Bayesian Analysis (2024) **TBA**, Number TBA, pp. 1–28

A Tree Perspective on Stick-Breaking Models in Covariate-Dependent Mixtures*

Akira Horiguchi†, Cliburn Chan‡, and Li Ma§

Abstract. Stick-breaking (SB) processes are often adopted in Bayesian mixture models for generating mixing weights. When covariates influence the sizes of clusters, SB mixtures are particularly convenient as they can leverage their connection to binary regression to ease both the specification of covariate effects and posterior computation. Existing SB models are typically constructed based on continually breaking a single remaining piece of the unit stick. We view this from a dyadic tree perspective in terms of a lopsided bifurcating tree that extends only on one side. We show that two unsavory characteristics of SB models are in fact largely due to this lopsided tree structure. We consider a generalized class of SB models with alternative bifurcating tree structures and examine the influence of the underlying tree topology on the resulting Bayesian analysis in terms of prior assumptions, posterior uncertainty, and computational effectiveness. In particular, we provide evidence that a balanced tree topology, which corresponds to continually breaking all remaining pieces of the unit stick, can resolve or mitigate these undesirable properties of SB models that rely on a lopsided tree.

Keywords: discrete random measure, Bayesian nonparametrics, tail-free process, clustering analysis, flow cytometry.

1 Introduction

Mixture models are a popular approach to clustering analysis. A mixture model's weights are often generated by stick-breaking processes (Sethuraman, 1994; Ishwaran and James, 2001) largely due to their tractable posterior computation. Stick-breaking processes are particularly well-suited when covariates influence cluster sizes; these models can readily incorporate covariate influence through binary regression at each stick break and can leverage efficient posterior computation strategies developed for binary regression models. A notable example is Rodríguez and Dunson (2011)'s probit stick-breaking whose Markov chain Monte Carlo algorithm relies on Albert and Chib (1993)'s truncated Gaussian data augmentation for probit regression. Similarly, for logit stick-breaking Rigon and Durante (2021) provide various posterior sampling methods that rely on Polson et al. (2013)'s Pólya-Gamma data augmentation for logistic regression.

*This research was supported by the Translating Duke Health Initiative (Immunology) [AH, CC, LM], the Duke University Center for AIDS Research (CFAR), an NIH funded program (5P30 AI064518) [SW, CC], the National Institute of General Medical Sciences grant R01-GM135440 [AH, LM], and the National Science Foundation grant DMS-2013930 [AH, LM].

†Department of Statistical Science, Duke University, Durham, NC 27708, akira.horiguchi@duke.edu

‡Department of Biostatistics and Bioinformatics and Center for Human Systems Immunology, Duke University School of Medicine, Durham, NC 27708, cliburn.chan@duke.edu

§Department of Statistical Science and Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27708, li.ma@duke.edu

 https://doi.org/10.1214/24-BA1462

Linderman et al. (2015) also employ Pólya-Gamma data augmentation for stick-breaking in dependent multinomial modeling. Alternatively, mixing-weight dependence could be induced by partially exchangeable observations (de Finetti, 1938; Diaconis and Freedman, 1980; Diaconis, 1988) or by splitting variables without explicitly incorporating covariates (e.g. Fuentes-Garcia et al., 2010; Favaro et al., 2012, 2016; Gil–Leyva et al., 2020; Gil-Leyva and Mena, 2023). Our paper will focus on dependence that is explicitly characterized by covariates in terms of their effects on the mixing weights.

While our main applied motivation is the need for covariate-dependent clustering in the analysis of flow cytometry data (see Section 4), stick-breaking models and their dependent variants have also been adopted in a wide range of other applied contexts. Such contexts include time-series data (Griffin and Steel, 2011; Bassetti et al., 2014), spatial data (Griffin and Steel, 2006; Duan et al., 2007; Reich and Fuentes, 2007; Dunson and Park, 2008), and space-time data (Hossain et al., 2013; Grazian, 2024) among many other examples. Deepening the understanding of this useful model class and improving its design and construction can thus lead to better statistical practice across these applications.

Although all discrete random measures in principle admit a stick-breaking representation, identifying the pairing between their stick-breaking construction and other constructive representations is an interesting ongoing endeavor in the Bayesian nonparametrics literature. Following the canonical construction of stick-breaking representations for the Dirichlet process (Ferguson, 1973; Sethuraman, 1994) and the Pitman-Yor process (Pitman and Yor, 1997), the stick-breaking representation has been successfully identified for the broad class of homogeneous normalized random measures with independent increments (Regazzini et al., 2003; Favaro et al., 2016). This class contains as special cases several processes whose stick-breaking representations were previously identified, including the normalized generalized gamma process (Lijoi et al., 2007), which in turn contains the normalized inverse Gaussian process (Lijoi et al., 2005; Favaro et al., 2012) and the Dirichlet process. Beyond random probability measures, stick-breaking constructions have also been identified for some completely random measures such as the beta process (Teh et al., 2007; Paisley et al., 2010, 2012; Broderick et al., 2012; Hjort, 1990), which has been applied in latent factor models (Ghahramani and Griffiths, 2005; Thibaux and Jordan, 2007) and mixed-membership models (Fox and Jordan, 2014).

While much of the existing literature investigates stick-breaking representations for known random measures, the reverse direction of identifying alternative representations for a given stick-breaking process is also of interest. Such a representation, either in the form of a random measure or otherwise, would provide both theoretical and practical insights into key properties of the stick-breaking model. The only representations for stick-breaking processes with independent splitting variables so far identified are the Dirichlet and Pitman-Yor processes (Pitman, 1996; Favaro et al., 2012). As a reviewer kindly pointed out, this appears to be due to the fact that the Dirichlet and the Pitman-Yor processes are the only stick-breaking processes with independent splitting variables being invariant under size-biased permutations as proved in Pitman (1996), and without such a property it seems impossible with current techniques to derive alternative representations for stick-breaking models with independent splitting variables.

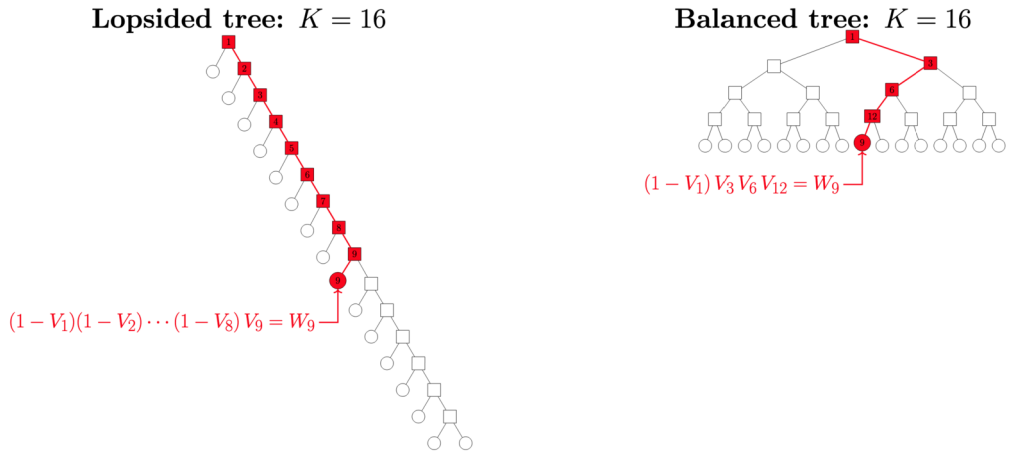

Figure 1: A binary tree representation of stick-breaking schemes. Each scheme's initial stick break is represented as a root node; each subsequent stick break is represented by an internal node. The mass of a stick-break's "left (right) piece" is sent to its left (right) child. The K leaf nodes represent the scheme's K weights. As an example, weight W_9 is a product of nine stick breaks in the lopsided-tree scheme but only four stick breaks in the balanced-tree scheme.

Existing stick-breaking models for random probability measures generate a set of unit-sum weights through continually breaking pieces off a unit stick, one piece at a time. This mechanism can be viewed through a dyadic tree perspective in which each stick-breaking step corresponds to a bifurcating split of the remaining stick resulting in two "children", one representing a leaf (i.e., terminal) node corresponding to the piece broken off from the stick, and the other a non-leaf (i.e., interior) node corresponding to the remaining stick to be further broken. Without loss of generality, we use the *left* child of each split to represent the leaf node (i.e., the piece broken off) and the *right* child as the remaining stick. Following this delineation, the tree has a lopsided shape which extends only on the right side as shown in Figure 1. Such a stick-breaking mechanism, which we shall refer to as lopsided-tree stick-breaking, has been widely applied in the Bayesian nonparametric literature in constructing discrete random measures.

Our work is motivated by the observation that some well-known properties of existing stick-breaking models, for example the stochastic ordering of the resulting weights, are in fact attributable to this lopsided-tree structure, whereas the key benefits of stick-breaking models, such as the ability to transform modeling and Bayesian computation into binary regression problems, require only a dyadic tree structure which need not be lopsided. It is therefore a promising strategy to resolve some limitations of stick-breaking models by generalizing beyond the lopsided tree to different dyadic tree structures, while maintaining the modeling and computational benefits. A natural question is whether such a generalization is practically useful. The short answer, as we shall demonstrate, is indeed yes. The choice of the tree structure can substantially influence statistical inference under stick-breaking models.

More specifically, we show that in the context of covariate-dependent mixture modeling, which is a major field of application for stick-breaking models, common model specifications with a lopsided tree structure induce two undesirable characteristics that can severely deteriorate the quality of Bayesian inference in terms of increased posterior uncertainty and reduced computational efficiency. Some of these properties have been noted previously in the literature, but they were never attributed to the underlying lopsided tree structure, which we show is a major culprit. In addition, we show through a combination of theoretical analysis, numerical experiments, and a case study that these two undesirable features of stick-breaking models can be resolved by simply adopting an alternative dyadic tree structure, namely a "balanced" tree (illustrated in Figure 1), corresponding to continually breaking off *both* remaining sticks at each stick-breaking step, without complicating the modeling and computational recipes that stick-breaking models enjoy.

What are these two undesirable characteristics of (lopsided-tree) stick-breaking processes in covariate-dependent mixture modeling? The first characteristic, which will be the focus of Section 3, is that under commonly adopted "default" specifications, covariate-dependent stick-breaking models can induce a strong positive prior correlation in the random measures over covariate values, which can cause excessive smoothing across the covariates even when the corresponding cluster sizes vary sharply across covariate values, thereby deprecating the clustering at each covariate value. This phenomenon was first explained in Rodríguez and Dunson (2011) in the context of probit-stick-breaking models as "a consequence of our use of a common set of atoms at every [covariate value]; even if the set of weights are independent from each other, the fact that the atoms are shared means that the distributions cannot be independent." Interestingly, we show that it is the lopsided-tree structure underlying standard stick-breaking that multiplies the effects of shared atoms on prior correlation.

The second characteristic, which will be the theme of Section 4, concerns the precision of the inference in terms of the posterior uncertainty of covariate effects on mixing weights. Common prior specifications of covariate effects in stick-breaking weights introduce many competing mechanisms in a mixture model. Such specifications include stochastic ordering of the weights and the large number of stick breaks the weights is a product of (on average); these both introduce mechanisms that, at best, add unnecessary layers of complexity and, at worst, actively degrade the quality of posterior inference.

To finish the introduction, we relate to some relevant papers in the literature that involve tree-structured Bayesian nonparametric models. Ghahramani et al. (2010) model hierarchical data by interleaving two lopsided stick-breaking processes to allow a wide range of latent tree structures (e.g. nodes with any number of children) to be inferred. On the other hand, stick-breaking models over a general tree structure are closely related to tree-structured random measures such as the Pólya tree (Ferguson, 1973; Lavine, 1992; Maudlin et al., 1992), and their generalizations involving covariate-dependence has a parallel development for tree-structured random measures, namely the covariate-dependent tail-free model introduced by Jara and Hanson (2011). To our knowledge, such processes have not been applied as a *discrete* prior for mixing distributions except

in Cipolli and Hanson (2017) who propose the finite Pólya tree as a prior for the mixing distributions but do not argue the use of their approach over lopsided stick-breaking processes. Stefanucci and Canale (2021) introduce another weight-generating mechanism along a balanced bifurcating tree that includes stick-breaking models as special cases, but this generalization assigns weights to every node of the tree, which makes this approach lose the main computational benefits of stick-breaking models in incorporating covariates. Finally, Ren et al. (2011) note a covariate-dependent stick-breaking process "may be viewed as a mixture-of-experts model" (Jordan and Jacobs, 1994; Peng et al., 1996; Bishop and Svensén, 2002) which uses a binary tree to define a covariate-dependent mixture distribution. Although we investigate the influence of the tree structure mainly in the context of stick-breaking models, the lessons drawn are also relevant for mixture-of-experts models given the connection between the two model classes.

2 Stick-breaking models from a tree perspective

We start by introducing our tree-based stick-breaking construction in the absence of covariates and then extend the construction to include covariates. This broader class of models contains existing stick-breaking models as a special case corresponding to a particular lopsided tree structure. This will let us later examine the impact of the tree structure on Bayesian inference through the lens of comparing models within this broader class.

This paper assumes conditionally independent and identically distributed (i.i.d.) observations $y_1, \ldots, y_n \in \mathbb{R}^d$, given a mixing measure G, from a sampling density of the mixture form $f(\cdot; G) = \int_\Theta h(\cdot; \theta)\, \mathrm{d}G(\theta)$, where $h\colon \mathbb{R}^d \times \Theta \to \mathbb{R}^+$ is a parametric mixing kernel density, e.g., Gaussian density $h(y; \theta) = (2\pi\sigma^2)^{-1/2} \exp\left\{-(y-\mu)^2/(2\sigma^2)\right\}$ with $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$.

If a mixing measure G on Θ is equipped with a stick-breaking prior, each realization of G is, with probability one, a discrete measure $G = \sum_{k=1}^\infty W_k \delta_{\theta_k}$ where the atoms θ_k are generated independently from a probability measure G_0 on Θ and where the nonnegative weights W_k, which are generated by breaking a unit stick, are independent of the atoms θ_k. This turns the mixture density $\int_\Theta h(\cdot; \theta)\, \mathrm{d}G(\theta)$ into the discrete sum $\sum_{k=1}^\infty W_k h(\cdot; \theta_k)$.

The remainder of this section examines the traditional stick-breaking scheme from a tree perspective, and then introduces a tree-based generalization to stick-breaking priors.

2.1 Stick-breaking models without covariates

A weight-generation process aims to produce nonnegative quantities that sum to unity. A stick-breaking process achieves this by successively breaking off pieces of a stick of initially unit length, where the lengths of the resulting pieces become the desired weights. In traditional stick-breaking, each stick break produces a piece that is untouched afterwards while the other piece—the "remaining stick"—breaks again and again. More

formally, to begin, a unit-length stick breaks at location $V_1 \in [0,1]$ so that the piece that breaks off has length V_1. The remaining stick, of length $1-V_1$, then breaks at (relative) location $V_2 \in [0,1]$ so that the new piece that breaks off has length $(1-V_1)V_2$. The remaining stick, now of length $(1-V_1)(1-V_2)$, then breaks at location $V_3 \in [0,1]$ so that the new piece that breaks off has length $(1-V_1)(1-V_2)V_3$, so on and so forth. For each $k \in \mathbb{N}$ the k^{th} stick break at location $V_k \in [0,1]$ produces a piece that breaks off with length $V_k \prod_{l=1}^{k-1}(1-V_l)$, which sets the value of weight W_k; the remaining stick then breaks at location $V_{k+1} \in [0,1]$. The process can either continue infinitely or terminate after a preset finite number of breaks, in which case the final remaining stick sets the last weight.

A stick-breaking scheme can be identified with a bifurcating tree whose nodes correspond to the pieces that arise during the stick breaking procedure. We construct the binary tree by first assigning the initial unit-length stick to the tree's root node, and iteratively apply the following steps. If a piece of the stick does not break further, the piece corresponds to a leaf node, i.e., a node with no children. If a piece is further broken instead, the two resulting pieces correspond to two children nodes. The traditional stick-breaking scheme's identified bifurcating tree (see the left of Figure 1) is the most "lopsided" as none of its nodes have children that both divide further. It is the deepest tree possible for generating a given number of pieces. We shall thus refer to the traditional stick-breaking strategy as *lopsided-tree stick-breaking*.

Given this tree view of stick-breaking, it is natural to consider stick-breaking schemes corresponding to binary tree topologies that differ from the lopsided one. To this end, it will be convenient to index the root node by the empty string $\emptyset$ and each other tree node $\boldsymbol{\varepsilon}$ by a finite string of 0s and 1s, where each digit indicates whether a node along the path from root to $\boldsymbol{\varepsilon}$ is a left (0) or right (1) child of its parent. In general, any node at level m of the tree (the root node is at level 0, its two children are at level 1, and so on) is indexed by a m-length binary string $\varepsilon_1\varepsilon_2\cdots\varepsilon_m$, where the string $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'$ denotes the concatenation of finite strings $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}'$. The set of all finite binary strings (including $\emptyset$) is denoted by $E^* := \cup_{m=0}^{\infty}\{0,1\}^m$. This machinery allows us to formally introduce stick-breaking strategies based on general tree topologies.

We give particular attention to the stick-breaking scheme corresponding to a "balanced" bifurcating tree, in which all nodes (up to a maximum level) are split into two children as illustrated at the right of Figure 1. We refer to this stick-breaking scheme as *balanced-tree stick-breaking*. Opposite of the lopsided-tree scheme, the balanced-tree scheme results in the most shallow tree structure needed to generate any given number of weights. In this scheme, each stick break produces two pieces that *both* break again until the total number of breaks exceeds a preset threshold (we discuss allowing infinitely many breaks in Section 3.3). More formally, the unit-length stick $I_\emptyset$ breaks according to $V_\emptyset$ so that the left piece I_0 has length $|I_0| = |I_\emptyset|V_\emptyset = V_\emptyset$ and the right piece has length $|I_1| = |I_\emptyset|(1-V_\emptyset) = 1-V_\emptyset$. These two pieces I_0 and I_1 then break similarly according to respective splitting variables V_0 and V_1 so that the resulting four pieces I_{00}, I_{01}, I_{10}, and I_{11} have lengths $|I_{00}| = V_\emptyset V_0$, $|I_{01}| = V_\emptyset(1-V_0)$, $|I_{10}| = (1-V_\emptyset)V_1$, and $|I_{11}| = (1-V_\emptyset)(1-V_1)$. These four pieces then break similarly according to respective splitting variables V_{00}, V_{01}, V_{10}, and V_{11} to produce the eight pieces I_{000}, I_{001}, I_{010}, I_{011}, I_{100}, I_{101}, I_{110}, and I_{111} with lengths defined similarly. If each piece is only allowed to

be a product of at most $m \in \mathbb{N}$ stick breaks, this recursive procedure produces 2^m leaf nodes each with a stick piece. The lengths of these pieces define the desired weights: $W_{\boldsymbol{\varepsilon}} := |I_{\boldsymbol{\varepsilon}}|$ for each m-length string $\boldsymbol{\varepsilon} \in \{0,1\}^m$.

The value of any stick-breaking weight at a node $\boldsymbol{\varepsilon} = \varepsilon_1 \cdots \varepsilon_m$ can be expressed as

$$W_{\varepsilon_1 \cdots \varepsilon_m} = \prod_{l=1}^{m} V_{\varepsilon_1 \cdots \varepsilon_{l-1}}^{1-\varepsilon_l} (1 - V_{\varepsilon_1 \cdots \varepsilon_{l-1}})^{\varepsilon_l} \qquad V_{\boldsymbol{\varepsilon}} \sim F_{\boldsymbol{\varepsilon}} \text{ for all } \boldsymbol{\varepsilon} \in E^*, \tag{1}$$

where by convention $\varepsilon_1 \cdots \varepsilon_{l-1} = \emptyset$ if $l = 1$, and the splitting variables are distributed according to a countable sequence $\{F_{\boldsymbol{\varepsilon}} \colon \boldsymbol{\varepsilon} \in E^*\}$ of distributions each with full support on $[0,1]$ and the tail-free condition $V_\emptyset \perp \{V_0, V_1\} \perp \{V_{00}, V_{01}, V_{10}, V_{11}\} \perp \cdots$ (Freedman, 1963a; Fabius, 1964a).

We can now define the *tree stick-breaking* class of priors, which includes traditional stick-breaking (where each level of the tree beyond level zero has exactly one weight and that node takes the form $1 \cdots 10$) and balanced-tree stick-breaking but also admits other tree structures.

Definition 1. *If G_0 is a probability measure, $\{F_{\boldsymbol{\varepsilon}} : \boldsymbol{\varepsilon} \in E^*\}$ is a sequence of distributions each with support $[0,1]$, and τ is a binary tree structure, we say a probability measure G is equipped with a tree stick-breaking prior with parameters G_0, $\{F_{\boldsymbol{\varepsilon}}\}$, and τ if it can be constructed as*

$$G = \sum_{\boldsymbol{\varepsilon} \in B(\tau)} W_{\boldsymbol{\varepsilon}} \delta_{\theta_{\boldsymbol{\varepsilon}}}, \tag{2}$$

where the set $B(\tau) \subset E^*$ indexes τ's leaf nodes, the random weights $\{W_{\boldsymbol{\varepsilon}} \colon \boldsymbol{\varepsilon} \in B(\tau)\}$ are constructed according to (1), the splitting variables $V_{\boldsymbol{\varepsilon}} \sim F_{\boldsymbol{\varepsilon}}$ and satisfy $V_\emptyset \perp \{V_0, V_1\} \perp \{V_{00}, V_{01}, V_{10}, V_{11}\} \perp \cdots$, and the atoms $\theta_{\boldsymbol{\varepsilon}} \overset{ind}{\sim} G_0$ and are generated independently of the splitting variables $V_{\boldsymbol{\varepsilon}}$. Such a measure G is denoted by $G \sim treeSB(G_0, \{F_{\boldsymbol{\varepsilon}}\}, \tau)$.

As with traditional stick-breaking, a random measure $G \sim treeSB(G_0, \{F_{\boldsymbol{\varepsilon}}\}, \tau)$ turns the mixture density $\int_\Theta h(\cdot;\theta)\, \mathrm{d}G(\theta)$ into the discrete sum $\sum_{\boldsymbol{\varepsilon} \in B(\tau)} W_{\boldsymbol{\varepsilon}} h(\cdot; \theta_{\boldsymbol{\varepsilon}})$.

2.2 Tree stick-breaking with covariates

Following the well-known strategy of MacEachern (2000), we next extend the covariate-independent tree stick-breaking prior (2) by replacing each splitting variable $V_{\boldsymbol{\varepsilon}}$ in (1) with a stochastic process $\{V_{\mathbf{x},\boldsymbol{\varepsilon}} \colon \mathbf{x} \in \mathcal{X}\}$, where $\mathcal{X}$ is a set of covariates and each splitting variable $V_{\mathbf{x},\boldsymbol{\varepsilon}}$ has distribution $F_{\mathbf{x},\boldsymbol{\varepsilon}}$. The resulting random measure $G_{\mathbf{x}} \sim treeSB(G_0, \{F_{\mathbf{x},\boldsymbol{\varepsilon}}\}, \tau)$ now depends on $\mathbf{x}$ through its weights: $G_{\mathbf{x}} = \sum_{\boldsymbol{\varepsilon} \in B(\tau)} W_{\mathbf{x},\boldsymbol{\varepsilon}} \delta_{\theta_{\boldsymbol{\varepsilon}}}$ where $\theta_{\boldsymbol{\varepsilon}} \overset{ind}{\sim} G_0$. As in the "common-atoms" or "single-atoms" model (see e.g., Quintana et al., 2022, for a review), here we do not incorporate dependence into the atoms (nor in the tree structure).

There are a number of possible strategies to incorporate covariate dependence on weights, including utilizing probit and logit transform on the weights as is commonly done for traditional stick-breaking. In particular, we consider the logit approach due

to the computational convenience for posterior inference that follows from the Pólya-Gamma augmentation technique detailed in Section 4, though the theoretical properties we establish in Section 3.1 do not assume this model choice and apply more generally. Specifically, we adopt the following logit-normal model on each splitting variable:

$$V_{\mathbf{x},\boldsymbol{\varepsilon}} = \text{logistic}(\eta_{\mathbf{x},\boldsymbol{\varepsilon}}), \qquad \eta_{\mathbf{x},\boldsymbol{\varepsilon}} = \boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\gamma}_{\boldsymbol{\varepsilon}}, \qquad \boldsymbol{\gamma}_{\boldsymbol{\varepsilon}} \sim N_R(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}) \tag{3}$$

with hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$, where $\text{logistic}(z) = \exp\{z/(1+z)\}$ and $\eta_{\mathbf{x},\boldsymbol{\varepsilon}}$ is a linear combination of selected functions of the covariates $\boldsymbol{\psi}(\mathbf{x}) = \{\psi_1(\mathbf{x}), \ldots, \psi_R(\mathbf{x})\}^\top$. Thus $\{\eta_{\mathbf{x},\boldsymbol{\varepsilon}} \colon \mathbf{x} \in \mathcal{X}\}$ is a Gaussian process with mean $\boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\mu}_{\boldsymbol{\gamma}}$ and covariance $\text{Cov}(\eta_{\mathbf{x},\boldsymbol{\varepsilon}}, \eta_{\mathbf{x}',\boldsymbol{\varepsilon}}) = \boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \boldsymbol{\psi}(\mathbf{x}')$. The remainder of the paper (except for Sections 3.1 and 3.3) assumes the logit-normal prior (3) or a mixed-effects version of it.

3 Impact of the tree on cross-covariate correlation

This section examines the impact of tree structure on the prior cross-covariate correlation between two random measures created by stick breaking with dependent mixture weights and independent atoms. We provide expressions for various moments of the covariate-dependent random measures and create a simulation study that explores the cross-covariate correlation between random measures.

3.1 Moments of random measures

The prior (3) satisfies the conditions Theorems 1 and 2 place on the splitting variables.

Theorem 1. *For some $K \in \{1, 2, \ldots, \infty\}$, suppose for any covariates $\mathbf{x}$ that the random measure $G_{\mathbf{x}} = \sum_{k=1}^{K} W_{\mathbf{x},k}\, \delta_{\theta_k}$ on Θ is constructed by drawing each θ_k independently from a base measure G_0 (that does not depend on $\mathbf{x}$) on Θ, and drawing a weight vector $(W_{\mathbf{x},1}, \ldots, W_{\mathbf{x},K})$ according to some distribution (that might depend on $\mathbf{x}$) on the probability simplex Δ_K.*

For any measurable sets $A, A' \in \mathcal{B}$ and covariates $\mathbf{x}, \mathbf{x}'$, we have

$$E(G_{\mathbf{x}}(A)) = G_0(A) \tag{4}$$

$$Var(G_{\mathbf{x}}(A)) = \left\{G_0(A) - G_0^2(A)\right\} a_{\mathbf{x},\mathbf{x}} \tag{5}$$

$$Cov(G_{\mathbf{x}}(A), G_{\mathbf{x}}(A')) = \{G_0(A \cap A') - G_0(A)G_0(A')\}\, a_{\mathbf{x},\mathbf{x}} \tag{6}$$

$$Cov(G_{\mathbf{x}}(A), G_{\mathbf{x}'}(A)) = \left\{G_0(A) - G_0^2(A)\right\} a_{\mathbf{x},\mathbf{x}'}, \tag{7}$$

where $a_{\mathbf{x},\mathbf{x}'} = \sum_{k=1}^{K} E(W_{\mathbf{x},k} W_{\mathbf{x}',k})$. If also $0 < G_0(A) < 1$ and $0 < G_0(A') < 1$, then

$$corr(G_{\mathbf{x}}(A), G_{\mathbf{x}}(A')) = \frac{G_0(A \cap A') - G_0(A)G_0(A')}{[G_0(A)\{1 - G_0(A)\}G_0(A')\{1 - G_0(A')\}]^{1/2}} \tag{8}$$

$$corr(G_{\mathbf{x}}(A), G_{\mathbf{x}'}(A)) = a_{\mathbf{x},\mathbf{x}'}(a_{\mathbf{x},\mathbf{x}} a_{\mathbf{x}',\mathbf{x}'})^{-1/2}. \tag{9}$$

We note that the structure of these prior moments in Theorem 1 is identical to that of homogeneous normalized random measures (cfr. Proposition 1 in James et al., 2006).

Also, the correlations (8) and (9) discussed in this paper are always nonnegative, but negative correlations can be produced by inducing repulsion into the atom-generating measure for exchangeable observations (Petralia et al., 2012) or partially exchangeable observations (Ascolani et al., 2023).

In Theorem 1, each moment factorizes into a function of the base measure G_0 and a function of the quantity $a_{\mathbf{x},\mathbf{x}'}$. The mean (4) implies G_0 can be viewed as the mean of the random measure $G_{\mathbf{x}}$ while the correlation (8) does not depend on tree depth. On the other hand (and of greater interest), the cross-covariate correlation (9) depends not on the base measure G_0 but rather on $a_{\mathbf{x},\mathbf{x}'}$, which Theorem 2 expresses as a function of τ and the splitting variables' mean and cross-covariate covariance. In particular, this theorem states that $a_{\mathbf{x},\mathbf{x}'}$ for a balanced tree approaches zero as $K \to \infty$ whereas $a_{\mathbf{x},\mathbf{x}'}$ for a lopsided tree (which is mostly derived in Appendix 2 of Rodríguez and Dunson, 2011) approaches a positive limit, which means a lopsided tree induces a baseline cross-covariate *covariance* value (7) between random measures that does not vanish as the number of weights approaches infinity. Corollary 2.1 says that these statements also apply to their cross-covariate *correlation* counterparts (9) if the splitting variables have mean 1/2 and nonnegative cross-covariate correlation.

Theorem 2. *Suppose a set of weights* $\{W_{\mathbf{x},\boldsymbol{\varepsilon}} : \boldsymbol{\varepsilon} \in B(\tau)\}$ *is constructed by stick-breaking according to tree structure* τ*, where* $B(\tau)$ *is the set of leaf nodes in* τ *and* $K = |B(\tau)|$ *is the number of leaf nodes. Also assume that the distribution of any splitting variable* $V_{\mathbf{x},\boldsymbol{\varepsilon}}$ *does not depend on* $\boldsymbol{\varepsilon}$*. Let* $a_{\mathbf{x},\mathbf{x}'} = \sum_{\boldsymbol{\varepsilon}\in B(\tau)} E(W_{\mathbf{x},\boldsymbol{\varepsilon}} W_{\mathbf{x}',\boldsymbol{\varepsilon}})$*. If* τ *is a lopsided tree, then, letting* $e_{\mathbf{x},\mathbf{x}'} = E(V_{\mathbf{x}}) + E(V_{\mathbf{x}'}) - E(V_{\mathbf{x}} V_{\mathbf{x}'})$*, we have*

$$a_{\mathbf{x},\mathbf{x}'} = \frac{E(V_{\mathbf{x}} V_{\mathbf{x}'})}{e_{\mathbf{x},\mathbf{x}'}} + \left(1 - \frac{E(V_{\mathbf{x}} V_{\mathbf{x}'})}{e_{\mathbf{x},\mathbf{x}'}}\right)(1 - e_{\mathbf{x},\mathbf{x}'})^{K-1} \xrightarrow[K\to\infty]{} \frac{E(V_{\mathbf{x}} V_{\mathbf{x}'})}{e_{\mathbf{x},\mathbf{x}'}}.$$

If instead τ *is a balanced tree and* $m = \log_2 K$ *is a nonnegative integer, then*

$$a_{\mathbf{x},\mathbf{x}'} = \{1 - E(V_{\mathbf{x}}) - E(V_{\mathbf{x}'}) + 2E(V_{\mathbf{x}} V_{\mathbf{x}'})\}^m \xrightarrow[m\to\infty]{} 0.$$

The two values of $a_{\mathbf{x},\mathbf{x}'}$ above agree for $K \in \{1, 2\}$, which reflects the equivalence of any binary tree for these two values of K.

Corollary 2.1 (Bounds for (9))**.** *Given the assumptions in Theorem 2, suppose also that* $E(V_{\mathbf{x}}) = 1/2$ *for any* $\mathbf{x}$ *and* $Cov(V_{\mathbf{x}}, V_{\mathbf{x}'}) \geq 0$ *for any* $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$*.*

(a) Lower bounds for (9) are $1/3 + (2/3)4^{1-K}$ *for a lopsided tree and* 2^{-m} *for a balanced tree.*

(b) If for some $(\mathbf{x}, \mathbf{x}')$ *the conditions* $Var(V_{\mathbf{x}})\, Var(V'_{\mathbf{x}}) > 0$ *and* $corr(V_{\mathbf{x}}, V_{\mathbf{x}'}) < 1$ *are also satisfied, then the cross-covariate correlation (9) for a balanced tree and this* $(\mathbf{x}, \mathbf{x}')$ *shrinks to zero as* $m \to \infty$*.*

(c) For such $(\mathbf{x}, \mathbf{x}')$ *as in part (b), the cross-covariate correlation (9) for the lopsided tree is strictly larger than that for the balanced tree with the same number of leaves when the number of leaves is sufficiently large.*

The statement in part (c) applies to sufficiently large trees and so in principle the lopsided tree could lead to weaker cross-covariate correlation (9) than that for the balanced tree, but we found empirically that this would require the correlation $\text{corr}(V_{\mathbf{x}}, V_{\mathbf{x}'})$ to be extremely close to 1 and the trees to have very few leaves (e.g., $K = 4$ or 8), and even then the lopsided tree would have only slightly smaller cross-covariate correlation. (See the Supplemental Material (Horiguchi et al., 2024) for a numerical experiment that validates this claim.)

We will further explain and discuss the above corollary in Sections 3.2 and 3.3.

Finally, the following theorem establishes the important property that the random measure $G_{\mathbf{x}}$ changes smoothly with respect to $\mathbf{x}$. This property requires a moment condition which is satisfied if the stochastic process $\{V_{\mathbf{x}} \colon \mathbf{x} \in \mathcal{X}\}$ is second-order stationary.

Theorem 3 (Smoothness)**.** *Given the assumptions in Theorem 2, if both* $E(V_{\mathbf{x}'}) \to E(V_{\mathbf{x}})$ *and* $E(V_{\mathbf{x}}V_{\mathbf{x}'}) \to EV_{\mathbf{x}}^2$ *as* $\mathbf{x}' \to \mathbf{x}$*, then the correlation* (9) $\to 1$ *as* $\mathbf{x}' \to \mathbf{x}$*.*

3.2 Numerical illustration on prior cross-covariate correlation

Here we explore the impact of the cross-covariate correlation between splitting variables on the cross-covariate correlation between random measures: given covariates $\mathbf{x}$ and $\mathbf{x}'$, how does $\text{corr}(V_{\mathbf{x}}, V_{\mathbf{x}'})$ affect $\text{corr}\left(G_{\mathbf{x}}(A), G_{\mathbf{x}'}(A)\right)$ for any measurable set $A \in \mathcal{B}$? We provide insight into this question through the following example.

Example 1. *We reduce the number of influences on* $corr(V_{\mathbf{x}}, V_{\mathbf{x}'})$ *by making the following assumptions:* $\boldsymbol{\mu}_{\boldsymbol{\gamma}} = 0_2$, $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = diag(\sigma_1^2, \sigma_2^2)$, $\sigma_1^2 > 0$, $\boldsymbol{\psi}(\mathbf{x}) = (1, 0)^\top$, *and* $\boldsymbol{\psi}(\mathbf{x}') = (1, 1)^\top$. *Though seemingly strict, these assumptions encompass a large class of scenarios. The mean-zero assumption is reasonable if no prior information is given. These assumptions also imply* $corr(V_{\mathbf{x}}, V_{\mathbf{x}'}) = corr(\eta_{\mathbf{x}}, \eta_{\mathbf{x}'}) = (1 + \tilde{\sigma}_2^2)^{-1/2}$, *which is a strictly decreasing function of* $\tilde{\sigma}_2^2 := \sigma_2^2/\sigma_1^2 \geq 0$ *whose image is* $(0, 1]$. *Thus, any positive value of* $corr(V_{\mathbf{x}}, V_{\mathbf{x}'})$ *can be achieved by using the appropriate* $\tilde{\sigma}_2^2$ *value if* $\mathbf{x} \neq \mathbf{x}'$. *Similarly, these assumptions reduce* $corr(G_{\mathbf{x}}(A), G_{\mathbf{x}'}(A))$ *(whose expression is provided by* (9) *and Theorem 2) to a function of* σ_1^2, σ_2^2, K, *and* τ.

Figure 2 shows the behavior of $corr(G_{\mathbf{x}}(A), G_{\mathbf{x}'}(A))$, *which by Corollary 2.1 has lower bounds of* $1/3$ *and* $1/K$ *for, respectively, a lopsided tree and balanced tree. These lower bounds hold regardless of the degree of correlation between splitting variables, which in this scenario is controlled by choice of* $\tilde{\sigma}_2^2$. *Thus, the lopsided-tree scheme always imposes a nontrivial baseline correlation between random measures while the balanced-tree scheme can achieve both large and small correlation values, which provides more flexibility in setting prior correlation values.*

3.3 Choice of number of leaves

This section considers how a practitioner should choose the number of leaves K in the tree τ used in a tree stick-breaking prior. First, we consider the scenario where the splitting variables follow the conditions in Theorem 2. As done in Argiento and De Iorio (2022), we emphasize the distinction between K and the number of inferred

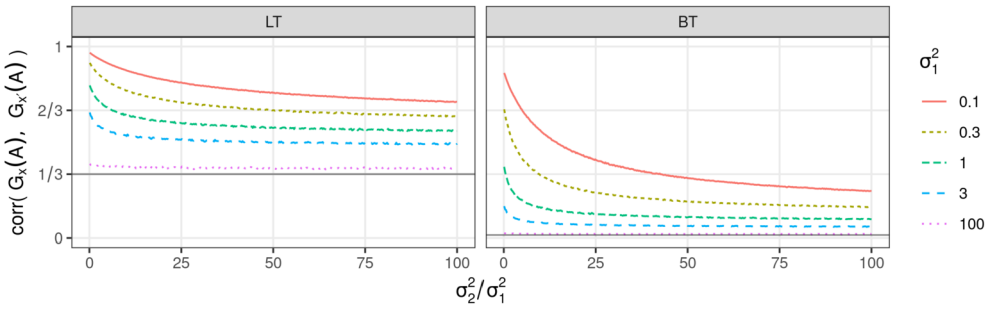

Figure 2: Simulated values of (9) as a function of $\tilde{\sigma}_2^2 := \sigma_2^2/\sigma_1^2$ for a lopsided and balanced tree each with $K = 64$ and various values of σ_1^2. Lower bounds are shown as solid horizontal lines.

clusters. The chosen K should be large enough to capture the "true" number of clusters, but not so large that the posterior inference algorithm in Section 4 becomes computationally intractable. We believe that K in the range between 16 and 64 should serve as recommended default values for most applications, which strikes a balance between flexibility and computational efficiency. For the example in Section 3.2, $K = 64$ allows the balanced-tree to achieve any prior cross-covariate correlation value (9) above $1/K \approx 0.016$. When relevant domain knowledge is available, one could alternatively set K to be the smallest power of two greater than a prior upperbound for the number of clusters. Regarding robustness of inference with respect to K, Theorem 2 suggests that (9) can be sensitive to choice of K for either tree structure if K is small. At the other extreme, for a lopsided tree the exponential bounds in the finite approximation theorems of Ishwaran and James (2002) imply the value of K beyond say 64 seems to affect inference by a trivial amount (assuming K upper bounds the "true" number of clusters). We introduce Theorem 4 to make a similar statement for a balanced tree.

Theorem 4. *Consider a sequence* $\{\mathcal{T}_j = \{A_\varepsilon : \varepsilon \in \{0,1\}^j\} : j \in 0, 1, 2, \ldots\}$ *of measurable partitions of the sample space obtained by splitting every set in the preceding partition into two new sets. Suppose the sample space is a metric space and* $diam(A_\varepsilon) = 2^{-|\varepsilon|}$ *for all* A_ε*, where* $|\varepsilon|$ *is the length of the string* ε *of 0s and 1s. If two tail-free processes agree on all subsets* A_ε *with* $|\varepsilon| \leq M$ *for some positive integer* M*, then the Wasserstein distance between these two processes is bounded above by* 2^{-M+1}*.*

Putting aside computational implications, it is natural to consider the theoretical feasibility of constructing a suitable stick-breaking model on a balanced-tree of infinite depth. We believe this is possible if the splitting variables are specified with extra care. To see this, note that if one simply adopts i.i.d. splitting variables as we have done so far, then the resulting model would have zero prior probability to generate any non-trivial cluster sizes and hence lead to the nonsensical inference that there are *always* as many distinct clusters as there are distinct observations in the data. This is in stark contrast to lopsided-tree stick-breaking models such as those that give rise to the Dirichlet process for which i.i.d. splitting variables suffice. Strategies to regularize infinitely-deep balanced

trees through prior specification have been well studied in the density estimation context for tail-free processes (Freedman, 1963b, 1965; Fabius, 1964b) and in particular for Pólya trees (Lavine, 1992, 1994). In those applications, the prior variance of the splitting variables must decrease sufficiently fast down the tree. Interestingly the consideration for specifying the splitting variable here is just the *opposite*. Prior specifications that produce well-defined densities will in fact again fall into the trap of generating no non-trivial clusters with probability 1. In fact, for the infinite balanced-tree stick-breaking process to work as a prior for cluster weights, the splits must increasingly resemble Bernoulli distributions sufficiently fast down the tree so that most stick pieces essentially remain unchanged in deep enough levels of the tree. We defer the study of the specific conditions for specifying the splitting variables to future work.

An alternative strategy for an infinite balanced tree is to treat the depth of the tree, or more generally the specific topology of the tree, as an unknown object of interest and place a prior on it, resulting in the counterpart of a mixture of finite mixtures (Antoniak, 1974; Richardson and Green, 1997; Nobile, 2004; Gnedin and Pitman, 2006; Gnedin, 2010; De Blasi et al., 2013; Miller and Harrison, 2018; Grazian et al., 2020; Frühwirth-Schnatter et al., 2021). In this case, inferring the number of clusters along with the allocation of the clusters over the leaves becomes in essence a problem of learning an unknown tree topology. Similar strategies have also been employed in the density estimation context based on Pólya trees and related models (Wong and Ma, 2010; Ma, 2017), but the computational task of learning the unknown tree is less demanding for those tree-based density models because in that context not only can the conditional model *given* the tree topology often be analytically integrated out but the Bayes factor between two slightly different tree topologies can often be computed analytically involving only a small subset of the training data. Neither is true in the context of modeling cluster sizes. Computational techniques developed for tree learning and regression in density models, though immediately available in theory, are not practically feasible and what type of computational strategies are effective remains an open and interesting question. We defer further investigation in this direction also to future work. The remainder of this paper will focus on the simple case of a finite fixed K along with splitting variables whose distribution depends on the covariates in a node-specific fashion but does not depend on the tree node it belongs to in any other way.

4 Impact of the tree on posterior uncertainty

This section begins by detailing the posterior computation of the mixture model $f(\cdot; G_{\mathbf{x}}) = \int_\Theta h(\cdot;\theta)\,\mathrm{d}G_{\mathbf{x}}(\theta)$ with a tree stick-breaking prior on the mixing measure $G_{\mathbf{x}}$. If τ is a lopsided tree with $K < \infty$ leaves, we call the resulting mixture a finite lopsided-tree mixture. We similarly define a finite balanced-tree mixture, where K must be a power of 2. This section then presents a case study that analyzes flow-cytometry data and illustrates the impact of tree structure on the posterior uncertainty of covariate effects in mixture weights.

For posterior computation, we generalize Rigon and Durante (2021)'s Gibbs sampler to admit any stick-breaking scheme. Their Gibbs step for the regression coefficients

$\boldsymbol{\gamma}$ relies on Polson et al. (2013)'s Pólya-Gamma data augmentation technique, which allows efficient posterior sampling of a Bayesian logistic regression. For this technique, Polson et al. (2013) carefully construct the Pólya-Gamma family of distributions to allow conditionally conjugate updating for the coefficient parameter and provide a fast, exact way to simulate Pólya-Gamma random variables. They show a posterior sampler for $\boldsymbol{\gamma}$ is obtained by iterating between a step that, conditional on $\boldsymbol{\gamma}$, samples the Pólya-Gamma data and a step that, conditional on the Pólya-Gamma data and regression responses, samples $\boldsymbol{\gamma}$ from a multivariate Gaussian distribution.

Our generalization of this Gibbs step is stated explicitly in Algorithm 1 in the Supplemental Material (we later introduce Algorithm 1 in the main text which is a further generalization that incorporates both fixed and random effects). Given a posterior draw, for all $i = 1, \ldots, n$ let $C_\tau(i)$ be τ's leaf node assigned to observation i. For each internal node ε, the update for the coefficient $\boldsymbol{\gamma}_\varepsilon$ relies on a Bayesian logistic regression with responses $Z_{i\varepsilon}$, defined as the indicator that leaf node $C_\tau(i)$ is a "left descendant" of node ε, for all i corresponding to a descendant node of ε. A "left descendant" of ε is either ε's left child or a descendant of ε's left child, and "right descendant" is defined similarly. In a lopsided tree the left child of any internal node is a leaf (see Figure 1) whereas for any internal node in a balanced tree the number of left descendants equals the number of right descendants. This formulation easily applies to any finite tree stick-breaking scheme. However, we find that training a balanced-tree mixture model takes less time than training its lopsided-tree counterpart for the data sets in Section 4.1 and Section 3 in the Supplementary Material; this observation is supported by the theoretical discussion in Section 2 in the Supplementary Material.

For many applications, it is crucial to include random effects into the splitting-variable model. Without random effects, the model (3) makes the strong assumption that mixture-weight differences between groups is due entirely to differences in covariates. A mixture model that assumes (3) will resolve any large difference in cluster proportions between the two individuals that share the same covariates by breaking up the would-be cluster into many smaller clusters; such a mixture model would thus infer many more clusters than actually exists in the data (as we have seen from experience). Section 4.2 provides details of the specific random effects and the Gibbs sampler for a flow-cytometry case study, but these can easily generalize to other contexts.

4.1 The impact of the tree on posterior inference: a case study

We conduct a case study involving covariate-dependent clustering to demonstrate the influence of the tree structure on the posterior inference of the covariate effects over the cluster sizes. The scientific objective in this case study is to quantify the impact of an African-American female's age on her proportions of T-cell types. Our analysis uses two groups of individuals, younger (aged 18–29) and older (aged 50–65), from a publicly available data set to establish normative ranges (Yi et al., 2019) using the Human Immunology Profiling Consortium T cell immunophenotyping panel (Maecker et al., 2012). This panel has antibodies to cell surface proteins, known as biomarkers, designed to identify CD4+ and CD8+ T cell activation and maturational status but is not specialized to resolve other immune cell types or degrees of immune senescence.

The sample is of all peripheral blood mononuclear cells, and clusters may include non-T cell subsets. In standard analysis, an expert is required to visually identify distinct cell subsets using a sequence of 2D boundaries known as gates. We will instead identify cell subsets using a mixture model.

Our analysis is based on flow cytometry data measured on blood samples from 15 healthy plasma donors, six of which are 18–29 years old and remaining nine of which are 50–65 years old. Each sample can be roughly considered a collection of exchangeable observations (each corresponding to a blood cell) from a seven-dimensional continuous sample space, with each dimension corresponding to the measurement from one marker. These 15 subjects together produce too many viable cells for the model to fit in a reasonable amount of time with Markov chain Monte Carlo. As such, we subset the data in a way that uses all 15 subjects while representing each age group by the same number of viable cells. Hence we subset a total of $n = 403200$ viable cells where each age group is represented by $n/2$ cells. Within each age group, each subject contributes the same number of cells, i.e. the six 18–29 subjects each contributes one-sixth of the $n/2$ cells and the nine 50–65 subjects each contributes one-ninth of the $n/2$ cells.

4.2 A mixed-effects model and a recipe for Bayesian computation

In this study, cells are grouped by the subject they come from and subjects are grouped by the laboratory in which their cells are collected. We account for any resulting group effects in the mixture weights by including random effects in the splitting variables. For each subject s, each splitting variable will include fixed effects $\boldsymbol{\gamma}$ for covariates $\boldsymbol{\psi}(\mathbf{x}_s) = (1, \text{age group of subject } s)^\top$, a random effect u_s for the subject, and a random effect v_j for the subject's batch j (we omit j's dependence on s to avoid visual clutter):

$$\text{logit } V_{s,\varepsilon} = \boldsymbol{\psi}(\mathbf{x}_s)^\top \boldsymbol{\gamma}_\varepsilon + v_{j,\varepsilon} + u_{s,\varepsilon}. \tag{10}$$

The covariates and fixed effects are treated as in (3), and the random effects have priors

$$v_{j,\varepsilon} | \phi_\varepsilon^{(v)} \overset{ind}{\sim} \text{N}(0, (\phi_\varepsilon^{(v)})^{-1}), \quad u_{s,\varepsilon} | \phi_\varepsilon^{(u)} \overset{ind}{\sim} \text{N}(0, (\phi_\varepsilon^{(u)})^{-1}), \quad \phi_\varepsilon^{(v)}, \phi_\varepsilon^{(u)} \overset{ind}{\sim} \text{Gamma}(1, 1).$$

Using Wang and Roy (2018)'s two-block Gibbs sampler, Algorithm 1 extends our Gibbs step from Algorithm 1 in the Supplementary Material to also update the random effects.

To the data we fit a lopsided-tree mixture model and a balanced-tree mixture model each with skew-normal kernels and hyperparameter values of $K = 16$, prior mean $\boldsymbol{\mu}_\gamma = 0_R$, and prior covariance $\boldsymbol{\Sigma}_\gamma = 10 I_R$ where $R = 2$. Each chain burns in 5000 steps before sampling every 10 steps to ultimately keep 1000 posterior draws. Both models use *cross-sample calibration* to account for subject-data being collected in different batches (Gorsky et al., 2023).

4.3 Posterior summaries of the covariate effects on mixing weights

Because our scientific interest concerns the *general* African-American female population rather than only the subjects in our study, population-level parameters such as the

Algorithm 1: Gibbs step to update fixed and random effects under any binary tree.

Result: Update each fixed-effects R-tuple $\boldsymbol{\gamma}_{\boldsymbol{\varepsilon}}$ and random-effects $(J+S)$-tuple $\mathbf{u}_{\boldsymbol{\varepsilon}}$.

for *each internal node* $\boldsymbol{\varepsilon}$ *in binary tree* τ **do**

Let $\mathcal{D}_{\boldsymbol{\varepsilon}} \subseteq \{1, \ldots, n\}$ be the set of indices i where $C_\tau(i)$ is a descendant of node $\boldsymbol{\varepsilon}$;

Update precision parameter $[\phi_{\boldsymbol{\varepsilon}}^{(v)} \mid \cdots] \sim \Gamma(a_{\boldsymbol{\varepsilon}}^{(v)} + 0.5, b_{\boldsymbol{\varepsilon}}^{(v)} + 0.5 \sum_{j \in \mathcal{J}_{\boldsymbol{\varepsilon}}} v_{j,\boldsymbol{\varepsilon}}^2)$, where $\mathcal{J}_{\boldsymbol{\varepsilon}} \subseteq \mathcal{J}$ is the set of batches corresponding to observations in $\mathcal{D}_{\boldsymbol{\varepsilon}}$;

Update precision parameter $[\phi_{\boldsymbol{\varepsilon}}^{(u)} \mid \cdots] \sim \Gamma(a_{\boldsymbol{\varepsilon}}^{(u)} + 0.5, b_{\boldsymbol{\varepsilon}}^{(u)} + 0.5 \sum_{s \in \mathcal{S}_{\boldsymbol{\varepsilon}}} u_{s,\boldsymbol{\varepsilon}}^2)$, where $\mathcal{S}_{\boldsymbol{\varepsilon}} \subseteq \mathcal{S}$ is the set of subjects corresponding to observations in $\mathcal{D}_{\boldsymbol{\varepsilon}}$;

for *every observation* $i \in \mathcal{D}_{\boldsymbol{\varepsilon}}$ **do**

Sample $[\omega_{i\boldsymbol{\varepsilon}} \mid \cdots] \sim$ Pólya-Gamma$(1, |\boldsymbol{\psi}(\mathbf{x}_s)^\top \boldsymbol{\gamma}_\varepsilon + u_{s,\varepsilon} + v_{j,\varepsilon}|)$, for the subject s and batch j associated with i ;

end

Let $\mathcal{L}_{\boldsymbol{\varepsilon}} \subset \mathcal{D}_{\boldsymbol{\varepsilon}}$ be the set of indices corresponding to left descendants of node $\boldsymbol{\varepsilon}$;

Update fixed and random effects by drawing from the full conditional

$$\left[\begin{pmatrix} \boldsymbol{\gamma}_{\boldsymbol{\varepsilon}} \\ \mathbf{u}_{\boldsymbol{\varepsilon}} \end{pmatrix} \mid \cdots \right] \sim \mathrm{N}_{R+J+S}(\boldsymbol{\mu}_{\boldsymbol{\gamma}_{\boldsymbol{\varepsilon}}, \mathbf{u}_{\boldsymbol{\varepsilon}}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_{\boldsymbol{\varepsilon}}, \mathbf{u}_{\boldsymbol{\varepsilon}}})$$

where, letting $\mathbf{M}_{\boldsymbol{\varepsilon}}$ be the $|\mathcal{D}_{\boldsymbol{\varepsilon}}| \times (R+J+S)$ matrix with row entries $(\boldsymbol{\psi}(\mathbf{x}_i)^\top, \mathbf{z}_i^\top)$ for only those $i \in \mathcal{D}_{\boldsymbol{\varepsilon}}$, $\boldsymbol{\kappa}_{\boldsymbol{\varepsilon}} = (1_{i \in \mathcal{L}_{\boldsymbol{\varepsilon}}} - 0.5)_{i \in \mathcal{D}_{\boldsymbol{\varepsilon}}}$, $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon}} = \mathrm{diag}(\omega_{i\boldsymbol{\varepsilon}} : i \in \mathcal{D}_{\boldsymbol{\varepsilon}})$, and $\boldsymbol{\Phi}_{\boldsymbol{\varepsilon}} = (\phi_{\boldsymbol{\varepsilon}}^{(v)} \mathbf{I}_J) \oplus (\phi_{\boldsymbol{\varepsilon}}^{(u)} \mathbf{I}_S)$,

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{\gamma}_{\boldsymbol{\varepsilon}}, \mathbf{u}_{\boldsymbol{\varepsilon}}} &= \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_{\boldsymbol{\varepsilon}}, \mathbf{u}_{\boldsymbol{\varepsilon}}} \left[\mathbf{M}_{\boldsymbol{\varepsilon}}^\top \boldsymbol{\kappa}_{\boldsymbol{\varepsilon}} + \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\gamma}} \\ 0_{(S+J)\times 1} \end{pmatrix} \right], \\ \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_{\boldsymbol{\varepsilon}}, \mathbf{u}_{\boldsymbol{\varepsilon}}}^{-1} &= \mathbf{M}_{\boldsymbol{\varepsilon}}^\top \boldsymbol{\Omega}_{\boldsymbol{\varepsilon}} \mathbf{M}_{\boldsymbol{\varepsilon}} + (\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \oplus \boldsymbol{\Phi}_{\boldsymbol{\varepsilon}}). \end{aligned}$$

end

effects of age should retain nontrivial posterior uncertainties due to the relatively small number of subjects in the study. Hence the effects of the study's subjects on the mixture weights should be captured by the mixture model but the covariate effects on the cluster sizes should only be quantified by the difference in mixture weights across covariates at the *population* level, not for the specific individuals in the study. We compute this *population-level* weight difference by first fitting a mixture model with the mixed-effects model (10). Using this fitted mixture model, for each subject s and internal node $\boldsymbol{\varepsilon}$ we then compute a *population-level* splitting variable $V_{s,\varepsilon}^*$ using the posterior draws of only the covariate fixed effects (i.e., we omit the posterior draws of the random effects u_s

and v_j in this computation):

$$\text{logit } (V^*_{s,\varepsilon})^{(d)} = \boldsymbol{\psi}(x_s)^\top \boldsymbol{\gamma}^{(d)}_{\varepsilon} + 0v^{(d)}_{j,\varepsilon} + 0u^{(d)}_{s,\varepsilon}, \quad d = 1, \ldots, 1000,$$

where d indexes the 1000 posterior draws; this population-level splitting variable hence depends on the subject s only through its age group. For each age group x, we convert x's splitting variables into *population-level* mixture weights $\mathbf{W}^*_x := (W^*_{x,\boldsymbol{\varepsilon}_1}, \ldots, W^*_{x,\boldsymbol{\varepsilon}_K}) \in \Delta_K$ using the usual tree-dependent conversion (1). We can then measure the age effect on the mixture weights at the population level as

$$\mathbf{W}^*_{older} - \mathbf{W}^*_{younger}. \tag{11}$$

4.4 The influence of the tree structure on posterior inference

After fitting the mixed-effects mixture model under both a lopsided tree and a balanced tree, we compare the inference between the corresponding posterior distributions. The two models infer very similar *sample-level* cluster sizes and shapes, see Figure 3, which indicates a degree of robustness in the inference and creates an approximate one-to-one correspondence between most of the lopsided-tree clusters and most of the balanced-tree clusters. This robustness in inferring *sample-level* cluster sizes and shapes, i.e., those for the specific samples collected in the study, are expected for flow cytometry given the massive number of cells in each sample. The difference in the lopsided-tree mixture and balanced-tree mixture is expected to be apparent on the *population-level* cluster sizes given the limited number of samples and the substantial sample-to-sample variability typically observed in flow cytometry.

Interesting differences show up in the posterior distributions of population-level parameters such as the covariate effects on cluster sizes. Given the approximate correspondence in the sample-level clusters, we can directly compare the inference on covariate effects (which are on the *population-level* cluster sizes) between the two models. Figure 4 shows both sets of credible intervals of (11) on each cluster in order of cluster size, which allows easier visual comparison between corresponding lopsided-tree and balanced-tree clusters. Even for cluster pairs with very similar size and shape, some of the corresponding credible intervals are noticeably different. In particular, consider the three cluster pairs whose 90% credible intervals are most away from zero. For one of these three cluster pairs (whose size is roughly 7.1% and whose biomarker values—i.e., the plotted values in Figure 3—align with activated monocytes), the two credible intervals are similar in length but differ in location. For the remaining two cluster pairs (whose sizes are roughly 15.5% and 2.8% and whose biomarker values respectively align with those of CD4+ naïve T cells and resting monocytes), the balanced-tree credible intervals have smaller skewness and spread than do the lopsided-tree credible intervals. Regarding CD4+ naïve T cells, CD4+ T cells generally coordinate the overall immune response by the secretion of signaling molecules, and naïve T cells are cells which have never previously encountered antigen but might become memory cells after encountering antigen. Hence it is highly plausible that naïve T cells would decrease with age as their production slows down markedly after adolescence (Mogilenko et al., 2022), as suggested by the two credible intervals of this cluster.

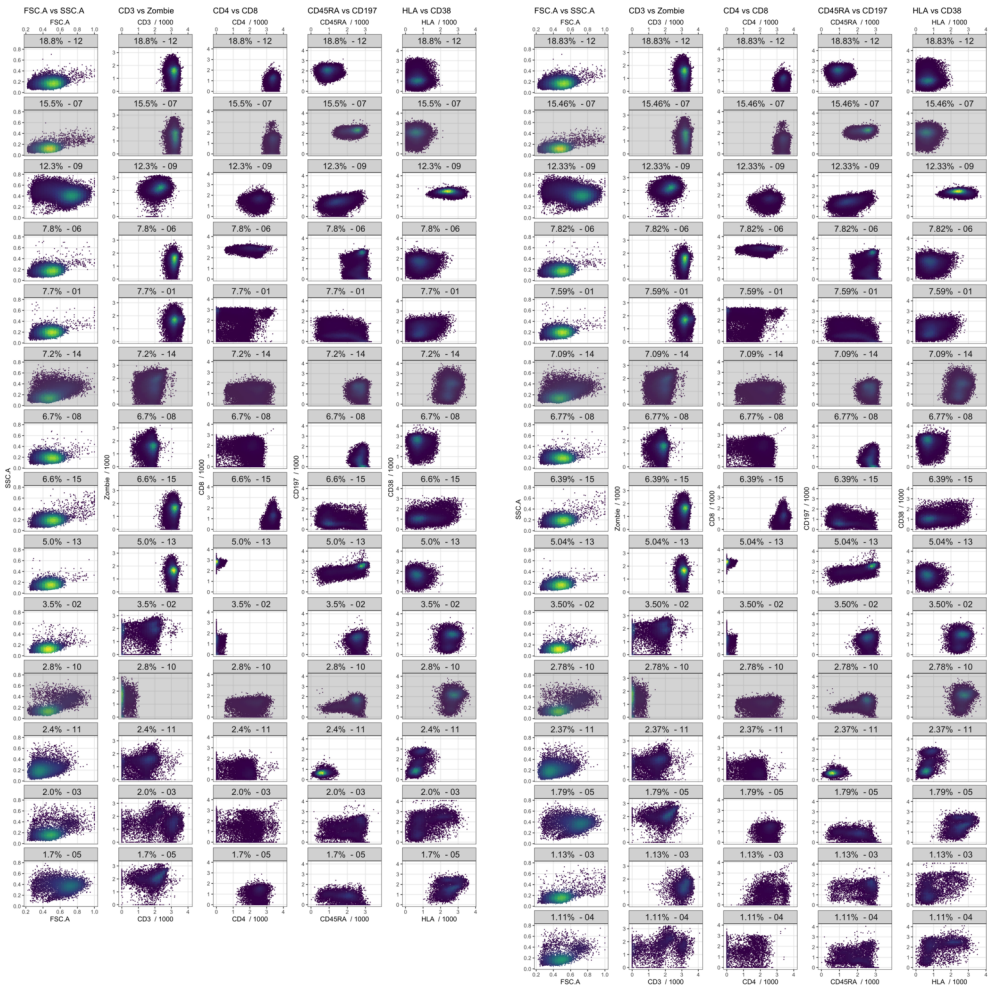

Figure 3: Clusters inferred by the lopsided-tree mixture model (left) and the balanced-tree mixture model (right). The highlighted panels indicate the clusters discussed in the text.

Why these credible intervals are so different is difficult to pinpoint exactly due to the many moving parts in a mixture model, but we can offer a few conjectures. For example, consider the prior assumption of mean-zero covariate effects in the splitting variables. For either tree, it pushes clusters with strong covariate effects toward leaf nodes graphically near each other because such a configuration would allow more splitting variables to maintain (near) zero-valued covariate effects. For this same reason, it also pushes such clusters away from the tree's root. But for the lopsided-tree mixture model, prior stochastic ordering in the mixture weights pushes larger clusters *toward* the root and hence competes with the previous mechanism over large clusters with strong

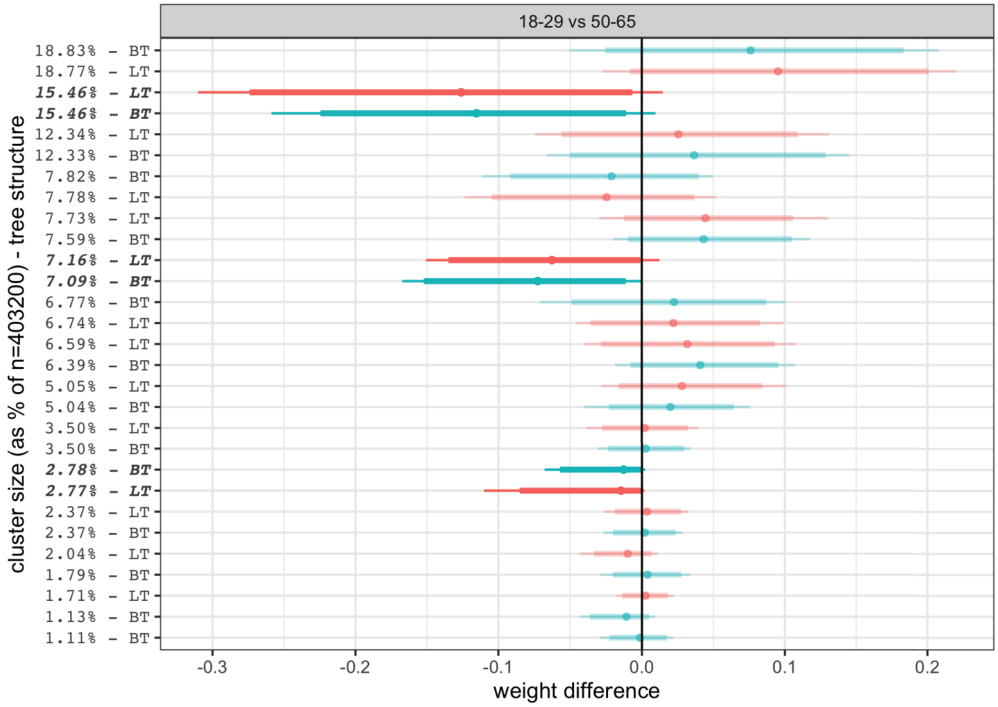

Figure 4: Credible intervals of (11) for both the fitted lopsided-tree mixture and balanced-tree mixture model. Thicker bar indicates 90% quantile and thinner bar indicates 95% quantile. The bolded italic labels indicate the clusters discussed in the text.

covariate effects (such as the cluster whose size is roughly 15.5%). In addition, for a mixture model with K components, the lopsided-tree weights are on average a product of $\approx K/2$ splitting variables whereas each balanced-tree weight is a product of $\log_2 K$ splitting variables, which implies the lopsided-tree weights are a more interdependent function of the splitting variables than are the balanced-tree weights. The exact way these mechanisms affect the quality of the posterior inference is unclear, but the seemingly fewer moving parts and more efficient model representation of the mixture weights in the balanced-tree mixture model are appealing.

In addition to offering the above conjectures, we also conduct the following experiment to gauge the trustworthiness of the lopsided-tree and balanced-tree inferences in the above GRIFOLS data analysis. Rather than use completely synthetic data, which would omit much of the complexity inherent in flow cytometry data, we instead modify the GRIFOLS data by injecting a covariate effect of known size into one of the clusters in Figure 3. Because we perform many runs of this experiment, for time and memory considerations we down sample from 403200 cells to 100800 cells and have each chain burn in 5000 steps before sampling every 20 steps to ultimately keep 500 posterior

	length median (sd)		**p truth in**		**p subset**	
CI level	LT	BT	LT	BT	pLTinBT	pBTinLT
0.50	0.036 (0.017)	0.033 (0.018)	0.595	0.550	0.080	0.165
0.80	0.073 (0.032)	0.066 (0.034)	0.740	0.790	0.100	0.235
0.90	0.098 (0.041)	0.086 (0.044)	0.910	0.925	0.120	0.325
0.95	0.125 (0.050)	0.106 (0.053)	0.940	0.935	0.130	0.380

Table 1: Summary of the credible intervals for the cluster labeled 2.8%/2.78% in Figure 3 in the GRIFOLS simulation study detailed in Section 4.4. LT stands for lopsided-tree and BT for balanced tree. pLTinBT stands for "proportion of LT intervals that lie within BT intervals" and pBTinLT for "proportion of BT intervals that lie within LT intervals".

draws. However, this downsampling results in some clusters not being inferred (e.g., they might be absorbed by one or many other clusters) and hence not every cluster in Figure 3 corresponds to a cluster in the downsampled regime, and vice versa. Furthermore, we would like the to-be-enhanced cluster to not consist of too many cells so that the induced perturbation to the data is as local as possible. Hence, we choose the cluster labelled 2.8%/2.78% in Figure 3, which, after downsampling, is still clearly identified and separated from other clusters.

Given this cluster, we aim to assess the quality of the inference to be performed. Although we know how many of the cluster's cells come from younger and older subjects, the "correct" value of the mixture-weight difference (11) is still unclear due to the various complex properties of the data and imposed model, such as misshapen clusters and subject random effects. Hence our data modification consists of enhancing the existing covariate effect by reassigning 99% of the older-subject cells to younger subjects. Given this artificially large imbalance, the difference (11) for this modified cluster should be close to the inferred cluster size divided by the total number of cells.

Given the modified data, we fit a lopsided-tree model and a balanced-tree model for 200 different starting seeds and compare the behavior of the credible intervals of the difference (11) for this cluster. Rather than simply show all 400 credible intervals here (such plots are relegated to Figure 7 in the Supplemental Material), we summarize their behavior by first describing some desirable properties for them to have. One such property is whether it contains the truth, which for our purposes we set to be the cluster size. Also, an interval that contains the truth is more informative if it is short. In light of these properties, Table 1 shows the median interval length, the proportion of intervals that contain the truth, the proportion of lopsided-tree intervals that are a subset of the corresponding balanced-tree interval, and vice versa. Here (and in Figure 5) we see that the lopsided-tree intervals tend to be longer than their balanced-tree counterparts. Furthermore, there are many more starting seeds where the balanced-tree interval is a subset of the lopsided-tree interval than vice versa. We note that under some random seeds (14 seeds under the lopsided-tree model and 13 under the balanced-tree model), the posterior distribution fails to correctly identify the cluster and hence produces a tiny and incorrect credible interval centered at zero which (slightly) deflates the median interval length and proportion that contain the truth, but we find that this phenomenon

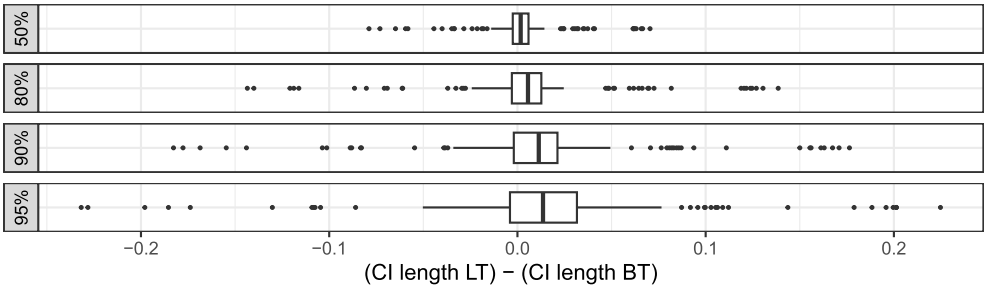

Figure 5: Each panel summarizes the distribution of 200 credible-interval length differences, where each difference is the difference between the credible-interval length for the lopsided-tree mixture model and that for the balanced-tree mixture model for some starting seed. Panel strips indicate credible-interval level.

influences the two different models to similar degrees and the general conclusions still hold even if these intervals are omitted. To avoid the suspicion of cherry picking, we did not exclude these seeds from our results.

5 Discussion

In addition to the comparisons made in this paper thus far between lopsided and balanced tree stick-breaking models, Section 4 of the Supplementary Materials presents yet another way the two tree structures behave differently from each other, this time regarding the phenomenon of component *label switching* (see Stephens, 2000; Papastamoulis, 2016, for an overview). Label switching has been observed to occur frequently in Markov chain Monte Carlo algorithms for mixture models that use lopsided-tree stick-breaking models. Despite label switching being seen as a way to achieve convergence to the true posterior, such convergence can often become an unattainable goal to pursue. The aforementioned section presents a simulation study where label switching occurs more often and to greater effect in a lopsided-tree mixture model than in a corresponding balanced-tree mixture model, and discusses reasons behind this behavior difference.

One limitation of the balanced tree model investigated in this work is that we assumed the tree is truncated at a fixed maximum depth. With a sufficiently large depth this causes little practical restriction in applications, but it does preclude theoretical analysis of such models as fully nonparametric processes with infinite-dimensional parameters. In particular, Pitman-Yor processes belong to the class of Gibbs-type priors (Gnedin and Pitman, 2006), which have explicit forms for the predictive and clustering structure. Such an extension to balanced trees is of future interest. Another area of future work is to explore posterior-inference algorithms that avoid finite-dimensional approximations, still allow covariates to be incorporated into mixture weights, and are computationally tractable. A Pólya-urn sampler is one such method of avoiding finite-dimensional approximations, but it is not clear if such an approach could be extended

to incorporate covariate dependence. On the other hand, Foti and Williamson (2012) introduces a slice sampler for the dependent Dirichlet process which perhaps could be modified for a dependent balanced-tree mixture model. Alternatively, we can still restrict K to be finite but to also place a prior distribution on it as done in the mixture of finite mixtures e.g. Gnedin and Pitman (2006); De Blasi et al. (2013); Miller and Harrison (2018); Frühwirth-Schnatter et al. (2021).

Another line of future work involves quantifying the prior dependence between random measures beyond the pairwise linear correlation used in Sections 3 and 3.2. Though our paper introduces dependence between random measures by stick breaking, much work has been done for the alternative approach of generating mixing weights using completely random measures (Kingman, 1967) in which dependence can be induced at the level of the underlying Poisson random measures. Under this framework, Catalano et al. (2021, 2023) provide a general approach of quantifying the dependence between groups of random measures by measuring the dependence as distance from exchangeability using Wasserstein distance. If we place our generalized tree stick-breaking prior under the framework of completely random measures, it might be possible to glean additional insight into its dependence structure.

Acknowledgments

We thank two anonymous referees, the associate editor, and the editor for their helpful suggestions and comments. We gratefully acknowledge Scott White for assistance in processing the flow cytometry data used in this study.

Supplementary Material

Supplementary material for "A tree perspective on stick-breaking models in covariate-dependent mixtures" (DOI: 10.1214/24-BA1462SUPP; .pdf).
Proofs. Proofs of Theorems 1, 2, 3, 4, and Corollary 2.1.
Numerical experiment. Numerical experiment regarding Corollary 2.1(c).
Computational cost of Gibbs step. Discussion on the computational cost of Gibbs step from the fixed-effects version of Algorithm 1.
Impact of the tree on Markov chain Monte Carlo sampling. Discussion on the impact of the tree on Markov chain Monte Carlo sampling regarding component label switching.
Plot of 400 credible intervals. Plot of the 400 credible intervals in the numerical study from Section 4.4.

References

Albert, J. H. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association*, 88(422): 669–679. MR1224394. 1

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *The annals of statistics*, 1152–1174. MR0365969. 12

Argiento, R. and De Iorio, M. (2022). "Is infinity that far? A Bayesian nonparametric perspective of finite mixture models." *The Annals of Statistics*, 50(5): 2641–2663. MR4505373. doi: https://doi.org/10.1214/22-aos2201. 10

Ascolani, F., Franzolini, B., Lijoi, A., and Prünster, I. (2023). "Nonparametric priors with full-range borrowing of information." *Biometrika*, asad063. URL https://doi.org/10.1093/biomet/asad063 9

Bassetti, F., Casarin, R., and Leisen, F. (2014). "Beta-product dependent Pitman–Yor processes for Bayesian inference." *Journal of Econometrics*, 180(1): 49–72. MR3188911. doi: https://doi.org/10.1016/j.jeconom.2014.01.007. 2

Bishop, C. M. and Svensén, M. (2002). "Bayesian hierarchical mixtures of experts." In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, 57–64. 5

Broderick, T., Jordan, M. I., and Pitman, J. (2012). "Beta Processes, Stick-Breaking and Power Laws." *Bayesian Analysis*, 7(2): 439 – 476. URL https://doi.org/10.1214/12-BA715 MR2934958. doi: https://doi.org/10.1214/12-BA715. 2

Catalano, M., Lavenant, H., Lijoi, A., and Prünster, I. (2023). "A Wasserstein index of dependence for random measures." *Journal of the American Statistical Association*, 1–11. 21

Catalano, M., Lijoi, A., and Prünster, I. (2021). "Measuring dependence in the Wasserstein distance for Bayesian nonparametric models." *The Annals of Statistics*, 49(5): 2916–2947. MR4338898. doi: https://doi.org/10.1214/21-aos2065. 21

Cipolli, W. and Hanson, T. (2017). "Computationally tractable approximate and smoothed Polya trees." *Statistics and Computing*, 27(1): 39–51. MR3598907. doi: https://doi.org/10.1007/s11222-016-9652-3. 5

De Blasi, P., Lijoi, A., and Prünster, I. (2013). "An asymptotic analysis of a class of discrete nonparametric priors." *Statistica Sinica*, 1299–1321. MR3114715. 12, 21

de Finetti, B. (1938). "Sur la condition d'Equivalence partielle." *Actualités Scientifiques et Industrielles*, (739): 5–18. 2

Diaconis, P. (1988). "Recent progress on de Finetti's notions of exchangeability." *Bayesian statistics*, 3(111-125): 13–14. MR1008047. 2

Diaconis, P. and Freedman, D. (1980). *De Finetti's generalizations of exchangeability*. Stanford University. Department of Statistics. MR0587993. 2

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). "Generalized spatial Dirichlet process models." *Biometrika*, 94(4): 809–825. MR2416794. doi: https://doi.org/10.1093/biomet/asm071. 2

Dunson, D. B. and Park, J.-H. (2008). "Kernel stick-breaking processes." *Biometrika*, 95(2): 307–323. MR2521586. doi: https://doi.org/10.1093/biomet/asn012. 2

Fabius, J. (1964a). "Asymptotic behavior of Bayes' estimates." *The Annals of Mathemat-*

ical Statistics, 34(4): 846–856. MR0162325. doi: https://doi.org/10.1214/aoms/1177703584. 7

Fabius, J. (1964b). "Asymptotic behavior of Bayes' estimates." *The Annals of Mathematical Statistics*, 846–856. 12

Favaro, S., Lijoi, A., Nava, C. R., Nipoti, B., Pruenster, I., Teh, Y. W., et al. (2016). "On the stick-breaking representation for homogeneous NRMIs." *Bayesian Analysis*, 11: 697–724. MR3498043. doi: https://doi.org/10.1214/15-BA964. 2

Favaro, S., Lijoi, A., and Prünster, I. (2012). "On the stick-breaking representation of normalized inverse Gaussian priors." *Biometrika*, 99(3): 663–674. MR2966776. doi: https://doi.org/10.1093/biomet/ass023. 2

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 209–230. MR0350949. 2, 4

Foti, N. and Williamson, S. (2012). "Slice sampling normalized kernel-weighted completely random measure mixture models." *Advances in Neural Information Processing Systems*, 25. 21

Fox, E. B. and Jordan, M. I. (2014). "Mixed Membership Models for Time Series." In Airoldi, E. M., Blei, D., Erosheva, E. A., and Fienberg, S. E. (eds.), *Handbook of Mixed Membership Models and Their Applications*, chapter 24. New York: Chapman and Hall/CRC. MR3380022. 2

Freedman, D. A. (1963a). "On the asymptotic behavior of Bayes' estimates in the discrete case." *The Annals of Mathematical Statistics*, 34(4): 1386–1403. MR0158483. doi: https://doi.org/10.1214/aoms/1177703871. 7

Freedman, D. A. (1963b). "On the asymptotic behavior of Bayes' estimates in the discrete case." *The Annals of Mathematical Statistics*, 34(4): 1386–1403. 12

Freedman, D. A. (1965). "On the asymptotic behavior of Bayes estimates in the discrete case II." *The Annals of Mathematical Statistics*, 36(2): 454–456. MR0174146. doi: https://doi.org/10.1214/aoms/1177700155. 12

Frühwirth-Schnatter, S., Malsiner-Walli, G., and Grün, B. (2021). "Generalized mixtures of finite mixtures and telescoping sampling." *Bayesian Analysis*, 16(4): 1279–1307. MR4381135. doi: https://doi.org/10.1214/21-BA1294. 12, 21

Fuentes-Garcia, R., Mena, R. H., and Walker, S. G. (2010). "A new Bayesian nonparametric mixture model." *Communications in Statistics—Simulation and Computation®*, 39(4): 669–682. MR2785657. doi: https://doi.org/10.1080/03610910903580963. 2

Ghahramani, Z. and Griffiths, T. (2005). "Infinite latent feature models and the Indian buffet process." In Weiss, Y., Schölkopf, B., and Platt, J. (eds.), *Advances in Neural Information Processing Systems*, volume 18. MIT Press. URL https://proceedings.neurips.cc/paper_files/paper/2005/file/2ef35a8b78b572a47f56846acbeef5d3-Paper.pdf MR2441315. 2

Ghahramani, Z., Jordan, M., and Adams, R. P. (2010). "Tree-structured stick breaking for hierarchical data." *Advances in neural information processing systems*, 23. 4

Gil-Leyva, M. F. and Mena, R. H. (2023). "Stick-breaking processes with exchangeable length variables." *Journal of the American Statistical Association*, 118(541): 537–550. MR4571140. doi: https://doi.org/10.1080/01621459.2021.1941054. 2

Gil–Leyva, M. F., Mena, R. H., and Nicoleris, T. (2020). "Beta-Binomial stick-breaking non-parametric prior." *Electronic Journal of Statistics*, 14(1): 1479 – 1507. URL https://doi.org/10.1214/20-EJS1694 MR4082475. doi: https://doi.org/10.1214/20-EJS1694. 2

Gnedin, A. (2010). "A Species Sampling Model with Finitely Many Types." *Electronic Communications in Probability*, 15(none): 79 – 88. URL https://doi.org/10.1214/ECP.v15-1532 MR2606505. doi: https://doi.org/10.1214/ECP.v15-1532. 12

Gnedin, A. and Pitman, J. (2006). "Exchangeable Gibbs partitions and Stirling triangles." *Journal of Mathematical sciences*, 138: 5674–5685. MR2160320. doi: https://doi.org/10.1007/s10958-006-0335-z. 12, 20, 21

Gorsky, S., Chan, C., and Ma, L. (2023). "Coarsened Mixtures of Hierarchical Skew Normal Kernels for Flow and Mass Cytometry Analyses." *Bayesian Analysis*, 1(1): 1–25. MR4729434. doi: https://doi.org/10.1214/22-ba1356. 14

Grazian, C. (2024). "Spatio-temporal stick-breaking process." *Bayesian Analysis*, 1(1): 1–32. MR2717937. 2

Grazian, C., Villa, C., and Liseo, B. (2020). "On a loss-based prior for the number of components in mixture models." *Statistics & Probability Letters*, 158: 108656. MR4025678. doi: https://doi.org/10.1016/j.spl.2019.108656. 12

Griffin, J. E. and Steel, M. F. (2011). "Stick-breaking autoregressive processes." *Journal of econometrics*, 162(2): 383–396. MR2795625. doi: https://doi.org/10.1016/j.jeconom.2011.03.001. 2

Griffin, J. E. and Steel, M. J. (2006). "Order-based dependent Dirichlet processes." *Journal of the American statistical Association*, 101(473): 179–194. MR2268037. doi: https://doi.org/10.1198/016214505000000727. 2

Hjort, N. L. (1990). "Nonparametric Bayes estimators based on beta processes in models for life history data." *the Annals of Statistics*, 1259–1294. MR1062708. doi: https://doi.org/10.1214/aos/1176347749. 2

Horiguchi, A., Chan, C., and Ma, L. (2024). "Supplementary Material for "A Tree Perspective on Stick-Breaking Models in Covariate-Dependent Mixtures"." *Bayesian Analysis*. doi: https://doi.org/10.1214/24-BA1462SUPP. 10

Hossain, M. M., Lawson, A. B., Cai, B., Choi, J., Liu, J., and Kirby, R. S. (2013). "Space-time stick-breaking processes for small area disease cluster estimation." *Environmental and ecological statistics*, 20: 91–107. MR3039588. doi: https://doi.org/10.1007/s10651-012-0209-0. 2

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96(453): 161–173. MR1952729. doi: https://doi.org/10.1198/016214501750332758. 1

Ishwaran, H. and James, L. F. (2002). "Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information." *Journal of Computational and Graphical statistics*, 11(3): 508–532. MR1938445. doi: https://doi.org/10.1198/106186002411. 11

James, L. F., Lijoi, A., and Prünster, I. (2006). "Conjugacy as a distinctive feature of the Dirichlet process." *Scandinavian Journal of Statistics*, 33(1): 105–120. MR2255112. doi: https://doi.org/10.1111/j.1467-9469.2005.00486.x. 8

Jara, A. and Hanson, T. E. (2011). "A class of mixtures of dependent tail-free processes." *Biometrika*, 98(3): 553–566. MR2836406. doi: https://doi.org/10.1093/biomet/asq082. 4

Jordan, M. I. and Jacobs, R. A. (1994). "Hierarchical mixtures of experts and the EM algorithm." *Neural computation*, 6(2): 181–214. 5

Kingman, J. (1967). "Completely random measures." *Pacific Journal of Mathematics*, 21(1): 59–78. MR0210185. 21

Lavine, M. (1992). "Some aspects of Polya tree distributions for statistical modelling." *The Annals of Statistics*, 20(3): 1222–1235. MR1186248. doi: https://doi.org/10.1214/aos/1176348767. 4, 12

Lavine, M. (1994). "More aspects of Polya tree distributions for statistical modelling." *The Annals of Statistics*, 22(3): 1161–1176. MR1311970. doi: https://doi.org/10.1214/aos/1176325623. 12

Lijoi, A., Mena, R. H., and Prünster, I. (2005). "Hierarchical mixture modeling with normalized inverse-Gaussian priors." *Journal of the American Statistical Association*, 100(472): 1278–1291. MR2236441. doi: https://doi.org/10.1198/016214505000000132. 2

Lijoi, A., Mena, R. H., and Prünster, I. (2007). "Controlling the reinforcement in Bayesian non-parametric mixture models." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(4): 715–740. MR2370077. doi: https://doi.org/10.1111/j.1467-9868.2007.00609.x. 2

Linderman, S., Johnson, M. J., and Adams, R. P. (2015). "Dependent Multinomial Models Made Easy: Stick-Breaking with the Polya-gamma Augmentation." In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/07a4e20a7bbeeb7a736682b26b16ebe8-Paper.pdf 2

Ma, L. (2017). "Adaptive shrinkage in Pólya tree type models." *Bayesian Analysis*, 12(3): 779–805. MR3655876. doi: https://doi.org/10.1214/16-BA1021. 12

MacEachern, S. N. (2000). "Dependent dirichlet processes." Technical report, Department of Statistics, The Ohio State University. 7

Maecker, H. T., McCoy, J. P., and Nussenblatt, R. (2012). "Standardizing immunophenotyping for the Human Immunology Project." *Nat Rev Immunol*, 12(3): 191–200. Maecker, Holden T McCoy, J Philip Nussenblatt, Robert eng Z99 HL999999/Intramural NIH HHS/ ZIC HL005905-02/Intramural NIH HHS/ ZIC HL005905-03/Intramural NIH HHS/ ZIC HL005905-04/Intramural NIH HHS/ Review England 2012/02/22 Nat Rev Immunol. 2012 Feb 17;12(3):191-200. doi: 10.1038/nri3158. URL https://www.ncbi.nlm.nih.gov/pubmed/22343568 13

Maudlin, R. D., Sudderth, W. D., and Williams S. C. (1992). "Polya trees and random distributions." *The Annals of Statistics*, 20(3): 1203-1221. MR1186247. doi: https://doi.org/10.1214/aos/1176348766. 4

Miller, J. W. and Harrison, M. T. (2018). "Mixture models with a prior on the number of components." *Journal of the American Statistical Association*, 113(521): 340–356. MR3803469. doi: https://doi.org/10.1080/01621459.2016.1255636. 12, 21

Mogilenko, D. A., Shchukina, I., and Artyomov, M. N. (2022). "Immune ageing at single-cell resolution." *Nat Rev Immunol*, 22(8): 484–498. Mogilenko, Denis A Shchukina, Irina Artyomov, Maxim N eng Research Support, Non-U.S. Gov't Review England 2021/11/25 Nat Rev Immunol. 2022 Aug;22(8):484-498. doi: 10.1038/s41577-021-00646-4. Epub 2021 Nov 23. URL https://www.ncbi.nlm.nih.gov/pubmed/34815556 16

Nobile, A. (2004). "On the posterior distribution of the number of components in a finite mixture." *The Annals of Statistics*, 32(5): 2044 – 2073. URL https://doi.org/10.1214/009053604000000788 MR2102502. doi: https://doi.org/10.1214/009053604000000788. 12

Paisley, J., Blei, D., and Jordan, M. (2012). "Stick-breaking beta processes and the Poisson process." In *Artificial Intelligence and Statistics*, 850–858. PMLR. 2

Paisley, J. W., Zaas, A. K., Woods, C. W., Ginsburg, G. S., and Carin, L. (2010). "A stick-breaking construction of the beta process." In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 847–854. 2

Papastamoulis, P. (2016). "label.switching: An R Package for Dealing with the Label Switching Problem in MCMC Outputs." *Journal of Statistical Software, Code Snippets*, 69(1): 1–24. URL https://www.jstatsoft.org/index.php/jss/article/view/v069c01 20

Peng, F., Jacobs, R. A., and Tanner, M. A. (1996). "Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition." *Journal of the American Statistical Association*, 91(435): 953–960. 5

Petralia, F., Rao, V., and Dunson, D. (2012). "Repulsive mixtures." *Advances in neural information processing systems*, 25. 9

Pitman, J. (1996). "Random discrete distributions invariant under size-biased permu-

tation." *Advances in Applied Probability*, 28(2): 525–539. MR1387889. doi: https://doi.org/10.2307/1428070. 2

Pitman, J. and Yor, M. (1997). "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator." *The Annals of Probability*, 855–900. MR1434129. doi: https://doi.org/10.1214/aop/1024404422. 2

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian inference for logistic models using Pólya–Gamma latent variables." *Journal of the American statistical Association*, 108(504): 1339–1349. MR3174712. doi: https://doi.org/10.1080/01621459.2013.829001. 1, 13

Quintana, F. A., Müller, P., Jara, A., and MacEachern, S. N. (2022). "The dependent Dirichlet process and related models." *Statistical Science*, 37(1): 24–41. MR4371095. doi: https://doi.org/10.1214/20-sts819. 7

Regazzini, E., Lijoi, A., and Prünster, I. (2003). "Distributional results for means of normalized random measures with independent increments." *The Annals of Statistics*, 31(2): 560–585. MR1983542. doi: https://doi.org/10.1214/aos/1051027881. 2

Reich, B. J. and Fuentes, M. (2007). "A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields." *The Annals of Applied Statistics*, 1(1): 249 – 264. URL https://doi.org/10.1214/07-AOAS108 MR2393850. doi: https://doi.org/10.1214/07-AOAS108. 2

Ren, L., Du, L., Carin, L., and Dunson, D. B. (2011). "Logistic stick-breaking process." *Journal of Machine Learning Research*, 12(1): 203–239. MR2773552. 5

Richardson, S. and Green, P. J. (1997). "On Bayesian analysis of mixtures with an unknown number of components (with discussion)." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(4): 731–792. MR1483213. doi: https://doi.org/10.1111/1467-9868.00095. 12

Rigon, T. and Durante, D. (2021). "Tractable Bayesian density regression via logit stick-breaking priors." *Journal of Statistical Planning and Inference*, 211: 131–142. MR4117446. doi: https://doi.org/10.1016/j.jspi.2020.05.009. 1, 12

Rodríguez, A. and Dunson, D. B. (2011). "Nonparametric Bayesian models through probit stick-breaking processes." *Bayesian Analysis (Online)*, 6(1): 145–178. MR2781811. doi: https://doi.org/10.1214/11-BA605. 1, 4, 9

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4(2): 639–650. MR1309433. 1, 2

Stefanucci, M. and Canale, A. (2021). "Multiscale stick-breaking mixture models." *Statistics and Computing*, 31(2): 1–13. MR4206653. doi: https://doi.org/10.1007/s11222-020-09991-1. 5

Stephens, M. (2000). "Dealing with label switching in mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4): 795–809. MR1796293. doi: https://doi.org/10.1111/1467-9868.00265. 20

Teh, Y. W., Grür, D., and Ghahramani, Z. (2007). "Stick-breaking construction for the Indian buffet process." In *Artificial intelligence and statistics*, 556–563. PMLR. 2

Thibaux, R. and Jordan, M. I. (2007). "Hierarchical beta processes and the Indian buffet process." In *Artificial intelligence and statistics*, 564–571. PMLR. 2

Wang, X. and Roy, V. (2018). "Analysis of the Pólya-Gamma block Gibbs sampler for Bayesian logistic linear mixed models." *Statistics & Probability Letters*, 137: 251–256. MR3776224. doi: https://doi.org/10.1016/j.spl.2018.02.003. 14

Wong, W. H. and Ma, L. (2010). "Optional Pólya tree and Bayesian inference." *The Annals of Statistics*, 38(3): 1433–1459. MR2662348. doi: https://doi.org/10.1214/09-AOS755. 12

Yi, J. S., Rosa-Bray, M., Staats, J., Zakroysky, P., Chan, C., Russo, M. A., Dumbauld, C., White, S., Gierman, T., Weinhold, K. J., and Guptill, J. T. (2019). "Establishment of normative ranges of the healthy human immune system with comprehensive polychromatic flow cytometry profiling." *PLoS One*, 14(12): e0225512. 13