Biometrics XX, 1–25 DOI: 10.1111/XXXX
XXXX XXX

Integrating Sample Similarities into Latent Class Analysis: A Tree-Structured Shrinkage Approach

Mengbing Li[1], Daniel E. Park[3], Maliha Aziz[3], Cindy M. Liu[3], Lance B. Price[3], Zhenke Wu[1,2]*

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

[2]Michigan Institute for Data Science (MIDAS), University of Michigan, Ann Arbor, MI 48109, USA

[3]Environmental and Occupational Health, Milken Institute School of Public Health,
The George Washington University, Washington, DC 20052, USA

**email:* zhenkewu@umich.edu

Summary: This paper is concerned with using multivariate binary observations to estimate the probabilities of unobserved classes with scientific meanings. We focus on the setting where additional information about sample similarities is available and represented by a rooted weighted tree. Every leaf in the given tree contains multiple samples. Shorter distances over the tree between the leaves indicate *a priori* higher similarity in class probability vectors. We propose a novel data integrative extension to classical latent class models (LCMs) with tree-structured shrinkage. The proposed approach enables 1) borrowing of information across leaves, 2) estimating data-driven leaf groups with distinct vectors of class probabilities, and 3) individual-level probabilistic class assignment given the observed multivariate binary measurements. We derive and implement a scalable posterior inference algorithm in a variational Bayes framework. Extensive simulations show more accurate estimation of class probabilities than alternatives that suboptimally use the additional sample similarity information. A zoonotic infectious disease application is used to illustrate the proposed approach. The paper concludes by a brief discussion on model limitations and extensions.

Key words: Gaussian Diffusion; Latent Class Models; Phylogenetic Tree; Spike-and-Slab Prior; Variational Bayes; Zoonotic Infectious Diseases.

1. Introduction

1.1 *Motivating Application*

The fields of infectious disease epidemiology and microbial ecology need better tools for tracing the transmission of microbes between humans and other vertebrate animals (i.e., zoonotic transmissions), especially for colonizing opportunistic pathogens (COPs). Unlike frank zoonotic pathogens (e.g., *Salmonella*, SARS-CoV-2), the epidemiology of COPs, such as *Escherichia coli* (*E. coli*), *Staphylococcus aureus* (*S. aureus*) and *Enterococcus spp.*, can be particularly cryptic due to their ability to asymptomatically colonize the human body for indefinite periods prior to initiating an infection, transmitting to another person, or being shed without a negative outcome (e.g., Price et al., 2017). Some COPs can colonize many different vertebrate hosts and cross-species transmissions can go unrecognized. Estimating the probability of zoonotic origin for a population of isolates and for each isolate would provide important insights into the natural history of infections and inform more effective intervention strategies, such as eliminating high-risk clones from livestock via vaccination.

Scientists often have two complementary sources of information: i) a phylogenetic tree constructed based on a few single nucleotide polymorphisms (SNPs) in the core genome shared by all isolates, where the leaves represent distinct core-genome multi-locus sequence types (STs, Maiden et al., 1998); the tree is useful for identifying a recent common ancestor for isolates that comprise an infectious disease outbreak; ii) presence or absence of multiple mobile genetic elements (MGE) that provide selective advantages in particular hosts and may be lost and gained as COPs transmit among hosts (e.g., Lindsay and Holden, 2004).

Recent research on two COP species, *E. coli* and *S. aureus*, has demonstrated the utility of complementing core-genome phylogenetic trees with host-associated MGEs to resolve host origins (e.g., Liu et al., 2018). However, in both cases only a single host-associated MGE was used. Analyses were largely limited to visual inspection of how each element fell on the scaffold of the evolutionary tree. For this approach to reach its full potential, we would need a statistical model that can 1) integrate phylogenetic information with the presence and absence of multiple host-associated MGEs, and 2) estimate the probability with which the isolates were derived from a particular host in each ST-specific population and for each individual isolate.

1.2 *Integrating Sample Similarities into Latent Class Analysis*

Based on multivariate binary data (e.g., presence or absence of multiple MGEs), we use latent class models (LCMs; e.g., Lazarsfeld, 1950; Goodman, 1974) to achieve the scientific goal of estimating the probabilities of unobserved host origins and perform individual-level probabilistic assignment of host origin. LCMs are examples of latent variable models that assume the observed dependence among multivariate discrete responses is induced by variation among unobserved or "latent" variables. It is well known that any multivariate discrete data distribution can be approximated arbitrarily closely by an LCM with a sufficiently large number of classes (Dunson and Xing, 2009, Corollary 1). The most commonly used LCMs assume the class membership indicators for the observations are drawn from a population with the same vector of class probabilities.

Trees or hierarchies are useful and intuitive for representing and reasoning about similarity or relation among objects in many real-world domains. We assume known entities at the leaves. In our context, each leaf may contain multiple observations or samples, each associated with the multivariate binary responses which are then combined to form the rows of a binary data matrix $\mathbf{Y}$. In the motivating application, the latent class indicates the unobserved host origin (human or non-human) to be inferred by the presence or absence of multiple MGEs. The additional sample similarity information is represented by a maximum likelihood phylogenetic tree (e.g., Scornavacca et al., 2020). The leaves represent distinct contemporary core-genome *E. coli* STs.

To integrate tree-encoded sample similarity information into a latent class analysis, ad hoc groupings of the leaves may be adopted. From the finest to the coarsest leaf grouping, one may 1) analyze data from distinct lineages one at a time, 2) manually form groups of at least one leaf node and fit separate LCMs, or 3) fit all the data by a single LCM. However, all these methods pose significant statistical challenges. First, separate latent class analyses may have low accuracy in estimating latent class probabilities and other model parameters for rare lineages. Second, observations of similar lineages may have similar propensities in host jump resulting in similar host origin class probabilities. Modeling these similarities could lead to gain in statistical efficiency. Third, approaches based on coarse ad hoc groupings may obscure the study of the variation in the latent class probabilities across different parts of the tree. Finally, based on a single LCM or other approaches that use ad hoc leaf groupings, individual-specific posterior class probabilities can be averaged within in each leaf to produce a local estimate of the $\boldsymbol{\pi}_v$. However, the ad hoc post-processing cannot fully address the issue of assessment of posterior uncertainty nor data-driven grouping of leaves, necessitating development of an integrative probabilistic modeling framework for uncertainty quantification and adaptive formation of leaf groups.

In this paper, we focus on integrating the tree-encoded sample similarity information into latent class analysis. We assume the tree information is given and not computed from the multivariate binary measurements. Observations in nearby leaves are assumed to have *a priori* similar propensities of being members of a particular class as characterized by the latent class probabilities. For example, higher similarities are indicated by shorter pairwise distances between observations. More generally, classical covariate-dependent latent class models (e.g., Bandeen-Roche et al., 1997; Formann, 1992) let the latent class probabilities vary explicitly as functions of observed covariates so that observations with more similar covariate values are assumed to have more similar latent class probabilities. Fully probabilistic tree-integrative methods have appeared in machine learning literature (e.g., Ghahramani et al., 2010; Roy et al., 2006; Ranganath et al., 2015) or in statistics for modeling hierarchical topic annotations (e.g., Airoldi and Bischof, 2016) or hierarchical outcome annotations based on given trees (e.g., Thomas et al., 2019). In epidemiology, Avila et al. (2014) proposed a two-stage approach to link patient clusters estimated from the tree and by the LCM results, which however remains ad hoc. However, current literature does not address probabilistic tree-integrative latent class analysis or adaptive formation of leaf groups for dimension reduction.

1.3 *Primary Contributions*

In this paper, we propose an unsupervised, tree-integrative LCM framework to 1) discover groups of leaves where multivariate binary measurements in distinct leaf groups have distinct vectors of latent class probabilities; And observations nested in any leaf group may belong to a pre-specified number of latent classes; 2) accurately estimate the latent class probabilities for each discovered leaf group and assign probabilities of an individual sample belonging to the latent classes; 3) leverage the relationship among the observations as encoded by the tree to boost the accuracy of the estimation of latent class probabilities. Without pre-specifying the leaf groups, the automatic data-driven approach enjoys robustness by avoiding potential mis-specification of the grouping structure. On the other hand, the discovered data-driven leaf groups dramatically reduce the dimension of leaves into fewer subgroups of leaves hence improving interpretation. In addition, the proposed approach shows better accuracy in estimating the latent class probabilities in terms of root mean squared errors, indicating the advantage of the shrinkage. On posterior computation, we derive a scalable inference algorithm based on variational inference (VI).

The rest of the paper is organized as follows. Section 2.2 defines tree-related terminologies and formulates LCMs. Section 3 proposes the prior for tree-structured shrinkage in LCMs. Section 4 derives a variational Bayes algorithm for inference. Section 5 compares the performances of the proposed and alternative approaches via simulations. Section 6 illustrates the approach by analyzing an *E. coli* data set. The paper concludes with a brief discussion.

2. Model

We first introduce necessary terminologies and notations to describe a rooted weighted tree. LCMs are then formulated for data on the leaves of the tree.

2.1 *Rooted Weighted Trees*

A rooted tree is a graph $\mathcal{T} = (\mathcal{V}, E)$ with node set $\mathcal{V}$ and edge set E where there is a root u_0 and each node has at most one parent node. Let $p = |\mathcal{V}|$ represent the total number of leaf and non-leaf nodes. Let $\mathcal{V}_L \subset \mathcal{V}$ be the set of leaves, and $p_L = |\mathcal{V}_L| < p$. We typically use u to denote any node ($u \in \mathcal{V}$) and v to denote any leaf ($v \in \mathcal{V}_L$). Each edge in a rooted tree defines a *clade*: the group of leaves below it. Splitting the tree at an edge creates a partition of the leaves into two groups. For any node $u \in \mathcal{V}$, the following notations apply: $c(u)$ is the set of offspring of u, $pa(u)$ is the parent of u, $d(u)$ is the set of descendants of u including u, and $a(u)$ is the set of ancestors of u including u. In Figure 3(a), if $u = 2$, then $c(u) = \{6, 7, 8\}$, $pa(u) = \{1\}$, $d(u) = \{2, 6, 7, 8\}$, and $a(u) = \{1, 2\}$. The phylogenetic tree in our motivating application is a nested hierarchy of 133 STs for $N = 2,663$ observations, where the $p_L = |\mathcal{V}_L| = 133$ leaves represent distinct STs and the $p - p_L = 132$ internal (non-leaf) nodes represent ancestral *E. coli* strains leading up to the observed leaf descendants.

Edge-weighted graphs appear as a model for numerous problems where nodes are linked with edges of different weights. In particular, the edges in $\mathcal{T}$ are attached with weights where $w : E \rightarrow \mathbb{R}^+$ is a weight function. Let $\mathcal{T}_w = (\mathcal{T}, w)$ be a rooted weighted tree. A path in a graph is a sequence of edges which joins a sequence of distinct vertices. For a path P in the tree connecting two nodes, $w(P)$ is defined as the sum of all the edge weights along the path, often referred to as the "length" of P. The distance between two vertices u and u', denoted by $dist_{\mathcal{T}_w}(u, u')$ is the length of a shortest (with minimum length) (u, u')-path. $dist_{\mathcal{T}_w}$ is a distance: it is symmetric and satisfies the triangle inequality. In our motivating application, the edge length represents the number of nucleotide substitutions per position; the distance between two nodes provides a measurement of the similarity or divergence between any two core-genome sequences of the input set. In this paper, we use w_u to represent the edge length between a node u and its parent node $pa(u)$. w_u is fully determined by $\mathcal{T}_w$. For the root u_0, there are no parents, i.e. $pa(u_0) = \emptyset$; we set $w_{u_0} = 1$.

2.2 *Latent Class Models for Data on the Leaves*

Although LCMs can deal with multiple categorical responses in general, for simpler presentations in this paper, we focus on presenting the model and algorithm using multivariate binary responses and their application to the motivating data.

Notations. Let $\mathbf{Y}_i^{(v)} = (Y_{i1}^{(v)}, \ldots, Y_{iJ}^{(v)})^\intercal \in \{0,1\}^J$ be the vector of binary responses for observation $i \in [n_v]$ that is nested within leaf node $v \in \mathcal{V}_L$, where n_v is the number of observations in leaf v. Throughout this paper, let $[Q] := \{1, \ldots, Q\}$ denote the set of positive integers smaller than or equal to Q, where Q is a positive integer. Let $\mathbf{Y}^{(v)} = \left(\mathbf{Y}_1^{(v)}, \ldots, \mathbf{Y}_{n_v}^{(v)}\right)^\intercal$ be the data from observations in leaf v. Let $\mathbf{Y} = \left(\left(\mathbf{Y}^{(1)}\right)^\intercal, \ldots, \left(\mathbf{Y}^{(p_L)}\right)^\intercal\right)^\intercal$ represent the binary data matrix with $N = \sum_{v \in \mathcal{V}_L} n_v$ rows and J columns. Let $\mathcal{L} = (v_1, \ldots, v_N)^\intercal$ be the "sample-to-leaf indicators" that map every row of data $\mathbf{Y}$ into a leaf in $\mathcal{T}_w$. Sample similarities are then characterized by between-leaf distances in $\mathcal{T}_w$. In this paper, we assume $\mathcal{L}$ and $\mathcal{T}_w$ are given and focus on incorporating $(\mathcal{L}, \mathcal{T}_w)$ into a statistical model for $\mathbf{Y}$.

LCM for Data on the Leaves. The LCM is specified in two steps:

$$\textsf{class indicator}: \quad I_i^{(v)} \mid \boldsymbol{\pi}_v \sim \textsf{Categorical}_K\left\{\boldsymbol{\pi}_v\right\}, \boldsymbol{\pi}_v \in \mathcal{S}_{K-1}, \tag{1}$$

$$\textsf{data}: \quad Y_{ij}^{(v)} \mid I_i^{(v)} \sim \textsf{Bernoulli}\left\{\theta_{j, I_i^{(v)}}\right\}, \text{ independently for feature } j \in [J], \tag{2}$$

and independently for observation $i \in [n_v]$ and leaf node $v \in \mathcal{V}_L$. Here K is a pre-specified number of latent classes in the context of the application, e.g., $K = 2$ for unobserved human and non-human hosts; see Section 4.1 for a simple strategy in applications where data-driven K is desired. In addition, $\mathbf{I} = \{I_i^{(v)} : i \in [n_v]; v \in \mathcal{V}_L\}$ represent the latent class indicators and $Z_{ik}^{(v)} = \mathbf{1}\{I_i^{(v)} = k\}$, $k \in [K]$, where $\mathbf{1}\{A\}$ is an indicator function which equals 1 if statement A is true and 0 otherwise; Let $\mathbf{Z} = \{Z_{ik}^{(v)}\}$. We have assumed observations in different leaves have potentially different vectors of class probabilities $\boldsymbol{\pi}_v = (\pi_{v1}, \ldots, \pi_{vK})^\intercal \in \mathcal{S}_{K-1}$, $v \in \mathcal{V}_L$, where $\mathcal{S}_{K-1} = \{\boldsymbol{r} \in [0,1]^K : \sum_{k=1}^K r_k = 1\}$ is the probability simplex. $\theta_{jk} \in [0,1]$ is the positive response probability for feature $j \in [J]$ in class $k \in [K]$. In our motivating application, the MGEs adapt to the unobserved type of host origin (i.e., latent class) which can be characterized by class-specific response probability profiles $\boldsymbol{\theta}_{\cdot k} = (\theta_{1k}, \ldots, \theta_{Jk})^\intercal$, $k \in [K]$; let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_{\cdot 1}, \ldots, \boldsymbol{\theta}_{\cdot K})^\intercal$. Because the latent class indicators $I_i^{(v)}$'s are assumed to be unobserved, the observed data likelihood for N observations is $\prod_{v\in\mathcal{V}_L} \prod_{i=1}^{n_v} \sum_{k=1}^K \pi_{vk} \mathbb{P}\left(\boldsymbol{Y}_i^{(v)} \mid I_i^{(v)} = k, \boldsymbol{\theta}_{\cdot k}\right)$.

Throughout this paper, we assume that we wish to classify individuals into K classes with the same set of $(\boldsymbol{\theta}_{\cdot 1}, \ldots, \boldsymbol{\theta}_{\cdot K})$ so classes have coherent interpretation. However, we do not assume that observations are drawn from a population with a single vector of latent class probabilities. Figure 1 provides a schematic of the data generating mechanism given $\boldsymbol{\pi}_v$ for three leaves.

[Figure 1 about here.]

3. Prior Distribution

We first specify a prior distribution for $\{\boldsymbol{\pi}_v : v \in \mathcal{V}_L\}$. Because leaf-specific sample sizes may vary, we propose a tree-structured prior to borrow information across nearby leaves. The prior encourages collapsing certain parts of the tree so that observations within a collapsed leaf group share the same vector of latent class probabilities. In particular, we extend Thomas et al. (2019) to deal with rooted weighted trees in an LCM setting. The prior specification is completed by priors for the class-specific response probabilities $\boldsymbol{\Theta}$.

Tree-structured prior for latent class probabilities $\boldsymbol{\pi}_v$. We specify a spike-and-slab Gaussian diffusion process prior along a rooted weighted tree based on a logistic stick-breaking parameterization of $\boldsymbol{\pi}_v$. We first reparameterize $\boldsymbol{\pi}_v$ with a stick-breaking representation: $\pi_{vk} = V_{vk}\prod_{s<k}(1-V_{vs})$, for $k \in [K]$, where $0 \leqslant V_{vk} \leqslant 1$, for $k \in [K-1]$ and $V_{vK} = 1$.

We further logit-transform $V_{vk}, k \in [K-1]$, to facilitate the specification of a Gaussian diffusion process prior without range constraints. In particular, let $\eta_{vk} = \sigma^{-1}(V_{vk})$, $k \in [K-1]$, $v \in \mathcal{V}_L$, where $\sigma(x) = 1/\{1+\exp(-x)\}$ is the sigmoid function. The logistic stick-breaking parameterization is completed by

$$\pi_{vk} = \{\sigma(\eta_{vk})\}^{\mathbf{1}\{k<K\}}\prod_{s<k}\sigma(-\eta_{vs}), k \in [K], \tag{3}$$

which affords simple and accurate posterior inference via variational Bayes (see Section 4).

For a leaf $v \in \mathcal{V}_L$, let

$$\eta_{vk} = \sum_{u \in a(v)} \xi_{uk}, k \in [K-1]. \tag{4}$$

Here η_{vk} is defined for leaves only and ξ_{uk} is defined for all the nodes. Suppose v and v' are leaves and siblings in the tree such that $pa(v) = pa(v')$, setting $\xi_{vk} = \xi_{v'k} = 0$ implies $\eta_{vk} = \eta_{v'k}$ for $k \in [K-1]$, and hence $\boldsymbol{\pi}_v = \boldsymbol{\pi}_{v'}$. More generally, a sufficient condition for M leaves η_{vk}, $v \in \{v_1, \ldots, v_M\}$ to fuse is to set $\xi_{uk} = 0$ for any u that is an ancestor of any of $\{v_1, \ldots, v_M\}$ but not common ancestors for all v_m. That is, to achieve grouping of observations that share the same vector of latent class probabilities, in our model, it is equivalent to parameter fusing. In the following, we specify a prior on the ξ_{uk} that *a priori* encourages sparsity, so that closely related observations are likely grouped to have the same vector of class probabilities. The fewer distinct ancestors two nodes have, the more likely the parameters η_{vk} are fused, because the prior would encourage fewer auxiliary variables ξ_{uk} to be set to zero. In particular, we specify

$$\xi_{uk} = s_u \alpha_{uk}, \forall\ u \in \mathcal{V}, \tag{5}$$

$$\alpha_{uk} \sim N(0, \tau_{1k\ell_u} w_u),\ \text{independently for } k \in [K-1], \forall\ u \in \mathcal{V}, \tag{6}$$

$$s_{u_0} = 1,\ \text{and } s_u \sim \mathsf{Bernoulli}(\rho_{\ell_u}),\ \text{independently for } u \in \mathcal{V} \setminus u_0, \tag{7}$$

$$\rho_\ell \sim \mathsf{Beta}(a_\ell, b_\ell),\ \text{independently for } \ell \in [L], \tag{8}$$

where $N(m, s)$ represents a Gaussian density function with mean m and variance s. $\tau_{1k\ell}$ is the unit-length variance and controls the degree of diffusion along the tree which may differ by dimension k and node level ℓ_u where $\ell_u \in [L]$ represents the "level" or "hyperparameter set indicator" for node u. For example, in simulations and data analysis, we will assume that the root for the diffusion process has a prior unit-length variance distinct from other non-root nodes. For the root u_0 with $s_{u_0} = 1$, $\alpha_{u_0 k}$ initializes the diffusion of η_{uk}.

Leaf groups are formed by selecting a subset of nodes in $\mathcal{V}$: $\mathcal{U} = \{u \in \mathcal{V} : s_u = 1\}$. Except a probability-zero set, two leaves v and v' are grouped, or "fused", if and only if $a(v) \cap \mathcal{U} = a(v') \cap \mathcal{U}$. In particular, the null set is $\{\eta_{vk} = \eta_{v'k}, k \in [K-1]\} \cap \{\sum_{u \in [a(v) \cap \mathcal{U}] \setminus [a(v') \cap \mathcal{U}]} \alpha_{uk} = \sum_{u \in [a(v') \cap \mathcal{U}] \setminus [a(v) \cap \mathcal{U}]} \alpha_{uk}\}$ where the latter has probability zero. In Section 4.1, we will estimate $\mathcal{U}$ using the posterior median model.

Remark 1: Equations (4)-(8) define a Gaussian diffusion process initiated at $\alpha_{u_0 k}$:

$$\eta_{uk} \mid \mathsf{others} \sim N\left(\sum_{u' \in a(u)} \xi_{u'k}, s_u \tau_{1k\ell_u} w_u\right),\ \mathsf{independently\ for}\ k \in [K-1], \tag{9}$$

for any non-root node $u \neq u_0$; also see the seminal formulation by Felsenstein (1985). To aid the understanding of this Gaussian diffusion prior, it is helpful to consider a special case of $s_u = 1$ and $\ell_u = 1$, $\forall u \in \mathcal{V}$. For two leaves $v, v' \in \mathcal{V}_L$, the prior correlation between η_{vk} and $\eta_{v'k}$ is

$$\mathrm{Corr}(\eta_{vk}, \eta_{v'k}) = \frac{\sum_{u \in a(v) \cap a(v')} w_u}{\{dist_{\mathcal{T}_w}(u_0, v) dist_{\mathcal{T}_w}(u_0, v')\}^{1/2}}, \tag{10}$$

When v and v' have the same number of ancesters ($|a(v)| = |a(v')|$) and all edges have identical weight $w_u = c, \forall u$, the prior correlation is the fraction of common ancestors. Note that $\boldsymbol{\eta}_v$ fully determines $\boldsymbol{\pi}_v$ in (3) and induces correlations among $\{\boldsymbol{\pi}_v, v \in \mathcal{V}_L\}$.

Remark 2: One reviewer raised an important question on the choice of encouraging prior correlation among $\{\boldsymbol{\pi}_v\}$ rather than among the latent class indicators $\{I_i^{(v)}\}$. In the present prior distribution, by integrating out $\{\boldsymbol{\pi}_v\}$, we have induced prior marginal correlation among $\{I_i^{(v)}\}$ for observation in nearby leaves. Additional prior correlation amongst the $\{I_i^{(v)}\}$ can be introduced via an additional layer of prior over the $\{I_i^{(v)}\}$ conditional on $\{\boldsymbol{\pi}_v\}$, e.g., through clustered samples. The absence of such clustered sampling structure in the motivating application points us towards the former simpler strategy.

Priors for class-specific response probabilities. Let $\gamma_{jk} = \log\{\theta_{jk}/(1-\theta_{jk})\}$. We specify

$$\gamma_{jk} \sim N(0, \tau_{2jk}), \text{ independently for feature } j \in [J] \text{ and class } k \in [K]. \tag{11}$$

Joint distribution. Let $\boldsymbol{\beta} = (\mathbf{Z}, \boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\varrho})$ collect all the unknown parameters where $\boldsymbol{s} = \{s_u : u \in \mathcal{V}\}$, $\boldsymbol{\gamma} = \{\gamma_{jk}, j \in [J]; k \in [K]\}$, $\boldsymbol{\alpha} = \{\alpha_{uk} : u \in \mathcal{V}, k \in [K-1]\}$, $\boldsymbol{\varrho} = (\rho_1, \ldots, \rho_L)^\intercal$, $\boldsymbol{a} = (a_1, \ldots, a_L)^\intercal$, and $\boldsymbol{b} = (b_1, \ldots, b_L)^\intercal$. Hereafter we use $\mathrm{pr}(A \mid B)$ to denote a probability density or mass function of quantities in A with parameters B; when B represents hyperparameters or given information in this paper, we simply use $\mathrm{pr}(A)$, e.g., we will use $\mathrm{pr}(\mathbf{Y}, \boldsymbol{\beta})$ to represent $\mathrm{pr}(\mathbf{Y}, \boldsymbol{\beta} \mid \boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{a}, \boldsymbol{b}, \mathcal{T}_w, \mathcal{L})$. The joint distribution of data and unknown quantities can thus be written as:

$$\mathrm{pr}(\mathbf{Y} \mid \boldsymbol{\beta})\mathrm{pr}(\boldsymbol{\beta}) = \prod_{v \in \mathcal{V}_L} \prod_{i=1}^{n_v} \prod_{k=1}^{K} \left[\{\sigma(\eta_{vk})\}^{\mathbf{1}\{k<K\}} \prod_{s<k} \{1 - \sigma(\eta_{vs})\} \prod_{j=1}^{J} \sigma\left(X_{ij}^{(v)} \gamma_{jk}\right) \right]^{Z_{ik}^{(v)}} \tag{12}$$

$$\times \prod_{u \in \mathcal{V}} \prod_{k=1}^{K-1} \left(\frac{1}{\sqrt{2\pi\tau_{1k\ell_u} w_u}} \exp\left\{ -\frac{1}{2\tau_{1k\ell_u} w_u} \alpha_{uk}^2 \right\} \right) \times \prod_{j=1}^{J} \prod_{k=1}^{K} \left(\frac{1}{\sqrt{2\pi\tau_{2jk}}} \exp\left\{ -\frac{1}{2\tau_{2jk}} \gamma_{jk}^2 \right\} \right)$$

$$\times \prod_{u \in \mathcal{V}} \rho_{\ell_u}^{s_u} (1 - \rho_{\ell_u})^{1-s_u} \cdot \prod_{\ell=1}^{L} \frac{1}{\mathsf{Beta}(a_\ell, b_\ell)} \rho_\ell^{a_\ell - 1} (1 - \rho_\ell)^{b_\ell - 1}, \tag{13}$$

where $X_{ij}^{(v)} = 2Y_{ij}^{(v)} - 1$. Tree information $\mathcal{T}_w$ enters the joint distribution in the definition of $\boldsymbol{\eta}_v$ (Equations (4)); sample-to-leaf indicators $\mathcal{L}$ choose among $\{\boldsymbol{\eta}_v, v \in \mathcal{V}_L\}$ for every observation in Equation (12). By setting $s_u = 0$ for all the non-root nodes in Equation (5), the classical LCM with a single $\boldsymbol{\pi} = \boldsymbol{\pi}_{u_0}$ results. Figure 2 shows a directed acyclic graph (DAG) that represents the model likelihood and prior specifications.

[Figure 2 about here.]

4. Variational Inference Algorithm

Calculating a posterior distribution often involves intractable high-dimensional integration over the unknowns in the model. Traditional sequential sampling approaches such as Markov chain Monte Carlo (MCMC) remains a widely used inferential tool based on approximate samples from the posterior distribution. They can be powerful in evaluating multidimensional integrals. However, they do not guarantee closed-form posterior distributions. Variational inference (VI) is a popular alternative to MCMC for approximating the posterior distribution and has been widely used in machine learning and gaining interest in statistics (e.g., Blei et al., 2017; Ormerod and Wand, 2010). In particular, VI has also been used for fitting the classical LCMs (e.g., Grimmer, 2011). VI requires a user-specified family of distributions that can be expressed in tractable forms while being flexible enough to approximate the true posterior; the approximating distributions and their parameters are referred to as "variational distributions" and "variational parameters", respectively. VI algorithms find the best variational distribution that minimizes the Kullback-Leibler (KL) distance between the variational family and the true posterior distribution. VI has been widely applied in Gaussian (Carbonetto and Stephens, 2012; Titsias and Lázaro-Gredilla, 2011) and binary likelihoods (e.g., Jaakkola and Jordan, 2000; Thomas et al., 2019). Also see Blei et al. (2017) for a detailed review. We use VI because it is fast, bypasses infeasible analytic integration or data augmentation that is otherwise needed for MCMC under Dirac spike components and prior-likelihood non-conjugacy (Tüchler, 2008), and enables data-driven selection of hyperparameters via approximate empirical Bayes (Equation (S8), Supporting Information). These advantages of VI are achieved at a cost of slight variance-covariance under-estimation, the degree of which we assess in Section 5.

We use VI algorithm to conduct inference using variational distributions factorized as:

$$q(\boldsymbol{\beta}) = q(\boldsymbol{\gamma}) \cdot \underbrace{\prod_{u \in \mathcal{V}} q(s_u, \boldsymbol{\alpha}_u)}_{q(\boldsymbol{s}, \boldsymbol{\alpha})} \cdot \underbrace{\prod_{v \in \mathcal{V}_L} \prod_{i=1}^{n_v} q(\boldsymbol{Z}_i^{(v)})}_{q(\mathbf{Z})} \cdot \underbrace{\prod_{\ell=1}^{L} q(\rho_\ell)}_{q(\boldsymbol{\varrho})}, \tag{14}$$

where $q(\boldsymbol{Z}_i^{(v)})$ is a multinomial distribution with variational parameters $\boldsymbol{r}_i^{(v)} = \left(r_{i1}^{(v)}, \ldots, r_{iK}^{(v)}\right)^{\mathsf{T}}$, and $r_{ik}^{(v)}$ represents the approximate posterior probability of observation i in leaf v belonging to class k and $\sum_{k=1}^{K} r_{ik}^{(v)} = 1$. Importantly, we make no other assumptions about the particular parametric form of variational distributions, which by the VI updating rules can be shown to take familiar distributional forms (see Appendix A).

VI finds q that minimizes the Kullback-Leibler (KL) distance between the variational family and the true posterior distribution: $\mathrm{KL}(q(\boldsymbol{\beta})||\mathrm{pr}(\boldsymbol{\beta} \mid \mathbf{Y})) = -\int q(\boldsymbol{\beta}) \log \left\{\frac{\mathrm{pr}(\boldsymbol{\beta}|\mathbf{Y})}{q(\boldsymbol{\beta})}\right\} d\boldsymbol{\beta}$. However, the KL distance depends on the intractable posterior distribution is not easily computed. Fortunately, based on a well-known equality $\log \mathrm{pr}(\mathbf{Y}) = \mathcal{E}(q) + \mathrm{KL}(q(\boldsymbol{\beta})\|\mathrm{pr}(\boldsymbol{\beta} \mid \mathbf{Y}))$, where $\mathcal{E}(q) = \int q(\boldsymbol{\beta}) \log \frac{\mathrm{pr}(\mathbf{Y},\boldsymbol{\beta})}{q(\boldsymbol{\beta})} d\boldsymbol{\beta}$ is referred to as evidence lower bound (ELBO) because $\log \mathrm{pr}(\mathbf{Y}) \geqslant \mathcal{E}(q)$. Because $\mathrm{pr}(\mathbf{Y})$ is a constant, minimizing the KL divergence is equivalent to maximizing $\mathcal{E}(q)$. The VI algorithm updates each component of $q(\boldsymbol{\beta})$ in turn while holding other components fixed. However, because of the nonlinear sigmoid functions in Equation (12), generic VI updating algorithms for $q(s_u, \boldsymbol{\alpha}_u)$ and $q(\boldsymbol{\gamma})$ involve integrating over random variables in the sigmoid function hence lack closed forms. To make the updates analytically tractable, we replace Equation (12) with an analytically tractable lower bound. In particular, we use a technique introduced by Jaakkola and Jordan (2000) which bounds the sigmoid function from below by a Gaussian kernel with a tuning parameter, hence affords closed-form VI updates; also see Durante et al. (2019) for a modern view of this technique as a bona fide mean-field approximation with Pòlya Gamma data augmentation. In particular, we will use the inequality

$$\sigma(x) \geqslant \sigma(\psi)\exp\{(x-\psi)/2 - g(\psi)(x^2-\psi^2)\} := h(x,\psi), \tag{15}$$

with $g(\psi) = \frac{1}{2\psi}[\sigma(\psi) - \frac{1}{2}]$ where ψ is a tuning parameter.

We approximate ELBO $\mathcal{E}(q)$ by $\mathcal{E}^*(q)$:

$$\mathcal{E}^*(q) := \int q(\boldsymbol{\beta}) \log \frac{h^*(\mathbf{X},\boldsymbol{\psi},\boldsymbol{\gamma},\mathbf{Z})h^{**}(\boldsymbol{\phi},\boldsymbol{s},\boldsymbol{\alpha},\mathbf{Z})\mathrm{pr}(\boldsymbol{s},\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\varrho})}{q(\boldsymbol{\beta})}\mathrm{d}\boldsymbol{\beta} \leqslant \mathcal{E}(q), \tag{16}$$

where $h^*(\mathbf{X},\boldsymbol{\psi},\boldsymbol{\gamma},\mathbf{Z}) = \prod_{v\in\mathcal{V}_L}\prod_{i=1}^{n_v}\prod_{k=1}^{K}\left\{\prod_{j=1}^{J} h\left(X_{ij}^{(v)}\gamma_{jk},\psi_{jk}\right)\right\}^{Z_{ik}^{(v)}}$, and $h^{**}(\boldsymbol{\phi},\boldsymbol{s},\boldsymbol{\alpha},\mathbf{Z}) = \prod_{v\in\mathcal{V}_L}\prod_{i=1}^{n_v}\prod_{k=1}^{K}\left\{\{h(\eta_{vk};\phi_k^{(v)})\}^{\mathbf{1}\{k<K\}}\prod_{m<k} h(-\eta_{vm};\phi_m^{(v)})\right\}^{Z_{ik}^{(v)}}$. The VI algorithm iterates until convergence to find the optimal variational distribution q that maximizes $\mathcal{E}^*(q)$. Because $\mathcal{E}^*(q) \leqslant \log\pi(\mathbf{Y})$, it can be viewed as an approximation to the marginal likelihood. We maximize over $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ to obtain the best approximation. In addition, we adopt an approximate empirical Bayes approach by optimizing the VI objective function $\mathcal{E}^*(q)$ over the hyperparameters $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$. Relative to specifying weakly informative but often non-conjugate hyperprior for the variance parameters, optimizing hyperparameter is more practically convenient (e.g., Thomas et al., 2019). Because updating the hyperparameters changes the prior, we need to update q, $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ again. This leads to an algorithm that alternates between maximizing $\mathcal{E}^*(q)$ in $(q,\boldsymbol{\psi},\boldsymbol{\phi})$ and in $(\boldsymbol{\tau}_1,\boldsymbol{\tau}_2)$ until convergence. We update the hyperparameters every d complete VI iterations. Pseudocode in Algorithm 1 outlines the VI updates; Appendix A1 details the exact updating formula.

4.1 *Posterior Summaries*

Two sets of point and interval estimates for $\{\boldsymbol{\pi}_v : v \in \mathcal{V}_L\}$ are available from the VI algorithm: 1) data-driven grouped ("fused") estimates $(\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}})$ that are formed by setting a subset of $\boldsymbol{s}$ to one and the rest to zero, and 2) leaf-specific estimates $(\widehat{\boldsymbol{\pi}}_v^{\mathsf{leaf}})$. For 1), we select the posterior median model by setting $s_u = 1$ for nodes in $\widehat{\mathcal{U}} = \{u : E_{q_t}[s_u] > 0.5\}$ (see Step 1b, Appendix A1). For leaves v and v', $\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}} = \widehat{\boldsymbol{\pi}}_{v'}^{\mathsf{dgrp}}$ if and only if $a(v) \cap \widehat{\mathcal{U}} = a(v') \cap \widehat{\mathcal{U}}$. Because no closed-form posterior distributions for $\boldsymbol{\pi}_v$ are readily available under logistic stick-breaking representation, we compute the approximate posterior mean and approximate 95% credible intervals (CrIs) by a Monte Carlo procedure after convergence of Algorithm 1. For $u \in \widehat{\mathcal{U}}$, we first draw $B = 10^5$ random independent samples of α_{uk} from $N(E_{q_t}[\alpha_{uk} \mid s_u = 1], V_{q_t}[\alpha_{uk} \mid s_u = 1])$, for $k \in [K-1]$. We then compute B corresponding $\boldsymbol{\pi}_v$ vectors based on Equations (3) to (5) with $s_u = \mathbf{1}\{u \in \widehat{\mathcal{U}}\}$ in (5). Finally, we compute the empirical means and 95% CrIs marginally for π_{vk}, $k \in [K]$. The above Monte Carlo procedure is extremely fast given only independent Gaussian samples are drawn. As a comparison, for 2), we define leaf-specific estimates $\widehat{\boldsymbol{\pi}}_v^{\mathsf{leaf}}$ by the mean of (3) where $\eta_{uk} \overset{d}{\sim} N(\sum_{u \in a(v)} E_{q_t}[s_u\alpha_{uk}], \sum_{u \in a(v)} V_{q_t}[s_u\alpha_{uk}])$, for $k \in [K]$. We also use Monte Carlo simulation to approximate the posterior means and 95% CrIs. In general, $\widehat{\boldsymbol{\pi}}_v^{\mathsf{leaf}}$ differ across the leaves. In contrast, the data-driven grouped estimates $\{\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}\}$ induce dimension reduction.

Prediction. The out-of-sample predictive probability of class k for a new observation nested in leaf v is $r_{i'k}^{(v)} := \mathrm{pr}(I_{i'}^{(v)} = k \mid Y_{i'}^{(v)}, \mathcal{D})$, where $\mathcal{D} = (\mathbf{Y}, \mathcal{T}_w, \mathcal{L}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\tau}_1, \boldsymbol{\tau}_2)$. We have

$$r_{i'k}^{(v)} = \int \underbrace{\mathrm{pr}(I_{i'}^{(v)} = k \mid \boldsymbol{\theta}_{\cdot k}, \boldsymbol{\pi}_v, Y_{i'}^{(v)}, \mathcal{D})}_{(i)} \underbrace{\mathrm{pr}(\boldsymbol{\theta}_{\cdot k}, \boldsymbol{\pi}_v \mid Y_{i'}^{(v)}, \mathcal{D})}_{(ii)} \mathrm{d}\boldsymbol{\theta}_{\cdot k} \mathrm{d}\boldsymbol{\pi}_v. \tag{17}$$

We approximate (17) by plug-in estimators: $\widehat{r}_{i'k}^{(v)} \propto \mathrm{pr}(Y_{i'}^{(v)} \mid I_{i'}^{(v)} = k, \widehat{\boldsymbol{\theta}}_{\cdot k}, \mathcal{T}_w) \cdot \widehat{\pi}_{vk}$, $k \in [K]$. This can be seen by noting that term $(i) \propto \mathrm{pr}(Y_{i'}^{(v)} \mid I_{i'}^{(v)} = k, \boldsymbol{\theta}_{\cdot k}, \mathcal{T}_w) \cdot \pi_{vk}$, and term $(ii) \approx \mathrm{pr}(\boldsymbol{\theta}_{\cdot k}, \boldsymbol{\pi}_v \mid \mathcal{D})$ which we approximate by a Dirac measure at $(\widehat{\boldsymbol{\theta}}_{\cdot k}, \widehat{\boldsymbol{\pi}}_v)$. Here $\widehat{\boldsymbol{\pi}}_v = \widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$.

Choice of K. In applications where data-driven selection of K is more desirable, we may follow Bishop (2006) and use criterion $\mathcal{E}_K^*(q) + \log(K!)$ where $\mathcal{E}_K^*(q)$ is the lower bound of log marginal data likelihood for a K-class model and the correction term is to make different models comparable (e.g., Grimmer, 2011, Section 5.2).

5. Simulation

5.1 *Design and Performance Metrics*

We conducted a simulation study to evaluate the performance of the proposed tree-integrative LCM. We compare our model to a few alternatives with ad hoc grouping of observations in terms of accuracy in estimating $\{\boldsymbol{\pi}_v, v \in V_L\}$. Data were generated under two scenarios with different class-specific response profiles $\boldsymbol{\Theta}$. Appendix A2 details the true parameter settings of the simulations. Figure 3(a) visualizes the tree $\mathcal{T}_w$ with equal edge weights and true leaf groups used in the simulation with $p_L = 11$ leaves and $G = 3$ groups.

We simulated $R = 200$ independent replicate data sets for different total sample sizes ($N = 1000, 4000$). For each N, we set $n_v \approx N/p_L$ for $\nu \in \mathcal{V}_L$ (with rounding where needed) to investigate balanced leaves and set n_v to be approximately $\frac{1}{5}N/p_L$ or $\frac{4}{5}N/p_L$ with equal chance for mimicking unbalanced observations across leaves. For observations in a leaf v, we simulate $\boldsymbol{Y}_i^{(v)}$ according to an LCM with class probabilities $\boldsymbol{\pi}_v$ and class-specific response probabilities $\boldsymbol{\Theta}$. We simulated data for different dimensions $J = 21, 84$, for $K = 3$ classes.

For each simulated data set, we fitted the proposed model, based on which we compute $\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$ and $\widehat{\boldsymbol{\pi}}_v^{\mathsf{leaf}}$ (see Section 4.1). Our primary interest is in $\{\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}\}$; $\{\widehat{\boldsymbol{\pi}}_v^{\mathsf{leaf}}\}$ are for comparisons. In addition, we also tested a few approaches based on ad hoc leaf node groupings: 1) True grouping analysis (fit separate LCMs to obtain estimates in each of the true groups); 2) Single group LCM analysis (omit sample-to-leaf indicators $\mathcal{L}$, hence the tree information); 3) Ad hoc grouping 1 (manual grouping coarser than the true grouping); 4) Ad hoc grouping 2: classical LCMs for data on each leaf. All analyses assume $\boldsymbol{\Theta}$ does not vary by leaves.

We used three model performance metrics. First, we computed the root mean squared errors (RMSE) for an estimate $\widehat{\boldsymbol{\pi}}_v$ where $\mathsf{RMSE}(\widehat{\boldsymbol{\pi}}_\nu) = \sqrt{(Kp_L)^{-1}\sum_{k=1}^K \sum_{v \in V_L}\{\widehat{\pi}_{vk} - \pi_{vk}\}^2}$. Second, we compared the true and the estimated leaf groupings via adjusted Rand Index (ARI, Hubert and Arabie, 1985). ARI is a chance-corrected index that takes value between

-1 and 1 with values closer to 1 indicating better agreement. Finally, we estimated the coverage probability of the approximate 95% CrIs. For each true group g, we compute the frequency of the approximate 95% CrI (computed along with $\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$) containing the truth, conditional on the event that an estimated partition of the leaf nodes includes g.

5.2 *Simulation Results*

Figure 3 shows comparisons among the RMSEs for different models under different scenarios. For sample sizes $N = 1000$ and $N = 4000$, the proposed methods with data-driven grouping ($\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$) produced similar or better RMSE than analyses based on ad hoc leaf groupings, which restrict leaves into incorrect groupings that are coarser (single LCM and ad hoc grouping 1) or finer (ad hoc grouping 2) than the truth. The proposed approach ($\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$) achieved similar RMSE as $\widehat{\boldsymbol{\pi}}_v^{\mathsf{leaf}}$, indicating little accuracy was lost in exchange for dimension reduction. The RMSEs of $\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$ were similar to estimates of $\boldsymbol{\pi}_v, v \in \mathcal{V}_L$ obtained from analyses based on the true leaf grouping. Indeed, the accuracy of group discovery increased with sample sizes with other settings fixed. Average ARIs across replications for each scenario were high (0.94 to 0.99) indicating good recovery of the true leaf groups. Although the groups discovered were not perfect, the comparable RMSEs suggest desirable adaptability of the proposed approach in effective collapsing of the leaves. The RMSE for $\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$ was smaller than analyses based on a refined leaf-level grouping: smaller sample sizes in the leaves resulted in loss of efficiency in separate estimations of $\boldsymbol{\pi}_v$ across leaves. RMSEs were further reduced under a larger J or balanced sample sizes in the leaves. However, we again observed similar relative advantage of the proposed $\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$. The relative comparisons of RMSEs under less discrepant true class-specific response profiles remained similar (see Appendix Figure S2).

The observed coverage rates of the approximate 95% CrIs achieved the nominal level satisfactorily (see Appendix Figure S1). Slight under-coverage occurred under smaller N, unbalanced sample sizes, smaller J and leaf groups with smaller number of observations.

This is partially a consequence of VI as an inner approximation to the posterior distribution which may underestimate the posterior uncertainty (e.g., Chapter 10, Bishop, 2006).

Finally, we also considered scenarios where only a single group of leaves is present in truth for which the classical LCM is perfectly appropriate. Appendix Figure S3 shows, by learning the posterior node-specific slab-versus-spike selection probabilities, the proposed model produces similar RMSEs as the classical LCM.

[Figure 3 about here.]

6. *E. Coli* Data Application

6.1 *Background and Data*

E. coli infections cause millions of urinary tract infections (UTIs) in the US each year (e.g., Johnson and Russo, 2002). Many studies have shown that extraintestinal pathogenic *E. coli* (ExPEC) strains routinely colonize food animals and contaminate the food supply chain serving as a likely link between food-animal *E. coli* and human UTIs (e.g., Johnson et al., 2005). The scientific team adopted a novel strategy of augmenting fine-scale core-genome phylogenetics with interrogation of accessory host-adaptive MGEs (see Section 1.1). The scientific goal is to accurately estimate the probabilities of *E. coli* isolates with human and non-human host-origins across genetically diverse but related *E. coli* sequence types (STs).

We restrict our analysis to $N = 2,663$ *E. coli* isolates in a well-defined collection from humans and retail meat obtained over a 12-month period in Flagstaff, Arizona, US. Each isolate belongs to one of $p_L = 133$ different STs (leaves in the phylogenetic tree) that are identified via a multilocus sequence typing scheme based on short-read DNA sequencing. A total of $J = 17$ MGEs were curated and associated with functional annotations. Each ST was represented by at least four isolates. We constructed rooted, maximum-likelihood phylogenies using core-genome SNP data for the 133 STs. Figure 4 shows the estimated phylogenetic tree for the STs where the edge lengths represents the substitution rate in the conserved core genome. Every ST is overlaid in the same row with the empirical frequencies of 1) $J = 17$ MGEs and 2) the observed sources (human clinical or meat samples) which may differ from the true host origin. The observed frequencies of the MGEs vary greatly across lineages. We apply the proposed tree-integrative LCM to 1) estimate the probabilities of *unobserved* human and non-human host-origins for all *E. coli* STs with data-driven groupings of the STs for dimension reduction; and 2) to produce isolate-level probabilistic host-origin assignment. The context of the study restricts us to assume the host origin of each isolate is in one of two unobserved class of human vs food animals. A subset of preliminary data is analyzed in this paper for illustrating the proposed method. Inclusion of additional samples and/or MGEs may change findings. The final results and the detailed workflow of MGE discovery will be reported elsewhere.

[Figure 4 about here.]

6.2 *Data Results*

The proposed approach produces estimated class-specific response profiles ($\widehat{\boldsymbol{\theta}}_{\cdot k}, k = 1, 2$) that exhibit differential enrichment of MGEs (Figure 5(b)). For example, MGEs 3, 10 to 17 are estimated with probability of between 0.15 and 0.71 being present in class 1, with log odds ratios (LORs; class 1 vs class 2: $LOR(\widehat{\theta}_{j1}, \widehat{\theta}_{j2})$) greater than one. The functional annotations of these MGEs reveal that class 1 is likely associated with food-animal hosts. In contrast, MGEs 4 to 9 are estimated to be present in class 2 with probability between 0.35 to 0.82 with LORs greater than one relative to the corresponding estimated response probabilities in class 1. The results suggest the MGEs are highly associated with different types of host-origins.

[Figure 5 about here.]

The proposed approach discovered 21 ST leaf groups, for which distinct estimated vectors of the latent class probabilities $\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$ are shown in Figure 5(a). For many estimated ST groups, the class probabilities are almost entirely dominated by one type of host-origin. For example, the estimated ST Group 1 (38 leaves; 649 samples; class 1 probability 0.98, 95%

CrI: $(0.97, 0.99)$) and Group 3 (31 leaves; 422 samples; class 1 probability 0.97, 95% CrI: $(0.96, 0.98)$) showed high probabilities of non-human (class 1) host-origin of *E. coli.* The results suggest recent cross-species transmissions were rare among multiple nearby lineages.

We also compared against results based on two fixed and more restrictive leaf groups, (a) classical LCM (one leaf group); (b) four leaf groups selected by the scientific team (Appendix Figure S4). The single LCM (a) estimated the probability of class 1 to be $0.60, 95\%$ CrI $: (0.58, 0.62)$. The ad hoc leaf grouping (b) produced coarser estimates relative to the proposed $\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$ which identified four local leaves (ST1141, ST10, ST744 and ST5996) comprising 116 samples that have estimated probability of class 1: 0.74 $(0.66, 0.82)$). This highlights the inability of potentially misspecified leaf groups to uncover subtle local variations in the latent class probabilities. We compared these models via 10-fold cross-validation based on the mean predictive log-likelihood (MPL) of the test data, which is computed by plugging in the estimated latent class probabilities and response probability profiles. Of note, because of small sample sizes in some leaves, a naive cross-validation may by chance result in a training set without any observation in some leaves. We therefore randomly keep two observations per leaf and use one random fold of the remaining samples as test data. The proposed approach (with posterior median node selection) achieves the highest MPL (-2015.48) compared to (a) (-2030.15) and (b)(-2162.45). The estimates of response probability profiles are similar.

On an individual isolate level, the proposed model can estimate the probability that an isolate was derived from a particular host. For example, by incorporating additional observed sample source information, we can compute "posterior concordance probability (PCP)" for each observation. In particular, PCP, $r^{(v)}_{i,S^{(v)}_i}$, is defined as the approximate posterior probability of the true host origin agreeing with the *observed* sample source category $S^{(v)}_i$ of the same *E. coli* isolate (e.g., $S^{(v)}_i = 1$ for meat and 2 for human clinical samples). Figure 5(c) shows the histogram of PCPs for all the isolates. Small PCPs, e.g., below a user-specified threshold of 0.5, indicate likely recent host jumps which may subject to further examination to estimate the timing of host transmissions based on *in vitro* stability data of each MGE.

7. Discussion

In this paper, we proposed a tree-integrative LCM for analyzing multivariate binary data. We formulated the motivating scientific question in terms of inferring latent class probabilities that may vary in different parts of a tree. We proposed a Gaussian diffusion prior for logistic stick-breaking parameterized latent class probabilities and designed a scalable approximate algorithm for posterior inference. Our *E. coli* data analysis revealed that multiple MGEs are disproportionately associated with specific host origins. Combined with external sample source information, the model can help identify isolates that underwent recent host jump, paving the way for further isolate-level host origin validation.

Our study has some limitations. First, the MGE data we analyzed may represent a fraction of the host-associated accessory elements. By design, additional accessory elements identified in future studies can be readily integrated and evaluated in the proposed framework. Second, host-associated accessory elements are lost and gained over time as *E. coli* strains transition across hosts. For infections that were zoonotic in nature, we did not observe how much time had lapsed between the cross-species host jump and the actual infection. Our model partly accounted for these uncertainties by the imperfect positive response probabilities. However, the timings may drive the presence or absence of multiple MGEs, resulting in potential statistical dependence given the true class of host-origin. Deviations from local independence assumption may impact model-based inference (e.g., Pepe and Janes, 2006; Albert and Dodd, 2004). In practice, a subset of samples with ascertained host-origins may provide critical information to estimate the conditional dependence structure.

Further model extensions may improve model applicability. First, when a subset of observations is not mapped in the tree at random, the algorithm can add additional unobserved leaf indicators to be inferred along with other parameters. Second, it is important to note that the tree integrated into LCM in general is estimated with uncertainty in the topological structure. Methods that use an additional layer of prior over the tree space centered around the estimated tree may account for the upstream uncertainty (e.g., Willis and Bell, 2018). Third, *E. coli* isolates may vary in additional factors such as the hosts' clinical characteristics. Regression extensions may refine the understanding of variation in latent class probabilities and positive response probabilities that are driven by covariates (e.g., Huang and Bandeen-Roche, 2004). Fourth, LCM is an example of probability tensor decomposition methods (e.g., Johndrow et al., 2017), the tree-integrative LCM motivates extensions to general graph-guided probability tensor decomposition methods. Finally, the truncated stick-breaking formulation in Equation (3) motivates connections to a broader class of covariate-indexed dependent process priors as K approaches infinity (e.g., Rodriguez and Dunson, 2011; Ren et al., 2011). Extensions along this line may also relax the present assumption of identical number of realized classes at additional computational cost.

Without relying on prior-likelihood conjugacy, neuronized priors for Bayesian sparse linear regression has been proposed (Shin and Liu, 2021). Comparative studies against spike-and-slab priors are warranted. One known drawback of mean field VI is that it tends to underestimate the marginal posterior variances of parameters. In our simulations, we showed near nominal coverages of the true $\boldsymbol{\pi}$ with slight undercoverages happening mostly for leaf groups with very small sample sizes. It is an interesting line of work to incorporate the methods of Giordano et al. (2015) to correct the variance-covariance matrices used in the component variational distributions. We leave these topics for future work.

Acknowledgment

The research is supported in part by a Precision Health Investigator Award from University of Michigan, Ann Arbor (ML, ZW); an award from Wellcome Trust (LBP, MA and CML;

award number 201866); and National Institutes of Health (NIH) grants R01AR073208 (ZW), P30CA04659 (ZW), and 1R01AI130066-01A1 (LBP). We also thank the Co-Editor, Associate Editor and two referees whose comments greatly improved the presentation of our work.

Data Availability Statement

An `R` package "`lotR`" is freely available at `https://github.com/zhenkewu/lotR`. The data that support the findings in this paper are available from the corresponding author upon reasonable request.

References

Airoldi, E. M. and Bischof, J. M. (2016). Improving and evaluating topic models and other models of text. *Journal of the American Statistical Association* **111,** 1381–1403.

Albert, P. S. and Dodd, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* **60,** 427–435.

Avila, D., Keiser, O., Egger, M., Kouyos, R., Böni, J., Yerly, S., et al. (2014). Social meets molecular: combining phylogenetic and latent class analyses to understand hiv-1 transmission in switzerland. *American journal of epidemiology* **179,** 1514–1525.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association* **92,** 1375–1386.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112,** 859–877.

Carbonetto, P. and Stephens, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis* **7,** 73–108.

Dunson, D. and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104,** 1042–1051.

Durante, D., Rigon, T., et al. (2019). Conditionally conjugate mean-field variational bayes for logistic models. *Statistical science* **34,** 472–485.

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist* **125,** 1–15.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association* **87,** 476–486.

Ghahramani, Z., Jordan, M. I., and Adams, R. P. (2010). Tree-structured stick breaking for hierarchical data. In *Advances in Neural Information Processing Systems*, pages 19–27.

Giordano, R., Broderick, T., and Jordan, M. (2015). Linear response methods for accurate covariance estimates from mean field variational bayes. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 1441–1449.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61,** 215–231.

Grimmer, J. (2011). An introduction to Bayesian inference via variational approximations. *Political Analysis* **19,** 32–47.

Huang, G.-H. and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika* **69,** 5–32.

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

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* **10,** 25–37.

Johndrow, J. E., Bhattacharya, A., and Dunson, D. B. (2017). Tensor decompositions and sparse log-linear models. *Annals of statistics* **45,** 1.

Johnson, J. R., Delavari, P., O'Bryan, T. T., Smith, K. E., and Tatini, S. (2005). Contamination of retail foods, particularly turkey, from community markets (Minnesota, 1999–2000) with antimicrobial-resistant and extraintestinal pathogenic Escherichia coli. *Foodbourne Pathogens & Disease* **2,** 38–49.

Johnson, J. R. and Russo, T. A. (2002). Extraintestinal pathogenic Escherichia coli:"the other bad E. coli". *Journal of Laboratory and Clinical Medicine* **139,** 155–162.

Lazarsfeld, P. F. (1950). The logical and mathematical foundations of latent structure analysis. In Stouffer, S., editor, *The American Soldier: Studies in Social Psychology in World War II*, volume IV, pages 362–412. Princeton University Press, Princeton, NJ.

Lindsay, J. A. and Holden, M. T. (2004). Staphylococcus aureus: superbug, super genome? *Trends in microbiology* **12,** 378–385.

Liu, C. M., Stegger, M., Aziz, M., Johnson, T. J., Waits, K., Nordstrom, L., et al. (2018). Escherichia coli ST131-H22 as a foodborne uropathogen. *mBio* **9,** e00470–18.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences* **95,** 3140–3145.

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *The American Statistician* **64,** 140–153.

Pepe, M. S. and Janes, H. (2006). Insights into latent class analysis of diagnostic test performance. *Biostatistics* **8,** 474–484.

Price, L. B., Hungate, B. A., Koch, B. J., Davis, G. S., and Liu, C. M. (2017). Colonizing opportunistic pathogens (cops): the beasts in all of us. *PLoS pathogens* **13,** e1006369.

Ranganath, R., Tang, L., Charlin, L., and Blei, D. (2015). Deep exponential families. In *Artificial Intelligence and Statistics*, pages 762–771. PMLR.

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

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

Roy, D. M., Kemp, C., K, M. V., and Tenenbaum, J. B. (2006). Learning annotated hierarchies from relational data. In *Advances in Neural Information Processing Systems*, pages 475–482.

Scornavacca, C., Delsuc, F., and Galtier, N. (2020). *Phylogenetics in the Genomic Era.* No commercial publisher — Authors open access book.

Shin, M. and Liu, J. S. (2021). Neuronized priors for bayesian sparse linear regression. *Journal of the American Statistical Association* pages 1–43.

Thomas, E. G., Trippa, L., Parmigiani, G., and Dominici, F. (2019). Estimating the effects of fine particulate matter on 432 cardiovascular diseases using multi-outcome regression with tree-structured shrinkage. *Journal of the American Statistical Association* pages 1–11.

Titsias, M. and Lázaro-Gredilla, M. (2011). Spike and slab variational inference for multi-task and multiple kernel learning. *Advances in Neural Information Processing Systems* **24,** 2339–2347.

Tüchler, R. (2008). Bayesian variable selection for logistic models using auxiliary mixture sampling. *Journal of Computational and Graphical Statistics* **17,** 76–94.

Willis, A. and Bell, R. (2018). Uncertainty in phylogenetic tree estimates. *Journal of Computational and Graphical Statistics* **27,** 542–552.

Supporting Information

Web Appendices and Figures referenced in Sections 4, 5 and 6, and R programs are available with this paper at the *Biometrics* website on Wiley Online Library.

Algorithm 1: Pseudocode for Variational Algorithm to Integrate Sample Similarities into Latent Class Analysis

```
Tree-Encoded Information and Data:
  (a) A weighted rooted tree $\mathcal{T}_w = (\mathcal{T} = (\mathcal{V}, E), w)$: leaves $\mathcal{V}_L \subset \mathcal{V}$, edge lengths $\boldsymbol{w} = (w_u)_{u \in \mathcal{V}}$;
  (b) The leaf id for each observation $\mathcal{L}$;
  (c) Multivariate binary data $\mathbf{Y}$ (organize the observations with the same leaf id into
      consecutive rows: a total of $n_v$ observations in leaf $v$, $v \in \mathcal{V}_L$; The leaf in the $s$-th row of
      $\mathbf{Y}$ is $v_s$, $s \in [N]$.)
Fixed Hyperparameters:
  (a') The number of classes $K \geqslant 2$; levels $\ell_u \in [L]$ for all nodes $u \in \mathcal{V}$;
  (b') Hyperparameters for the prior probability of $s_u = 1$: $(a_\ell, b_\ell)$, $\ell \in [L]$.
Initialize:
  (a'') $t \longleftarrow 0$; Initialize $q_t(\boldsymbol{s}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$        // (see Step 0 in Appendix A1)
  (b'') Set an initial ELBO $\mathcal{E}^*_0 \longleftarrow 0$
$t \longleftarrow 1$; $\mathcal{E}^*_1 \longleftarrow \mathcal{E}^*_0 + 2\epsilon$
while $|\mathcal{E}^*_t - \mathcal{E}^*_{t-1}| > \epsilon$ do
    $q_t(\boldsymbol{s}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \longleftarrow q_{t-1}(\boldsymbol{s}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$
    $\boldsymbol{\phi}^{(t)} \longleftarrow \boldsymbol{\phi}^{(t-1)}$; $\boldsymbol{\psi}^{(t)} \longleftarrow \boldsymbol{\psi}^{(t-1)}$
    $\boldsymbol{\tau}_1^{(t)} \longleftarrow \boldsymbol{\tau}_1^{(t-1)}$; $\boldsymbol{\tau}_2^{(t)} \longleftarrow \boldsymbol{\tau}_2^{(t-1)}$
    for $v \in \mathcal{V}_L$ do
        for $i \in [n_v]$ do
            for $k \in [K]$ do
                $r_{ik}^{(v),(t)} \longleftarrow \operatorname{argmax}_{r_{ik}^{(v)}} \mathcal{E}^*_t(q)$        // (See Step 1a in Appendix A1)
    $q_t(\boldsymbol{\gamma}) \longleftarrow \operatorname{argmax}_{q_t(\boldsymbol{\gamma})} \mathcal{E}^*_t(q)$        // (see Step 1b in Appendix A1)
    for $u \in \mathcal{V}$ do
        $q_t(s_u, \boldsymbol{\alpha}_u) \longleftarrow \operatorname{argmax}_{q_t(s_u, \boldsymbol{\alpha}_u)} \mathcal{E}^*_t(q)$        // (see Step 1b in Appendix A1)
    for $\ell \in [L]$ do
        $q_t(\rho_\ell) \longleftarrow \operatorname{argmax}_{q_t(\rho_\ell)} \mathcal{E}^*_t(q)$        // (see Step 1c in Appendix A1)
    for $k \in [K]$ do
            // update local variational parameters for tighter lower bounds
        for $v \in \mathcal{V}_L$ do
            $\phi_k^{(v),(t)} \longleftarrow \operatorname{argmax}_{\phi_k^{(v)}} \mathcal{E}^*_t(q)$
        for $j \in [J]$ do
            $\psi_{jk}^{(t)} \longleftarrow \operatorname{argmax}_{\psi_{jk}} \mathcal{E}^*_t(q)$        // (see Step 2 in Appendix A1)
    if $t \ mod \ d = 0$ then
        for $k \in [K]$ do
            for $\ell \in [L]$ do
                $\tau_{1kl}^{(t)} \longleftarrow \operatorname{argmax}_{\tau_{1kl}} \mathcal{E}^*_t(q)$
            for $j \in [J]$ do
                $\tau_{2jk}^{(t)} \longleftarrow \operatorname{argmax}_{\tau_{2jk}} \mathcal{E}^*_t(q)$        // (see Step 3 in Appendix A1)
    $\mathcal{E}^*_t \longleftarrow ELBO(q_t)$        // (see Step 4 in Appendix A1)
    $t \longleftarrow t + 1$
Return: $q_{t-1}(\boldsymbol{\gamma})$, $q_{t-1}(\boldsymbol{s}, \boldsymbol{\alpha})$, $\{q_{t-1}(\boldsymbol{Z}_i^{(v)})\}$, $q_{t-1}(\boldsymbol{\varrho})$, $\{\mathcal{E}^*_1, \ldots, \mathcal{E}^*_{t-1}\}$
```

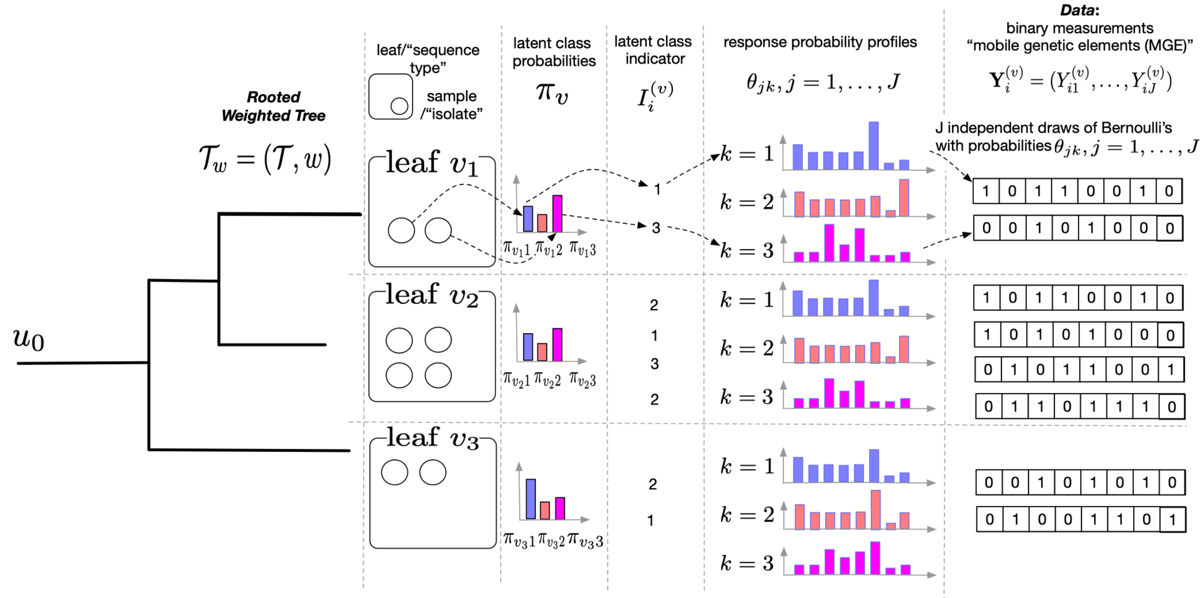

Rooted Weighted Tree
$\mathcal{T}_w = (\mathcal{T}, w)$
u_0
leaf/"sequence type"
sample /"isolate"
latent class probabilities
π_v
latent class indicator
$I_i^{(v)}$
response probability profiles
$\theta_{jk}, j = 1, \ldots, J$
Data:
binary measurements
"mobile genetic elements (MGE)"
$\mathbf{Y}_i^{(v)} = (Y_{i1}^{(v)}, \ldots, Y_{iJ}^{(v)})$
J independent draws of Bernoulli's with probabilities $\theta_{jk}, j = 1, \ldots, J$
leaf v_1
$\pi_{v_1 1}$ $\pi_{v_1 2}$ $\pi_{v_1 3}$
1
3
$k = 1$
$k = 2$
$k = 3$
1 0 1 1 0 0 1 0
0 0 1 0 1 0 0 0
leaf v_2
$\pi_{v_2 1}$ $\pi_{v_2 2}$ $\pi_{v_2 3}$
2
1
3
2
$k = 1$
$k = 2$
$k = 3$
1 0 1 1 0 0 1 0
1 0 1 0 1 0 0 0
0 1 0 1 1 0 0 1
0 1 1 0 1 1 1 0
leaf v_3
$\pi_{v_3 1}$ $\pi_{v_3 2}$ $\pi_{v_3 3}$
2
1
$k = 1$
$k = 2$
$k = 3$
0 0 1 0 1 0 1 0
0 1 0 0 1 1 0 1

Figure 1: Schematic representation of a hypothetical rooted weighted tree with three leaves and data generated based on the proposed model with $K = 3$ latent classes, $n_{v_1} = 2$, $n_{v_2} = 4$ and $n_{v_3} = 2$, $J = 8$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

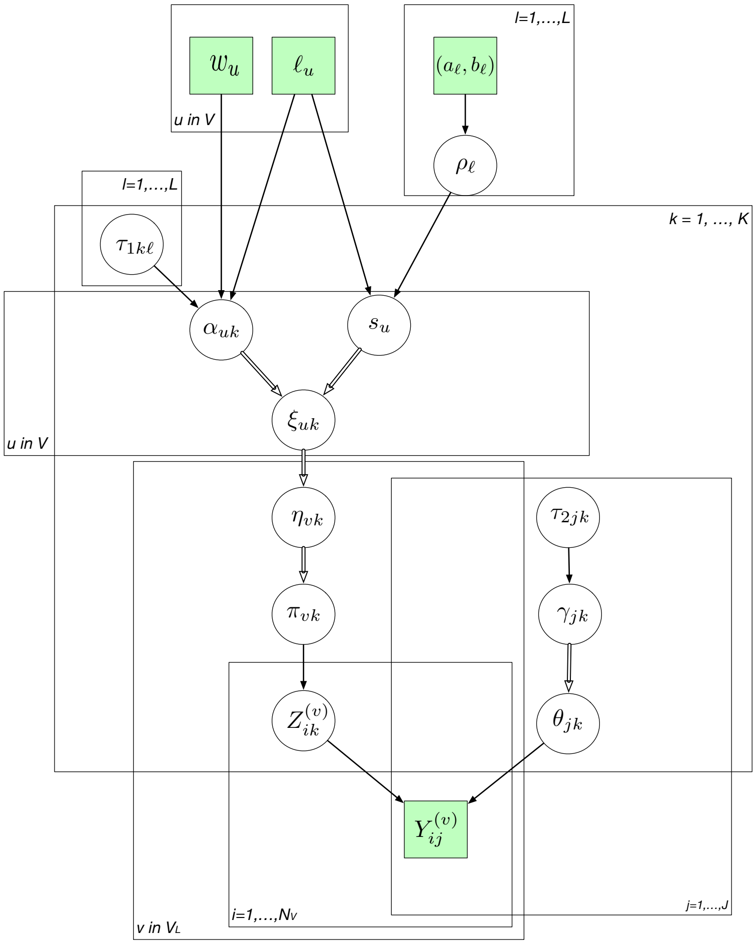
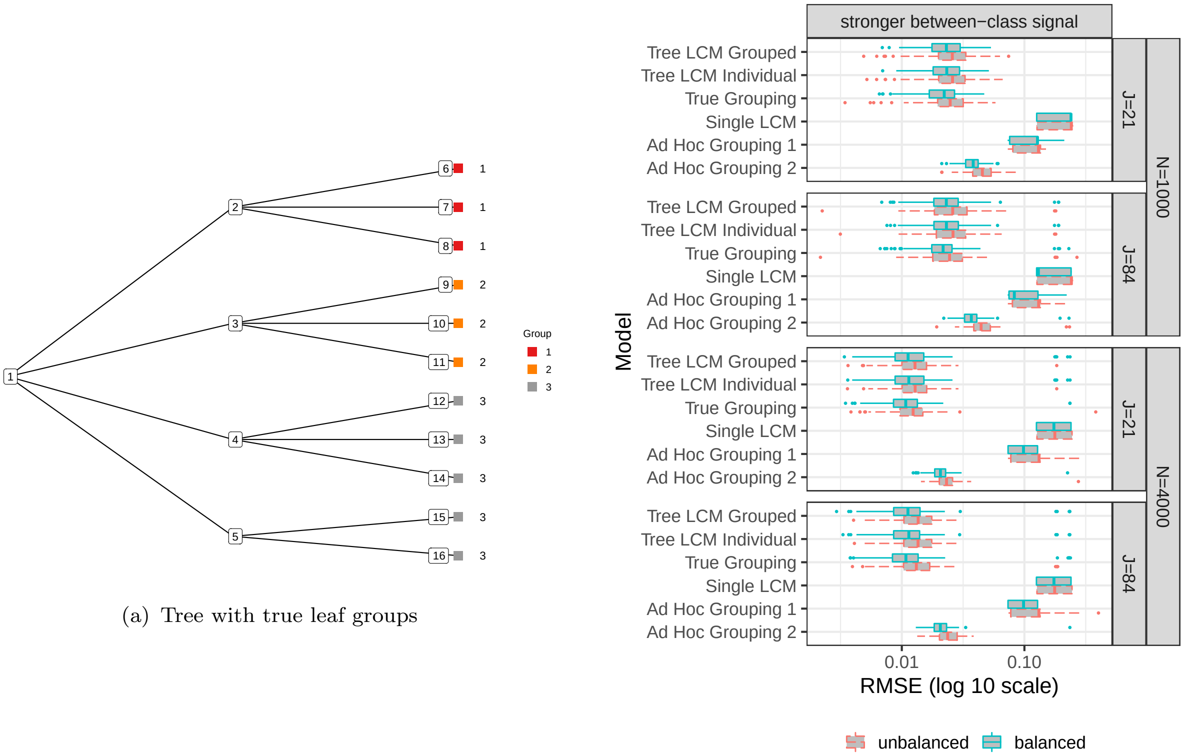


Figure 2: The directed acyclic graph (DAG) representing the structure of the model likelihood and priors. The quantities in squares are either data or hyperparameters; the unknown quantities are shown in the circles. The arrows connecting variables indicate that the parent parameterizes the distribution of the child node (solid lines) or completely determines the value of the child node (double-stroke arrows). The rectangular "plates" where the variables are enclosed indicate that a similar graphical structure is repeated over the index; The index in a plate indicate nodes, hyperparameter levels, leaves, subjects, classes and features. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.



(a) Tree with true leaf groups

(b) RMSE comparisons across multiple models and scenarios

Figure 3: Simulation studies show the proposed model produces grouped estimates $\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$ with similar or smaller RMSEs compared to alternatives (see Section 5). This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

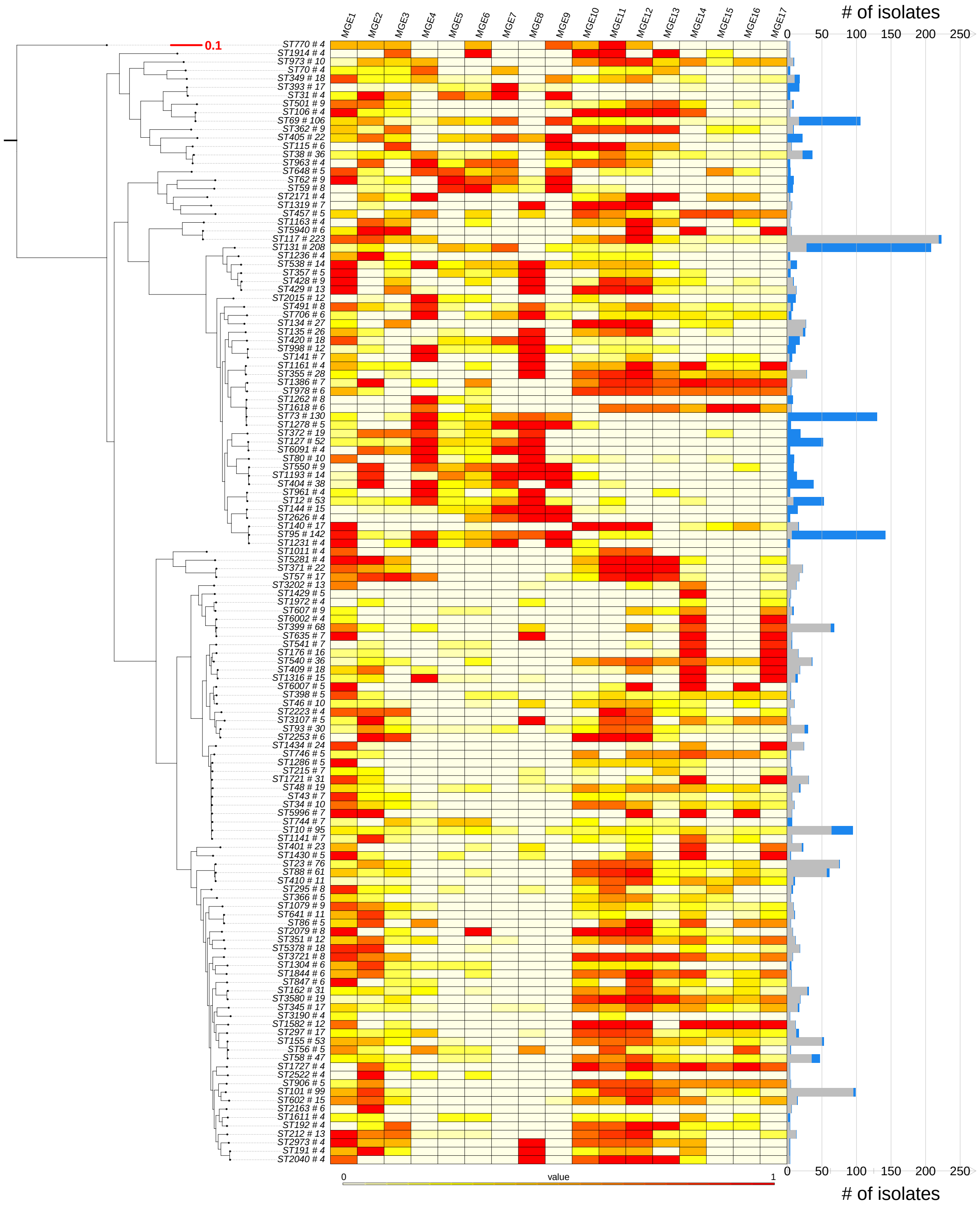
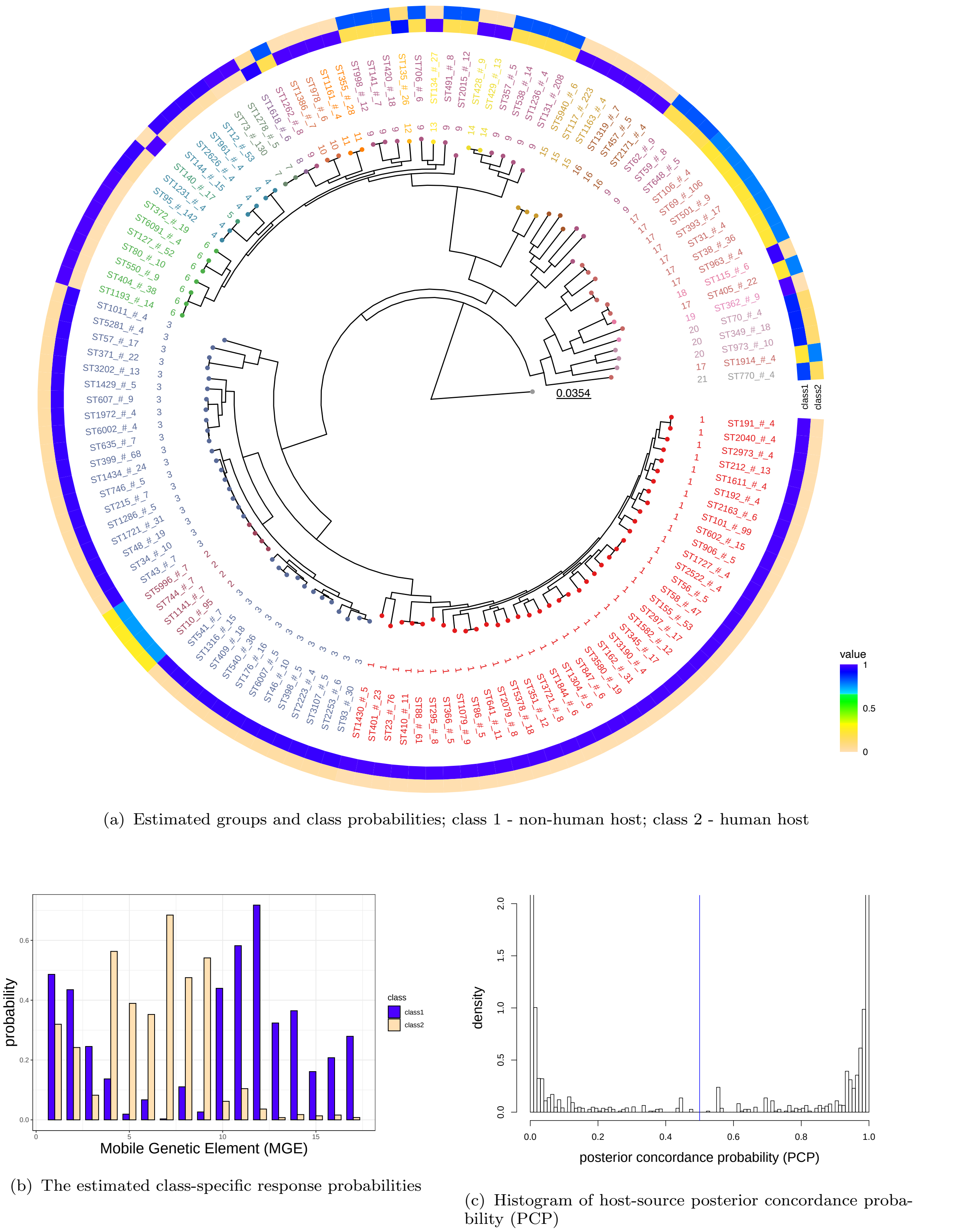


Figure 4: The empirical frequencies for $J = 17$ MGEs within each ST mapped in the core-genome phylogenetic tree. The red scale bar represents the substitution rate in the conserved core genome. The bars on the right indicate the total number isolates of each ST; the gray and blue bars represent the number of isolates obtained from apparent non-human and human sources, respectively. The core-genome phylogenetic tree on the left margin maps $N = 2,663$ *E. coli* isolates into $p_L = 133$ STs (leaves). This figure appears in color in the electronic version of this article, and any mention of color refers to that version.



(a) Estimated groups and class probabilities; class 1 - non-human host; class 2 - human host

(b) The estimated class-specific response probabilities

(c) Histogram of host-source posterior concordance probability (PCP)

Figure 5: a) Data results with estimated leaf groups and latent class probabilities by group. ST names (ST_#_isolates) are aligned to the tips of the circular tree, which are colored by discovered leaf groups. The scale bar represents the substitution rate in the conserved core genome. The circular heatmap shows the estimated latent class probabilities ($\widehat{\boldsymbol{\pi}}_v^{\mathsf{dgrp}}$, $v \in \mathcal{V}_L$); b) and c): see the captions of the subfigures. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.