


Augmentation Samplers for Multinomial Probit Bayesian Additive Regression Trees

Yizhen Xu, Joseph Hogan, Michael Daniels, Rami Kantor & Ann Mwangi

To cite this article: Yizhen Xu, Joseph Hogan, Michael Daniels, Rami Kantor & Ann Mwangi (05 Aug 2024): Augmentation Samplers for Multinomial Probit Bayesian Additive Regression Trees, Journal of Computational and Graphical Statistics, DOI: [10.1080/10618600.2024.2388605](https://doi.org/10.1080/10618600.2024.2388605)

To link to this article: <https://doi.org/10.1080/10618600.2024.2388605>

 View supplementary material [↗](#)

 Accepted author version posted online: 05 Aug 2024.

 Submit your article to this journal [↗](#)

 Article views: 18

 View related articles [↗](#)

 View Crossmark data [↗](#)

Augmentation Samplers for Multinomial Probit Bayesian Additive Regression Trees

Yizhen Xu^{a,*,#}, Joseph Hogan^b, Michael Daniels^c, Rami Kantor^d, Ann Mwangi^e

^aDivision of Biostatistics, University of Utah

^bDepartment of Biostatistics, Brown University

^cDepartment of Statistics, University of Florida

^dDivision of Infectious Diseases, Brown University

^eCollege of Health Sciences, School of Medicine, Moi University

#xuyizhen00@gmail.com

*The authors gratefully acknowledge that *funding of this work is provided by the US National Institutes of Health (NIH) under R01 AI136664, R01 AI108441, R01 AI167694, and P30 AI 42853.*

Accepted Manuscript

Abstract

The multinomial probit (MNP) (Imai and van Dyk, 2005) framework is based on a multivariate Gaussian latent structure, allowing for natural extensions to multilevel modeling. Unlike multinomial logistic models, MNP does not assume independent alternatives. Kindo et al. (2016) proposed multinomial probit BART (MPBART) to accommodate Bayesian additive regression trees (BART) formulation in MNP. The posterior sampling algorithms for MNP and MPBART are collapsed Gibbs samplers. Because the collapsing augmentation strategy yields a geometric rate of convergence no greater than that of a standard Gibbs sampling step, it is recommended whenever computationally feasible (Liu, 1994a; Imai and van Dyk, 2005). While this strategy necessitates simple sampling steps and a reasonably fast converging Markov chain, the complexity of the stochastic search for posterior trees may undermine its benefit. We address this problem by sampling posterior trees conditional on the constrained parameter space and compare our proposals to that of Kindo et al. (2016), who sample posterior trees based on an augmented parameter space. We also compare to the approach by Sparapani et al. (2021) that specified the multinomial model in terms of conditional probabilities. In terms of MCMC convergence and posterior predictive accuracy, our proposals are comparable to the conditional probability approach and outperform the augmented tree sampling approach. We also show that the theoretical mixing rates of our proposals are guaranteed to be no greater than the augmented tree sampling approach. Appendices and codes for simulations and demonstrations are available online.

Keywords: keyword: Data Augmentation, Categorical Outcomes, Latent Models

1 Introduction

Bayesian additive regression trees (BART) (Chipman et al., 2010) is a flexible nonparametric Bayesian approach for regression on a recursively binary-partitioned predictor space; it uses sum-of-trees to model the mean function such that nonlinearities and interactions along with additive effects are naturally accounted for, and regularization priors are imposed to favor shallow trees to reduce over-fitting. There has been considerable literature on extending BART to various types of outcome variables (Bonato et al., 2011; Low-Kam et al., 2015; Sparapani et al., 2016; Waldmann, 2016; Henderson et al., 2020; Linero et al., 2021; Um et al., 2022). We consider the extension of BART to multinomial probit models (Imai and van Dyk, 2005) (MNP). Existing BART-related work has developed efficient Markov chain Monte Carlo (MCMC) algorithms for Gaussian likelihoods, which naturally adapt to frameworks with Gaussian-distributed latent variables. However, careful consideration of data augmentation (DA) schemes is needed to ensure the computational efficiency of implementing BART under the multinomial probit framework. The main contributions of this paper are to provide a detailed review of sampling algorithms for parameter expansion that are based on DA schemes and to introduce a set of new MCMC algorithms for multinomial probit BART (MPBART).

The motivation for this work stems from the necessity to develop accurate predictive models for patient engagement in HIV care (Gardner et al., 2011; WHO, 2012). This requires taking into account death and transfer out of care as competing endpoints (Lee et al., 2017). These models are used to characterize patient transition through the HIV cascade, which describes essential stages of the HIV care continuum: (a) HIV diagnosis through testing, (b) linkage to care, (c) engagement in care, (d) initiation of antiviral therapy (ART) through retention, and (e) sustained suppression of viral load. The care cascade framework has been widely used as a monitoring and evaluation tool for improving and managing HIV healthcare systems. In Section 4, we will demonstrate and compare different algorithms for using multinomial BART models to characterize engagement and retention in HIV care.

MNP (Imai and van Dyk, 2005) and multinomial logistic (McFadden, 1974) (MNL) regression models are widely used tools for predicting and describing the relationships of explanatory variables to multinomial outcomes. Kindo et al. (2016) proposed the MPBART framework that fits BART to the multivariate Gaussian latent variables in the MNP. Related work incorporating BART into categorical response models is introduced by Murray (2020), where BART is extended to log-linear models that include multinomial logistic BART (MLBART). Both MNP and MNL regression can be derived from a latent variable framework, where each outcome category is a manifestation of a latent utility that depends on covariates. The observed categorical outcome is the utility-maximizing category. MNP and MNL regression assume the latent utility distribution to be multivariate Gaussian and independent extreme-value distribution, respectively. The MNP formulation is appealing because it incorporates between-category dependence, a feature that extends naturally to MPBART. We will show that allowing non-zero correlations between latent variables can have a substantial impact on predictive accuracy.

There are two difficulties in sampling from posterior distributions of MNP. First, a closed-form expression for the multinomial outcome's marginal distribution is not available; second, the identifiability of the MNP model requires constraints on the covariance matrix of the latent variables, hindering specification of conjugate distributions and making posterior sampling challenging. There has been considerable work on Bayesian sampling techniques to

address these computational issues based on DA-related methods (Albert and Chib, 1993; McCulloch and Rossi, 1994; Nobile, 1998; McCulloch et al., 2000; Imai and van Dyk, 2005). The original DA algorithm (Tanner and Wong, 1987) is a stochastic generalization of the EM algorithm (Dempster et al., 1977). Marginal data augmentation (MDA) (Meng and Van Dyk, 1999; Liu and Wu, 1999; Van Dyk and Meng, 2001) generalizes and accelerates the DA algorithm via parameter expansion such that full conditionals are easier to sample from and expansion parameter(s) are subsequently marginalized over. Heuristically, the MDA Gibbs sampler can traverse the parameter space more efficiently with the extra variation induced by the expansion parameter(s), resulting in possible computational gains, including a faster mixing rate (Meng and Van Dyk, 1999; Liu and Wu, 1999). The MDA scheme circumvents the difficulties in sampling from a constrained parameter space and allows an easier and more efficient joint sampling of expansion parameters and transformed model parameters. Imai and van Dyk (2005) unified several previous proposals under the umbrella of MDA, examined different prior specifications of the model parameters, and outlined two adaptations of the MDA scheme for posterior sampling of the MNP based on parameter expansion.

Building upon the work of Imai and van Dyk (2005), Kindo et al. (2016) proposed an algorithm, which we refer to as KD, for fitting the MPBART. Our own implementation of KD yielded oversized posterior trees from overfitting and difficulty in posterior convergence. We therefore propose two alternative procedures for fitting the MPBART that have simpler algorithmic structure, improved convergence in the sum-of-trees and the covariance matrix, and a mixing rate at least as good as the original procedure when the Markov chain reaches equilibrium. Our algorithms show better out-of-sample accuracy and stability in predictive tasks under various settings when evaluated in terms of posterior predictive distribution and posterior mode. The posterior mode accuracy is commonly used as an evaluation metric in supervised learning literature (Kindo et al., 2016). Our proposals are based on the idea of fitting the sum of trees in a normalized parameter space to reduce disruptions to the stochastic search of posterior trees, resulting in a less difficult convergence of the Markov chain.

In every step of the Gibbs sampler, the MDA scheme requires (1) the joint sampling of expansion parameter(s) and transformed model parameters, and (2) the marginalization over the expansion parameter. However, the two actions are not always feasible for complicated Gibbs sampling problems. For example, sampling the functional mean component jointly with an expansion parameter in an MPBART algorithm is difficult because posterior trees are sampled by stochastic search. Thus, instead of MDA schemes, algorithms for MNP and MPBART are in fact partially marginalized augmentation (PMA) samplers (van Dyk, 2010), which relaxes the fully marginalized structure of the MDA and can lead to improvements in convergence rate when more steps involve joint sampling and marginalization of the expansion parameter(s)' components. Contrary to the intuition regarding PMA samplers that more augmented posterior sampling steps are associated with improved posterior convergence, we illustrate that when sophisticated Metropolis-Hastings or stochastic search is involved in complex samplers, certain steps may be sensitive to or undermined by the incorporation of expansion parameters. This motivates the need for new algorithm design considerations.

This paper is structured as follows. Section 2.1 describes the formulation of MNP and MPBART frameworks; Section 2.2 reviews sampling schemes for the MNP, including DA and MDA; Section 2.3 connects the sampling schemes to the algorithms for fitting the MNP; Section 2.4 describes the existing algorithms and introduces our new proposals for fitting the

MPBART; and Section 2.5 provides a theoretical evaluation of different MPBART algorithms in terms of the mixing rate under stationarity. Section 3 compares multiple BART-related multinomial outcome models, including our proposals, on simulated datasets under different settings, and Section 4 demonstrates the comparison on a real-world dataset from a large HIV care program in Kenya. Section 5 summarizes the conclusions.

2 Method

2.1 General Background

For the categorical outcome S , which takes value in $\{0, \dots, C\}$, the general latent variable framework for multinomial models assumes that S is a manifestation of unobserved latent utilities $Z = (Z_0, \dots, Z_C)^T \in \mathbb{R}^{C+1}$, where $S = S(Z) = \arg \max_l Z_l$, i.e. $S = k$ if $Z_k > Z_l$ for all $l \neq k$. In general, C is the number of outcome categories minus one. The framework requires normalization for identifiability because S is invariant to a translation or a scaling (by a positive constant) of Z . Without loss of generality, we assume that the reference outcome category is 0; the normalization is achieved by first characterizing S as a function of latent variables $W = (W_1, \dots, W_C)^T \in \mathbb{R}^C$, such that $W_l = Z_l - Z_0$ and

$$S(W) = \begin{cases} l & \text{if } \max(W) = W_l \geq 0 \\ 0 & \text{if } \max(W) < 0. \end{cases} \quad (1)$$

The MNP models W in terms of covariates X and accounts for correlation across outcome levels by assuming W follows a multivariate normal model

$$W(X) \sim \text{MVN}(G(X; \theta), \Sigma), \quad (2)$$

where $G(X; \theta) = (G_1(X; \theta_1), \dots, G_C(X; \theta_C))^T$, $\theta = (\theta_1, \dots, \theta_C)^T$ and $\Sigma = \{\sigma_{ij}\}$ is a $C \times C$ positive definite symmetric matrix.

Identifiability of the model requires normalizing the scale of W because by definition the outcome S is invariant to a multiplication of W by any positive constant. From (2), the normalization for scale occurs by imposing a constraint on the covariance matrix Σ , i.e. $\text{trace}(\Sigma) = C$ (Burgette and Nordheim, 2012). To illustrate, suppose there are latent variables \tilde{W} such that

$$\tilde{W}(X) \sim \text{MVN}(G(X; \tilde{\theta}), \tilde{\Sigma}), \quad (3)$$

where $\tilde{W}(X) = \bar{\alpha}W(X)$, $G(X; \tilde{\theta}) = \bar{\alpha}G(X; \theta)$, $\tilde{\Sigma} = \bar{\alpha}^2\Sigma$, and $\bar{\alpha} > 0$. By (1), \tilde{W} and W yield the same S . However, if Σ satisfies the trace constraint, W is the normalized counterpart of \tilde{W} and $\bar{\alpha}^2 = \text{trace}(\Sigma) / C$ is a positive scalar that ensures a one-to-one mapping from W to \tilde{W} .

Direct posterior sampling of parameters in (2) is difficult due to the constraint on Σ . A technique for easier sampling is to augment the parameter space such that it is possible to specify a conjugate prior so that target parameters can be obtained by converting samples

back to the normalized scale. The obvious choice of augmented parameter space is the one without the normalization for scale, i.e. $(\tilde{w}, \tilde{\theta}, \tilde{\Sigma})$ in (3). Imai and van Dyk (2005) suggested a constrained inverse Wishart prior for Σ such that its joint distribution with α^2 is equivalent to the unconstrained covariance matrix having prior distribution $\tilde{\Sigma} \sim \text{inv-Wishart}(\nu, \Psi)$. This makes it possible to sample easily from the conditional posterior of $\tilde{\Sigma}$. Setting $\nu = C + 1$ and Ψ to be an identity matrix is equivalent to sampling the corresponding correlations of $\tilde{\Sigma}$ from a uniform distribution. When $\nu > C + 1$, the expectation of $\tilde{\Sigma}$ has a closed form $E(\tilde{\Sigma}) = \Psi / (\nu - C - 1)$.

The standard framework for MNP regression assumes a linear model specification for each $w_l(X)$, i.e. $G_l(X; \theta_l) = X \theta_l$ for $l = 1, \dots, C$. Kindo et al. (2016) proposed MPBART to increase the predictive power and the flexibility in dealing with complicated nonlinear and interaction effects. The innovative idea is to approximate each mean component of $w(X)$

using a sum of m trees, $G_l(X; \theta_l) = \sum_{k=1}^m g(X; \theta_{lk})$, where $l = 1, \dots, C$ and θ_{lk} is the set of

parameters corresponding to the k th binary tree for the l th latent variable, $w_l(X)$. MPBART uses the same Bayesian regularization prior on the trees to restrict over-fitting as in Chipman et al. (2010). An important contribution of Kindo et al. (2016) is deriving from (2) the conditional distribution for Gibbs sampling of each individual tree, and embedding it into the backfitting procedure of BART. See Chipman et al. (1998) and Chipman et al. (2010) for details on the BART backfitting procedure.

2.2 Review of Data Augmentation

The goal of data augmentation (DA) schemes is to draw samples of (Y, ϕ) , where Y and ϕ represent the augmented data and model parameters, respectively. The sampling algorithm Kindo et al. (2016) have for MPBART heavily relies on Imai & van Dyk's (Imai and van Dyk, 2005) work on fitting the MNP, which explores different Gibbs samplers of (w, θ, Σ) under the umbrella of marginal data augmentation (MDA) (Meng and Van Dyk, 1999; Liu and Wu, 1999), an extension and improvement of the DA algorithm (Tanner and Wong, 1987). This section provides a brief overview of relevant developments on the DA algorithm for fitting the MPBART.

Basic data augmentation. For any variable X , let $f(X)$ denote the density function of X . We illustrate the task of sampling (Y, ϕ) under the DA algorithm of Tanner and Wong (1987):

Scheme [DA]

1. Draw $Y \sim f(Y | \phi)$.
2. Draw $\phi \sim f(\phi | Y)$.

Marginalized data augmentation (MDA). The basic idea of MDA versus DA is to expand the model and overparameterize $f(Y, \phi)$ to $f(Y, \phi, \alpha)$; the expansion parameter α often corresponds to a transformation of Y and/or ϕ . For example, α may index a transformation

of Y to $\tilde{Y} = t_\alpha(Y)$ where t_α is one-to-one and differentiable, thereby expanding the model from $f(Y, \phi)$ to $f(\tilde{Y}, \phi, \alpha)$. The choice to sample from $f(Y, \phi, \alpha)$ or $f(\tilde{Y}, \phi, \alpha)$ depends on the specific model, and they are usually interchangeable. This approach is appealing when sampling from $f(Y, \alpha | \phi)$ or $f(\tilde{Y}, \alpha | \phi)$ is easier than the sampling of Y alone. Liu and Wu (1999) and Meng and Van Dyk (1999) simultaneously developed MDA. Liu and Wu (1999) provided theoretical results on the convergence rate of the MDA. Meng and Van Dyk (1999) introduced the MDA under two augmentation schemes, *grouping* and *collapsing* (Liu, 1994a; Liu et al., 1994); both procedures lead to the same distribution of (Y, ϕ) as Scheme [DA].

MDA with grouping. The grouping scheme samples conditionally on the expansion parameter α , while the collapsing scheme integrates α out from the joint distribution. MDA under the grouping scheme is preferred when the sampling of Y or ϕ jointly with α is easier than that in Scheme [DA]. For example, when $f(\phi | Y, \alpha)$ is easier to sample than $f(\phi | Y)$, and $f(Y, \alpha | \phi)$ is easy to sample, the sampler can “group” Y and α together and treats them as a single component,

Scheme [MDA-G]

1. Draw $(Y, \alpha) \sim f(Y, \alpha | \phi)$.
2. Draw $\phi \sim f(\phi | Y, \alpha)$.

MDA with collapsing. MDA under the collapsing scheme “collapses down” α by integrating it out from the joint distributions, i.e. $Y \sim f(Y | \phi) = \int f(Y | \phi, \alpha) f(\alpha | \phi) d\alpha$ and $\phi \sim f(\phi | Y) = \int f(\phi | Y, \alpha) f(\alpha | Y) d\alpha$. The implementation is as follows:

Scheme [MDA-C]

1. Draw $(Y, \alpha) \sim f(Y, \alpha | \phi)$ by $\alpha \sim f(\alpha | \phi)$ and $Y \sim f(Y | \phi, \alpha)$.
2. Draw $(\phi, \alpha) \sim f(\phi, \alpha | Y)$ by $\alpha \sim f(\alpha | Y)$ and $\phi \sim f(\phi | Y, \alpha)$.

Notice that the newly sampled α is discarded in each step of the Scheme [MDA-C]. In practice, it may be reasonable to assume a priori independence between ϕ and α because ϕ are parameters identified from the observed data, which does not contain information on α . Furthermore, given that transforming the augmented data Y is of interest, it may be true that the conditional sampling of model parameters ϕ is more plausible under \tilde{Y} than Y . Accordingly, Scheme [MDA-C] can be rewritten as:

Scheme [MDA-C']

1. Draw (\tilde{Y}, α) by drawing $\alpha \sim f(\alpha)$ and then $\tilde{Y} \sim f(\tilde{Y} | \phi, \alpha)$, and compute $\tilde{Y} = t_\alpha(Y)$.
2. Draw (ϕ, α) by drawing $\alpha \sim f(\alpha | \tilde{Y})$ and then $\phi \sim f(\phi | \tilde{Y}, \alpha)$.

The $f(\alpha)$ and $f(\alpha | \tilde{Y})$ are the prior and posterior (under the transformed augmented data) of α , respectively. The optimality of MDA under the collapsing scheme (Scheme [MDA-C]) over the DA algorithm (Scheme [DA]) in terms of convergence rate is proven in Meng and Van Dyk (1999) and Liu and Wu (1999). Liu and Wu (1999) also introduced Scheme [MDA-LW], which is equivalent to Scheme [MDA-C'] in terms of the sampling distribution and rate of convergence. This scheme is implicitly applied in the algorithms for fitting the MNP and MPBART, typically in the normalization of model parameters after each round of Gibbs sampling. Structurally, Scheme [MDA-LW] is in the form of Scheme [DA] with an additional intermediate step, which makes more clear the connection between the MDA and the DA algorithm:

Scheme [MDA-LW]

1. Draw $Y \sim f(Y | \phi)$.
2. Draw $\alpha_1 \sim f(\alpha)$, compute $\tilde{Y} = t_{\alpha_1}(Y)$; draw $\alpha_2 \sim f(\alpha | \tilde{Y})$, compute $Y' = t_{\alpha_2}^{-1}(\tilde{Y})$.
3. Draw $\phi \sim f(\phi | Y')$.

Note that Y and Y' follow the same distribution. The intuition behind the improvement of Scheme [MDA-LW] compared to the DA algorithm is that the intermediate step of sampling from Y' allows the sampler for ϕ to explore the expanded model space with more freedom.

2.3 Data Augmentation for the MNP

For fitting the MNP, Imai and van Dyk (2005) introduced two algorithms for the Gibbs sampling of (w, θ, Σ) , which we refer to as IvD1 and IvD2. The IvD1 modifies Scheme [MDA-C'] by expanding the model to $(\tilde{w}, \tilde{\theta}, \tilde{\Sigma}, \alpha)$ such that \tilde{w} and $(\tilde{\theta}, \tilde{\Sigma})$ correspond to \tilde{Y} and ϕ , respectively, and $\alpha = (\alpha_1, \alpha_2, \alpha_3)$:

Algorithm [IvD1]

1. Draw (\tilde{w}, α_1) by drawing $\alpha_1 \sim f(\alpha | \Sigma)$ and $w \sim f(w | \theta, \Sigma)$, and compute $\tilde{w} = \alpha_1 w$.
2. Draw $(\tilde{\theta}, \alpha_2)$ by drawing $\alpha_2 \sim f(\alpha | \tilde{w}, \Sigma)$ and then $\tilde{\theta} \sim f(\tilde{\theta} | \alpha_2, \tilde{w}, \Sigma)$, and compute $\theta = \tilde{\theta} / \alpha_2$.
3. Draw $(\tilde{\Sigma}, \alpha_3)$ by $\tilde{\Sigma} \sim f(\tilde{\Sigma} | \tilde{w} - X \tilde{\theta})$ and compute $\alpha_3 = \sqrt{\text{trace}(\tilde{\Sigma}) / C}$.

Using $\tilde{\Sigma}$ and α_3 from Step 3, we can compute the normalized covariance matrix $\Sigma = \tilde{\Sigma} / \alpha_3^2$ and use it in Steps 1 and 2 of the next round of posterior sampling; this is analogous to having α_3 index a one-to-one mapping from the expanded model space $(\tilde{\Sigma})$ to the normalized space (Σ) . Steps 1 and 3 in Algorithm [IvD1] collapse down α_1 and α_3 , but Algorithm [IvD1] is not a direct implementation of the MDA as in Scheme [MDA-C'] because Step 1 is

conditional on θ , or equivalently $(\tilde{\theta}, \alpha_2)$ where $\theta = \tilde{\theta} / \alpha_2$. Hence, Step 2 does not integrate out (collapse down) α_2 .

Standard MDA (Schemes [MDA-C] and [MDA-C']) are collapsed Gibbs samplers that integrate out expansion parameter(s) by redrawing and discarding α in every step. Algorithm [IvD1] is a partially marginalized augmentation (PMA) (van Dyk, 2010) procedure that relaxes the restrictive structure of full marginalization in MDA. PMA allows the conditional distribution in a k th step of the Gibbs sampler to depend on expansion parameter(s) drawn in other steps. Algorithms for fitting the MPBART in Section 2.4 are also PMA procedures.

IvD1 can also be viewed from a different perspective. Due to the linearity in model specification of the MNP, i.e. $G_l(X; \theta_l) = X \theta_l$ for $l = 1, \dots, C$, the linear relationship between θ and $\tilde{\theta}$ holds in Step 2 of Algorithm [IvD1], and it is equivalent to direct sampling of θ from $f(\theta | \tilde{W} / \alpha_2, \Sigma)$. Hence, IvD1 can be rearranged as follows:

Algorithm [IvD1']

1. Draw $w \sim f(w | \theta, \Sigma)$.
2. Draw $\alpha_1 \sim f(\alpha | \Sigma)$, compute $\tilde{w} = \alpha_1 w$; draw $\alpha_2 \sim f(\alpha | \tilde{w}, \Sigma)$, compute $w' = \tilde{w} / \alpha_2$.
3. Draw $\theta \sim f(\theta | w', \Sigma)$.
4. Draw Σ by $\tilde{\Sigma} \sim f(\tilde{\Sigma} | \tilde{w} - X \tilde{\theta})$, compute

$$\alpha_3 = \sqrt{\text{trace}(\tilde{\Sigma}) / C}, \text{ and } \Sigma = \tilde{\Sigma} / \alpha_3^2, \text{ where } \tilde{\theta} = \alpha_2 \theta.$$

The first three steps are equivalent to sampling $f(w, \theta | \Sigma)$ in Scheme [MDA-LW]. Step 4 collapses down α_3 , but the fact that Step 4 is conditional on (α_1, α_2) through $(\tilde{w}, \tilde{\theta})$ makes IvD1 not a collapsed Gibbs sampler collectively. IvD2 is given as follows:

Scheme [IvD2]

1. Draw $(\tilde{\epsilon}, \alpha_1)$ by $\alpha_1 \sim f(\alpha | \Sigma)$ and $w \sim f(w | \theta, \Sigma)$, compute $\tilde{\epsilon} = \alpha_1 [w - G(X; \theta)]$.
2. Draw (Σ, α_3) by $\tilde{\Sigma} \sim f(\tilde{\Sigma} | \tilde{\epsilon})$, compute

$$\alpha_3 = \sqrt{\text{trace}(\tilde{\Sigma}) / C}, \text{ and } \Sigma = \tilde{\Sigma} / \alpha_3^2.$$

3. Draw $\theta \sim f(\theta | w, \Sigma)$.

IvD2 separates the sampling into two parts, $(\tilde{\epsilon}, \Sigma)$ and θ ; the first part utilizes the MDA under Scheme [MDA-C] and the second part is a standard Gibbs sampling draw.

Theoretically, as stated in Imai and van Dyk (2005), IvD1 and IvD2 have the same lag-one

autocorrelation when the MCMC chain is stationary. However, they showed through numerical experiments that IvD1 is better than IvD2 in estimating the MNP in terms of being less sensitive to the starting values of (θ, Σ) . In the next section, we describe Kindo et al.'s algorithm (KD) and our two new proposals and connect them to the schemes reviewed here.

2.4 Algorithms for Posterior Sampling Algorithms of MPBART

For ease of notation, let $W_{i,-l} = (W_{i1}, \dots, W_{i,l-1}, W_{i,l+1}, \dots, W_{iC})$ and let $\mu = G(X; \theta) \in \mathbb{R}^c$ be the sum-of-trees component under the normalization of scale. We start with describing Algorithm [KD] and the two proposals, Algorithms [P1] and [P2], under the expanded model (W, μ, Σ, α) as follows, where W is the normalized latent variables with distribution $MVN(\mu, \Sigma)$ and α is the vector of expansion parameters indexed as in Algorithm [IvD1].

Steps	1	2	3
Algorithm [KD]	$(W, \alpha_1) \mu, \Sigma$	$\mu (W, \alpha_1), \Sigma$	$(\Sigma, \alpha_3) (W, \alpha_1), \mu$
Algorithm [P1]	$(W, \alpha_1) \mu, \Sigma$	$\mu W, \Sigma$	$(\Sigma, \alpha_3) (W, \alpha_1), \mu$
Algorithm [P2]	$W \mu, \Sigma$	$\mu W, \Sigma$	$(\Sigma, \alpha_3) W, \mu$

We make a few observations about these three algorithms: (a) Algorithm [KD] groups W and α_1 together, as in Scheme [MDA-G]; (b) Algorithm [P1] is structurally equivalent to Scheme [IvD2]; and (c) the sampling of the normalized covariance matrix in all three algorithms integrates out α_3 as in Scheme [MDA-C], i.e. $\Sigma \sim \int f(\Sigma, \alpha_3 | W, \mu) d\alpha_3$ in Algorithm [P2], and $\Sigma \sim \int f(\Sigma, \alpha_3 | W, \alpha_1, \mu) d\alpha_3$ in Algorithms [KD] and [P1]. In detail, Kindo et al.'s algorithm for fitting the MPBART can be summarized as the following augmented Gibbs sampler:

Algorithm [KD]

1. Sample $(\tilde{W}, \alpha_1^2) | (\mu, \Sigma, S)$.

(a) Draw α_1^2 from its conditional prior $f(\alpha^2 | \Sigma) = \text{trace}[\Psi \Sigma^{-1}] / \chi_{vc}^2$;

(b) for each l , update w_{il} conditional on $w_{i,-l}$, μ , Σ , and the observed outcome s_i , from a truncated normal distribution; and

(c) transform w_i and Σ to $\tilde{w}_i = \alpha_1 w_i$ and $\tilde{\Sigma}^* = \alpha_1^2 \Sigma$.

2.

Sample $\tilde{\theta} | (\tilde{W}, \alpha_1^2, \Sigma)$.

(a) Draw $\tilde{\theta} \sim f(\tilde{\theta} | \tilde{W}, \tilde{\Sigma}^*)$; and

(b) set $\tilde{\mu} = G(X; \tilde{\theta})$ and $\mu = \tilde{\mu} / \alpha_1$.

3.

Sample $(\Sigma, \alpha_3^2) | (\tilde{W}, \tilde{\theta})$.

(a) Draw $\tilde{\Sigma} \sim \text{Inv-Wishart}(N + \nu, \Psi + \sum_{i=1}^N \tilde{\epsilon}_i \tilde{\epsilon}_i^T)$, where $\tilde{\epsilon}_i = \tilde{W}_i - \tilde{\mu}_i$;

(b) set $\alpha_3^2 = \text{trace}(\tilde{\Sigma}) / C$; and

(c) set $\Sigma = \tilde{\Sigma} / \alpha_3^2$ and $W = \mu + \tilde{\epsilon} / \alpha_3$.

Step 1 jointly samples from $f(\tilde{W}, \alpha_1^2 | \mu, \Sigma, S)$ by first drawing the expansion parameter α_1^2 from its prior distribution $f(\alpha^2 | \Sigma)$, and then computing $\tilde{W} = \alpha^2 W$ where W is sampled from $f(W | \mu, \Sigma, S)$. Step 1(a) samples α_1^2 such that $\alpha_1^2 / \text{trace}[\Psi \Sigma^{-1}]$ follows an inverse-chi-squared distribution with νC degrees of freedom. Step 1(b) samples each w_{ii} from a truncated normal distribution described in Appendix D.1 based on (1), as the observed outcome s_i imposes an interval constraint on w_i , e.g. if s_i equals the reference level 0, then w_{ii} 's are right truncated at 0. Step 2 samples posterior trees across multivariate mean components by Gibbs sampling and each posterior tree is sampled as in regular BART. Step 3 computes α_3 using the sampled $\tilde{\Sigma}$ and then normalizes the scale of the model by Step 3(c).

Notice that the sampling of model parameters $\tilde{\theta}$ is conditional on $(\tilde{W}, \tilde{\Sigma}^*)$, which is equivalent to conditioning on $(\tilde{W}, \alpha_1^2, \Sigma)$ or (W, α_1^2, Σ) ; this observation is essential to the analysis of Algorithm [KD] in Section 2.5. Algorithm [KD] is closely related to IvD1 (Algorithm [IvD1]) but different in that it does not update the expansion parameter α_2 as in Step (b) of IvD1. This is analogous to having α_2 in IvD1 set to the sampled value of α_1 from Step (a). The reason for this modification is that the posterior tree parameters in BART, denoted by θ , are drawn via stochastic search; it would be extremely challenging to derive an analytical expression for $f(\alpha | \tilde{W}, \Sigma)$ from $\int f(\alpha, \theta | \tilde{W}, \Sigma) d\theta$ as in MNP because the specification is no longer linear in θ .

In the first step, \tilde{w} is a scaled version of W through $\tilde{w} = \alpha_1 W$. From (3), fitting the sum-of-trees component to \tilde{w} is analogous to sampling the parameters in an un-normalized space. We adopt four proposals for the posterior sampling of trees: GROW, PRUNE, CHANGE, and SWAP. Other tree sampling proposals such as perturbation and rotation (Pratola, 2016) are not considered here. Stochastic search in a massive space of possible tree structures can

be challenging when \tilde{w} , the quantity to which the sum-of-trees is fitting, is unstable. Heuristically, we would expect fitting the sum-of-trees component to W , which is a normalized quantity, instead of \tilde{w} to be more stable, induce better posterior convergence, and improve the prediction accuracy. Given these considerations, we modify Algorithm [KD] and propose the following:

Algorithm [P1]

1. Sample $(W, \alpha_1^2) | (\mu, \Sigma, S)$.

(a) Draw α_1^2 from its conditional prior $f(\alpha^2 | \Sigma) = \text{trace}[\Psi \Sigma^{-1}] / \chi_{\nu_C}^2$;

(b) for each l , update w_{il} conditional on $w_{i,l-1}$, μ , Σ , and s_i , from a truncated normal distribution; and

(c) transform w_i to $\tilde{w}_i = \alpha_1 w_i$.

2.

Sample $\theta | (W, \Sigma)$. Draw $\theta \sim f(\theta | W, \Sigma)$ and then set $\mu = G(X; \theta)$.

3.

Sample $(\Sigma, \alpha_3^2) | (W, \alpha_1, \theta)$.

(a) Draw $\tilde{\Sigma} \sim \text{Inv-Wishart}(N + \nu, \Psi + \sum_{i=1}^N \tilde{\epsilon}_i \tilde{\epsilon}_i^T)$, where $\tilde{\epsilon}_i = \tilde{w}_i - \alpha_1 \mu_i$;

(b) set α_3^2 to $\text{trace}(\tilde{\Sigma}) / C$; and

(c) set $\Sigma = \tilde{\Sigma} / \alpha_3^2$ and $W = \mu + \tilde{\epsilon} / \alpha_3$.

In the first proposal (Algorithm [P1]), the expansion parameters (α_1, α_3) do not affect the sampling of the trees in Step 2. If the order of Steps 2 and 3 are swapped, it becomes Scheme [IvD2] in Section 2.2. Algorithms [KD] and [P1] are the respective MPBART analogs of IvD1 and IvD2 for the MNP. Imai and van Dyk (2005) expected IvD1 to outperform IvD2 for the MNP and demonstrated through simulations. While for MPBART, we find Algorithm [P1] to be equal or superior to Algorithm [KD] theoretically (Section 2.5) and computationally (Sections 3 and 4).

As an alternative to Algorithm [P1], we introduce another proposal, Algorithm [P2], which “abandons” the MDA framework. The only augmentation involved in Algorithm [P2] is Step 3, which adopts a Scheme [MDA-LW]-like strategy in the constrained parameter space. If we fix α_1 to be 1, both Algorithms [KD] and [P1] simplify to Algorithm [P2]. We show in Appendix B that Algorithms [P1] and [P2] draw Σ from approximately the same sampling distribution under certain conditions.

Algorithm [P2]

1. Sample $w | (\mu, \Sigma, S)$. For each l , update w_{il} conditional on $w_{i,-l}$, μ , Σ , and s_i from a truncated normal distribution.

2. Sample $\theta | (w, \Sigma)$. Draw $\theta \sim f(\theta | w, \Sigma)$ and then set $\mu = G(X; \theta)$.

3. Sample $(\Sigma, \alpha_3^2) | (w, \theta)$.

(a) Draw $\tilde{\Sigma} \sim \text{Inv-Wishart}(N + \nu, \Psi + \sum_{i=1}^N \epsilon_i \epsilon_i^T)$, where $\epsilon_i = w_i - \mu_i$;

(b) set α_3^2 to $\text{trace}(\tilde{\Sigma}) / C$; and

(c) set $\Sigma = \tilde{\Sigma} / \alpha_3^2$ and $w = \mu + \epsilon / \alpha_3$.

Appendix D provides more details on the implementation of the algorithms. Software for fitting all three algorithms is available at <https://github.com/yizhenxu/GcompBART>.

2.5 Theoretical Comparison of Algorithms for MPBART

In what follows, we assume the Markov chain of (w, θ, Σ) has reached equilibrium. Liu (1994b) introduced the usage of diagrams that show dependency structures between two consecutive iterations for analyzing Bayesian algorithms. We do this for Algorithms [KD], [P1], and [P2], and derive their mixing rate in terms of autocorrelations. We prove the dependency structure as diagrams in Figure 1, which summarizes the three algorithms based on their sampling schemes.

A common measure for quantifying the mixing rate of a Markov chain is the lag-1 autocorrelation; lower autocorrelation indicates a better mixing rate. Using the dependency diagrams, we argue that Algorithms [P1] and [P2] have an ideal mixing rate when the Markov chain is stationary.

Theorem 1

Assuming the chain of MPBART parameters (w, μ, Σ, α) has reached equilibrium. For μ , Algorithms [P1] and [P2] have the same lag-1 autocorrelation, which is no larger than that from Algorithm [KD].

Proof: Appendix A.

3 Simulation

The simulation study will compare the predictive accuracy of three algorithms, Algorithms [KD], [P1], and [P2]. Denote the posterior sample of model parameters by $\{\theta^{(j)}, \Sigma^{(j)} \mid j = 1, \dots, J\}$. The posterior predictive distribution for s_i can be represented by its J posterior predictions, $\{\hat{s}_i^{(1)}, \dots, \hat{s}_i^{(J)}\}$, where

$$\hat{s}_i^{(j)} = \begin{cases} l & \text{if } \max(\hat{W}_i^{(j)}) = \hat{W}_{il}^{(j)} \geq 0 \\ 0 & \text{if } \max(\hat{W}_i^{(j)}) < 0, \end{cases} \quad (4)$$

$\hat{W}_i^{(j)} = (\hat{W}_{i1}^{(j)}, \dots, \hat{W}_{iC}^{(j)})$ is the vector of latent variables, $\hat{W}_i^{(j)} \sim \text{MVN}(G(X_i; \theta^{(j)}), \Sigma^{(j)})$, and

$$G(X_i; \theta^{(j)}) = (G_1(X_i; \theta_1^{(j)}), \dots, G_C(X_i; \theta_C^{(j)})).$$

Recall that each mean component is parameterized as the sum of trees,

$$G_l(X_i; \theta_l^{(j)}) = \sum_{k=1}^m g(X_i; \theta_{lk}^{(j)}), \text{ where } l = 1, \dots, C.$$

We consider two metrics of predictive accuracy: posterior percent agreement and posterior mode accuracy. While posterior mode accuracy compares the observed outcome s_i and the maximum a posteriori (MAP) estimate of the outcome, posterior percent agreement measures the concordance between s_i and the sampled posterior predictive distribution. Under the multinomial probit framework, Algorithms [KD], [P1] and [P2] directly sample posterior predicted outcomes, $\{\hat{s}_i^{(1)}, \dots, \hat{s}_i^{(J)}\}$, e.g. the j th posterior draw is $\hat{s}_i^{(j)}$; the posterior percent agreement is averaged over N subjects as follows,

$$\frac{1}{N} \sum_{i=1}^N \left\{ \frac{1}{J} \sum_{j=1}^J 1\{\hat{s}_i^{(j)} = s_i\} \right\}. \quad (5)$$

Posterior mode accuracy summarizes the agreement between the observed s_i and the

posterior mode prediction, $S_i = \arg \max_{l \in \{0, \dots, C\}} \sum_{j=1}^J 1\{\hat{s}_i^{(j)} = l\}$, via

$$\frac{1}{N} \sum_{i=1}^N 1\{S_i = s_i\}. \quad (6)$$

The two accuracy measures are different in that the posterior mode accuracy ignores the infrequent categories in MCMC sampling, whereas the posterior percent agreement accounts for all posterior predictive draws.

Numerical experiments for all simulations use 3,000 posterior draws after a burn-in of 5,000 for each model and parameterize the mean component of each latent variable as the sum of 100 trees. The tree priors for the three algorithms are specified as recommended in Chipman

et al. (2010), where the prior probabilities for the posterior tree search are 0.25, 0.25, 0.4, and 0.1 for tree GROWTH, PRUNE, CHANGE, and SWAP, respectively. Prior specification of the latent variable covariance matrix assumes the scale matrix Ψ has diagonal elements equal to 1.

For each simulation replicate, we create a training set \mathcal{D}_1 and a test set \mathcal{D}_2 , each of size 5000. Under different prior specifications and data-generating settings, we apply the three algorithms on \mathcal{D}_1 . Each experiment is performed on 100 simulation replications and the out-of-sample performances are evaluated by calculating the two accuracy metrics using \mathcal{D}_2 . We simulate \mathcal{D}_1 and \mathcal{D}_2 similar to Kang and Schafer (2007). We set the number of latent utilities

$C = 2$ and assume a set of covariates (U, V) where $U = (U_1, \dots, U_5) \sim \text{Uniform}(0, 1)$ and $V \sim \text{Uniform}(0, 2)$, and set $G_1 = 15 \sin(\pi U_1 U_2) + (U_3 - 0.5)^2 - 10U_4 - 5U_5$. We set $G_2 = (U_3 - 0.5)^3 - 20U_4 U_5 + 4V$ in Setting 1 for a relatively balanced distribution of the outcome categories and $G_2 = (U_3 - 0.5)^2 - U_4 U_5 + 4V$ in Setting 2 for highly unbalanced outcomes. The covariance matrix is $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.

Averaged across the 100 simulation replicates, the distribution of the outcome alternatives is (0.45, 0.25, 0.30) and (0.32, 0.65, 0.03) for Settings 1 and 2, respectively. Figure 2 compares the out-of-sample posterior predictive accuracy of the algorithms, under different priors for the augmented latent covariance, $\tilde{\Sigma} \sim \text{Inv-Wishart}(\nu, \Psi)$. Assuming $\Psi_{11} = \Psi_{22} = 1$, we consider uniform ($\nu = C + 1, \Psi_{12} = 0$), negatively tilted ($\nu = C + 3, \Psi_{12} = -0.5$), and positively tilted ($\nu = C + 3, \Psi_{12} = 0.5$) priors. Algorithm [KD] performs well under the posterior mode accuracy but is relatively more sensitive to the prior specifications and tends to have large variation across posterior predictions, resulting in a sub-optimal performance under the posterior agreement accuracy, which reflects the posterior predictive distribution.

We also investigate how the multinomial probit algorithms behave in estimating Σ under different prior specifications and simulation settings. As with data generation, the reference level used in estimations is set to zero. Figure 3 summarizes the posterior mean of the normalized covariance matrix Σ . For σ_{11} and σ_{12} , $E[\cdot | D]$ is the posterior mean based on a simulation replicate D ; $E\{E[\cdot | D]\}$ and $S\{E[\cdot | D]\}$ are the mean and standard deviation of $E[\cdot | D]$ across the 100 replicates. Note that σ_{22} is not displayed in the Figure 3 because Σ is normalized, satisfying $\sigma_{22} = \text{trace}(\Sigma) - \sigma_{11}$. The true conditional correlation, $\text{corr}(W_1, W_2 | G)$, equals 0.5; for the posterior mean of the covariance, σ_{12} , Algorithm [KD] returns negative estimates while our proposals generate positive estimates, agreeing with the true correlation in sign. Appendix C shows how σ_{12} affects the outcome distribution, given $\sigma_{11} = \sigma_{22} = 1$. Conditional on the additive trees, σ_{12} has a substantial effect on the outcome predictive distribution, for example, Appendix C illustrated that a negative σ_{12} induces smaller reference level outcome probability $P(S = 0)$. Having a negative estimated posterior mean of σ_{12} may lead to posterior tree estimates that are systematically different from the simulation truth, where σ_{12} is set to be positive.

The autocorrelation for the average tree depth with lags ranging from 1 to 10 is shown in Figure 4. We summarize the sum-of-trees component with its average tree depth because trees are nonparametric and cannot be directly analyzed. The figure shows that Algorithm [KD] has stronger short-term autocorrelation in the average tree depths than the two proposals, reflecting the conclusion in Theorem 1 on lag-1 autocorrelation. It indicates that Algorithms [P1] and [P2] may be mixing better than Algorithm [KD] for the sum-of-trees component. Our proposals converge faster than Algorithm [KD] because the latter updates the sum-of-trees component conditional on latent utilities that are augmented / not normalized, which makes posterior convergence for trees more challenging. When the outcome is unbalanced, posterior convergence is more difficult.

4 Application

In this application, we investigate patients' retention in HIV care after enrollment as a function of their baseline characteristics and treatment status. The data were extracted from electronic health records of adults enrolled in HIV care between June 1st 2008 and August 23rd 2016 in AMPATH, an HIV care program in Kenya. We look at a 200-days window after the initial care encounter and split the data into training and test sets of sample sizes 49,942 and 26,714, respectively. Outcome is defined as disengagement, engagement, and reported death, where engagement in care means there was at least one visit to the clinic for HIV care during the first 200 days after a patient's initial encounter, and disengagement otherwise if the person was not reported dead. The outcome distribution is extremely imbalanced, such that the frequency of disengagement, engagement, and death is 16% , 80% , and 4% , respectively. Covariates are summarized in Table 1, including baseline age, gender, year of enrollment, travel time to clinic, marriage status, weight, height, baseline treatment status, whether CD4 count was measured at enrollment, whether CD4 count was updated post-enrollment before day 200, and the most recent CD4 count by day 200. To handle missingness, we use a separate category and a missing values indicator for categorical and continuous covariates, respectively.

We use 10,000 posterior draws after a burn-in of 10,000 and keep other settings the same as in simulations. Table 2 compares the posterior accuracy for Algorithms [KD], [P1], and [P2]. Algorithm [KD] has posterior mode accuracy comparative to, but not as good as that from our proposals. Algorithm [KD] does not separate the latent utilities of the true outcome level and those for the other outcome alternatives well, resulting in a less ideal posterior agreement accuracy. In terms of the stability in accuracy measures with respect to the choice of reference level, the performance of the proposals is more stable than the Algorithm [KD].

Under the reference level being disengagement, the first row of Figure 5 presents the MCMC convergence plots of the average tree depth corresponding to latent variables $w_1 = Z_{\text{eng}} - Z_{\text{diseng}}$ and $w_2 = Z_{\text{death}} - Z_{\text{diseng}}$, and the histogram of the posteriors of $\sigma_{12} = \text{Cov}(w_1, w_2)$, where $(Z_{\text{eng}}, Z_{\text{diseng}}, Z_{\text{death}})$ are latent utilities corresponding to each of the outcome levels and σ_{12} is the normalized conditional covariance of w_1 and w_2 . The plots show that the average tree depths are around 6 and 9 respectively for w_1 and w_2 under Algorithm [KD], and approximately 2 for those under Algorithms [P1] and [P2]. The Bayesian regularization priors that favor shallow trees do not work well for Algorithm [KD], as a tree depth of 6 allows up to 2^6 leaves, which increases the risk of over-fitting and makes the stochastic tree search inefficient. The second and third rows of Figure 5 set engagement

and reported death as the reference level, respectively, and the latent variables are defined accordingly. Similar conclusions are observed for tree depth. Under all choices of the reference level, the histogram of σ_{12} from Algorithms [P1] and [P2] agree on the sign of σ_{12} , which was demonstrated in previous simulations to match the sign of the true value of the underlying σ_{12} .

5 Concluding Remarks

While computational performance is an important criterion in building Gibbs sampler for complicated models, the dependency structure and sampling schemes are as crucial for devising an algorithm that generates a Markov chain with computational efficiency and fast mixing rates. We explore the data augmentation schemes involved in the Bayesian estimation of multinomial probit models and propose two alternative algorithms that improve the computational and theoretical properties of the estimating procedure of MPBART proposed in Kindo et al. (2016). We showed that KD and one of our proposals are, respectively, the MPBART-generalization of Algorithms [IvD1] and [IvD2] proposed in Imai and van Dyk (2005) for estimating MNP. The primary distinction between the two MNP algorithms is that the former uses augmentation in the sampling of model coefficients for the mean of latent variables, while the latter does not. Imai and van Dyk (2005) recommended [IvD1] over [IvD2] because the geometric rate of convergence of [IvD1] is at least as good as [IvD2]. One of our key contributions is to demonstrate that the same recommendation does not apply to MPBART.

We evaluate the algorithms' computational performance under the same parameter specifications using two accuracy measures: posterior percent agreement and posterior mode accuracy. Posterior mode accuracy, which compares observed categorical outcomes to the mode in posterior predictions, is widely used in machine learning literature, particularly in cross-sectional supervised learning studies such as Kindo et al. (2016). Alternatively, posterior percent agreement accounts for the posterior predictive probabilities of the outcome labels, so the estimated distribution of the non-dominant levels also influences the metric. In applications where multinomial models are used as generative components, posterior predictive distribution is more relevant than posterior mode predictions and it is crucial to examine the posterior predictive distribution of the categorical outcomes.

Through simulations and application, we compare our proposals to the estimating procedure in Kindo et al. (2016) (Algorithm [KD]). While Algorithm [KD] performs well in terms of posterior mode, its posterior percent agreement is less ideal. One possible explanation is that Algorithm [KD] samples posterior trees conditional on augmented latent variables, making posterior convergence of the trees more challenging; this may undermine the Bayesian regularization priors in BART, resulting in larger trees and higher computational costs, and lead to exploration of the latent correlation structure in a parameter space different from the truth. In Appendix C we further explore how the correlation of the latent variables affects the marginal outcome distribution, demonstrating that an estimated covariance of the wrong sign may be associated with a sum-of-trees component with values that are systematically different from the true data-generating mechanism.

Acknowledgment

We would like to acknowledge support for this project from the National Institutes of Health (NIH grants R01 AI 108441, R01 CA 183854, GM 112327, AI 136664, K24 AI 134359). The authors report there are no competing interests to declare.

Disclosure Statement: The authors report here are no competing interests to declare.

References

- Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):669–679.
- Bonato, V., Baladandayuthapani, V., Broom, B. M., Sulman, E. P., Aldape, K. D., and Do, K.-A. (2011). Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics*, 27(3):359–367.
- Burgette, L. F. and Nordheim, E. V. (2012). The Trace Restriction: An Alternative Identification Strategy for the Bayesian Multinomial Probit Model. *Journal of Business & Economic Statistics*, 30(3):404–410.
- Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART Model Search. *Journal of the American Statistical Association*, 93(443):935–948.
- Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.
- Gardner, E. M., McLees, M. P., Steiner, J. F., Del Rio, C., and Burman, W. J. (2011). The spectrum of engagement in HIV care and its relevance to test-and-treat strategies for prevention of HIV infection. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 52(6):793–800.
- Henderson, N. C., Louis, T. A., and Rosner, Gary L and Varadhan, R. (2020). Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. *Biostatistics*, 21(1):50–68.
- Imai, K. and van Dyk, D. A. (2005). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics*, pages 311 – 334.

- Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, pages 523–539.
- Kindo, B. P., Wang, H., and Peña, E. A. (2016). Multinomial probit Bayesian additive regression trees. *Stat (International Statistical Institute)*, 5(1):119–131.
- Lee, H., Genberg, B. L., Nyambura, M., Hogan, J., Braitstein, P., and Sang, E. (2017). State-Space Models for Engagement, Retention, and Reentry in the HIV Care Cascade. *CROI Conference*.
- Linero, A. R., Basak, P., Li, Y., and Sinha, D. (2021). Bayesian survival tree ensembles with submodel shrinkage. *Bayesian Analysis*, 1(1):1–24.
- Liu, J. S. (1994a). The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem. *Journal of the American Statistical Association*, 89(427):958–966.
- Liu, J. S. (1994b). Fraction of Missing Information and Convergence Rate of Data Augmentation. Research Triangle Park, North Carolina.
- Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance Structure of the Gibbs Sampler with Applications to the Comparisons of Estimators and Augmentation Schemes. *Biometrika*, 81(1):27–40.
- Liu, J. S. and Wu, Y. N. (1999). Parameter Expansion for Data Augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274.
- Low-Kam, C., Telesca, D., Ji, Z., Zhang, H., Xia, T., Zink, J. I., and Nel, A. E. (2015). A Bayesian regression tree approach to identify the effect of nanoparticles' properties on toxicity profiles. *Annals of Applied Statistics*, 9(1):383–401. Publisher: Institute of Mathematical Statistics.
- McCulloch, R. and Rossi, P. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1-2):207–240.
- McCulloch, R. E., Polson, N. G., and Rossi, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99(1):173–193.
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in econometrics*.
- Meng, X.-L. and Van Dyk, D. A. (1999). Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation. *Biometrika*, 86(2):301–320.
- Murray, J. S. (2020). Log-Linear Bayesian Additive Regression Trees for Multinomial Logistic and Count Regression Models. *Journal of the American Statistical Association*, 116(534).

- Nobile, A. (1998). A hybrid Markov chain for the Bayesian analysis of the multinomial probit model. *Statistics and Computing*, 8(3):229–242.
- Pratola, M. T. (2016). Efficient Metropolis–Hastings proposal mechanisms for Bayesian regression tree models. *Bayesian Analysis*, 11(3):885–911.
- Sparapani, R., Spanbauer, C., and McCulloch, R. (2021). Nonparametric machine learning and efficient computation with bayesian additive regression trees: the bart r package. *Journal of Statistical Software*, 97:1–66.
- Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016). Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Statistics in Medicine*, 35(16):2741–2753.
- Tanner, M. A. and Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398):528–540.
- Um, S., Linero, A. R., and Sinha, Debajyoti and Bandyopadhyay, D. (2022). Bayesian additive regression trees for multivariate skewed responses. *Statistics in Medicine*.
- Van Dyk, D. and Meng, X.-L. (2001). The Art of Data Augmentation. *The Journal of Computational and Graphical Statistics*, 10:1–111.
- van Dyk, D. A. (2010). MARGINAL MARKOV CHAIN MONTE CARLO METHODS. *Statistica Sinica*, 20(4):1423–1454.
- Waldmann, P. (2016). Genome-wide prediction using Bayesian additive regression trees. *Genetics, Selection, Evolution : GSE*, 48.
- WHO (2012). Meeting report on Framework for metrics to support effective treatment as prevention.

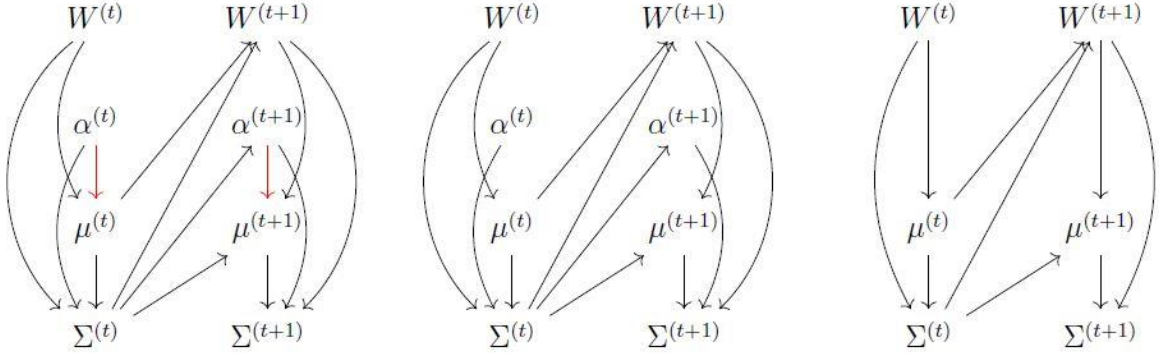


Figure 1: Above diagrams from left to right correspond to Algorithms [KD], [P1], and [P2], respectively.

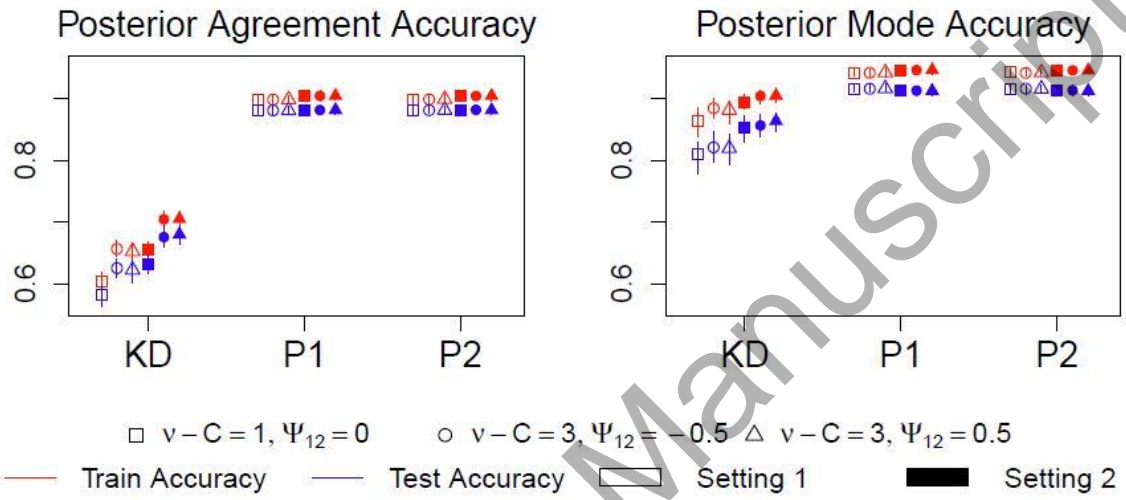


Figure 2: Posterior predictive accuracy measured by (5) and (6) are compared under Algorithms [KD], [P1], and [P2]. Based on 100 simulation replicates, the averages of the accuracies are displayed as squares, circles, and triangles, with empty and solid symbols indicating different simulation settings, and the corresponding 95% confidence intervals are represented as bars. The prior of Σ is $\text{Inv-Wishart}(\nu, \Psi)$, where $\Psi_{11} = \Psi_{22} = 1$, with $(\nu - C, \Psi_{12})$ being $(1, 0)$, $(3, -0.5)$, and $(3, 0.5)$.

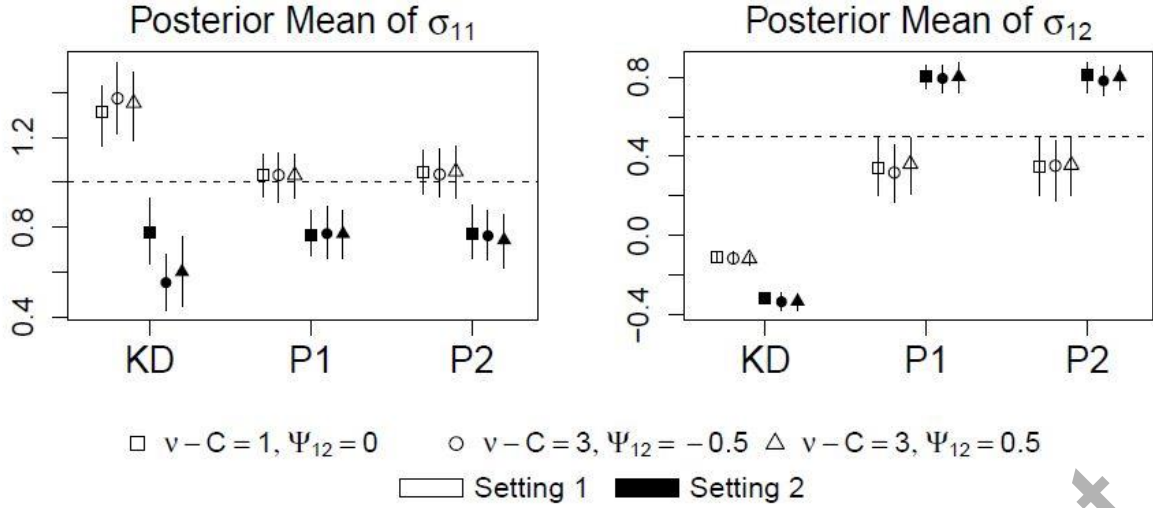


Figure 3: Comparison of mean and standard deviation of $E[\sigma_{11} | D]$ and $E[\sigma_{12} | D]$ under Algorithms [KD], [P1], and [P2]. $E[\cdot | D]$ indicates the estimated posterior mean on one simulated data D . Figures display the means (squares, circles, and triangles) and 95% confidence intervals (bars) from 100 simulation replicates. Settings 1 and 2 are represented by empty and solid symbols, respectively. The prior of $\tilde{\Sigma}$ is $\text{Inv-Wishart}(\nu, \Psi)$, where $\Psi_{11} = \Psi_{22} = 1$, with $(\nu - C, \Psi_{12})$ being $(1, 0)$, $(3, -0.5)$, and $(3, 0.5)$. Dashed horizontal lines show the corresponding true values.

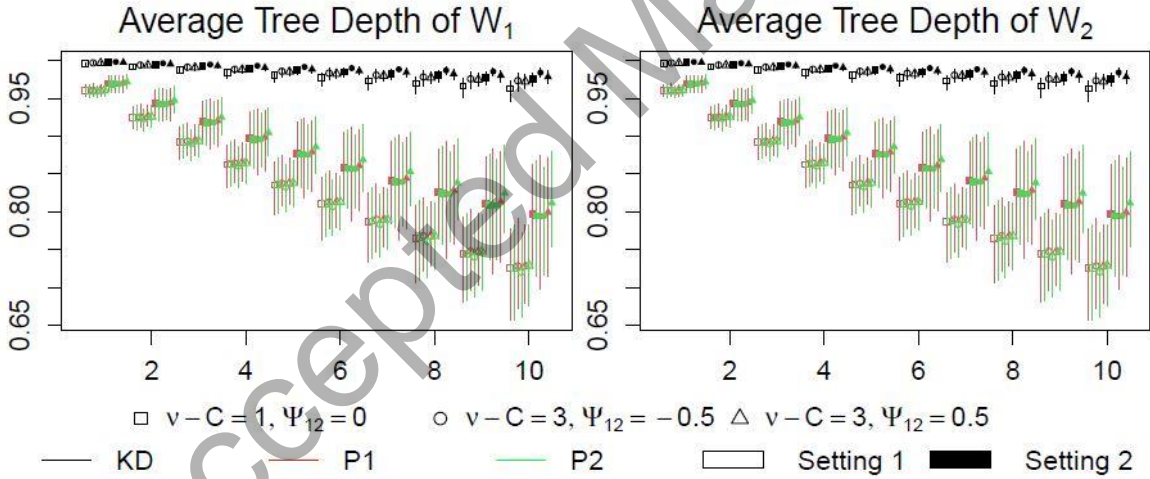


Figure 4: Comparing the autocorrelation of the average tree depth for the two latent utilities under Algorithms [KD], [P1], and [P2]. Figures display the means (squares, circles, and triangles) and 95% confidence intervals (bars) from 100 replicates. Settings 1 and 2 are represented by empty and solid symbols, respectively. The prior of $\tilde{\Sigma}$ is $\text{Inv-Wishart}(\nu, \Psi)$, where $\Psi_{11} = \Psi_{22} = 1$, with $(\nu - C, \Psi_{12})$ being $(1, 0)$, $(3, -0.5)$, and $(3, 0.5)$.

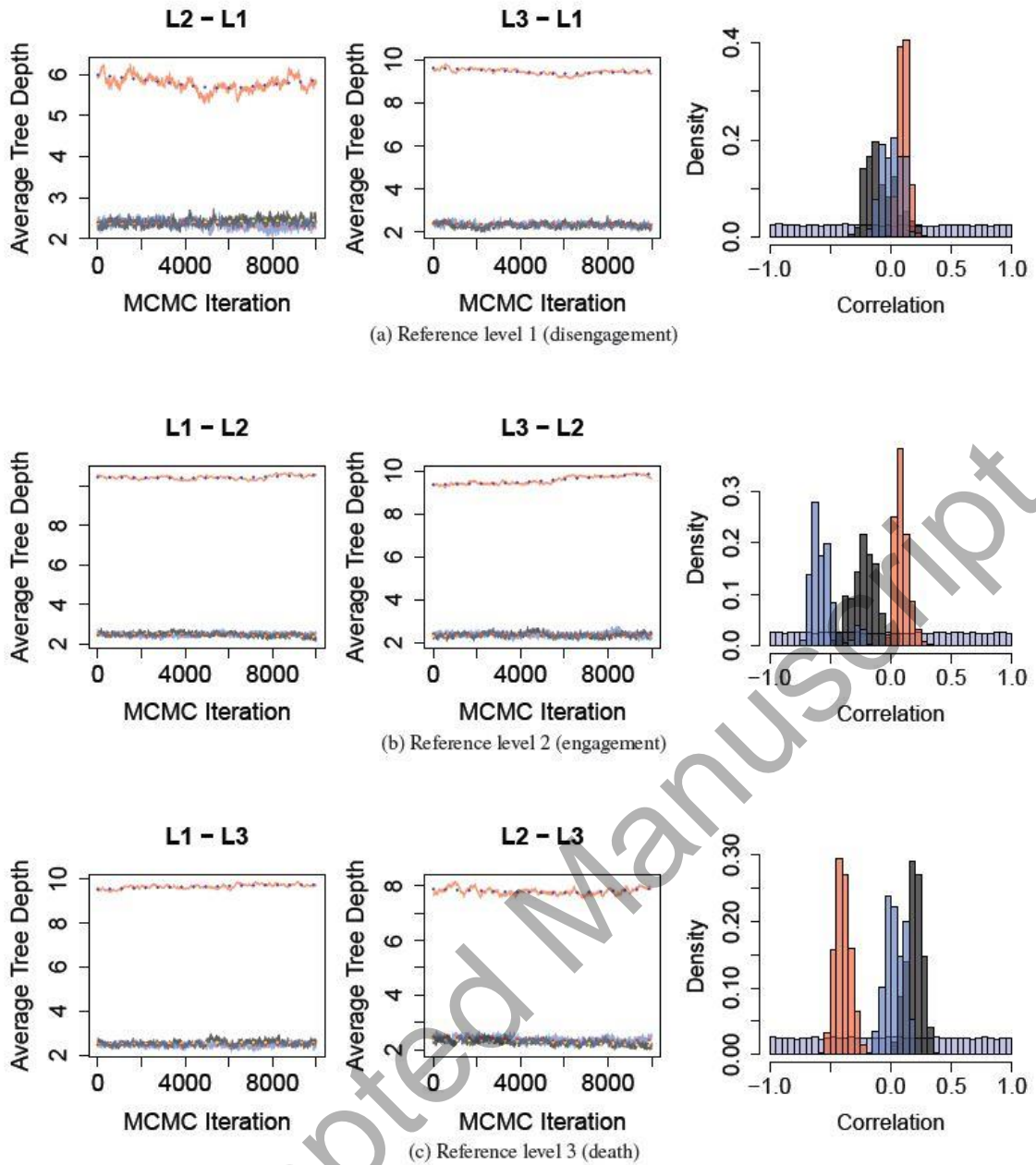


Figure 5: Traceplot of posterior average tree depth for each latent utility in the application to AMPATH data (left), and histogram of the σ_{12} (right) under its prior (purple), posterior from Algorithms [KD] (red), [P1] (black), and [P2] (blue); same color specification applies to the left plot. Posterior inference is under $\nu = C + 1$, $\Psi_{12} = 0$, with reference level as indicated in the plot labels.

Table 1: Summary table of covariates stratified by outcome. The table reports “median (25th percentile, 75th percentile)” for continuous variables and percentage of true for binary variables or each level of categorical variables. ENRL is the abbreviation for enrollment.

		Disengaged (6497)	Engaged (67462)	Died (2697)
Male		22.5	34	51.3
Year of ENRL	2008	5.1	9.7	11.2
	2009	8.3	18.7	17.1
	2010	9.3	17.3	17.6
	2011	9.2	15.8	17
	2012	17.9	11.5	14
	2013	18.5	8.9	11.3
	2014	18.8	9.0	8.2
	2015	12.8	8.3	3.3
	2016	0.3	0.8	0.3
Travel Time	< 30 min	17.4	24	23.6
	30 min - 1 h	19.4	26.9	29.4
	1 h - 2 h	8.2	14.6	16.5
	> 2 h	5.2	7.7	7.8
	Missing	49.9	26.8	22.6
WHO Stage	1	13.7	4.7	1.0
	2	1.8	2.0	1.1
	3	2.3	2.2	4.3
	4	0.6	0.3	0.7
	Missing	81.7	90.7	92.9
Married		57.2	52.3	49.7
	Missing	13.6	8.3	6.2
On ART		39.9	14.1	14.1

CD4 Measured at ENRL		64.8	80.9	74.7
CD4 Updated after ENRL		6.6	26	7.6
Most Recent CD4 Count		327 (144, 525)	279.77 (137, 462)	59 (18, 152)
Age		29.91 (24.66, 36.51)	35.56 (28.93, 43.65)	37.97 (31.7, 45.7)
Height		163 (158, 169)	165 (159.1, 171)	167 (160, 173)
	Missing	24.6	16.2	17.7
Weight		57.5 (51, 65)	56 (50, 63)	50 (44, 57)
	Missing	7.9	3.9	6.9

Accepted Manuscript

Table 2: Accuracy comparison of Algorithms [KD], [P1], and [P2] on the AMPATH data. Posterior predictive accuracy measured by (5) and (6) are reported under reference levels 1, 2, and 3. The prior of $\tilde{\Sigma}$ is Inv-Wishart($3, I_3$).

Posterior Agreement Accuracy						
	Train			Test		
Ref Level	KD	P1	P2	KD	P1	P2
1	0.67	0.82	0.82	0.67	0.81	0.81
2	0.55	0.82	0.82	0.54	0.81	0.81
3	0.66	0.82	0.82	0.66	0.81	0.81
Posterior Mode Accuracy						
	Train			Test		
Ref Level	KD	P1	P2	KD	P1	P2
1	0.88	0.89	0.89	0.88	0.89	0.89
2	0.85	0.89	0.89	0.84	0.89	0.89
3	0.88	0.89	0.89	0.88	0.89	0.89

Accepted Manuscript