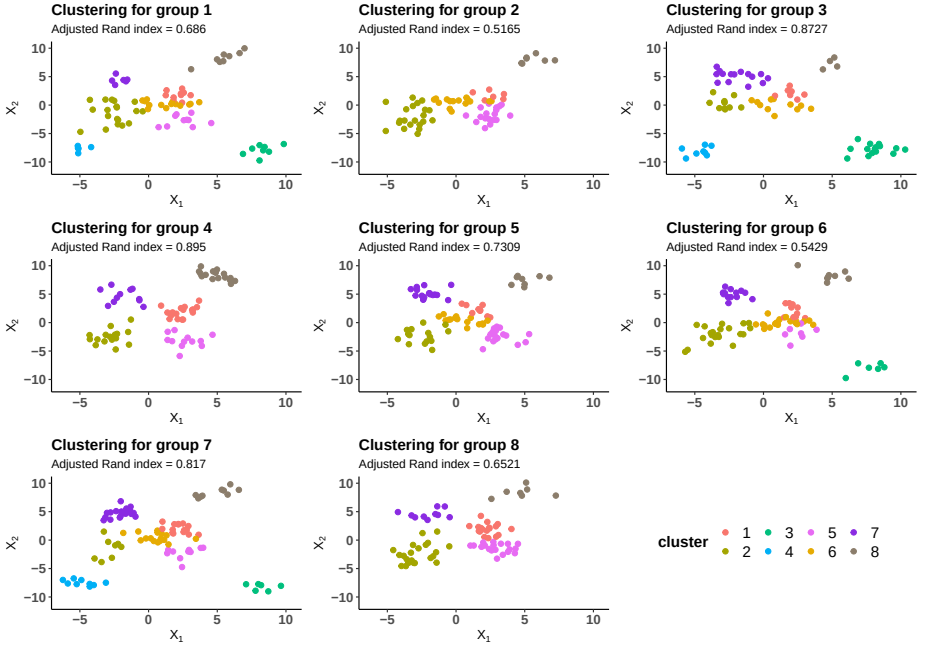


(b) Large sample size in each group

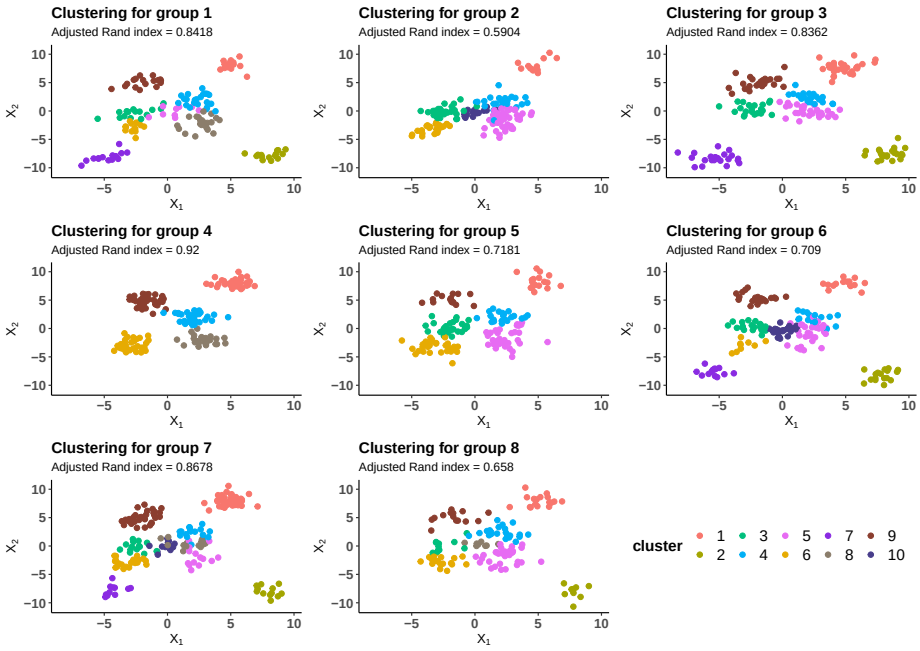
Figure 13: Clustering performance of GDP for additional sample sizes. The colors indicate the estimated clusters by GDP. Adjusted Rand index is reported at the top of each panel.



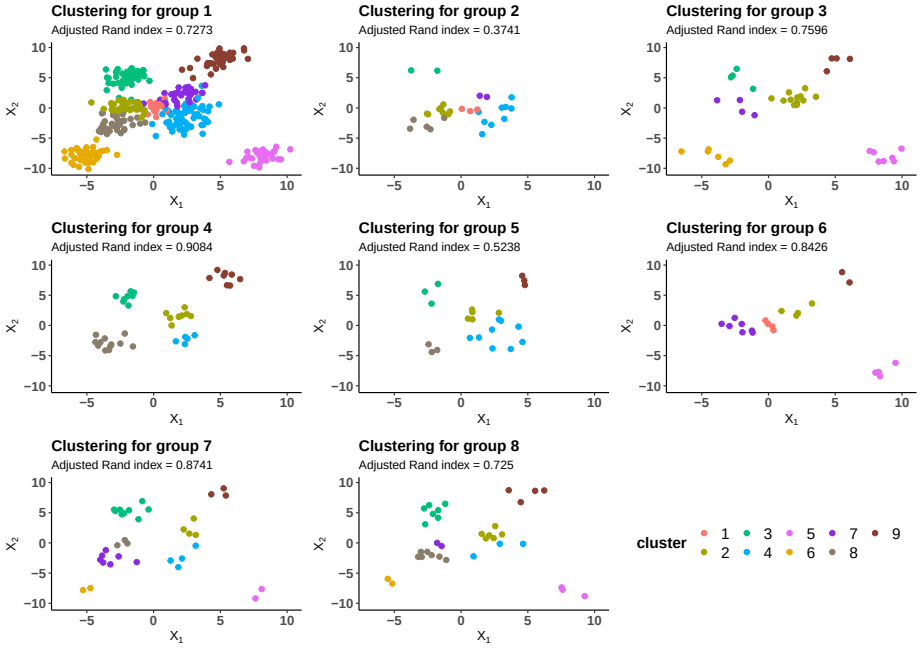
(a) Small sample size in each group



(b) Moderate sample size in each group



(c) Large sample size in each group



(d) Unbalanced sample size in each group

Figure 14: Clustering performance of GDP for various sample sizes and difficult simulation scenario. The colors indicate the estimated clusters by GDP. Adjusted Rand index is reported at the top of each panel.

performance. Figure 15 shows that boxplot of Adjusted Rand indices for 50 replicates, comparing GDP, HDP, and k-means with α_0 taken to be 6. In all situations, GDP uniformly out-performed the other two methods.

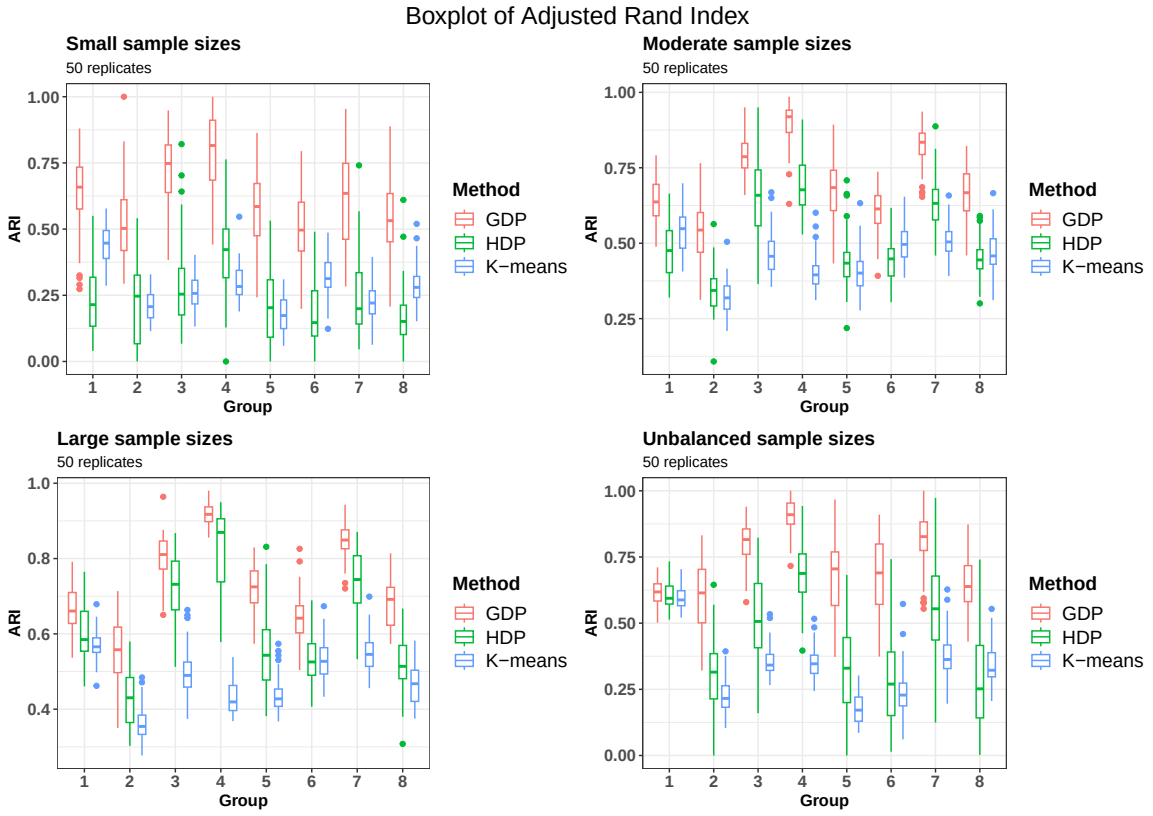


Figure 15: The boxplots of the Adjusted Rand indices for GDP, HDP, and k-means for all sample sizes. In all simulations α_0 was taken to be 6.

Appendix F. Additional simulations

In the main manuscript, we presented simulations to mimick the motivating application where we have 8 experimental groups. However, there are several motivating applications of the proposed GDP. One such example is time-series data. One might be interested in clustering stocks based on daily prices for each year. Each calendar year is then a group. The groups naturally have time dependence (i.e., one does not expect the clustering of stocks to change dramatically in consecutive years), which may be represented by an autoregressive (AR) model. AR model is one type of DAG model. Particularly, for an AR model with lag 1, the time dependencies can be represented by a simple DAG (see Figure 16), which may be analyzed by the GDP. With this specific DAG (chain DAG), the GDP is given by,

$$G_1 \mid \alpha_1, G_0 \sim DP(\alpha_1, G_0), \tag{48}$$

$$G_t \mid \alpha_t, G_{t-1} \sim DP(\alpha_t, G_{t-1}), \quad t = 2, \ldots, T, \tag{49}$$

where T denotes the total number of observed time points. Let x_{ti} denote the observation i from time point t and θ_{ti} denote the parameter specifying the mixture component associated with the corresponding observation. Let $F(\theta_{ti})$ denote the distribution of x_{ti} given θ_{ti} and G_t denote a prior distribution for θ_{ti}. The group-specific mixture model is given by,

$$\begin{aligned} \theta_{ti} \mid G_t &\overset{ind}{\sim} G_t, \\ x_{ti} \mid \theta_{ti} &\overset{ind}{\sim} F(\theta_{ti}), \end{aligned} \tag{50}$$

where G_t follows (48) and (49). The observations $x_{t1}, x_{t2}, \ldots, x_{tn_t}$ are exchangeable at each observed time point t but the groups (formed at the different time points) are not exchangeable due to the time-dependency between the groups. The corresponding GDP mixture model can be derived as the infinite limit of a finite mixture model. Let us denote the mixture component associated with the observation x_{ti} from time point t, by z_{ti}. Suppose $\boldsymbol{\beta}_1$ is the vector of mixing weights for the root node (corresponding to time point $t = 1$). Denoting by $\boldsymbol{\beta}_t$ the mixing weights of node t (corresponding to time point t), we consider a finite mixture version of the proposed GDP,

$$\begin{aligned} \alpha_1 \mid \alpha_0 &\sim Gamma(\alpha_0, 1), \\ \boldsymbol{\beta}_1 \mid \alpha_1 &\sim Dir\left(\alpha_1/L, \ldots, \alpha_1/L\right), \\ \alpha_t \mid \alpha_{t-1} &\sim Gamma(\alpha_{t-1}, 1), \\ \boldsymbol{\beta}_t \mid \alpha_t, \boldsymbol{\beta}_{t-1} &\sim Dir\left(\alpha_t \left(\beta_{t1}, \ldots, \beta_{tL}\right)\right), \\ \phi_l \mid G_0 &\sim G_0, \\ z_{ti} \mid \boldsymbol{\beta}_t &\sim \boldsymbol{\beta}_t, \\ x_{ti} \mid z_{ti}, (\phi_l)_{l=1}^{L} &\sim F\left(\phi_{z_{ti}}\right), \quad i = 1, \ldots, n_t, \ t = 1, \ldots T. \end{aligned} \tag{51}$$

We considered simple simulation examples, where we analyzed time-dependent observations (Figure 16) for $T = 10, 20$, and 50 time points. We generated data within each of the T time points (groups) from a five-component mixture of bivariate Gaussian distributions. Taking $\alpha_0 = 15$, we drew the concentration parameters for the different groups α_t's, the mixture model weights, $\boldsymbol{\beta}_t$'s, and the true cluster indicators z_{ti}'s for each of the different groups using (51). Given the cluster indicators, the data were generated from the Gaussian distribution with the true cluster-specific means ϕ_l's given in Table 9 and the same covariance matrix $\begin{bmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{bmatrix}$ across clusters, which was assumed to be known for simplicity. Furthermore, we considered 100 observations at each time point t, i.e., $n_t = 100$, $t = 1, \ldots, T$. In our Gibbs sampler, the truncation level of the finite mixture model was set to $L = 10$, and the base measure for GDP, G_0, was specified as the normal distribution, $\mathcal{N}(\mathbf{0}, 0.01^{-1}\mathbb{I}_2)$. We ran our MCMC for $50,000$ iterations, discarded the first $35,000$ iterations as burn-in, and retained every 15th posterior sample. Upon the completion of the Gibbs sampler, the clusters were estimated by using the least squares criterion (Dahl, 2006), and they were compared with the true cluster labels for evaluation. Figures 17 - 19 show the clustering plots for $T = 10, 20$, and 50 time points respectively. Clearly, our model was able to identify the clusters within each group and link them across groups (time points) with good accuracy as measured by adjusted Rand indices (ARI) for each group (shown in the plots). Furthermore, the traceplots of the log-likelihood showed no lack of convergence (Figure 20). Additionally, we considered 50 independent replications to investigate the runtime of our MCMC for varying number of nodes, T. Figure 21 shows that the runtime is approximately linear in the number of nodes T, for fixed truncation level L of the GDP. We further compared the clustering performance for the time-dependent grouped data using HDP. Figures 22 - 24 show the corresponding clustering plots corresponding to $T = 10, 20$, and 50 time points respectively. As before, HDP fails to capture meaningful clusters across the non-exchangeable groups, as indicated by the low ARI (shown in the plots).

Cluster	Mean
1	(-2, -5)
2	(0, 0)
3	(-3, 3)
4	(3, -3)
5	(8, 5)

Table 9: True cluster-specific means.

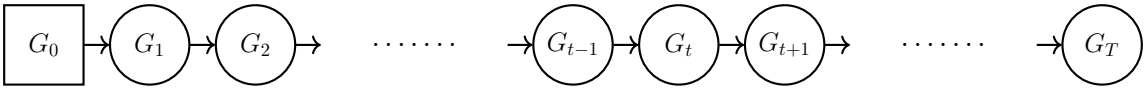


Figure 16: The DAG denoting time-dependency.

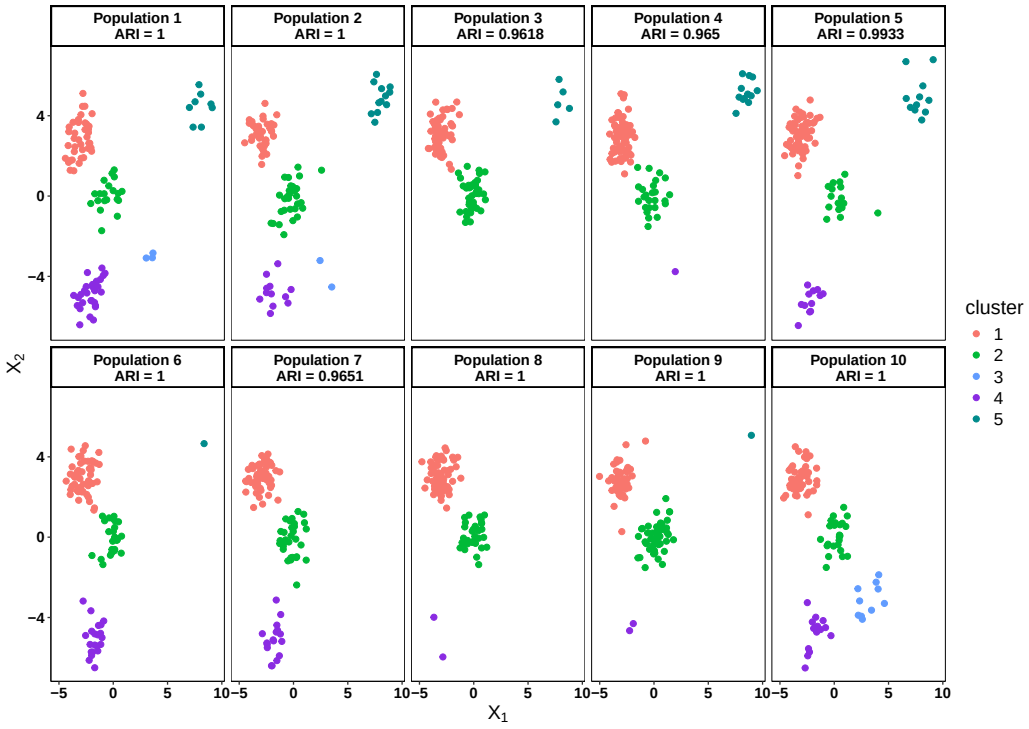


Figure 17: Clustering performance of time-dependent GDP for $T = 10$ time points. Population t refers to the observed group at time point t. The colors indicate the estimated clusters by GDP. Adjusted Rand index is reported at the top of each panel.

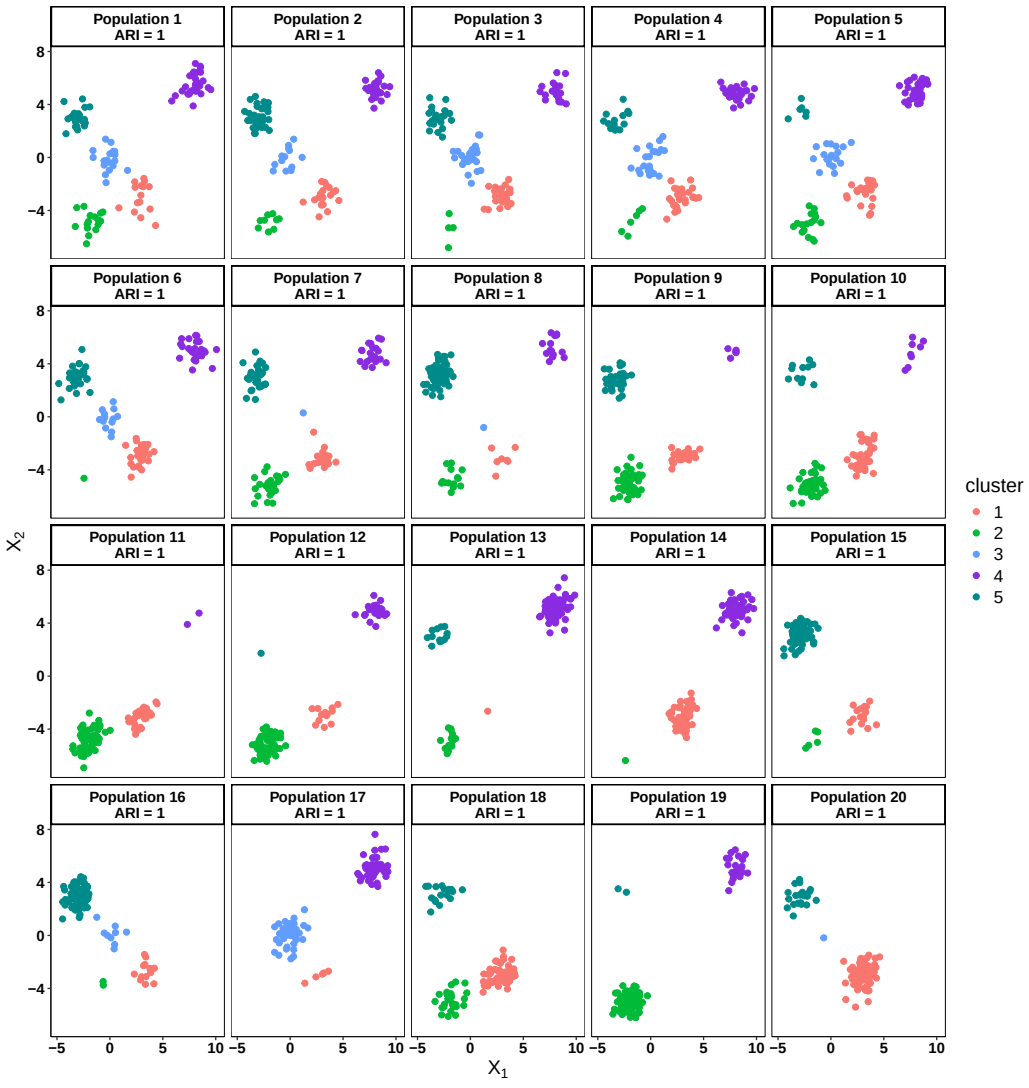


Figure 18: Clustering performance of time-dependent GDP for $T = 20$ time points. Population t refers to the observed group at time point t. The colors indicate the estimated clusters by GDP. Adjusted Rand index is reported at the top of each panel.

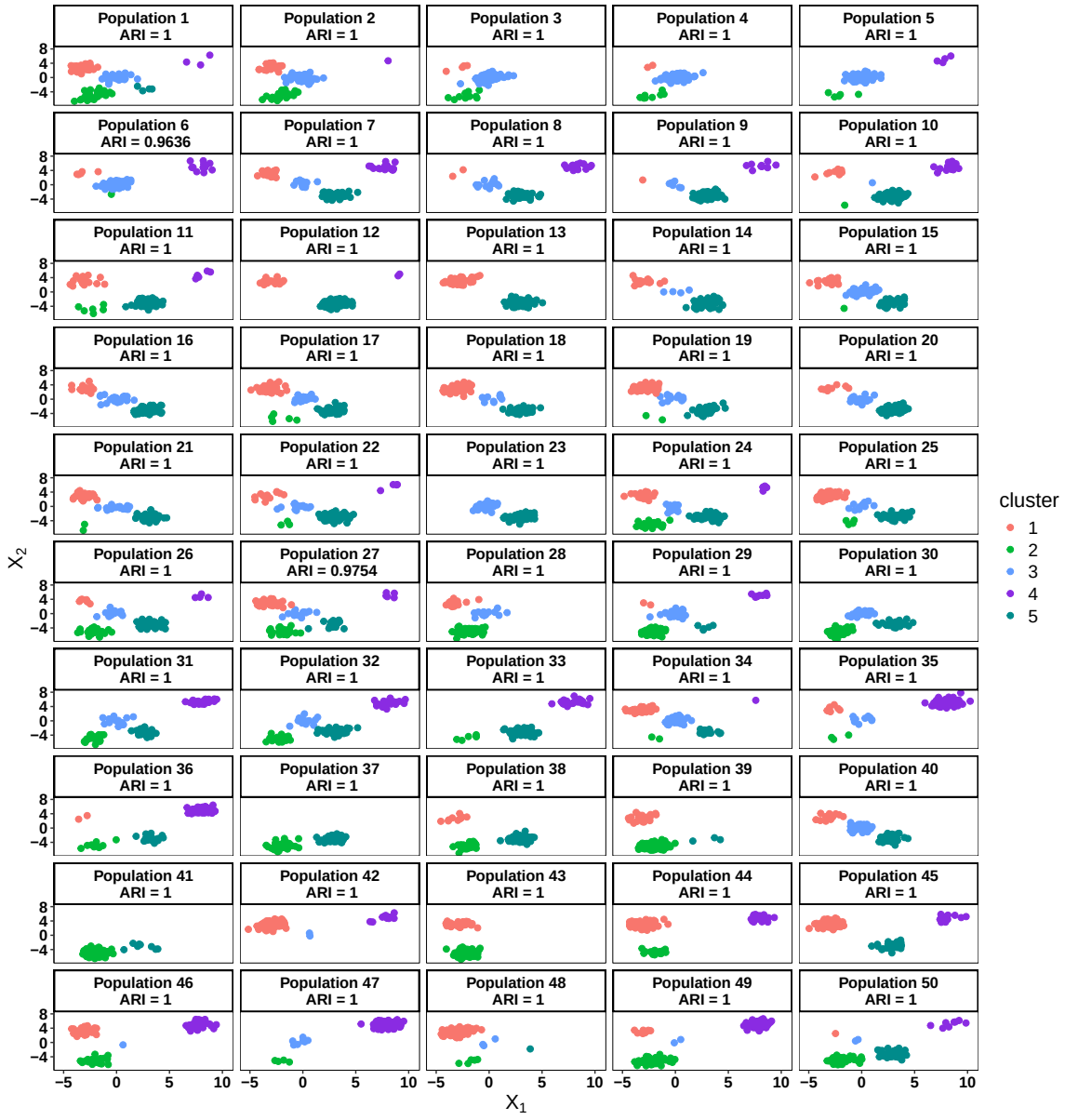
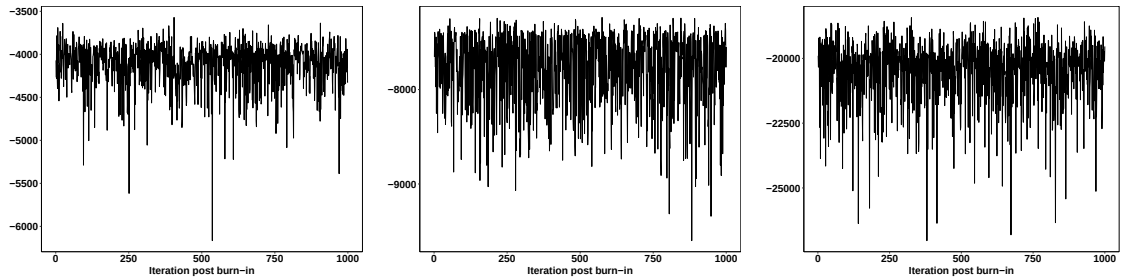


Figure 19: Clustering performance of time-dependent GDP for $T = 50$ time points. Population t refers to the observed group at time point t. The colors indicate the estimated clusters by GDP. Adjusted Rand index is reported at the top of each panel.



(a) $T = 10$ time points. (b) $T = 20$ time points. (c) $T = 50$ time points.

Figure 20: Traceplot of log-likelihood post burn-in and thinning for the varying number of nodes (T) in the chain DAG.

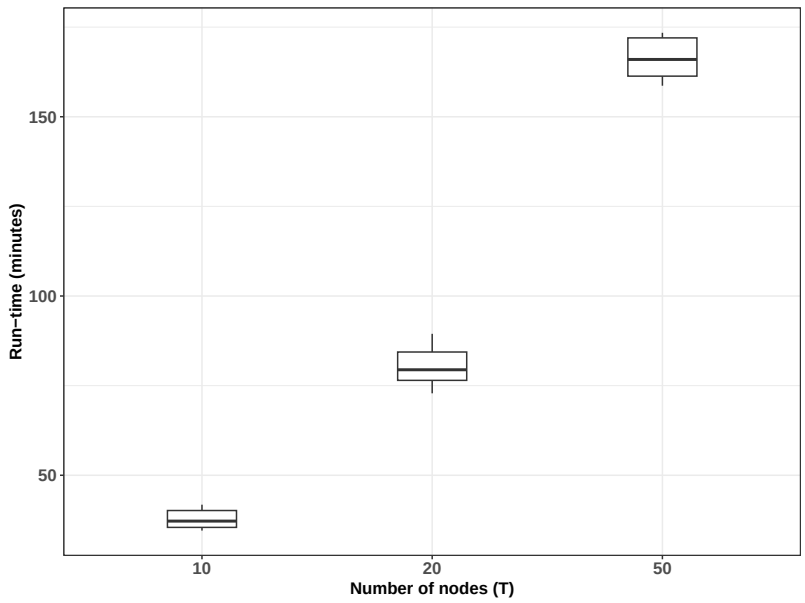


Figure 21: Runtime of time-dependent GDP for varying number of time points (nodes T). The truncation level of proposed GDP is fixed at $L = 10$. Boxplots show variation across 50 independent replicates.

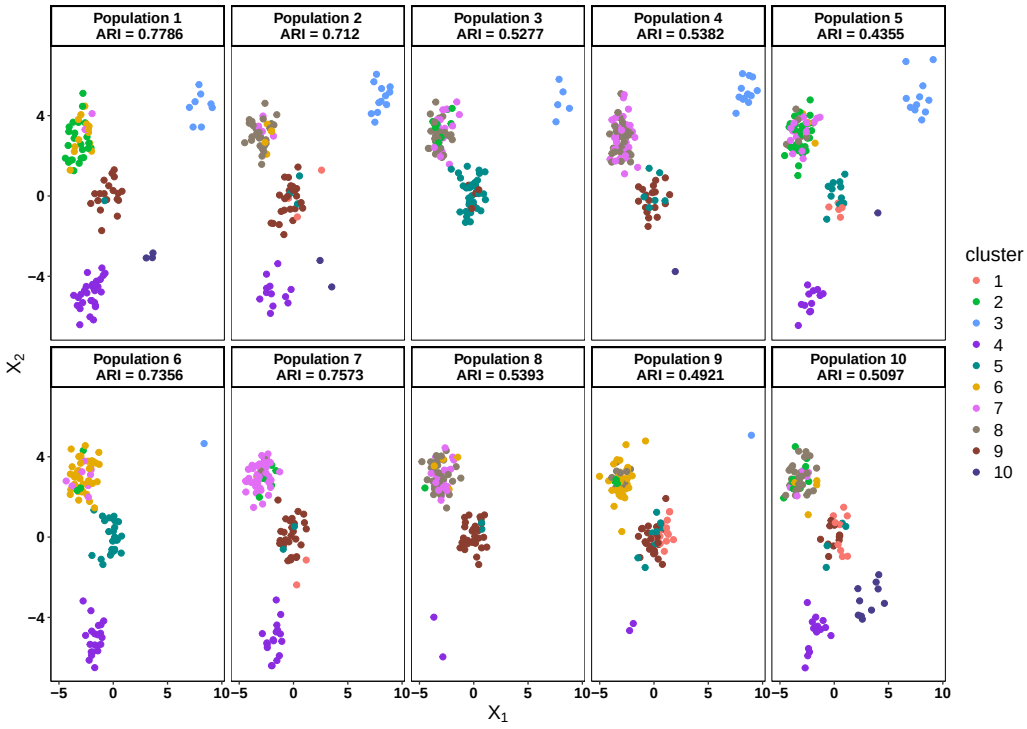


Figure 22: Clustering performance of time-dependent data using HDP for $T = 10$ time points. Population t refers to the observed group at time point t. The colors indicate the estimated clusters by HDP. Adjusted Rand index is reported at the top of each panel.

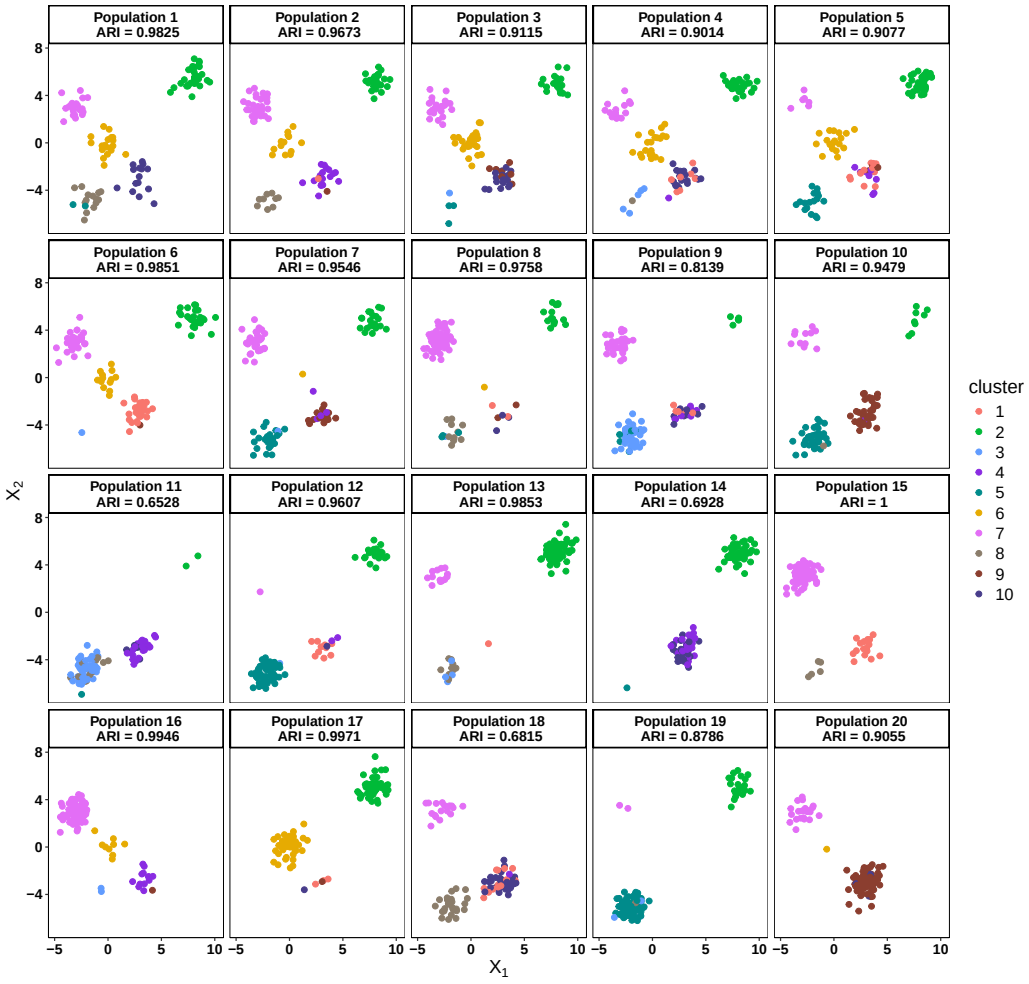


Figure 23: Clustering performance of time-dependent data using HDP for $T = 20$ time points. Population t refers to the observed group at time point t. The colors indicate the estimated clusters by HDP. Adjusted Rand index is reported at the top of each panel.

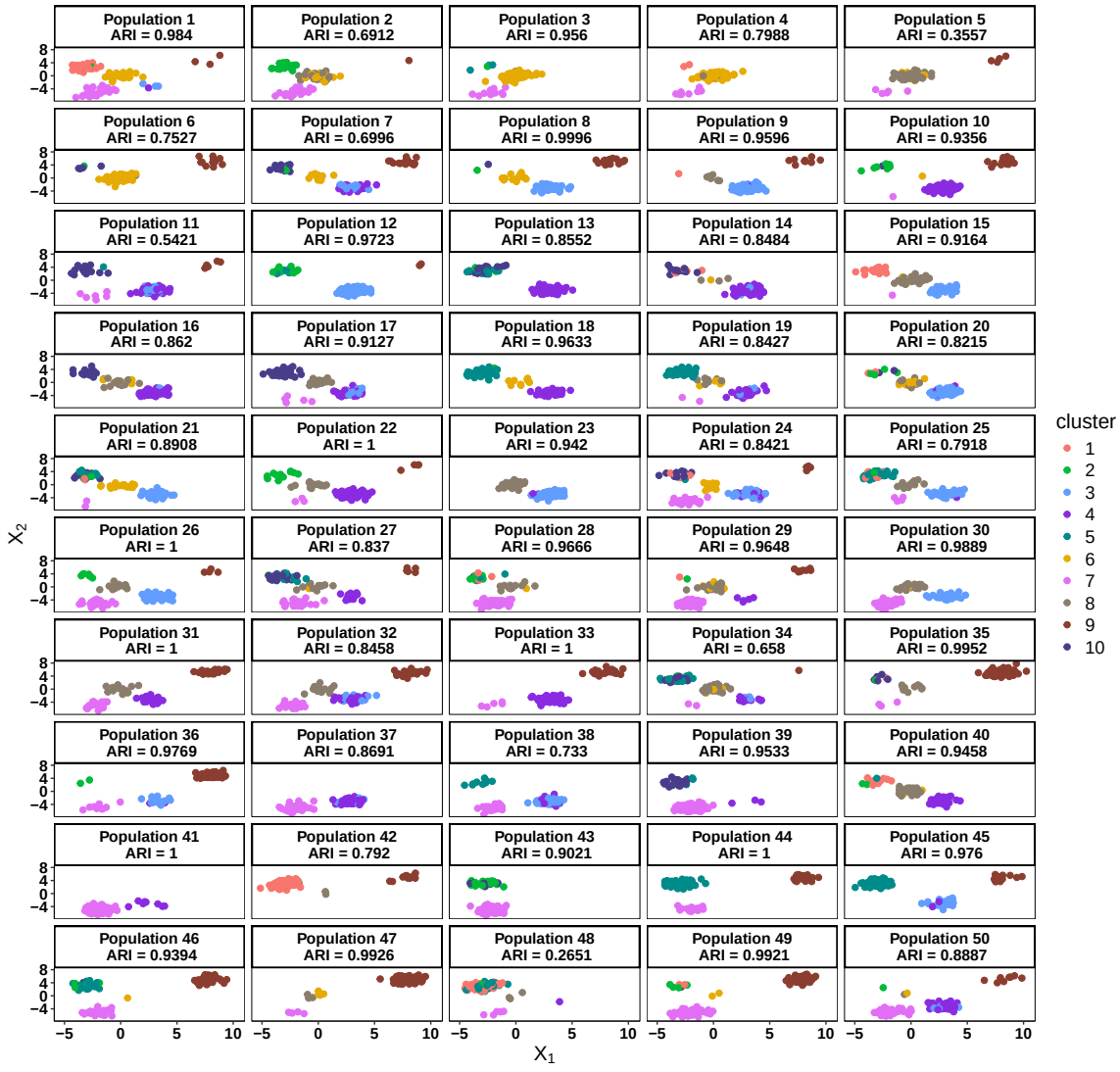
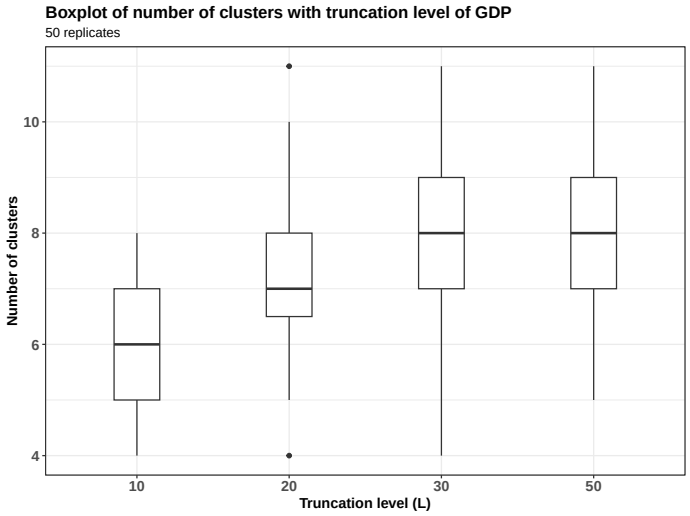


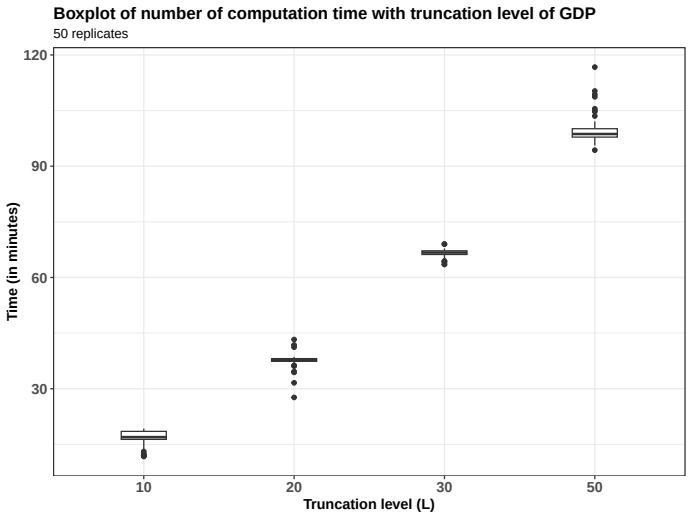
Figure 24: Clustering performance of time-dependent data using HDP for $T = 50$ time points. Population t refers to the observed group at time point t. The colors indicate the estimated clusters by HDP. Adjusted Rand index is reported at the top of each panel.

Appendix G. Real Data Analysis plots

Sensitivity. To study the effect of the truncation level of GDP, we varied $L = 10, 20, 30,$ and 50. We considered 50 independent replications and studied the estimated number of clusters for the different choices of the truncation level, L. The boxplots of the number of estimated clusters in Figure 25a shows that our method is relatively robust with respect to the truncation level, especially for $L = 30, 50$. Furthermore, Figure 25b shows that the runtime of our sampler is approximately linear in the truncation level of GDP. These results led us to consider the truncation level $L = 30$ for the real data analysis using GDP (the estimated number of clusters is well below 30), as reported in the main manuscript.



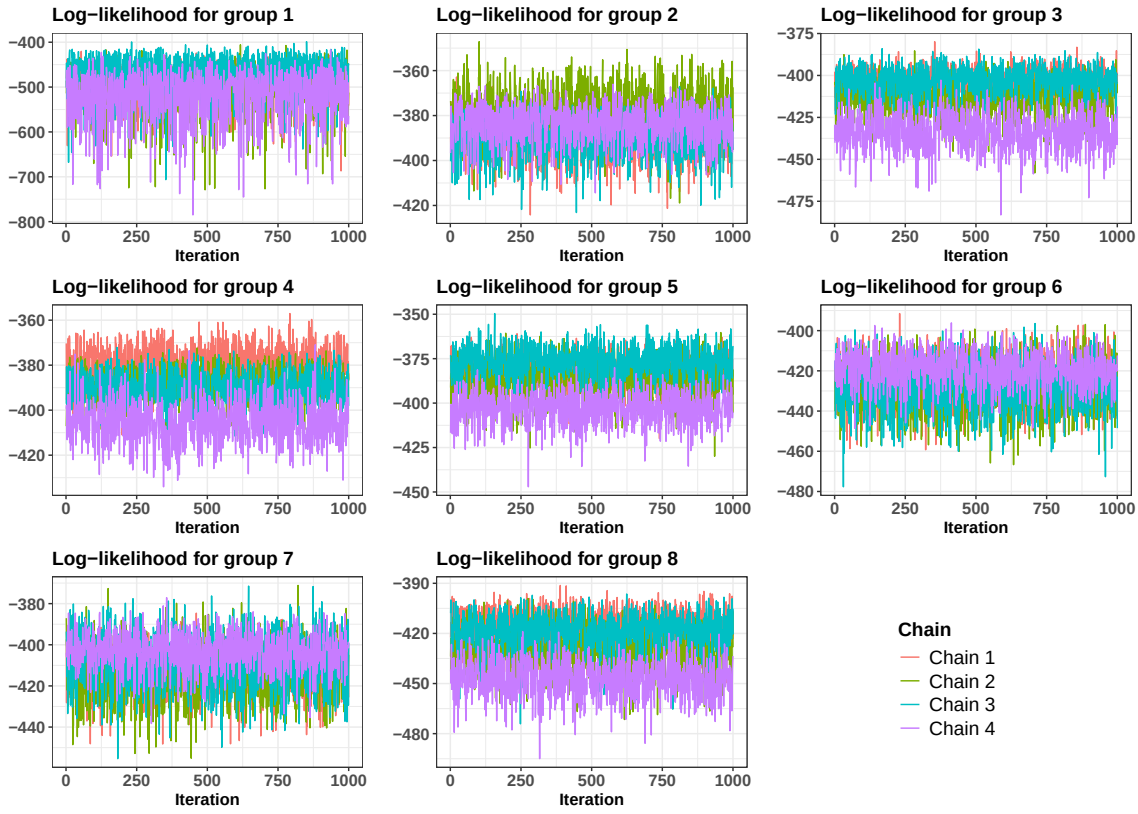
(a) Number of estimated clusters.



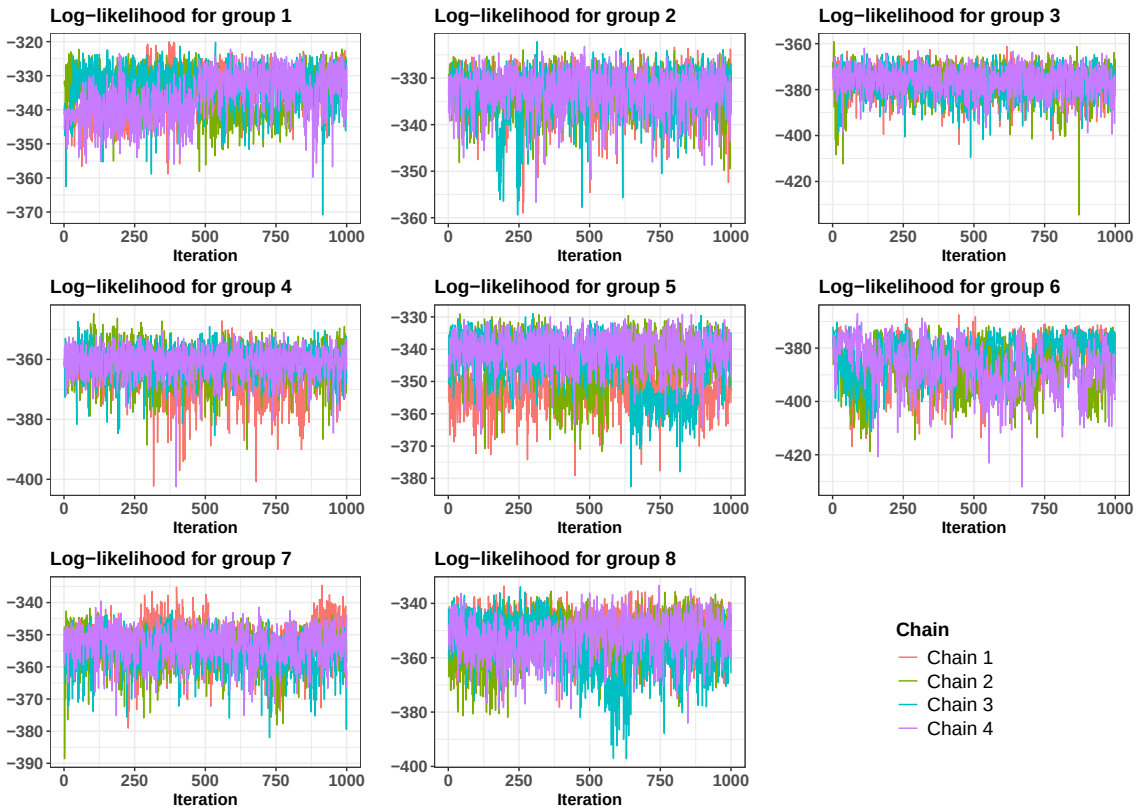
(b) Computational time.

Figure 25: Robustness and scalability of GDP for various choices of truncation level (L) for the real data over 50 independent replications.

For our real data analysis, we ran four parallel chains of the Gibbs sampler for 50,000 iterations. The traceplots (Figure 26) of the log-likelihood for each of the four parallel chains of our sampler, after discarding the initial 35,000 samples and thinning the samples by a factor of 15 indicated no lack of convergence of our sampler. Furthermore, the traceplots indicate the presence of local modes, necessitating the need to concatenate posterior samples across these chains for more efficient and reliable inference.



(a) GDP



(b) HDP

Figure 26: Group-specific traceplots of log-likelihood for four parallel chains of our MCMC for (a) GDP and (b) HDP.

References

Md. Hijbul Alam, Jaakko Peltonen, Jyrki Nummenmaa, and Kalervo Järvelin. Tree-structured hierarchical dirichlet process. In Sara Rodríguez, Javier Prieto, Pedro Faria, Sławomir Kłos, Alberto Fernández, Santiago Mazuelas, M. Dolores Jiménez-López, María N. Moreno, and Elena M. Navarro, editors, *Distributed Computing and Artificial Intelligence, Special Sessions, 15th International Conference*, pages 291–299, Cham, 2019. Springer International Publishing. ISBN 978-3-319-99608-0.

Charles E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974. ISSN 00905364. URL `http://www.jstor.org/stable/2958336`.

Ernesto Barrios, Antonio Lijoi, Luis E Nieto-Barajas, and Igor Prünster. Modeling with normalized random measure mixture models. *Statistical Science*, 28(3):313–334, 2013.

D. Basu. On statistics independent of a complete sufficient statistic. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 15(4):377–380, 1955. ISSN 00364452. URL `http://www.jstor.org/stable/25048259`.

Mario Beraha, Alessandra Guglielmi, and Fernando A Quintana. The semi-hierarchical dirichlet process and its application to clustering homogeneous distributions. *Bayesian Analysis*, 16(4): 1187–1219, 2021.

Dehua Bi and Yuan Ji. A class of dependent random distributions based on atom skipping, 2023.

Annie Bouchard-Mercier, Ann-Marie Paradis, Iwona Rudkowska, Simone Lemieux, Patrick Couture, and Marie-Claude Vohl. Associations between dietary patterns and gene expression profiles of healthy men and women: a cross-sectional study. *Nutrition Journal*, 12(1):24, 2013. doi: 10.1186/1475-2891-12-24. URL `https://doi.org/10.1186/1475-2891-12-24`.

Federico Camerlenghi, David B. Dunson, Antonio Lijoi, Igor Prünster, and Abel Rodríguez. Latent Nested Nonparametric Priors (with Discussion). *Bayesian Analysis*, 14(4):1303 – 1356, 2019a. doi: 10.1214/19-BA1169. URL `https://doi.org/10.1214/19-BA1169`.

Federico Camerlenghi, Antonio Lijoi, Peter Orbanz, and Igor Prünster. Distribution theory for hierarchical processes. *The Annals of Statistics*, 47(1):67 – 92, 2019b. doi: 10.1214/17-AOS1678. URL `https://doi.org/10.1214/17-AOS1678`.

Noirrit Kiran Chandra, Abhra Sarkar, John F. de Groot, Ying Yuan, and Peter Müller. Bayesian nonparametric common atoms regression for generating synthetic controls in clinical trials. *Journal of the American Statistical Association*, 118(544):2301–2314, 2023. doi: 10.1080/01621459.2023.2231581. URL `https://doi.org/10.1080/01621459.2023.2231581`.

Junsouk Choi, Robert Chapkin, and Yang Ni. Bayesian causal structural learning with zero-inflated poisson bayesian networks. *Advances in neural information processing systems*, 33:5887–5897, 2020.

D Cifarelli and E Regazzini. Problemi statistici non parametrici in condizioni di scambiabilita parziale e impiego di medie associative. Technical report, Tech. rep., Quaderni Istituto Matematica Finanziaria dell'Universita di Torino, 1978.

David. B. Dahl. Model-based clustering for expression data via a dirichlet process mixture model. *Bayesian Inference for Gene Expression and Proteomics*, 2006.

David B. Dahl, Ryan Day, and Jerry W. Tsai. Random partition distribution indexed by pairwise information. *Journal of the American Statistical Association*, 112(518):721–732, 2017. doi: 10.1080/01621459.2016.1165103. URL https://doi.org/10.1080/01621459.2016.1165103. PMID: 29276318.

Laura D'Angelo and Francesco Denti. A finite-infinite shared atoms nested model for the bayesian analysis of large grouped data, 2024. URL https://arxiv.org/abs/2406.13310.

Laura D'Angelo, Antonio Canale, Zhaoxia Yu, and Michele Guindani. Bayesian nonparametric analysis for the detection of spikes in noisy calcium imaging data. *Biometrics*, 79(2):1370–1382, 2023. doi: https://doi.org/10.1111/biom.13626. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13626.

Snigdha Das, Yabo Niu, Yang Ni, Bani K. Mallick, and Debdeep Pati. Blocked gibbs sampler for hierarchical dirichlet processes. *Journal of Computational and Graphical Statistics*, 0(ja):1–31, 2024. doi: 10.1080/10618600.2024.2388543. URL https://doi.org/10.1080/10618600.2024.2388543.

Pierpaolo De Blasi, Stefano Favaro, Antonio Lijoi, Ramsés H Mena, Igor Prünster, and Matteo Ruggiero. Are gibbs-type priors the most natural generalization of the dirichlet process? *IEEE transactions on pattern analysis and machine intelligence*, 37(2):212–229, 2013.

B. de Finetti. Sur la condition d'equivalence partielle. *Actual. Sci. Ind.*, 739:5–18, 1938.

Maria De Iorio, Wesley O Johnson, Peter Müller, and Gary L Rosner. Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*, 65(3):762–771, 2009.

Francesco Denti, Federico Camerlenghi, Michele Guindani, and Antonietta Mira. A common atoms model for the bayesian nonparametric analysis of nested data. *Journal of the American Statistical Association*, 118(541):405–416, 2023.

Debangan Dey, Abhirup Datta, and Sudipto Banerjee. Graphical gaussian process models for highly multivariate spatial data. *Biometrika*, 109(4):993–1014, 2022.

Hannah M. Director, James Gattiker, Earl Lawrence, and Scott Vander Wiel. Efficient sampling on the simplex with a self-adjusting logit transform proposal. *Journal of Statistical Computation and Simulation*, 87(18):3521–3536, 2017. doi: 10.1080/00949655.2017.1376063. URL https://doi.org/10.1080/00949655.2017.1376063.

David Dunson and Amy Herring. Bayesian latent variable models for mixed discrete outcomes. *Biostatistics (Oxford, England)*, 6:11–25, 02 2005. doi: 10.1093/biostatistics/kxh025.

Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995. doi: 10.1080/01621459.1995.10476550. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476550.

Xiuqin Fan, Hongyang Yao, Xuanyi Liu, Qiaoyu Shi, Liang Lv, Ping Li, Rui Wang, Tiantian Tang, and Kemin Qi. High-fat diet alters the expression of reference genes in male mice. *Frontiers in Nutrition*, 7, 2020. ISSN 2296-861X. doi: 10.3389/fnut.2020.589771. URL https://www.frontiersin.org/articles/10.3389/fnut.2020.589771.

Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973. doi: 10.1214/aos/1176342360. URL https://doi.org/10.1214/aos/1176342360.

Nicholas J Foti and Sinead A Williamson. A survey of non-exchangeable priors for bayesian nonparametric models. *IEEE transactions on pattern analysis and machine intelligence*, 37(2): 359–371, 2013.

Alan E. Gelfand, Athanasios Kottas, and Steven N. MacEachern. Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association*, 100(471): 1021–1035, 2005. ISSN 01621459. URL http://www.jstor.org/stable/27590632.

Rebecca Graziani, Michele Guindani, and Peter F Thall. Bayesian nonparametric estimation of targeted agent effects on biomarker change to predict clinical outcome. *Biometrics*, 71(1):188–197, 2015.

J. E. Griffin and M. F. J. Steel. Order-based dependent dirichlet processes. *Journal of the American Statistical Association*, 101(473):179–194, 2006. ISSN 01621459. URL http://www.jstor.org/stable/30047448.

Yuqi Gu and David B Dunson. Bayesian pyramids: identifiable multilayer discrete latent structure models for discrete data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):399–426, 2023.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zagar, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar B. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 2021. doi: 10.1016/j.cell.2021.04.048. URL https://doi.org/10.1016/j.cell.2021.04.048.

Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G Walker. *Bayesian nonparametrics*, volume 28. Cambridge University Press, 2010.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. doi: 10.1007/BF01908075. URL https://doi.org/10.1007/BF01908075.

Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell rna sequencing technologies and bioinformaticspipelines. *Experimental & Molecular Medicine*, 50(8):1–14, 2018. doi: 10.1038/s12276-018-0071-8. URL https://doi.org/10.1038/s12276-018-0071-8.

Maria De Iorio, Peter Müller, Gary L Rosner, and Steven N MacEachern. An anova model for dependent random measures. *Journal of the American Statistical Association*, 99 (465):205–215, 2004. doi: 10.1198/016214504000000205. URL https://doi.org/10.1198/016214504000000205.

Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001. doi: 10.1198/016214501750332758. URL https://doi.org/10.1198/016214501750332758.

Hemant Ishwaran and Mahmoud Zarepour. Exact and approximate sum representations for the dirichlet process. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 30 (2):269–283, 2002. ISSN 03195724. URL http://www.jstor.org/stable/3315951.

Donald B. Jump and Steven D. Clarke. Regulation of gene expression by dietary fat. *Annual Review of Nutrition*, 19(1):63–90, 1999. doi: 10.1146/annurev.nutr.19.1.63. URL https://doi.org/10.1146/annurev.nutr.19.1.63. PMID: 10448517.

Olav Kallenberg. *Probabilistic symmetries and invariance principles.* Probability and its Applications (New York). Springer, New York, 2005. ISBN 978-0387-25115-8; 0-387-25115-4.

Ken P. Kleinman and Joseph G. Ibrahim. A semiparametric bayesian approach to the random effects model. *Biometrics*, 54(3):921–938, 1998. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2533846.

Antonio Lijoi, Igor Prünster, and Giovanni Rebaudo. Flexible clustering via hidden hierarchical dirichlet priors. *Scandinavian Journal of Statistics*, feb 2022. doi: 10.1111/sjos.12578. URL https://doi.org/10.1111%2Fsjos.12578.

Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining*, pages 911–916, 2010. doi: 10.1109/ICDM.2010.35.

S. N. MacEachern. Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA, 1999. American Statistical Association.

S. N. MacEachern. Dependent dirichlet processes. Technical report, Department of Statistics, The Ohio State University, 2000.

Steven N. MacEachern and Peter Müller. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998. ISSN 10618600. URL http://www.jstor.org/stable/1390815.

Bani K. Mallick and Stephen G. Walker. Combining information from several experiments with nonparameter priors. *Biometrika*, 84(3):697–706, 1997. ISSN 00063444. URL http://www.jstor.org/stable/2337589.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018. URL https://arxiv.org/abs/1802.03426.

Patrice J. Morin, Andrew B. Sparks, Vladimir Korinek, Nick Barker, Hans Clevers, Bert Vogelstein, and Kenneth W. Kinzler. Activation of β-catenin-tcf signaling in colon cancer by mutations in β-catenin or apc. *Science*, 275(5307):1787–1790, 1997. doi: 10.1126/science.275.5307.1787. URL https://www.science.org/doi/abs/10.1126/science.275.5307.1787.

Peter Müller, Fernando Andrés Quintana, Alejandro Jara, and Tim Hanson. *Bayesian nonparametric data analysis*, volume 1. Springer, 2015.

Peter Müller, Fernando Quintana, and Gary Rosner. A method for combining inference across related nonparametric bayesian models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66(3):735–749, 2004. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/3647503.

Luis E Nieto-Barajas and Alberto Contreras-Cristán. A bayesian nonparametric approach for time series clustering. *Bayesian Analysis*, 9(1):147–170, 2014.

Judea Pearl. *Causality.* Cambridge university press, 2009.

Jim Pitman. Poisson–dirichlet and gem invariant distributions for split-and-merge transformations of an interval partition. *Comb. Probab. Comput.*, 11(5):501–514, sep 2002. ISSN 0963-5483. doi: 10.1017/S0963548302005163. URL https://doi.org/10.1017/S0963548302005163.

Fernand A. Quintana, Peter Mueller, Alejandro Jara, and Steven N. MacEachern. The dependent dirichlet process and related models, 2020. URL https://arxiv.org/abs/2007.06129.

Lu Ren, David B Dunson, and Lawrence Carin. The dynamic hierarchical dirichlet process. In *Proceedings of the 25th international conference on machine learning*, pages 824–831, 2008.

Abel Rodriguez and David B. Dunson. Functional clustering in nested designs: Modeling variability in reproductive epidemiology studies. *The Annals of Applied Statistics*, 8(3):1416 – 1442, 2014. doi: 10.1214/14-AOAS751. URL https://doi.org/10.1214/14-AOAS751.

Abel Rodríguez, David B Dunson, and Alan E Gelfand. The nested dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154, 2008. doi: 10.1198/016214508000000553. URL https://doi.org/10.1198/016214508000000553.

Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24305538.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Nathan Srebro and Sam Roweis. Time-varying topic models using dependent dirichlet processes. *UTML, TR# 2005*, 3, 2005.

Yee Whye Teh. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992, 2006.

Yee Whye Teh and Michael I Jordan. Hierarchical bayesian nonparametric models with applications. *Bayesian nonparametrics*, 1:158–207, 2010.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. doi: 10.1198/016214506000000302. URL https://doi.org/10.1198/016214506000000302.

Romain Thibaux and Michael I Jordan. Hierarchical beta processes and the indian buffet process. In *Artificial intelligence and statistics*, pages 564–571. PMLR, 2007.

Conor Tillinghast, Zheng Wang, and Shandian Zhe. Nonparametric sparse tensor factorization with hierarchical gamma processes. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 21432–21448. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/tillinghast22a.html.

Sinead Williamson, Avinava Dubey, and Eric Xing. Parallel markov chain monte carlo for nonparametric mixture models. In *International Conference on Machine Learning*, pages 98–106. PMLR, 2013.

Jianwen Zhang, Yangqiu Song, Changshui Zhang, and Shixia Liu. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1079–1088, 2010.

Mingyuan Zhou. Nonparametric bayesian negative binomial factor analysis. *Bayesian Analysis*, 2016. URL https://api.semanticscholar.org/CorpusID:51995146.

Daiane Aparecida Zuanetti, Peter Müller, Yitan Zhu, Shengjie Yang, and Yuan Ji. Clustering distributions with the marginalized nested dirichlet process. *Biometrics*, 74(2):584–594, 2018.