THE UNIVERSITY OF CHICAGO

BAYESIAN PARAMETRIC AND NONPARAMETRIC MODELS FOR CLINICAL TRIAL

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF PUBLIC HEALTH SCIENCES

BY
DEHUA BI

CHICAGO, ILLINOIS
JUNE 2024

To my parents, Fengxiang and Hong, and my love, Yanqing

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

There are many people who helped and supported me to make it possible for me to complete this dissertation. First and foremost, I would like to thank my advisor, Dr. Yuan Ji, for his deep insight, constant guidance, and dedication to my training and study. He not only introduced to me the academic world of biostatistics and helped me developing all the necessary skills and expertises that are required as a researcher in this field, he also helped me on training my mental and physical strength so I can live a healthy live while working on the interesting projects. Without him, I cannot imagine myself finishing this degree.

I also thank my committee members, Dr. Issam A. Awad, Dr. Mei-yin Polley, and Dr. Tianjian Zhou, for their continued guidance and constructive feedbacks on my projects. I would also like to acknowledge the help I received from Dr. Yitan Zhu and Dr. Wei Zhong, who helped me and provides me with insights in the projects we collaborated with. I am also thankful to all my fellow graduate students, as well as the faculty and staff members in the Department of Public Health Sciences for their companionship and support over the past five years.

Last but not least, I would also like to thank my parents, Fengxiang Bi and Hong Bi, for their steadfast love and encouragement throughout my life. They provide me with all the necessities I needed to pursue my dreams and goals. Finally, I would like to thank my future wife Yanqing Shen. This dissertation would not have been possible without her unconditional love, continued patience, and endless support.

ABSTRACT

With the advancement of computer-based technology, progress in computation has enabled effective real-life application of sampling methods. This has led to the adoption of Bayesian models in clinical trials. To this end, this dissertation comprises three papers that develop and apply Bayesian parametric and nonparametric models for the planning and analysis of clinical trials.

The first paper focuses on developing a statistical clustering method that clusters subjects across multiple groups through Bayesian nonparametric modeling. This method, named the Plaid Atoms Model (PAM), is built on the concept of "atom-skipping", which allows the model to stochastically assign zero weights to atoms in an infinite mixture. By implementing atom-skipping across different groups, PAM establishes a dependent clustering pattern, identifying both common and unique clusters among these groups. This approach further provides interpretable posterior inference such as the posterior probability of cluster being unique to a single group or common across a subset of groups. The paper also discusses the theoretical properties of the proposed and related models. Minor extensions of the model for multivariate or count data are presented. Simulation studies and applications using real-world datasets illustrate the performance of the new models with comparison to existing models.

The second paper delves into leveraging information from external data to augment the control arm of a current randomized clinical trial (RCT), aiming to borrow information while addressing potential heterogeneity in subpopulations between the external data and the current trial. To achieve this, we employ the PAM model introduced in the first paper. This method is used to identify overlapping and unique subpopulations across datasets, enabling us to limit information borrowing to those subpopulations common to both the external data and the current trial. This strategy establishes a Hybrid Control (HC) that results in a more precise estimation of treatment effects. Through simulation studies, we

validate the robustness of the proposed method. Additionally, its application to an Atopic Dermatitis dataset shows the method's improved treatment effect estimation.

The third paper introduces a Bayesian Estimator of Sample Size (BESS) method and its application in oncology dose optimization clinical trials. BESS seeks a balance among three factors: **S**ample size, **E**vidence from observed data, and **C**onfidence in posterior inference. It uses a simple logic of "given the evidence from data, with a specific sample size one is guaranteed to achieve a degree of confidence in the posterior inference." This approach contrasts with traditional sample size estimation (SSE), which typically relies on frequentist inference: BESS assumes a possible outcome from the observed data rather than utilizing the true parameters values in SSE method's sample size calculation. As a result, BESS does not calibration sample size based on type I or II error rates but on posterior probabilities, offering a more interpretable statement for investigators. The proposed method can easily accommodates sample size re-estimation and the incorporation of prior information. We demonstrate its performance through case studies via oncology optimization trials. However, BESS can be applied in general hypothesis tests which we discuss at the end.

CHAPTER 1

INTRODUCTION

This collection of papers focuses on the development and application of Bayesian parametric and nonparametric models in clinical trials. The first two papers explore the integration of external data into current trials to enhance tral analysis through Bayesian nonparametric (BNP) models. In contrast, the final paper introduces a Bayesian tool for sample size estimation based on a parametric model, designed to assist practitioners in planning sample sizes for dose-optimization trials, an essential phase in early phase trial stages. Here, we provide a briefly overview of these three papers.

The first paper addresses a statistical clustering challenge in grouped data, aiming to identity common and unique clusters of subjects within and across groups. The proposed Plaid Atoms Model (PAM) is a Bayesian nonparametric (BNP) model that is capable of generating flexible clustering patterns across groups *a priori*. Initially assuming all clusters to be common across groups, PAM then selectively eliminates clusters from each group through a novel “atom-skipping" mechanism, leading to a mix of overallping and distinct clusters. PAM offers interpretable posterior inferences, such as reporting the posterior probability of a cluster being exclusive to a single group or being common among a subset of groups. The paper also discusses the theoretical properties of PAM and its related models. For detail, refer to Chapter 2

The second paper develops a method to enhance the control arm of a Randomized Clinical Trial (RCT) by borrowing information from external data such as real-world data (RWD) or past trials. This approach, named PAM-HC, uses PAM method to cluster patients based on covariates, identifying those similar enough to warrant information borrowing. This targeted borrowing, facilitated by the power prior approach, aims to increase the accuracy of treatment effect estimation while mitigating bias from heterogeneous external patient subpopulations. The efficacy of PAM-HC is validated through simulations and a real-world

case study. The detail is in Chapter 3.

The third paper introduces the Bayesian Estimator of Sample Size (BESS), a novel method for estimating sample size in oncology dose optimization trial. Motivated by the current ad-hoc approach to determining patient numbers and the desire to quantify treatment effectiveness probabilistically, BESS seeks a balance between sample size, to-be-observed data evidence, and confidence in posterior inference.Unlike traditional frequentist sample size method calculations that rely on assumed true parameters values, BESS bases its sample size calibration on assumed evidence from data and the posterior probabilities, offering a clear, more interpretable framework. BESS also can be easily extended to accommodate sample size re-estimation strategies and to incorporate prior information. We dscribe the method in detail in Chapter 4.

CHAPTER 2

A CLASS OF DEPENDENT RANDOM DISTRIBUTIONS BASED ON ATOM SKIPPING

2.1 Introduction

Clustering, or unsupervised learning, is a primary tool for data analysis and scientific exploration. Representative clustering methods include algorithmic approaches like K-Means [MacQueen, 1967] and model-based clustering like MClust [Fraley and Raftery, 1998]. Alternatively, Bayesian nonparametric (BNP) models like the Dirichlet process (DP) [Ferguson, 1973] naturally induce clusters by allowing ties among observations. These "tied" values, which are random locations in the random probability measure of the BNP models, are also referred to as atoms in some literature, e.g., in Denti et al. [2021]. Hereafter, we use "clusters" and "atoms" interchangeably.

For complex problems and data structures where multiple datasets are analyzed together , dependent clustering is often necessary. For example, in linguistic research, it is of interest to discover common themes across multiple documents [Teh et al., 2004], where the themes are modeled as shared clusters. In drug development, oftentimes different studies and corresponding data are pooled to increase the precision of statistical inference. However, drug effects might be heterogeneous and therefore subpopulations (clusters) of patients must be identified to better characterize the treatment effects. A common question for many dependent clustering problems is whether a clustering method can capture shared clusters across all or some groups while also identify unique ones that belong to a single group.

Various dependent clustering approaches have been proposed in the BNP literature. Early pioneering work of dependent Dirichlet process (DDP) is initiated in MacEachern [1999, 2000]. These BNP models generate different patterns of atoms *a priori* on a spectrum that ranges from "common-atoms model" to "distinct-atoms model." Subsequently, common-

atoms models, such as hierarchical DP (HDP) [Teh et al., 2004], Common Atoms Model (CAM) [Denti et al., 2021], and hidden-HDP [Lijoi et al., 2022] assume all groups share the same set of atoms *a priori*. In contrast, distinct-atoms models, such as the nested DP (NDP) [Rodriguez et al., 2008], assume that groups with different distributions all have unique and distinctive atoms. Other methods in literature like the hierarchical mixture of DP [Müller et al., 2004b], the latent nested process (LNP) [Camerlenghi et al., 2019], and the semi-HDP [Beraha et al., 2021] take the middle ground by mixing distinct- and common-atoms processes. Consequently, a pattern of shared and unique clusters can be generated across groups under these models.

We consider a new approach to generate dependent clustering structure using a simple idea called atom skipping. Instead of mixing a distinct-atoms model with a common-atoms model, we construct random distributions by removing atoms from a common-atoms model in a group-specific fashion. This is realized by stochastically assigning the weight of an atom to be zero for each group. This effectively skips (removes) the atom from the group. For a single group or a single dataset, atom-skipping results in a new model called the Atom-Skipping Process (ASP). For multiple groups, it leads to the main proposal of the paper, the Plaid Atoms Model (PAM). When an atom is removed in all but one group, that atom becomes a unique cluster for that group. On the other hand, if an atom is not removed in any groups, it induces a common cluster shared for all groups. In-between is an atom that is removed in a fraction of groups, and the corresponding cluster is only shared by a subset of groups. The resulting dependent clustering pattern is slightly more flexible than some existing models. For example, when there are three or more groups, the set of overlapping clusters may vary between a pair of groups.

Furthermore, due to group-specific atom skipping, PAM defines a generative model that explicitly defines overlapping (common) and non-overlapping (unique) clusters across groups, which leads to more interpretable posterior inference. For example, PAM can perform infer-

ence on whether a cluster is absent in a group by reporting the posterior probability that the corresponding cluster has zero weight in the group. In contrast, common-atoms models like HDP or CAM always produce a positive cluster weight for any cluster in any group. An interesting by-product of atom-skipping is that the marginal mean of ASP and PAM follows a stochastic process that is called the Fractional Stick-Breaking Process (FSBP). This process is a simple modification of the stick-breaking representation of DP, and is linked to many random probability measures (RPM) and processes in the literature.

The remainder of the article are organized as follows. In Section 2.2, we review BNP models that are closely linked to PAM. In Section 2.3, we introduce three related new models, ASP, PAM, and FSBP. We discuss theoretical properties of the three new processes in Section 2.4, highlighting their interconnections. In Section 2.5, we discuss posterior inference and outline the slice sampler algorithm for PAM and FSBP. Section 2.6 presents comparative simulation results of the proposed models and Section 2.7 describes application of PAM to publicly available datasets. Lastly, Section 2.8 concludes the paper with some discussion.

2.2 Review of Some BNP Models for Clustering

2.2.1 Methods for Clustering a Single Study or Dataset

We review related BNP models to set the stage for the proposed new models. Figure 2.1 provides a graph illustrating the BNP models considered in our work. Specifically, nodes are BNP models, directed edges describe the extension of the parent node to the child node, black and red color represent existing and novel models respectively.

Consider a dataset with n observations of q-dimensional vectors ($q \geq 1$), with the ith observation denoted as $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iq})$, $i = 1, \ldots, n$. Denote the entire dataset $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$. Assume $\boldsymbol{y}_i$ takes a value from a suitable Polish space X that is endowed with the respective Borel σ-field $\mathcal{X}$. The observations are assumed to arise from a nonparametric

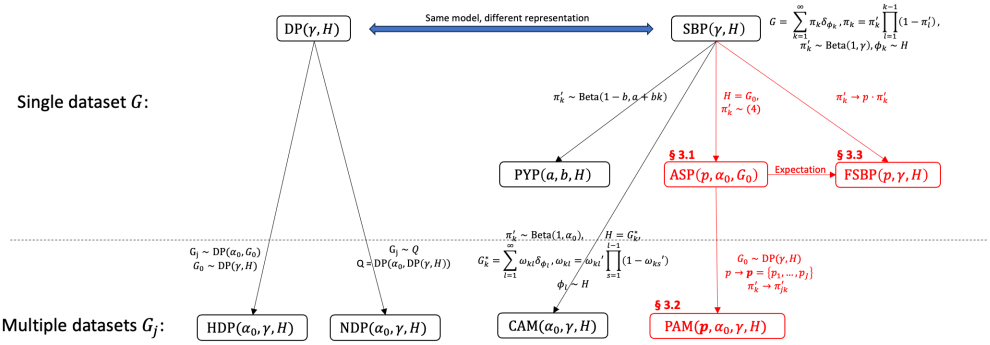

Figure 2.1: A graphic illustration of relationship of selected BNP models. A directed edge connecting two processes implies that the child process is an extension of the parent process. The red nodes and edges represent the contribution of this work. Section numbers of the manuscript are placed on the red nodes.

mixture model indexed by parameter $\boldsymbol{\theta}_i$ and a random distribution G as follows:

$$\boldsymbol{y}_i|\boldsymbol{\theta}_i \sim F(\boldsymbol{y}_i|\boldsymbol{\theta}_i), \quad \boldsymbol{\theta}_i|G \sim G, \quad i = 1, \ldots, n, \tag{2.1}$$

where $F(.|\boldsymbol{\theta}_i)$ is a parametric distribution for $\boldsymbol{y}_i$ with parameter $\boldsymbol{\theta}_i$, and G is assumed to have a nonparametric prior.

Review of DP: The DP prior is denoted as $G \sim DP(\gamma, H)$, where $\gamma > 0$ is the concentration parameter, and H is the base measure. Sethuraman [1994] shows that DP generates random distributions with the stick-breaking representation:

$$G = \textstyle\sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\phi}_k}, \ \pi_k \sim \text{GEM}(\gamma), \ \text{and } \boldsymbol{\phi}_k \sim H, \tag{2.2}$$

where GEM is the Griffiths-Engen-McCloskey distribution [Pitman, 2002]. Specifically, $\pi_k \sim \text{GEM}(\gamma)$ means that $\pi_k = \pi'_k \prod_{l=1}^{k-1}(1-\pi'_l)$, $\pi'_k \sim \text{Beta}(1, \gamma)$, where $\text{Beta}(a, b)$ denotes the beta distribution with mean $a/(a+b)$. This equivalent representation of DP is also referred

to as the Stick-Breaking Process (SBP), denoted as SBP(γ, H).

When applied to clustering, DP is often known for its “rich-get-richer" characteristic in that DP tends to produce few large clusters with many tiny ones or singletons. To address this, an alternative model is proposed known as the PYP.

Review of PYP: PYP(a, b, H) extends and modifies DP by assuming the atom weight follows

$$\pi'_k \sim \text{Beta}(1-b, a+b\cdot k),$$

where $a > -b$ and $b \in [0, 1)$. The construction of G and the distribution of $\boldsymbol{\phi}_k$ remain the same as in equation (2.2). PYP reduces to DP if $b = 0$ and $a = \gamma$. Compare to DP, PYP has two desirable properties: 1) the expected number of clusters of PYP grows more rapidly (with a rate of n^b) than that of DP (which grows with a rate of $\log(n)$), and 2) the rate of decay in terms of cluster-size follows a power law for PYP, but has an exponential tail in DP. However, due to non-i.i.d stick-breaking weights in PYP, many theoretical results of PYP are not available in closed form. For example, while the mean of DP is known to be the base-measure H in equation (2.2), PYP does not have a closed-form mean.

2.2.2 Methods for Clustering Multiple Studies or Datasets

Extend the previous setting to $J > 1$ studies or groups, each of which has a dataset of n_j observations. The ith observation in group j is denoted as a q-dimensional ($q \geq 1$) vector $\boldsymbol{y}_{ij}$. Let $\boldsymbol{y}_j = \{\boldsymbol{y}_{ij};\ i = 1, \ldots, n_j\}$ represent the entire dataset for the jth group. Assume

$$\boldsymbol{y}_{ij}|\boldsymbol{\theta}_{ij} \sim F(\boldsymbol{y}_{ij}|\boldsymbol{\theta}_{ij}), \quad \boldsymbol{\theta}_{ij}|G_j \sim G_j, \quad i = 1, \ldots, n_j;\ j = 1, \ldots, J, \tag{2.3}$$

where $F(\cdot|\boldsymbol{\theta}_{ij})$ is a parametric distribution for $\boldsymbol{y}_{ij}$. The models reviewed below assign priors to G_j for $j = 1, \ldots, J$. These models induce dependent partitions of $\boldsymbol{y}_j$'s, allowing for

information borrowing between groups. Most BNP models differ in their construction of common or distinct atoms across groups. While one school chooses to build common-atoms models that share a common set of atoms for all groups, another school allows groups to have non-overlapping atoms known as distinct-atoms models. A third school mixes the two ideas so that more flexible patterns of atoms can be modeled. Our work belongs to the third school.

Review of HDP: HDP [Teh et al., 2004] is a common-atoms model. In this model, each G_j is assigned a DP prior with a common base measure G_0, which itself is an instance of DP, i.e.,

$$G_j|\alpha_0, G_0 \sim \text{DP}(\alpha_0, G_0), \quad G_0|\gamma, H \sim \text{DP}(\gamma, H).$$

This model is denoted as $\text{HDP}(\alpha_0, \gamma, H)$. Using the stick-breaking representation [Sethuraman, 1994] of DP, HDP can be rewritten as

$$\begin{aligned} &G_j = \textstyle\sum_{k=1}^{\infty} \pi_{jk}\delta_{\boldsymbol{\phi_k}}, \pi_{jk} = \pi'_{jk}\prod_{l=1}^{k-1}(1-\pi'_{jl}) \\ &\pi'_{jk} \sim \text{Beta}\left(\alpha_0\beta_k, \alpha_0\left(1-\textstyle\sum_{l=1}^{k}\beta_l\right)\right) \\ &\boldsymbol{\phi}_k \sim H \text{ and} \qquad\qquad \beta_k \sim \text{GEM}(\gamma) \end{aligned} \tag{2.4}$$

where $\delta_{\{.\}}$ is the indicator function. Note that although G_0 is not shown in the stick-breaking construction of HDP in equation (2.4), it can be reconstructed from $\boldsymbol{\beta} = \{\beta_k; k \geq 1\}$ and $\boldsymbol{\Phi} = \{\boldsymbol{\phi_k}; k \geq 1\}$, i.e., $G_0 = \sum_{k=1}^{\infty}\beta_k\delta_{\boldsymbol{\phi_k}}$. Appropriate prior distributions, like the gamma distribution, can be specified for α_0 and γ to complete HDP.

It is clear from equation (2.4) that HDP is a common-atoms model, because all groups share the same set of atoms in $\boldsymbol{\Phi}$, and the atom weights $\pi_{jk} \neq 0$. While the atoms are shared, their weights $\boldsymbol{\pi}_j = \{\pi_{jk}; k \geq 1\}$ are distinct for different groups. Consequently, for G_j and $G_{j'}$ where $j \neq j'$, $G_j \neq G_{j'}$ with probability 1.

Review of NDP: Rodriguez et al. [2008] introduce the NDP, a model capable of clustering both subjects and groups. In NDP, the group-level clusters are referred to as distributional clusters, and the group-specific distribution G_j in NDP is defined as follows:

$$G_j|Q \sim Q, \quad Q = \mathrm{DP}(\alpha_0, \mathrm{DP}(\gamma, H)),$$

where the distribution of each group follows a DP with its base measure being another DP, rather than being a realization of DP as in HDP. We use $\mathrm{NDP}(\alpha_0, \gamma, H)$ to denote this model. NDP is a distinct-atoms model, which can be seen in its stick-breaking representation:

$$\begin{aligned} &G_j = \textstyle\sum_{k=1}^{\infty} \pi_k \delta_{G_k^*}, \;\; \pi_k \sim \mathrm{GEM}(\alpha_0) \\ &G_k^* = \textstyle\sum_{l=1}^{\infty} \omega_{kl} \delta_{\boldsymbol{\phi}_{kl}}, \omega_{kl} \sim \mathrm{GEM}(\gamma), \\ &\boldsymbol{\phi}_{kl} \sim H. \end{aligned} \tag{2.5}$$

For $j \neq j'$, if G_j and $G_{j'}$ are not equal to the same G_k^*, none of their atoms will be the same. In contrast, if they are equal to the same G_k^*, meaning G_j and $G_{j'}$ are two identical distributions, their atoms and atom weights will be identical. This phenomenon is known as "degeneracy" [Camerlenghi et al., 2019], where if two groups share just one atom they share all atoms and weights. Otherwise, the atoms and weights must all be distinct for these two groups. This presents a challenge if we aim to find common clusters for two groups belonging to different distributional clusters.

Review of CAM: Denti et al. [2021] extend NDP and introduce the Common Atoms Model, abbreviated as CAM. By definition, CAM is a common-atoms model. Building upon NDP, CAM provides distributional clustering similar to NDP. Specifically, CAM restricts the atoms in all G_k^*, $k \geq 1$ in NDP to a common set. In other words, rather than assuming that for each group k there is a distinct set of atoms $\{\boldsymbol{\phi}_{kl}\}$, CAM instead assumes all the

groups share a common set of atoms $\{\boldsymbol{\phi}_l\}$. Mathematically, the G_j's in CAM are defined as follows:

$$\begin{aligned} G_j &= \textstyle\sum_{k=1}^{\infty} \pi_k \delta_{G_k^*}, \pi_k \sim \text{GEM}(\alpha_0) \\ G_k^* &= \textstyle\sum_{l=1}^{\infty} \omega_{kl} \delta_{\boldsymbol{\phi}_l}, \omega_{kl} \sim \text{GEM}(\gamma), \boldsymbol{\phi}_l \sim H. \end{aligned} \tag{2.6}$$

We use the notation $\text{CAM}(\alpha_0, \gamma, H)$ to denote this model. Another recent development building upon the modeling of common atoms can be found in hidden-HDP [Lijoi et al., 2022]. Due to limited space, a review of this model is omitted.

Other BNP models: The aforementioned common-atoms and distinct-atoms models represent the two extremes of a spectrum of BNP priors for dependent random distributions. In the literature, many other models target the space in between, where the prior is allowed to contain both common and unique atoms *a priori*. Such models include the hierarchical mixture of DP [Müller et al., 2004b], the latent nested process (LNP) [Camerlenghi et al., 2019], and more recently, the semi-HDP [Beraha et al., 2021]. All these models construct flexible priors by adding or mixing distinct-atoms and common-atoms models together, in a nonparametric or semi-parametric fashion. For a comprehensive review, refer to Quintana et al. [2022]. In this work, we take a different approach. We start from a common-atoms model, and using an idea of atom skipping in a probabilistic fashion for each group. The resulting model provides common and unique atoms *a priori* but with interesting theoretical properties and behavior in statistical inference.

2.3 Proposed BNP Models

2.3.1 Atom-Skipping Process

The proposed models utilize a simple idea of atom skipping by probabilistically setting the weight of an atom to be exactly zero. We first consider a model for a single random

distribution (i.e., a single study or dataset). We denote such a model the ASP, standing for the atom-skipping process. Using the HDP in (2.4) as an example, atom skipping is implemented by assuming the prior for π'_{jk} to be

$$f(\pi'_{jk}) = p \times \underbrace{f_{\text{Beta}}\left(\alpha_0\beta_k, \alpha_0\left(1-\sum_{l=1}^{k}\beta_l\right)\right)}_{(2.4)} + (1-p) \times \underbrace{\delta_0}_{\text{atom skipping}}, \tag{2.7}$$

where $f_{\text{Beta}}(a, b)$ is the probability density function (p.d.f) of the beta distribution, and δ_0 is the indicator function at 0. Then we define the atom-skipping process (ASP) for a single dataset as

$$\begin{aligned} &G = \textstyle\sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\phi_k}}, \pi_k = \pi'_k \prod_{l=1}^{k-1}(1-\pi'_l), \\ &f(\pi'_k|\boldsymbol{\beta}, p, \alpha_0) = p \times f_{\text{Beta}}\left(\alpha_0\beta_k, \alpha_0\left(1-\textstyle\sum_{l=1}^{k}\beta_l\right)\right) + (1-p) \times \delta_0\,, \end{aligned} \tag{2.8}$$

where β_k and $\boldsymbol{\phi}_k$ are assumed to be given. We let $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\boldsymbol{\phi}_k}$ and denote model (2.8) as $G|p, \alpha_0, G_0 \sim \text{ASP}(p, \alpha_0, G_0)$. According to (2.8), since each π'_k or π_k has a probability to be zero, the corresponding atom $\boldsymbol{\phi}_k$ may be skipped when sampling from G.

2.3.2 Plaid Atoms Model

Adding back the DP prior on G_0 and extending ASP to multiple datasets, we propose the Plaid Atoms Model (PAM). Specifically, PAM is given in a hierarchical model as

$$G_j|p_j, \alpha_0, G_0 \sim \text{ASP}(p_j, \alpha_0, G_0),\ G_0|\gamma, H \sim \text{DP}(\gamma, H). \tag{2.9}$$

Using a stick-breaking representation, PAM can be shown to be equivalent to

$$
\begin{aligned}
&G_j = \textstyle\sum_{k=1}^{\infty} \pi_{jk} \delta_{\boldsymbol{\phi_k}}, \\
&\pi_{jk} = \pi'_{jk} \textstyle\prod_{l=1}^{k-1} (1 - \pi'_{jl}), \\
&f(\pi'_{jk} | \boldsymbol{\beta}, p, \alpha_0) = p_j \times f_{\text{Beta}} \left(\alpha_0 \beta_k, \alpha_0 \left(1 - \textstyle\sum_{l=1}^{k} \beta_l \right) \right) + (1 - p_j) \times \delta_0 \,, \\
&\boldsymbol{\phi}_k \sim H \text{ and } \beta_k \sim \text{GEM}(\gamma).
\end{aligned}
\tag{2.10}
$$

The proof of the equivalence between (2.9) and (2.10) is omitted and follows the same derivation in Teh et al. [2004]. Note that when $p_j = 1,\ \forall j$, PAM is equivalent to HDP. This can be trivially shown by comparing models (2.4) and (2.10).

Let $\boldsymbol{p} = \{p_1, \ldots, p_J\}$. We denote this model as $G_1, \ldots, G_J \sim \text{PAM}(\boldsymbol{p}, \alpha_0, \gamma, H)$. Additional pirors can be placed on the parameters of $\boldsymbol{p}$, α_0, and γ, for example,

$$
p_j | a, b \sim \text{Beta}(a, b), \alpha_0 \sim \text{Gamma}(a_\alpha, b_\alpha), \gamma \sim \text{Gamma}(a_\gamma, b_\gamma). \tag{2.11}
$$

By construction, PAM is more versatile as a generative model. It allows different G_j's to share some atoms but also possesses unique ones. A comparison of PAM and other dependent random distributions like HDP and CAM is given in Supplement A.1 as a reference.

Continuous Data: PAM in (2.9) can be used as a prior for the random distribution in model (2.3). If observations y_{ij} are continuous and univariate ($q = 1$), we use a Gaussian kernel by setting $\boldsymbol{\phi}_k = (\mu_k, \sigma_k^2)$ and $F(\cdot | \boldsymbol{\phi}_k) = N(\cdot | \mu_k, \sigma_k^2)$. To complete model specification for $\text{PAM}(\boldsymbol{p}, \alpha_0, \gamma, H)$, the base measure H is modeled as the conjugate prior of normal-inverse-gamma (NIG), where $H = \text{NIG}(\mu_0, \kappa_0, \alpha_0, \beta_0)$, i.e., $\mu_k | \sigma_k^2 \sim N(\mu_0, \sigma_k^2 / \kappa_0)$ and $\sigma_k^2 \sim \text{IG}(\alpha_0, \beta_0)$. For multivariate observations ($q > 1$), the related model components are changed to multivariate normal and normal-inverse-Wishart distributions. The detail is ommitted for simplicity.

Count Data: Following Denti et al. [2021], we extend the proposed PAM to count data and refer to it as the Discrete Plaid Atoms Model (DPAM). We only consider univariate count data and hence $q = 1$. Let $x_{ij} \in \mathbb{N}$ be the observed count data for observation $i = 1, \ldots, n_j$ in group $j = 1, \ldots, J$, where $\mathbb{N}$ denotes the natural numbers. Thus the data vector $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{n_j j})$ is the set of counts observed for the jth group. We apply the data augmentation framework in Canale and Dunson [2011] and introduce latent continuous variables y_{ij} so that

$$\Pr(x_{ij} = \omega) = \int_{a_\omega}^{a_{\omega+1}} g(y_{i,j}) dy_{ij}, \quad \omega = 0, 1, 2, \cdots \tag{2.12}$$

where $a_0 < a_1 < \cdots < a_\infty$ is a fixed sequence of thresholds that take values $\{a_\omega; \omega \geq 0\} = \{-\infty, 0, 1, 2, \ldots, +\infty\}$, and $g(y_{ij})$ follows the PAM mixture model as in equations (2.3) and (2.10). This construction allows posterior inference for y_{ij} since it is trivial to see that

$$x_{ij} | y_{ij} = \sum_{\omega=0}^{\infty} \mathbf{1}_\omega(x_{ij}) \cdot \mathbf{1}_{[a_\omega, a_{\omega+1})}(y_{ij}),$$

where $\mathbf{1}_a(b)$ equals 1 if $b = a$ or $b \in a$, and 0 otherwise.

2.3.3 Fractional Stick-Breaking Process

Taking expected value of π_k in (2.8), we derive a new process for a single group called the Fractional Stick-Breaking Process (FSBP). This new process gives an interesting and new solution for modeling a random distribution, and induces a clustering structure that is different from existing models like the DP or PYP.

Let $p \in (0, 1)$, and $a, b > 0$ be fixed constants. The FSBP is an extension of the DP (or

equivalently the SBP) and given by

$$\begin{aligned} &G = \textstyle\sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\phi}_k},\ \pi_k = p \cdot \pi_k{}' \prod_{l=1}^{k-1}(1 - p \cdot \pi_l{}'), \\ &\pi_k{}'|\gamma \sim \text{Beta}(a = 1, b = \gamma),\ \boldsymbol{\phi}_k|H \sim H. \end{aligned} \tag{2.13}$$

We denote this model as $G \sim \text{FSBP}(p, \gamma, H)$. When $p = 1$, $\text{FSBP}(p, \gamma, H)$ reduces to $\text{SBP}(\gamma, H)$ or equivalently $\text{DP}(\gamma, H)$. We show in Section 2.4 that the FSBP is the mean of the ASP and has more expected number of clusters than that of DP with the same concentration parameter γ.

2.4 Properties of ASP, PAM, and FSBP

2.4.1 Properties of ASP and PAM

We start by showing that the cluster weights in ASP and PAM sum to 1.

Proposition 1. *Assume $\boldsymbol{\beta} = \{\beta_k; k \geq 1\}$, $\beta_k \sim GEM(\gamma)$, $f(\pi_k'|\boldsymbol{\beta}, p, \alpha_0)$ is given in* (2.8) *and $f(\pi_{jk}'|\boldsymbol{\beta}, p_j, \alpha_0)$ in* (2.10)*. Furthermore, assume $p, p_j \sim Beta(a, b)$. Then*

1. $\sum_{k\geq 1} \pi_k = 1$, $\sum_{k\geq 1} \pi_{jk} = 1$,
2. $E[\pi_k|\boldsymbol{\beta}, p] = p \cdot \beta_k' \prod_{l=1}^{k-1}(1 - p \cdot \beta_l')$, $E[\pi_{jk}|\boldsymbol{\beta}, p_j] = p_j \cdot \beta_k' \prod_{l=1}^{k-1}(1 - p_j \cdot \beta_l')$, *and*
3. $E[\pi_k] = E[\pi_{jk}] = \frac{1}{1+\gamma'}\left(\frac{\gamma'}{1+\gamma'}\right)^{k-1}$ *where* $\gamma' = \frac{1+\gamma-\bar{p}}{\bar{p}}$, $\bar{p} = \frac{a}{a+b}$.

The proof is in Supplement A.2. This result shows that the random distributions G from ASP and G_j from PAM are proper discrete random distributions. Next, we show that the mean process of ASP is FSBP.

Theorem 1. *For an arbitrary set $A \subseteq X$, let $\alpha_0, \gamma > 0$, H be a fixed probability measure, $G_0 \sim DP(\gamma, H)$, and $G|G_0, p \sim ASP(p, \alpha_0, G_0)$ as in* (2.8)*. Then, conditional on G_0 and*

p, the conditional mean of G is

$$E[G(A)|G_0, p] = G^*(A),$$

where $G^* \sim FSBP(p, \gamma, H)$.

The proof of Theorem 1 is in Supplement A.3. Combining the results of Theorem 1 and Theorem 2, the following corollary gives the maginal mean of the ASP.

Corollary 1. *If* $G_0 \sim DP(\gamma, H)$, $p \sim Beta(a, b)$, *and* $G|G_0, p \sim ASP(p, \alpha_0, G_0)$, *then* $E[G(A)] = E[E[G(A)|G_0, p]] = E[G^*(A)] = H(A)$.

Lastly, we look at properties related to PAM. Since PAM is an extension of ASP to multiple groups, the results in Theorem 1 apply to the group-specific random distribution G_j of PAM as well. Corollary 1 also applies to a random distribution G_j from PAM. Moreover, in the next proposition, we show that *a priori*, there is a positive probability for two observations from two different groups to be clustered together in PAM.

Proposition 2. *Let* $G_1, \ldots, G_J \sim PAM(\boldsymbol{p}, \alpha_0, \gamma, H)$. *Without loss of generality, for two groups* G_1 *and* G_2, *let* $\boldsymbol{\theta}_{i1}|G_1 \sim G_1$ *and* $\boldsymbol{\theta}_{i'2}|G_2 \sim G_2$, *then*

$$Pr(\boldsymbol{\theta}_{i1} = \boldsymbol{\theta}_{i'2}) > 0. \tag{2.14}$$

The proof of Proposition 2 is given in Supplement A.4.

Unfortunately, closed-form results are unavailable for the variance, correlation structure, and partition probability functions of PAM. The expected number of clusters for PAM is not available in closed form either. Consequently, we investigate the clustering properties of PAM through a small simulation and compare it to CAM and HDP.

We assume there are 500 groups ($j = 1, \ldots, 500$) and within each group G_j, we generate a random sample of $1,000$ observations from CAM, HDP, or PAM. This leads to a total of

$500,000$ observations for each process. When sampling an observation from these processes, computationally it is not feasible to sample from an infinite mixture. Instead, we consider a finite mixture of $1,000$ atoms, which are sampled from the base measure H: $\phi_k \sim H$ for $k = 1, \ldots, 1,000$, where $H = N(0,1)$. We set the concentration parameters $\alpha_0 = \gamma = 1$ for CAM, HDP, and PAM. Therefore, we use notation CAM$(1,1,H)$ and HDP$(1,1,H)$. We consider two versions of PAM, with $p_{j1} \sim$ Beta$(80,20)$ or $p_{j2} \sim$ Beta$(20,80)$, for $j = 1, \ldots, 500$. This leads to PAM$(\boldsymbol{p}_1,1,1,H)$ and PAM$(\boldsymbol{p}_2,1,1,H)$, where $\boldsymbol{p}_1 = \{p_{j1}; j = 1, \ldots, 500\}$ and $\boldsymbol{p}_2 = \{p_{j2}; j = 1, \ldots, 500\}$. We sample the atom weights π's in each group based on their corresponding stick-breaking processes, model (2.4) for HDP, (2.6) for CAM, and (2.10) for PAM. At the end, we obtain $1,000$ observations per group for 500 groups under each of the four processes, CAM$(1,1,H)$, HDP$(1,1,H)$, PAM$(\boldsymbol{p}_1,1,1,H)$ and PAM$(\boldsymbol{p}_2,1,1,H)$, with $H = N(0,1)$.

Figure 2.2 summarizes the number of clusters and the relative cluster size, either for a single group or for the entire 500,000 observations across all 500 groups, under each of the four processes. The processes exhibit quite different behavior. First, the average number of clusters in a group is 7.62 (SD 2.56), 3.00 (SD 0.86), 2.49 (SD 0.92), and 1.24 (SD 0.47) for CAM$(1,1,H)$, HDP$(1,1,H)$, PAM$(\boldsymbol{p}_1,1,1,H)$ and PAM$(\boldsymbol{p}_2,1,1,H)$, respectively. This can be observed based on the average length of the grey lines in the subplots of Figure 2.2, each grey line representing a group. However, aggregating all the observations, the total number of clusters is 18, 10, 13, and 43, for the four processes, respectively, corresponding to the length of the blue line in the figure. Therefore, HDP (top right) generates the smallest number of clusters while PAM$(\boldsymbol{p}_2,1,1,H)$ (bottom right) generates the largest number of clusters. Interestingly, PAM$(\boldsymbol{p}_2,1,1,H)$ (bottom right plot) also generates on average the smallest number of clusters per group (shortest grey lines). This means that for PAM many clusters across groups are unique, a feature that is different from the other three processes. Lastly, PAM$(\boldsymbol{p}_1,1,1,H)$ (bottom left) behaves similar to HDP (top right), which

is expected, since when p_j approaches 1, PAM is identical to HDP. Lastly, CAM (top left) generates on average the largest number of clusters per group (longest grey lines) without producing a large number of total clusters. This means CAM is inclined to generate more and overlapping clusters across groups. Additional results comparing the clustering behavior of the three models can be found in Supplement A.5.

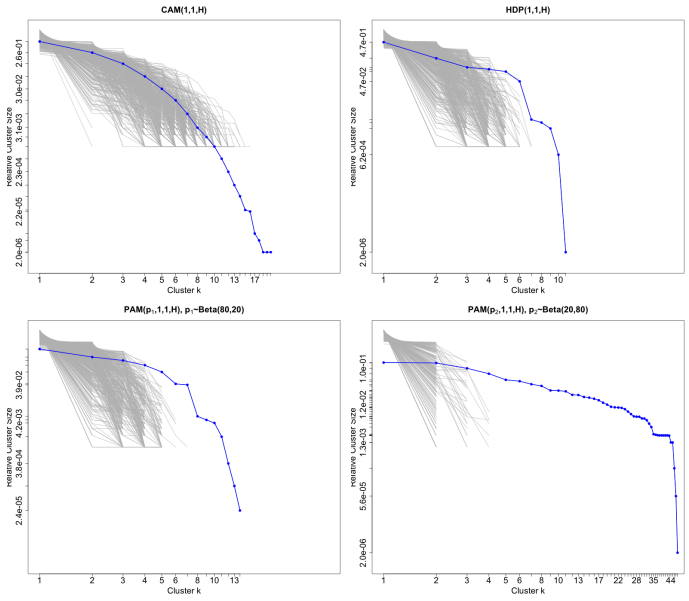

Figure 2.2: Clustering pattern of CAM, HDP, and PAM. The four subplots present the relative cluster size against the number of clusters for the four processes, CAM$(1, 1, H)$, HDP$(1, 1, H)$, PAM$(\boldsymbol{p}_1, 1, 1, H)$ and PAM$(\boldsymbol{p}_2, 1, 1, H)$, with $H = N(0, 1)$. The grey lines in each subplot correspond to the observations within each group and the blue lines correspond to the relative cluster size of all the observations aggregated across 500 groups.

2.4.2 Properties of FSBP

We first show that the mean and variance of FSBP are available in closed forms and is related to DP.

Theorem 2. *For an arbitrary set $A \subseteq X$, let $p \in (0, 1)$, $\gamma > 0$ be fixed constants, and H be a fixed probability measure. For $G^* \sim FSBP(p, \gamma, H)$, the mean and variance of G^* on A*

are

$$E[G^*(A)] = H(A), \;\; Var(G^*(A)) = \frac{H(A)\{1-H(A)\}}{v}, \;\; \textit{where } v = \frac{1+\gamma}{p} + \frac{1-p}{p}.$$

The proof of the theorem is in Supplement A.6.

Remark 1. *The mean and variance of G^* match the mean and variance of a DP $G' \sim DP(v-1, H)$, respectively.*

We next derive the exchangeable partition probability function (EPPF) of FSBP G^*. Let $\boldsymbol{z} = \{z_1, \ldots, z_n\}$ represent the vector of cluster memberships for n observations sampled from G^*. Without loss of generality, suppose $z_i \in \{1, \ldots, K\}$ which means there is a total of K clusters indexed from 1 to K. Then $\boldsymbol{z}$ defines a partition of the n observations, denoted as $C(\boldsymbol{z}) = \{c_1, \ldots, c_K\}$ where $\cup_{k=1}^{K} c_k = \{1, \ldots, n\}$ and $c_k = \{i; z_i = k\}$. For any partition C of $\{1, \ldots, n\}$, the EPPF of G^* is defined as $\Pr(C(\boldsymbol{z}) = C)$ [Pitman, 1995]. Following the work of Miller [2019], we derive the expression for the EPPF of G^*.

Theorem 3. *Let $p \in (0,1)$ be a fixed constant, and let H be a fixed probability measure. Suppose $G^* \sim FSBP(p, \gamma, H)$ and that n observations are sampled from G^*. Without loss of generality, denote $C = \{c_1, \ldots, c_K\}$ a partition of the n observations, with $1 \leq K \leq n$. Furthermore, let $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_K\}$ be a permutation of $\{1, \ldots, K\}$ and S_K denote the set of all $K!$ possible permutations of $\{1, \ldots, K\}$. The EPPF of G^* for n observations is given by*

$$\prod_{k=1}^{K} \Gamma(\gamma+1) p^{|c_k|} \frac{\Gamma(|c_k|+1)}{\Gamma(\gamma+|c_k|+1)} \times \left(\sum_{\boldsymbol{\lambda} \in S_K} \prod_{k=1}^{K} \left\{ \frac{{}_2F_1(-\alpha_{k+1}(\boldsymbol{\lambda}), |c_{\lambda_k}|+1; \gamma+|c_{\lambda_k}|+1; p)}{1 - {}_2F_1(-\alpha_{k+1}(\boldsymbol{\lambda}), |c_{\lambda_k}|+1; \gamma+|c_{\lambda_k}|+1; p)} \right\} \right)$$

where $\Gamma(\cdot)$ is the gamma function, $|c|$ denotes the cardinality of the set c, ${}_2F_1(a, b; c; d)$ is the hypergeometric function with parameters a, b, c and d, $\alpha_k(\boldsymbol{\lambda}) = |c_{\lambda_k}| + |c_{\lambda_{k+1}}| + \cdots + |c_{\lambda_K}|$, and c_{λ_k} is the λ_k's component of C. When $p \to 1$, the EPPF of G^* converges to the EPPF

of $G_0 \sim DP(\gamma, H)$*, which is given by*

$$\frac{\gamma^{|C|}\Gamma(\gamma)}{\Gamma(n+\gamma)} \prod_{k=1}^{K} \Gamma(|c_k|).$$

The proof of theorem 3 is given in Supplement A.7. Details of the hypergeometric function can be found in Abramowitz et al. [1988].

We next explore the clustering property of the FSBP to show that the expected number of clusters in G^* is greater than the corresponding DP with $G_0 \sim DP(\gamma, H)$. The first lemma derives the probability of forming a new cluster under FSBP.

Lemma 1. *Let* $p \in (0,1)$ *and* $\gamma > 0$ *be fixed constants, and let* H *be a fixed probability measure. Let* $G^* \sim FSBP(p, \gamma, H)$*, and let* $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_i | G^* \sim G^*$*. Denote* w_i *as a binary indicator for the* i*th sample* $\boldsymbol{\theta}_i$*, such that*

$$w_i = \begin{cases} 1 & \text{if } \boldsymbol{\theta}_i \notin \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{i-1}\} \\ 0 & \text{o.w.} \end{cases}.$$

Then, for $i \geq 2$*,*

$$Pr(w_i = 1 | p, \gamma) = 1 - \sum_{k=2}^{i} (-1)^k \binom{i-1}{k-1} \frac{(k-1)!}{\prod_{l=1}^{k}(l+\gamma)} \frac{(\gamma+1)p^{k-1}}{{}_2F_1(1, 1-k; \gamma+2; p)}$$

where ${}_2F_1(a, b; c; d)$ *is the hypergeometric function.*

The proof of Lemma 1 is in Supplement A.8.

Lemma 2. *It follows that*

$$\lim_{p \to 1} Pr(w_i = 1 | p, \gamma) = \frac{\gamma}{\gamma + i - 1}.$$

The proof of Lemma 2 is in Supplement A.9. Note that the right hand side of the equation

in Lemma 2 is the probability of forming a new cluster under DP [Müller et al., 2015]. Based on Lemma 1 and 2, we have the following theorem.

Theorem 4. *Let $p \in (0,1)$, $\gamma > 0$ be fixed constants, and w_i be defined as in Lemma 1. Then*

$$Pr(w_i = 1|p,\gamma) > \frac{\gamma}{\gamma + i - 1}.$$

The proof of Theorem 4 is shown in Supplement A.10. The following corollary follows directly from Theorem 4.

Corollary 2. *Let n^* be the expected number of clusters of $G^* \sim FSBP(p,\gamma,H)$ on n samples. Then*

$$E[n^*|p,\gamma] = 1 + \sum_{i=2}^{n} Pr(w_i = 1|p,\gamma).$$

Let n_0 be the expected number of clusters of $G_0 \sim DP(\gamma,H)$ on n samples. Then

$$E[n_0|\gamma] = \sum_{i=1}^{n} \frac{\gamma}{\gamma + i - 1}.$$

Additionally, we have

$$E[n^*|p,\gamma] > E[n_0|\gamma] \approx \gamma \log\left(\frac{\gamma + n}{\gamma}\right).$$

Remark 2. *The FSBP has a larger expected number of clusters than DP with the same concentration parameter .*

In summary, FSBP in (2.13) can be considered as a “truncated" DP with a factor of p. When $p = 1$, FSBP is the same as DP.

2.5 Posterior Inference

2.5.1 Overview

We develop computational algorithms for sampling PAM and FSBP. We do not consider sampler for ASP since ASP can be viewed as PAM for a single group and thus can be sampled similarly as PAM. For PAM, we modify an efficient slice sampler in Denti et al. [2021] and illustrate the new algorithm using univariate data. The modified sampler can be easily extended to accommodate multivariate observations (i.e., $q > 1$) and discrete data. Alternative approaches like the Gibbs sampler based on the Chinese restaurant franchise process Teh et al. [2004] or blocked-Gibbs sampler Rodriguez et al. [2008] by truncating the infinity mixture in PAM are not considered, as they are either not feasible or prone to inferential errors.

2.5.2 Slice Sampler for PAM and FSBP

To facilitate the development of the slice sampler for PAM, we adopt the parametrization in Denti et al. [2021] and Teh et al. [2004], adding the sampling model for observation y_{ij}. Specifically, the proposed PAM can be represented using a set of latent indicator variables $\boldsymbol{Z} = \{z_{ij};\, i \geq 1,\, j = 1, \ldots, J\}$ as cluster memberships for the observations. In other words, $z_{ij} = k$ if observation i in group j is assigned to cluster k. Denoting $\boldsymbol{\pi}_j = \{\pi_{jk}; k \geq 1\}$ and adding the sampling model for y_{ij}, we consider a PAM mixture model as

$$
\begin{aligned}
&y_{ij}|z_{ij}, \boldsymbol{\Phi} \sim F(\boldsymbol{y}_{ij}|\boldsymbol{\phi}_{z_{ij}}),\\
&z_{ij}|\boldsymbol{\pi}_j \sim \textstyle\sum_{k=1}^{\infty} \pi_{jk}\delta_k(z_{ij}), \quad \pi_{jk} = \pi'_{jk}\prod_{l=1}^{k-1}(1-\pi'_{jl}),\\
&f(\pi'_{jk}|\boldsymbol{\beta}, p_j, \alpha_0) = p_j \times f_{\text{Beta}}\left(\alpha_0\beta_k, \alpha_0\left(1-\textstyle\sum_{l=1}^{k}\beta_l\right)\right) + (1-p_j)\times\delta_0.
\end{aligned}
\tag{2.15}
$$

The other components of PAM are the same as (2.10) and (2.11). This reparameterization is routinely used to facilitate posterior inference [Denti et al., 2021, Teh et al., 2004].

By integrating out z_{ij} in model (2.15), we can rewrite the density function for y_{ij} as an infinite mixture as

$$f(y_{ij}|\boldsymbol{\Phi}, \boldsymbol{\pi}_j) = \sum_{k\geq 1} \pi_{jk} \cdot p(y_{ij}|\boldsymbol{\phi}_k), \tag{2.16}$$

where $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_k; k \geq 1\}$. Following Kalli et al. [2011], we use a set of uniformly distributed random variables $\boldsymbol{u} = \{u_{i,j}; i = 1, \ldots, n_j, j = 1, \ldots, J\}$ to separate the "active" mixture components from the other "inactive" components, which will become clear next. By definition, each $u_{ij} \sim \text{Unif}(0, 1)$. Additionally, we consider J deterministic probabilities $\boldsymbol{\xi}_j = \{\xi_{jk}; k \geq 1\}$ for a fixed j, where $\xi_{jk} \equiv \xi_k = (1-\zeta)\zeta^{k-1}$, $\zeta \in (0, 1)$ is a fixed parameter with a default value of 0.5, and $\boldsymbol{\xi}_j \equiv \boldsymbol{\xi} = \{\xi_k; k \geq 1\}$. A more complicated construction may allow different ζ_j for different groups j, which we do not consider here. As a result, the augmented likelihood for observation y_{ij} can be expressed as:

$$f_{\boldsymbol{\xi}}(y_{ij}, u_{ij}|\boldsymbol{\Phi}, \boldsymbol{\pi}_j) = \sum_{k\geq 1} 1_{\{u_{ij}<\xi_k\}} \frac{\pi_{jk}}{\xi_k} p(y_{ij}|\boldsymbol{\phi}_k), \tag{2.17}$$

where $1_{\{A\}}$ equals 1 if condition A is satisfied, and 0 otherwise. Integrating out u_{ij} in (2.17) returns $f(y_{ij}|\boldsymbol{\Phi}, \boldsymbol{\pi}_j)$ in (2.16). Now adding the cluster indicator z_{ij} in (2.15), we express (2.17) as

$$f_{\boldsymbol{\xi}}(y_{ij}, u_{ij}|z_{ij}, \boldsymbol{\Phi}, \boldsymbol{\pi}_j) = \sum_{k\geq 1} 1_{\{z_{ij}=k\}} 1_{\{u_{ij}<\xi_{z_{ij}}\}} \frac{\pi_{jz_{ij}}}{\xi_{z_{ij}}} p(y_{ij}|\boldsymbol{\phi}_{z_{ij}}) \tag{2.18}$$

The proposed slice sampler follows a Gibbs-sampler style, in which it iteratively samples the following parameters,

1. $u_{ij}|\cdots \propto \text{Unif}(0, \xi_{z_{ij}})$,
2. the stick-breaking weights β'_k, π'_{jk}, and p_j,
3. the indicator z_{ij} with $\Pr(z_{ij} = k|\cdots) \propto 1_{\{u_{ij}<\xi_k\}} \frac{\pi_{jk}}{\xi_k} p(y_{ij}|\boldsymbol{\phi}_k)$, and

4. the atom location parameter $\boldsymbol{\phi}_k|\cdots \propto \prod_{z_{ij}=k} N(y_{ij}|\boldsymbol{\phi}_k)p_H(\boldsymbol{\phi}_k)$.

In the last step, since $\boldsymbol{\phi}_k \sim H$, $p_H(\boldsymbol{\phi}_k)$ denotes the prior density of H. The entire sampler is presented in Algorithm 1. Below we describe the details of sampling π'_{jk} in step 2 above. The other details of the entire slice sampler are in Supplement A.11.

In each iteration of the slice sampler, due to the introduction of the latent uniform variate u_{ij} and the truncation on ξ_k, the infinite summation in equation (2.17) can be reduced to a finite sum through "stochastic truncation". To see this, first notice that $\{\xi_k; k \geq 1\}$ is a descending sequence, and therefore only finitely many ξ_k's can meet the condition $u_{ij} < \xi_k$. In other words, given $\boldsymbol{u}$, there exists a $K' \geq 1$ such that when $k \geq K'$, $\min(u_{ij}) \geq \xi_k$, where the min is taken over all i and j. This means that up to K' of the ξ_k's will be larger than u_{ij}. Let $K^* = K' - 1$. Then, noticing that $\xi_{K^*} = (1-\zeta)\zeta^{K^*}$, we can easily show that

$$K^* = \left\lfloor \frac{\log(\min(\boldsymbol{u})) - \log(1-\zeta)}{\log(\zeta)} \right\rfloor. \tag{2.19}$$

Here, K^* is called the "stochastic truncation" in the slice sampler. Given K^*, sampling β'_k is straightforward but requires a Metropolis-Hastings (MH) step (See Supplement A.11 for details). To sample π'_{jk}, again conditional on K^*, let

$\boldsymbol{Z}_j = \{z_{ij}; i = 1, \ldots, n_j\}$, $m_{jk} = \sum_{i=1}^{n_j} 1(z_{ij} = k)$, and refer to the stick-breaking representation. The full conditional distribution of π'_{jk} is given by

$$p(\pi'_{jk}|\cdots) = p(\pi'_{jk}|\boldsymbol{Z}_j, \boldsymbol{\beta}, p_j, \alpha_0) \propto \left[{\pi'_{jk}}^{m_{jk}} (1-\pi'_{jk})^{\sum_{s=k+1}^{K^*} m_{js}} \right] f(\pi'_{jk})$$

where $f(\pi'_{jk})$ is defined in equation (2.7). When $m_{jk} > 0$, it means cluster k in group j is not empty, and therefore $\pi'_{jk} \neq 0$ (otherwise, it would not be possible to have a non-empty

cluster k in group j). Hence, the full conditional of π'_{jk} is

$$p(\pi'_{jk}|\cdots) = f_{\text{Beta}}\left(\alpha_0\beta_k + m_{jk}, \alpha_0\left(1 - \sum_{l=1}^{k}\beta_l\right) + \sum_{s=k+1}^{K^*} m_{js}\right). \tag{2.20}$$

Recall $f_{\text{Beta}}(,)$ denotes a beta distribution density. When $m_{jk} = 0$, which could mean $\pi'_{jk} = 0$ or $\pi'_{jk} \neq 0$ but the atom is not sampled, we have

$$p(\pi'_{jk}|\cdots) \propto (1 - \pi'_{jk})^{\sum_{s=k+1}^{K^*} m_{js}} f(\pi'_{jk}).$$

This can be expressed as

$$p(\pi'_{jk}|\cdots) = p_j^* \times f_{\text{Beta}}\left(\alpha_0\beta_k, \alpha_0\left(1 - \sum_{l=1}^{k}\beta_l\right) + \sum_{s=k+1}^{K^*} m_{js}\right) + (1 - p_j^*) \times \delta_0 \tag{2.21}$$

where

$$p_j^* = \frac{p_j}{p_j + (1-p_j) \times \frac{B\left(\alpha_0\beta_k, \alpha_0\left(1-\sum_{l=1}^{k}\beta_l\right)\right)}{B\left(\alpha_0\beta_k, \alpha_0\left(1-\sum_{l=1}^{k}\beta_l\right) + \sum_{s=k+1}^{K^*} m_{js}\right)}}$$

and $B(a, b)$ is the beta function.

Lastly, sampling p_j and the concentration parameters follow standard MCMC simulation [Escobar and West, 1995], details of which is provided in Supplement A.11.

Additional step for count data Finally, for DPAM an additional step is added to update the latent continuous variable. Denote $\text{TN}(\mu, \sigma^2; a, b)$ the truncated normal distribution with mean μ, variance σ^2, and boundaries a and b, the full conditional distribution of y_{ij} is

$$y_{ij}|\cdots \sim \text{TN}(\mu_{z_{ij}}, \sigma^2_{z_{ij}}; a_{x_{ij}}, a_{x_{ij}+1}). \tag{2.22}$$

Computation Algorithm Algorithm 1 introduces the proposed slice sampler. For multivariate observations, step 9 of Algorithm 1 can be replaced with a conjugate NIW prior, and multivariate normal can be used for $p(y_{ij}|\boldsymbol{\phi}_k)$ in step 8. On the other hand, the extension to DPAM can be achieved by adding steps to sample the latent y_{ij} according to equation (2.22) after step 7, and modifying the likelihood $p(y_{ij}|\boldsymbol{\phi}_k)$ in step 8 with

$$p(x_{ij}|\boldsymbol{\phi}_k) = \Delta\Phi(a_{x_{ij}}|\boldsymbol{\phi}_k) = \Phi(a_{x_{ij}+1}|\boldsymbol{\phi}_k) - \Phi(a_{x_{ij}}|\boldsymbol{\phi}_k),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function (c.d.f) of the Gaussian distribution.

Algorithm 1 Slice-Efficient Sampler for PAM

1: **for** $m = 1, \ldots, M$ **do**
2: Sample each u_{ij} from $u_{ij} \sim \text{Unif}(0, \xi_{z_{ij}})$ and find K^* in (2.19).
3: Sample all β'_k for $k = 1, \cdots, K^*$ with MH step.
4: **for** each π'_{jk} for $j = 1, \cdots, J$ and $k = 1, \cdots, K^*$ **do**
5: **if** $m_{jk} > 0$, sample π'_{jk} from (2.20). **otherwise**, sample π'_{jk} from (2.21).
6: **end for**
7: Sample $\boldsymbol{p} = \{p_j; j = 1, \ldots, J\}$: denote $m_{j0} = \sum_{k=1}^{K^*} 1(\pi'_{jk} = 0)$,

$$p_j|\cdots \sim \text{Beta}(a + K^* - m_{j0}, b + m_{j0})$$

8: Sample $\boldsymbol{Z} = \{z_{ij}; i = 1, \ldots, n_j, j = 1, \ldots, J\}$ from the following full condition:

$$p(z_{ij} = k|\cdots) \propto 1_{\{u_{ij} < \xi_k\}} \frac{\pi_{jk}}{\xi_k} p(y_{ij}|\boldsymbol{\phi}_k)$$

9: Sample $\boldsymbol{\phi}_k$ from a conjugate NIG.
10: **end for**

Label Switching As PAM involves an infinite mixture model, the issue of label switching can arise in MCMC samples [Papastamoulis, 2015]. To address the problem of label switching, we use the Equivalence Classes Representatives (ECR) algorithm described in Papastamoulis and Iliopoulos [2010]. Details of label-switching with the ECR method are in Supplement A.11.

Slice sampler for FSBP The slice sampler for FSBP follows the same flow as the one for PAM above. We simply need to add the sampling model (2.1) and rewrite the FSBP in (2.13) using latent indicator variables $\boldsymbol{Z} = \{z_i; i \geq 1\}$ in a mixture model given by

$$\begin{aligned}
&\boldsymbol{y}_i | z_i, \boldsymbol{\Phi} \sim F(\boldsymbol{\phi}_{z_i}),\ z_i | \boldsymbol{\pi} \sim \textstyle\sum_{k\geq 1} \pi_k \delta_k, \\
&\pi_k = p \cdot \pi_k{}' \textstyle\prod_{l=1}^{k-1} (1 - p \cdot \pi_l{}'), \\
&\pi_k' \sim \text{Beta}(1, \gamma),\ \boldsymbol{\phi}_k \sim H,
\end{aligned}$$

where $\boldsymbol{\pi} = \{\pi_k; k \geq 1\}$. The detail of the sampler is almost identical to PAM and left for Supplement A.12.

2.5.3 Inference on Clusters

Like all BNP models, both PAM and FSBP produce random clusters and their associated posterior distributions. The slice sampler in the previous section produces Markov chain Monte Carlo (MCMC) samples that eventually converge to the true joint posterior distribution of all the parameters. These samples are used for posterior inference, including estimating a single clustering outcome of the observations, even though the posterior distribution of the clusters is available. We discuss the corresponding inference under PAM next. We consider two approaches but only present one of them below, leaving the other approach to the Supplement A.11.

First, for the mth MCMC sample, denote the matrix of cluster memberships of all the observations as $\boldsymbol{Z}^{(m)} = \{z_{ij}^{(m)};\ i = 1, \ldots, n_j, j = 1, \ldots, J\}$, and the vector of observations in the jth group as $\boldsymbol{Z}_j^{(m)} = \{z_{1j}^{(m)}, \ldots, z_{n_j j}^{(m)}\}$. These z values can be sampled in Step 8 of Algorithm 1. Let $\boldsymbol{t}_j^{(m)} = \left\{ t_1^{(m)}, \ldots, t_{K_j^{(m)}}^{(m)} \right\}$ denote the labels of these clusters, which are the unique values of the cluster memberships in $\boldsymbol{Z}_j^{(m)}$. Here $K_j^{(m)}$ represents the number of clusters in group j for the mth sample. Then the set and number of common clusters

between groups j and j' are given by $\boldsymbol{t}_j^{(m)} \cap \boldsymbol{t}_{j'}^{(m)}$ and its cardinality, respectively, and the set and number of unique clusters for group j are given by $\boldsymbol{t}_j^{(m)}$ *mod* $\boldsymbol{Z}^{(m)} \backslash \boldsymbol{Z}_j^{(m)}$ and its cardinality, respectively. Here, operation A *mod* B for two sets A and B is redefined as the unique elements in A but not B, and $\boldsymbol{Z} \backslash \boldsymbol{Z}_j$ means the set after removing $\boldsymbol{Z}_j$ from $\boldsymbol{Z}$. Through these operations, for every MCMC sample m we obtain clustering results for the observations. Together, all the MCMC samples constitute an approximation of the posterior distributions of the clusters.

To produce a point estimate of the clustering result, we follow the approach in Wade and Ghahramani [2018] to estimate an optimal partition through a decision-theoretic approach that minimizes the variation of information [Meilă, 2007]. This optimal partition is then used as a "point estimate" of the random clusters obtained from PAM or FSBP posterior inference.

2.6 Simulation Study

2.6.1 Simulation Setup

We assess the performance of PAM and FSBP via simulation. The ASP model is not evaluated since it is simply a PAM for a single group. In the simulation, we generate data from a Gaussian finite mixture model with specific clustering patterns, and apply BNP models as a prior for data analysis. Posterior inference from the BNP models is then compared to the simulation truth. We compare PAM with CAM and HDP in Scenarios 1 and 2, and FSBP with DP in Scenario 3. In all simulations, the variance is $\sigma^2 = 0.6$ in the Gaussian mixture.

Scenario 1 - PAM Univariate data We consider three cases under Scenario 1 to assess the performance of PAM under various clustering patterns.

Case 1: Unique Clusters We generate data from groups that have non-overlapping clus-

ters. This extreme case provides an evaluation of models' performance to capture unique clusters. We assume $J = 2$ groups, each with $n_j = n = 200$ samples. Within each group, the observations are generated from a mixture of four Gaussian distributions with distinct means. In mathematical terms, we have

$$f(y_{ij}) \propto \sum_{k=1}^{4} \frac{1}{4} N(m_{jk}, \sigma^2),\ i = 1, \ldots, n,\ j = 1, 2,$$

where m_{jk} represents the cluster mean for Group j and cluster k. There is a total of 8 clusters across two groups. For Group 1, the cluster means are $m_{1k} \in \{0, 4, 8, 12\}$, and for Group 2, the cluster means are $m_{2k} \in \{-16, -12, -8, -4\}$.

Case 2: A Single Common Cluster In this case, we assume the presence of one common cluster between groups. Specifically, we consider $J = 3$ groups, each comprising $n_j = n = 100$ samples. The observations in each group again follow a mixture of Gaussian distributions. A common cluster with mean 0 is shared across all three groups, while each group possesses its own unique clusters. Details regarding the cluster means and weights in each group can be found in Table A.2 in Supplement A.13.

Case 3: Nested Clusters In this case, taken from Denti et al. [2021], nested clusters are generated across groups. Specifically, let $J = 6$ groups. Ascending number and overlapping clusters are generated via the mixture of Gaussian distributions given by

$$f(y_{ij}) \propto \sum_{k=1}^{j} \frac{1}{j} N(m_k, \sigma^2), \quad i = 1, \ldots, n_j,\ j = 1, \cdots, 6,$$

where the cluster means $m_k \in \{0, 5, 10, 13, 16, 20\}$ for $j = 1, \ldots, 6$. Therefore, there are j true clusters in group j and clusters in group j is nested in group $(j + 1)$, with only the first cluster $N(m_1, \sigma^2)$ shared across all six groups. We test two sub-cases of Case 3 by setting the number of observations in group $n_j = n_A$, where $n_A \in \{50, 100, 150\}$, or by setting $n_j = n_B \times j$, where $n_B \in \{10, 20, 40\}$.

Scenario 2 - PAM Multivariate data In this scenario, each observation y_{ij} is assumed to be a 3-dimensional vector. Additionally, we consider $J = 3$ groups, each with $n_j = n$ subjects, where $n \in \{50, 100, 200\}$. The multivariate observations are generated from a mixture of multivariate Gaussian distributions, with the cluster means and weights shown in Table A.3 in Supplement A.13. The true covariance matrix is assumed to be the identity matrix. There are a total of five clusters across all groups: Group 1 possesses all five clusters, Group 2 has three clusters (clusters 1, 3 and 4), and Group 3 has two clusters (clusters 2 and 3). Note that cluster 3 is the only common cluster across all three groups.

In both scenarios, we compare the performance of PAM with HDP and CAM. We obtain a point-estimate of clustering results based on the procedure in Wade and Ghahramani [2018] and assess the models' performance based on the following criteria.

1. The total number of clusters, number of common clusters, and number of unique clusters based on the estimated clustering results.

2. The adjusted Rand index (ARI) [Hubert and Arabie, 1985] between the estimated clustering results and the ground truth, with a value closer to 1 indicating better performance.

3. The normalized Frobenius distance (NFD) [Horn and Johnson, 1990] between the estimated posterior pairwise co-clustering matrices and the true co-clustering structure, with a value closer to 0 indicating better performance.

These metrics have been routinely adopted in the literature, e.g., in Denti et al. [2021].

Scenario 3 - FSBP Univariate data In this scenario, we evaluate the performance of FSBP. We consider $n = 300$ observations, each following a mixture of five Gaussian

distributions with distinct means, given by

$$f(y_i) \propto \sum_{k=1}^{5} \frac{1}{5} N(m_k, \sigma^2),\ i = 1, \ldots, n,$$

where $m_k \in \{0, 3, 6, 9, 12\}$. FSBP is then compared to DP, and the performance is assessed based on the estimated posterior density function, as well as the number of clusters inferred with each method.

2.6.2 Simulation Results for PAM and FSBP

Scenario 1 We generate 30 datasets for each available sample size in each case. For each simulated dataset, we adopt standard prior settings for the hyperparameters in model (2.10). Specifically, we use the NIG distribution as the base measure H, with hyperparameters $\mu_0 = 0$, $\kappa_0 = 0.1$, $\alpha_0 = 3$ and $\beta_0 = 1$. We use Jeffrey's prior for p_j's, i.e., $a = b = 0.5$. Lastly, we set $a_{\alpha_0} = b_{\alpha_0} = a_\gamma = b_\gamma = 3$ for the gamma priors of the concentration parameters α_0 and γ. We collect an MCMC sample of 10,000 iterations after 10,000 iterations of burn-in. The Markov chains mix well.

We present the simulation results for all three cases in Table 2.1. The winning performance is highlighted in bold font. We use notation G_j to denote the group j. Full results in terms of cluster numbers are presented for cases 1 and 2. For Case 3, results from one sample size $n = 150$ is presented, and we selected clustering results for G5 and G6 as they are representative of the models' distinct behavior. Full results are reported in Supplement A.13. Overall, PAM exhibits competitive performance in terms of identifying the correct number of clusters and ARI/NFD scores. PAM also is the most stable method consistently producing the smallest standard deviations. In Case 1, PAM is superior in capturing the special clustering structure where no clusters are shared across groups. In contrast, HDP seems to struggle in identifying the unique clusters in this case. These can be found in

"Number of clusters" for "All groups" in the table. To further examine the model fitting in Case 1, Figure A.6 in Supplement A.13 shows that HDP (middle panel) sometimes merge two different clusters in the posterior inference, leading to under-estimated cluster numbers. CAM and PAM appear to be able to avoid this and report mostly the correct clustering structure. In Case 2, CAM and HDP are the better methods, both able to capture the sole common cluster more often than PAM. These two cases seem to show distinct behavior of PAM vs CAM and HDP. We confirm this in case 3. In particular, PAM is more likely to identify the correct number of clusters across all groups as the average number of clusters under PAM is 5.97 compared to 4.97 for CAM and 4.27 for HDP. However, CAM is better at identifying common clusters, say between G5 and G6, while PAM is more capable of finding the unique cluster in G6.

Since by definition PAM allows the weight π_{jk} of cluster k in group j to be zero, it can output $\Pr(\pi_{jk} = 0|\text{Data})$ which can be interpreted as cluster k is absent from group j. In addition, $\Pr(\pi_{jk} > 0|\text{Data})$ describes the posterior probability that cluster k is present in group j, and $\Pr(\pi_{jk} > 0, \pi_{j'k} > 0|\text{Data})$ the posterior probability that cluster k is shared between groups j and j'. More generally, a posterior probability of different configurations of π's can be used to estimate more complex clustering patterns. In contrast, common-atoms models like HDP and CAM assign $\Pr(\pi_{jk} = 0|\text{Data}) \equiv 0$ and $\Pr(\pi_{jk} > 0|\text{Data}) \equiv 1$ by definition. A work-around might be to report the frequency of a cluster sampled in the MCMC iterations in a group, which can be used as an approximation to the probability a cluster belongs to the group.

To illustrate our point, in Table 2.2, we present posterior summaries of PAM and CAM using a simulated dataset under Case 1, in which clusters 1-4 belong to group 2 (G2) and 5-8 to group 1 (G1). Both PAM and CAM report small, < 0.01, but non-zero estimated cluster weights (posterior mean). However, PAM reports large posterior probabilities of "Unique in G1" for clusters 5-8 and of "Unique in G2" for clusters 1-4, while those posterior probabilities

Case	Metrics		CAM	HDP	PAM	Truth
Case 1	# of clusters	All groups	7.87 (0.35)	5.80 (0.66)	**7.97 (0.18)**	8
		G1	4.07 (0.25)	3.37 (0.56)	**3.97 (0.18)**	4
		G2	3.87 (0.35)	2.43 (0.50)	**4.00 (0.00)**	4
	# of common clusters		0.07 (0.25)	**0.00 (0.00)**	**0.00 (0.00)**	0
	# of unique clusters	G1	**4.00 (0.00)**	3.37 (0.56)	3.97 (0.18)	4
		G2	3.80 (0.48)	2.43 (0.50)	**4.00 (0.00)**	4
		ARI	0.96 (0.04)	0.67 (0.09)	**0.97 (0.02)**	
		NFD	**0.01 (0.01)**	0.09 (0.03)	**0.01 (0.01)**	
Case 2	# of clusters	All groups	**5.00 (0.00)**	5.03 (0.18)	5.07 (0.25)	5
		G1	**2.27 (0.91)**	1.70 (0.47)	1.67 (0.48)	2
		G2	3.80 (0.71)	**2.83 (0.53)**	2.60 (0.50)	3
		G3	2.97 (0.76)	2.13 (0.35)	**2.00 (0.00)**	2
	# of common clusters	All groups	**1.20 (0.71)**	0.47 (0.51)	0.30 (0.47)	1
		G1 and G2	1.80 (0.96)	**0.60 (0.56)**	0.33 (0.48)	1
		G1 and G3	1.40 (0.77)	**0.80 (0.61)**	0.63 (0.49)	1
		G2 and G3	2.03 (0.93)	**0.70 (0.47)**	0.53 (0.51)	1
	# of unique clusters	G1	0.27 (0.45)	0.77 (0.50)	**1.00 (0.26)**	1
		G2	1.17 (0.65)	**2.00 (0.00)**	2.03 (0.18)	2
		G3	0.73 (0.52)	**1.10 (0.31)**	1.13 (0.35)	1
		ARI	0.85 (0.04)	**0.87 (0.05)**	**0.87 (0.04)**	
		NFD	0.04 (0.01)	**0.03 (0.01)**	**0.03 (0.01)**	
Case 3 ($n_j = 150$)	# of clusters	All groups	4.97 (0.49)	4.27 (0.58)	**5.97 (0.62)**	6
		G5	**4.43 (0.50)**	3.33 (0.48)	4.13 (0.57)	5
		G6	**4.60 (0.62)**	3.17 (0.46)	4.53 (0.68)	6
	# of common clusters	G5 and G6	**4.43 (0.50)**	3.13 (0.35)	3.47 (0.68)	5
	# of unique clusters	G5	**0.00 (0.00)**	**0.00 (0.00)**	0.53 (0.51)	0
		G6	0.13 (0.35)	0.03 (0.18)	**0.90 (0.40)**	1
		ARI	**0.95 (0.02)**	0.90 (0.04)	0.95 (0.03)	
		NFD	0.07 (0.02)	0.03 (0.01)	**0.02 (0.01)**	

Table 2.1: Simulated univariate data in Scenario 1. Clustering performance of CAM, HDP, and PAM is evaluated based on the following metrics: number of clusters across all and individual groups, number of common clusters across all groups and pairwise groups, number of unique clusters within each group, Adjusted Rand Index (ARI), and normalized Forbenius distance (NFD). Entries represent the Mean (SD) over 30 datasets. Bold entries mean the corresponding model performs the best with the corresponding metric. Note that the notation G1 to G6 refers to Group 1 to Group 6, respectively.

are 0's for CAM. Therefore, PAM gives a more interpretable summary based on the posterior probability of atom weights equal to or greater than 0.

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
True mean		-16	-12	-8	-4	0	4	8	12
True weight	G1	0.00				0.25	0.25	0.25	0.25
	G2	0.25	0.25	0.25	0.25	0.00			
		PAM Estimates (CAM Estimates)							
Mean		-15.99 (-15.98)	-11.81 (-11.81)	-7.74 (-4.82)	-4.07 (-3.74)	0.11 (0.13)	3.91 (3.92)	7.85 (7.84)	11.96 (11.93)
Weight	G1	< 0.01(< 0.01)				0.19 (0.24)	0.24 (0.24)	0.23 (0.21)	0.25 (0.25)
	G2	0.22 (0.22)	0.27 (0.26)	0.22 (0.12)	0.29 (0.22)	< 0.01(< 0.01)			
Unique in G1 $\widehat{\Pr}(\pi_{1k} > 0, \pi_{2k} = 0 \mid \text{Data})$		0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.75 (0.00)	0.84 (0.00)	0.81 (0.00)	0.86 (0.00)
Unique in G2 $\widehat{\Pr}(\pi_{1k} = 0, \pi_{2k} > 0 \mid \text{Data})$		0.78 (0.00)	0.78 (0.00)	0.68 (0.00)	0.79 (0.00)	0.02 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)

Table 2.2: Posterior summaries of CAM and PAM for a randomly selected dataset in Case 1 of Scenario 1. Reported estimates for Mean and Weight are posterior means. The last two rows correspond to MCMC-estimated posterior probabilities of a cluster has zero weight in one group and positive in the other.

Scenario Two For the multivariate data scenario, we use the following prior settings for the hyperparameters in (2.10). The NIW distribution (NIW($\boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{\Psi}$)) is used as the base measure H, with hyperparameters $\boldsymbol{\mu}_0 = \mathbf{0} = \{0, 0, 0\}$, $\kappa_0 = 0.1$, $\nu_0 = 4$ and $\boldsymbol{\Psi} = I_3$, where I_3 is the 3×3 identity matrix. Similar to Scenario 1 , we use Jeffrey's prior for p_j. We also set $a_{\alpha_0} = b_{\alpha_0} = a_\gamma = b_\gamma = 3$ in the gamma priors for the concentration parameters α_0 and γ. For simplicity, we only report the model accuracy on the number of clusters, ARI, and NFD for this simulation. We generate 30 datasets for each sample size, and summarize the results in Table A.4 in Supplement A.13. The results indicate that all three methods have high accuracy in the multivariate data simulation. PAM performs competitively with the other two methods in terms of the ARI and NFD metrics when the sample size is large ($n \geq 100$).

Scenario 3 We generate 30 datasets for the simulation of FSBP, each with 300 observations. Standard priors are adopted for the hyperparameters. Specifically, we use NIG distribution as the base measure H, set $\mu_0 = 0$, $\kappa_0 = 0.1$, $\alpha_0 = 3$, $\beta_0 = 1$, use Jeffrey's prior for p ($p \sim \text{Beta}(0.5, 0.5)$), and use Gamma$(3, 3)$ for the hyperparameter γ. We collect an

MCMC sample of 5,000 iterations after 5,000 iterations of burn-in. The Markov chains mix well. The results are presented in Figure A.7 in Supplement A.13. On average, both methods successfully capture the five true clusters. However, FSBP provides more accurate posterior inference on the cluster distributions (top panel) and cluster numbers (bottom panel).

2.7 Case Studies

We apply PAM to two real-life datasets, one from a microbiome project and the other related to treatment of warts. The microbiome data demonstrate PAM's performance for count data and the warts data consists of multivariate observations.

2.7.1 Microbiome Dataset

The microbiome dataset, reported in O'Keefe et al. [2015], measures microbiota abundance for 38 healthy middle-aged African Americans (AA) and rural Africans (AF). The study aims to investigate the effect of diet swap between individuals of AF and AA, as traditional foods for these populations differ. The 38 study participants are instructed to follow their characteristic diet, such as a low-fat and high-fiber diet for AF and a high-fat and low-fiber diet for AA, for two weeks, and then swap diets for another two weeks. We consider cluster the subjects based on the measured microbiota abundance, in terms of counts of operational taxonomic units (OTUs), which reflect the recurrences of the corresponding OTUs in a particular ecosystem [Jovel et al., 2016, Kaul et al., 2017]. For more background, refer to O'Keefe et al. [2015] and Section 4 of Denti et al. [2021]. Hereafter, we use the term "expression" and "counts" interchangeably in this application.

To apply PAM, or more specifically DPAM (due to the discrete data of OTU counts), we treat each subject as a group, and counts of different OTUs as observations within a group. Following the same data-preprocessing steps as in Denti et al. [2021], we obtain 38 subjects (17 AF and 21 AA) with 119 OTUs. Note that all the OTUs are the same, and so a

cluster here refers to a group of OTU counts, just like in Denti et al. [2021]. When applying to demonstrate the CAM model, Denti et al. [2021] use the entire dataset with a goal to generate nested clusters of subjects and OTU counts within subjects. The proposed PAM cannot cluster subjects and therefore we randomly select four subjects from the dataset for analysis. In a future work, we will consider extend PAM to allow nested clustering and will apply the new model to the full dataset.

We randomly select four subjects as four groups, two AAs (individuals 5 and 22) and two AFs (individuals 13 and 14), from the dataset. We remove the OTUs that had zero expression in all four individuals from the selected data. In the end, we obtain a dataset with $J = 4$ individuals (groups) and $n_j = 109$ OTUs (observations). The histograms of the microbiome populations of the four selected individuals are shown in Supplement A.14.

Let x_{ij} denote the observed OTU count for OTU i from individual j. For inference, similar to Denti et al. [2021], we incorporate the average OTU frequencies for subject j, denoted as $\eta_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, as a scaling factor in the latent variable y_{ij} of the DPAM model in (2.12). This leads to the following distribution for the change of variables:

$$y_{ij}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 \sim N(\eta_j \mu_{z_{ij}}, \eta_j^2 \sigma^2_{z_{ij}}) \leftrightarrow \frac{y_{ij}}{\eta_j}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 \sim N(\mu_{z_{ij}}, \sigma^2_{z_{ij}}) \tag{2.23}$$

The prior hyperparameters follow the same settings as in Scenario 1 of the simulation study, and we present the analysis results in Table 2.3.

PAM reports a total of eight estimated clusters across the four individuals: clusters 1 and 2 are shared by all four individuals (with posterior probabilities of 1.00 and 0.95, respectively), cluster 7 is shared among individuals 5 (AF), 13 (AA), and 14 (AA) (with posterior probability of 0.96), and cluster 8 is shared among individuals 5 and 22 (both from AF, with posterior probability of 0.93). The other clusters are unique to a specific individual (with posterior probabilities of 0.60, 0.60, 0.42, and 0.51, respectively, for clusters 3 to 6). Based on the optimal partition of OTUs, we plot the taxa counts (TC) of OTUs grouped

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Mean		0.07(0.01)	0.53(0.04)	1.75(0.20)	1.50(0.26)	2.21(0.27)	3.73(0.36)	9.89(1.21)	74.21(8.99)
Weight	ID 5	0.56	0.26	0.11	< 0.01	< 0.01	< 0.01	0.05	0.02
	ID 22	0.84	0.12	< 0.01	< 0.01	< 0.01	0.02	< 0.01	0.02
	ID 13	0.77	0.11	< 0.01	< 0.01	0.10	< 0.01	0.02	< 0.01
	ID 14	0.74	0.10	< 0.01	0.11	< 0.01	< 0.01	0.05	< 0.01
Unique in	ID 5	0.00	0.00	0.60	0.00	0.00	0.00	0.00	0.06
	ID 22	0.00	0.00	0.00	0.00	0.00	0.51	0.00	0.00
	ID 13	0.00	0.00	0.00	0.00	0.42	0.00	0.00	0.00
	ID 14	0.00	0.00	0.00	0.60	0.00	0.00	0.00	0.00

Table 2.3: Estimated clusters based on Wade and Ghahramani [2018] using posterior samples from PAM. A total of eight OTU count clusters is estimated. “Mean" and “Weight" are the posterior mean estimates of the cluster mean and weight. Parantheses are the standard deviations. An entry in a row corresponding to “Unique in" is the posterior probability that a cluster (column) is only present in the individual (row) but not in other individuals (rows).

by all eight estimated clusters as well as by both clusters and individuals in Figure A.9 in Supplement A.14. Note that for easy demonstration of clusters across individuals, we have manually reordered the clusters in ascending order based on the cluster mean. The boxplots illustrate the clusters and their distributions across individuals.

We report an interesting finding related to the PAM clustering of OTU counts. Specifically, the counts of the OTU *Prevotella melaninogenica* is in cluster 8, which has the highest expression and is shared (both the cluster and the OTU counts) only by AF individuals 5 and 22. This finding is consistent with previous studies that have shown that the individuals with a predominance of *Prevotella spp.* are more likely to consume fiber, which is a typical component of an African diet [Graf et al., 2015, Preda et al., 2019].

2.7.2 Warts Dataset

In this example, we consider a publicly available dataset reporting treatment of patients with warts. Two groups of patients are considered, treated with immunotherapy or cryotherapy. Each treatment group contains medical records for 90 patients, and for each patient, six baseline characteristics (covariates) are reported, including the patient’s gender, age (Age), time elapsed before treatment (Time), the number of warts (NW), the type of warts, and the

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Mean	Age	18.53	31.66	23.68	27.36	19.64	24.51	16.55
	Time	6.19	6.71	8.63	6.96	7.38	4.41	3.80
	NW	2.44	7.13	8.44	2.75	7.98	7.54	4.28
	Area	68.41	40.82	195.16	389.20	312.65	87.78	6.41
Weight	Immunotherapy (G1)	0.15	0.68	0.02	0.12	0.03	< 0.01	< 0.01
	Cryotherapy (G2)	0.31	0.17	0.06	< 0.01	< 0.01	0.36	0.10
$\widehat{\Pr}(\pi_{1k} > 0, \pi_{2k} > 0 \mid \text{Data})$		1.00	0.81	0.66	0.31	0.32	0.28	0.40
$\widehat{\Pr}(\pi_{1k} > 0, \pi_{2k} = 0 \mid \text{Data})$		0.00	0.19	0.00	0.69	0.68	0.00	0.00
$\widehat{\Pr}(\pi_{1k} = 0, \pi_{2k} > 0 \mid \text{Data})$		0.00	0.00	0.34	0.00	0.00	0.72	0.60

Table 2.4: Estimated clusters based on Wade and Ghahramani [2018] using posterior samples from PAM. Reported are the cluster means, weights, probabilities of common, $\widehat{\Pr}(\pi_{1k} > 0, \pi_{2k} > 0 \mid \text{Data})$, and unique clusters in either the immunotherapy (G1) or the cryotherapy (G2), $\widehat{\Pr}(\pi_{jk} > 0, \pi_{j'k} = 0 \mid \text{Data})$, for the inferred seven clusters. Observed data consist of 4-dimensional covariate vectors for all the patients. The covarites are, "Age", "Time" referring to the time elapsed before treatment, "NW" referring to number of warts, and "Area" referring to the surface area of warts of the patient.

surface area of warts in mm^2 (Area). Additionally, patients' responses to the corresponding treatments are also recorded.

To better understand potential differences between responders to the two treatments, we use PAM to cluster the covariate values of the responders. The sets of responders contain 71 patients for the immunotherapy group and 48 for the cryotherapy group. We construct an observation y_{ij} as $q = 4$ -dimensional vector including four continuous baseline covariates, Age, Time (time of sickness before treatment) , NW (number of warts) , and Area (surface area of warts) . We set the hyperparameters of the priors to be the same as in Scenario 2 of the simulation and apply PAM to the dataset of two groups of warts patients.

Table 2.4 reports inference results. PAM identifies a total of seven clusters, three of which are shared between the immunotherapy and cryotherapy groups, and the remaining four unique to a group. The table reveals that, among all responders, individuals with younger age, a short time elapsed from treatment (less than five months), and small surface area of warts form unique clusters in the cryotherapy group. On the other hand, those who were not treated for a longer time and had a large surface area of warts (over 300 mm^2) form distinct clusters in the immunotherapy group. Furthermore, it seems that the number

of warts does not provide much information in determining a better treatment option for warts patients.

These findings are consistent with results from previously published studies. For instance, Khozeimeh et al. [2017b] found that patients younger than 24 years old showed a better response to cryotherapy, and patients who received cryotherapy within six months had a very high probability of being cured. This is consistent with the information implied by clusters 6 and 7, which are unique to the cryotherapy group. Moreover, another study by Khozeimeh et al. [2017a] developed an expert system with fuzzy rules, and one such rule for immunotherapy is "If (types of wart is Plantar) and (time elapsed before treatment is VeryLate) then (response to treatment is Yes)." In Khozeimeh et al. [2017a]'s expert system, time elapsed before treatment longer than six months is considered "VeryLate". This rule echoes the common and unique clusters for the immunotherapy group found by PAM. In the unique clusters 4 and 5, and the common clusters 1 to 3, the time before treatment was 6.96, 7.38, 6.19, 6.71 and 8.63 months, respectively, all larger than six months. Additional results are illustrated in Supplement A.15, which shows the cluster membership of each patient. The figure indicates that patients with a large area of warts are unique to the immunotherapy group, while those with a younger age are mostly from the cryotherapy group.

2.8 Discussion

We have introduced a novel BNP model constructed with a novel technique called Atom Skipping. A stochastic process that uses atom-skipping on single datasest is ASP, which has a mean process of FSBP, an extension of DP that has higher expected number of clusters than DP with the same concentration parameter. Extending ASP to multiple groups forms the proposed PAM, where the weights of clusters in PAM are allowed to be exactly zero in some groups, effectively removing these clusters from those groups. Thus, PAM generates an interpretable clustering structure. Additionally, PAM accommodates count data and

multivariate observations. Efficient slice samplers are developed for PAM, with substantial modifications due to atom-skipping. In simulation studies, PAM demonstrated its robustness across different simulation scenarios. In particular, it performed the best when there are many unique clusters with little or no common ones among the groups. In the case studies, PAM also produces sensible results.

There are some limitations to our current work. Firstly, the model is unable to cluster groups (i.e., distributional clusters), unlike NDP and CAM. However, we are currently working on a separate model that extends PAM to cluster nested data at both group and observational levels. Secondly, the model has not been applied to real datasets consisting of different types of covariates, such as binary and multinomial covariates. Finally, longitudinal data is another interesting direction for extending the model.

CHAPTER 3

PAM-HC: A BAYESIAN NONPARAMETRIC CONSTRUCTION OF HYBRID CONTROL FOR RANDOMIZED CLINICAL TRIALS USING EXTERNAL DATA

3.1 Introduction

Randomized clinical trials (RCTs) are the gold standard to objectively assess the superiority of a new drug over a control. It is widely acknowledged that RCTs with a 1:1 randomization ratio yield the highest statistical power. Nevertheless, enrolling patients under such a design can sometimes be challenging like in rare diseases, pediatric trials, or settings where an r:1 (r >1) randomization ratio is used to enhance patient enrollment. With the availability of historical trial data or real-world data (RWD) like the electronic health records, statistical models have been proposed to borrow information in these data to estimate treatment effects more accurately. For example, when a standard of care has been widely tested or administered in a patient population, the available response data could be used to augment the control arm in a clinical trial and form a hybrid control (HC). Due to the augmented information in the HC, a more precise estimation of the treatment effect could be achieved.

In drug development, information borrowing from external data for an RCT is regulated. The U.S. FDA recently has released a guidance document on the design and conduct of external controlled trials for Drug and Biological products [FDA, 2023b]. The document emphasizes the importance of ensuring that the trial eligibility criteria can be applied to the external control arm in order to obtain a population comparable to that of the clinical trial. Thus, it is critical to ensure the similarity of patient baseline characteristics between the external data and the current RCT. Another issue discussed is the extent to which one may borrow information [Chen et al., 2020]. Historical trial data and RWD can be larger than the current RCT data, and one must be cautious not to let the borrowed information

dominate the results of the current trial. Therefore, oftentimes information from external data is discounted to avoid overwhelming the statistical inference of the current study.

In the literature, many statistical methods have been proposed to borrow information from external data for RCTs. Bayesian models like the power prior (PP) [Ibrahim and Chen, 2000], commensurate prior (CP) [Hobbs et al., 2012], robust meta-analytic-predictive prior (RMAP) [Schmidli et al., 2014], and the latent exchangeability prior (LEAP) [Alt et al., 2023] all construct hierarchical models for external and current trial data. Specifically, the PP method assumes that the treatment outcome parameters are the same between the current trial and the external data. It utilizes a discounting factor to reflect the user's prior belief regarding the similarity between the historical data and the current trial, thereby discounting the likelihood of the historical data. CP uses different parameters for the current trial and the external data, but assuming the parameters of the trial follow a prior distribution with a mean equal to the parameters for the external data. RMAP employs a mixture of a meta-analytic-predictive (MAP) prior, which is an informative prior, and a vague prior (robust component) to mitigate the potential issue of over-borrowing. However, these three methods do not consider situations in which only a subset of patients in the external data are comparable to the current study. Recently, Alt et al. [2023] propose the LEAP prior, which dynamically borrows information from historical trials assuming a subset of individuals in the historical data are exchangeable with the current study.

Another class of methods utilizes propensity scores (PS) to identify matched patients between the external data and the current trial data. For example, Chen et al. [2020] proposed the propensity-score integrated composite likelihood (PSCL) method to address the situation where only a subset of patients in the external data are comparable to the current trial. However, as noted by Chandra et al. [2023], King and Nielsen [2019], and Zhao [2004], matching patients based on their PS does not necessarily imply matching of covariates. In addition, PS-based methods are often sensitive to model specification for

estimating the PS.

A recent study by Chandra et al. [2023] introduces a third class of methods that utilizes Bayesian nonparametric models (BNP) to identify “common clusters" of patients across the current trial and the external data. The BNP model has the ability to automatically cluster patients in the current trial and the external data based on baseline covariates. Their method, called CA-PPMx, assumes the external data consists of all the subpopulations that are present in the current trial.

Motivated by CA-PPMx, we propose a BNP approach, called PAM-HC, for constructing an HC arm for an RCT using external data. Here, PAM refers to a BNP model in Bi and Ji [2023] that generates overlapping clusters. Different from Chandra et al. [2023], we assume that the current RCT and external data may share common subpopulations of patients, while each may consist of unique ones as well. In other words, PAM can identify common and unique subpopulations across observations arranged in groups. Using PAM, the HC is constructed by only borrowing information from the common subpopulations between the external data and the control in the RCT. We employ a power prior to discount the information borrowing. In addition, the BNP models in the proposed PAM-HC method generate random clusters characterized by a posterior distribution. Therefore, the entire statistical inference is model based and variabilities on the clustering and treatment effect estimates are properly accounted for.

In the subsequent sections, we first review the PAM method in Section 3.2. We provide a detailed description to the proposed PAM-HC method in Section 3.3. We present the simulation setup and results in Section 3.4 comparing our method to the PSCL method and a baseline method that does not involve information borrowing. In Section 3.5, we showcase an application of PAM-HC to real-life trial data. Finally, we conclude our work in Section 3.6.

3.2 Review Plaid Atoms Model (PAM)

We assume that there is an ongoing RCT with an $r:1\,(r>1)$ randomization ratio between the treatment and control arms. In addition, assume there exists an external dataset comprising patients with the same disease that have been treated by the same control in the current RCT. For example, the control arm in the RCT and external data could be a standard chemotherapy and the treatment arm in the RCT could be a new immunotherapy. We denote the treatment arm of the current RCT as group 1 ($j=1$), the control arm as group 2 ($j=2$), and the external data as group 3 ($j=3$). Assume there are n_j patients in group j, for $j=1,2$, and 3. Therefore, $N=n_1+n_2$ is the sample size of the RCT, and $n_1/n_2\approx r$ due to the $r:1$ randomization ratio. We denote $\boldsymbol{y}_j=\{y_{i,j}\}_{i=1}^{n_j}$ the patients outcome data in group j, and $\boldsymbol{x}_{i,j}=\{x_{i,j,1},\ldots,x_{i,j,p}\}$ a p-dimensional random vector consisting of the patient i's baseline covariates in group j.

We assume that the current trial and the external data consist of heterogeneous subpopulations of patients based on their baseline characteristics (covarivates), with patients from the same subpopulation forming a cluster. For the rest of the discussion, we use the terms "cluster" and "subpopulation" of patients interchangeable. Lastly, we refer to overlapping subpopoulations of patients that are present in multiple groups as "common clusters". Conversely, we use the term "unique cluster" to describe the subpopulation of patients that is only present in one but not other groups.

To find patients clusters in the current RCT and the external data, we adopt PAM in Bi and Ji [2023]. An example of the clustering structures identified by PAM is shown in Figure 3.1.

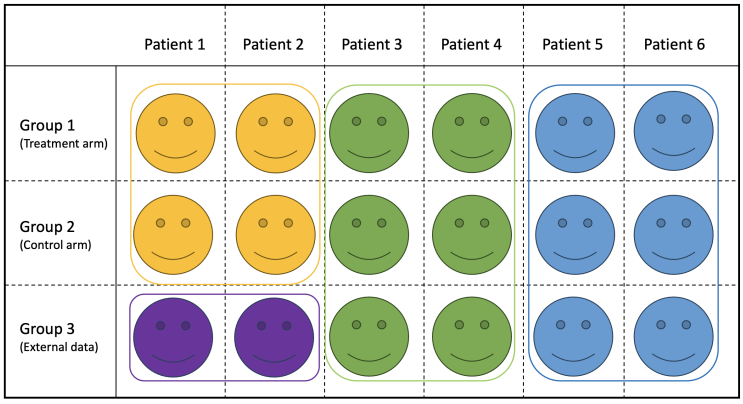

Figure 3.1: An illustration of clustering pattern under PAM. Rows represent groups and columns are patients within each group. The three groups correspond to the current RCT's treatment and control arms, and the external data. There are four homogeneous subpopoulations of patients (clusters) represented by colored smiley faces in blue, green, purple, and yellow. The boxes represent the common or unique clusters. For example, the green cluster is common and shared across all three groups, while purple is unique to group 3.

In this illustration, each color represents a cluster. The blue and green clusters are common and the purple cluster is unique to group 3. If one knows the clustering pattern in Figure 3.1, one would only borrow information from the green and blue clusters in the external data because they are shared with the trial data. However, one should not borrow information from the purple cluster since it is unique to the external data.

A brief review of the statistical model is provided next. For more detail refer to Bi and Ji [2023]. Denote $Z_{i,j}$ as the cluster membership indicator for patient i in group j, where $\{Z_{i,j} = k\}$ indicates that patient i in group j is assigned to cluster k. The Bayesian

nonparametrics model in PAM is given by a hierarchical structure as follows:

$$
\begin{aligned}
&\boldsymbol{x}_{i,j}|Z_{i,j},\{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}_{k=1}^{\infty} \sim MVN\left(\boldsymbol{\mu}_{Z_{i,j}},\boldsymbol{\Sigma}_{Z_{i,j}}\right)\\
&Z_{i,j}|\{\pi_{j,k}\}_{k=1}^{\infty} \sim \textstyle\sum_{k=1}^{\infty}\pi_{j,k}\delta_k(Z_{i,j}),\ \pi_{j,k}=\pi'_{j,k}\prod_{l=1}^{k-1}(1-\pi'_{j,l})\\
&f(\pi'_{j,k}|\boldsymbol{\beta},\alpha_0,p_j)=p_j\times f_{\text{Beta}}\left(\alpha_0\beta_k,\alpha_0\left(1-\textstyle\sum_{l=1}^{k}\beta_l\right)\right)+\underbrace{(1-p_j)\times I(\pi'_{j,k}=0)}_{(*)} \qquad (3.1)\\
&p_j|a,b\sim Beta(a,b),\ \boldsymbol{\beta}|\gamma\sim GEM(\gamma),\\
&(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)|\boldsymbol{\mu}_0,\lambda,\boldsymbol{\Psi},\nu\sim NIW(\boldsymbol{\mu}_0,\lambda,\boldsymbol{\Psi},\nu),
\end{aligned}
$$

where MVN stands for the multivariate normal distribution, $\delta_A(B)$ is the indicator function ($\delta_A(B)=1$ if $B\in A$ or $B=A$, and $\delta_A(B)=0$ otherwise), $f_{\text{Beta}}(a,b)$ is the density function of the Beta(a,b) distribution with mean $a/a+b$, $GEM(\gamma)$ represents the Griffths, Engen and McCloskey distribution [Pitman, 2002], distribution NIW stands for to the normal-inverse-Wishart distribution, parameter $\pi_{j,k}$ represents the cluster weights of cluster k in group j, and parameters $(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ denote (mean, covaraince matrix) of the kth cluster. Additional priors can be assigned to hyperparameters γ and α_0. Model (3.1) in PAM largely resembles the well known hierarchical Dirichlet Process (HDP) model [Teh et al., 2004], which induces common clusters across groups. PAM adds a unique model component $(*)$, which allows some common cluster to have zero weight in group j, thereby producing unique clusters.

Through (3.1), PAM generates a joint posterior distribution of all the parameters including the cluster membership $\boldsymbol{Z}=\{Z_{i,j}\}_{\forall i,j}$. Through $\boldsymbol{Z}$ we can easily find the common and unique clusters. Since a posterior distribution of $\boldsymbol{Z}$ is generated, the number of clusters and clustering memberships themselves are random. In PAM-HC, we utilize the features of PAM that identifies common and unique clusters, upon which we build models and inference for constructing a hybrid control arm and estimating treatment effects.

3.3 Methodology

3.3.1 Clustering of patients

In PAM, the cluster membership matrix $\boldsymbol{Z} = \{Z_{i,j}\}_{\forall i,j}$ indicates which clusters are common and which are unique. Since patients in the same cluster are believed to be "similar" in their covariates, they are expected to react similarly to the control treatment, under the assumption that the covariates have captured all the factors that are related to treatment response. Of course, when unmeasured confounders are present, the proposed method will be inadequate. Such investigation is beyond the scope of this paper and left for future work.

Let $A_{j,k} = \{i : Z_{i,j} = k, i = 1, \ldots, n_j\}$ represent the set of patients in the k-th cluster in group j. Also denote the set of cluster labels in each group j as $O_j = \{k : A_{j,k} \neq \emptyset\}$. We define the set of common cluster labels between a pair of groups as $C_{j,j'} = \{k : k \in O_j \cap O_{j'} \text{ for } j \neq j',\ j, j' \in \{1, 2, 3\}\}$. PAM clusters the treatment arm, control arm, as well as the external data simultaneously, and our focus is on set $C_{2,3}$, which consists of the common cluster labels between the current trial control arm and the external data. The proposed PAM-HC method augments $A_{2,k}$ by borrowing information from patients in $A_{3,k}$ for cluster(s) $k \in C_{2,3}$. Figure 3.2 provides a schematic overview of PAM-HC.

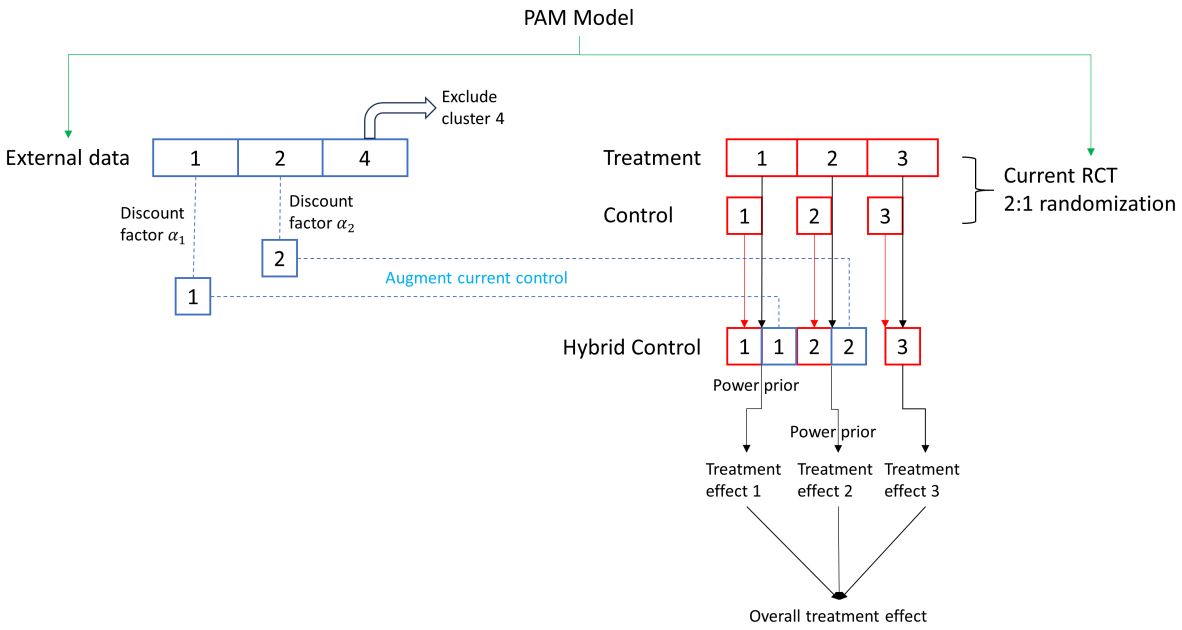

Figure 3.2: A stylized illustration of PAM-HC. Numbers in the boxes denote cluster labels. Boxes in red color represent patients in the RCT and those in blue represent patients in the external data. Cluster 4 is unique to the external data and therefore is not used for forming the HC. Cluster 3 is unique to the RCT and therefore is not augmented.

To further illustrate our idea, Consider a hypothetical cluster membership $\boldsymbol{Z}$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} 1 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 3 & & & \\ 1 & 1 & 2 & 2 & 4 & 4 \end{bmatrix},$$

where the rows are groups and columns are patients. Based on $\boldsymbol{Z}$, there are four clusters, three ($k = 1, 2, 3$) for groups 1 and 2, and three ($k = 1, 2, 4$) for group 3. Clusters 1 and 2 are shared across groups 1 and 2, while cluster 4 is unique for group 3. Also, we have for group 1: $A_{1,1} = \{1, 2\}$, $A_{1,2} = \{3, 4\}$, $A_{1,3} = \{5, 6\}$, and $O_1 = \{1, 2, 3\}$; for group 2: $A_{2,1} = \{1\}$, $A_{2,2} = \{2\}$, $A_{2,3} = \{3\}$, and $O_2 = \{1, 2, 3\}$; and for group 3: $A_{3,1} = \{1, 2\}$, $A_{3,2} = \{3, 4\}$, $A_{3,4} = \{5, 6\}$, and $O_3 = \{1, 2, 4\}$. The set of common clusters

are $C_{1,2} = \{1, 2, 3\}$, $C_{1,3} = \{1, 2\}$, and $C_{2,3} = \{1, 2\}$ between the treatment and control arms, between the treatment arm and the external data, and between the control arm and the external data, respectively. We focus on the set $C_{2,3} = \{1, 2\}$, and for clusters $k \in C_{2,3}$, i.e., $k = 1$ and $k = 2$, construct an HC by borrowing information from patients in the external data belonging to clusters 1 and 2, but not cluster 4. For illustrative purposes, the stylized example assumes all the clustering memberships are fixed. In actual modeling, PAM-HC generates random clustering memberships which allows for assessment of variabilities in subsequent inference of treatment effects. This will be clear in Section 3.3.4 later.

3.3.2 Information borrowing across common clusters

We use the power prior to borrow information across the common clusters between the control and external data in order to form an HC. Similar to Chandra et al. [2023], our approach involves performing a regression analysis of the outcome variables $y_{i,j}$ on the corresponding covariates $\boldsymbol{x}_{i,j}$ through the clusters $A_{j,k}$, $j = 1, 2, 3$. Denote $\theta_{1,k}$ and $\theta_{2,k}$ the cluster-specific response parameter in the treatment and control arms, respectively, for cluster k. We use a simple hierarchical model for $\theta_{1,k}$ for $k \in O_1$, the clusters in the treatment group. Recall $\boldsymbol{y}_1 = \{y_{i,1}\}_{i=1}^{n_1}$ are the observed patient responses in group 1, the treatment group. We assume

$$y_{i,1}|Z_{i,1} = k, \theta_{1,k} \sim F(\theta_{1,k}),$$

$$\theta_{1,k} \sim \pi_0(\theta_{1,k}),$$

where $F(\cdot)$ denotes the likelihood of y. For continuous outcome, $F(\theta) = N(\mu, \sigma^2)$, and $\theta_{1,k} = (\mu_{1,k}, \sigma^2_{1,k})$, and for binary outcome, $F(\theta) = Bern(q)$, with $\theta_{1,k} = q_{1,k}$. In addition, $\pi_0(\theta_{1,k})$ is a vague prior for $\theta_{1,k}$. For $\theta_{2,k}$, the response parameter for group 2, the control arm, we use

$$y_{i,2}|Z_{i,2} = k, \theta_{2,k} \sim F(\theta_{2,k}),$$

and the power prior for $\theta_{2,k}$. Recall $\boldsymbol{y}_2 = \{y_{i,2}\}_{i=1}^{n_2}$ and $\boldsymbol{y}_3 = \{y_{i,3}\}_{i=1}^{n_3}$ are the observed responses in groups 2 (the current control) and 3 (external data), respectively. We assume the prior of $\theta_{2,k}$ is given by

$$p(\theta_{2,k}|\boldsymbol{y}_3, A_{3,k}, \alpha_k) \propto \left[\prod_{i \in A_{3,k}} f(y_{i,3}|\theta_{2,k})\right]^{\alpha_k} \pi_0(\theta_{2,k})$$

where $f(\cdot|\theta)$ is the p.d.f of $F(\theta)$, $\pi_0(\theta_{2,k})$ is a vague prior for $\theta_{2,k}$, and $\alpha_k \in [0, 1]$ is a discount factor (or power parameter) for cluster k. We estimate α_k as a deterministic function of cluster weights $\pi_{j,k}$ and cluster membership $\boldsymbol{Z}$. Specifically, when $k \in O_3$ but $k \notin C_{2,3}$, we set $\alpha_k = 0$. In words, for unique clusters in the external data, there is no borrowing and the power parameter $\alpha_k = 0$. Otherwise, the cluster is shared between the external data and control, and the discount factor α_k is given by Chen et al. [2020]:

$$\alpha_k = \min\left(\pi^*_{2,k} \cdot I,\ n_{3,k}\right) / n_{3,k}, \tag{3.2}$$

where

$$I = \frac{r-1}{r+1} N,$$

is the total number of patients to be borrowed from external data so that the information in the HC is of the same amount as the treatment arm. The value of I is easily derived based on the $r:1$ randomization ratio of the RCT and the desired 1:1 matching between the treatment and HC. In addition, $n_{3,k} = |A_{3,k}|$, where $|.|$ denotes the cardinality of the set, and $\pi^*_{2,k}$ is the proportion with which we want to borrow from the I patients for cluster k. For example, if $N = 300$ and $r = 2$, then $I = 100$ which means one would borrow information from up to 100 patients from the external data to form an HC so that the amount of information in the HC matches that of information in the treatment arm. Within each cluster k, we use the following steps to compute $\pi^*_{2,k}$. Recall the current trial is randomized with a ratio of $r:1$,

$r > 1$, and $\pi_{j,k}$ is the probability (or weights) of cluster k in group j (PAM model (3.1)). We want to augment the control to form an HC in which the information is worth the same number of patients as the treatment arm in each cluster k. Mathematically, this means

$$N \cdot \pi_{1,k} \cdot \frac{r}{r+1} = N \cdot \pi_{2,k} \cdot \frac{1}{r+1} + I \cdot \pi^*_{2,k}, \quad \text{where } I = \frac{r-1}{r+1}N.$$

Solving for $\pi^*_{2,k}$, we have

$$\pi^*_{2,k} = \frac{1}{r-1}(r\pi_{1,k} - \pi_{2,k}). \tag{3.3}$$

Equation (3.3) leads to a solution for (3.2), and hence a value for α_k. In practice, to prevent negative values of $\pi^*_{2,k}$ (when the control arm already has a larger number of patients in cluster k than the treatment arm), we use $\pi^*_{2,k+}$, i.e., $\pi^*_{2,k+} = \pi^*_{2,k}$ if $\pi^*_{2,k} > 0$, and $\pi^*_{2,k+} = 0$ otherwise.

The construction of α_k in (3.2) and $\pi^*_{2,k}$ in (3.3) adaptively borrows more or less information for cluster k based on the imbalance in the patient assignment between the treatment and control in cluster k. This is perhaps more clear in (3.3). Due to the $r:1$ randomization, the term $(r\pi_{1,k} - \pi_{2,k})$ in (3.3) reflects the difference in the expected sample sizes between treatment and control for cluster k. When the term has a larger value, there is a larger difference (imbalance) of information between the two arms, which leads to a large $\pi^*_{2,k}$, and therefore larger α_k. In other words, when the treatment arm has more patients than the control arm, PAM-HC borrows more to augment the control.

3.3.3 Estimate treatment effects

Conditional on $\boldsymbol{Z}$, the cluster membership, we assume treatment effects are cluster specific. Due to randomization, we assume $C_{1,2} = O_1 = O_2$, i.e., the treatment and control arms share all the clusters and there are no unique clusters in each of the two arms. Under this

setting, denote Δ_k the cluster-specific treatment effect, for $k \in C_{1,2}$, given by

$$\Delta_k = \theta_{1,k} - \theta_{2,k}.$$

In rare cases where there are unique clusters in the treatment or control arms, we use an ad-hoc rule to merge the unique clusters to a common cluster in $C_{1,2}$ that has the shortest distance in terms of L2-norm between the cluster means. The overall treatment effect can be computed as a weighted average of the cluster-specific treatment effects Δ_k. Conditional on $\boldsymbol{Z}$, we let

$$\Delta(\boldsymbol{Z}) = \sum_{k \in O_1} \pi_{1,k} \Delta_k = \sum_{k \in O_1} \pi_{1,k} (\theta_{1,k} - \theta_{2,k}) \tag{3.4}$$

be the conditional overall treatment effects. We could either use $\{\pi_{1,k}\}$ or $\{\pi_{2,k}\}$ as the weights, which are in principle close to each other due to randomization. However, we decide to use $\{\pi_{1,k}\}$ since the treatment arm (group $j = 1$) is expected to have more patients and therefore lead to more stable estimates of clustering weights. The (unconditional) overall treatment effect is given by $\Delta = E[\Delta(\boldsymbol{Z})]$.

3.3.4 Inference

Bi and Ji [2023] develop a slice sampler to generate posterior samples via Markov chain Monte Carlo (MCMC) simulations. We use $m = 1, \ldots, M$ to index the M MCMC samples and use a generic notation $\hat{X}^{(m)}$ to denote the m-th sample for random variable X. Also, for simplicity, let $\boldsymbol{D} = (\boldsymbol{y}_1, \boldsymbol{x}_1, \boldsymbol{y}_2, \boldsymbol{x}_2, \boldsymbol{y}_3, \boldsymbol{x}_3)$ denote the entire data, including the data from the RCT and external source. Note that the posterior mean of overall treatment effect can

be expressed as an integration of (3.4) over the posterior distributions of θ's, π, and $\boldsymbol{Z}$, i.e.,

$$\begin{aligned}
&\mathrm{E}[\Delta|\boldsymbol{D}] \equiv \mathrm{E}[\mathrm{E}[\Delta(\boldsymbol{Z})|\boldsymbol{Z},\boldsymbol{D}]] \\
&= \int \left\{ \int \textstyle\sum_{k\in O_1} \pi_{1,k}(\theta_{1,k}-\theta_{2,k}) p(\theta_{1,k}|\boldsymbol{D},\boldsymbol{Z}) p(\theta_{2,k}|\boldsymbol{D},\boldsymbol{Z}) p(\boldsymbol{\pi}_1|\boldsymbol{D}) d\boldsymbol{\theta} d\boldsymbol{\pi}_1 \right\} p(\boldsymbol{Z}|\boldsymbol{D}) d\boldsymbol{Z}
\end{aligned} \tag{3.5}$$

where O_1, $\pi_{1,k}$ and k are all functions of $\boldsymbol{Z}$, $\boldsymbol{\theta} = \{(\theta_{1,k}, \theta_{2,k}) : k \in O_1\}$, and $\boldsymbol{\pi}_1 = \{\pi_{1,k} : k \in O_1\}$. Using the MCMC samples, the posterior mean (3.5) is estimated as follows. For the m-th sample, let $\hat{O}_j^{(m)}$ be the cluster labels for group j. Let cluster $k^{(m)} \in \hat{O}_1^{(m)}$, then we compute the m-th posterior sample of the cluster-specific treatment effect as

$$\hat{\Delta}^{(m)}_{k^{(m)}} = \hat{\theta}^{(m)}_{1,k^{(m)}} - \hat{\theta}^{(m)}_{2,k^{(m)}}.$$

Finally, the overall treatment effect $\hat{\Delta}^{(m)}$ can be obtained with $\hat{\Delta}^{(m)}_{k^{(m)}}$ and the weights $\hat{\boldsymbol{\pi}}_1^{(m)}$ using equation (3.4):

$$\hat{\Delta}^{(m)} = \sum_{k^{(m)} \in \hat{O}_1^{(m)}} \hat{\pi}^{(m)}_{1,k^{(m)}} \hat{\Delta}^{(m)}_{k^{(m)}},$$

and the posterior mean treatment effect is estimated as $\sum_{m=1}^{M} \hat{\Delta}^{(m)}/M$. Also, given the posterior sample $\{\hat{\Delta}^{(m)},\ m = 1, \ldots, M\}$, we can easily compute various quantities of interest, such as the standard deviation of the overall treatment effect. Additionally, it allows us to determine the posterior probability of a significant treatment effect, denoted as

$$\Pr(\Delta > \epsilon | \text{Data}),$$

for some minimal clinically meaningful treatment effect ϵ.

3.4 Simulation Studies

3.4.1 Simulation Setup

We generate covariates values $\boldsymbol{x}_{i,1}$ and $\boldsymbol{x}_{i,2}$ for the current RCT by simulating from a mixture of three multivariate normal distributions. Specifically,

$$\boldsymbol{x}_{i,j} \sim 0.3 \times MVN(\boldsymbol{2}_3, \boldsymbol{I}) + 0.4 \times MVN(\boldsymbol{0}_3, \boldsymbol{I}) + 0.3 \times MVN(\boldsymbol{-2}_3, \boldsymbol{I}),\ j = 1, 2,$$

where notation $\boldsymbol{a}_3 = [a, a, a]^T$ and $\boldsymbol{I}$ is the 3 by 3 identity matrix. We generate the covariates of the external data, denoted as $\boldsymbol{x}_{i,3}$, under three different scenarios:

- Scenario 1 (Superset): we introduce an *additional cluster* with a cluster mean of $\boldsymbol{-4}_3$ and a covariance matrix of $\boldsymbol{I}$. The weights assigned to the clusters are also different from those used in the RCT:

$$\boldsymbol{x}_{i,3} \sim 0.2 \times MVN(\boldsymbol{2}_3, \boldsymbol{I}) + 0.3 \times MVN(\boldsymbol{0}_3, \boldsymbol{I}) + 0.3 \times MVN(\boldsymbol{-2}_3, \boldsymbol{I}) + \underbrace{0.2 \times MVN(\boldsymbol{-4}_3, \boldsymbol{I})}_{\text{unique}}.$$

- Scenario 2 (Overlap): the external data shares some clusters with the current RCT while also having a unique cluster. Specifically, we *remove* the cluster with mean $\boldsymbol{-2}_3$ from the RCT, and similar to Scenario 1, we *add a cluster* with mean $\boldsymbol{-4}_3$ and covariance $\boldsymbol{I}$ to the external data:

$$\boldsymbol{x}_{i,3} \sim 0.5 \times MVN(\boldsymbol{2}_3, \boldsymbol{I}) + 0.3 \times MVN(\boldsymbol{0}_3, \boldsymbol{I}) + \cancel{0.3 \times MVN(\boldsymbol{-2}_3, \boldsymbol{I})} + \underbrace{0.2 \times MVN(\boldsymbol{-4}_3, \boldsymbol{I})}_{\text{unique}}.$$

- Scenario 3 (Subset): In this scenario, the external data is a subset of the current RCT. Similar to Scenario 2, we *remove* the cluster with mean $\boldsymbol{-2}_3$ from the RCT, but do

not add any clusters to the external data:

$$\boldsymbol{x}_{i,3} \sim 0.5 \times MVN(\mathbf{2}_3, \boldsymbol{I}) + 0.5 \times MVN(\mathbf{0}_3, \boldsymbol{I}) + 0.3 \times MVN(-\mathbf{2}_3, \boldsymbol{I}).$$

For outcomes $\boldsymbol{y}$, we assume that they are associated with the baseline covariates. We consider both continuous and binary outcomes.

For continuous outcomes, we simulate the outcome $y_{i,j}$ using a linear regression given by:

$$y_{i,j} \sim \beta_{0,j} + \boldsymbol{\beta}^T \boldsymbol{x}_{i,j} + \epsilon_{i,j},\ \epsilon_{i,j} \sim N(0,1),\ j = 1, 2, 3.$$

For binary outcomes, we generate $P(y_{i,j} = 1|\boldsymbol{x}_{i,j})$ using a logistic regression given by:

$$\text{logit } P(y_{i,j} = 1|\boldsymbol{x}_{i,j}) = \beta_{0,j} + \boldsymbol{\beta}^T \boldsymbol{x}_{i,j},$$

and then simulate $y_{i,j}$ from the Bernoulli distribution with the success probability $P(y_{i,j} = 1|\boldsymbol{x}_{i,j})$. We fix $\beta_{0,2} = \beta_{0,3} = 0$, and set $\boldsymbol{\beta} = \mathbf{1}_3$. We consider four cases $\beta_{0,1} = 0$, 1, 2, or 3, which is the intercept (also treatment effect) in the outcome regression models in the treatment arm. Furthermore, we assume that the current RCT uses a randomization ratio of $r = 2:1$. We simulate the RCT data with two different sample sizes, $N = \{300, 450\}$. For the case where $N = 300$, we place 200 patients in the treatment arm and 100 in the control. We generate a total of 300 patients for the external data. For the case where $N = 450$, we place 300, 150, and 450 patients in the treatment arm, control arm, and the external data, respectively.

The parameter of interest is the treatment effect Δ, which represents the difference in mean responses between the treatment and control arms. Consequently to $\beta_{0,1} = 0$, 1, 2, and 3, the true values of Δ are 0, 1, 2, and 3, respectively, for the continuous outcome, and 0.00%, 9.42%, 15.76%, and 19.52% , respectively, for the binary outcome.

We compare the proposed PAM-HC method with two other methods. The first method, referred to as the baseline method, estimates the treatment effect using only the RCT data and does not borrow information from the external data. Inference of treatment effect is based on the maximum likelihood estimation (MLE). The second method is the PSCL method proposed by Chen et al. [2020]. We implement the PSCL method using three different strategies, corresponding to the ways propensity scores are estimated. Namely, they are PSCL 1, PSCL 2, and PSCL 3, referring to the PSCL method with propensity scores estimated using a first-order logistic regression, random forest, and a logistic regression with model selection, respectively. All three versions of PSCL are implemented in the R package "PSRWE" [Wang and Chen, 2022]. Finally, for each scenario, we simulate 100 datasets.

We use the following priors and hyperparameters for the proposed PAM-HC method. For PAM model (3.1), we set $a = b = 0.5$, and utilize a $Gamma(3, 3)$ prior for the hyperparameters γ and α_0. Additionally, we set $\boldsymbol{\mu}_0 = \mathbf{0}_3$, $\Psi = \boldsymbol{I}$, $\lambda = 0.1$, and $\nu = 3$. For the power priors in PAM-HC, we adopt a normal distribution $N(\mu, \sigma^2)$ for the continuous outcome and a Bernoulli distribution Bern(p) for the binary outcome. Conjugate priors are chosen for the parameters in the sampling model. Specifically, we use a normal-inverse-gamma prior $NIG(0, 0.1, 3, 3)$ for (μ, σ^2) and $Beta(0.5, 0.5)$ for p. These are standard priors that are not informative and routinely applied in the literature. Lastly, we run an MCMC simulation of 10,000 iterations, with burn-in period of 5,000 iterations.

We summarize the posterior cluster membership using an optimal clustering method [Meilă, 2007] to obtain a point estimate. To assess the clustering accuracy in comparison to the ground truth cluster membership of each patient, we use the adjusted Rand index (ARI) [Hubert and Arabie, 1985] and the normalized Frobenius distance (NFD) [Horn and Johnson, 1990]. More detail can be found in Bi and Ji [2023]. Lastly, we assess the performance of PAM-HC in terms of the estimated overall treatment effect, including the mean, standard deviation, bias, and mean squared error (MSE) across all simulated datasets.

3.4.2 Simulation results

We assess similarity in the distributions of covariates between the treatment and HC. We first check that under PAM-HC, if the distribution of covariates of the treatment arm is similar to the distribution of the hybrid control arm. Specifically, within each estimated cluster, the distributions of covariates should be similar between the two arms. We randomly selected one dataset in each scenario with $N = 300$. We plot the density of the covariates by estimated clusters in Figure 3.3 below for Scenario 1, and in Figures A.2.1 and A.2.2 in Appendix for scenarios two and three, respectively.

Figure 3.3 shows that the distribution of each covaraite are indeed similar between the treatment arm, control arm, and the external data, for each inferred cluster k. Furthermore, the estimated cluster centers are shown in Table 3.1 below. In addition, the cluster-specific treatment effects for the three selected datasets are reported in Tables A.2.1 and A.2.2 in Appendix. In the selected examples, we see that PAM-HC is able to correctly identify the number of clusters in these selected examples. The estimated cluster centers are also close to the true values in their corresponding scenarios. In addition, Tables A.2.1 and A.2.2 suggest that treatment effects for clusters are well estimated.

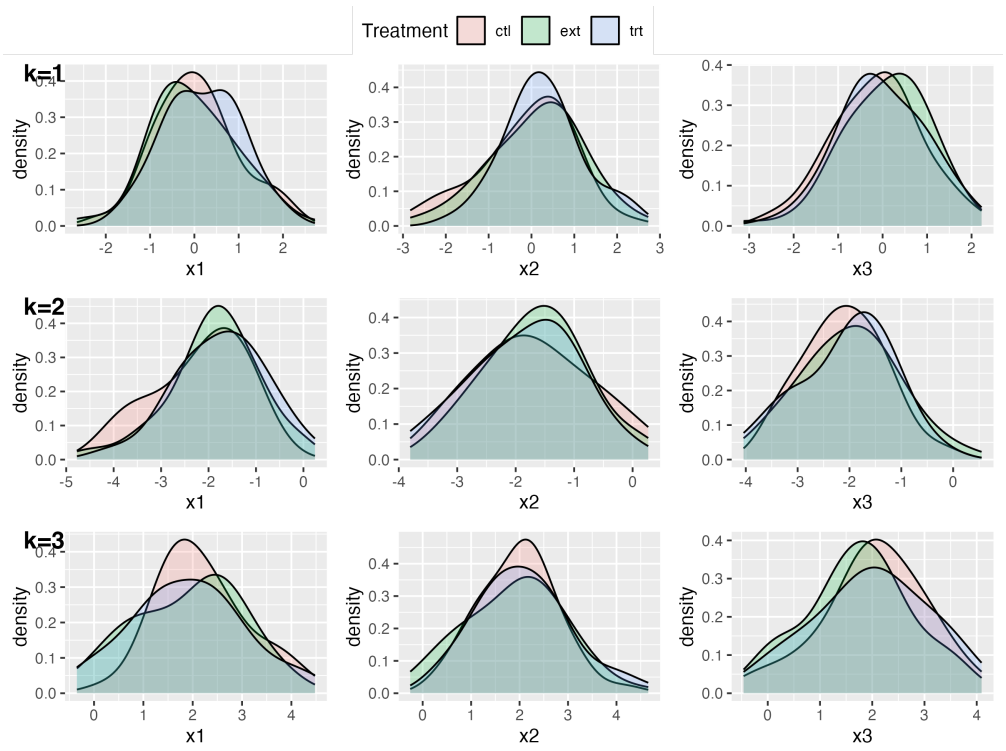

Figure 3.3: The covariate density plots of one simulated data in Scenario 1. The rows represent three clusters estimated by PAM-HC.

Sc	Cluster	Cluster mean			Groups (j)
		x_1	x_2	x_3	
Sc 1	1	$0.08_{0.91}$ (0)	$0.14_{1.03}$ (0)	$0.01_{0.96}$ (0)	1,2,3 (1,2,3)
	2	$-1.89_{0.95}$ (-2)	$-1.72_{0.88}$ (-2)	$-2.03_{0.90}$ (-2)	1,2,3 (1,2,3)
	3	$1.98_{1.05}$ (2)	$1.99_{0.93}$ (2)	$1.91_{1.02}$ (2)	1,2,3 (1,2,3)
	4	$-4.13_{0.96}$ (-4)	$-4.03_{0.90}$ (-4)	$-4.15_{0.96}$ (-4)	3 (3)
Sc 2	1	$0.01_{0.90}$ (0)	$0.19_{1.07}$ (0)	$-0.01_{0.95}$ (0)	1,2,3 (1,2,3)
	2	$-1.88_{0.97}$ (-2)	$-1.95_{1.14}$ (-2)	$-1.96_{0.81}$ (-2)	1,2 (1,2)
	3	$2.02_{0.98}$ (2)	$2.05_{1.11}$ (2)	$1.94_{0.98}$ (2)	1,2,3 (1,2,3)
	4	$-3.85_{1.05}$ (-4)	$-4.08_{0.90}$ (-4)	$-3.88_{0.85}$ (-4)	3 (3)
Sc 3	1	$-0.09_{0.92}$ (0)	$-0.66_{0.75}$ (0)	$-0.38_{0.93}$ (0)	1,2,3 (1,2,3)
	2	$-1.95_{0.90}$ (-2)	$-2.00_{1.12}$ (-2)	$-1.99_{0.81}$ (-2)	1,2 (1,2)
	3	$1.29_{1.38}$ (2)	$1.61_{1.10}$ (2)	$1.33_{1.28}$ (2)	1,2,3 (1,2,3)

Table 3.1: Cluster mean estimated by PAM-HC on selected examples from each of the three scenarios. The entries for columns x_1, x_2, and x_3 are posterior mean_{SD} (truth), and estimated clusters (truth) for the last column.

To further assess the similarity of covariate distributions between treatment and the hybrid control arms, we follow the procedure outlined in Chandra et al. [2023] and apply the Bayesian Additive Regression Tree (BART) model [Chipman et al., 2010]. For each estimated cluster k, we aggregate data from all three groups and create a dummy variable T_i indicating whether patient i belongs to the treatment arm ($T_i = 1$) or the hybrid control arm ($T_i = 0$). We then carry out a 10-fold cross-validation, with 9-folds used as the training data and 1-fold as the testing data. We apply BART to predict whether an observation in the testing data belongs to the treatment arm or not. The results are reported as the Area Under the ROC Curve (AUC). A value around 0.5 indicates no difference between the patients in the current treatment arm and the patients in the hybrid control. We randomly select 10 datasets in

each scenario, and the corresponding results show that across all scenarios, the range of mean AUC values is between 0.525 and 0.544, all around 0.5. This indicates that the covariate distributions for the treatment and hybrid control arms are similar and indistinguishable by BART.

Next, we report the clustering results of PAM-HC for all simulated datasets. The true number of clusters is four in Scenario 1 and Scenario 2, and three in Scenario 3. For ARI and NFD, the closer the value of ARI is to 1 or the value of NFD to 0, the better the clustering result of the method. Table A.2.3 in Appendix shows the estimated total number of clusters across all groups, as well as the ARI and the NFD of the estimated clusters compared to the true cluster membership. On average, the number of estimated cluster is accurate, close to its truth in all cases. The ARI and NFD values are satisfactory, improving with increasing sample size.

Lastly, we present the estimated treatment effect, its standard deviation, and the mean squared error (MSE). We compare these results with the baseline method and PSCL 1-3. Table A.2.4 in Appendix provides a summary of the results. In the case of a smaller sample size ($N = 300$), PAM-HC shows the lowest MSE in Scenario 2 and lowest bias in Scenario 3, for all four values of Δ. The performance of the PAM-HC design improves further with a larger sample size ($N = 450$). PAM-HC is also comparable to the PSCL methods in MSE and much smaller than the baseline method in Scenarios 1 and 2. We also evaluate the performance of PAM-HC with binary outcomes, and the results are shown in Table A.2.5 in Appendix. Similar to the continuous outcome, PAM-HC exhibits desirable performance.

Scenario 2 is an interesting case in which some but not all clusters are shared across groups. This is where PAM-HC excels in its performance. Since PAM-HC is designed to capture the pattern of overlapping clusters, it leads to more precise information borrowing and better performance. To see this, we assess the “inclusion probability", defined as the probability of each patients in the external data being borrowed for HC. Mathematically, this

is equal to $\Pr(Z_{i,3} \notin C_{2,3} \mid \boldsymbol{D})$. In words, if a patient i in the external data group ($j = 3$) is not in a common cluster with the control, the patient is not "borrowed" for forming the HC. In Scenario 2, for patients in the unique cluster 4 in external data, the mean and SD inclusion probability across the patients are 3.47% (SD = 0.17) and 2.51% (SD = 0.15) for sample sizes $N = 300$ and 450, respectively. For patients in other common clusters in the external data, the inclusion probability are all greater than 91%. These results demonstrate that PAM-HC is able to adaptively borrow based on the overlapping status of each cluster.

3.5 Application

3.5.1 Background and Dataset

We consider clinical trials for patients with Atopic Dermatitis (AD). AD is a significant contributor to skin-related disability globally, characterized by recurrent eczematous lesions and intense itch [Simpson et al., 2022]. In this application, we analyze data from the control arms (placebo) of three historical trials (with NCT numbers NCT03569293, NCT03607422 and NCT03568318) for AD. The treatment arms and their data are not available for analysis due to confidentiality. Specifically, the three control arms of the three historical trials share similar inclusion and exclusion criteria, and the patients in the three control arms all receive a placebo. The control arms consist of 263, 265, and 306 patients.

Each trial reports several baseline characteristics of the patients, including their gender, age, race, ethnicity, body mass index (BMI), baseline body surface area affected (BSA), and baseline Eczema Area and Severity Index (EASI). Additionally, the trials record the EASI score at the 16-week mark to assess the progression of the patients' disease. The primary outcome is the percent change in the EASI score from baseline to 16 weeks, denoted as:

$$\text{EASI}_{\text{pc}} = \frac{\text{Baseline EASI} - \text{16-week EASI}}{\text{Baseline EASI}} \times 100\%$$

The binary response to the treatment is defined as $\text{EASI}_{\text{pc}} \geq 75\%$. In other words, the binary outcome, denoted as y, is defined as

$$y = \begin{cases} 1, & \text{if } \text{EASI}_{\text{pc}} \geq 75\%; \\ 0 & \text{otherwise.} \end{cases}$$

To illustrate PAM-HC, we pretend the control arm of trial one is the treatment arm of a hypothetical RCT and randomly select 131 patients from the control arm of trial two to serve as the RCT control. Therefore, we construct a hypothetical RCT of 394 patients with a randomization ratio of $2:1$. We examine the distributions of the covaraites between the hypothetical treatment and control arms and find no major differences (results not shown). Lastly, we use the 306 patients of trial three as the external data with which we build a hybrid control for the RCT. The observed response rates are of 25%, 21%, and 34%, for trials one, two, and three, respectively. And the overall mean responses is roughly 28% across all three trials. We use these data for PAM-HC in a null scenario.

Alternatively, to construct a trial with an actual treatment effect (the alternative case), we follow the findings of Simpson et al. [2022] that reports a response rate of 80% under the treatment arm. We spike in response data in trial one and use it as the treatment arm in the hypothetical RCT. Specifically, we generate a treatment arm (consisting of the 263 patients from trial one) based on the following procedure. To make sure that the outcome is related to the covariates, we first fit a logistic regression model with the original outcome of trial one ($y_{i,1}$) as the dependent variable and the four covariates ($\boldsymbol{x}_{i,1}$) as the independent variables. We then fixed the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ and conducted a grid search to find a value of $\tilde{\beta}_{0,1} = 2.33$ that satisfied the condition

$$\tilde{y}_{i,1} \sim Bern(p_i), \text{ logit } p_i = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_{i,1} + \tilde{\beta}_{0,1},$$

$\tilde{\boldsymbol{y}}_1 = \{\tilde{y}_{i,1}\}_{i=1}^{263}$, and $\Pr(\tilde{\boldsymbol{y}}_1 = 1) \approx 80\%$. The true treatment effect is roughly $80\% - 28\% \approx 52\%$ after spike-in. These data form the alternative scenario.

3.5.2 Analysis Results

The posterior mean number of clusters by PAM-HC is 4.15 (SD = 0.36), and PAM-HC generates a point-estimate of cluster structure that consists of four common clusters that are shared across all three arms without a unique cluster. Table 3.2 below summarizes the cluster means as well as the cluster-specific treatment effects of PAM-HC for the null and alternative scenarios.

Cluster k	Weights			Cluster mean				Cluster-specific treatment effect	
	$\pi_{1,k}$	$\pi_{2,k}$	$\pi_{3,k}$	Age	Baseline EASI	BSA	BMI	Null	Alternative
Cluster 1	0.14	0.18	0.16	$19.86_{0.70}$	$18.55_{0.46}$	$24.56_{1.07}$	$22.78_{0.50}$	$0.077_{0.113}$	$0.659_{0.102}$
Cluster 2	0.20	0.18	0.23	$37.79_{1.37}$	$39.04_{1.53}$	$66.30_{2.52}$	$29.11_{0.79}$	$-0.031_{0.074}$	$0.561_{0.076}$
Cluster 3	0.42	0.40	0.32	$41.26_{1.54}$	$21.68_{0.40}$	$31.79_{1.26}$	$27.30_{0.74}$	$-0.081_{0.068}$	$0.500_{0.065}$
Cluster 4	0.24	0.24	0.29	$21.56_{0.88}$	$33.82_{1.87}$	$57.06_{2.85}$	$22.12_{0.40}$	$-0.080_{0.091}$	$0.519_{0.091}$

Table 3.2: Estimated cluster mean and cluster-specific treatment effect using the data of the AD trial. The entries are posterior mean_{SD}.

We report the estimated treatment effects using the PAM-HC method as well as the baseline and PSCL 1-3 methods. The results are shown in Table 3.3. We observe that all methods report small treatment effects that are not statistically significant under the null case. However, when borrowing information from external data, the PAM-HC and PSCL methods report negative treatment effects as opposed to a positive treatment effect reported by the baseline method which does not borrow information from external data. This is expected since the response rate of the external data is 34%, higher than those of the hypothetical treatment (25%) and control (21%) arms. For the alternative case, all methods find non-zero treatment effects, although PAM-HC and the PSCL methods report a lower treatment effect compared to the baseline. Again, this is expected since when borrowing

from the external data with 34% response, the control response rate is expected to increase from 21% in the hybrid control. Lastly, the proposed PAM-HC method reports an accurate estimation of the treatment effect in the alternative case, which is around the ground true of 52%.

	Method	$\Delta_{\text{SD}(\Delta)}$	Significance
Null Case	Baseline	$0.026_{0.044}$	P-value: 0.281
	PSCL1	$-0.045_{0.149}$	P-value: 0.618
	PSCL2	$-0.034_{0.149}$	P-value: 0.590
	PSCL3	$-0.041_{0.315}$	P-value: 0.551
	PAM-HC	$-0.049_{0.037}$	$\Pr(\Delta > 0 \mid \text{data}) = 0.09$
Alternative Case	Baseline	$0.627_{0.042}$	P-value: 0.000
	PSCL1	$0.596_{0.149}$	P-value: 8.4e-6
	PSCL2	$0.606_{0.139}$	P-value: 6.5e-6
	PSCL3	$0.598_{0.140}$	P-value: 1.0e-5
	PAM-HC	$0.539_{0.036}$	$\Pr(\Delta > 0 \mid \text{data}) = 1.00$

Table 3.3: Estimated treatment effects for the AD trial, using the proposed PAM-HC method, the baseline method, and three versions of PSCL method.

3.6 Discussion

In this study, we introduce the PAM-HC method to augment the control arm of an RCT using external data and improve the estimation of treatment effects. A key innovation is to identify common subpopulations of patients between the RCT and the external data and allow information to be borrowed only across these common subpopulations. We find that PAM-HC performs well when compared to existing methods in the simulation and case study, especially when not all patient subpopulations are shared between RCT and external data.

Thanks to the model-based inference on all the unknown parameters using BNP models, PAM-HC is powerful in reporting posterior distributions of cluster-specific treatment effects, overall treatment effects, and the random clusters themselves.

However, it is important to acknowledge the limitations of the current method. Firstly, the assumption that covariates are continuous variables restricts the applicability of PAM to handle binary and categorical variables. Another limitation lies in the underlying assumption that the covariates used for clustering and inference includes all the relevant confounders. Future work is ongoing to address these issues.

CHAPTER 4

A BAYESIAN ESTIMATOR OF SAMPLE SIZE

4.1 Introduction

4.1.1 Motivation

In novel drug development, the clinical objective is to establish the drug's effectiveness and safety. Randomized clinical trials (RCTs) are the gold standard to achieve the objective. As it is usually impractical to include all patients from a disease population, RCTs take a random sample of certain size to address the clinical question through statistical inference. Due to the law of large numbers, the larger the sample size, the more precise the estimated treatment effect is but the more costly the trial is as well. Therefore, in practice a sample size estimation (SSE) approach is applied to compute an appropriate sample size for a trial to balance the tradeoff between precision in statistical inference and its cost.

We consider a Bayesian framework for SSE. The research is motivated by the recent change in early-phase oncology drug development that recommends randomized comparison of multiple doses. In the past two decades, oncology drug development has benefited from biological and genomics breakthroughs and novel cancer drugs are no longer based on cytotoxic one-size-fits-all mechanism like chemicals or radiations. Instead, targeted, immune, and gene and cell therapies combat tumor cells by precisely altering oncogenic cellular or molecular pathways. Consequently, the traditional monotonic dose-response relationship is no longer valid for many novel cancer drugs. Instead, efficacy often plateaus or even decreases after dose rises above certain level. To this end, US FDA launched Project Optimus [FDA, 2023a, Shah et al., 2021, Blumenthal et al., 2021] aiming to transform the early-phase oncology drug development. The main objective of the project is to identify an optimal dose, potentially lower than the maximum tolerated dose (MTD) but with at least comparable anti-tumor effect. The new dose optimization paradigm involves a randomized trial component com-

paring two or more doses following an initial dose escalation. While randomized trial design and SSE are routinely conducted in drug clinical trials, they are new in dose-optimization trials, usually coupled with limited resources. Traditional sample size estimation (SSE) is based on Frequentist Type I/II error rates, which is rarely the main objective in early-phase of drug development. Instead, investigators are more concerned about the accuracy in the trial decision making, such as selecting the right dose to start confirmatory studies which may be very costly. Current practice for SSE in dose optimization trials is often based on ad-hoc choices, such as using a small sample size of 20 patients per dose. This leads to a question of how much is expected to learn from the 20 patients, or whatever the number may be. Moreover, clinical trials that use human subjects for clinical research should always provide a reasonable justification for the number of subjects to be enrolled. We attempt to fill the gap by considering a Bayesian approach for sample size estimation.

4.1.2 Review of SSE Methods

In the literature, most SSE methods [Adcock, 1997, Wittes, 2002, Desu, 2012] are based on Frequentist hypothesis testing. A comprehensive review can be found in Wang and Ji [2020]. Standard SSE methods determine the sample size by controlling the type I error rate α to achieve a desirable power $(1-\beta)$, assuming the true population parameters are known. For example, a sample size statement for a two-arm RCT based on binary outcome is as follows:

> **Statement 1:** At type I error rate of α, with a clinically minimum effect size θ^*, S subjects are needed to achieve $(1-\beta)$ power when the response rates for the treatment and control arms are θ_1 and θ_0.

In addition, sample size may be determined based on hybrid Frequentist-Bayesian inference as in Ciarleglio and Arendt [2017], Berry et al. [2010]. These methods use Bayesian models but calibrate sample size based on type I error rate and power via simulation. In contrast, some methods use Bayesian properties for sample size estimation such as the length and

coverage of posterior credible intervals [Pham-Gia et al., 1993, Joseph and Bélisle, 1997, M'Lan et al., 2008]. Another type of Bayesian SSE methods determines sample size through Bayes factor with sampling and fitting priors [Lin et al., 2022], or loss functions like $k \cdot$ False Positive Rate + False Negative Rate [Müller et al., 2004a].

Several works in the literature attempt to bridge the SSE philosophies between Frequentist and Bayesian methods. Notable contributions include Kunzmann et al. [2021], Inoue et al. [2005], and Lee and Zelen [2000]. In particular, Lee and Zelen [2000] propose to link posterior estimates with Type I/II error rates, and estimate sample size by providing posterior probabilities rather than the Frequentist error rates. However, the sample size calculation still follows standard SSE approach. Notably the authors advocate defining "statistical significance" as the trial outcome having a large posterior probability of being correct.

4.1.3 Main idea

To this end, we propose a new Bayesian estimator of sample size (BESS) based on a general hierarchical modeling framework and posterior inference. Let $\boldsymbol{y}$ denote the observed data and $\boldsymbol{\theta}$ the unknown parameters as generic notation. We assume a Bayesian hierarchical model is to be used for data analysis. Our main idea is originated from noticing a simple trend between the sample size and observed effect size from data. Suppose in a two-arm RCT with sample size n per arm and a binary outcome, we are interested in testing if the effect size (difference in response rates), defined as $\theta = (\theta_1 - \theta_0)$ between the treatment (1) and control (0) is greater than a clinically minimum effect size θ^*. This can be formulated as testing the null hypothesis $H_0 : \theta \leq \theta^*$ versus the alternative $H_1 : \theta > \theta^*$. Suppose the observed response rates are $\bar{y}_1$ and $\bar{y}_0$ for the treatment and control arms, respectively. A Frequentist z-test is used to test whether θ is greater than the minimum effect size θ^*, given by

$$z^* = \frac{(\bar{y}_1 - \bar{y}_0) - \theta^*}{\sqrt{[\bar{y}_1(1-\bar{y}_1) + \bar{y}_0(1-\bar{y}_0)]/n.}}$$

If we assume, for the sake of argument, that $\bar{y}_1 = 0.3$, we find in the table below the relationship between the sample size n and $(\bar{y}_1 - \bar{y}_0)$, for reaching the same z^* value of 1.64, i.e., a one-sided p-value of 0.05. We call $(\bar{y}_1 - \bar{y}_0)$ the "Evidence." In order to reach the same Frequentist statistical significance, sample size increases if Evidence decreases in order to achieve $z^* = 1.64$ (p-value $= 0.05$).

Sample size (n)	10	20	30	50	100	500	1000
Evidence ($\bar{y}_1 - \bar{y}_0$)	0.29	0.24	0.21	0.18	0.15	0.10	0.08

Instead of using a Frequentist inference like the z-test, we consider Bayesian hypothesis testing based on posterior probability of the alternative hypothesis, which, in the setting of clinical trials, refers to the treatment being more effective than the control. Decision of accepting the alternative hypothesis is by thresholding the posterior probability of H_1 at a relatively large value $c \in (0, 1)$. Assume a Bayesian hierarchical model is given by $f(\boldsymbol{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid H)\pi(H)$, where $H = H_0$ or H_1 is a binary indicator of the null and alternative hypotheses. The proposed BESS aims to find a balance between 1) "Sample size" S of the clinical trial, 2) "Evidence" defined as the observed treatment effect $e = (\bar{y}_1 - \bar{y}_0)$, and 3) "Confidence" defined as the posterior probability of the alternative hypothesis.

Consider the previous example of testing if the treatment effect $\theta = (\theta_1 - \theta_0)$ is greater than θ^*, a minimum effect size. BESS provides a sample size statement as follows:

> **Statement 2:** Assuming the evidence is at least e, S subjects are needed to declare with confidence c that the treatment effect is at least θ^*.

Note that "confidence" here refers to $\Pr(H_1|\boldsymbol{y})$, the posterior probability of the alternative hypothesis, i.e., the treatment effect is at least θ^*. Comparing **Statement 1** and **Statement 2**, one can see that the two statements are based on different statistical properties, Frequentist type I/II error rates for **Statement 1** and Bayesian posterior probabilities for

Statement 2. We will show that the BESS and associated **Statement 2** is easier to interpret in practice since it directly addresses the uncertainty in the decision to be made for the trial at hand, measured by posterior probabilities.

The remainder of the article is organized as follows: Section 4.2 presents the proposed probability model. Section 4.3 describes the BESS method. Section 4.4 illustrates some of the proprieties of BESS, including the relationships among sample size, evidence, and confidence, as well as the coherence between BESS and Bayesian inference. Section 4.5 reports the operating characteristics of BESS with comparison to the standard SSE method. Section 4.6 illustrates the applications of BESS to a hypothetical dose optimization trial. Finally, we conclude the article in Section 4.7.

4.2 Probability Model

We consider BESS for both one-arm and two-arm trials. Denote y_{ij} the outcome of patient i in arm j, where $i = 1, \ldots, n$, index the patients, and $j = 0$ and 1 index the control and treatment arm, respectively. When needed, we drop index j and use y_i for one-arm trials. Let θ_1 denote the true response parameter for the treatment arm, and let θ_0 be the true response parameter for the control arm in a two-arm trial or the reference response parameter in a one-arm trial. Let $\theta = (\theta_1 - \theta_0)$ be the true treatment effect. Consider hypotheses

$$H_0 : \theta \leq \theta^* \text{ vs. } H_1 : \theta > \theta^*, \tag{4.1}$$

where θ^* is a minimum size for treatment effect deemed clinically relevant. Let H be the binary random variable taking H_0 or H_1 with probability $(1 - q)$ and q, respectively.

We propose a Bayesian hierarchical model for testing the hypotheses (4.1). Let

$$\begin{aligned} y_{ij}|\theta_j &\sim f(\theta_j), \\ (\theta_0,\theta_1)|\tilde{\boldsymbol{\theta}}, H = H_j &\sim \pi(\theta_0,\theta_1|\tilde{\boldsymbol{\theta}})I(\theta \in H_j), \\ \Pr(H = H_1) &= q, \end{aligned} \tag{4.2}$$

where $f(.)$ represents the likelihood function, $\pi(\theta_0,\theta_1|\tilde{\boldsymbol{\theta}})$ is a joint prior distribution for (θ_0,θ_1), $\tilde{\boldsymbol{\theta}}$ are hyper-parameters, and $I(x \in A)$ is the indicator function which equals 1 if $x \in A$ and 0 otherwise. For simplicity, we consider $\pi(\theta_0,\theta_1|\tilde{\boldsymbol{\theta}}) = \pi_0(\theta_0|\tilde{\boldsymbol{\theta}})\pi_1(\theta_0|\tilde{\boldsymbol{\theta}})$, where $\pi_0 = \pi_1$. In this work, we consider three specific types of outcome y_{ij}: binary, continuous, and count. A summary of the parameter θ_j, likelihood function $f(.)$, and prior distribution π_j for θ_j is shown in Table 4.1. In all three cases, the observed sample mean for each arm j, denoted as $\bar{y}_j$, is the sufficient statistics for θ_j. In Table 4.1 we consider conjugate prior for simplicity and computational speed. In general, different priors can be considered. While it is not the focus of this work to discuss choice of priors, we note the flexibility of BESS to incorporate various priors in real-life applications. For example, when little prior information is known vague priors like Jeffrey's prior may be considered; in contrast, it is also possible to use informative priors when prior information is available. For instance, assume a previous trial with binary outcome has completed with a sample size of n_0 patients of whom their outcome data are available, denoted as $\boldsymbol{y}^0 = \{y_i^0; i = 1,\ldots,n_0\}$. Then, one could consider an informative prior $\pi_1(\theta_1) = \text{Beta}(a,b)$ and set $a = a^* + \sum_{i=1}^{n_0} y_i^0,\ b = b^* + n_0 - \sum_{i=1}^{n_0} y_i^0$, where a^* and b^* are small (e.g., $a^* = b^* = 0.5$).

Table 4.1: Summary of parameter, likelihood function, and prior distribution for different outcome types.

Outcome type	Parameter θ_j	Likelihood $f(\cdot)$	Prior distribution $\pi_j(\tilde{\boldsymbol{\theta}})$
Binary	Response rate	Bern(θ_j)	Beta(a, b)
Continuous	Mean response	$N(\theta_j, \sigma^2)$, σ known	$N(a, b)$
Count-data	Event rate	Poi(θ_j)	Gamma(a, b)

4.3 Confidence, Evidence, and Sample Size

4.3.1 Confidence

We introduce the three pillar of BESS, Confidence, Evidence, and Sample Size. We start with "C", the confidence, which refers to the confidence of posterior inference expressed mathematically as the posterior probability of the alternative hypothesis $\Pr(H = H_1|\boldsymbol{y}_n)$, where $\boldsymbol{y}_n$ denotes the data with sample size n. The optimal decision rule under a variety of loss functions [Müller et al., 2004a] is to reject the null H_0 and accept H_1 if $\Pr(H = H_1|\boldsymbol{y}_n) \geq c$ for cutoff $c \in (0, 1)$. The value of c measures the least confidence of the decision to reject H_0 and accept H_1. The higher c is, the more confident is the decision. To see this, one simply observes that $(1 - c)$ is the upper bound of posterior probability of a wrong rejection since when H_1 is rejected, $\Pr(H = H_0|\boldsymbol{y}_n) < c$.

According to model (4.2), it is straightforward to show that

$$\Pr(H = H_1|\boldsymbol{y}_n) = \frac{q \cdot \int_{\boldsymbol{\theta} \in H_1} f(\boldsymbol{\theta}|\boldsymbol{y}_n) d\boldsymbol{\theta}}{1 - q + [(2q - 1) \cdot \int_{\boldsymbol{\theta} \in H_1} f(\boldsymbol{\theta}|\boldsymbol{y}_n) d\boldsymbol{\theta}]}. \tag{4.3}$$

If we assume *a priori*, both hypotheses are equally likely, i.e., $q = 0.5$, then equation (4.3)

may be further reduced to

$$\Pr(H = H_1|\boldsymbol{y}_n) = \int_{\boldsymbol{\theta}\in H_1} f(\boldsymbol{\theta}|\boldsymbol{y}_n)d\boldsymbol{\theta}. \tag{4.4}$$

If conjugate priors in Table 4.1 are used, $f(\theta_j|\boldsymbol{y}_{jn})$ have closed-form solutions. Otherwise, numerical evaluation of (4.3) or (4.4) is needed.

4.3.2 Evidence

Evidence is the main metric that differentiates BESS from a standard SSE. In short, we define evidence as a function of data, $e(\boldsymbol{y}_n)$, that reflects the strength of treatment effect. In the settings listed in Table 4.1 we consider evidence defined as

$$e = \bar{y} - \theta_0 \text{ for one-arm trials, and } e = \bar{y}_1 - \bar{y}_0 \text{ for two-arm trials.} \tag{4.5}$$

In simple words, evidence e is the observed effect size from the trial data before they are observed. This means that in order to apply BESS, investigator needs to pre-specify (and calibrate) the potential observed effect size before the trial is conducted. This is analogous to the requirement of specifying the true parameter values in standard SSE, except BESS assumes what might be observed rather than what might be true. We consider evidence e as a function of the sufficient statistic in Table 4.1. For the three outcomes in one-arm trials, $(e + \theta_0)$ is exactly the sufficient statistic. For two-arm trials, e is the difference between the sufficient statistics $(\bar{y}_0, \bar{y}_1)$ of the two arms. Being a function of sufficient statistics allows for a search algorithm to find an appropriate sample size under BESS, which will be clear next.

4.3.3 Sample Size of BESS

We first briefly review the standard SSE based on Frequentist inference. Consider a z-test for a two-arm trial with binary outcome, the standard sample size approach assumes true values of θ_1 and θ_0, and solves for n based on desirable Type I/II error rates α/β given by

$$n = \frac{(z_\alpha + z_\beta)^2}{(\theta - \theta^*)^2}[\theta_1(1-\theta_1) + \theta_0(1-\theta_0)]. \tag{4.6}$$

In BESS, we find the sample size through a similar argument but using posterior inference instead. Investigators specify the confidence c, so that

$$\Pr(H = H_1|\boldsymbol{y}_n) \geq c, \tag{4.7}$$

where $\Pr(H = H_1|\boldsymbol{y}_n)$ is computed by equation (4.3). In Müller et al. [2004a] decision rule (4.7) is shown to be optimal for a variety of common loss functions, such as the posterior expected loss of $k \cdot \text{FPR} + \text{FNR}$, where FPR and FNR are false positive and negative rates, respectively. Next, we provide three algorithms for sample size calculation based on BESS and models in Table 4.1.

One-arm trial For one-arm trials, with the settings of likelihood and prior in Table 4.1, we can show that $\Pr(H = H_1|\boldsymbol{y}_n) = \Pr(H = H_1|e, n)$. See Appendix A.3.1 for detail. Therefore, by fixing e and c, we can find the smallest sample size n that satisfies (4.7). This leads to the proposed BESS Algorithm 1 and the corresponding sample size statement **BESS 1.**

BESS 1: Assuming the evidence is e, n subjects are needed to declare with confidence c that the treatment effect is larger than θ^*.

Two-arm trial; Continuous data, known variance Second, for two-arm trials and continuous outcome with normal likelihood and known variance, we still have $\Pr(H = H_1|\boldsymbol{y}_n) =$

$\Pr(H = H_1|e, n)$. See Appendix A.3.2 for detail. Therefore, BESS Algorithm 1 applies. And we have the sample size statement BESS 2.1.

BESS 2.1: Assuming the evidence is e, n subjects per arm are needed to declare with confidence c that the treatment effect is larger than θ^*, assuming a known variance of σ^2 in each arm.

BESS Algorithm 1 One-Arm Trials; Two-Arm Trials with Continuous Data and Known Variance

Input: The hierarchical models in Table 4.1.
Input: Clinically meaningful effect size θ^*, evidence e, confidence c, prior probability q, reference response rate θ_0 (for one-arm trials).
Set $n_{\min}$ and $n_{\max}$ ($n_{\min} < n_{\max}$) the smallest and largest candidate sample sizes.
Set $n = n_{\min}$.
while $n \leq n_{\max}$ **do**
  Compute equation (4.3).
  if condition (4.7) is true **then**
    Stop and return the sample size n.
  else
    $n = n + 1$
  end if
end while
if $n > n_{\max}$ **then**
  Return sample size is larger than $n_{\max}$.
end if

Two-arm trial; Binary and Count data For binary and count data in two-arm trials, $\Pr(H = H_1|\boldsymbol{y}_n) = \Pr(H = H_1|\bar{y}_1, \bar{y}_0, n)$. See Appendix A.3.3 for detail. Since $e = \bar{y}_1 - \bar{y}_0$, fixing evidence e does not uniquely define $\Pr(H = H_1|\boldsymbol{y}_n)$. Hence, we propose to specify evidence e and find all pairs of $\bar{y}_1$ and $\bar{y}_0$ that satisfies $\bar{y}_1 - \bar{y}_0 = e$. We then find the minimum posterior probability of H_1 among these pairs of $\bar{y}_1$ and $\bar{y}_0$, i.e.,

$$\Pr(H = H_1|e, n) = \min_{\bar{y}_1, \bar{y}_0} \{\Pr(H = H_1|(\bar{y}_1, \bar{y}_0, n)); \forall \bar{y}_1 - \bar{y}_0 = e\}. \tag{4.8}$$

We propose BESS Algorithm 2 and the corresponding sample size statement BESS 2.2.

BESS 2.2 Assuming the evidence is e, at least n subjects per arm are needed to declare with confidence c that the treatment effect is larger than θ^*.

BESS Algorithm 2 Two-Arm Trials; Binary or Count Data

Input: The hierarchical models in Table 4.1.
Input: Clinically meaningful effect size θ^*, evidence e, confidence c, prior probability q.
Set $n_{\min}$ and $n_{\max}$ ($n_{\min} < n_{\max}$) the smallest and largest candidate sample sizes.
Set $n = n_{\min}$.
while $n \leq n_{\max}$ **do**
    Find all pairs of $(\bar{y}_1, \bar{y}_0)$ where $\bar{y}_1 - \bar{y}_0 = e$.
    for each pair of $(\bar{y}_1, \bar{y}_0)$ **do**
        Compute equation (4.3).
    end for
    Compute equation (4.8).
    if $\Pr(H = H_1|e, n) \geq c$ **then**
        Stop and return the sample size n.
    else
        n = n+1.
    end if
end while
if $n > n_{\max}$ **then**
    Return sample size is larger than $n_{\max}$.
end if

Alternatively, one may wish to specify the values of $\bar{y}_1$ and $\bar{y}_0$ directly instead of their difference e. Since $(\bar{y}_1, \bar{y}_0)$ is sufficient, the sample size search algorithm becomes easier. To this end, we propose a simple variation BESS Algorithm 2' in Appendix A.3.4 and statement BESS 2.2' assuming $(\bar{y}_1, \bar{y}_0)$ are given.

BESS 2.2' Assuming the response parameters in treatment and control arms are $\bar{y}_1$ and $\bar{y}_0$, respectively, n subjects per arm are needed to declare with confidence c that the treatment effect is larger than θ^*.

Discrete Data Lastly, since binary and count data are integer-valued, not all specified evidence value e can be achieved for a given sample size in the proposed BESS Algorithms

1 and 2. For example, when the outcome is binary and sample size $n = 10$, it is impossible to observe evidence $e = 0.15$ since this requires having $n \cdot e = 1.5$ more responders in the treatment arm than the control. The number of responders cannot be a fraction. To this end, we propose to round down the specified e to the nearest possible value for a given n by $e' = \frac{1}{n}\lfloor n \cdot e \rfloor$, where $\lfloor . \rfloor$ is the floor function. This gives a conservative estimate of the sample size since even with smaller evidence, the sample size would still ensure the needed confidence.

4.4 Properties of BESS

4.4.1 Correlation between sample size, evidence, and confidence

We explore the relationship of the three pillars of BESS, sample size, evidence, or confidence. Specifically, we fix one and report the correlation of the remaining two using a two-arm trial with binary outcome. Similar results can be achieved for other types of data or one-arm trials.

Positive correlation between confidence and evidence, Figure 4.1(a) We compute confidence using equation (4.3) for various values of evidence e, while keeping the sample size n constant. Figure 4.1(a) presents a line plot of the posterior probability of H_1 (the confidence) against various values of evidence, when sample size is set at $n = 10, 20$, or 30, assuming $\theta^* = 0.05$. This plot demonstrates that confidence increases monotonically from 0 to 1 as evidence shifts from -1 to 1. When evidence is negative, the data does not support the alternative hypothesis and therefore the confidence (that the alternative is true) drops to zero. When evidence is positive, the confidence improves. Interestingly, the order of confidence across different sample sizes flips when evidence is near 0, giving smaller sample size more confidence. This is expected since when evidence is small, a larger sample size should imply less confidence of the alternative hypothesis.

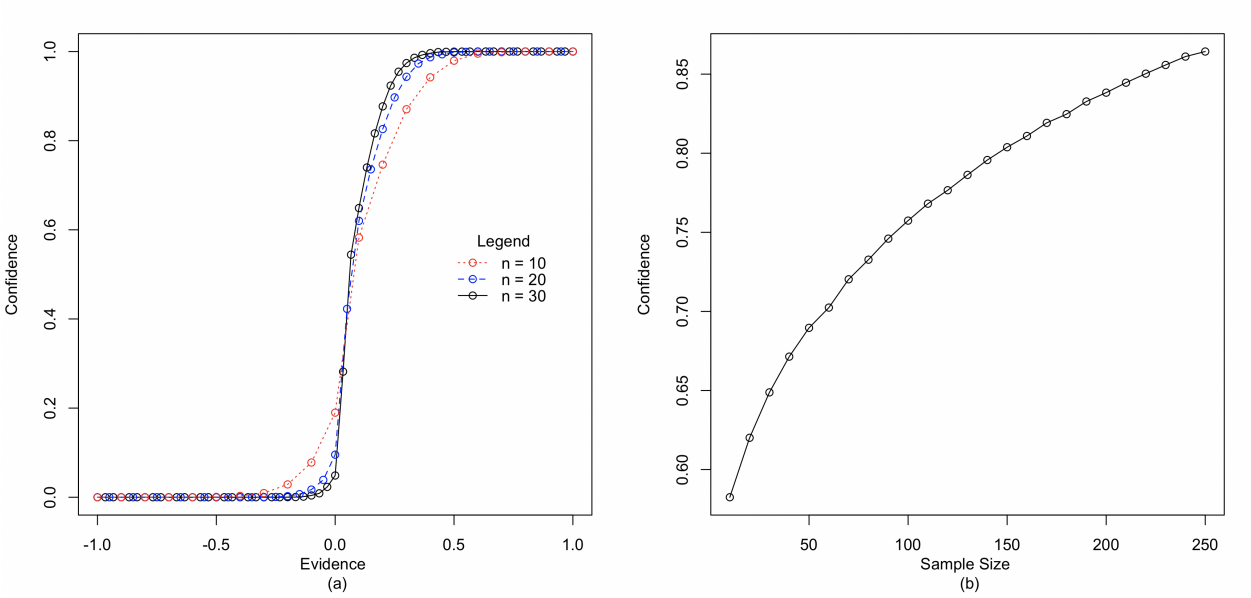

Figure 4.1: The line plots of (left) confidence vs. evidence when sample size is fixed to be $n = 10, 20,$ and 30, and (right) confidence vs. sample size when $e = 0.1$. The result assumes binary outcome for two-arm trial, with $\theta^* = 0.05$.

Positive correlation between sample size and confidence, Figure 4.1(b) We again compute confidence using equation (4.3), varying the sample size n while keeping the evidence e constant. Figure 4.1(b) demonstrates the positive relationship between sample size and confidence, when $e = 0.1$ and $\theta^* = 0.05$. This is expected when evidence is supportive of the alternative hypothesis since the larger the sample size , the larger the posterior probability of H_1.

Negative correlation between evidence and sample size We next explore the correlation between evidence and sample size, maintaining a fixed confidence level at $\Pr(H = H_1|e, n) = 0.6$. For this demonstration, we calculate confidence using (4.3) across a range of evidence and sample sizes. For example, with $n = 10$ patients per arm, the potential range of evidence e - reflecting the difference in percent responders between the treatment and control arms - spans from -1 to 1, with increments of 0.1.

Given the discrete nature of the outcomes, it's not always possible to find an evidence

level for each sample size n that exactly matches $\Pr(H = H_1|e, n) = 0.6$. Thus, we document the smallest e for each n where $\Pr(H = H_1|e, n) \geq 0.6$. In this case, for $\theta^* = 0.05$, we report $n(e)$, the different sample sizes n with their corresponding evidence e: 50(0.080), 100(0.070), 150(0.067), 200(0.065), and 1000(0.057). These values clearly show that as the sample size increases, the evidence needed to maintain the confidence at 0.6 decreases, approaching $\theta^* = 0.05$. Similar results can be obtained for continuous and count-data outcomes in two-arm trial, also the three outcomes in one-arm trial.

4.4.2 Coherence between BESS and Bayesian Inference

In the proposed BESS approach, the evidence $e(\boldsymbol{y}_n)$ is a function of the trial data $\boldsymbol{y}_n$. In addition, the confidence, defined as the posterior probability $\Pr(H_1|\boldsymbol{y}_n)$, is also a function of $\boldsymbol{y}_n$. Let's take a look at the BESS sample size statement again. It can be generalized as

> For assumed evidence e, a sample size of n will provide confidence c that the alternative hypothesis is true.

After the trial is conducted and data is observed, a Bayesian analysis of the observed data using posterior probability will be **coherent** with the BESS statement. To explain this, we first denote $\boldsymbol{y}_n^\star$ the observed data after the trial is completed, $e^\star$ the observed evidence, and $c^\star = \Pr(H_1|\boldsymbol{y}_n^\star)$ the posterior probability of H_1 conditional on the observed data $\boldsymbol{y}_n^\star$. Then the coherence between BESS and Bayesian inference means that if $e^\star > e$, $c^\star > c$.

This type of coherence is important for BESS to be adopted in practice since investigators of clinical trials are usually not statisticians, and the coherence property of BESS allows them to connect the design (i.e., sample size statement) of the trial with the statistical analysis of the observed data once the trial is carried out.

To verify the coherence numerically, we perform a simple simulation. In Appendix Table A.3.1 we set up the true parameters for data simulation for each type of clinical trials. For example, for two-arm binary data, we assume the minimum treatment effect $\theta^* = 0.05$,

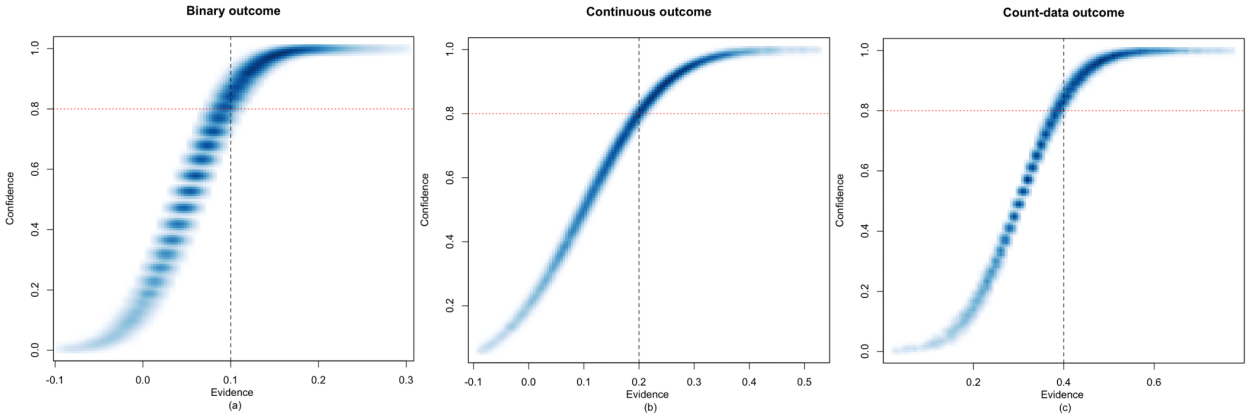

Figure 4.2: Plots of observed confidence $c^\star$ vs. observed evidence $e^\star$ for binary, continuous, and count-data outcomes with two-arm trial. The black vertical dashed line shows the location of $e^\star = e$, and the red horizontal dotted line shows the location of $c^\star = c$.

evidence $e = 0.1$, confidence $c = 0.8$. Using a prior distribution $\theta_j \sim \text{Beta}(0.5, 0.5)$ BESS leads to a sample size of 150. We then repeatedly generate trial data assuming different true values of θ_1 and θ_0 in Table A.1, and report the inference results of $e^\star$ and $c^\star$. Figure 4.2 summarizes the simulation results. In all three subplots (a)-(c), when $e^\star$ is larger than e (the black vertical line), $c^\star$ is larger than c, the red horizontal line.

4.5 Comparison with Standard SSE

4.5.1 Simulation Setup

Through simulation, we compare BESS and standard SSE for a two-arm trial with binary outcomes. For fair comparison, we match the Type I/II error rates for both methods. The matching is realized by three steps. Step 1: we obtain sample size estimated via the proposed BESS approach for a trial. Step 2: we repeatedly simulate trial data under the null and alternative hypotheses, perform the Bayesian inference using the same model as in BESS, and record the Type I and Type II error rates from the simulated trials. Step 3: using the two error rates, we apply the standard SSE to arrive at a Frequentist sample size,

and compare it with the BESS estimate. See Appendix Figure A.3.1 for an illustration.

Step 1: BESS sample size We consider a two-arm trial with binary outcome and let θ_1 and θ_0 be the response rates, and θ^* the clinical minimum effect size. The trial aims to test $H_0 : \theta_1 - \theta_0 \leq \theta^*$ vs. $H_1 : \theta_1 - \theta_0 > \theta^*$. For BESS, we assume a binomial likelihood, an improper prior $(\theta_1, \theta_0)|H_j \sim \text{Beta}(0,0) \cdot \text{Beta}(0,0) I((\theta_1, \theta_0) \in H_j)$, and $\Pr(H = H_1) \equiv q = 0.5$. We apply BESS Algorithm 2 to obtain a sample size for a pair of desirable evidence e and confidence cutoff c. We try several different pairs of e and c as well, listed in Table 4.2.

Step 2: Estimate Type I/II error rates Next, for each BESS sample size in Step 1, we repeatedly simulate trials under the null H_0 and the alternative H_1 to numerically compute the Type I/II error rates. Specifically, under the null we let $\theta_1 = 0.3$ and $\theta_0 = 0.25$, and under the alternative, $\theta_1 = 0.4$ and $\theta_0 = 0.25$. We set $\theta^* = 0.05$.

Based on a set of θ_1 and θ_0 values, we generate the binary responses of 150 patients each in the treatment arm and control arm. Denote the simulated outcomes $\boldsymbol{y}_n = \{\boldsymbol{y}_{n1} \text{ and } \boldsymbol{y}_{n0}\}$, $\boldsymbol{y}_{nj} = \{y_{ij}; i = 1, \ldots, n\}$, $j = 0, 1$. We generate $y_{ij} \sim \text{Bern}(\theta_j)$ where $\text{Bern}(\theta)$ is a Bernoulli distribution with mean θ. The null is rejected if $\Pr(H = H_1|\boldsymbol{y}_n) > 0.8$. Here, 0.8 is arbitrarily selected without specific intention. A larger or smaller cutoff than 0.8 arbitrarily affects the computed Type I/II error rates, which does not affect the objective of our simulation. We simulate 10,000 trials, each under the null and alternative. A rejection of the null for a trial simulated under the null is recorded as an incidence of Type I error and a non-rejection of the null for a trial under the alternative is recorded as an incidence of Type II error. The Type I/II error rates (α/β) are then computed as the frequencies of the corresponding incidences over the 10,000 trials.

Step 3: Estimate Frequentist sample size and compare with BESS Based on the computed Type I/II error rates (α/β) from Step 2, denoted as α and β, we estimate a

sample size using a standard SSE approach. For example, we consider a superiority z-test for comparing θ_1 and θ_0, and a sample size can be estiamted via (4.6). In the estimation, we assume the true values for θ_1 and θ_0, which gives the standard approach an "oracle" performance. In other words, the estimated sample size is guaranteed to achieve the target Type I/II error rates α and β, since the assumed θ_1 and θ_0 in the sample size estimation match the true values. Even though this is typically not achievable in reality, we decide to compare the oracle Frequentist sample size with the BESS sample size nevertheless.

4.5.2 Simulation Result

Table 4.2 presents the simulation results. They are both surprising and reassuring that BESS and the standard SSE produce similar sample sizes across a variety of settings. It is surprising since the two approaches are based on different statistical metrics. For BESS, it's aiming to balance between the anticipated evidence in the observed data and the confidence expressed as posterior probability; for the standard approach, it's trading off among the Type I/II error rates and assumed true parameter values. The results are reassuring since despite using different metrics, when matching the Type I/II error rates, both approaches produce highly similar sample size estimates.

Several trends are worth noting in Table 4.2. First, increase confidence cutoff c leads to lower Type I error rate for fixed evidence e. This is because increase in cutoff c makes it harder to reject the null for BESS under (4.7), and hence fewer simulated trials under H_0 will be rejected, i.e., decrease in the Type I error rate. Second, note that the true effect size under H_1 equals $\theta \equiv \theta_1 - \theta_0 = 0.4 - 0.25 = 0.15$. We observe that 1) power increases when the confidence cutoff c increases and evidence e is less than the the true effect size, i.e., $e = 0.1$, 2) power stays the same if evidence e equals the true effect size, and 3) power decreases when confidence increases and e is greater than the true effect size. This complicated trend demonstrates an interaction between c and power conditional on

whether e is greater than, equal to, or less than the true effect size θ. To understand this, first recall in Section 4.4.2 we used notation $e^\star$ to denote the observed evidence after the trial is completed and data observed. When c increases, the BESS-estimated sample size also increases (see Section 4.4.1). Therefore, the observed $e^\star$ will be closer to θ due to large number theory. Consequently, more simulated trials under the alternative will see $e^\star$ close to θ. If the assumed evidence e is smaller than θ, more likely it will be smaller than $e^\star$ as well. This means that the posterior probability $\Pr(H_1|\boldsymbol{y}_n)$ will be higher (since there is stronger evidence supporting H_1), and hence more rejections, i.e., higher power. Therefore, when e is smaller than θ, a larger c leads to higher power. Same logic applies to the case when e is larger than θ, in which a larger c leads to lower power. Lastly, when $e = \theta$, increasing sample size (as a result of increasing c) makes $e^\star$ approach e, and therefore there is no obvious impact on the power.

Results in Table 4.2 implies that if one wishes to have high power and a low Type I error rate using BESS, one may want to specify an evidence that is smaller than the true effect size and a high confidence cutoff c. As the true effect size is typically unknown, one may either construct an informative prior of θ_1 and θ_0 for BESS if prior information is available, or conduct interim analysis and sample size re-estimation to better plan and conduct a trial. We will explore the latter option in the next section.

Finally, from a true Bayesian perspective, BESS is concerned about the trial at hand, rather than hypothetical trials generated from the null or alternative. Therefore, while Table 4.2 illustrates a connection between BESS and standard SSE, it does not imply that BESS needs to be calibrated based on Type I/II error rates in practice. On the contrary, BESS focuses on the probability of making a right decision given the observed data, which can be measured by the false positive rate (FPR) and false negative rate (FNR). In Table 4.2 the reported FPR and FNR for BESS and standard SSE are the same since 1) the prevalence of trials under the H_0 and H_1 is 50% and 2) the standard SSE is oracle since it assumes

the true θ_0 and θ_1 values in its computation. In Appendix Table A.3.2 we show that when the θ_0 and θ_1 are mis-specified in the standard SSE, the estimated sample sizes may be too large or too small, leading to over- or under-power, and deflated or inflated FPR/FNR's.

Table 4.2: Simulation results compare BESS with Standard SSE when the type I error rate and power are matched between both methods in a two-arm trial with binary outcome. The results show the estimated sample sizes, false positive rates (FPR), and false negative rates (FNR) of the two methods across various levels of evidence e and confidence c.

Evidence	Confidence	type I	power	BESS			Standard SSE		
e	c	error rate α	$1-\beta$	n	FPR	FNR	n	FPR	FNR
0.10	0.7	0.31	0.76	60	0.29	0.26	62	0.29	0.26
	0.8	0.20	0.84	150	0.19	0.17	145	0.19	0.17
	0.9	0.10	0.94	340	0.10	0.07	344	0.10	0.07
0.15	0.7	0.29	0.56	20	0.34	0.38	22	0.34	0.38
	0.8	0.21	0.56	40	0.27	0.36	40	0.27	0.36
	0.9	0.10	0.56	87	0.16	0.33	88	0.16	0.33
0.20	0.7	0.44	0.57	5	0.43	0.43	5	0.43	0.43
	0.8	0.24	0.46	15	0.34	0.41	16	0.34	0.41
	0.9	0.11	0.38	35	0.23	0.41	36	0.23	0.41

Sensitivity of Prior We demonstrate the sensitivity of incorporating prior information through simulation, assuming there exists prior data of n_0 patients per arm. The simulation details are presented in Appendix 4.5.1 , and the average sample size from the 1,000 simulated trials is 31.42 with a standard deviation of 28. Recall with a Beta$(0, 0)$ prior the BESS sample size was 40 in Table 4.2 for $e = 0.15$ and $c = 0.8$. The results show that BESS is able to properly borrow prior information for sample size estimation. In practice, one may use different informative priors, e.g, power prior [Ibrahim and Chen, 2000] or commensurate prior [Hobbs et al., 2011], for information borrowing. We leave these topics for future research.

4.6 Demonstration of BESS with Dose Optimization Trial

4.6.1 Fixed Sample Size

Lastly, we consider a randomized comparison of two selected doses in an oncology phase I trial as part of FDA's Project Optimus initiative for dose optimization in oncology drug development. Suppose two doses are compared via a $1:1$ randomized design. We apply BESS to estimate the sample size of the comparison. In dose optimization, the goal is to test if the lower dose is no worse than the higher dose in terms of efficacy, i.e., non-inferiority. Denoting θ_H and θ_L the response rates for the higher and lower doses, respectively, we want to test the following non-inferiority hypotheses

$$H_0 : \theta_H - \theta_L \geq \theta^* \text{ vs. } H_1 : \theta_H - \theta_L < \theta^*, \tag{4.9}$$

where $\theta^* \in (0, 1)$ is the non-inferiority margin. To fit the setting in (4.1), we rewrite the hypotheses as $H_0 : \theta_L - \theta_H \leq -\theta^*$ vs. $H_1 : \theta_L - \theta_H > -\theta^*$. We then apply BESS algorithm 2 to estimate the sample size for the dose-optimization trial. Assuming $\theta^* = 0.05$, we consider two related objectives: 1) find sample size given evidence and confidence, and 2) find evidences and the corresponding confidences for a fixed sample size.

Objective 1 BESS provides the following sample size statement: Assuming evidence $e = 0$, a sample size of 57 patients per arm is needed to declare with 70% confidence that the response rate of the higher dose is no higher than the lower dose by 0.05.

Objective 2 Assume 20 patients per dose are randomized and denote $\bar{y}_L$ and $\bar{y}_H$ the observed response rates for the lower and higher doses, respectively. We compute the confidence, $\Pr(\theta_H - \theta_L < \theta^* | \bar{y}_L, \bar{y}_H, n)$, for various values of $(\bar{y}_L - \bar{y}_H)$ and $\theta^* = 0.05$, shown in Table 4.3. For example, if one observes the response rate of the lower dose is the same

Table 4.3: List of various evidence and confidence for $\theta^* = 0.05$ with $n = 20$ patients per arm.

	Noninferiority margin $\theta^* = 0.05$, Sample size $n = 20$									
Evidence $\bar{y}_L - \bar{y}_H$	≤ -0.20	-0.15	-0.10	-0.05	0.00	0.05	0.10	0.15	0.20	≥ 0.25
Confidence c	< 0.05	0.12	0.28	0.50	0.62	0.74	0.84	0.90	0.94	> 0.95

as the higher dose, i.e., $(\bar{y}_L - \bar{y}_H) = 0$, with 20 patients per dose one gets about 62.49% confidence to declare the lower dose response rate is within the non-inferiority margin 0.05 of the higher dose.

4.6.2 Sample Size for Adaptive Designs

Setup We further consider adaptive designs that allow interim analysis and early stopping for the randomized dose comparison in the previous section. We set the non-inferiority margin to $\theta^* = 0.07$ so that the standard SSE gives a sample size of 100 patients per dose. We consider four different designs based on either the standard SSE or the BESS.

1. BESS SSR The first design is BESS with sample size re-estimation (SSR). BESS SSR estimates a sample size n for the entire trial first using input of evidence e and confidence c. Then when $n/2$ patients are enrolled at each dose, an interim analysis is performed to allow trial stopping or SSR if the trial is not stopped. Denoting the interim patients outcome as $\boldsymbol{y}_{n/2}$, the trial is stopped early if $\Pr(H_1|\boldsymbol{y}_{n/2}) \geq c$ or $\Pr(H_1|\boldsymbol{y}_{n/2}) \leq c^*$, where c close to 1 and c^* close to 0 are probability thresholds for early stopping due to success or failure, respectively. If neither condition is met, the trial proceeds with an SSR as follows.

In the SSR, we again use BESS to re-estimate the sample size based on updated evidence from the interim data, defined as $e_{\text{int}} = \mathrm{E}[\theta_L|\boldsymbol{y}_{n/2}] - \mathrm{E}[\theta_H|\boldsymbol{y}_{n/2}]$, which corresponds to the difference of posterior means. We use the same cutoff c for the SSR. However, we use the posterior distribution $\pi(\theta_L, \theta_H|\boldsymbol{y}_{n/2})$ as the prior in the SSR using the BESS algorithm 2. Denote the additional sample size as n^* based on SSR. At the end of the trial after n^* more

patients are randomized to each dose, BESS SSR rejects the null and accepts the alternative if $\Pr(H_1|\boldsymbol{y}_{n/2+n^*}) \geq c$. Otherwise, it accepts the null and rejects the alternative.

2. BESS SSR Cap The design is the same as BESS SSR, except when $n^* > n/2$, we cap $n^* = n/2$. That is, we restrict the maximum sample size at n for the entire trial.

3. Standard SSE The standard SSE estimates a sample size n based on a z-test for (4.9), with a significance level α, and a desired power $(1-\beta)$. The null hypothesis is rejected if $1 - \Phi(z) \leq \alpha$. Formula (4.6) is used to compute n.

4. Standard SSE with interim This design follows the previous standard SSE to estimate n, but at $n/2$ it stops the trial if $\Pr(H_1|\boldsymbol{y}_{n/2}) \geq c$ or $\Pr(H_1|\boldsymbol{y}_{n/2}) \leq c^*$. This is the same Bayesian interim analysis in design 1, BESS SSR. If the trial is not stopped at $n/2$, the trial adds additional $n/2$ patients and use the z-test to make a final decision.

The four designs are compared through simulated trials. For each trial, we generate the alternative and null hypotheses indicators $H = 1$ or 0 with probabilities q or $(1-q)$, respectively. Then given H, we generate true model parameters θ_L and θ_H based on two scenarios. In scenario 1, θ_L and θ_H are random variables under null or alternative, where $\theta_H \sim \text{Unif}(\theta^* = 0.07, 0.6)$, and

$$\theta_L|\theta_H, H \sim \begin{cases} \text{Unif}(0, \theta_H - \theta^*) & H = 0 \\ \text{Unif}(\theta_H - \theta^*, 0.28)I(\theta_H - \theta^* \leq 0.28) + \delta_{\theta_H} I(\theta_H - \theta^* > 0.28) & H = 1 \end{cases}$$

This setting ensures $E(\theta_L) = E(\theta_H) = 0.335$ when $H = 1$. In scenario 2, we assume $\theta_H = 0.335$ and $\theta_L = 0.265$ under H_0 or $\theta_L = 0.335$ under H_1, i.e., they are fixed. The two scenarios reflect two different practical considerations. Scenario 1 aims to assess the performance of the four designs assuming they are applied to different programs and trials in which the true responses of the drugs and doses are different. Scenario 2 is a classical

Frequentist setting to assess the Type I/II error rates of a design assuming true response rates are fixed. Finally, given θ_L and θ_H, for each trial we simulate patients outcome data based on Binomial distributions and the corresponding designs.

We assume $q = 0.5$. For Designs 1 & 2, we let $e = 0.02$, $c = 0.7$ and $c^* = 0.3$. For designs 3 & 4, we set $\alpha = 0.3$ and $\beta = 0.3$, so that BESS and standard SSE produce the same sample size for the trial in the beginning, which is $n = 100$. Lastly, we assume $(\theta_L, \theta_H) \sim \text{Beta}(0.05, 0.05)\text{Beta}(0.05, 0.05)I(\theta \in H)$, where $\text{Beta}(0.05, 0.05)$ is chosen to reduce the prior effective sample size. A total of 2,000 simulated trials are generated, 1,000 each under H_0 or H_1, for each design in each scenario. Four metrics are used to compare the designs, Type I/II error rates, false positive rate (FPR) and false negative rate (FNR). See Appendix A.3.7 for detail.

Results We first present the operating characteristics of the four designs by evaluating their Type I and II error rates across each scenario. To facilitate the comparison of designs, motivated by Kim and Choi [2021] we consider a combined error rate defined as

$$\text{Combined Error Rate (CER)} = \text{Type I error rate} + k \cdot \text{Type II error rate},$$

where the weight $k \in [0, \infty)$ is a pre-determined factor that quantifies the relative weight between the Type I and Type II error rates. We let $k \in \{0.5, 1, 1.5\}$. A lower CER is more desirable. Top two rows in Figure 4.3 reports the results for the four designs. Across all designs, as the sample size increases, the rate at which CER declines, indicating potential diminishing return when sample size gets larger. In other words, the gain in reduction of error rates may lessen when sample size continues to increase. Considering the substantial costs associated with patient enrollment, these findings suggest finding a “sweet spot" of sample size that achieves a desirable tradeoff between cost and statistical properties. Comparing across designs, we find that except for the Standard SSE design, the performance of the other

designs are similar. Notably, the BESS SSR Cap design (design 2) seems to be in general the best across most cases. In Scenario 2 when $k = 0.5$, the standard SSE is the winning design. In contrast, it is the losing design in the same scenario but with $k = 1.5$. This seems to suggest the standard SSE is a better design if the Frequentist Type I error rate is of important consideration in design evaluation. However, in early-phase dose optimization trials, Type I error rate is not the primary concern. In fact, one may argue it is of the least concern. For example, one would be much more concerned if a "GO" decision that recommend a wrong dose to further clinical development, or a "No Go" decision that fails to recommend a promising dose. These are measured by the FPR and FNR in the bottom two rows of Figure 4.3.

Following Müller et al. [2004a], we employ a loss function that integrates FPR and FNR into a single metric given by

$$\text{Combined False Rate (CFR)} = k \cdot \text{FPR} + \text{FNR}.$$

Again, $k \in \{0.5, 1, 1.5\}$. The results, shown in Figure 4.3, again suggest the BESS SSR Cap design (design 2) is the overall most desirable method. Therefore, to summarize, we recommend using the BESS with a sample size re-estimation capping the max sample size for the dose optimization trial.

4.7 Discussion

In this work we propose BESS as a new and simple Bayesian sample size estimator. BESS leads to a straightforward interpretation of estimated sample size: if the observed data exhibits certain level of evidence e that supports the alternative hypothesis, with sample size n one can conclude the alternative with confidence c, measured by posterior probability of the alternative. The statement is coherent with a subsequent Bayesian analysis of the trial using

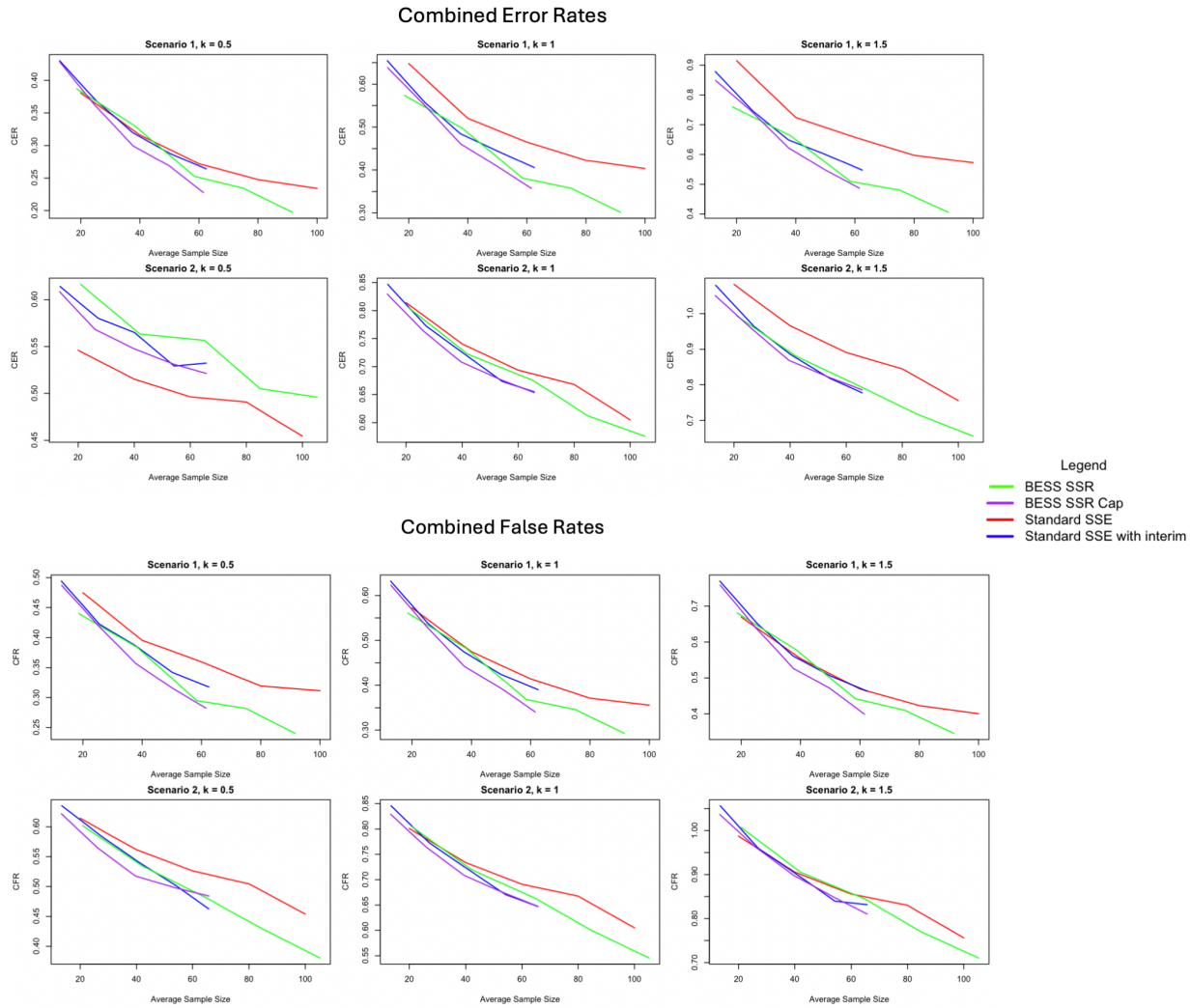

Figure 4.3: Combined Error Rates (CER) and Combined False Rates (CFR) across various sample sizes for the four designs under comparison. Different k values are used to illustrate the importance of type I error rate over the Type II error rate in CER or FPR over FNR in CFR.

n as the sample size. If the observed trial data exhibits evidence e, the posterior probability of alternative will be greater than c, corroborating with the sample size estimation. For dose optimization trials, the BESS SSR Cap design shows a superior performance under both Frequentist and Bayesian properties.

Simulation results show that for matched Type I/II error rates and using vague priors, BESS produces similar sample size estimates as the standard Frequentist sample size estimation, even though the two are based on different philosophies and metrics. In addition, the preliminary comparison of various designs based on BESS suggests a slight advantage of the Bayesian approach.

Many future directions may be considered to further develop BESS and corresponding adaptive designs, such as sample size re-estimations, downgrade historical data in prior construction, decision rules for futility stopping, etc.

APPENDIX A

APPENDIX OF THREE PAPERS

A.1 Appendix to "A Class of Dependent Random Distributions Based on Atom Skipping"

A.1.1 Features of BNP models

Table A.1.1 summarizes the features of some BNP models, along with the proposed PAM. A feature is checked based on the definition of the model, not the posterior inference.

BNP Models	Common Atoms / Common Weights	Common Atoms / Distinct Weights	Distinct Atoms / Distinct Weights	Plaid* Atoms / Distinct Weights
CAM	✓	✓		
HDP		✓		
LNP	✓			✓
NDP	✓		✓	
PAM		✓	✓	✓

∗ "Plaid" atoms means groups can share common atoms but can also possess unique atoms.

Table A.1.1: Features supported by various BNP models. A check-mark means the model supports such feature.

A.1.2 Proof of Proposition 1

We show the results for π'_{jk} as the result for π'_k is the same, with index j removed. Conditional on $\boldsymbol{\beta} = \{\beta_k; k \geq 1\}$ (which is equivalent as conditional on $\boldsymbol{\beta}' = \{\beta'_k; k \geq 1\}$ because β_k is constructed from β'_k deterministically), we have

$$E[\pi'_{jk}|\boldsymbol{\beta}, p_j] = \frac{p_j \beta_k}{1 - \sum_{l=1}^{k-1} \beta_l}. \tag{A.1.1}$$

Then

$$E[E[\pi_{jk}|\boldsymbol{\beta}, p_j]] = E\left[E\left[\pi'_{jk}\prod_{l=1}^{k-1}(1-\pi'_{jl})|\boldsymbol{\beta}, p_j\right]\right]$$

$$= E\left[\frac{p_j\beta_k}{1-\sum_{l=1}^{k-1}\beta_l}(1-p_j\beta_1)\prod_{l=2}^{k-1}\left(\frac{1-\sum_{w=1}^{l-1}\beta_w - p_j\beta_l}{1-\sum_{w=1}^{l-1}\beta_w}\right)\right]$$

$$= E\left[p_j\beta_k\prod_{l=1}^{k-1}\left(\frac{1-\sum_{w=1}^{l-1}\beta_w - p_j\beta_l}{1-\sum_{w=1}^{l}\beta_w}\right)\right]$$

$$= E\left[p_j\beta_k\prod_{l=1}^{k-1}\left(\frac{1-\sum_{w=1}^{l}\beta_w + \beta_l - p_j\beta_l}{1-\sum_{w=1}^{l}\beta_w}\right)\right]$$

$$= E\left[p_j\beta_k\prod_{l=1}^{k-1}\left\{\frac{\sum_{w=l+1}^{\infty}\beta_w + (1-p_j)\beta_l}{\sum_{w=l+1}^{\infty}\beta_w}\right\}\right]$$

$$= E\left[p_j\beta_k\prod_{l=1}^{k-1}\left\{1+\frac{(1-p_j)\beta_l}{\sum_{w=l+1}^{\infty}\beta_w}\right\}\right]$$

Expanding the term in the expectation, we have

$$p_j\beta_k\prod_{l=1}^{k-1}\left\{1+\frac{(1-p_j)\beta_l}{\sum_{w=l+1}^{\infty}\beta_w}\right\}$$

$$= p_j\beta'_k\prod_{l=1}^{k-1}(1-\beta'_l)\prod_{l=1}^{k-1}\left\{1+\frac{(1-p_j)\beta'_l\prod_{s=1}^{l-1}(1-\beta'_s)}{\sum_{w=l+1}^{\infty}\beta'_w\prod_{s=1}^{w-1}(1-\beta'_s)}\right\}$$

$$= p_j\beta'_k\prod_{l=1}^{k-1}(1-\beta'_l)\times$$

$$\prod_{l=1}^{k-1}\left\{1+\frac{(1-p_j)\beta'_l\prod_{s=1}^{l-1}(1-\beta'_s)}{\beta'_{l+1}\prod_{s=1}^{l}(1-\beta'_s)+\beta'_{l+2}\prod_{s=1}^{l+1}(1-\beta'_s)+\beta'_{l+3}\prod_{s=1}^{l+2}(1-\beta'_s)+\cdots}\right\}$$

$$= p_j\beta'_k\prod_{l=1}^{k-1}(1-\beta'_l)\times$$

$$\prod_{l=1}^{k-1}\left\{1+\frac{(1-p_j)\beta'_l}{\beta'_{l+1}(1-\beta'_l)+\beta'_{l+2}\prod_{s=l}^{l+1}(1-\beta'_s)+\beta'_{l+3}\prod_{s=l}^{l+2}(1-\beta'_s)+\cdots}\right\}$$

$$=p_j\beta'_k\prod_{l=1}^{k-1}(1-\beta'_l)\times$$

$$\prod_{l=1}^{k-1}\left\{1+\frac{(1-p_j)\beta'_l}{(1-\beta'_l)}\frac{1}{\beta'_{l+1}+\beta'_{l+2}(1-\beta'_{l+1})+\beta'_{l+3}\prod_{s=l+1}^{l+2}(1-\beta'_s)+\cdots}\right\}$$

$$=p_j\beta'_k\prod_{l=1}^{k-1}(1-\beta'_l)\prod_{l=1}^{k-1}\left\{1+\frac{(1-p_j)\beta'_l}{(1-\beta'_l)}\frac{1}{\sum_{w=l+1}^{\infty}\beta'_w\prod_{s=l+1}^{w-1}(1-\beta'_s)}\right\} \tag{A.1.2}$$

Denote $\Gamma=\sum_{w=l+1}^{\infty}\beta'_w\prod_{s=l+1}^{w-1}(1-\beta'_s)$ in (A.1.2). Then it follows

$$1-\Gamma=(1-\beta'_{l+1})(1-\beta'_{l+2})\cdots=\prod_{w=l+1}^{\infty}(1-\beta'_w)=0.$$

Therefore, $\Gamma=1$ and the expectation of (A.1.2) becomes

$$E[E[\pi_{jk}|\boldsymbol{\beta'},p_j]]=E\left[p_j\beta'_k\prod_{l=1}^{k-1}(1-\beta'_l)\prod_{l=1}^{k-1}\left\{\frac{1-\beta'_l+(1-p_j)\beta'_l}{(1-\beta'_l)}\right\}\right]$$

$$=E\left[p_j\beta'_k\prod_{l=1}^{k-1}(1-p_j\beta'_l)\right]=E[p_j]E[\beta'_k]\prod_{l=1}^{k-1}(1-E[p_j]E[\beta'_l]) \tag{A.1.3}$$

Since $\beta'_k\sim\text{Beta}(1,\gamma)$ and $p_j\sim\text{Beta}(a,b)$, we have

$$E[\pi_{jk}]=\frac{\bar{p}}{1+\gamma}\left(\frac{1+\gamma-\bar{p}}{1+\gamma}\right)^{k-1}=\frac{1}{1+\gamma'}\left(\frac{\gamma'}{1+\gamma'}\right)^{k-1}$$

where $\gamma'=\frac{1+\gamma-\bar{p}}{\bar{p}}$, $\bar{p}=\frac{a}{a+b}$. This proves the second and third claims in Proposition 1.

To show the first claim, we first show $E\left[\sum_{k\geq1}\pi_{jk}\right]=1$. Notice that

$$E\left[\sum_{k\geq1}\pi_{jk}\right]=\sum_{k\geq1}E[\pi_{jk}]=\sum_{k\geq1}\frac{\bar{p}}{1+\gamma}\left(\frac{1+\gamma-\bar{p}}{1+\gamma}\right)^{k-1}$$

$$= \sum_{k^* \geq 0} \frac{\bar{p}}{1+\gamma} \left(1 - \frac{\bar{p}}{1+\gamma}\right)^{k^*} = \frac{\bar{p}}{1+\gamma} \times \frac{1+\gamma}{\bar{p}} = 1.$$

Next, we show $0 < \sum_{k \geq 1} \pi_{jk} \leq 1$. It is trivial to see that $\sum_{k \geq 1} \pi_{jk} > 0$. We now show $\sum_{k \geq 1} \pi_{j,k} \leq 1$. Notice

$$1 - \sum_{k \geq 1} \pi_{jk} = 1 - \pi'_{j1} - \pi'_{j2}(1 - \pi'_{j1}) - \pi'_{j3}(1 - \pi'_{j1})(1 - \pi'_{j2}) - \cdots = \prod_{k=1}^{\infty} (1 - \pi'_{jk}) \geq 0$$

since $0 \leq \pi'_{jk} < 1$. Therefore, $\sum_{k \geq 1} \pi_{jk} \leq 1$. Thus, we have shown $0 < \sum_{k \geq 1} \pi_{jk} \leq 1$ and $E\left[\sum_{k \geq 1} \pi_{jk}\right] = 1$, and we conclude $\sum_{k \geq 1} \pi_{jk} = 1$ almost surely. This proves the first claim of Proposition 1. □

A.1.3 Proof of Theorem 1

For $G|G_0, p \sim ASP(p, \alpha_0, G_0)$, we derive the mean of G. Recall $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\boldsymbol{\phi}_k}$. Conditional on G_0 is equivalent as conditional on $\boldsymbol{\beta'} = \{\beta'_k; k \geq 1\}$ and $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_k; k \geq 1\}$. From equation (A.1.3) in subsection A.1.2, we have

$$\mathrm{E}[G(A)|G_0, p] = \mathrm{E}[G(A)|\boldsymbol{\beta'}, \boldsymbol{\Phi}, p] = \sum_{k=1}^{\infty} \mathrm{E}[\pi_k|\boldsymbol{\beta'}, p] \delta_{\boldsymbol{\phi}_k}(A)$$

$$= \sum_{k=1}^{\infty} p\beta'_k \prod_{l=1}^{k-1} (1 - p\beta'_l) \delta_{\boldsymbol{\phi}_k}(A) = G^*(A),$$

where $G^* = \sum_{k=1}^{\infty} \omega_k \delta_{\phi_k}$, $\omega_k = \omega'_k \prod_{l=1}^{k-1}(1 - \omega'_l)$, $\omega'_k = p \cdot \beta'_k$. As $G_0 \sim DP(\gamma, H)$, we have $\beta'_k \sim \mathrm{Beta}(1, \gamma)$, and $\boldsymbol{\phi}_k \sim H$. Plugging in the priors for β'_k and $\boldsymbol{\phi}_k$, we see that the stick-breaking construction of G^* is equal to that of FSBP in Section 3.3. □

A.1.4 Proof of Proposition 2

Let $\boldsymbol{\theta}_{i1}|G_1 \sim G_1$ and $\boldsymbol{\theta}_{i'2}|G_2 \sim G_2$, without loss of generality,

$$\Pr(\boldsymbol{\theta}_{i1} = \boldsymbol{\theta}_{i'2}) = \int \Pr(\boldsymbol{\theta}_{i1} = \boldsymbol{\theta}_{i'2}|G_1, G_2)p(G_1)p(G_2)dG_1dG_2 > \int 0p(G_1)p(G_2)dG_1dG_2 = 0$$

if and only if $\Pr(\boldsymbol{\theta}_{i1} = \boldsymbol{\theta}_{i'2}|G_1, G_2) > 0$. We next show $\Pr(\boldsymbol{\theta}_{i,1} = \boldsymbol{\theta}_{i'2}|G_1, G_2) > 0$. Denote the set $A^s = \{\boldsymbol{\phi}_k; \pi_{jk} \neq 0 \text{ and } \pi_{j'k} \neq 0\}$ and $A^j = \{\boldsymbol{\phi}_k; \pi_{jk} \neq 0\}$ for $j \neq j'$, $j = 1, 2$. Then

$$\Pr(\boldsymbol{\theta}_{i1} = \boldsymbol{\theta}_{i'2}|G_1, G_2) = \Pr(\boldsymbol{\theta}_{i1} = \boldsymbol{\theta}_{i'2}|\boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s)\Pr(\boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s|G_1, G_2) \tag{A.1.4}$$

The second term in (A.1.4) is

$$\Pr(\boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s|G_1, G_2)$$

$$= \Pr(\boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s|A^s \neq \emptyset, G_1, G_2)\Pr(A^s \neq \emptyset) +$$

$$\Pr(\boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s|A^s = \emptyset, G_1, G_2)\Pr(A^s = \emptyset)$$

$$= \Pr(\boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s|A^s \neq \emptyset, G_1, G_2)\Pr(A^s \neq \emptyset)$$

Then $\Pr(A^s \neq \emptyset) = 1 - \Pr(A^s = \emptyset) = 1 - \prod_{k=1}^{\infty}\{p_1(1-p_2) + p_2(1-p_1)\} = 1$. This is because at each atom k, G_1 selects the atom with probability p_1 and G_2 does not select the atom, with probability $(1-p_2)$, or vice versa. Denote $K^s = \{k; \boldsymbol{\phi}_k \in A^s\}$ and $K^j = \{k; \boldsymbol{\phi}_k \in A^j\}$. The term $\Pr(\boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s|A^s \neq \emptyset, G_1, G_2)$ is evaluated as

$$\Pr(\boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s|A^s \neq \emptyset, G_1, G_2) = \left[\sum_{k \in K^s} \pi_{1k}\right]\left[\sum_{k \in K^s} \pi_{2k}\right].$$

Since $\Pr(A^s \neq \emptyset) = 1$, $|K^s| \geq 1$, and since $\pi_{1k} > 0$ and $\pi_{2k} > 0$ for $k \in K^s$, for some arbitrary $k^* \in K^s$, we have

$$\Pr(\boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s | A^s \neq \emptyset, G_1, G_2) \geq \pi_{1k^*}\pi_{2k^*} > 0$$

Therefore,

$$\Pr(\boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s | G_1, G_2) = \Pr(\boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s | A^s \neq \emptyset, G_1, G_2) \times 1 > 0.$$

And the first term in (A.1.4) is

$$\Pr(\boldsymbol{\theta}_{i1} = \boldsymbol{\theta}_{i'2} | \boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s) = E[\Pr(\boldsymbol{\theta}_{i1} = \boldsymbol{\theta}_{i'2} | G_1, G_2, \boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s)]$$

$$= E\left[\left(\sum_{\boldsymbol{\phi}_k \in A^s} I(\boldsymbol{\theta}_{i1} = \boldsymbol{\theta}_{i'2} = \boldsymbol{\phi}_k) p(\boldsymbol{\phi}_k)\right) | G_1, G_2, \boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s\right]$$

$$= E\left[\left(\sum_{\boldsymbol{\phi}_k \in A^s} \pi_{1k}\pi_{2k} p(\boldsymbol{\phi}_k)\right) | G_1, G_2, \boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s\right]$$

$$\overset{(a)}{=} \sum_{k \in K^s} E[\pi_{1k}]\mathrm{E}[\pi_{2k}] = \sum_{k \in K^s} E\left\{\pi'_{1k} \prod_{l \in K^1, l<k} (1 - \pi'_{1l})\right\} E\left\{\pi'_{2k} \prod_{l \in K^2, l<k} (1 - \pi'_{2,l})\right\}$$

$$\geq \sum_{k \in K^s} \left[E\left\{\pi'_{1k} \prod_{l \in K^1, l<k} (1 - \pi'_{1l}) \prod_{l \in K^{1c}, l<k} (1 - {\pi'_{1l}}^*)\right\}\right.$$

$$\left. E\left\{\pi'_{2k} \prod_{l \in K^2, l<k} (1 - \pi'_{2l}) \prod_{l \in K^{2c}, l<k} (1 - {\pi'_{2l}}^*)\right\}\right]$$

$$\overset{(b)}{=} \sum_{k \in K^s} [E[\beta_k]^2] \overset{(c)}{=} \sum_{k \in K^s} \left[\frac{1}{1+\gamma}\left(\frac{\gamma}{1+\gamma}\right)^{k-1}\right]^2$$

where I(A) is the indicator function that equals to 1 if condition A is satisfied, ${\pi'_{jl}}^* \sim \text{Beta}\left(\alpha_0\beta_k, \alpha_0\left(1-\sum_{l=1}^{k}\beta_l\right)\right)$, and K^{j^c}s are the complement sets of K^j, for $j = 1, 2$. In addition, (a) is true because

$$p(\boldsymbol{\phi}_k|G_j) = \begin{cases} 1 & \text{if } \boldsymbol{\phi}_k \in G_j \\ 0 & \text{o.w.} \end{cases},$$

and (b) is true because the term $\pi'_{jk}\prod_{l\in K^j, l<k}(1-\pi'_{jl})\prod_{l\in K^{j^c}, l<k}(1-{\pi'_{jl}}^*) = {\pi'_{jk}}^*\prod_{l<k}(1-{\pi'_{jl}}^*)$ for $k \in K^s$ (i.e., equation (4)), with conditional expectation (conditional on $\boldsymbol{\beta}$) equals to β_k, and (c) is true because $\beta_k = \beta'_k\prod_{l<k}(1-\beta'_k)$, $\beta'_k \sim \text{Beta}(1, \gamma)$.

Again since $|K^s| \geq 1$, for some arbitrary $k^* \in K^s$, we have

$$\sum_{k\in K^s}\left[\frac{1}{1+\gamma}\left(\frac{\gamma}{1+\gamma}\right)^{k-1}\right]^2 \geq \left[\frac{1}{1+\gamma}\left(\frac{\gamma}{1+\gamma}\right)^{k^*-1}\right]^2 > 0.$$

Thus, we have

$$\Pr(\boldsymbol{\theta}_{i,1} = \boldsymbol{\theta}_{i'2}|\boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s) > 0.$$

Combine with $\Pr(\boldsymbol{\theta}_{i1} \in A^s, \boldsymbol{\theta}_{i'2} \in A^s|A^s \neq \emptyset, G_1, G_2) > 0$, we have now shown that

$$\Pr(\boldsymbol{\theta}_{i1} = \boldsymbol{\theta}_{i'2}|G_1, G_2) > 0,$$

which completes the proof. □

A.1.5 Additional Simulation Plots of Expected Number of Clusters for CAM, HDP, and PAM

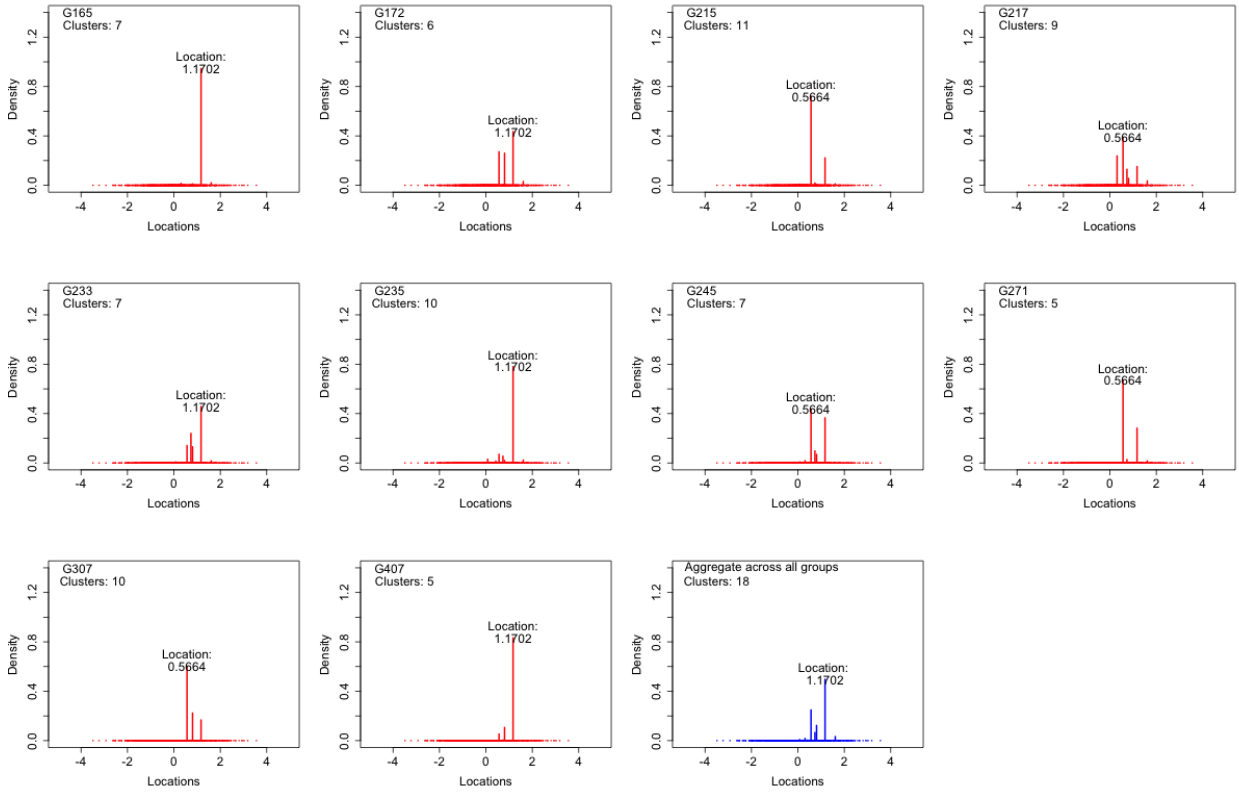

Figure A.1.1: Plots of simulated G_j for 10 randomly selected samples (subplots with red sticks) and the random distribution aggregating all 500 groups (bottom right subplot with blue sticks) for CAM$(1, 1, H)$, $H = N(0, 1)$. In each plot, the text "Gj" represents group j for $j \in \{1, \ldots, 500\}$, "Cluster:K_j" represents the number of clusters K in group j, and "Location:ϕ_k" represents the location that has the highest probability in the random discrete distribution G_j.

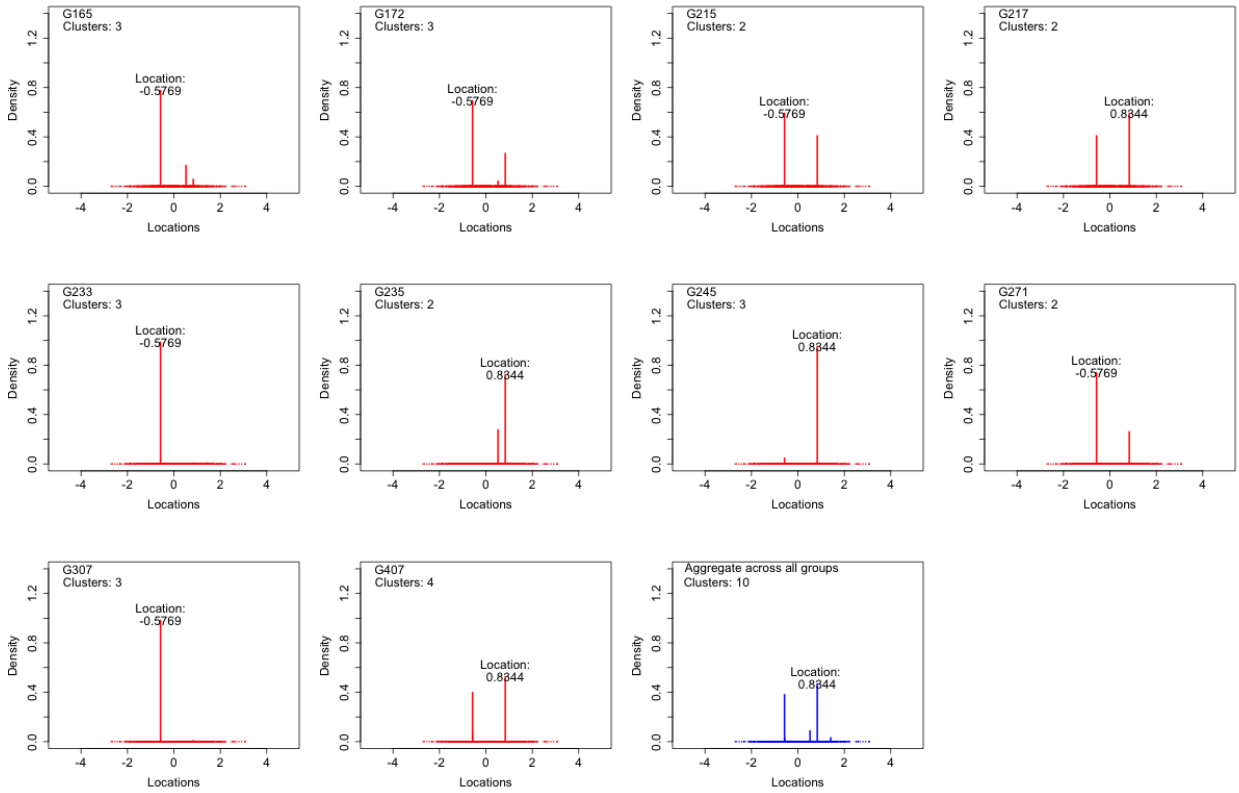

Figure A.1.2: Plots of simulated G_j for 10 randomly selected samples (subplots with red sticks) and the random distribution aggregating all 500 groups (bottom right subplot with blue sticks) for HDP$(1, 1, H)$, $H = N(0, 1)$. In each plot, the text "Gj" represents group j for $j \in \{1, \ldots, 500\}$, "Cluster:K_j" represents the number of clusters K in group j, and "Location:ϕ_k" represents the location that has the highest probability in the random discrete distribution G_j.

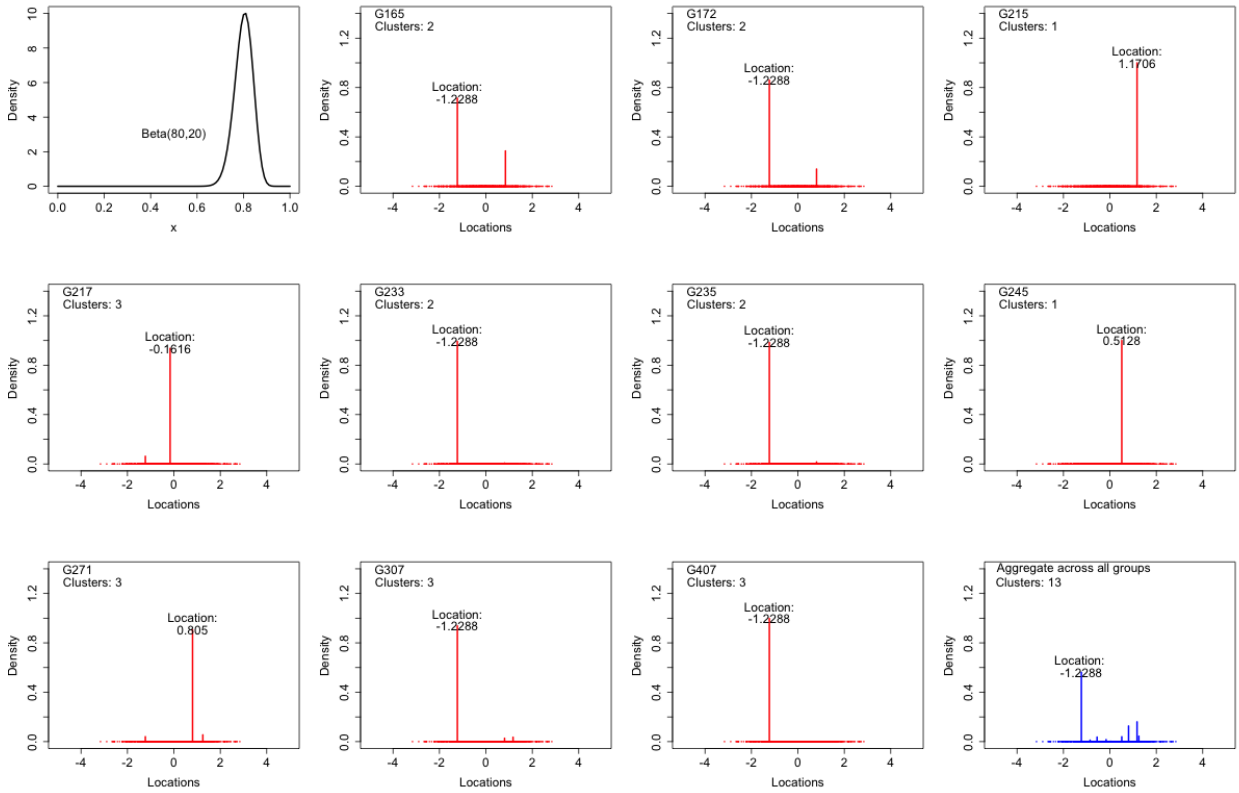

Figure A.1.3: Plots of prior for $\boldsymbol{p}_1$ (top left subplot), simulated G_j for 10 randomly selected samples (subplots with red sticks), and the random distribution aggregating all 500 groups (bottom right subplot with blue sticks) for PAM$(\boldsymbol{p}_1, 1, 1, H)$, $H = N(0, 1)$, $p_{j1} \sim$ Beta$(80, 20)$. In each plot, the text "Gj" represents group j for $j \in \{1, \ldots, 500\}$, "Cluster:K_j" represents the number of clusters K in group j, and "Location:ϕ_k" represents the location that has the highest probability in the random discrete distribution G_j.

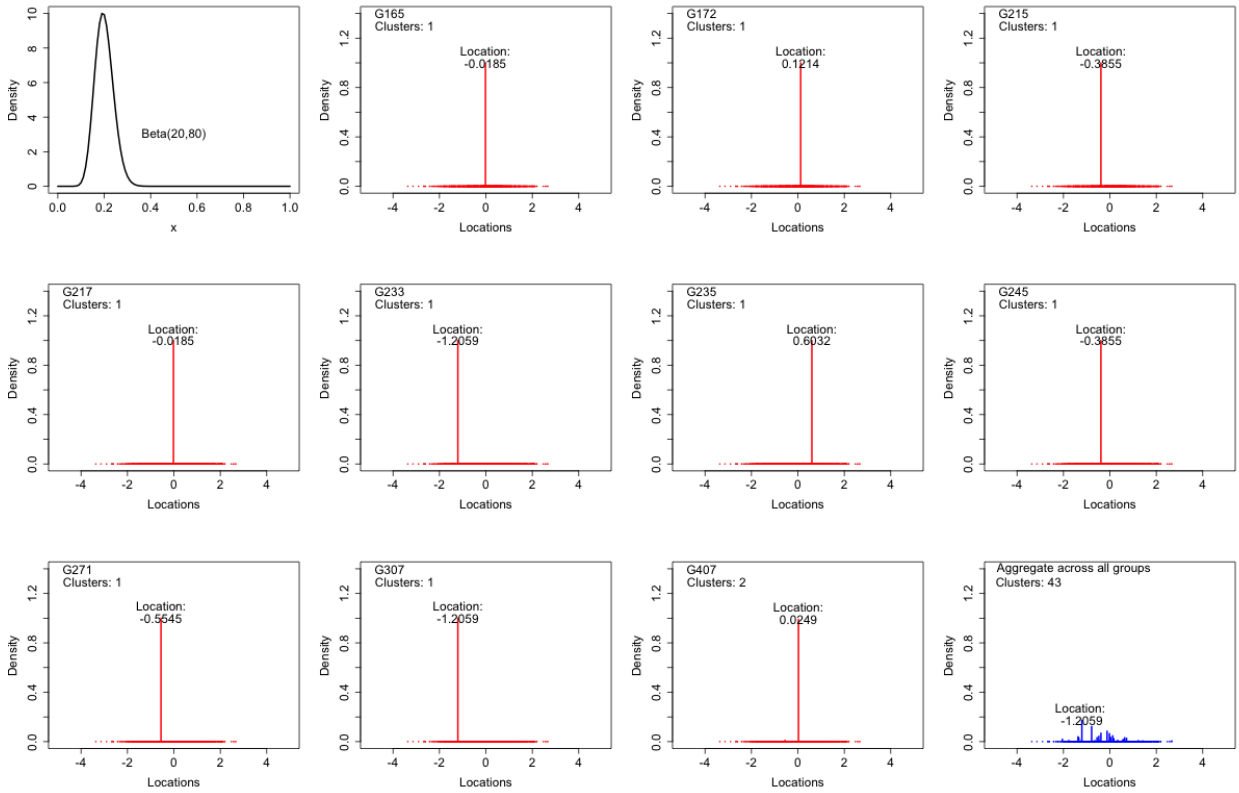

Figure A.1.4: Plots of prior for $\boldsymbol{p}_2$ (top left subplot), simulated G_j for 10 randomly selected samples (subplots with red sticks), and the random distribution aggregating all 500 groups (bottom right subplot with blue sticks) for PAM$(\boldsymbol{p}_2, 1, 1, H)$, $H = N(0, 1)$, $p_{j2} \sim$ Beta$(20, 80)$. In each plot, the text "Gj" represents group j for $j \in \{1, \ldots, 500\}$, "Cluster:K_j" represents the number of clusters K in group j, and "Location:ϕ_k" represents the location that has the highest probability in the random discrete distribution G_j.

A.1.6 Proof of Theorem 2

The FSBP is a special case of the kernel stick-breaking process of Dunson and Park [2008]. Using their notation, the kernel function $K(\boldsymbol{x}, \Gamma_k) = p$, i.e., constant over k and independent of covariates. Thus, their theoretical results are applicable in our case. From equation (4) of Dunson and Park [2008], the mean of G^* is immediate and given by

$$E[G^*(A)] = E\left[E[G^*(A)|\boldsymbol{\beta'}, p]\right] = E[H(A)] = H(A),$$

where $\boldsymbol{\beta'} = \{\beta'_k; k \geq 1\}, \beta'_k \sim \text{Beta}(1, \gamma)$. To find the variance of G^*, apply equation (7) of Theorem 1 of Dunson and Park [2008]

$$\text{Var}(G^*(A)) = \frac{\mu^{(2)} Var_{Q(A)}}{2\mu - \mu^{(2)}} \tag{A.1.5}$$

where

$$Var_{Q(A)} = Var_H\{\delta_{\boldsymbol{\phi}_k}(A)\} = H(A)(1 - H(A)),$$

$$\mu = p \cdot E[\beta'_k] = \frac{p}{1+\gamma},$$

and

$$\mu^{(2)} = p^2 \cdot E[{\beta'_k}^2] = \frac{2p^2}{(1+\gamma)(2+\gamma)}.$$

Substituting the expression for $Var_{Q(A)}$, $\mu(x)$, and $\mu^{(2)}(x)$ into equation (A.1.5), we obtain

$$Var(G^*(A)) = \frac{H(A)(1 - H(A))}{\frac{1+\gamma}{p} + \frac{1-p}{p}}.$$

□

A.1.7 Proof of Theorem 3

Denote $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots\}$ the atoms in G^*. Consider n samples generated from G^*, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_i; i = 1, \ldots, n\}$, $\boldsymbol{\theta}_i | G^* \sim G^*$, and $\boldsymbol{\theta}_i$ takes a value in $\boldsymbol{\Phi}$ with a probability. Assume there are K clusters, denote the atoms associated with the K clusters by $\boldsymbol{\Phi}_K = \{\boldsymbol{\phi}_{r_1}, \ldots, \boldsymbol{\phi}_{r_K}\}$ where each r_k indexes the kth cluster and $r_k \in \mathbb{N}$, where $\mathbb{N}$ denotes the set of all natural numbers, i.e., $\mathbb{N} = \{1, 2, \ldots\}$. Denote $\nabla = \{r_1, \ldots, r_K\}$ the index set in ascending order of the K clusters, i.e., $r_1 < r_2 < \ldots < r_K$. Let $\boldsymbol{z} = \{z_1, \ldots, z_n\}$ be the cluster label where $\{z_i = k\}$ means observation $\boldsymbol{\theta}_i$ belongs to cluster k, i.e., $\{\boldsymbol{\theta}_i = \boldsymbol{\phi}_{r_k}\}$. Further, denote $c_k = \{i; z_i = k\}$ the indices of $\boldsymbol{\theta}_i$'s belonging to cluster k. It is important to note that the cluster label k's do not need to be consecutive integers. For example, $K = 3$ and $\nabla = \{1, 3, 5\}$ or $K = 5$ and $\nabla = \{2, 5, 6, 20, 100\}$. Lastly, assume the unique value of the kth cluster is the atom $\boldsymbol{\phi}_k$, i.e., $\{\boldsymbol{\theta}_i = \boldsymbol{\phi}_k\}$ if $\{z_i = k\}$, for $k \in \nabla$.

Let $m = \max(z_1, \cdots, z_n)$. It follows that $K \leq m$ due to the fact that the cluster labels do not need to be consecutive integers. A partition $\boldsymbol{z}$ of the n samples $\boldsymbol{\Theta}$ is then denoted as $C(\boldsymbol{z}) = \{c_k; k \in \nabla\}$, the collection of c_k's, where $c_k \cap c_{k'} = \emptyset$ for $k \neq k'$, $|C(\boldsymbol{z})| = K$, and $\cup_{k \in \nabla} c_k = \{1, \ldots, n\}$. Here, $|.|$ refers to the cardinality of a set. The EPPF of G^* evaluated at a specific partition C is given by

$$\Pr(C(\boldsymbol{z}) = C) = \sum_{\boldsymbol{z}^* \in \mathbb{N}^n} \Pr(C(\boldsymbol{z}^*) = C | \boldsymbol{z} = \boldsymbol{z}^*) \Pr(\boldsymbol{z} = \boldsymbol{z}^*) = \sum_{\boldsymbol{z}^* \in \mathbb{N}^n} I(C(\boldsymbol{z}^*) = C) \Pr(\boldsymbol{z} = \boldsymbol{z}^*) \tag{A.1.6}$$

where $\mathbb{N}^n$ is the $n-$dimensional space of positive integers. The second equality is true since given $\boldsymbol{z} = \boldsymbol{z}^*$, $C(\boldsymbol{z}^*)$ is fixed and is either equal to C or not.

We first find $\Pr(\boldsymbol{z} = \boldsymbol{z}^*)$. For a specific $\boldsymbol{z}^* = \{z_1^*, \ldots, z_n^*\}$, denote $e_k(\boldsymbol{z}^*) = |\{i; z_i^* = k\}|$, $f_k(\boldsymbol{z}^*) = |\{i; z_i^* > k\}|$, and $g_k(\boldsymbol{z}^*) = |\{i; z_i^* \geq k\}|$. Also let $m(\boldsymbol{z}^*) = \max(z_1^*, \ldots, z_n^*)$. Recall

the definition of FSBP in Section 3.3, with $a = 1$ and $b = \gamma$, $\pi'_k \sim \text{Beta}(1, \gamma)$, and we have

$$\Pr(\boldsymbol{z} = \boldsymbol{z}^*) = \int \Pr(\boldsymbol{z}^* | \pi'_1, \cdots, \pi'_{m(\boldsymbol{z}^*)}) p(\pi'_1) \cdots p(\pi'_{m(\boldsymbol{z}^*)}) d\pi'_1 \cdots d\pi'_{m(\boldsymbol{z}^*)}$$

$$= \int \left[\prod_{k=1}^{m(\boldsymbol{z}^*)} \left\{ p\pi'_k \prod_{l<k} (1 - p\pi'_l) \right\}^{e_k(\boldsymbol{z}^*)} \right] p(\pi'_1) \cdots p(\pi'_{m(\boldsymbol{z}^*)}) d\pi'_1 \cdots d\pi'_{m(\boldsymbol{z}^*)}$$

$$= \int \left[\prod_{k=1}^{m(\boldsymbol{z}^*)} (p\pi'_k)^{e_k(\boldsymbol{z}^*)} (1 - p\pi'_k)^{f_k(\boldsymbol{z}^*)} \right] p(\pi'_1) \cdots p(\pi'_{m(\boldsymbol{z}^*)}) d\pi'_1 \cdots d\pi'_{m(\boldsymbol{z}^*)}$$

$$= \prod_{k=1}^{m(\boldsymbol{z}^*)} \left\{ \frac{p^{e_k(\boldsymbol{z}^*)}}{B(1, \gamma)} \int {\pi'_k}^{e_k(\boldsymbol{z}^*)} (1 - p\pi'_k)^{f_k(\boldsymbol{z}^*)} (1 - \pi'_k)^{\gamma - 1} d\pi'_k \right\}$$

where $B(a, b)$ is the Beta function with parameters a and b. If we re-write the integral of the last step as the follows:

$$\int {\pi'_k}^{(e_k(\boldsymbol{z}^*)+1)-1} (1 - p\pi'_k)^{-(-f_k(\boldsymbol{z}^*))} (1 - \pi'_k)^{(\gamma + e_k(\boldsymbol{z}^*)+1)-(e_k(\boldsymbol{z}^*)+1)-1} d\pi'_k,$$

it is easy to see that this integration can be written as the Euler type hypergeometric function. Thus, we have

$$\int {\pi'_k}^{e_k(\boldsymbol{z}^*)} (1 - p\pi'_k)^{f_k(\boldsymbol{z}^*)} (1 - \pi'_k)^{\gamma - 1} d\pi'_k$$

$$= B(e_k(\boldsymbol{z}^*) + 1, \gamma) {}_2F_1(-f_k(\boldsymbol{z}^*), e_k(\boldsymbol{z}^*) + 1; \gamma + e_k(\boldsymbol{z}^*) + 1; p)$$

where ${}_2F_1(a, b; c; d)$ is the hypergeometric function with parameters a, b, c and d. Consequently, we have

$$\Pr(\boldsymbol{z} = \boldsymbol{z}^*) = \prod_{k=1}^{m(\boldsymbol{z}^*)} \left\{ \frac{p^{e_k(\boldsymbol{z}^*)}}{B(1, \gamma)} B(e_k(\boldsymbol{z}^*) + 1, \gamma) {}_2F_1(-f_k(\boldsymbol{z}^*), e_k(\boldsymbol{z}^*) + 1; \gamma + e_k(\boldsymbol{z}^*) + 1; p) \right\}$$

$$= \prod_{k=1}^{m(\boldsymbol{z}^*)} \left\{ \frac{p^{e_k(\boldsymbol{z}^*)}\Gamma(\gamma+1)}{\Gamma(\gamma)} \frac{\Gamma(e_k(\boldsymbol{z}^*)+1)\Gamma(\gamma)}{\Gamma(\gamma+e_k(\boldsymbol{z}^*)+1)} \, {}_2F_1(-f_k(\boldsymbol{z}^*), e_k(\boldsymbol{z}^*)+1; \gamma+e_k(\boldsymbol{z}^*)+1; p) \right\}$$

$$= \prod_{k=1}^{m(\boldsymbol{z}^*)} \left\{ p^{e_k(\boldsymbol{z}^*)} \frac{\Gamma(\gamma+1)\Gamma(e_k(\boldsymbol{z}^*)+1)}{\Gamma(\gamma+e_k(\boldsymbol{z}^*)+1)} \, {}_2F_1(-f_k(\boldsymbol{z}^*), e_k(\boldsymbol{z}^*)+1; \gamma+e_k(\boldsymbol{z}^*)+1; p) \right\}$$

$$= \left\{ \prod_{k=1}^{m(\boldsymbol{z}^*)} \Gamma(\gamma+1) p^{e_k(\boldsymbol{z}^*)} \frac{\Gamma(e_k(\boldsymbol{z}^*)+1)}{\Gamma(\gamma+e_k(\boldsymbol{z}^*)+1)} \right\} \times$$

$$\left\{ \prod_{k=1}^{m(\boldsymbol{z}^*)} {}_2F_1(-f_k(\boldsymbol{z}^*), e_k(\boldsymbol{z}^*)+1; \gamma+e_k(\boldsymbol{z}^*)+1; p) \right\}$$

$$= \left\{ \prod_{c \in C(\boldsymbol{z}^*)} \Gamma(\gamma+1) p^{|c|} \frac{\Gamma(|c|+1)}{\Gamma(\gamma+|c|+1)} \right\} \times$$

$$\left\{ \prod_{k=1}^{m(\boldsymbol{z}^*)} {}_2F_1(-g_{k+1}(\boldsymbol{z}^*), e_k(\boldsymbol{z}^*)+1; \gamma+e_k(\boldsymbol{z}^*)+1; p) \right\} \tag{A.1.7}$$

where $f_k(\boldsymbol{z}^*) = g_{k+1}(\boldsymbol{z}^*)$.

Back to equation (A.1.6) and substituting in equation (A.1.7), we have

$$\Pr(C(\boldsymbol{z}) = C) = \left\{ \prod_{c \in C} \Gamma(\gamma+1) p^{|c|} \frac{\Gamma(|c|+1)}{\Gamma(\gamma+|c|+1)} \right\} \times$$

$$\underbrace{\sum_{\boldsymbol{z}^* \in \mathbb{N}^n} I(C(\boldsymbol{z}^*) = C) \left\{ \underbrace{\prod_{k=1}^{m(\boldsymbol{z}^*)} {}_2F_1(-g_{k+1}(\boldsymbol{z}^*), e_k(\boldsymbol{z}^*)+1; \gamma+e_k(\boldsymbol{z}^*)+1; p)}_{(A)} \right\}}_{(B)}. \tag{A.1.8}$$

Now, recall $K = |C|$ is the number of unique clusters in the n samples, and $C = \{c_1, \ldots, c_K\}$. Denote S_K the set of all $K!$ permutations of $\{1, \ldots, K\}$, and denote $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_K\} \in S_K$ a permutation of $\{1, \ldots, K\}$. For any $\boldsymbol{\lambda} \in S_K$, define $\alpha_k(\boldsymbol{\lambda}) =$

$|c_{\lambda_k}| + \cdots + |c_{\lambda_K}|$. By definition, $\alpha_{K+1}(\boldsymbol{\lambda}) = 0$. Consider a given $\boldsymbol{z}^*$ such that $C(\boldsymbol{z}^*) = C$, recall that $r_1, \ldots, r_K$ are the distinct values of $\boldsymbol{z}^*$ in ascending order, i.e., $r_1 < r_2 < \cdots < r_k < \cdots < r_K$, $r_k \in \mathbb{N}$, we can rewrite the (A) term in (A.1.8) as

$$\begin{aligned}
&\prod_{k=1}^{m(\boldsymbol{z}^*)} {}_2F_1(-g_{k+1}(\boldsymbol{z}^*), e_k(\boldsymbol{z}^*) + 1; \gamma + e_k(\boldsymbol{z}^*) + 1; p) \\
&= [{}_2F_1(-g_{r_2}(\boldsymbol{z}^*), e_{r_1}(\boldsymbol{z}^*) + 1; \gamma + e_{r_1}(\boldsymbol{z}^*) + 1; p)]^{r_1} \times \\
&[{}_2F_1(-g_{r_3}(\boldsymbol{z}^*), e_{r_2}(\boldsymbol{z}^*) + 1; \gamma + e_{r_2}(\boldsymbol{z}^*) + 1; p)]^{r_2 - r_1} \times \cdots \times \\
&\left[{}_2F_1(-g_{r_K+1}(\boldsymbol{z}^*), e_{r_K}(\boldsymbol{z}^*) + 1; \gamma + e_{r_K}(\boldsymbol{z}^*) + 1; p)\right]^{r_K - r_{K-1}} \\
&= \left[{}_2F_1(-\alpha_2(\boldsymbol{\lambda}), |c_{\lambda_1}| + 1; \gamma + |c_{\lambda_1}| + 1; p)\right]^{d_1} \times \\
&\left[{}_2F_1(-\alpha_3(\boldsymbol{\lambda}), |c_{\lambda_2}| + 1; \gamma + |c_{\lambda_2}| + 1; p)\right]^{d_2} \times \cdots \times \\
&\left[{}_2F_1(-\alpha_{K+1}(\boldsymbol{\lambda}), |c_{\lambda_K}| + 1; \gamma + |c_{\lambda_K}| + 1; p)\right]^{d_K} \\
&= \prod_{k=1}^{K} \left[{}_2F_1(-\alpha_{k+1}(\boldsymbol{\lambda}), |c_{\lambda_k}| + 1; \gamma + |c_{\lambda_k}| + 1; p)\right]^{d_k}
\end{aligned}$$

where $\boldsymbol{d} = (d_1, \ldots, d_K)$, $d_1 = r_k$, and $d_k = r_k - r_{k-1}$ for $k = 2, \ldots, K$. For any $\boldsymbol{z}^* \in \mathbb{N}^n$, note that the definition of $\boldsymbol{d}$ and $\boldsymbol{\lambda}$ sets up a one-to-one correspondence, which is a bijection, between $\{\boldsymbol{z}^* \in \mathbb{N}^n; C(\boldsymbol{z}^*) = C\}$ and $\{(\boldsymbol{\lambda}, \boldsymbol{d}); \boldsymbol{\lambda} \in S_K, \boldsymbol{d} \in \mathbb{N}^K\}$, and the expression in (B) in (A.1.8) can then be rewritten as

$$\begin{aligned}
&\sum_{\boldsymbol{z}^* \in \mathbb{N}^n} I(C(\boldsymbol{z}^*) = C) \left\{ \prod_{k=1}^{m(\boldsymbol{z}^*)} {}_2F_1(-g_{k+1}(\boldsymbol{z}^*), e_k(\boldsymbol{z}^*) + 1; \gamma + e_k(\boldsymbol{z}^*) + 1; p) \right\} \\
&= \sum_{\boldsymbol{\lambda} \in S_K} \sum_{\boldsymbol{d} \in \mathbb{N}^K} \prod_{k=1}^{K} \left[{}_2F_1(-\alpha_{k+1}(\boldsymbol{\lambda}), |c_{\lambda_k}| + 1; \gamma + |c_{\lambda_k}| + 1; p)\right]^{d_k}
\end{aligned}$$

$$\overset{(a)}{=} \sum_{\boldsymbol{\lambda}\in S_K} \prod_{k=1}^{K} \sum_{d_k\in\mathbb{N}} \left[{}_2F_1(-\alpha_{k+1}(\boldsymbol{\lambda}), |c_{\lambda_k}|+1; \gamma+|c_{\lambda_k}|+1; p)\right]^{d_k}$$

$$\overset{(b)}{=} \sum_{\boldsymbol{\lambda}\in S_K} \prod_{k=1}^{K} \left\{ \frac{{}_2F_1(-\alpha_{k+1}(\boldsymbol{\lambda}), |c_{\lambda_k}|+1; \gamma+|c_{\lambda_k}|+1; p)}{1-\ {}_2F_1(-\alpha_{k+1}(\boldsymbol{\lambda}), |c_{\lambda_k}|+1; \gamma+|c_{\lambda_k}|+1; p)} \right\} \tag{A.1.9}$$

where the second equality (a) can be shown as the follows: let $f(\alpha_k(\boldsymbol{\lambda})) = {}_2F_1(-\alpha_{k+1}(\boldsymbol{\lambda}), |c_{\lambda_k}|+1; \gamma+|c_{\lambda_k}|+1; p)$, then

$$\sum_{\boldsymbol{d}\in\mathbb{N}^K} \prod_{k=1}^{K} f(\alpha_k(\boldsymbol{\lambda}))^{d_k} = \sum_{d_1\in\mathbb{N}} \cdots \sum_{d_K\in\mathbb{N}} \left[f(\alpha_1(\boldsymbol{\lambda}))^{d_1} \cdots f(\alpha_K(\boldsymbol{\lambda}))^{d_K} \right]$$

$$= \sum_{d_1\in\mathbb{N}} \cdots \sum_{d_{K-2}\in\mathbb{N}} \left\{ \sum_{d_{K-1}\in\mathbb{N}} f(\alpha_1(\boldsymbol{\lambda}))^{d_1} \cdots f(\alpha_{K-1}(\boldsymbol{\lambda}))^{d_{K-1}} \left(\sum_{d_K\in\mathbb{N}} f(\alpha_K(\boldsymbol{\lambda}))^{d_K} \right) \right\}$$

$$= \left(\sum_{d_K\in\mathbb{N}} f(\alpha_K(\boldsymbol{\lambda}))^{d_K} \right) \left\{ \sum_{d_1\in\mathbb{N}} \cdots \sum_{d_{K-2}\in\mathbb{N}} \left\{ \sum_{d_{K-1}\in\mathbb{N}} f(\alpha_1(\boldsymbol{\lambda}))^{d_1} \cdots f(\alpha_{K-1}(\boldsymbol{\lambda}))^{d_{K-1}} \right\} \right\}$$

$$= \left(\sum_{d_1\in\mathbb{N}} f(\alpha_1(\boldsymbol{\lambda}))^{d_1} \right) \times \cdots \times \left(\sum_{d_K\in\mathbb{N}} f(\alpha_K(\boldsymbol{\lambda}))^{d_K} \right) = \prod_{k=1}^{K} \sum_{d_k\in\mathbb{N}} f(\alpha_k(\boldsymbol{\lambda}))^{d_k}.$$

And the last equality (b) of equation (A.1.9) is due to geometric series: $\sum_{d=1}^{\infty}(r^d) = 1/(1-r) - 1 = r/(1-r)$. Moreover, ${}_2F_1(-\alpha_{k+1}(\boldsymbol{\lambda}), |c_{\lambda_k}|+1; \gamma+|c_{\lambda_k}|+1; p)$ is between 0 and 1. This can be seen from the derivative of the hypergeometric function:

$$\frac{d}{dp}{}_2F_1(-\alpha_{k+1}(\boldsymbol{\lambda}), |c_{\lambda_k}|+1; \gamma+|c_{\lambda_k}|+1; p)$$

$$= -\frac{\alpha_{k+1}(\boldsymbol{\lambda})(|c_{\lambda_k}|+1)}{(\gamma+|c_{\lambda_k}|+1)}\, {}_2F_1(-\alpha_{k+1}(\boldsymbol{\lambda})+1, |c_{\lambda_k}|+2; \gamma+|c_{\lambda_k}|+2; p)$$

$$= -\frac{\alpha_{k+1}(\boldsymbol{\lambda})(|c_{\lambda_k}|+1)}{(\gamma+|c_{\lambda_k}|+1)}(1-p)^{\gamma+\alpha_{k+1}(\boldsymbol{\lambda})}\, {}_2F_1(\gamma+|c_{\lambda_k}|+\alpha_{k+1}(\boldsymbol{\lambda})+1, \gamma; \gamma+|c_{\lambda_k}|+2; p) < 0.$$

Since the derivative is less than zero, the function monotonically decrease with p. For

$p \in (0,1]$, the hypergeometric function ${}_2F_1(-\alpha_{k+1}(\boldsymbol{\lambda}), |c_{\lambda_k}| + 1; \gamma + |c_{\lambda_k}| + 1; p)$ equals 1 when $p = 0$ and equals

$$0 < \frac{(\gamma)_{\alpha_{k+1}(\boldsymbol{\lambda})}}{(\gamma + |c_{\lambda_k}| + 2)_{\alpha_{k+1}(\boldsymbol{\lambda})}} < 1$$

when $p = 1$, where $(a)_b$ is the rising Pochhammer symbol defined as $(a)_b = 1$ if $b = 0$ and $(a)_b = a(a+1)\cdots(a+b-1)$ if $b > 0$. Substituting (A.1.9) into (B) of (A.1.8), we have proved the EPPF of Theorem 3.

Lastly, for the claim of the EPPF of G^* converging to the EPPF of $G_0 \sim DP(1, H)$ when $p \to 1$, the hypergeometric function

$${}_2F_1(-f_k(\boldsymbol{z}^*), e_k(\boldsymbol{z}^*) + 1; \gamma + e_k(\boldsymbol{z}^*) + 1; 1) = \frac{\Gamma(\gamma + e_k(\boldsymbol{z}^*) + 1)\Gamma(\gamma + f_k(\boldsymbol{z}^*))}{\Gamma(\gamma)\Gamma(\gamma + e_k(\boldsymbol{z}^*) + f_k(\boldsymbol{z}^*) + 1)}$$

$$= \frac{\Gamma(\gamma + e_k(\boldsymbol{z}^*) + 1)\Gamma(\gamma + f_k(\boldsymbol{z}^*))}{\Gamma(\gamma)\Gamma(\gamma + g_k(\boldsymbol{z}^*) + 1))}$$

where $g_k(\boldsymbol{z}^*) = f_k(\boldsymbol{z}^*) + e_k(\boldsymbol{z}^*)$. And equation (A.1.7) becomes

$$\prod_{k=1}^{m(\boldsymbol{z}^*)} \frac{\Gamma(\gamma+1)\Gamma(e_k(\boldsymbol{z}^*)+1)\Gamma(\gamma + e_k(\boldsymbol{z}^*) + 1)\Gamma(\gamma + f_k(\boldsymbol{z}^*))}{\Gamma(\gamma + e_k(\boldsymbol{z}^*) + 1)\Gamma(\gamma)\Gamma(\gamma + g_k(\boldsymbol{z}^*) + 1)}$$

$$= \prod_{k=1}^{m(\boldsymbol{z}^*)} \frac{\gamma\Gamma(e_k(\boldsymbol{z}^*)+1)\Gamma(f_k(\boldsymbol{z}^*) + \gamma)}{\Gamma(g_k(\boldsymbol{z}^*) + \gamma + 1)},$$

which then equals the right-hand side of the sixth equal sign of equation $\Pr(\boldsymbol{z} = z)$ in the proof of Lemma 2.2 in Miller [2019]. Then the author shows that (Proof of Theorem 2.1 therein) the EPPF of $G_0 \sim DP(\gamma, H)$ can be written as

$$\Pr(C(\boldsymbol{z}) = C) = \frac{\gamma^{|C|}\Gamma(\gamma)}{\Gamma(n+\gamma)} \prod_{c \in C} \Gamma(|c|).$$

A.1.8 Proof Lemma 1

Since $G^* \sim FSBP(p, \gamma, H)$, consider the following prediction rule for samples $\boldsymbol{\theta}_i|\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{i-1}$, where $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_i|G^* \sim G^*$:

$$\Pr(\boldsymbol{\theta}_i|\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{i-1}) = W_{\text{base}_i} H + \sum_{l=1}^{i-1} W_{i_l} \delta_{\boldsymbol{\theta}_l}$$

where W_{base_i} corresponds to the probability $\boldsymbol{\theta}_i$ sampled from the base probability measure H (and not equal to any $\boldsymbol{\theta}_l \in \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}\}$) when there are i samples, and W_{i_l} corresponds to the probability of $\boldsymbol{\theta}_i$ sampled from a previously seen $\boldsymbol{\theta}_l$ for $l = 1, \ldots, i-1$. Then, we have

$$\Pr(w_i = 1|p, \gamma) = \Pr(\boldsymbol{\theta}_i \notin \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}\}|G^*) = W_{\text{base}_i}.$$

W_{base_i} can be evaluated by (using the prediction rule in Theorem 2 of Dunson and Park [2008])

$$W_{\text{base}_i} = \left\{ 1 - \sum_{k=2}^{i} (-1)^k \sum_{I \in N_i^{(k,i)}} \omega_I \right\},$$

where $N_i^{(k,i)}$ is a set contains all possible k-dimensional subsets of $\{1, \cdots, i\}$ that includes index i, with I an element (a set) in the set, $\omega_I = \mu_I \cdot \left(\sum_{l=1}^{|I|} (-1)^{l-1} \sum_{m \in I_l} \mu_m \right)^{-1}$, $\mu_I = E[\prod_{k \in I} p\pi_k{}']$, and I_l the set of length-l subsets of the set I. The cardinality of the sets $N_i^{(k,i)}$, I, and I_l are $|N_i^{(k,i)}| = \binom{i-1}{k-1}$, $|I| = k$, and $|I_l| = \binom{k}{l}$, respectively. For example, let $i = 3$, $k = 2$, and $l = 1$. $N_{i=3}^{(k=2,i=3)} = \{I_1, I_2\} = \{\{1, 3\}, \{2, 3\}\}$, with $|N_{i=3}^{(k=2,i=3)}| = 2$. Also, $|I_1| = |I_2| = 2$. And when $I = I_1$, $I_{l=1} = \{\{1\}, \{3\}\}$, and when $I = I_2$, $I_{l=2} = \{\{2\}, \{3\}\}$. Both have cardinality $|I_l| = 2$, $l = 1, 2$.

For G^*, recall $\pi'_k \sim \text{Beta}(1, \gamma)$. For a set I, $\mu_I = E\left[\prod_{k \in I} p\pi'_k\right]$, which can be shown to

be

$$\mu_I = p^{|I|} \prod_{l=1}^{|I|} \frac{l}{l+\gamma}.$$

Thus, μ_I depends on the cardinality of the set I only. Furthermore, for $\sum_{m\in I_l} \mu_m$ in the denominator of ω_I, μ_m can be similarly computed, and the values are the same for all $m \in I_l$ (since μ_m depends only on $|m|$, and all $m \in I_l$ are of the same cardinality that is equal to l). Plugging in μ_I and $\sum_{m\in I_l} \mu_m$ to the theorem, we have

$$\omega_I = \frac{p^{|I|} \prod_{l=1}^{|I|} \frac{l}{l+\gamma}}{\sum_{l=1}^{|I|} (-1)^{l-1} \binom{|I|}{l} p^l \prod_{m=1}^{l} \frac{m}{m+\gamma}},$$

which again only depends on the cardinality of the set I, i.e., $|I|$. Let $|N_i^{(k,i)}| = \binom{i-1}{k-1} = B$. Further notice that the sets in $N_i^{(k,i)}$, denoted as $I_1, \ldots, I_{b'}, \ldots, I_B$, have the same cardinality for a given k, i.e., $|I_{b'}| = k$ for all $b' \in \{1, ..., B\}$. Thus, we have

$$W_{\text{base}_i} = 1 - \sum_{k=2}^{i} (-1)^k \sum_{I \in N_i^{(k,i)}} \omega_I = 1 - \sum_{k=2}^{i} (-1)^k \binom{i-1}{k-1} \frac{p^k \prod_{l=1}^{k} \frac{l}{l+\gamma}}{\sum_{l=1}^{k} (-1)^{l-1} \binom{k}{l} p^l \prod_{m=1}^{l} \frac{m}{m+\gamma}}$$

$$= 1 - \sum_{k=2}^{i} (-1)^k \binom{i-1}{k-1} \frac{k!}{\prod_{l=1}^{k}(l+\gamma)} \frac{p^{k-1}}{\sum_{l=1}^{k} (-1)^{l-1} \binom{k}{l} p^{l-1} \frac{l!}{\prod_{m=1}^{l}(m+\gamma)}}$$

$$= 1 - \sum_{k=2}^{i} (-1)^k \binom{i-1}{k-1} \frac{k!}{\prod_{l=1}^{k}(l+\gamma)} \frac{(\gamma+1)p^{k-1}}{k \times {}_2F_1(1, 1-k; \gamma+2; p)}$$

$$= 1 - \sum_{k=2}^{i} (-1)^k \binom{i-1}{k-1} \frac{(k-1)!}{\prod_{l=1}^{k}(l+\gamma)} \frac{(\gamma+1)p^{k-1}}{{}_2F_1(1, 1-k; \gamma+2; p)}. \tag{A.1.10}$$

where ${}_2F_1(a, b; c; z)$ is the hypergeometric function.

Since FSBP is a special case of KSBP, and in KSBP, $W_{\text{base}_i} \in (0,1)$, we have

$$0 < \sum_{k=2}^{i} (-1)^k \binom{i-1}{k-1} \frac{(k-1)!}{\prod_{l=1}^{k}(l+\gamma)} \frac{(\gamma+1)p^{k-1}}{{}_2F_1(1,1-k;\gamma+2;p)} < 1.$$

□

A.1.9 Proof Lemma 2

Setting let $p \to 1$ in equation (A.1.10), we have

$$\lim_{p\to 1} \Pr(w_i = 1|p,\gamma) = 1 - \sum_{k=2}^{i} (-1)^k \binom{i-1}{k-1} \frac{(k-1)!}{\prod_{l=1}^{k}(l+\gamma)} \frac{(\gamma+1)}{{}_2F_1(1,1-k;\gamma+2;1)}$$

$$\stackrel{(a)}{=} 1 - \sum_{k=2}^{i} (-1)^k \binom{i-1}{k-1} \frac{(k-1)!}{\prod_{l=1}^{k}(l+\gamma)} \frac{(\gamma+1)}{\frac{\gamma+1}{\gamma+k}}$$

$$= 1 - \sum_{k=2}^{i} (-1)^k \frac{\Gamma(i)}{\Gamma(i-k+1)} \frac{\Gamma(\gamma+1)}{\Gamma(\gamma+k)} = 1 - \frac{i-1}{\gamma+i-1} = \frac{\gamma}{\gamma+i-1},$$

where the second equality (a) is because

$${}_2F_1(1,1-k;\gamma+2;1) = \frac{\Gamma(\gamma+2)\Gamma(\gamma+k)}{\Gamma(\gamma+1)\Gamma(\gamma+1+k)} = \frac{(\gamma+1)\Gamma(\gamma+1)\Gamma(\gamma+k)}{\Gamma(\gamma+1)(\gamma+k)\Gamma(\gamma+k)} = \frac{\gamma+1}{\gamma+k}.$$

Notice that $\frac{\gamma}{\gamma+i-1}$ is the probability of generating a new sample $\boldsymbol{\theta}_i \notin \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{i-1}\}$, i.e., from the base measure, in DP. □

A.1.10 Proof Theorem 4

To show $\Pr(w_i = 1|p,\gamma) > \frac{\gamma}{\gamma+i-1}$, it is sufficient to show that $1 - \frac{\gamma}{\gamma+i-1} > 1 - \Pr(w_i = 1|p,\gamma)$, or

$$\frac{i-1}{\gamma+i-1} > \sum_{k=2}^{i} (-1)^k \binom{i-1}{k-1} \frac{(k-1)!}{\prod_{l=1}^{k}(l+\gamma)} \frac{(\gamma+1)p^{k-1}}{{}_2F_1(1,1-k;\gamma+2;p)}.$$

First, notice that the hypergeometric function ${}_2F_1(1, 1-k; \gamma+2; p)$ is monotonically decreasing with respect to p since

$$\frac{d}{dp}{}_2F_1(1, 1-k; \gamma+2; p) = -\frac{(k-1){}_2F_1(2, 2-k; \gamma+3; p)}{\gamma+2}$$

$$= -\frac{(k-1)(1-p)^{\gamma+k-1}{}_2F_1(\gamma+1, \gamma+k+1; \gamma+3; p)}{\gamma+2} < 0,$$

with ${}_2F_1(1, 1-k; \gamma+2; 0) = 1$ and ${}_2F_1(1, 1-k; \gamma+2; 1) = \frac{\gamma+1}{\gamma+k}$. As a result, $\frac{1}{{}_2F_1(1,1-k;\gamma+2;p)}$ is monotonically increasing with p, with maximum at $p \to 1$, and

$$\lim_{p\to 1} \frac{1}{{}_2F_1(1, 1-k; \gamma+2; p)} = \frac{\gamma+k}{\gamma+1}.$$

Next, when substituting this maximum for ${}_2F_1(1, 1-k; \gamma+2; p)$, it can be shown that

$$\frac{i-1}{\gamma+i-1} - \sum_{k=2}^{i}(-1)^k \binom{i-1}{k-1} \frac{(k-1)!}{\prod_{l=1}^{k}(l+\gamma)} \frac{(\gamma+1)p^{k-1}}{{}_2F_1(1, 1-k; \gamma+2; p)}$$

$$> \frac{i-1}{\gamma+i-1} - \sum_{k=2}^{i}(-1)^k \binom{i-1}{k-1} \frac{(k-1)!}{\prod_{l=1}^{k}(l+\gamma)} \frac{(\gamma+1)p^{k-1}(\gamma+k)}{\gamma+1}$$

$$= \frac{i-1}{\gamma+i-1} - \sum_{k=2}^{i}(-1)^k \frac{\Gamma(i)}{\Gamma(i-k+1)} \frac{\Gamma(\gamma+1)}{\Gamma(\gamma+k)} p^{k-1}$$

$$= \frac{i-1}{\gamma+i-1} - \frac{(i-1){}_2F_1(1, 2-i; \gamma+2; p)p}{\gamma+1}.$$

Now, for $p \cdot {}_2F_1(1, 2-i; \gamma+2; p)$, from the property of hypergeometric function, we have $0 \cdot {}_2F_1(1, 2-i; \gamma+2; 0) = 0 \cdot 1 = 0$, and $1 \cdot {}_2F_1(1, 2-i; \gamma+2; 1) = \frac{\Gamma(\gamma+2)\Gamma(\gamma+i-1)}{\Gamma(\gamma+1)\Gamma(\gamma+i)} = \frac{\gamma+1}{\gamma+i-1}$. In addition, we have

$$\frac{d}{dp}{}_2F_1(1, 2-i; \gamma+2; p)p = {}_2F_1(2, 2-i; \gamma+2; p)$$

$$= (1-p)^{\gamma+i-2}{}_2F_1(\gamma, \gamma+i; \gamma+2; p) > 0,$$

and therefore, $p \cdot {}_2F_1(1, 2-i; \gamma+2; p)$ monotonically increases with p, is equal to 0 if $p \to 0$, and is equal to $\frac{\gamma+1}{\gamma+i-1}$ if $p \to 1$. Consequently, for $p \in (0,1)$, we have

$$\frac{i-1}{\gamma+i-1} - \frac{(i-1){}_2F_1(1, 2-i; \gamma+2; p)p}{\gamma+1} > \frac{i-1}{\gamma+i-1} - \frac{i-1}{\gamma+i-1} = 0.$$

□

A.1.11 Additional Details on Posterior Inference

More details on the slice-efficient sampler To sample β'_k conditional on the other parameters and data, we use an Metropolis-Hastings (MH) step to sample from

$$p(\beta'_k|\cdots) \propto \prod_{\{(j,l);j=1,\ldots,J,\, l\geq k,\, \pi'_{jl}\neq 0\}} \left[\frac{{\pi'_{jl}}^{\alpha_0\beta_l - 1}(1-\pi'_{jl})^{\alpha_0\left(1-\sum_{s=1}^{l}\beta_s\right)-1}}{B(\alpha_0\beta_l, \alpha_0\left(1-\sum_{s=1}^{l}\beta_s\right)}\right] \times (1-\beta'_k)^{\gamma-1} \tag{A.1.11}$$

where $\beta_k = \beta'_k \prod_{l=1}^{k-1}(1-\beta'_l)$. In addition, we use a uniform distribution as the proposal density function: $\beta'_{k_{\text{prop}}} \sim \text{Unif}(\beta'_{k_{\text{curr}}} - \epsilon, \beta'_{k_{\text{curr}}} + \epsilon)$, where $\beta'_{k_{\text{prop}}}$ is the proposal, $\beta'_{k_{\text{curr}}}$ is the β'_k in current iteration, and $\epsilon \in (0,1)$ is the step size. If $\beta'_{k_{\text{prop}}} < 0$, we set $\beta'_{k_{\text{prop}}} = |\beta'_{k_{\text{prop}}}|$, and if $\beta'_{k_{\text{prop}}} > 1$, we set $\beta'_{k_{\text{prop}}} = 2 - \beta'_{k_{\text{prop}}}$. It can be shown the proposal density is symmetric.

To sample p_j with a prior of $p_j \sim \text{Beta}(a, b)$, we have

$$p(p_j|\cdots) \propto p_j^{\sum_k 1(\pi'_{jk}\neq 0)+a-1}(1-p_j)^{\sum_k I(\pi'_{jk}=0)+b-1}.$$

Denoting $m_{j0} = \sum_{k=1}^{K^*} I(\pi'_{jk} = 0)$ the number of zero weights, we can sample p_j as

$$p_j|\cdots \sim \text{Beta}(a + K^* - m_{j0}, b + m_{j0})$$

If we assume that the concentration parameters α_0 and γ are random with gamma priors, we can sample them using the procedure described in Escobar and West [1995] and Teh et al. [2004].In Teh et al. [2004], the authors show that the full conditional of α_0 and γ is based on a matrix $\boldsymbol{W} = \{w_{jk}; j = 1, \ldots, J, k \geq 1\}$ that records the number of tables in restaurant j serving dish k according to the Chinese restaurant franchise process, and the posterior of this matrix depends only on $\boldsymbol{Z}$ and $\boldsymbol{\beta}$. We use equation (40) of Teh et al. [2004] to construct a latent matrix $\boldsymbol{W}$ and then follow the same method as the HDP to sample both concentration parameters.

Label switching As shown in the manuscript, we use the ECR algorithm of Papastamoulis and Iliopoulos [2010] to resolve the issue of label switching. This algorithm post-processes the MCMC samples using label permutations. The idea behind ECR is based on the invariance of likelihood with respect to the permutation of component labels.

For each MCMC iteration with label matrix $\boldsymbol{Z}^{(m)} = \{z_{ij}^{(m)}; i = 1, \ldots, n_j, j = 1, \ldots, J\}$, $z_{ij}^{(m)} \in \{1, \cdots, K^{(m)}\}$, where the superscript (m) denotes the mth MCMC iteration, we can form a partition of the $N = \sum_{j=1}^{J} n_j$ observations based on $\boldsymbol{Z}^{(m)}$. With slightly abuse of notation, we denote the corresponding unique labels of $\boldsymbol{Z}^{(m)}$ as $\boldsymbol{t}^{(m)} = \{t_1^{(m)}, \cdots, t_{K^{(m)}}^{(m)}\}$, $t_k^{(m)} \in \{1, \cdots, K^{(m)}\}$. For example, suppose we have a sample of $N = 7$ observations across $J = 2$ groups, $\boldsymbol{y} = \begin{bmatrix} y_{11} & y_{21} & y_{31} & \\ y_{12} & y_{22} & y_{32} & y_{42} \end{bmatrix}$, and two iterations of MCMC samples, i.e., $m = 1$ and $m = 2$. Assume in the MCMC samples, both partition the observations into the same 3 clusters, i.e., $K^{(1)} = K^{(2)} = 3$, Cluster A $= \{y_{11}, y_{12}, y_{22}\}$, Cluster B $= \{y_{21}, y_{31}\}$, and Cluster C $= \{y_{32}, y_{42}\}$, according to their corresponding $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$. However, in each of the two MCMC iterations, different labels of $\boldsymbol{t}^{(1)} = \{1, 2, 3\}$, with $\boldsymbol{Z}^{(1)} = \begin{bmatrix} 1 & 2 & 2 & \\ 1 & 1 & 3 & 3 \end{bmatrix}$, and $\boldsymbol{t}^{(2)} = \{2, 1, 3\}$, with $\boldsymbol{Z}^{(2)} = \begin{bmatrix} 2 & 1 & 1 & \\ 2 & 2 & 3 & 3 \end{bmatrix}$ are assigned to the observations. Thus, there is a switched label of Cluster A and Cluster B through $m = 1$

and $m = 2$. To resolve the label-switching issue, the method finds a permutation of labels at each MCMC iteration, denote as $\boldsymbol{\tau}^{(m)}(\boldsymbol{t}^{(m)})$, such that, compare to a reference label, say $\boldsymbol{t}^{(1)}$, $\boldsymbol{\tau}^{(2)}(\boldsymbol{t}^{(2)}) = \boldsymbol{t}^{(1)} = \{1, 2, 3\}$.

Specifically, the ECR method first picks an MCMC sample from one iteration (e.g., one close to MAP) as the reference label. Then, the method iterates over each MCMC sample of parameters of interest to find a random permutation of labels corresponding to the equivalent allocation of the reference label. We then switch the labels accordingly for all model parameters related to the cluster labels, i.e., label matrix $\boldsymbol{Z}$, MCMC samples of cluster weights $\{\pi_{jk}\}$, and cluster means $\{\boldsymbol{\phi}_k\}$. The ECR method is implemented in **R** package **label.switching** [Papastamoulis, 2015]. We use ECR to relabel the MCMC samples of the weights. After permuting the weights according to the result of ECR, we then explore the MCMC samples of the permuted weights for all j groups to learn the common and unique clusters in the groups.

Additional method to summarize common and unique clusters The second approach to summarize common and unique clusters is to use the posterior sample of the group-specific weights $\boldsymbol{\pi}_j^{(m)}$, $j = 1, \ldots, J$. Specifically, in the mth MCMC iteration, denote the number of common clusters between groups j and j' as $n_{\text{comm}}(\{\boldsymbol{\pi}_j^{(m)}, \boldsymbol{\pi}_{j'}^{(m)}\})$, and denote the number of unique clusters in group j as $n_{\text{uniq}}(\boldsymbol{\pi}_j^{(m)})$, we have

$$n_{\text{comm}}(\{\boldsymbol{\pi}_j^{(m)}, \boldsymbol{\pi}_{j'}^{(m)}\}) = \sum_{k=1}^{|\boldsymbol{\pi}_j^{(m)}|} 1(\pi_{jk}^{(m)} \neq 0 \text{ and } \pi_{j'k}^{(m)} \neq 0),$$

$$n_{\text{uniq}}(\boldsymbol{\pi}_j^{(m)}) = \sum_{k=1}^{|\boldsymbol{\pi}_j^{(m)}|} 1\left(\pi_{jk}^{(m)} \neq 0 \text{ and } \sum_{j' \in \{1,\cdots,j-1,j+1,\cdots,J\}} \pi_{j'k}^{(m)} = 0\right)$$

where $|\cdot|$ denotes the cardinality of the corresponding vector. Thus, the weight approach is able to learn the same information as the $\boldsymbol{Z}$ matrix method.

A.1.12 Slice Sampler for FSBP

Follow Kalli et al. [2011], we derive the slice sampler for PAM. From the model in manuscript, the density function for observation y_i can be rewritten as an infinite mixture

$$f_{\boldsymbol{\xi}}(y_i, u_i | z_i, \{\phi_k; k \geq 1\}, \{\pi_k; k \geq 1\}) = \sum_{k \geq 1} 1_{\{z_i = k\}} 1_{\{u_i < \xi_k\}} \frac{\pi_{z_i}}{\xi_{z_i}} p(y_i | \phi_{z_i}),$$

where u_i is the latent variable for observation i, and ξ_k is the same quantity as defined in the slice sampler of PAM. Thus, stochastic truncation K^* can be similarly computed following that of PAM.

To sample from FSBP, we iteratively sample the following parameters:

1. $u_i \sim \text{Unif}(0, \xi_{z_i})$,
2. stick-breaking weight π'_k for $k = 1, \ldots, K^*$,
3. the indicator z_i with $\Pr(z_i = k | \ldots) \propto 1_{\{u_i < \xi_k\}} \frac{\pi_k}{\xi_k} p(y_i | \boldsymbol{\phi}_k)$, and
4. the atom locations $\boldsymbol{\phi}_k | \cdots \propto \prod_{\{i; z_i = k, i = 1, \ldots, n\}} N(y_i | \boldsymbol{\phi}_k) p_H(\boldsymbol{\phi}_k)$.

To sample π'_k, we can use a MH step, where the full condition is

$$p(\pi'_k | \cdots) \propto \left[(p\pi'_k)^{m_k} (1 - p\pi'_k)^{\sum_{s=k+1}^{K^*} m_s} \right] f(\pi'_k)$$

$$\propto (\pi'_k)^{m_k + a - 1} (1 - p\pi'_k)^{\sum_{s=k+1}^{K^*} m_s} (1 - \pi'_k)^{b-1}.$$

Here, $m_k = \sum_{i=1}^{n} 1(z_i = k)$, and $f(\pi'_k)$ is the density function of the prior for π'_k as defined in Section 3.3. The same proposal density for β'_k (discussed in subsection A.1.11) can be used.

Lastly, if we place a Beta prior on p, then conditional on $\{\pi'_k\}$, p can be similarly sampled with another MH step and the same proposal. The other hyperparameter, γ, can also be

straight-forwardly sampled as in PAM (discussed in subsection A.1.11). The entire sampler is presented in Algorithm 2.

Algorithm 2 Slice Sampler for FSBP

1: **for** $m = 1, \cdots, M$ **do**

2: Sample each u_i from $u_i \sim \text{Unif}(0, \xi_{z_i})$ and find K^*.

3: Sample all π'_k for $k = 1, \cdots, K^*$ with MH step.

4: Sample p with MH step.

5: Sample z_i from the following full condition:

$$p(z_i = k| \cdots) \propto 1_{\{u_i < \xi_k\}} \frac{\pi_k}{\xi_k} p(y_i|\boldsymbol{\phi}_k)$$

6: Sample $\boldsymbol{\phi}_k$ from a conjugate NIG.

7: **end for**

A.1.13 Additional Distributions of Simulated Data in Section 6

Additional simulation data and results Table A.1.2 shows the cluster mean and weight in the simulation setup for Case 2 of Scenario 1.

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Mean		-4	-2	0	2	4
Weight	Group 1	0	0.8	0.2	0	0
	Group 2	0.3	0	0.1	0.6	0
	Group 3	0	0	0.2	0	0.8

Table A.1.2: Simulation truth of cluster means and weights for Case 2 of Scenario 1. Note that cluster 3 is the common cluster among all groups, while the other clusters are unique to their corresponding groups.

Figure A.1.5 shows the data distribution of one randomly selected sample, with a sample

size of 150, in Case 3 of Scenario 1. Note that the titles G1 to G6 refer to Groups 1 to 6, respectively.

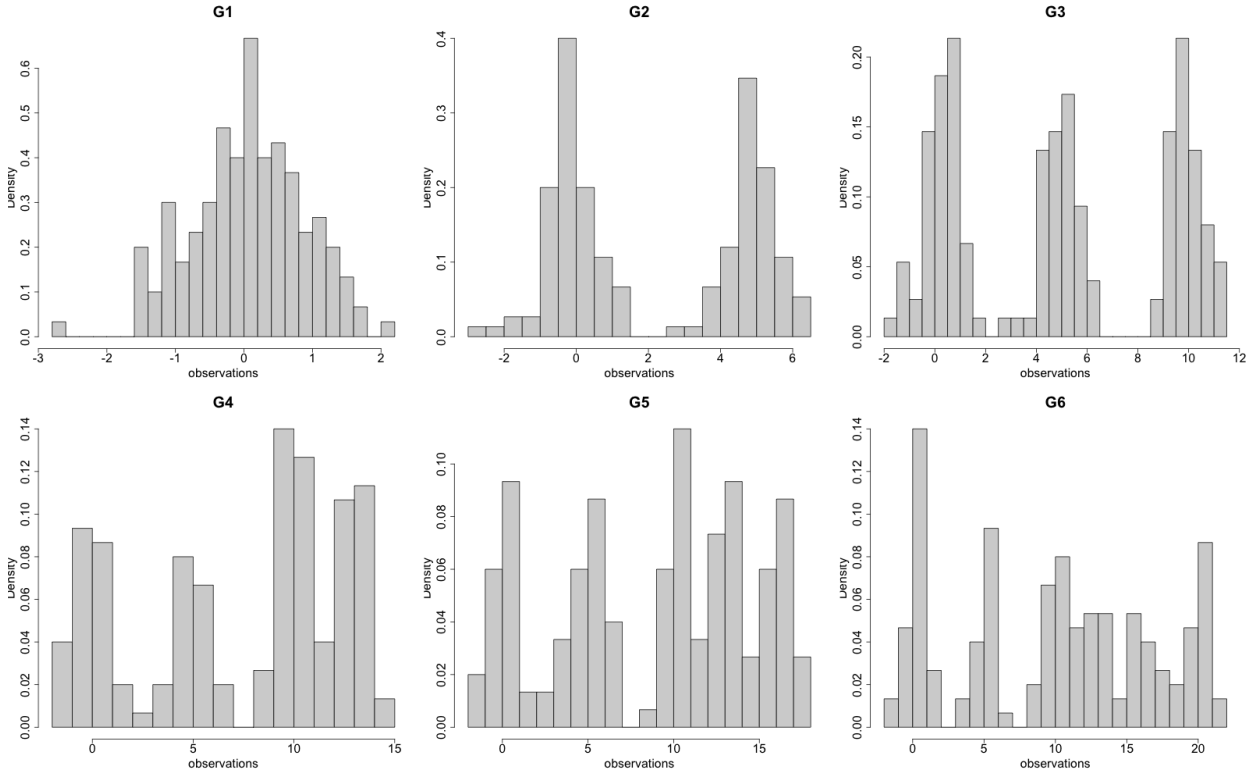

Figure A.1.5: Histogram of data distribution for randomly selected sample with sample size of 150 in Case 3 of Scenario 1. G1 to G6 means Groups 1 to 6, respectively.

Table A.1.3 shows the cluster means and weights in the simulation setup for the multivariate data in Scenario 2.

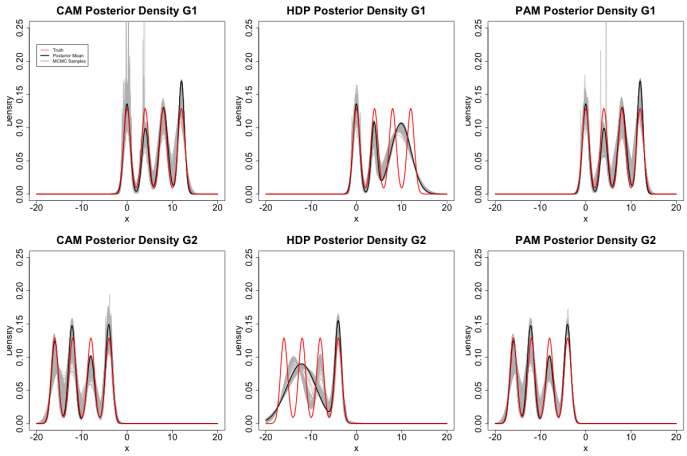

Figure A.1.6: Posterior density estimation using CAM (first column), HDP (second column), and PAM (third column) for a randomly selected dataset in Case 1 of Scenario 1. Each row corresponds to a specific group. The red lines represent the truth, the grey lines indicate the posterior density estimated in each MCMC iteration, and the black lines represent the point-estimate of the posterior density.

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Mean		$\begin{pmatrix} -6 \\ 4 \\ -6 \end{pmatrix}$	$\begin{pmatrix} -3 \\ 2 \\ -3 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 3 \\ -2 \\ -3 \end{pmatrix}$	$\begin{pmatrix} 6 \\ -4 \\ -6 \end{pmatrix}$
Weight	Group 1	0.2	0.2	0.2	0.2	0.2
	Group 2	0.3	0	0.5	0.2	0
	Group 3	0	0.6	0.4	0	0

Table A.1.3: Simulation truth of cluster means and weights for Scenario 2 in simulation. Here, cluster 3 is the common cluster shared among all groups.

Table A.1.6 shows the estiamted posterior density of a randomly selected sample for Case 1 in Scenario 1.

Table A.1.4 shows the performance of CAM, HDP, and PAM over 30 datasets for each sample sizes of Case 3 in Scenario 1.

Sample sizes	Metrics	CAM	HDP	PAM	Truth
$n_A = 50$	Number of clusters in all groups	4.03 (0.49)	3.93 (0.53)	4.93 (0.87)	6
	ARI	0.90 (0.05)	0.87 (0.05)	0.87 (0.07)	
	NFD	0.07 (0.03)	0.04 (0.02)	0.06 (0.02)	
$n_A = 100$	Number of clusters in all groups	4.67 (0.61)	4.00 (0.59)	5.67 (0.71)	6
	ARI	0.93 (0.04)	0.87 (0.05)	0.91 (0.04)	
	NFD	0.07 (0.02)	0.04 (0.02)	0.04 (0.02)	
$n_A = 150$	Number of clusters in all groups	4.97 (0.49)	4.27 (0.58)	5.97 (0.62)	6
	ARI	0.95 (0.02)	0.90 (0.04)	0.95 (0.03)	
	NFD	0.07 (0.02)	0.03 (0.01)	0.02 (0.01)	
$n_B = 10$	Number of clusters in all groups	4.17 (0.75)	4.23 (0.50)	5.77 (0.82)	6
	ARI	0.79 (0.08)	0.76 (0.08)	0.73 (0.06)	
	NFD	0.08 (0.02)	0.07 (0.03)	0.09 (0.02)	
$n_B = 20$	Number of clusters in all groups	4.40 (0.56)	4.30 (0.65)	6.00 (0.59)	6
	ARI	0.83 (0.08)	0.78 (0.09)	0.82 (0.06)	
	NFD	0.08 (0.02)	0.06 (0.03)	0.06 (0.02)	
$n_B = 40$	Number of clusters in all groups	5.43 (0.50)	4.33 (0.61)	6.17 (0.38)	6
	ARI	0.93 (0.05)	0.82 (0.07)	0.94 (0.05)	
	NFD	0.05 (0.02)	0.05 (0.02)	0.02 (0.01)	

Table A.1.4: Simulated univariate data in Case 3 of Scenario 1. Clustering performance for CAM, HDP, and PAM are evaluated according to the number of total estimated clusters (truth = 6 clusters), the Adjusted Rand Index (ARI), and the normalized Frobenius distance (NFD). Entries are Mean (SD) over 30 datasets.

Table A.1.5 shows the estimated number of clusters, common clusters, and unique clusters for the sample size of $n_A = 150$ in Case 3 of Scenario 1. For simplicity, except for all groups, the common clusters reported use Group 6 as a reference group, and measures the common clusters between Groups 1 to 5 with Group 6.

Metrics		CAM	HDP	PAM	Truth
Number of clusters	G1	1.00 (0.00)	1.00 (0.00)	1.03 (0.18)	1
	G2	2.00 (0.00)	2.00 (0.00)	2.00 (0.00)	2
	G3	3.30 (0.47)	3.00 (0.00)	3.03 (0.18)	3
	G4	4.00 (0.53)	3.33 (0.48)	3.93 (0.25)	4
	G5	4.43 (0.50)	3.33 (0.48)	4.13 (0.57)	5
	G6	4.60 (0.62)	3.17 (0.46)	4.53 (0.68)	6
Common clusters	All Groups	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1
	G6 and G5	4.43 (0.50)	3.13 (0.35)	3.47 (0.68)	5
	G6 and G4	3.70 (0.38)	3.00 (0.26)	3.10 (0.66)	4
	G6 and G3	3.17 (0.38)	2.67 (0.48)	2.90 (0.31)	3
	G6 and G2	2.00 (0.00)	2.00 (0.00)	1.96 (0.18)	2
	G6 and G1	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1
Uniqe clusters	G1	0.00 (0.00)	0.00 (0.00)	0.03 (0.18)	0
	G2	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0
	G3	0.00 (0.00)	0.00 (0.00)	0.03 (0.18)	0
	G4	0.17 (0.38)	0.00 (0.00)	0.63 (0.56)	0
	G5	0.00 (0.00)	0.00 (0.00)	0.53 (0.51)	0
	G6	0.13 (0.35)	0.03 (0.18)	0.90 (0.40)	1

Table A.1.5: The estimated number of clusters, common, and unique clusters for simulated univariate data in Case 3 of Scenario 1, when the sample size is $n_A = 150$. Note that except all groups, the estimated number of common clusters use Group 6 as a reference. Entries are Mean (SD) over 30 datasets.

Table A.1.6 shows the models' performance on the multivariate data in Scenario 2.

Sample sizes	Metrics	CAM	HDP	PAM	Truth
	Number of clusters in all groups	5.40 (1.13)	5.50 (0.97)	4.93 (0.64)	5
$n_j = 50$	ARI	0.90 (0.06)	0.86 (0.11)	0.89 (0.08)	
	NFD	0.05 (0.03)	0.05 (0.02)	0.03 (0.03)	
	Number of clusters in all groups	5.37 (0.71)	4.90 (0.76)	5.07 (0.26)	5
$n_j = 100$	ARI	0.95 (0.04)	0.91 (0.07)	0.96 (0.02)	
	NFD	0.04 (0.02)	0.04 (0.02)	0.01 (0.01)	
	Number of clusters in all groups	5.04 (0.19)	4.93 (0.47)	5.03 (0.18)	5
$n_j = 200$	ARI	0.97 (0.01)	0.96 (0.02)	0.97 (0.01)	
	NFD	0.03 (0.02)	0.01 (0.01)	0.01 (0.00)	

Table A.1.6: Simulated multivariate data in Scenario 2. Clustering performance for CAM, HDP, and PAM evaluated according to the number of total estimated clusters (truth = 5 clusters), the Adjusted Rand Index (ARI), and the normalized Frobenius distance (NFD). The entries are Mean (SD) over 30 datasets.

Table A.1.7 shows the performance of FSBP and DP on the univariate data in Scenario 3.

A.1.14 Additional Distributions and Results of Microbiome Population in Section 7.1

Figure A.1.8 shows the histogram of OTU counts for the four randomly selected individuals in the analysis of the microbiome dataset.

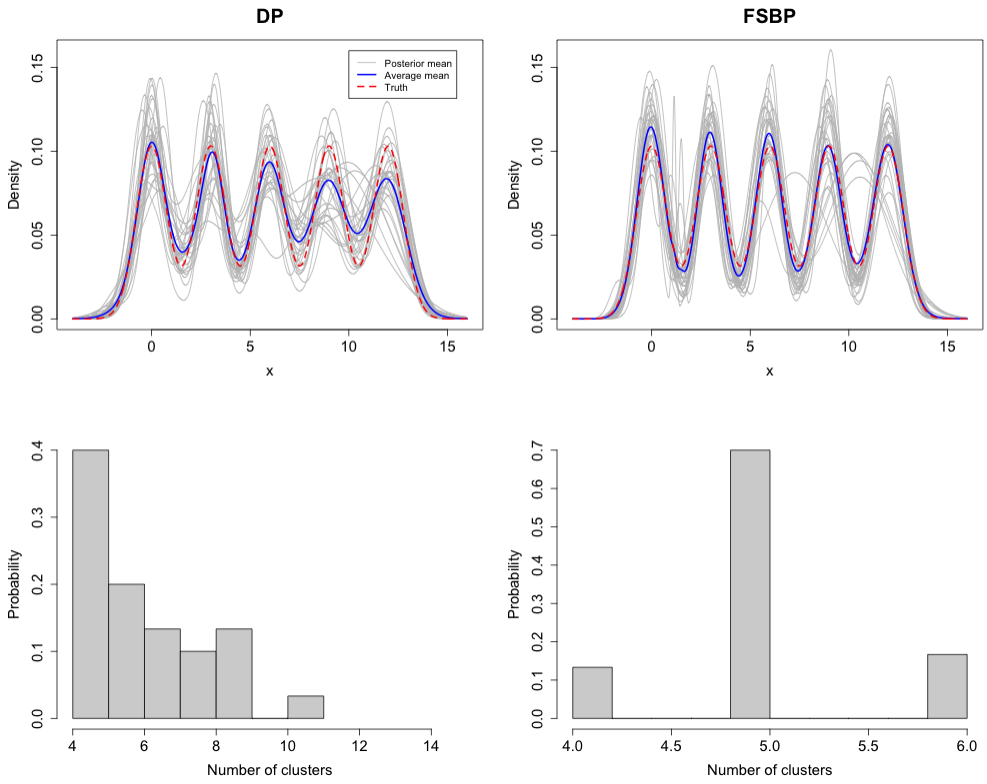

Figure A.1.7: Estimated posterior density for DP (top-left) and FSBP (top-right), along with histograms depicting the estimated number of clusters (bottom plots). FSBP estimates are based on Wade and Ghahramani [2018] using posterior samples. Grey lines represent the posterior mean for each simulated dataset, blue lines show the average of the posterior means across the 30 simulated datasets, and the red dashed lines indicate the truth.

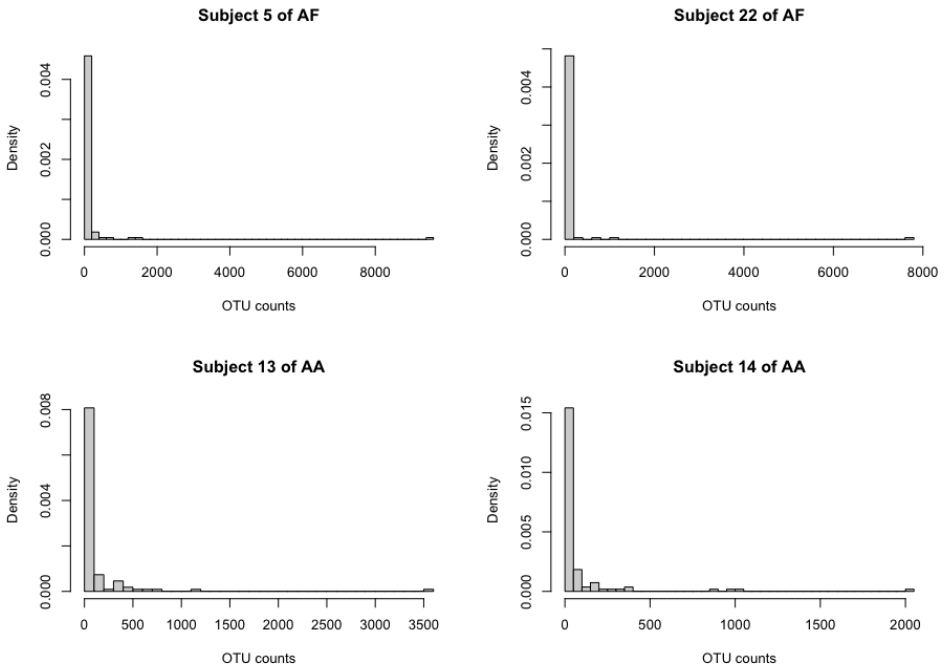

Figure A.1.8: Histograms of the microbiome population of the four selected individuals.

Figure A.1.9 shows barplots of the taxa counts (TC) of OTUs grouped by eight estiamted clusters as well as by both cluster and individuals.

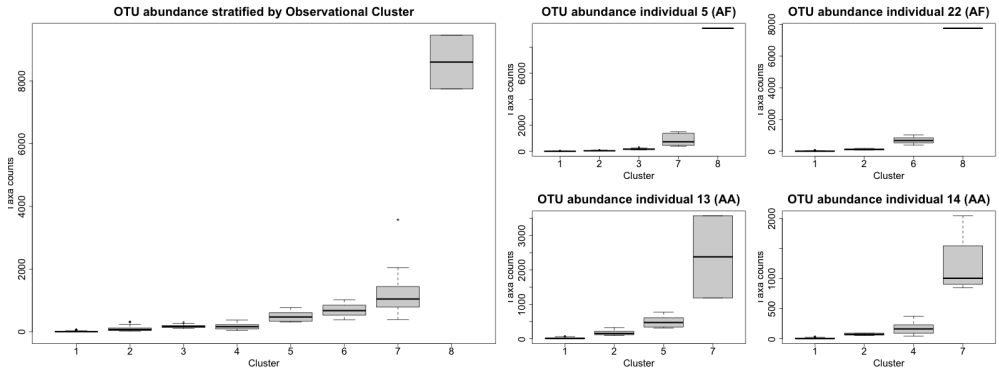

Figure A.1.9: Boxplots of microbiome abundance counts stratified by clusters (Left subplot) and by both clusters and individuals (Four right subplots).

A.1.15 Additional Results of Warts Dataset Analysis in Section 7.2

Figure A.1.10 below shows the cluster membership of each patient of the warts dataset.

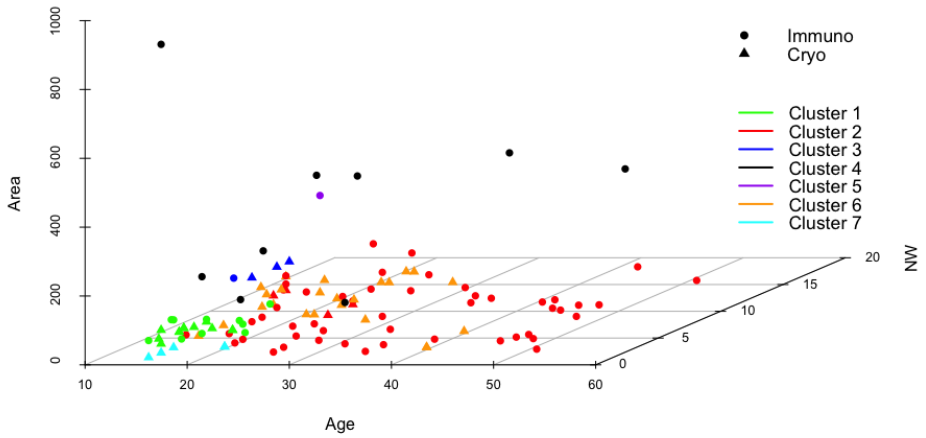

Figure A.1.10: Estimated cluster membership of patients in the warts dataset. The cluster labels are shown with different colors, across two groups indicated by the circles and triangles. The clustering result is based on four covariates of area, age, number of warts, and time elaspsed until treatment. We plot three of them: Area, Age, and number of warts (NW).

A.2 Appendix to "PAM-HC: A Bayesian Nonparametric Construction of Hybrid Control for Randomized Clinical Trials Using External Data"

A.2.1 Additional Simulation Results

Figures A.2.1 and A.2.2 are the density plots of covariates for the randomly selected example dataset from Scenarios 2 and 3, respectively. Each row represents a cluster, each column represents a dimension of the multivariate covariate, and each color represent a different treatment group (treatment arm, control arm, or the external data).

Tables A.2.1 and A.2.2 show the cluster-specific treatment effects for each scenario for the continuous and binary outcomes, respectively.

Table A.2.3 shows the estimated number of clusters, the ARI, and the NFD by PAM for each scenario in the simulation study.

Tables A.2.4 and A.2.5 show the estimated overall treatment effects with the baseline model, PSCL, and PAM-HC for each simulation scenario for the continuous and binary outcomes, respectively.

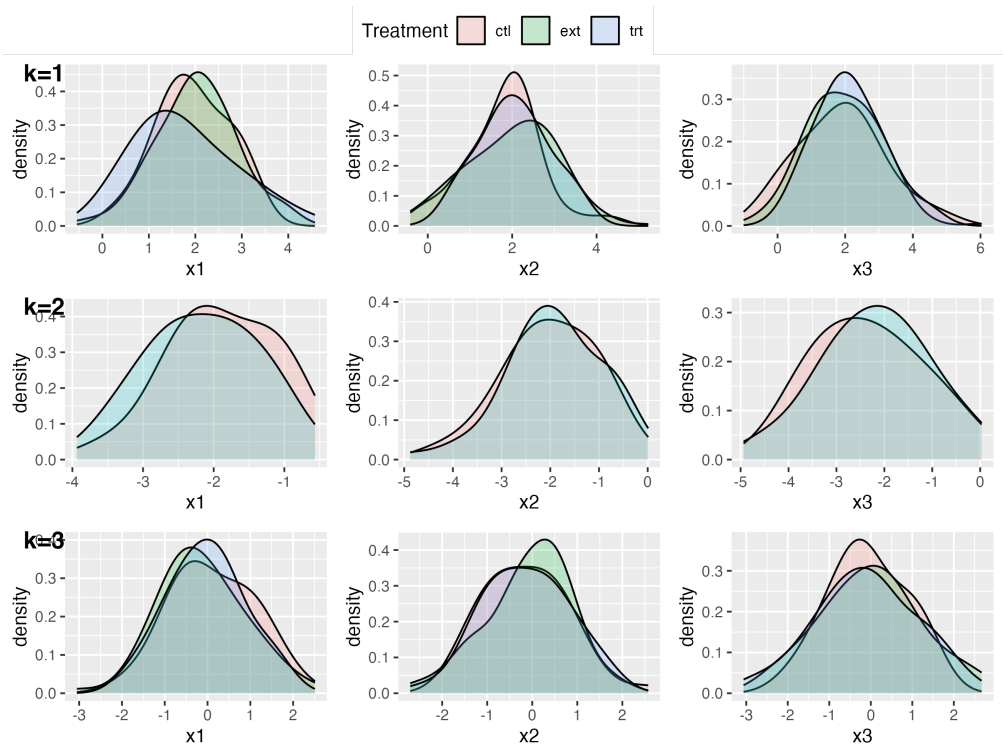

Figure A.2.1: The covariate density plots of one simulated data in Scenario 2. The rows represent three clusters estimated by PAM-HC.

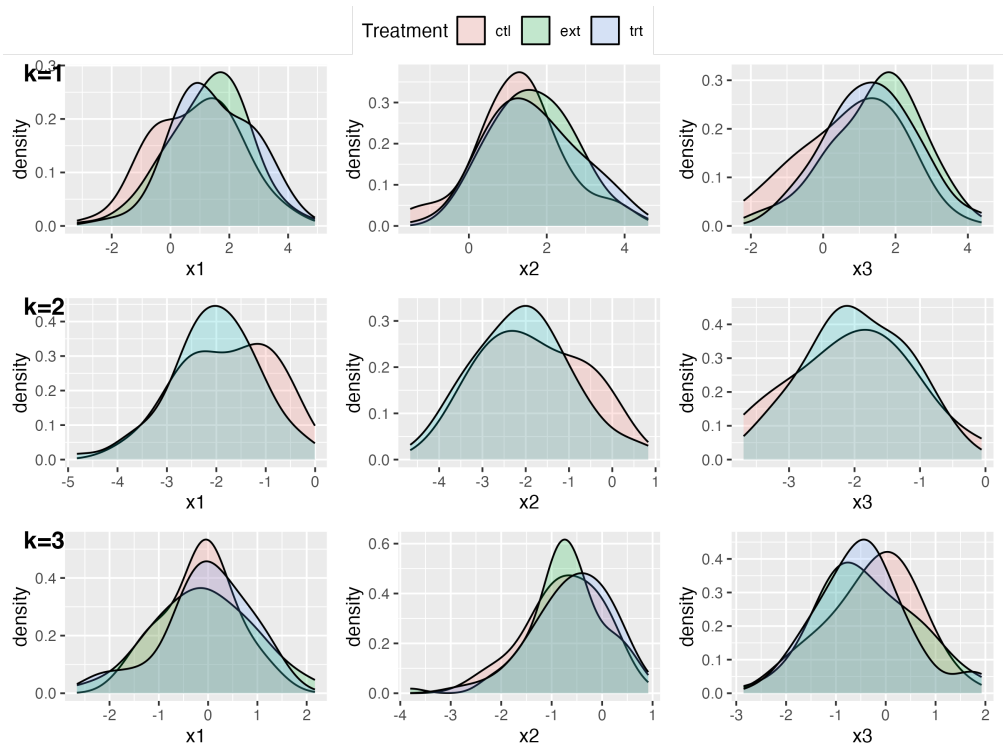

Figure A.2.2: The covariate density plots of one simulated data in Scenario 3. The rows represent three clusters estimated by PAM-HC.

Sc	Cluster specific treatment effect			Overall
	Cluster 1	Cluster 2	Cluster 3	treatment effect
Sc 1	$0.18_{0.18}$ (0.03)	$-0.09_{0.21}$ (0.01)	$0.23_{0.18}$ (0.26)	$0.12_{0.11}$ (0)
	$1.27_{0.18}$ (1.08)	$-0.02_{0.22}$ (0.05)	$2.22_{0.18}$ (2.24)	$1.22_{0.12}$ (1)
	$1.99_{0.18}$ (1.97)	$0.81_{0.21}$ (0.86)	$2.62_{0.19}$ (2.60)	$1.87_{0.12}$ (2)
	$3.14_{0.17}$ (3.14)	$2.04_{0.21}$ (2.02)	$4.17_{0.19}$ (4.13)	$3.17_{0.12}$ (3)
Sc 2	$-0.10_{0.17}$ (-0.01)	$-0.16_{0.23}$ (-0.19)	$-0.14_{0.23}$ (-0.14)	$-0.13_{0.11}$ (0)
	$0.95_{0.18}$ (0.93)	$-0.08_{0.24}$ (-0.22)	$1.87_{0.22}$ (1.82)	$0.76_{0.13}$ (1)
	$1.99_{0.20}$ (1.83)	$1.33_{0.21}$ (1.28)	$3.16_{0.21}$ (3.10)	$2.01_{0.13}$ (2)
	$3.30_{0.19}$ (3.24)	$2.04_{0.21}$ (1.91)	$4.04_{0.22}$ (3.94)	$2.98_{0.12}$ (3)
Sc 3	$-0.11_{0.23}$ (-0.12)	$-0.18_{0.23}$ (-0.19)	$-0.11_{0.17}$ (0.03)	$-0.13_{0.11}$ (0)
	$0.74_{0.24}$ (0.82)	$-0.20_{0.23}$ (-0.22)	$1.52_{0.20}$ (1.99)	$0.76_{0.13}$ (1)
	$1.76_{0.26}$ (1.92)	$1.32_{0.22}$ (1.28)	$2.71_{0.21}$ (3.06)	$2.02_{0.12}$ (2)
	$2.92_{0.25}$ (3.12)	$1.95_{0.21}$ (1.91)	$3.85_{0.20}$ (4.08)	$2.97_{0.13}$ (3)

Table A.2.1: Estimated cluster-specific treatment effects for selected examples with different values of true Δ in each of the three scenarios using continuous outcome in the simulation. The entries for the three columns, Cluster 1, Cluster 2, and Cluster 3, are posterior mean_{SD} (observed cluster-specific treatment effects). The entries for the last column are posterior mean_{SD} (truth).

Sc	Cluster specific treatment effect			Overall
	Cluster 1	Cluster 2	Cluster 3	treatment effect
Sc 1	$0.06_{0.09}$ (0.07)	$-0.01_{0.04}$ (0.00)	$0.00_{0.03}$ (0.00)	$0.02_{0.04}$ (0.00)
	$0.25_{0.09}$ (0.23)	$0.04_{0.04}$ (0.03)	$0.01_{0.03}$ (0.01)	$0.11_{0.04}$ (0.09)
	$0.33_{0.08}$ (0.35)	$0.04_{0.05}$ (0.05)	$0.01_{0.02}$ (0.00)	$0.14_{0.03}$ (0.16)
	$0.48_{0.07}$ (0.50)	$0.18_{0.07}$ (0.12)	$0.02_{0.03}$ (0.01)	$0.24_{0.04}$ (0.20)
Sc 2	$0.04_{0.09}$ (0.01)	$-0.01_{0.03}$ (0.00)	$0.02_{0.03}$ (0.01)	$0.02_{0.03}$ (0.00)
	$0.22_{0.09}$ (0.19)	$0.00_{0.06}$ (0.01)	$0.01_{0.02}$ (0.02)	$0.08_{0.04}$ (0.09)
	$0.36_{0.08}$ (0.33)	$0.09_{0.05}$ (0.08)	$0.02_{0.03}$ (0.01)	$0.16_{0.04}$ (0.16)
	$0.41_{0.08}$ (0.39)	$0.14_{0.07}$ (0.12)	$0.01_{0.02}$ (0.02)	$0.19_{0.04}$ (0.20)
Sc 3	$-0.09_{0.11}$ (-0.10)	$-0.03_{0.04}$ (-0.04)	$0.08_{0.08}$ (0.01)	$0.00_{0.04}$ (0.00)
	$0.15_{0.12}$ (0.10)	$0.00_{0.03}$ (0.00)	$0.12_{0.08}$ (0.03)	$0.08_{0.04}$ (0.09)
	$0.36_{0.11}$ (0.27)	$-0.01_{0.05}$ (-0.01)	$0.16_{0.08}$ (0.01)	$0.14_{0.04}$ (0.16)
	$0.47_{0.10}$ (0.36)	$0.09_{0.05}$ (0.03)	$0.20_{0.07}$ (0.10)	$0.22_{0.04}$ (0.20)

Table A.2.2: Estimated cluster-specific treatment effects for selected examples with different values of true Δ in each of the three scenarios using binary outcome in the simulation. The entries for the three columns, Cluster 1, Cluster 2, and Cluster 3, are posterior mean_{SD} (observed cluster-specific treatment effects). The entries for the last column are posterior mean_{SD} (truth).

N	Sc	Clusters	ARI	NFD
300	Sc 1	$3.91_{0.46}$	$0.74_{0.13}$	$0.09_{0.06}$
	Sc 2	$4.01_{0.46}$	$0.77_{0.16}$	$0.08_{0.05}$
	Sc 3	$3.03_{0.36}$	$0.75_{0.16}$	$0.09_{0.07}$
450	Sc 1	$4.04_{0.20}$	$0.81_{0.03}$	$0.06_{0.01}$
	Sc 2	$4.02_{0.20}$	$0.84_{0.07}$	$0.05_{0.02}$
	Sc 3	$3.09_{0.32}$	$0.83_{0.05}$	$0.06_{0.02}$

Table A.2.3: Clustering performance for PAM-HC evaluated according to the number of total detected clusters (truth = 4 clusters for Sc 1 and Sc 2; 3 clusters for Sc 3) based on the estimated optimal clustering, the adjusted Rand index (ARI), and the normalized Frobenius distance (NFD). The entries are mean_{SD} over 100 datasets.

N	Sc	Method	True $\Delta = 0$			True $\Delta = 1$			True $\Delta = 2$			True $\Delta = 3$		
			$\hat{\Delta}$	SD	MSE	$\hat{\Delta}$	SD	MSE	$\hat{\Delta}$	SD	MSE	$\hat{\Delta}$	SD	MSE
300	Sc 1	Baseline	-0.08	0.58	0.35	0.92	0.56	0.32	1.92	0.58	0.35	2.92	0.56	0.32
		PSCL1	0.01	0.20	0.04	1.02	0.17	0.03	2.01	0.19	0.04	3.02	0.17	0.03
		PSCL2	0.44	0.45	0.40	1.46	0.44	0.40	2.44	0.44	0.40	3.45	0.43	0.39
		PSCL3	0.08	0.33	0.12	1.09	0.31	0.11	2.08	0.33	0.12	3.09	0.31	0.11
		PAM-HC	0.03	0.31	0.10	1.04	0.30	0.10	2.03	0.31	0.10	3.03	0.31	0.10
	Sc 2	Baseline	-0.09	0.58	0.34	0.92	0.57	0.34	1.91	0.58	0.34	2.92	0.57	0.34
		PSCL1	-0.04	0.36	0.13	0.97	0.34	0.12	1.96	0.36	0.13	2.97	0.34	0.12
		PSCL2	-0.53	0.31	0.38	0.47	0.30	0.37	1.46	0.30	0.38	2.46	0.31	0.38
		PSCL3	-0.45	0.35	0.33	0.55	0.35	0.32	1.56	0.35	0.33	2.55	0.35	0.32
		PAM-HC	-0.08	0.28	0.09	0.94	0.30	0.09	1.92	0.28	0.09	2.93	0.30	0.09
	Sc 3	Baseline	-0.09	0.58	0.34	0.91	0.56	0.32	1.91	0.58	0.34	2.91	0.56	0.32
		PSCL1	-0.09	0.16	0.03	0.92	0.16	0.03	1.91	0.16	0.03	2.92	0.16	0.03
		PSCL2	-0.49	0.31	0.33	0.54	0.29	0.30	1.51	0.32	0.34	2.53	0.29	0.31
		PSCL3	-0.15	0.23	0.08	0.86	0.22	0.07	1.85	0.23	0.08	2.86	0.22	0.07
		PAM-HC	-0.03	0.22	0.05	0.98	0.22	0.05	1.97	0.22	0.05	2.98	0.22	0.05
450	Sc 1	Baseline	-0.09	0.45	0.21	0.90	0.48	0.24	1.89	0.47	0.23	2.89	0.47	0.24
		PSCL1	0.05	0.12	0.02	1.03	0.14	0.02	2.03	0.14	0.03	3.02	0.13	0.02
		PSCL2	0.48	0.33	0.34	1.46	0.35	0.34	2.46	0.34	0.32	3.46	0.35	0.33
		PSCL3	0.14	0.22	0.07	1.13	0.23	0.07	2.12	0.22	0.06	3.12	0.23	0.07
		PAM-HC	0.02	0.14	0.02	1.00	0.16	0.02	2.00	0.14	0.02	2.99	0.15	0.02
	Sc 2	Baseline	-0.08	0.46	0.22	0.91	0.48	0.24	1.90	0.48	0.24	2.90	0.48	0.24
		PSCL1	-0.04	0.24	0.06	0.95	0.25	0.07	1.94	0.25	0.07	2.94	0.26	0.07
		PSCL2	-0.46	0.28	0.29	0.53	0.28	0.30	1.52	0.30	0.32	2.52	0.30	0.32
		PSCL3	-0.40	0.28	0.24	0.58	0.30	0.26	1.58	0.30	0.27	2.57	0.30	0.27
		PAM-HC	-0.00	0.17	0.03	0.99	0.18	0.03	1.97	0.20	0.04	2.97	0.19	0.04
	Sc 3	Baseline	-0.09	0.49	0.25	0.90	0.50	0.27	1.88	0.50	0.27	2.89	0.51	0.27
		PSCL1	-0.08	0.11	0.02	0.90	0.12	0.02	1.90	0.13	0.03	2.90	0.12	0.03
		PSCL2	-0.44	0.28	0.25	0.55	0.28	0.26	1.54	0.31	0.27	2.54	0.27	0.27
		PSCL3	-0.16	0.18	0.06	0.82	0.19	0.07	1.82	0.20	0.07	2.81	0.23	0.07
		PAM-HC	0.02	0.24	0.07	1.00	0.23	0.07	1.99	0.24	0.08	2.99	0.23	0.06

Table A.2.4: Simulation results based on continuous outcome for PAM-HC, baseline method, and three versions of PSCL methods. Here $\hat{\Delta}$ is the average posterior mean of the overall treatment effect across 100 simulated trials.

N	Sc	Method	True $\Delta = 0.00$*			True $\Delta = 9.42$*			True $\Delta = 15.67$*			True $\Delta = 19.52$*		
			$\hat{\Delta}$*	SD	MSE*	$\hat{\Delta}$*	SD	MSE*	$\hat{\Delta}$*	SD	MSE*	$\hat{\Delta}$*	SD	MSE*
300	Sc 1	Baseline	-0.37	0.06	0.35	5.69	0.06	0.50	12.60	0.06	0.50	17.73	0.06	0.38
		PSCL1	-0.24	0.02	0.05	6.18	0.03	0.20	12.77	0.03	0.16	18.25	0.03	0.11
		PSCL2	3.15	0.04	0.26	9.57	0.05	0.22	16.13	0.04	0.19	21.57	0.04	0.23
		PSCL3	0.41	0.03	0.11	6.90	0.04	0.20	13.48	0.03	0.17	19.02	0.03	0.12
		PAM-HC	0.29	0.03	0.11	6.65	0.04	0.23	13.14	0.04	0.22	18.52	0.04	0.16
	Sc 2	Baseline	-0.34	0.06	0.37	5.85	0.06	0.49	12.72	0.06	0.48	18.04	0.06	0.39
		PSCL1	-4.56	0.04	0.37	1.76	0.04	0.76	8.40	0.04	0.73	13.94	0.04	0.48
		PSCL2	-4.54	0.04	0.32	1.78	0.03	0.68	8.50	0.03	0.64	13.84	0.03	0.42
		PSCL3	-6.08	0.04	0.56	0.25	0.04	1.01	6.99	0.04	0.97	12.42	0.04	0.70
		PAM-HC	-0.61	0.03	0.12	5.67	0.03	0.26	12.25	0.03	0.26	17.69	0.03	0.16
	Sc 3	Baseline	-0.48	0.06	0.34	6.73	0.06	0.51	12.50	0.06	0.46	17.78	0.06	0.39
		PSCL1	-0.34	0.02	0.06	6.17	0.03	0.18	12.68	0.03	0.16	18.23	0.03	0.09
		PSCL2	-4.01	0.03	0.25	2.60	0.03	0.58	9.01	0.03	0.55	14.65	0.03	0.33
		PSCL3	-0.50	0.02	0.06	6.08	0.03	0.18	12.51	0.03	0.17	18.14	0.03	0.08
		PAM-HC	-0.30	0.03	0.08	6.08	0.03	0.20	12.49	0.03	0.20	17.95	0.03	0.13
450	Sc 1	Baseline	-0.51	0.05	0.23	5.93	0.05	0.37	12.68	0.05	0.31	17.89	0.05	0.28
		PSCL1	-0.01	0.02	0.06	6.52	0.02	0.14	13.22	0.02	0.10	18.61	0.02	0.07
		PSCL2	3.36	0.04	0.25	9.92	0.04	0.14	16.51	0.03	0.12	21.96	0.03	0.17
		PSCL3	0.75	0.03	0.09	7.35	0.03	0.14	14.02	0.02	0.09	19.42	0.03	0.07
		PAM-HC	0.32	0.03	0.08	6.86	0.02	0.13	13.41	0.02	0.10	18.73	0.03	0.07
	Sc 2	Baseline	-0.57	0.05	0.24	6.04	0.05	0.37	12.77	0.05	0.31	17.95	0.05	0.28
		PSCL1	-4.33	0.04	0.31	2.26	0.03	0.62	9.06	0.03	0.56	14.23	0.04	0.41
		PSCL2	-4.30	0.03	0.28	2.42	0.03	0.59	9.17	0.03	0.53	14.36	0.03	0.35
		PSCL3	-6.20	0.04	0.52	0.49	0.04	0.92	7.36	0.04	0.85	12.39	0.04	0.64
		PAM-HC	-0.22	0.03	0.08	6.40	0.03	0.17	13.08	0.03	0.14	18.18	0.03	0.10
	Sc 3	Baseline	-0.40	0.05	0.26	5.70	0.05	0.41	12.58	0.05	0.35	17.85	0.05	0.31
		PSCL1	-0.23	0.02	0.06	6.18	0.02	0.15	12.87	0.02	0.12	18.32	0.02	0.06
		PSCL2	-3.73	0.03	0.23	2.65	0.03	0.54	9.33	0.02	0.47	14.82	0.03	0.29
		PSCL3	-0.43	0.02	0.05	5.96	0.02	0.18	12.64	0.02	0.14	18.06	0.02	0.08
		PAM-HC	0.16	0.03	0.10	6.49	0.03	0.20	13.18	0.02	0.15	18.49	0.03	0.11

* True Δ, $\hat{\Delta}$, and MSE values times 100.

Table A.2.5: Simulation results based on binary outcome for PAM-HC, baseline method, and three versions of PSCL methods. Here $\hat{\Delta}$ is the average posterior mean of the overall treatment effect across 100 simulated trials.

A.3 Appendix to "A Bayesian Estimator of Sample Size"

A.3.1 Posterior Probability of H_1 for One-arm Trial

In this subsection, we show for binary, continuous with known variance, and count-data outcomes in one-arm trial, the posterior probability of H_1 is a function of evidence e as defined in equation (4.5) and sample size n, i.e., $\Pr(H = H_1|\boldsymbol{y}_n) = \Pr(H = H_1|e, n)$.

Assume θ_0 is a known constant, from Table (4.1) and equation (4.5), we have the posterior $f(\theta_1|\boldsymbol{y}_n)$ in equation (4.3) for the three outcomes as follows:

- Binary:

$$f(\theta_1|\boldsymbol{y}_n) = \frac{\theta_1^{a+n\bar{y}}(1-\theta_1)^{b+n(1-\bar{y})}}{B(a+n\bar{y}, b+n(1-\bar{y}))} = \frac{\theta_1^{a+n[e+\theta_0]}(1-\theta_1)^{b+n(1-[e+\theta_0])}}{B(a+n[e+\theta_0], b+n(1-[e+\theta_0]))} = f(\theta_1|e,n),$$

  where $B(\cdot,\cdot)$ is the beta function;

- Continuous with known σ^2:

$$f(\theta_1|\boldsymbol{y}_n) = \frac{1}{\sqrt{\frac{2\pi}{\frac{1}{b}+\frac{n}{\sigma^2}}}} \exp\left\{-\frac{1}{2}\left[\frac{1}{b}+\frac{n}{\sigma^2}\right]\left(\theta_1 - \frac{1}{\frac{1}{b}+\frac{n}{\sigma^2}}\left[\frac{a}{b}+\frac{n\bar{y}}{\sigma^2}\right]\right)^2\right\}$$

$$= \frac{1}{\sqrt{\frac{2\pi}{\frac{1}{b}+\frac{n}{\sigma^2}}}} \exp\left\{-\frac{1}{2}\left[\frac{1}{b}+\frac{n}{\sigma^2}\right]\left(\theta_1 - \frac{1}{\frac{1}{b}+\frac{n}{\sigma^2}}\left[\frac{a}{b}+\frac{n[e+\theta_0]}{\sigma^2}\right]\right)^2\right\} = f(\theta_1|e,n),$$

- Count-data:

$$f(\theta_1|\boldsymbol{y}_n) = \frac{(b+n)^{a+n\bar{y}}}{\Gamma(a+n\bar{y})}\theta_1^{a+n\bar{y}} \exp\{-(b+n)\theta_1\}$$

$$= \frac{(b+n)^{a+n[e+\theta_0]}}{\Gamma(a+n[e+\theta_0])}\theta_1^{a+n[e+\theta_0]} \exp\{-(b+n)\theta_1\} = f(\theta_1|e,n),$$

  where $\Gamma(\cdot)$ is the gamma function.

Hence, for all three outcomes in one-arm trial, we have $f(\theta_1|\boldsymbol{y}_n) = f(\theta_1|e,n)$. As a result, when both θ_0 and θ^* are given as fixed constants, we see that the integration term in equation (4.3) becomes a function of e and n, i.e.,

$$g(e,n) = \int_{\theta_1-\theta_0>\theta^*} f(\theta_1|e,n)d\theta_1.$$

Consequently, equation (4.3) is a function of e and n:

$$\Pr(H = H_1|\boldsymbol{y}_n) = \Pr(H = H_1|e,n).$$

Note that e here is the sufficient statistic for $\boldsymbol{y}_n$ for the three outcome types specified in this work.

A.3.2 Posterior Probability of H_1 for Two-arm Trial with Continuous outcome and known variance

Let $\boldsymbol{y}_n^d = \{y_i^d; i = 1, \ldots, n\}$ where $y_i^d = y_{i1} - y_{i0}$. Since $y_{ij}|\theta_j \sim N(\theta_j, \sigma^2)$, $j = 0, 1$ in model (4.2) for continuous outcome with known variance σ^2, which is common between $j = 0$ and $j = 1$, the likelihood of y_i^d follows a normal distribution index by the parameters θ and $2\sigma^2$:

$$y_i^d|\theta \sim N(\theta, 2\sigma^2),\ i = 1, \ldots, n$$

Let $\pi = N(a, b)$ for θ, we have:

$$\theta|H = H_j \sim N(a, b)I(\theta \in H_j),\ j = 0, 1,$$

and we have $\Pr(H = H_1) = q$ and $\Pr(H = H_0) = 1 - q$ as in model (4.2).

The posterior probability of H_1 is again given by equation (4.3), which we have

$$f(\boldsymbol{\theta}|\boldsymbol{y}_n) = f(\theta|\boldsymbol{y}_n^d) \propto f(\boldsymbol{y}_n^d|\theta)\pi(\theta).$$

Since the variance is known, and both $f(.)$ and $\pi(.)$ follows normal distribution, we have $f(\theta|\boldsymbol{y}_n^d)$ follows a normal distribution and is similar to the continuous with known variance outcome in one-arm trial. Let $\bar{y}^d = \frac{1}{n}\sum_{i=1}^n y_i^d = \bar{y}_1 - \bar{y}_2$, we have

$$f(\theta|\boldsymbol{y}_n^d) = \frac{1}{\sqrt{\frac{2\pi}{\frac{1}{b}+\frac{n}{\sigma^2}}}} \exp\left\{-\frac{1}{2}\left[\frac{1}{b} + \frac{n}{\sigma^2}\right]\left(\theta_1 - \frac{1}{\frac{1}{b}+\frac{n}{\sigma^2}}\left[\frac{a}{b} + \frac{n\bar{y}^d}{\sigma^2}\right]\right)^2\right\}$$

$$= \frac{1}{\sqrt{\frac{2\pi}{\frac{1}{b}+\frac{n}{\sigma^2}}}} \exp\left\{-\frac{1}{2}\left[\frac{1}{b} + \frac{n}{\sigma^2}\right]\left(\theta_1 - \frac{1}{\frac{1}{b}+\frac{n}{\sigma^2}}\left[\frac{a}{b} + \frac{ne}{\sigma^2}\right]\right)^2\right\} = f(\theta_1|e, n),$$

where by equation (4.5), $\bar{y}^d = e$. With an argument similar to section A.3.1, we have

$$\Pr(H = H_1|\boldsymbol{y}_n) = \Pr(H = H_1|e, n)$$

for the continuous with known variance outcome for two-arm trial.

A.3.3 Posterior Probability of H_1 for Two-arm Trial with Binary and Count-data Outcomes

From section A.3.1, we see that for θ_j, $j = 0, 1$, we have $f(\theta_j|\boldsymbol{y}_n) = f(\theta_j|\bar{y}_j, n)$ for both the binary and count-data outcomes. Hence, we have

$$f(\boldsymbol{\theta}|\boldsymbol{y}_n) = f(\theta_0|\boldsymbol{y}_{0n})f(\theta_1|\boldsymbol{y}_{1n}) = f(\theta_0|\bar{y}_0, n)f(\theta_1|\bar{y}_1, n) = f(\boldsymbol{\theta}|\bar{y}_1, \bar{y}_0, n).$$

A.3.4 BESS Algorithm 2'

Below is BESS Algorithm 2', where $\bar{y}_1$ and $\bar{y}_0$ are provided instead of evidence e.

BESS Algorithm 2' Two-Arm Trials; Binary or Count Data

Input: The hierarchical models in Table 4.1.
Input: Clinically meaningful effect size θ^*, response parameters for the treatment and control $\bar{y}_1$ and $\bar{y}_0$, confidence c, prior probability q.
Set $n_{\min}$ and $n_{\max}$ ($n_{\min} < n_{\max}$) the smallest and largest candidate sample sizes.
Set $n = n_{\min}$.
while $n \leq n_{\max}$ **do**
    Compute equation (4.3).
    if condition (4.7) is true **then**
        Stop and return the sample size n.
    else
        n = n+1
    end if
end while
if $n > n_{\max}$ **then**
    Return sample size is larger than $n_{\max}$.
end if

A.3.5 Simulation Parameters for Coherence in Section 4.4.2

Table A.3.1 shows the true parameters for data simulation in section 4.4.2 to demonstrate coherence between BESS and Bayesian Inference.

Trial	Outcome	θ^*	σ^2	e	c	n	a	b	True θ_1	True θ_0
Two-arm	Binary	0.05	-	0.1	0.8	150	0.5	0.5	0.3	0.2
	Continuous	0.1	0.5	0.2	0.8	71	0	10	0.8	0.6
	Count-data	0.3	-	0.4	0.8	100	1	2	0.7	0.3

Table A.3.1: Parameters for BESS, estimated sample size, and simulation truth for two-arm trial with all three outcome types. "-" means not applicable.

A.3.6 Additional Simulation Setup and Results in Section 4.5.1

Additional simulation details and results Figure A.3.1 shows a flowchart for the simulation process in Section 4.5.1.

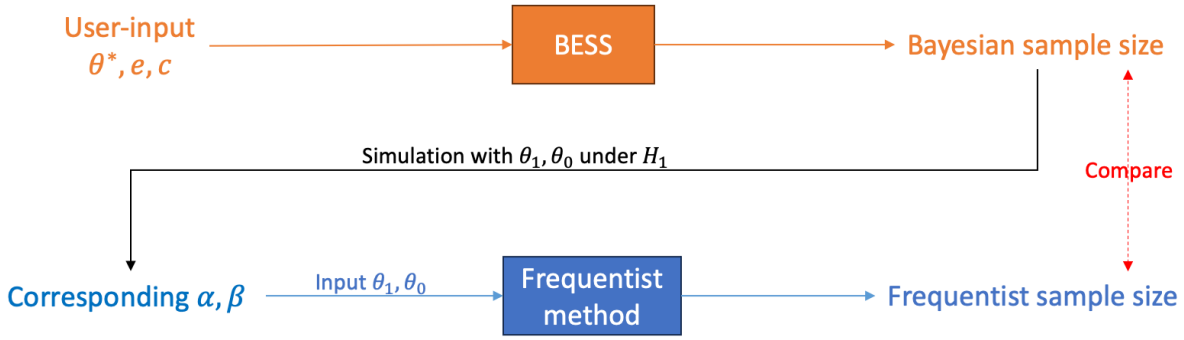

Figure A.3.1: Flowchart of simulation process to compare sample sizes estimated under BESS and that of the Frequenitst method.

Table A.3.2 show the results when θ_0 and θ_1 are mis-specified in the standard SSE.

Evidence	Confidence	Planned				Standard SSE	Simulation			
e	c	θ_1	θ_0	α	$1-\beta$	sample size n	α	$1-\beta$	FPR	FNR
	0.7	0.35	0.25	0.31	0.76	240	0.31	0.96	0.25	0.05
0.1	0.8	0.35	0.25	0.20	0.84	560	0.20	1.00	0.17	0.00
	0.9	0.35	0.25	0.10	0.94	1136	0.10	1.00	0.09	0.00
	0.7	0.45	0.25	0.44	0.57	3	0.37	0.49	0.43	0.45
0.2	0.8	0.45	0.25	0.24	0.46	8	0.26	0.44	0.37	0.42
	0.9	0.45	0.25	0.11	0.38	17	0.14	0.30	0.31	0.44

Table A.3.2: Standard SSE sample size and simulated Type I error rate, power, false positive rate, and false negative rate when $\theta_1 - \theta_0$ is mis-specified by the Frequentist method to match e in BESS for two-arm trial with binary outcome.

Additional simulation details in sensitivity of prior In the previous simulation, we assume a flat prior Beta$(0, 0)$ for BESS. In particular, this prior is used to find the sample size as well as to compute $\Pr(H = H_1|\boldsymbol{y}_n)$ for each simulated trial. In Morita et al. [2008], the authors show that the prior effective sample size of Beta(a, b) for a binomial likelihood

is quantified as $(a+b)$. Therefore, Beta$(0,0)$ is considered noninformative in that it includes no prior information for the posterior inference.

However, BESS can accommodate prior information, if available, as part of sample size estimation. Assuming there exist n_0 patients per arm as prior data, we demonstrate the sensitivity of incorporating these prior information through simulation. First consider the simulation process for a single trial with the following two steps: 1) generating the prior data, 2) assume evidence $e = 0.15$ and $c = 0.8$, estimate the sample size via BESS using the informative priors constructed from the prior data. For step 1, assuming there are $n_0 = 10$ external patients data per arm that is available to be incorporated, we generate these prior data similar to the simulation process in section 4.5.1 with $\theta_1 = 0.4$ and $\theta_0 = 0.25$. Denote the generated data $\boldsymbol{y}_j^0 = \{y_{ij}^0; i = 1, \ldots, n_0\}$, $j = 1, 0$ as the binary outcomes of these 10 patients' external data for the treatment and the control arms, respectively. We consider an informative prior for BESS as

$$(\theta_1, \theta_0)|H = H_j \sim \prod_{j=1}^{2} \text{Beta}\left(\sum_{i=1}^{n_0} y_{ij}^0, n_0 - \sum_{i=1}^{n_0} y_{ij}^0\right) I\{(\theta_1, \theta_0) \in H_j\}.$$

We find sample size via BESS's algorithm 2 with the informative prior, e, and c. Following this process, we simulate 1,000 trials and compute the average sample size to compare to BESS with vague prior.

A.3.7 Metrics Used in Section 4.6.2

The metrics include: 1) Type I error rate, 2) Type II error rate, 3) False positive rate, and 4) False negative rate. These rates are defined below:

Type I error rate: proportion of simulated trials in which the null is true but falsely rejected:

$$\text{Type I error rate} = \frac{\text{\# reject simulation trials in null}}{\text{\# simulation trials in null}}.$$

Type II error rate: proportion of simulated trials in which the alternative is true but falsely rejected:

$$\text{Type II error rate} = \frac{\text{\# accept simulation trials in alternative}}{\text{\# simulation trials in alternative}}.$$

False positive rate (FPR): proportion of simulated trials in which the null is rejected but true:

$$\text{FPR} = \frac{\text{\# reject simulation trials in null}}{\text{\# reject simulation trials}}.$$

False negative rate (FNR): proportion of simulated trials in which the null is accepted but not true:

$$\text{FNR} = \frac{\text{\# accept simulation trials in alternative}}{\text{\# accept simulation trials}}.$$

REFERENCES

Milton Abramowitz, Irene A Stegun, and Robert H Romer. Handbook of mathematical functions with formulas, graphs, and mathematical tables, 1988.

CJ Adcock. Sample size determination: a review. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(2):261–283, 1997.

Ethan M Alt, Xiuya Chang, Xun Jiang, Qing Liu, May Mo, H Amy Xia, and Joseph G Ibrahim. Leap: The latent exchangeability prior for borrowing information from historical data. *arXiv preprint arXiv:2303.05223*, 2023.

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.

Scott M Berry, Bradley P Carlin, J Jack Lee, and Peter Muller. *Bayesian adaptive methods for clinical trials*. CRC press, 2010.

Dehua Bi and Yuan Ji. Pam: Plaid atoms model for bayesian nonparametric analysis of grouped data. *arXiv preprint arXiv:2304.14954*, 2023.

Gabriel Blumenthal, Lakhmir Jain, Amy L Loeser, Yogesh K Pithaval, Arshad Rahman, Mark J Ratain, Manish Shah, Laurie Strawn, and Marc R Theoret. Optimizing dosing in oncology drug development. *Friends Cancer Res*, pages 1–14, 2021.

Federico Camerlenghi, David B Dunson, Antonio Lijoi, Igor Prünster, and Abel Rodríguez. Latent nested nonparametric priors (with discussion). *Bayesian Analysis*, 14(4):1303–1356, 2019.

Antonio Canale and David B Dunson. Bayesian kernel mixtures for counts. *Journal of the American Statistical Association*, 106(496):1528–1539, 2011.

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*, (just-accepted):1–30, 2023.

Wei-Chen Chen, Chenguang Wang, Heng Li, Nelson Lu, Ram Tiwari, Yunling Xu, and Lilly Q Yue. Propensity score-integrated composite likelihood approach for augmenting the control arm of a randomized controlled trial by incorporating real-world data. *Journal of Biopharmaceutical Statistics*, 30(3):508–520, 2020.

Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 2010.

Maria M Ciarleglio and Christopher D Arendt. Sample size determination for a binary response in a superiority clinical trial using a hybrid classical and bayesian procedure. *Trials*, 18:1–21, 2017.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

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*, pages 1–12, 2021.

MM Desu. *Sample size methodology*. Elsevier, 2012.

David B Dunson and Ju-Hyun Park. Kernel stick-breaking processes. *Biometrika*, 95(2): 307–323, 2008.

John Eng. Sample size estimation: how many individuals should be studied? *Radiology*, 227 (2):309–313, 2003.

Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

US FDA. Project optimus: Reforming the dose optimization and dose selection paradigm in oncology. *Food and Drug Administration*, 2023a.

US FDA. Considerations for the design and conduct of externally controlled trials for drug and biological products, 2023b.

US FDA. Optimizing the dosage of human prescription drugs and biological products for the treatment of oncologic diseases, 2023c.

Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

US Food, Drug Administration, et al. Use of real-world evidence to support regulatory decision-making for medical devices: guidance for industry and food and drug administration staff. *Food and Drug Administration*, 2017.

Chris Fraley and AE Raftery. Mclust: Software for model-based cluster and discriminant analysis. *Department of Statistics, University of Washington: Technical Report*, 342, 1998.

Daniela Graf, Raffaella Di Cagno, Frida Fåk, Harry J Flint, Margareta Nyman, Maria Saarela, and Bernhard Watzl. Contribution of diet to the composition of the human gut microbiota. *Microbial ecology in health and disease*, 26(1):26164, 2015.

Brian P Hobbs, Bradley P Carlin, Sumithra J Mandrekar, and Daniel J Sargent. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3):1047–1056, 2011.

Brian P Hobbs, Daniel J Sargent, and Bradley P Carlin. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis (Online)*, 7(3):639, 2012.

Paul G Hoel et al. Introduction to mathematical statistics. *Introduction to mathematical statistics.*, (2nd Ed), 1954.

Roger A Horn and Charles R Johnson. Norms for vectors and matrices. *Matrix analysis*, pages 313–386, 1990.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1): 193–218, 1985.

Joseph G Ibrahim and Ming-Hui Chen. Power prior distributions for regression models. *Statistical Science*, pages 46–60, 2000.

Lurdes YT Inoue, Donald A Berry, and Giovanni Parmigiani. Relationship between bayesian and frequentist sample size determination. *The American Statistician*, 59(1):79–87, 2005.

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

Lawrence Joseph and Patrick Bélisle. Bayesian sample size determination for normal means and differences between normal means. *Journal of the Royal Statistical Society Series D: The Statistician*, 46(2):209–226, 1997.

Juan Jovel, Jordan Patterson, Weiwei Wang, Naomi Hotte, Sandra O'Keefe, Troy Mitchel, Troy Perry, Dina Kao, Andrew L Mason, Karen L Madsen, et al. Characterization of the gut microbiome using 16s or shotgun metagenomics. *Frontiers in microbiology*, 7:459, 2016.

Maria Kalli, Jim E Griffin, and Stephen G Walker. Slice sampling mixture models. *Statistics and computing*, 21(1):93–105, 2011.

Vytautas Kasiulevičius, V Šapoka, and R Filipavičiūtė. Sample size calculation in epidemiological studies. *Gerontologija*, 7(4):225–231, 2006.

Abhishek Kaul, Siddhartha Mandal, Ori Davidov, and Shyamal D Peddada. Analysis of microbiome data in the presence of excess zeros. *Frontiers in microbiology*, 8:2114, 2017.

Fahime Khozeimeh, Roohallah Alizadehsani, Mohamad Roshanzamir, Abbas Khosravi, Pouran Layegh, and Saeid Nahavandi. An expert system for selecting wart treatment method. *Computers in biology and medicine*, 81:167–175, 2017a.

Fahime Khozeimeh, Farahzad Jabbari Azad, Yaghoub Mahboubi Oskouei, Majid Jafari, Shahrzad Tehranian, Roohallah Alizadehsani, and Pouran Layegh. Intralesional immunotherapy compared to cryotherapy in the treatment of warts. *International journal of dermatology*, 56(4):474–478, 2017b.

Jae H Kim and In Choi. Choosing the level of significance: A decision-theoretic approach. *Abacus*, 57(1):27–71, 2021.

Gary King and Richard Nielsen. Why propensity scores should not be used for matching. *Political analysis*, 27(4):435–454, 2019.

Kevin Kunzmann, Michael J Grayling, Kim May Lee, David S Robertson, Kaspar Rufibach, and James MS Wason. A review of bayesian perspectives on sample size derivation for confirmatory trials. *The American Statistician*, 75(4):424–432, 2021.

Sandra J Lee and Marvin Zelen. Clinical trials and sample size considerations: another perspective. *Statistical science*, 15(2):95–110, 2000.

Antonio Lijoi, Igor Prünster, and Giovanni Rebaudo. Flexible clustering via hidden hierarchical dirichlet priors. *Scandinavian Journal of Statistics*, 2022.

Xiaolei Lin, Jiaying Lyu, Shijie Yuan, Dehua Bi, Sue-Jane Wang, and Yuan Ji. Bayesian sample size planning tool for phase i dose-finding trials. *JCO Precision Oncology*, 6: e2200046, 2022.

Steven N. MacEachern. Dependent nonparametric processes. *ASA Proceedings of the Section on Bayesian Statistical Science, VA.*, 1999.

Steven N. MacEachern. Dependent dirichlet processes technical report. *Department of Statistics, The Ohio State University*, 2000.

J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297, 1967.

Ella Marley-Zagar, Ian R White, Mahesh KB Parmar, Patrick Royston, Abdel G Babiker, et al. A unified stata package for calculating sample sizes for trials with binary outcomes (artbin). In *London Stata Conference 2021*, number 3. Stata Users Group, 2021.

Matthew S Mayo and Byron J Gajewski. Bayesian sample size calculations in phase ii clinical trials using informative conjugate priors. *Controlled clinical trials*, 25(2):157–167, 2004.

Marina Meilă. Comparing clusterings—an information based distance. *Journal of multivariate analysis*, 98(5):873–895, 2007.

Ramsés H Mena, Matteo Ruggiero, and Stephen G Walker. Geometric stick-breaking processes for continuous-time bayesian nonparametric modeling. *Journal of Statistical Planning and Inference*, 141(9):3217–3230, 2011.

Jeffrey W Miller. An elementary derivation of the chinese restaurant process from sethuraman's stick-breaking process. *Statistics & Probability Letters*, 146:112–117, 2019.

Satoshi Morita, Peter F Thall, and Peter Müller. Determining the effective sample size of a parametric prior. *Biometrics*, 64(2):595–602, 2008.

Peter Müller, Giovanni Parmigiani, Christian Robert, and Judith Rousseau. Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, 99(468):990–1001, 2004a.

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, 2004b.

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

Cyr E M'Lan, Lawrence Joseph, and David B Wolfson. Bayesian sample size determination for binomial proportions. *Bayesian Analysis*, 3(2):269–296, 2008.

Marlies Noordzij, Giovanni Tripepi, Friedo W Dekker, Carmine Zoccali, Michael W Tanck, and Kitty J Jager. Sample size calculations: basic principles and common pitfalls. *Nephrology dialysis transplantation*, 25(5):1388–1393, 2010.

Stephen JD O'Keefe, Jia V Li, Leo Lahti, Junhai Ou, Franck Carbonero, Khaled Mohammed, Joram M Posma, James Kinross, Elaine Wahl, Elizabeth Ruder, et al. Fat, fibre and cancer risk in african americans and rural africans. *Nature communications*, 6(1):1–14, 2015.

Panagiotis Papastamoulis. label. switching: An r package for dealing with the label switching problem in mcmc outputs. *arXiv preprint arXiv:1503.02271*, 2015.

Panagiotis Papastamoulis and George Iliopoulos. An artificial allocations based solution to the label switching problem in bayesian analysis of mixtures of distributions. *Journal of Computational and Graphical Statistics*, 19(2):313–331, 2010.

Thu Pham-Gia, Noyan Turkkan, and P Eng. Bayesian analysis of the difference of two proportions. *Communications in Statistics-Theory and Methods*, 22(6):1755–1771, 1993.

Jim Pitman. Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2):145–158, 1995.

Jim Pitman. Poisson–dirichlet and gem invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing*, 11(5): 501–514, 2002.

Mădălina Preda, Mircea Ioan Popa, Mara Mădălina Mihai, Teodora Cristiana Oţelea, and Alina Maria Holban. Effects of coffee on intestinal microbiota, immunity, and disease. *Caffeinated and Cocoa Based Beverages*, pages 391–421, 2019.

Fernando A Quintana, Peter Müller, Alejandro Jara, and Steven N MacEachern. The dependent dirichlet process and related models. *Statistical Science*, 37(1):24–41, 2022.

Abel Rodriguez, David B Dunson, and Alan E Gelfand. The nested dirichlet process. *Journal of the American statistical Association*, 103(483):1131–1154, 2008.

Tushar Vijay Sakpal. Sample size estimation in clinical trial. *Perspectives in clinical research*, 1(2):67, 2010.

Heinz Schmidli, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O'Hagan, David Spiegelhalter, and Beat Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, 2014.

Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

Mirat Shah, Atiqur Rahman, Marc R Theoret, and Richard Pazdur. The drug-dosing conundrum in oncology-when less is more. *The New England journal of medicine*, 385(16): 1445–1447, 2021.

Eric L Simpson, Kim A Papp, Andrew Blauvelt, Chia-Yu Chu, H Chih-ho Hong, Norito Katoh, Brian M Calimlim, Jacob P Thyssen, Albert S Chiou, Robert Bissonnette, et al. Efficacy and safety of upadacitinib in patients with moderate to severe atopic dermatitis: analysis of follow-up data from the measure up 1 and measure up 2 randomized clinical trials. *JAMA dermatology*, 158(4):404–413, 2022.

Say-Beng Tan and David Machin. Bayesian two-stage designs for phase ii clinical trials. *Statistics in medicine*, 21(14):1991–2012, 2002.

Yee Teh, Michael Jordan, Matthew Beal, and David Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. *Advances in neural information processing systems*, 17, 2004.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

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

Chenguang Wang and Wei-Chen Chen. psrwe: Ps-integrated methods for incorporating rwe in clinical studies. r package, version 3.1, 2022.

Xiaofeng Wang and Xinge Ji. Sample size estimation in clinical research: from randomized controlled trials to observational studies. *Chest*, 158(1):S12–S20, 2020.

Janet Wittes. Sample size calculations for randomized controlled trials. *Epidemiologic reviews*, 24(1):39–53, 2002.

Zhong Zhao. Using matching to estimate treatment effects: Data requirements, matching metrics, and monte carlo evidence. *Review of economics and statistics*, 86(1):91–107, 2004.