Tree stick-breaking priors for covariate-dependent mixture models

Akira Horiguchi\textsuperscript{1}, Cliburn Chan\textsuperscript{2,1}, and Li Ma*\textsuperscript{1,2}

\textsuperscript{1}Department of Statistical Science, Duke University
\textit{214 Old Chemistry, Box 90251, Durham, NC 27708}
\textsuperscript{2}Department of Biostatistics and Bioinformatics, Duke University
\textit{2424 Erwin Road, Suite 1102, Hock Plaza, Box 2721, Durham, NC 27710}

August 5, 2022

Abstract

Stick-breaking priors are often adopted in Bayesian nonparametric mixture models for generating mixing weights. When covariates influence the sizes of clusters, stick-breaking mixtures can leverage various computational techniques for binary regression to ease posterior computation. Existing stick-breaking priors are typically based on continually breaking a single remaining piece of the unit stick. We demonstrate that this “single-piece” scheme can induce three highly undesirable behaviors; these behaviors are circumvented by our proposed model which continually breaks off all remaining pieces of a unit stick while keeping posterior computation essentially identical. Specifically, the new model provides more flexibility in setting cross-covariate prior correlation among the generated random measures, mitigates the impact of component label switching when posterior simulation is performed using Markov chain Monte Carlo, and removes the imposed artificial decay of posterior uncertainty on the mixing weights according to when the weight is “broken off” the unit stick. Unlike previous works on covariate-dependent mixtures, which focus on estimating covariate-dependent distributions, we instead focus on inferring the effects of individual covariates on the mixture weights in a fashion similar to classical regression analysis, and propose a new class of posterior predictives for summarizing covariate effects.

\textit{Keywords:} Discrete random measure, Bayesian nonparametric, tail-free process, clustering analysis, flow cytometry.

\*Send correspondence to li.ma@duke.edu
1 Introduction

Mixture models are a popular approach to clustering analysis. Stick-breaking (SB) processes are widely used in the Bayesian literature as priors on mixture models (Sethuraman 1994; Ishwaran and James 2001) largely due to their tractable posterior computation. SB priors are particularly well-suited when covariates influence cluster sizes, as these priors can achieve efficient posterior computation by incorporating computational strategies involving latent variable representation and data augmentation techniques originally designed for binary outcome regression models. Notable examples include Rodriguez and Dunson (2011)’s probit SB whose Markov chain Monte Carlo (MCMC) algorithm relies on Albert and Chib (1993)’s truncated Gaussian data augmentation for probit regression, and Rigon and Durante (2021)’s logit SB whose various posterior sampling methods rely on Polson et al. (2013)’s Polya-gamma data augmentation for logistic regression. SB priors have also been applied in generalizations of mixture models such as mixed-membership models (Linderman et al. 2015).

Covariate-dependent clustering applications range widely; one motivating example for this paper is flow cytometry (FCM), a technique whose uses include identifying cells according to their type (e.g. B-cell, T-cell), in which the lack of existing nonparametric mixture models based on SB priors become apparent. In FCM, a question of interest is to see how cell populations change across covariate values. For example, one might wish to detect cell-type population differences between healthy and sick patients. Very often, the covariate values influence relatively small cell clusters.

Existing SB priors are based on continually breaking a single remaining piece of the unit stick. As we shall demonstrate through both theoretical investigation and numerical experiments, this “single-piece” scheme can induce three highly undesirable properties that
severely impair posterior inference in the presence of covariates.

The first is that under common default hyperprior values, “single-piece” stick breaking induces a rather strong positive prior correlation in the random measures over covariate values, which can cause excessive borrowing of information across the covariates even when the corresponding cluster sizes vary sharply across covariate values, thereby deprecating the clustering at each covariate value. More concretely, Corollary 2.1 in Section 2.3 states that if the stick breaks have prior mean 0.5 with non-negative pairwise correlations, then the prior cross-predictor correlation between random measures is lower-bounded by 1/3. Rodriguez and Dunson (2011) noticed this lower bound and explained it as “a consequence of our use of a common set of atoms at every [covariate]; 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.” We show that the large size of this baseline correlation is determined largely by the particular way the weights are constructed from the stick breaks.

Another undesirable property involves the long-standing problem of component label switching. In a Bayesian mixture model with $K$ components under a prior with exchangeable mixture components, the prior and the resulting posterior cannot distinguish the $K!$ permutations of the mixture component labels and hence are symmetric between the corresponding $K!$ regions of the parameter space $\Theta$. If a posterior simulation stays in a single region, the simulation’s outputs can be interpreted, but the more a simulation travels between these regions, the more difficult the interpretation becomes. Section 3 contains a simulation study in which the MCMC simulation travels between eight of these regions even after a long burn-in. We find that the ease with which this MCMC simulation travels between so many regions is not inherent to SB priors; rather, it is specific to the “single-piece” approach to generating weights. We also find that the simulation’s interpretation
cannot in general be "fully restored" by Cron and West (2011)'s relabeling algorithm.

A third undesirable behavior is that a weight’s variance tends to decrease according to the order in which the weight was created during stick breaking. To intuit why this phenomenon occurs, note that any weight \( W_k = V_k \prod_{l<k}(1-V_l) \) must follow

\[
\text{Var}W_k + [\mathbb{E}W_k]^2 = \mathbb{E}W_k^2 = \mathbb{E}V_k^2 \mathbb{E}R_{k-1}^2
\]

where \( R_{k-1} := \prod_{l<k}(1-V_l) \) is the length of the remaining right piece after \( k-1 \) stick breaks. Each \( \text{Var}W_k \) is bounded above by \( \mathbb{E}R_{k-1}^2 = \prod_{l<k} \mathbb{E}(1-V_l)^2 \) which is a product of \( k-1 \) quantities strictly between zero and one. So in the likely scenario that \( \mathbb{E}R_{k-1}^2 \) shrinks to zero (or to a very small positive number) as \( k \) increases, then \( \text{Var}W_k \) must also. Although this artificial variance decay is easily unnoticed in our simulation study in Section 3 where all of the weight variances are small (likely because we are able to perfectly specify the weights model), it is strikingly present in our analysis of real-world FCM data in Section 4 (see Figure 8a), where a few of the weights have large variances (likely because it is often impossible to impose a perfectly specified model on real-world data). Seeing as this variance decay is a product of this “single-piece” stick-breaking construction rather than of any particular qualities of the data, it becomes difficult to trust the uncertainty quantification of these weights.

We aim to construct new SB priors for covariate-dependent mixture modeling while maintaining its desirable computational properties. To this end, we adopt a dyadic tree perspective on stick-breaking models. Specifically, starting from the whole unit stick as the root of a tree, we view each piece of the stick that arises during the stick-breaking as a node in a bifurcating tree. The traditional “single-piece” stick breaking then corresponds to a special tree topology—which is “lopsided” on one side. See Figure 1. The issues of
Figure 1: 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. Thus, the $K - 1$ stick breaks distribute the initial unit mass to the tree’s $K$ leaf nodes which represent the scheme’s $K$ weights. As an example, the filled leaf in each tree represents weight $W_9$ whose value is determined by the stick breaks whose representative internal nodes are the filled leaf’s ancestors. This weight is a product of nine stick breaks in the lopsided-tree scheme but only four stick breaks in the balanced-tree scheme.

Traditional stick-breaking processes become clear under the tree perspective. We show that the lopsided structure lies behind the aforementioned issues of SB models.

We instead propose a new class of SB priors based on an alternative tree topology—namely a “balanced” one (shown in Figure 1), which corresponds to the continual breaking of all remaining pieces of the unit stick at each stage of the stick breaking. Under this scheme, each weight becomes a multiplicative product of the same (small) number of stick breaks. No weight is a product of just one stick break, which allows this alternative scheme to offer a much wider range of prior cross-covariate correlation in the resulting random measures. Furthermore, this tree topology mitigates the impact of label switching as shown empirically in Section 3.

Posterior computation under our new covariate-dependent SB model remains essentially identical to that of existing SB mixtures. In particular, this scheme can easily adapt exist-
ing (and possibly future) algorithms for tractable posterior inference, e.g., Rodriguez and Dunson (2011)'s Markov chain Monte Carlo (MCMC) algorithm for probit SB and Rigon and Durante (2021)'s MCMC, variational Bayes, and expectation-maximization algorithms for logistic SB. In this work, we shall demonstrate using Pólya-gamma augmentation for existing logistic SB priors under the new balanced-tree SB mixture models. We note that our new SB prior for mixing weights along a balanced bifurcating tree is a tail-free process, and in the presence of covariates, it becomes a covariate-dependent tail-free model introduced by Jara and Hanson (2011). According 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 and Stefanucci and Canale (2021) whose multiscale SB method generalizes our approach to generating weights. Both works consider only univariate response data with no covariate dependence, and neither work provides theoretical or methodological justification for using their respective method over an approach based on traditional SB.

Previous works (e.g., Rigon and Durante 2021) on covariate-dependent random measures have mostly evaluated the “goodness” of the models based on posterior predictive densities. While this might be appropriate when density estimation or prediction is the main objective, they are not suitable for applications such as flow cytometry in which the key scientific interest lies in quantifying the impact of a covariate (e.g., treatments or patient characteristics) on the clusters (e.g., immune cell types). For the latter purpose, we instead utilize the posterior distributions of covariate effects, in the spirit of classical regression analysis, to examine model performance. One complication is that under stick-breaking models (both existing and our new model), covariate effects are parameterized on the stick-breaking proportions rather than on the cluster weights, which is typically of
direct scientific interest. As such, we propose a strategy to translate posterior covariate effects on stick-breaking proportions onto that on the weights. While not the major focus of this paper, our strategy presents a practical solution to practitioners who wish to more intuitively interpret covariates.

We end the introduction by reviewing several works that also utilize tree structures to construct SB models in ways different from ours. To model hierarchical data, Ghahramani et al. (2010) interleave two “lopsided” SB processes to allow a wide range of latent tree structures (e.g. nodes with any number of children) to be inferred, and Dubey et al. (2014)’s extension incorporates time information into weights. In a different context, Nassar et al. (2018) introduce what is essentially a generalization of our “balanced” SB approach to generating covariate-dependent weights and even utilize PG augmentation in their Gibbs sampler, but do so to model nonlinear dynamical systems and hence do not consider its use in mixture models.

The paper is structured as follows. Section 2 introduces our SB scheme without covariate dependence. Section 2.3 extends the model to incorporate covariate dependence. Section 2.6 considers posterior computation. Section 3 showcases a numerical example where tree structure exacerbates label switching on posterior inference for weights. Section 4 applies our method to the GRIFOLS FCM data set. Section 5 provides concluding remarks.
2 Method

2.1 Basic setup and notation

We start by introducing our tree-based stick-breaking construction in the absence of covariates. Throughout we assume independent and identically distributed (i.i.d.) observations $y_1, \ldots, y_n \in \mathbb{R}^d$ from a sampling density of the following mixture form

$$f(\cdot; G) = \int_{\Theta} h(\cdot; \theta) \, dG(\theta).$$  \hspace{1cm} (1)

where $h: \mathbb{R}^d \times \Theta \rightarrow \mathbb{R}^+$ is a parametric mixing kernel density, e.g., Gaussian density $h(y; \theta) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{(y - \mu)^2}{2\sigma^2}\right\}$ with $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+.$

We consider a mixing measure $G$ on $\Theta$ in (1) with SB priors. Thus each realization of $G$ is a discrete measure of the following form with probability one:

$$G = \sum_{k=1}^{\infty} W_k \delta_{\theta_k} \quad \theta_k \overset{\text{ind}}{\sim} G_0$$

where $G_0$ is a base measure on $\Theta$ and the nonnegative weights $W_k$, which are generated through breaking a unit stick, are independent of the atoms $\theta_k$. This turns the mixture density (1) into $f(\cdot; G) = \sum_{k=1}^{\infty} W_k h(\cdot; \theta_k)$.

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

2.2 Stick-breaking priors from a tree perspective

A weight-generation process aims to produce nonnegative quantities that sum to unity. A SB 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 the traditional SB scheme, after 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 stick of unit-length 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 general $k \in \mathbb{N}$, the $k^{th}$ stick break at location $V_k \in [0, 1]$ produces a piece that breaks off with length $\left(\prod_{l=1}^{k-1}(1 - V_l)\right)V_k$, which sets the value of weight $W_k$; the remaining stick then breaks at location $V_{k+1} \in [0, 1]$. If the process could either continue infinitely or terminate after a preset finite number of breaks, in which case the final remaining stick sets the last weight.

The SB 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, then this piece corresponds to a leaf node, i.e., a node with no children. If a piece is further broken, then we use two children nodes to represent the two resulting pieces. The traditional SB scheme’s identified with a bifurcating tree (see the top right of Figure 2), which is the most “lopsided” as only no nodes in the tree can have both children that are further divided. It is the deepest tree possible for generating a given number of pieces. We shall thus refer to the traditional SB strategy as the lopsided-tree (LT) SB.

Given this tree perspective of stick-breaking, it is natural to consider alternative SB schemes that correspond to different tree topologies than the lopsided one. To this end, it
will be convenient to index each binary tree node \( \varepsilon \) by a finite string of 0s and 1s, where each digit indicates whether a node along the path from root to \( \varepsilon \) is a left (0) or right (1) child of its parent. For notational ease, the root node corresponds to the empty string \( \emptyset \). Examples are shown in Figure 2. In general, if a node \( \varepsilon \) is at level \( m \) of the tree (the root node is at level 0, its two children are at level 1, and so on), then \( \varepsilon \) is the \( m \)-length string \( \varepsilon_1 \varepsilon_2 \cdots \varepsilon_m \), where each \( \varepsilon_i \in \{0, 1\} \) and the string \( \varepsilon \varepsilon' \) denotes the concatenation of any two finite strings \( \varepsilon \) and \( \varepsilon' \). The set of all finite strings (including \( \emptyset \)) of 0s and 1s is 
\[ E^* = \bigcup_{m=0}^{\infty} \{0, 1\}^m \]

With this notation, we can now formally introduce a SB strategies based on general tree topologies. In particular, we give particular attention to the SB 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 bottom of Figure 2. We refer to this particular SB scheme as a balanced-tree (BT) SB. Opposite to the lopsided-tree scheme, the balanced-tree scheme results in the most shallow tree topology needed to generate any given number of weights.

Specifically, in this new scheme, after each stick break both pieces break again until the total number of breaks exceeds a preset threshold (we argue against allowing infinitely many breaks in Section 2.2.1). The unit-length stick \( I_\emptyset \) breaks according to \( V_\emptyset \) so that the left piece \( I_0 \) and right piece \( I_1 \) have lengths

\[ |I_0| = |I_\emptyset|V_\emptyset = V_\emptyset, \quad |I_1| = |I_\emptyset|(1 - V_\emptyset) = 1 - V_\emptyset. \]

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

\[ |I_{00}| = V_\emptyset V_0, \quad |I_{01}| = V_\emptyset(1 - V_0), \quad |I_{10}| = (1 - V_\emptyset)V_1, \quad |I_{11}| = (1 - V_\emptyset)(1 - V_1). \]
These four pieces $I_{00}, I_{01}, I_{10},$ and $I_{11}$ 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 we preset the number of such collections of stick breaks to $m \in \mathbb{N}$, this recursive procedure produces $2^m$ leaf nodes each with a stick piece. The lengths of these pieces define the desired weights: $W_\varepsilon := |I_\varepsilon|$ for each $\varepsilon \in \{0, 1\}^m$.

The value of any SB weight at a node $\varepsilon = \varepsilon_1 \cdots \varepsilon_m$ can be expressed as

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

(2)

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

$$\text{supp}(V_\varepsilon) = [0, 1] \text{ for all } \varepsilon \in E^*$$

(3)

$$V_\emptyset \perp \{V_0, V_1\} \perp \{V_{00}, V_{01}, V_{10}, V_{11}\} \perp \cdots .$$

(4)

With this tree view, we can now define the tree stick-breaking (treeSB) class of priors, which includes traditional SB but also admits more general tree structures. We say that a probability measure $G$ is endowed with a treeSB prior with parameters $G_0, \{F_\varepsilon\},$ and binary tree structure $\tau$, denoted by $G \sim \text{treeSB}(G_0, \{F_\varepsilon\}, \tau)$, if it can be constructed as

$$G = \sum_{\varepsilon \in B(\tau)} W_\varepsilon \delta_{\theta_\varepsilon}, \quad \theta_\varepsilon \overset{\text{ind}}{\sim} G_0$$

(5)

where the set $B(\tau) \subset E^*$ indexes $\tau$’s leaf nodes, and the random weights $\{W_\varepsilon : \varepsilon \in B(\tau)\}$
are constructed according to (2) where the distributions $F_\varepsilon$ follow the full-support and tail-free conditions (3) and (4). As with traditional SB, this turns the mixture density (1) into $f(\cdot; G) = \sum_{\varepsilon \in B(\tau)} W_\varepsilon h(\cdot; \theta_\varepsilon)$.

2.2.1 Choice of the number of leaves

In real-life applications, a practitioner will use only finitely many stick breaks (i.e. use $\tau$ with finite depth), in which case the treeSB construction (5) is entirely valid. Special consideration must be taken if $\tau$ has infinite depth. In the following we avoid these technical considerations by assuming $\tau$ has large finite depth, which suffices for practical uses of the model. We note that a similar simplification is often adopted through truncation of traditional SB priors (Ishwaran and James, 2001).
How should a practitioner set the number of leaves $K$? For LT, the finite approximation theorems of Ishwaran and James (2002) imply that the value of $K$ beyond say 64 seems to affect inference by at most a trivial amount. This statement is supported by the moment expressions (8) and (9) to be introduced in Theorem 1 and by Theorem 2. For BT, however, Theorem 2 implies the possibility of setting $K$ so large that the variance of the random measure $G$ shrinks to almost zero. This danger is based on identically distributed splitting variables and hence might possibly be avoided by choosing splitting variables whose mean and/or variance increase with tree level. In any case, a rule of thumb that still allows “room to breathe” is to set $K$ to be the smallest power of two greater than or equal to twice the prior expected number of clusters.

2.3 Tree stick-breaking with covariates

Following the well-known strategy of MacEachern (2000), we next extend the covariate-independent treeSB prior (5) by replacing the splitting variables $\{V_\varepsilon\}_{\varepsilon \in E^\ast}$ in (2) with a sequence of stochastic processes $\{V_{x,\varepsilon}: x \in \mathcal{X}\}_{\varepsilon \in E^\ast}$, where $\mathcal{X}$ is a set of covariates and each splitting variable $V_{x,\varepsilon}$ has distribution $F_{x,\varepsilon}$. The resulting random probability measure $G_x \sim treeSB(G_0, \{F_{x,\varepsilon}\}, \tau)$ now depends on $x$ through its weights:

$$G_x = \sum_{\varepsilon \in B(\tau)} W_{x,\varepsilon} \delta_{\theta_{\varepsilon}}, \quad \theta_{\varepsilon} \sim G_0.$$  

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 SB. In particular, let us consider the logit approach due to the particular computational convenience for posterior inference that follows from the Pólya-Gamma augmentation technique, to be detailed in Section 2.6. (We note that the theoretical properties
we establish in Section 2.4 do not assume this model choice and apply more generally.)

Specifically, we adopt the following logit-normal model on each splitting variable:

\[ V_{x,e} = \text{logistic}(\eta_{x,e}), \quad \eta_{x,e} = \psi(x)^T \gamma_e, \quad \gamma_e \sim N_R(\mu_\gamma, \Sigma_\gamma) \]  

(7)

with hyperparameters \( \mu_\gamma \) and \( \Sigma_\gamma \), where \( \text{logistic}(z) = \exp[z/(1 + z)] \) and \( \eta_{x,e} \) is a linear combination of selected functions of the covariates \( \psi(x) = [\psi_1(x), \ldots, \psi_R(x)]^T \). Thus, \( \{\eta_{x,e}: x \in \mathcal{X}\} \) is a Gaussian process (GP) with mean \( \mu_{x,e} = \psi(x)^T \mu_\gamma \) and covariance \( \text{Cov}(\eta_{x,e}, \eta_{x',e}) = \psi(x)^T \Sigma_\gamma \psi(x') \). The remainder of the paper (with the exception of Section 2.4) assumes the logit-normal prior (7).

### 2.4 Moments of random measures

Here we provide the mean, variance, and correlation properties of the random measure \( G_x \) and compare these properties between the LT and BT SB schemes. Proofs can be found in the Supplementary Materials. We note that logit-normal prior (7) satisfies the conditions that Theorems 1 and 2 place on the splitting variable distributions.

**Theorem 1.** Suppose for any (vector of) covariates \( x \) that the atomic random measure

\[ G_x = \sum_{k=1}^{K} W_{x,k} \delta_{\theta_k}, \quad K \in \{1, 2, \ldots, \infty\} \]

on \( \Theta \) is constructed by drawing each \( \theta_k \) independently (of each other and of \( x \)) from some base measure \( G_0 \) on \( \Theta \), and drawing a weight vector \( (W_{x,1}, \ldots, W_{x,K}) \) according to some distribution (that may or may not depend on \( x \)) on the probability simplex \( S_K \).
For any measurable sets $A, A' \in \mathcal{B}$ and covariates $x, x'$, we have

\begin{align*}
\mathbb{E}G_x(A) &= G_0(A) \\
\text{Var}G_x(A) &= [G_0(A) - G_0^2(A)] a_{x,x} \\
\text{Cov}(G_x(A), G_x(A')) &= [G_0(A \cap A') - G_0(A)G_0(A')] a_{x,x} \\
\text{Cov}(G_x(A), G_{x'}(A)) &= [G_0(A) - G_0^2(A)] a_{x,x'}
\end{align*}

where $a_{x,x'} = \sum_{k=1}^{K} E W_{x,k} W_{x',k}$. If also the quantities $G_0(A), [1 - G_0(A)], G_0(A'),$ and $[1 - G_0(A')]$ are all nonzero, then

\begin{align*}
corr(G_x(A), G_x(A')) &= \frac{G_0(A \cap A') - G_0(A)G_0(A')}{\sqrt{G_0(A)[1 - G_0(A)]G_0(A')[1 - G_0(A')]} } \\
corr(G_x(A), G_{x'}(A)) &= a_{x,x'}/\sqrt{a_{x,x}a_{x',x'}}.
\end{align*}

Theorem 1 states that each of these various moments can be factorized into a function of the base measure $G_0$ and a function of the quantity $a_{x,x'}$. As functions of only $G_0$, the mean (8) implies $G_0$ can be viewed as the mean of the random measure $G_x$ while the correlation (12) does not depend on tree depth. On the other hand (and of greater interest), the cross-covariate correlation (13) depends not on the base measure $G_0$ but rather on $a_{x,x'}$, which Theorem 2 expresses as a function of $\tau$ and the splitting variables' mean and cross-covariate covariance. In particular, this theorem states that $a_{x,x'}$ for BT approaches zero as $K \to \infty$ whereas $a_{x,x'}$ for LT approaches a positive limit, which means LT induces a baseline cross-covariate covariance value (11) between random measures that does not vanish as the number of weights approaches infinity. Corollary 2.1 posits that these statements also apply to their cross-covariate correlation counterparts (13) if also
assumed are splitting variables with mean 1/2 and nonnegative cross-covariate correlation.

**Theorem 2.** Suppose weights \((W_{x,\varepsilon} : \varepsilon \in B(\tau))\) are constructed by SB according to tree topology \(\tau\). Also assume that at any \(x\) the splitting variables \(\{V_{x,\varepsilon} : \varepsilon \in E^*\}\) are identically distributed. Let \(a_{x,x'} = \sum_{\varepsilon \in B(\tau)} E W_{x,\varepsilon} W_{x',\varepsilon}\) and \(K = |B(\tau)| \in \mathbb{N}\). If \(\tau\) is a LT, then

\[
a_{x,x'} = \frac{EV_xV_{x'}(1 - [1 - EV_x - EV_{x'} + EV_xV_{x'}]^K)}{EV_x + EV_{x'} - EV_xV_{x'}}
\]

which equals \(EV_xV_{x'}\) if \(K = 1\) and approaches \(EV_xV_{x'}(EV_x + EV_{x'} - EV_xV_{x'})^{-1}\) as \(K \to \infty\).

If instead \(\tau\) is a BT and \(m = \log_2 K\), then

\[
a_{x,x'} = (1 - EV_x - EV_{x'} + 2EV_xV_{x'})^m
\]

which approaches zero as \(m \to \infty\).

**Corollary 2.1** (Lower bounds). If \(EV_x = 1/2\) for any \(x\) and \(\text{Cov}_V(x, x') \geq 0\) for any \(x, x' \in X\), lower bounds for \([13]\) are 1/3 for LT and \(2^{-m}\) for BT. These lower bounds are strict if \(\text{Var}_V(x) > 0\) for all \(x \in X\).

Finally, the following theorem establishes the important property that the random measure \(G_x\) changes smoothly with respect to \(x\). This property, expressed as the cross-covariate correlation \([13]\) between random measures approaching unity as \(x' \to x\), requires a moment condition which is satisfied if the stochastic process \(\{V_x : x \in X\}\) is second-order stationary.

**Theorem 3** (Smoothness). Given the assumptions in Theorem 2, if weights are either finite LT or BT and if \(EV_xV_{x'} \to EV_x^2\) as \(x' \to x\), then \([13]\) \(\to 1\).
2.5 Numerical illustration on prior cross-covariate correlation

Here we wish to explore the impact of the cross-covariate correlation between splitting variables on the cross-covariate correlation between random measures: given covariates $x$ and $x'$, how does $\text{corr}(V_x, V_{x'})$ affect $\text{corr}(G_x(A), G_{x'}(A))$ 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 levers that influence $\text{corr}(V_x, V_{x'})$ by making the following simplifying assumptions:

$$
\mu_\gamma = 0, \quad \Sigma_\gamma = \text{diag}(\sigma_1^2, \sigma_2^2), \quad \psi(x) = (1, 0)^T \text{ and } \psi(x') = (1, 1)^T
$$

where $\sigma_1^2 > 0$. 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 $\text{corr}(V_x, V_{x'}) = \text{corr}(\eta_x, \eta_{x'}) = (1 + \tilde{\sigma}_2^2)^{-1/2}$, 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 $\text{corr}(V_x, V_{x'})$ can be achieved by using the appropriate $\tilde{\sigma}_2^2$ value if $x \neq x'$. Similarly, these assumptions reduce $\text{corr}(G_x(A), G_{x'}(A))$ (whose expression is provided by (13) and Theorem 2) to a function of just $\sigma_1^2$, $\sigma_2^2$, $K$, and $\tau$.

Figure 3 explores the behavior of the correlation $\text{corr}(G_x(A), G_{x'}(A))$, which by Corollary 2.1 has lower bounds of $1/3$ and $1/K$ for, respectively, LT and BT. 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 LT scheme always imposes a nontrivial baseline correlation between random measures while the BT scheme can achieve both large and small correlation values, which provides more flexibility in setting prior correlation values.
2.6 Posterior sampling

This section details the posterior computation of the mixture model

\[ f(\cdot; G_x) = \int_{\Theta} h(\cdot; \theta) \, dG_x(\theta). \]  

(14)

with a treeSB prior on the mixing measure \( G_x \). If \( \tau \) is LT’s tree structure with \( K < \infty \) leaves, we call the resulting mixture a finite LT mixture (LTM). We similarly define a finite BT mixture (BTM), where \( K \) must be a power of 2.

For posterior computation, we generalize Rigon and Durante (2021)’s Gibbs sampler to admit any SB scheme. Their Gibbs step for \( \gamma \) relies on Polson et al. (2013)’s Pólya–Gamma (PG) data augmentation technique, which allows efficient posterior sampling of a Bayesian logistic regression. For this technique, Polson et al. (2013) carefully construct the PG family of distributions to allow conditionally conjugate updating for the coefficient parameter and provide a fast, exact way to simulate PG random variables. Let \( Z_i \) be the number of
Algorithm 1: Gibbs step to update logit-normal coefficient under any binary tree.

Result: Update $\gamma_\varepsilon$ for each internal node $\varepsilon$ in binary tree $\tau$.

for each internal node $\varepsilon$ in $\tau$ do
  Let $S_\varepsilon$ be the set of indices $i \in [n]$ where $C_\tau(i)$ is a descendant of node $\varepsilon$;
  for every $i \in S_\varepsilon$ do
    Sample Pólya-gamma data $[\omega_i | \cdots] \sim PG(1, x_i^T \gamma_\varepsilon)$;
  end
  Let $L_\varepsilon \subset S_\varepsilon$ be the set of indices corresponding to “left descendants” of node $\varepsilon$;
  Update $\gamma_\varepsilon$ by drawing from full conditional $[\gamma_\varepsilon | \cdots] \sim N_R(\mu_\gamma, \Sigma_\gamma)$
    where
    \[
    \mu_\gamma = \Sigma_\gamma \left[ X_\varepsilon^T \kappa_\varepsilon + \Sigma_\gamma^{-1} \mu_\gamma \right] \quad \text{with } \kappa_\varepsilon = (1_{i \in L_\varepsilon} - 0.5)_{i \in S_\varepsilon},
    \]
    \[
    \Sigma_\gamma = \left[ X_\varepsilon^T \Omega X_\varepsilon + \Sigma_\gamma^{-1} \right]^{-1} \quad \text{with } \Omega = \text{diag}(\omega_1, \ldots, \omega_{|S_\varepsilon|}),
    \]
    and $X_\varepsilon$ is the $|S_\varepsilon| \times R$ matrix with row entries $x_i^T$ for only those $i \in S_\varepsilon$.
end

successes, $n_i$ the number of trials, and $x_i$ be a $p$-tuple of regressors for observation $i \in [n]$.

Also suppose $Z_i \sim \text{Binom}(n_i, [1 + \exp(-x_i^T \gamma)]^{-1})$ and $\gamma \sim N(b, B)$. They show that a posterior sampler for $\gamma$ is obtained by iterating the two steps

$$
(\omega_i | \gamma) \sim PG(n_i, x_i^T \gamma) \\
(\gamma | \omega, Z_1, \ldots, Z_n) \sim N(m_\omega, V_\omega)
$$

where

$$
\begin{align*}
m_\omega &= V_\omega \left[ X^T \kappa + B^{-1} b \right] \quad \text{with } \kappa^T = (Z_1 - n_1/2, \ldots, Z_n - n_n/2), \\
V_\omega &= \left[ X^T \Omega X + B^{-1} \right]^{-1} \quad \text{with } \Omega = \text{diag}(\omega_1, \ldots, \omega_n),
\end{align*}
$$

and $X$ is the matrix with row entries $x_i^T$.

Our generalization of this Gibbs step is stated explicitly in Algorithm 1. Given a
posterior draw, for all $i \in [n] := \{1, \ldots, n\}$, let $C_\tau(i)$ be $\tau$’s leaf node assigned to observation $i$. For each internal node $\varepsilon$, the update for the coefficient $\gamma_\varepsilon$ relies on a Bayesian logistic regression with responses

$$Z_{i\varepsilon} = \begin{cases} 
1 & \text{if leaf node } C_\tau(i) \text{ is a “left descendant” of node } \varepsilon \\
0 & \text{if leaf node } C_\tau(i) \text{ is a “right descendant” of node } \varepsilon
\end{cases}$$

for all $i$ corresponding to a descendant node of $\varepsilon$. Here, a “left descendant” of $\varepsilon$ is either the left child or a descendant of the left child of $\varepsilon$, and “right descendant” is defined similarly. For example, in LT the left child of any internal node is a leaf (see Figure 2) whereas for any internal node in BT the number of left descendant nodes equals the number of right descendant nodes. This formulation easily applies to any finite treeSB scheme.

Next we compare between LT and BT the computational cost of this Gibbs step for some arbitrarily fixed MCMC iteration. In this case, each tree has the same number of internal nodes and hence their corresponding mixture models perform the same number of regressions. From this viewpoint, the computational cost should be the same between the two SB schemes. However, we find from experience that for the data sets in Sections 3 and 4 training a BTM model takes less time than training its LTM counterpart, which is supported by the theoretical discussion in Section 6.5 in the Supplementary Materials.

3 Simulation study

This example explores how tree structure affects MCMC fitting. We create two data sets $\mathcal{D}_{indep}$ and $\mathcal{D}_{dep}$, whose cluster proportions respectively do not and do depend on the observations’ covariate values. The two data sets share the same $n = 8000$ observations
Figure 4: Data generated by 20 skew-normal distributions. The boxed numbers are the number of observations generated from each distribution. The triangles indicate the observations whose covariate assignment differs between the two data sets.

shown in Figure 4. Each of the 20 skew-normal (SN) distributions generates some number of observations; these cluster counts are listed as the 20-tuple

$$n = 8 \cdot (200, 170, 130, 100, 80, 70, 50, 30, 28, 22, 20, 18, 16, 14, 12, 10, 9, 8, 7, 6).$$

For each $i \in \mathbb{N}_{20}$, each data set assigns a covariate $x_i$ in the space $\mathcal{X} := \{1\} \times \{0, 1\}^3$ to observation $i$. Let $n_{x_{indep}}$ and $n_{x_{dep}}$ be the cluster counts for all observations assigned the covariates $x$ by, respectively, $\mathcal{D}_{indep}$ and $\mathcal{D}_{dep}$. The covariate-specific cluster counts $n_{x_{indep}}$ are the same for all covariates, i.e. $n_{x_{indep}} = n/8$ for all $x \in \mathcal{X}$. Thus, we dub the data set $\mathcal{D}_{indep}$ “covariate-independent.” In contrast, the “covariate-dependent” data set $\mathcal{D}_{dep}$ has six clusters whose covariate-specific counts do depend on $x$:

$$n_{x_{dep}} = (n/8) + 20(2x_1 - 1)(e_8 - e_3) + 20(2x_2 - 1)(e_10 - e_6) + 20(2x_3 - 1)(e_9 - e_7),$$
where $\mathbf{x} = (1, x_1, x_2, x_3)$ and each $\mathbf{e}_j$ is the $j$th standard basis vector.

We fit a dependent LTM model and a dependent BTM model to each data set using the Gibbs sampler in Section 2.6. Each model uses $K = 32$, prior mean $\mu_\gamma = \mathbf{0}_p$, and prior covariance $\Sigma_\gamma = 10\mathbf{I}_p$. Thus, four MCMC chains are run. Each chain burns in 100,000 steps before sampling every 100 steps to ultimately keep 10,000 posterior draws.

Figure 5 shows for each chain the Jaccard distance $[1 \text{ } - \text{ } J(c^{(i)}, c_0)] \in [0, 1]$ between the chain’s clustering $c^{(i)}$ at the $i$th posterior draw (for all $i \in N_{10000}$) and the true clustering $c_0$ (Jaccard, 1912). These distances provide a sense of a chain’s clustering performance and its mixing behavior. For each data set, LT and BT have roughly the same median Jaccard distance, but the mean Jaccard distance is slightly larger for LT. The LT chains also seem to get stuck in local modes for longer than the BT chains do.

What accounts for this difference in clustering behavior? Figure 6 shows selected mixture-weight inference for $\mathbf{x} = 1010$. The results are similar for the other covariate values; “full” versions of these plots are in the Supplementary Materials. In Figure 6a, s
(LT, D^{indep}) panel, none of the credible intervals (CIs) are centered around the second largest weight value of 0.17 (this is more easily seen in Figure 6b), which might foster the belief that this LTM model fails to capture this weight, but upon closer inspection, the heights of the eight curiously wide 95% CIs are roughly that same value. This observation together with Figure 6c imply that the eight wide LT CIs are a result of label switching in the MCMC chain. Similarly, Figure 6d implies that the wide CIs in Figure 6a’s (BT, D^{dep}) panel also happen because of label switching. Indeed, Figure 7 shows that these wide CIs disappear and the data’s weights are captured if we “resolve” the label switching by performing the following post-hoc sorting of weights: for each posterior draw $j$ we rearrange (in decreasing order) the elements of the chain’s corresponding weight vector $w^{(j)}$, then for each newly rearranged weight index $\tilde{k}$ we compute 95% CIs for the values in the set $\{w^{(j)}_{1010, \tilde{k}} : j \in \mathbb{N}_{10000}\}$.

Although this post-hoc sorting seems to resolve the wide CIs for this example, the procedure requires several assumptions to work and hence cannot be applied generally. Until a general solution is created, it seems prudent to “nip the problem in the bud” by mitigating the impact of label switching on inference. But why does label switching seem to affect LT’s inference more than BT’s? We conjecture that the behavior difference stems from two tree-related mechanisms.

First, consider how much friction there is to label switch between two leaf nodes (for either tree structure). The amount of friction seems to increase with the number of splitting variables affected by the proposed switch. If the two leaves are graphically far apart, the proposed switch would affect many splitting variables and hence be unlikely to occur. Under this reasoning, switching is most likely to occur between two leaves that are (graphically) close to each other. In a BT, this means two leaves that share the same parent. In a LT,
Figure 6: Selected mixture-weight inference for $x = 1010$.

this means two leaves separated by at most one tree level. For conceptual simplicity, we consider only such pairs and call these leaves adjacent to each other. Note that LT has
Figure 7: Inferred weights that are sorted before taking pointwise 95% CIs. Points indicate the data’s true weight values.

$K$ adjacent pairs whereas BT has $K/2$. Thus, LT has more adjacent pairs from which switching can occur.

Second, a label switch between two adjacent LT leaves can create a “chain reaction” of label switches either up or down the entire tree. If a switch occurs between adjacent LT leaf-pair $(\varepsilon_0, \varepsilon_{10})$ for some internal node $\varepsilon$, the adjacent pair $(\varepsilon_{10}, \varepsilon_{110})$ becomes “newly eligible” for switching, and if that switch eventually occurs, the adjacent pair $(\varepsilon_{110}, \varepsilon_{1110})$ becomes “newly eligible” for switching, and so on, which creates the possibility of a chain reaction in the direction of increasing tree level (though a chain reaction can just as easily occur in the other tree direction). This might explain the behavior in Figure 6c where the 0.17 mass moves almost sequentially through eight nodes. In contrast, because any BT leaf
is adjacent to exactly one other leaf and because adjacency is a symmetric binary relation, any switch between two adjacent BT leaves will likely stay localized i.e. such a switch will likely not create any “newly eligible” pairs. This BT conjecture is supported by the chain behavior in Figure 6d.

These two proposed mechanisms seem to explain the observed MCMC chain behavior difference between LT and BT. Overall, our aforementioned inference observations indicate that if we consider the parameter space $\Theta$ and posterior local modes corresponding to label permutations, the regions between these local modes appear to be deeper valleys for the BT posterior than they are for the LT posterior, making it more difficult to jump between different labeling for BT than for LT.

4 Case study: flow cytometry data

The goal of this analysis is to quantify the impact of a patient’s age, race, and sex on their proportions of T-cell types. The FCM data comes from 127 healthy plasma donors used to establish normative ranges for immune cell subsets enumerated by flow cytometry, as described in Yi et al. (2019). Donors were selected to be equally represented across the following age cohorts: 18–29 years, 30–39, 40–49, and 50–66. Of the 127 donors, 51 are Caucasian, 69 are African American, and 7 are classified as Other. Regarding sex, 65 are male and 62 are female. We focus our analysis on the T cell “panel,” a collection of custom antidotes designed to identify maturational and functional subsets of T cells developed by the Human ImmunoPhenotyping Consortium (HIPC) (Finak et al., 2016). Although the panel is designed to make it possible to identify multiple predefined T cell subsets, the sample is of all peripheral blood mononuclear cells, and clusters may include non-T cell subsets.
We subset the available data in the following way. Of the 127 donors, we use the 120 whose race is listed as either Caucasian or African-American and omit the remaining 7 subjects to avoid conflating effects attributed to the Other race label with subject-specific effects. Table 1 contains contingency tables for the subjects used in our analysis. Because these 120 subjects produce too many viable cells for the model to train on in a reasonable amount of time, we subset the data in a way that uses all 120 subjects while representing each covariate combination \( \mathbf{x} \) by the same number \( n_x \) of viable cells. First, we partition the subjects according to their age group, sex, and race. For each of the \( 4 \times 2 \times 2 = 16 \) covariate combinations \( \mathbf{x} \), we then subset \( n_x/|\mathcal{D}_x| \) viable cells from each subject with covariates \( \mathbf{x} \), where \( |\mathcal{D}_x| \) is the number of subjects that share covariates \( \mathbf{x} \) – see Table 1 for all values of \( |\mathcal{D}_x| \) – and \( n_x := 12600 = 2^3 \times 3^2 \times 5^2 \times 7 \) is divisible by all possible values of \( |\mathcal{D}_x| \). Thus, the training data consists of \( n = \sum_x n_x = 16n_x = 201600 \) viable cells.

| (Caucasian) | Male | Female | Total | (African-American) | Male | Female | Total |
|-------------|------|--------|-------|---------------------|------|--------|-------|
| 18 – 29     | 4    | 8      | 12    | 18 – 29             | 10   | 6      | 16    |
| 30 – 39     | 7    | 8      | 15    | 30 – 39             | 9    | 9      | 18    |
| 40 – 49     | 5    | 6      | 11    | 40 – 49             | 10   | 7      | 17    |
| 50 – 65     | 8    | 5      | 13    | 50 – 65             | 9    | 9      | 18    |
| Total       | 24   | 27     | 51    | Total               | 38   | 31     | 69    |

Table 1: Contingency tables for subjects used to analyze data from the GRIFOLS study.

To the data we fit a LTM model and a BTM model each with hyperparameter values of \( K = 64 \), prior mean \( \mu_\gamma = \mathbf{0}_p \), and prior covariance \( \Sigma_\gamma = 100I_p \), where \( p = 7 \). Each chain burns in 10,000 steps before sampling every 50 steps to ultimately keep 1,000 posterior draws. We also incorporate two statistical methodologies to address features common to FCM data. First, for both models we use \textit{cross-sample calibration} to account for subject-data being collected in different batches (Gorsky et al., 2020). Second, it is well known that FCM clusters do not match the shape of skew-normal distributions. To address this
kernel misspecification, we incorporate coarsening \cite{Miller2018} into our mixture models. In particular, we use the coarsening-parameter value of $\zeta = 0.1$, which is strong enough to combine similar-looking clusters but weak enough to avoid creating many multimodal clusters.

We use the following strategy to translate posterior covariate effects of stick-breaking proportions onto those of the weights. For any given MCMC chain and each of its posterior draws, each internal node $\varepsilon$ of the tree corresponds to a vector of regression coefficients $\gamma_\varepsilon = (\gamma_0, \gamma_{age1}, \gamma_{age2}, \gamma_{age3}, \gamma_{race}, \gamma_{sex})^T$ and each leaf $\delta$ of the tree corresponds to a weight $W_{x,\delta}$ created according to (7) and (2) for each vector of binary covariates $x \in X_{\text{GRIFOLS}}$ where $X_{\text{GRIFOLS}} = \{(1, x_{age1}, x_{age2}, x_{age3}, x_{race}, x_{sex}) \in \{0, 1\}^6 : x_{age1} \geq x_{age2} \geq x_{age3}\}$.

As an example, we examine the posterior effect of the race covariate on the weights by computing differences $W_{(1, x_{age1}, x_{age2}, x_{age3}, 1, x_{sex}), \delta} - W_{(1, x_{age1}, x_{age2}, x_{age3}, 0, x_{sex}), \delta}$ for various shared combinations of covariate values $x_{age1}, x_{age2}, x_{age3}, x_{sex}$ with $x_{age1} \geq x_{age2} \geq x_{age3}$ at each leaf $\delta$. The effect of the other covariates on the weights are computed in the same manner.

4.1 Trustworthiness of LT’s credible intervals

As mentioned in Section 1, Figure 8a showcases a disturbing dependence of a LTM model weight’s CI length on the order in which the weight was created during stick breaking. Indeed, cluster 01’s CI lengths are at least $5 \times$ many other CI lengths with larger weight indices. It now becomes difficult to trust how a LTM model quantifies a weight’s uncertainty in light of its strong artificial dependence on whichever leaf node the weight happens to belong to. This is particularly true if the model on the weights is misspecified to some degree (which is likely the case in any analysis of real data), in which case a trustworthy inference should retain a nontrivial amount of variance in the weights regardless of how
Figure 8: Pointwise 95% CIs of various weight differences inferred by a mixture model.

large \( n \) is. In contrast, each weight constructed by BT SB is a product of the same number
\((\log_2 K = 6)\) of stick breaks and thus does not share this artificial dependence. This
assertion is supported by Figure 8b whose CIs show no obvious relationship to weight
index and are roughly as wide as the CIs of cluster 05 and 07 from the LTM model.

4.2 BTM model inference

In light of Section 4.1, we provide the results from only the BTM model. Figure 12 shows
the salient clusters inferred by the BTM model. Figures 13, 14, 15, and 16 show weight
differences for various pairs of covariate-combinations that differ by exactly one covariate
value. For space considerations, we relegate them to the Supplementary Materials, but
here we present how a practitioner might interpret these results.

Changes with age Two cell subsets, corresponding to clusters 20 and 25, are over-
represented in older age groups (30-39, 40-49, 50-65) compared with the youngest cohort
(18-29) across both sexes and races. Cluster 20 has the phenotype of a central mem-
ory CD4+ T cell (CCR7+/CD45RA-), which arises after antigen encounter and hence
is expected to increase over time. Cluster 25 has the phenotype of a pro-inflammatory
monocyte/macrophage and suggests aging is associated with increased inflammation. This is not unexpected, as many diseases of aging are associated with chronic and excessive inflammation.

The cell subset corresponding to cluster 6 is over-represented in the second-youngest age group (30-39) compared with the youngest cohort (18-29) across both sexes and races. Cluster 6 has the phenotype of a naïve CD4+ T cell, and as naïve T cells are traditionally considered antigen-inexperienced T cells, this is counter-intuitive as thymic generation of new T cells drops precipitously after childhood. However, there is evidence that in adult humans, naïve T cells are maintained by cell division \cite{den Braber et al., 2012}, and hence while the diversity of the naïve T cell repertoire might be decreasing, the numbers may not.

Finally, the cell subset corresponding to cluster 24 is over-represented in older age groups (30-39, 40-49, 50-65) compared with the youngest cohort (18-29) across Caucasian males and females. Cluster 24 has the phenotype of a transitional cell moving from naïve to memory.

**Race-related differences** All three mentioned CD4+ T cell types (naïve, central memory, and transitional) show similar race-related differences in the 30-39 age cohort. Namely, they are predicted to be less abundant for African Americans relative to Caucasians. The same can also be said for both naïve and transitional CD4+ T cells in the oldest (50-65) age cohort, but not for the central memory CD4+ T cells.

**Sex-related differences** Only subtle sex differences in immune cell cluster are observed. In the oldest (50-65) age cohort, transitional cells are predicted to be moderately less abundant in males relative to females, whereas in the 30-39 age cohort, central memory
cells are predicted to be less abundant in males relative to females.

5 Conclusion

The focus of this paper has been to propose a new stick-breaking procedure that circumvents several undesirable behaviors of LT stick-breaking models. First, Section 2.5 shows that the LT induces a strong prior cross-covariate correlation between random measures whereas the BT allows more flexibility. Second, Section 3 details how the impact of label switching on posterior inference is stronger for LTM models than it is for BTM models, as well as conjectures as to the mechanisms that cause this behavior. Third, Section 4 illustrates the presence of artificial variance decay in weights constructed from LT SB as well as its absence in weights constructed from BT SB. We also provide a Gibbs sampler for the BTM model that either matches or beats the efficiency of that for the LTM model.

The model specified on the splitting variables can be tailored according to the nature of the given data. The splitting variables in (7) depend on covariates via a fixed-effects model, but a practitioner may wish to also allow random effects (e.g. corresponding to subject-specific idiosyncrasies in a FCM data set) to reduce posterior uncertainty and hence improve covariate-effect inference on the mixture weights. To achieve posterior computation with a mixed-effects model on the splitting variables, Wang and Roy (2018)’s two-block Gibbs sampler can be adapted to work for a BTM model.

Acknowledgements

We gratefully acknowledge Scott White for assistance in processing flow cytometry data.

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 [LM].

References

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

Cipolli, W. and T. Hanson (2017). Computationally tractable approximate and smoothed polya trees. *Statistics and Computing* 27(1), 39–51.

Cron, A. J. and M. West (2011). Efficient classification-based relabeling in mixture models. *The American Statistician* 65(1), 16–20.

den Braber, I., T. Mugwagwa, N. Vrisekoop, L. Westera, R. Mögling, A. B. de Boer, N. Willems, E. H. Schrijver, G. Spierenburg, K. Gaiser, et al. (2012). Maintenance of peripheral naive t cells is sustained by thymus output in mice but not humans. *Immunity* 36(2), 288–297.

Dubey, K. A., Q. Ho, S. A. Williamson, and E. P. Xing (2014). Dependent nonparametric trees for dynamic hierarchical clustering. *Advances in Neural Information Processing Systems* 27.

Fabius, J. (1964). Asymptotic behavior of bayes’ estimates. *The Annals of Mathematical Statistics* 34(4), 846–856.

Finak, G., M. Langweiler, M. Jaimes, M. Malek, J. Taghiyar, Y. Korin, K. Raddassi,
L. Devine, G. Obermoser, M. L. Pekalski, et al. (2016). Standardizing flow cytometry immunophenotyping analysis from the human immunophenotyping consortium. *Scientific reports* 6(1), 1–11.

Freedman, D. A. (1963). On the asymptotic behavior of bayes’ estimates in the discrete case. *The Annals of Mathematical Statistics* 34(4), 1386–1403.

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

Gorsky, S., C. Chan, and L. Ma (2020). Coarsened mixtures of hierarchical skew normal kernels for flow cytometry analyses. *arXiv preprint arXiv:2001.06451*.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* 96(453), 161–173.

Ishwaran, H. and L. F. James (2002). Approximate dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical statistics* 11(3), 508–532.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New phytologist* 11(2), 37–50.

Jara, A. and T. E. Hanson (2011). A class of mixtures of dependent tail-free processes. *Biometrika* 98(3), 553–566.

Linderman, S., M. J. Johnson, and R. P. Adams (2015). Dependent multinomial models made easy: Stick-breaking with the pólya-gamma augmentation. *Advances in Neural Information Processing Systems* 28.
MacEachern, S. N. (2000). Dependent dirichlet processes. *Unpublished manuscript, Department of Statistics, The Ohio State University*, 1–40.

Miller, J. W. and D. B. Dunson (2018). Robust bayesian inference via coarsening. *Journal of the American Statistical Association 114*(527), 1113–1125.

Nassar, J., S. W. Linderman, M. Bugallo, and I. M. Park (2018). Tree-structured recurrent switching linear dynamical systems for multi-scale modeling. *arXiv preprint arXiv:1811.12386*.

Polson, N. G., J. G. Scott, and J. Windle (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association 108*(504), 1339–1349.

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

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

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

Stefanucci, M. and A. Canale (2021). Multiscale stick-breaking mixture models. *Statistics and Computing 31*(2), 1–13.

Wang, X. and V. Roy (2018). Analysis of the pólya-gamma block gibbs sampler for bayesian logistic linear mixed models. *Statistics & Probability Letters 137*, 251–256.

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

Proofs: Proofs of Theorems 1, 2, 3, and Corollary 2.1.

Discussion: Discussion on computational cost of Gibbs step from Algorithm 1.

Plots: Plots of results from Sections 3 and 4.

6 Proofs

6.1 Proof of Theorem 1

*Proof of Theorem* 1 This proof follows closely from Appendix 2 of Rodriguez and Dunson (2011). To prove the mean result (8), we have

$$
\mathbb{E}G_x(A) = \mathbb{E} \left\{ \sum_{k=1}^{K} W_{x,k} \delta_{\theta_k}(A) \right\} = \sum_{k=1}^{K} \left[ \mathbb{E}W_{x,k} \right] Pr(\theta_1 \in A) = G_0(A).
$$

To prove the remaining results, let $X = G_x(A)$ and $Y = G_{x'}(A')$ and note that

$$
\mathbb{E}XY = \mathbb{E} \left[ \left\{ \sum_{k=1}^{K} W_{x,k} \delta_{\theta_k}(A) \right\} \left\{ \sum_{k=1}^{K} W_{x',k} \delta_{\theta_k}(A') \right\} \right]
$$

$$
= \sum_{k=1}^{K} \mathbb{E} [W_{x,k}W_{x',k} \delta_{\theta_k}(A)\delta_{\theta_k}(A')] + \sum_{k=1}^{K} \sum_{k' \neq k} \mathbb{E} [W_{x,k}W_{x',k'} \delta_{\theta_k}(A)\delta_{\theta_{k'}}(A')]
$$

$$
= G_0(A \cap A')a_{x,x'} + G_0(A)G_0(A')b_{x,x'}
$$

where $a_{x,x'} = \sum_{k=1}^{K} \mathbb{E}W_{x,k}W_{x',k}$ and $b_{x,x'} = \sum_{k=1}^{K} \sum_{k' \neq k} \mathbb{E}W_{x,k}W_{x',k'}$. Because $a_{x,x'} + b_{x,x'} = 1$, we have $\mathbb{E}XY = G_0(A)G_0(A')(a_{x,x'} + b_{x,x'})$, and thus

$$
\text{Cov}(X,Y) = \mathbb{E}XY - \mathbb{E}XEY = [G_0(A \cap A') - G_0(A)G_0(A')] a_{x,x'}.
$$

36
The variance and covariances (9), (10), and (11) result from appropriately setting $x = x'$ or $A = A'$ (or both). The correlations (12) and (13) result from the definition $\text{corr}(X, Y) = (\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y)/\sqrt{\text{Var}X\text{Var}Y}$, which requires the denominator to be nonzero.

6.2 Proof of Theorem 2

Proof of Theorem 2 The LT result comes from Appendix 2 of Rodriguez and Dunson (2011). For the BT result, let $m = |B(\tau)|$. For any $\varepsilon \in B(\tau)$,

$$
\mathbb{E}W_{x, \varepsilon}W_{x', \varepsilon} = \prod_{l=1}^{m} \mathbb{E} \left[ V_{x, \varepsilon_1 \cdots \varepsilon_{l-1}} V_{x', \varepsilon_1 \cdots \varepsilon_{l-1}} \right]^{1 - \varepsilon_l} \mathbb{E} \left[ (1 - V_{x, \varepsilon_1 \cdots \varepsilon_{l-1}}) (1 - V_{x', \varepsilon_1 \cdots \varepsilon_{l-1}}) \right]^{\varepsilon_l}.
$$

Because the splitting variables are identically distributed, the quantity $\mathbb{E}W_{x, \varepsilon}W_{x', \varepsilon}$ becomes the product $\prod_{l=1}^{m} c_{x, x'}^{1 - \varepsilon_l} d_{x, x'}^{\varepsilon_l} = c_{x, x'}^{m - \sum_{l=1}^{m} \varepsilon_l} d_{x, x'}^{\sum_{l=1}^{m} \varepsilon_l}$ where $c_{x, x'} = \mathbb{E} [V_x V_{x'}]$ and $d_{x, x'} = \mathbb{E} [(1 - V_x) (1 - V_{x'})]$. For any $k \in \{0, 1, \ldots, m\}$, the set $B(\tau)$ has exactly $\binom{m}{k}$-many leaves $\varepsilon_1 \cdots \varepsilon_m$ with the property $\sum_{l=1}^{m} \varepsilon_l = k$. Thus, the desired sum $\sum_{\varepsilon \in B(\tau)} \mathbb{E}W_{x, \varepsilon}W_{x', \varepsilon}$ is then $\sum_{k=0}^{m} \binom{m}{k} c_{x, x'}^{m-k} d_{x, x'}^k = (c_{x, x'} + d_{x, x'})^m$. 

6.3 Proof of Theorem 3

Proof of Theorem 3 In general, as $x' \to x$, we have $\mathbb{E}V_{x'} \to \mathbb{E}V_x$. But if also $\mathbb{E}V_x V_{x'} \to \mathbb{E}V_x^2$, then $a_{x, x'} \to a_{x, x}$ by Theorem 2. Thus, (13) $\to 1$. 

37
6.4 Proof of Corollary 2.1

Proof of Corollary 2.1. We begin by proving the result for BT. If \( EV_x = 1/2 \) for any \( x \), then for BT we have

\[
(13) = \left[ \frac{EV_x V_{x'}}{\sqrt{EV_x^2 EV_{x'}^2}} \right]^m = \left[ \frac{\text{Cov}(x, x') + 1/4}{\sqrt{(\text{Var}(x) + 1/4)(\text{Var}(x') + 1/4)}} \right]^m.
\]

Under the condition \( \text{Cov}(x, x') \geq 0 \), (13) is minimized when \( \text{Cov}(x, x') = 0 \) whenever \( x \neq x' \). Setting this condition results in

\[
(13) = \left( \frac{(4\text{Var}(x) + 1)(4\text{Var}(x') + 1)}{1 - \text{Var}(x) - \frac{1}{4}} \right)^{m/2} \geq 2^{-m}.
\]

This lower bound is strict if \( \text{Var}(x) > 0 \) for all \( x \in X \).

Now we prove the result for LT. If \( EV_x = 1/2 \) for any \( x \), then for LT we have

\[
(13) = \left[ \frac{EV_x V_{x'}}{\sqrt{EV_x^2 EV_{x'}^2}} \right]^m = \left[ \frac{1 - EV^2_x V_{x'}}{1 - EV^2_x V_{x'}} \right]^{m/2}.
\]

If we fix \( K \), \( \text{Var}(x) \) and \( \text{Var}(x') \), then (13) strictly increases in \( \text{corr}(x, x') \). So under the condition \( \text{Cov}(x, x') \geq 0 \), (13) is minimized when \( \text{Cov}(x, x') = 0 \) whenever \( x \neq x' \). Setting this condition results in

\[
(13) = \frac{1}{3} \sqrt{\left( \frac{1}{\text{Var}(x) + 1/4} - 1 \right) \left( \frac{1}{\text{Var}(x') + 1/4} - 1 \right)} \times \frac{1 - [1/4]^K}{\sqrt{(1 - [\text{Var}(x) + 1/4]^K)(1 - [\text{Var}(x') + 1/4]^K)}}.
\]

38
Because $\text{Var}_V(x) \leq 1/4$ for all $x \in \mathcal{X}$, we have (13) $\geq \frac{1}{3} \times \frac{1-\frac{4}{1-2^{-K}}}{1}$. This lower bound is a decreasing function of $K$. At $K = \infty$, it is $1/3$. This lower bound is also strict if $\text{Var}_V(x) > 0$ for all $x \in \mathcal{X}$.

6.5 Computational cost of Gibbs step

Next we compare between LT and BT the computational cost of this Gibbs step for some arbitrarily fixed MCMC iteration. In this case, each tree has the same number of internal nodes and hence their corresponding mixture models perform the same number of regressions. From this viewpoint, the computational cost should be the same between the two SB schemes.

However, we find from experience that when LTM and BTM models are trained on the same data set, the BTM model usually finishes first. To see why, we note that for a tree $\tau$ and leaf-assignment $C_\tau$, the computational cost of the Gibbs step can be roughly identified with the sum

$$\sum_{i=1}^{n} \left( \text{# of } C_\tau \text{'s regressions involving obs } i \right) = \sum_{i=1}^{n} \left( \text{# of ancestors of leaf node } C_\tau(i) \right).$$

For BT, all observations have $\log_2 K$ ancestors, which means the sum equals $n \log_2 K$. The sum for LT, however, decreases as observations are allocated to leaf nodes closer to root (meaning the observations are involved in fewer regressions) and increases as observations are allocated further from root. We can push this reasoning to the extreme to get the sum’s minimum and maximum values for LT. If $K^+$ clusters are inferred, then the minimum is achieved when the root’s left-child leaf node contains all but $(K^+-1)$ observations and $(K^+-1)$ other leaf nodes each contain just one observation, which results in the sum

39
equaling

\[(n - K^+ + 1) + (2 + 3 + 4 + \cdots + K^+) = n + K^+(K^+ - 1)/2.\]

The maximum value is similarly achieved if it is the nonempty leaf node furthest from root that contains the \((n - K^+ + 1)\) observations, which results in the sum equaling

\[K^+(n - K^+ + 1) + (1 + 2 + \cdots + (K^+ - 1)) = K^+(n - K^+ + 1) + K^+(K^+ - 1)/2.\]

If \(n \gg K^+\) and \(n\) is fixed, then the former quantity is roughly constant in \(K^+\) whereas the latter quantity is roughly linear in \(K^+\), as shown in Figure 9.

Though these extreme scenarios provide useful computational cost bounds for LT, seldom will they be even remotely close to how a leaf assignment \(C_\tau\) will allocate observations. On the other hand, what exactly is a more “typical” allocation, and can we find an analytical representation for it or for a reasonable approximation of it? The answer to the former question (and hence also the latter question) depends on, among other things, the size of the data’s clusters and how well the mixture model infers these sizes. For simplicity, in the following exposition we will assume the mixture model infers \(K^+\) leaf nodes each containing \(n/K^+\) observations, in which case the sum for LT becomes \(n(K^+ + 1)/2\), which again is linear in \(K^+\).

The final hurdle for this computational cost comparison between LT and BT is that our three sum quantities for LT — \(n + K^+(K^+ - 1)/2\), \(K^+(n - K^+ + 1) + K^+(K^+ - 1)/2\), and \(n(K^+ + 1)/2\) — are functions of \(n\) and \(K^+\) while the sum quantity for BT — \(n \log_2 K\) — is a function of \(n\) and \(K\). How can we relate \(K\) to \(K^+\)? We might optimistically assume that the practitioner has chosen \(K\) to be the smallest power of two greater than or equal to \(K^+\), but we will “leave room for error” by instead assuming \(K\) is the smallest power of two
Figure 9: Scaled sum values of \( (1 + n^{-1}K^+(K^+ - 1)/2) \), \( n^{-1}(K^+(n - K^+ + 1) + K^+(K^+ - 1)/2) \), \( (K^+ + 1)/2 \), and \( (1 + \lceil \log_2 K^+ \rceil) \) for various combinations of \( n \) and \( K^+ \). The first three expressions are displayed as ‘L’ while the fourth expression is displayed as ‘B’.

greater than or equal to \( 2K^+ \), which implies \( \log_2 K = (1 + \lceil \log_2 K^+ \rceil) \). Now all four sum quantities can be expressed as functions of \( n \) and \( K^+ \) as shown in Figure 9. The (scaled) sum for BT grows logarithmically in \( K^+ \) whereas the (scaled) sum for the “typical” LT scenario grows linearly, which explains why we find that training a BTM model is usually faster than training its LTM counterpart.

7 Simulation study plots

8 GRIFOLS plots
Figure 10: Weights for various covariate combinations.

Figure 11: Weights for various covariate combinations.
Figure 12: Salient clusters inferred by BTM model with $\zeta = 0.1$. 
| Weight Index (Cluster Size) |
|-----------------------------|
| 0.0%                        |
| 0.1%                        |
| 0.2%                        |
| 0.3%                        |
| 0.4%                        |
| 0.5%                        |
| 0.6%                        |
| 0.7%                        |
| 0.8%                        |
| 0.9%                        |
| 1.0%                        |
| 1.1%                        |
| 1.2%                        |
| 1.3%                        |
| 1.4%                        |
| 1.5%                        |
| 1.6%                        |
| 1.7%                        |
| 1.8%                        |
| 1.9%                        |
| 2.0%                        |
| 2.1%                        |
| 2.2%                        |
| 2.3%                        |
| 2.4%                        |
| 2.5%                        |
| 2.6%                        |
| 2.7%                        |
| 2.8%                        |
| 2.9%                        |
| 3.0%                        |
| 3.1%                        |
| 3.2%                        |
| 3.3%                        |
| 3.4%                        |
| 3.5%                        |
| 3.6%                        |
| 3.7%                        |
| 3.8%                        |
| 3.9%                        |
| 4.0%                        |
| 4.1%                        |
| 4.2%                        |
| 4.3%                        |
| 4.4%                        |
| 4.5%                        |
| 4.6%                        |
| 4.7%                        |
| 4.8%                        |
| 4.9%                        |
| 5.0%                        |
| 5.1%                        |
| 5.2%                        |
| 5.3%                        |
| 5.4%                        |
| 5.5%                        |
| 5.6%                        |
| 5.7%                        |
| 5.8%                        |
| 5.9%                        |
| 6.0%                        |
| 6.1%                        |
| 6.2%                        |
| 6.3%                        |
| 6.4%                        |
| 6.5%                        |
| 6.6%                        |
| 6.7%                        |
| 6.8%                        |
| 6.9%                        |
| 7.0%                        |
| 7.1%                        |
| 7.2%                        |
| 7.3%                        |
| 7.4%                        |
| 7.5%                        |
| 7.6%                        |
| 7.7%                        |
| 7.8%                        |
| 7.9%                        |
| 8.0%                        |
| 8.1%                        |
| 8.2%                        |
| 8.3%                        |
| 8.4%                        |
| 8.5%                        |
| 8.6%                        |
| 8.7%                        |
| 8.8%                        |
| 8.9%                        |
| 9.0%                        |
| 9.1%                        |
| 9.2%                        |
| 9.3%                        |
| 9.4%                        |
| 9.5%                        |
| 9.6%                        |
| 9.7%                        |
| 9.8%                        |
| 9.9%                        |
| 10.0%                       |

Figure 13: Weight differences between adjacent age groups for various baseline race/sex.
Figure 14: Weight differences between age groups and the youngest age group for various baseline race/sex.
Figure 15: Weight differences between race for various baseline sex and age group.
Comparison: Male vs Female
Model: BT with zeta=0.1

Figure 16: Weight differences between sex for various baseline race and age group.