Electronic Journal of Statistics
Vol. 19 (2025) 2876–2922
ISSN: 1935-7524
https://doi.org/10.1214/25-EJS2403

Bayesian variable selection on structured logistic-normal mixture models for subgroup analysis

Ruqian Zhang[1], **Naveen N. Narisetty**[2], **Xuming He**[3] **and Juan Shen**[*1]

[1]*Department of Statistics and Data Science, Fudan University,*
e-mail: rqzhang20@fudan.edu.cn*;* shenjuan@fudan.edu.cn

[2]*e-mail:* naveennn@umich.edu

[3]*Department of Statistics and Data Science, Washington University in St. Louis, e-mail:* hex@wustl.edu

Abstract: Subgroup analysis has emerged as an important tool to identify unknown subgroup memberships in the presence of heterogeneity. However, much of the existing work focused on the low-dimensional scenario where only a few candidate variables are considered for modeling the subgroup membership. In this paper, we propose a two-component structured mixture model with a Bayesian variable selection approach for identifying predictive and prognostic variables separately in the high-dimensional setting. By employing spike and slab priors, we achieve the selection of predictive and prognostic variables and the estimation of the treatment effect in the selected subgroup simultaneously. We establish theoretical properties by showing strong variable selection consistency and posterior contraction behavior of our method, and demonstrate its performance using simulation studies. Finally, we apply the proposed method to data from the National Supported Work and the AIDS Clinical Trials Group 320 study, identifying predictive and prognostic variables associated with subgroups exhibiting differential treatment effects.

Keywords and phrases: Bayesian variable selection, mixture models, predictive variable, prognostic variable, subgroup analysis.

Received October 2024.

1. Introduction

Subgroup analysis is a powerful tool for identifying heterogeneous treatment effects in various areas, including clinical trials and market segmentation. Traditional subgroup analysis focuses on the case where subgroup membership is determined by one or a few known covariates of interest, such as gender. Such a variable is said to be a "predictive" variable in subgroup analysis and helps to assign better treatment [20]. Other relevant variables are said to be "prognostic" when they contain information on the response regardless of the treatment. However, in recent years, researchers have also considered the case where the subgroup membership is unknown and the task is to target the potential subgroup. Seibold et al. [39], Huang et al. [17], Loh et al. [28], and Liu et al. [26] used tree-based methods to find the subgroup iteratively. Imai and Ratkovic [18] used a support vector machine model with lasso penalties to select subgroup variables. Chen et al. [8] proposed a search procedure to find patient stratification and described a resampling scheme to select the splitting variables. Shen and He [40] and Shen and Qu [41] proposed a mixture model to simultaneously model subgroup membership and response distributions within two distinct subgroups. Li et al. [24] and Wang et al. [45] used

[*]Corresponding author.

change plane models to identify unknown subgroups. Guo and He [14] and Guo et al. [15] made inference of the treatment effect on selected subgroups. Ma et al. [30] and Pedone et al. [33] clustered patients with similar predictive biomarkers and predicted the response based on both cluster results and prognostic variables. However, all these methods only assume a fixed number of covariates in their asymptotic studies and do not account for challenges arising in high-dimensional settings, where the number of candidate variables can be large relative to the sample size. As a result, their performance may deteriorate in such scenarios. In this paper, building on the work of Shen and He [40], we develop a two-component structured mixture model that extends to high-dimensional covariates.

When the design matrix is high-dimensional, especially when the number of variables exceeds the number of observations, the estimation problem is ill-posed. Moreover, as discussed by Ghosh et al. [13], the identification of subgroup membership will be subject to larger uncertainty without the exclusion of inactive covariates. These challenges can be remedied by variable selection under the assumption of sparsity. For variable selection under high-dimensional settings, one common approach is to add a penalty to the negative log-likelihood in the objective function, including the popular LASSO penalty [44], SCAD penalty [9], and MCP penalty [51], among others. However, Wang [48] observed several limitations of penalty-based methods for mixture models in terms of both theoretical properties and computational feasibility as they require optimization of non-convex objective functions.

In this paper, we consider a Bayesian alternative for high-dimensional subgroup analysis, which aims to alleviate the theoretical and computational challenges. In the framework of Bayesian approaches, suitable choices of prior distributions on the parameters can be used to perform estimation and variable selection [12, 21, 38, 4, 29]. With appropriate priors on the parameters involved in the model, the resultant posterior of the Bayesian method can be asymptotically similar to the L_0 penalized likelihood function [31, 25, 32]. A comprehensive overview of Bayesian variable selection methods can be found in Tadesse and Vannucci [43]. While previous works have mainly focused on the theoretical properties on variable selection [31, 50, 32], in this paper, we also study the posterior contraction properties on parameter estimation. Such properties have gained increasing interest in recent literature [35, 47, 10]. For computations, Markov Chain Monte Carlo (MCMC) techniques can be used for sampling from the posterior, which avoids the difficulties with optimization, especially in situations where the objective function is non-convex, such as censored regression models [37] and mixture models [49, 2]. Lu and Lou [29] proposed a Bayesian method to identify important variables for subgroup assignment using only predictive covariates, without considering high-dimensional scenarios.

The contributions of this article are summarized as follows. Firstly, we propose a structured mixture model that captures the subgroup membership and the within-subgroup information simultaneously and provides estimates of the treatment effect in the selected subgroup without ad hoc analysis. Secondly, we allow the variables in both parts of the model to be high-dimensional and provide variable selection methods to separately select predictive and prognostic variables. From our model, the “predictive” variables are directly used to predict the subgroup membership, while the selected variables from the tree-based methods or interaction models are not necessarily predictive. Thirdly, we establish strong selection consistency of variable selection and obtain posterior contraction rates for parameter estimation in the

ℓ_2 loss, and lastly, we provide a computationally scalable algorithm for high-dimensional settings.

In view of our contributions, we also acknowledge the broader context of subgroup analysis. While our method assumes a model-based framework for subgroup membership, another common approach follows a rule-based paradigm, where subgroups are defined by explicit covariate thresholding. This structure is prevalent in clinical applications, where subgroups are determined based on interpretable criteria, and many existing methods, such as GUIDE [27] and MOB [39], fall into that category. In this paper, we consider the rule-based setting as a form of model misspecification and evaluate the robustness of our method. Further, we recognize some limitations of our method in the rule-based setting. If predictive covariates exhibit high collinearity, model-based methods may become less robust than the tree-based approaches. We provide a critical analysis, along with simulation studies, to showcase this aspect of our method and discuss potential improvements.

The rest of the paper is organized as follows. Section 2 introduces the structured mixture model, the prior specifications, the posterior distribution, and the corresponding Gibbs sampler. Section 3 provides the theoretical justification of the proposed method. We provide comprehensive simulation studies in Section 4. We analyze data from the National Supported Work study and the AIDS Clinical Trials Group 320 study in Section 5 and conclude the paper with a discussion in Section 6. The R implementation of our method is publicly available at https://github.com/RuqianZhang/BVSA.

2. Methodology

In this section, we propose our model for simultaneous prognostic and predictive variable selection. We first introduce the structured logistic-normal mixture model conditional on the model indicator. Subsequently, we specify the variable selection priors accordingly.

2.1. Structured logistic-normal mixture models

Suppose we have n independent observations $\{(y_i, z_i, x_i, t_i)\}_{i=1}^n$ where $y_i \in \mathbb{R}$ is the continuous response, $z_i \in \mathbb{R}^{p_{1n}}$ and $x_i \in \mathbb{R}^{p_{2n}}$ denote the candidate prognostic and predictive covariates, respectively, and $t_i \in \{0, 1\}$ is the treatment indicator. The subscript n in p_{1n} and p_{2n} highlights that the model dimensions may depend on the sample size n, and we often omit this subscript unless necessary. Let $\delta_i \in \{0, 1\}$ be the latent subgroup indicator for the ith observation.

Let $\boldsymbol{\beta} \in \mathbb{R}^{p_1}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{p_2}$ be the corresponding coefficients for z_i and x_i, respectively. To facilitate variable selection, we introduce a binary model indicator $I \in \{0, 1\}^p$ with $p = p_1 + p_2$. The model indicator $I = (I^{\beta}, I^{\gamma})$ specifies model components included, where $I^{\beta} = (I_1^{\beta}, \cdots, I_{p1}^{\beta}) \in \{0, 1\}^{p_1}$ with each I_j^{β} indicating whether the jth component of $\boldsymbol{\beta}$ is included in the model, i.e., $\beta_j \neq 0$ if $I_j^{\beta} = 1$, and $\beta_j = 0$ if $I_j^{\beta} = 0$ for $j = 1, \ldots, p_1$, and similarly, $I^{\gamma} = (I_1^{\gamma}, \cdots, I_{p2}^{\gamma}) \in \{0, 1\}^{p_2}$ with each I_ℓ^{γ} indicating whether the ℓth component of $\boldsymbol{\gamma}$ is included in the model for $\ell = 1, \ldots, p_2$. Vectors with subscript I denote the sub-vectors corresponding to the nonzero components of I. Let $|\cdot|$ denote the L_0 norm. Then $\boldsymbol{\beta}_I$ and z_{iI} are the sub-vectors of $\boldsymbol{\beta}$ and z_i of length $|I^{\beta}|$ corresponding to the nonzero components of

I^{β}, while $\boldsymbol{\gamma}_I$ and x_{iI} for $\boldsymbol{\gamma}$ and x_i are sub-vectors corresponding to I^{γ}. Consider the following two-component structured logistic-normal mixture model:

$$\begin{aligned} y_i \mid (\delta_i, z_i, x_i, t_i, I) &= z_{iI}^T \boldsymbol{\beta}_I + \delta_i t_i \alpha_1 + (1-\delta_i) t_i \alpha_2 + \epsilon_i, \\ \delta_i \mid (z_i, x_i, I) &\overset{\text{ind}}{\sim} \text{Bernoulli}\left(\frac{\exp(x_{iI}^T \boldsymbol{\gamma}_I)}{1+\exp(x_{iI}^T \boldsymbol{\gamma}_I)} \right) \end{aligned} \tag{1}$$

for a given model I, where α_1 and α_2 represent the treatment effects in the two latent subgroups, and ϵ_i's are the random Gaussian noises with mean zero and variance σ_y^2. Without loss of generality, we assume that $\alpha_1 > \alpha_2$ for identifiability. Model (1) focuses on sub-models indicated by the indicator I only, and $|I^{\beta}|$ and $|I^{\gamma}|$ are the sizes of the prognostic and predictive models, respectively.

For further analysis, we denote the $n \times p_1$ and $n \times p_2$ design matrices by Z and X, and denote the treatment vector $(t_1, \ldots, t_n)^T$ by T. We assume that both Z and X include an intercept as the first column and further allow for overlapping components in Z and X. Matrices with subscript I denote the sub-matrices corresponding to the nonzero components of I, that is, Z_I and X_I are used to denote the $n \times |I^{\beta}|$ and $n \times |I^{\gamma}|$ sub-matrices of Z and X corresponding to the nonzero components in I^{β} and I^{γ}.

2.2. *Variable selection priors and joint posterior*

We now specify the prior distributions used in our Bayesian framework. We choose the commonly used Gaussian spike and slab priors on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ for variable selection. Conditional on I_j^{β}, the priors on β_j for $j = 1, \ldots, p_1$ are specified as:

$$\beta_j \mid (I_j^{\beta} = 1) \sim N\left(0, \sigma_y^2 \tau_{\beta 1n}^2\right), \;\; \beta_j \mid (I_j^{\beta} = 0) \sim N\left(0, \sigma_y^2 \tau_{\beta 0n}^2\right),$$

while, similarly, the priors on γ_ℓ for $\ell = 1, \ldots, p_2$ are specified as:

$$\gamma_\ell \mid (I_\ell^{\gamma} = 1) \sim N\left(0, \tau_{\gamma 1n}^2\right), \;\; \gamma_\ell \mid (I_\ell^{\gamma} = 0) \sim N\left(0, \tau_{\gamma 0n}^2\right),$$

where $\tau_{\beta 1n}^2$ and $\tau_{\gamma 1n}^2$ are the hyperparameters related to the variances of the slab distributions, and $\tau_{\beta 0n}^2$ and $\tau_{\gamma 0n}^2$ are the hyperparameters related to the variance of the spike distributions. The factor σ_y^2 is incorporated in the priors of $\boldsymbol{\beta}$ to naturally adapt the shrinkage effect to the scale differences between the linear and logistic components. The priors on I_j^{β} and I_ℓ^{γ} are independent Bernoulli distributions:

$$\begin{aligned} P(I_j^{\beta} = 1) &= 1 - P(I_j^{\beta} = 0) = q_{\beta n}, \\ P(I_\ell^{\gamma} = 1) &= 1 - P(I_\ell^{\gamma} = 0) = q_{\gamma n}, \end{aligned}$$

where $q_{\beta n}$ and $q_{\gamma n}$ are the prior inclusion probabilities. The choices of the hyperparameters in prior distributions may depend on the sample size n, which will be specified in Section 3.2. For conciseness, we omit the subscript n in the priors in the following. For α_1, α_2 and σ_y^2, we assume weakly informative prior distributions $\alpha_1 \sim N(0, \sigma_y^2 \sigma_\alpha^2)$, $\alpha_2 \sim N(0, \sigma_y^2 \sigma_\alpha^2)$, and $\sigma_y^2 \sim \text{IG}(a_0, b_0)$, where $\text{IG}(a, b)$ denotes the inverse gamma distribution with mean $b/(a-1)$, and a_0, b_0 and σ_α^2 are hyperparameters.

While the previously assigned prior distributions are conditionally conjugate and can deduce a closed-form Gibbs sampler for linear models, they are not conjugate for the logistic models. To address such difficulty, we adopt the Pólya-Gamma data-augmentation strategy [34]. For each binary subgroup indicator δ_i, a Pólya-Gamma latent variable ω_i is introduced, and the likelihood of the logistic model in (1) can be rewritten as:

$$\frac{\exp(x_{iI}^T\boldsymbol{\gamma}_I)^{\delta_i}}{1+\exp(x_{iI}^T\boldsymbol{\gamma}_I)} = \frac{1}{2}\int_0^\infty \exp\left\{\left(\delta_i - \frac{1}{2}\right)x_{iI}^T\boldsymbol{\gamma}_I - \frac{1}{2}\omega_i(x_{iI}^T\boldsymbol{\gamma}_I)^2\right\} p_{\mathrm{PG}}(\omega_i)d\omega_i,$$

where $p_{\mathrm{PG}}(\cdot)$ denotes the density of the Pólya-Gamma distribution $\mathrm{PG}(1,0)$. The Gaussian prior thus becomes conjugate for $\boldsymbol{\gamma}$ and the resulting posteriors of ω_i's and $\boldsymbol{\gamma}_I$ are as follows:

$$\begin{aligned}
&\omega_i \mid \boldsymbol{\gamma}_I \sim \mathrm{PG}(1, x_{iI}^T\boldsymbol{\gamma}_I),\\
&\boldsymbol{\gamma}_I \mid \Delta, \Omega \sim \mathrm{N}((X_I^T\Omega X_I + \tau_{\gamma 1}^{-2}\boldsymbol{I})^{-1}X_I^T\Omega\kappa, (X_I^T\Omega X_I + \tau_{\gamma 1}^{-2}\boldsymbol{I})^{-1}),
\end{aligned}$$

where $\kappa = ((\delta_1 - 1/2)/\omega_1, \ldots, (\delta_n - 1/2)/\omega_n)^T, \Delta = \mathrm{diag}(\delta_1, \ldots, \delta_n), \Omega = \mathrm{diag}(\omega_1, \ldots, \omega_n)$, and $\boldsymbol{I}$ is the identity matrix of suitable dimension. With the introduced ω_i's, the joint posterior density of $\boldsymbol{\gamma}, \boldsymbol{\beta}, \alpha_1, \alpha_2, \sigma_y$, Δ, Ω, I^{β}, and I^{γ} can be obtained by Bayes' formula as follows:

$$\begin{aligned}
&f(\boldsymbol{\gamma}, \boldsymbol{\beta}, \alpha_1, \alpha_2, \sigma_y, \Delta, \Omega, I^{\beta}, I^{\gamma} \mid Y)\\
&\propto \left(\sigma_y^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma_y^2}\sum_{i=1}^n \left(y_i - z_{iI}^T\boldsymbol{\beta}_I - t_i\alpha_1\delta_i - t_i\alpha_2(1-\delta_i)\right)^2\right\}\\
&\times \prod_{i=1}^n \exp\left\{(\delta_i - 1/2)\, x_{iI}^T\boldsymbol{\gamma}_I - \omega_i\left(x_{iI}^T\boldsymbol{\gamma}_I\right)^2/2\right\} p_{\mathrm{PG}}(\omega_i)\\
&\times \prod_{j=1}^{p_1} \left[q_{\beta}\pi_N\left(\boldsymbol{\beta}_j/\sigma_y\tau_{\beta 1}\right)\right]^{I_j^{\beta}} \left[(1-q_{\beta})\pi_N\left(\boldsymbol{\beta}_j/\sigma_y\tau_{\beta 0}\right)\right]^{1-I_j^{\beta}}\\
&\times \prod_{\ell=1}^{p_2} \left[q_{\gamma}\pi_N\left(\boldsymbol{\gamma}_{\ell}/\tau_{\gamma 1}\right)\right]^{I_{\ell}^{\gamma}} \left[(1-q_{\gamma})\pi_N\left(\boldsymbol{\gamma}_{\ell}/\tau_{\gamma 0}\right)\right]^{1-I_{\ell}^{\gamma}}\\
&\times \pi_N\left(\alpha_1/\sigma_y\sigma_{\alpha}\right)\pi_N\left(\alpha_2/\sigma_y\sigma_{\alpha}\right)\pi_{\mathrm{IG}}(\sigma_y^2; a_0, b_0),
\end{aligned} \tag{2}$$

where $\pi_N(\cdot)$ denotes the density of the standard Gaussian distribution, and $\pi_{\mathrm{IG}}(\cdot; a_0, b_0)$ denotes the density of an inverse gamma distribution with parameters a_0 and b_0. The posterior is conditional on X and Z which are excluded from the notation in the density function for simplicity.

2.3. Gibbs sampling algorithm

Since the likelihood depends only on the active part indicated by I in Model (1), the resultant Gibbs sampler enjoys an independent structure for active and inactive components when updating $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, which makes it scalable for large p_1 and p_2. We decompose $\boldsymbol{\beta} = (\boldsymbol{\beta}_I, \boldsymbol{\beta}_{I^C})$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_I, \boldsymbol{\gamma}_{I^C})$. Based on the joint posterior density (2), the Gibbs sampler draws samples from the following full conditional posteriors:

1. The conditional distributions of $\boldsymbol{\gamma}_I$ and $\boldsymbol{\gamma}_{I^C}$ are independent with

$$\begin{aligned}
\boldsymbol{\gamma}_I \mid (\cdots) &\sim N((X_I^T\Omega X_I + \tau_{\gamma 1}^{-2}\boldsymbol{I})^{-1}X_I^T\Omega\kappa, (X_I^T\Omega X_I + \tau_{\gamma 1}^{-2}\boldsymbol{I})^{-1}),\\
\boldsymbol{\gamma}_{I^C} \mid (\cdots) &\sim N\left(0, \tau_{\gamma 0}^2\boldsymbol{I}\right).
\end{aligned}$$

2. For $\ell = 1, \dots, p_2$, we generate $I_\ell^\gamma \in \{0, 1\}$ sequentially based on

$$\frac{P[I_\ell^\gamma = 1 \mid I_{-\ell}^\gamma, \cdots]}{P[I_\ell^\gamma = 0 \mid I_{-\ell}^\gamma, \cdots]} = \frac{q_\gamma \pi_N(\gamma_\ell/\tau_{\gamma 1})}{(1 - q_\gamma)\pi_N(\gamma_\ell/\tau_{\gamma 0})} \times \exp\left\{\left(\kappa - X_{C^\gamma(\ell)}\boldsymbol{\gamma}_{C^\gamma(\ell)}\right)^T \Omega X_\ell \gamma_\ell - \frac{1}{2} X_\ell^T \Omega X_\ell \gamma_\ell^2\right\},$$

   where $I_{-\ell}^\gamma$ represents the components of I^γ with I_ℓ^γ excluded and $C^\gamma(\ell) = \{k : k \neq \ell, I_k^\gamma = 1\}$.
3. For $i = 1, \dots, n$, the conditional distributions of ω_i's are $\mathrm{PG}(1, x_{iI}^T \boldsymbol{\gamma}_I)$.
4. For $i = 1, \dots, n$, we generate δ_i based on

$$\frac{P[\delta_i = 1 \mid \cdots]}{P[\delta_i = 0 \mid \cdots]} = \exp\left\{-\frac{1}{2\sigma_y^2}\left[\left(\alpha_1 + \alpha_2 - 2y_i + 2z_{iI}^T \boldsymbol{\beta}_I\right) t_i(\alpha_1 - \alpha_2)\right] + x_{iI}^T \boldsymbol{\gamma}_I\right\}.$$

5. Similar to $\boldsymbol{\gamma}_I$ and $\boldsymbol{\gamma}_{I^C}$, the conditional distributions of $\boldsymbol{\beta}_I$ and $\boldsymbol{\beta}_{I^C}$ are independent with

$$\begin{aligned} \boldsymbol{\beta}_I \mid (\cdots) &\sim N((Z_I^T Z_I + \tau_{\beta 1}^{-2}\boldsymbol{I})^{-1} Z_I^T \tilde{Y}_\beta, (W_\sigma(Z_I^T Z_I + \tau_{\beta 1}^{-2}\boldsymbol{I}))^{-1}), \\ \boldsymbol{\beta}_{I^C} \mid (\cdots) &\sim N(0, \sigma_y^2 \tau_{\beta 0}^2 \boldsymbol{I}), \end{aligned}$$

   where $W_\sigma = \mathrm{diag}(1/\sigma_y^2, |I^\beta|)$ and $\tilde{Y}_\beta = Y - T_\Delta \alpha_1 - T_{I-\Delta}\alpha_2$ with $T_\Delta = \Delta T$ and $T_{I-\Delta} = (\boldsymbol{I} - \Delta)T$.
6. The conditional distributions of α_1 and α_2 are given by

$$\begin{aligned} \alpha_1 \mid (\cdots) &\sim N((T_\Delta^T T_\Delta + \sigma_\alpha^{-2})^{-1} T_\Delta^T \tilde{Y}_1, \sigma_y^2 (T_\Delta^T T_\Delta + \sigma_\alpha^2)^{-1}), \\ \alpha_2 \mid (\cdots) &\sim N((T_{I-\Delta}^T T_{I-\Delta} + \sigma_\alpha^{-2})^{-1} T_{I-\Delta}^T \tilde{Y}_2, \sigma_y^2 (T_{I-\Delta}^T T_{I-\Delta} + \sigma_\alpha^{-2})^{-1}), \end{aligned}$$

   where $\tilde{Y}_1 = Y - Z_I^T \boldsymbol{\beta}_I - T_{I-\Delta}\alpha_2$, and $\tilde{Y}_2 = Y - Z_I^T \boldsymbol{\beta}_I - T_\Delta \alpha_1$.
7. For $j = 1, \dots, p_1$, we generate I_j^β sequentially based on

$$\frac{P[I_j^\beta = 1 \mid I_{-j}^\beta, \cdots]}{P[I_j^\beta = 0 \mid I_{-j}^\beta, \cdots]} = \frac{q_\beta \pi_N(\beta_j/\sigma_y \tau_{\beta 1})}{(1 - q_\beta)\pi_N(\beta_j/\sigma_y \tau_{\beta 0})} \times \exp\left\{\beta_j Z_j^T W_\sigma \left(\tilde{Y}_\beta - Z_{C^\beta(j)}\boldsymbol{\beta}_{C^\beta(j)}\right) - \frac{1}{2} W_\sigma Z_j^T Z_j \beta_j^2\right\},$$

   where I_{-j}^β represents the components of I^β with I_j^β excluded and $C^\beta(j) = \{k : k \neq j, I_k^\beta = 1\}$.
8. We generate σ_y^2 from $\mathrm{IG}(a_0 + (n + 2 + p_1)/2, b_1)$ with

$$b_1 = b_0 + (\tilde{Y}_\beta - Z_I \boldsymbol{\beta}_I)^T (\tilde{Y}_\beta - Z_I \boldsymbol{\beta}_I)/2 + (\alpha_1^2 + \alpha_2^2)/2\sigma_\alpha^2 + \boldsymbol{\beta}^T D_{I\beta} \boldsymbol{\beta}/2,$$

   where $D_{I\beta} = \mathrm{diag}(\tau_{\beta 1}^{-2} I^\beta + \tau_{\beta 0}^{-2}(1 - I^\beta))$.

The use of conjugate priors for the regression coefficients facilitates stable posterior updates, contributing to good mixing properties. Notably, the updates of the high-dimensional parameters are decomposed into two independent steps involving one dense but small precision matrix and one large but diagonal precision matrix. Similar model structure and decomposition of the precision matrix can be seen in Wang et al. [46] and in Narisetty et al. [32] where a sparse approximation of the precision matrix is used.

While we adopt a logit link function and leverage the Pólya-Gamma augmentation method to achieve efficient Gibbs sampling, a similar approach using a probit link is also feasible through the latent normal variable augmentation proposed by Albert and Chib [1].

3. Theoretical results

In this section, we investigate the theoretical properties of our proposed method. Specifically, we focus on the strong selection consistency of both prognostic and predictive variables, as well as the posterior contraction behavior with respect to the ℓ_2 error.

3.1. Reparameterization and marginal posterior distribution

Under the M-Closed assumption [3], the true model structure is assumed to be within the set of candidate models and is denoted by I_0, while I represents a candidate model. Let $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_1}$ and $\boldsymbol{\gamma}_0 \in \mathbb{R}^{p_2}$ be the true coefficient vectors, with $\boldsymbol{\beta}_{0I}$ and $\boldsymbol{\gamma}_{0I}$ denoting the sub-vectors of the true coefficients under model I. Given a model I, the likelihood of $\{y_i\}_{i=1}^n$ can be written as

$$L_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \alpha_1, \alpha_2, \sigma_y, I) = \prod_{i=1}^n \left[\frac{\pi(x_{iI}^T \boldsymbol{\gamma}_I)}{\sqrt{2\pi}\sigma_y} \exp\left\{ -\frac{(y_i - z_{iI}^T \boldsymbol{\beta}_I - t_i \alpha_1)^2}{2\sigma_y^2} \right\} + \frac{1 - \pi(x_{iI}^T \boldsymbol{\gamma}_I)}{\sqrt{2\pi}\sigma_y} \exp\left\{ -\frac{(y_i - z_{iI}^T \boldsymbol{\beta}_I - t_i \alpha_2)^2}{2\sigma_y^2} \right\} \right],$$

where $\pi(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$. As suggested by Städler et al. [42], we adopt a similar reparameterization and denote $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\phi} = \boldsymbol{\beta}/\sigma_y, s_1 = \alpha_1/\sigma_y, s_2 = \alpha_2/\sigma_y, \rho = 1/\sigma_y)$. Then the log-likelihood is

$$l_n(\boldsymbol{\theta}_I, I) = \sum_{i=1}^n \log \left(\pi(x_{iI}^T \boldsymbol{\gamma}_I) \frac{\rho}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}(\rho y_i - z_{iI}^T \boldsymbol{\phi}_I - t_i s_1)^2 \right\} + (1 - \pi(x_{iI}^T \boldsymbol{\gamma}_I)) \frac{\rho}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}(\rho y_i - z_{iI}^T \boldsymbol{\phi}_I - t_i s_2)^2 \right\} \right). \tag{3}$$

We assume the pairs $(\tau_{\beta 1}, \tau_{\gamma 1})$, $(\tau_{\beta 0}, \tau_{\gamma 0})$, and (q_{β}, q_{γ}) are of the same orders (as functions of n or p), respectively, so from now on we ignore the subscripts $\boldsymbol{\beta}$ or $\boldsymbol{\gamma}$ in these parameters. With the notation $\boldsymbol{b} = (\boldsymbol{\gamma}, \boldsymbol{\phi}) \in \mathbb{R}^p$, the joint prior distribution is given by

$$\pi(\boldsymbol{\theta}_I, \boldsymbol{\theta}_{IC}, I) \propto \exp\left\{ -\frac{1}{2}(\tau_1^{-2} \boldsymbol{b}_I^T \boldsymbol{b}_I + \tau_0^{-2} \boldsymbol{b}_{IC}^T \boldsymbol{b}_{IC}) \right\} \left(\frac{\tau_1(1-q)}{\tau_0 q} \right)^{-|I|} \pi(s_1, s_2, \rho), \tag{4}$$

where $\boldsymbol{\theta}_I = (\boldsymbol{b}_I, s_1, s_2, \rho)$, $\boldsymbol{\theta}_{IC} = \boldsymbol{b}_{IC}$, and

$$\pi(s_1, s_2, \rho) = \pi_N(s_1/\sigma_\alpha)\pi_N(s_2/\sigma_\alpha)\pi_\rho(\rho; a_0, b_0),$$

with $\pi_\rho(\cdot; a_0, b_0)$ denoting the density deduced from the Gamma distribution $\rho^2 \sim \Gamma(a_0, b_0)$. Then the joint posterior probability can be derived from (3) and (4) as:

$$\pi(\boldsymbol{\theta}_I, \boldsymbol{\theta}_{IC}, I \mid Y) \propto \exp\{l_n(\boldsymbol{\theta}_I, I)\} \pi(\boldsymbol{\theta}_I, \boldsymbol{\theta}_{IC}, I),$$

and the marginal posterior probability for model I is given by

$$\Pi(I \mid Y) \propto \int_{\Theta_I} \int_{\Theta_{IC}} \exp\{l_n(\boldsymbol{\theta}_I, I)\} \pi(\boldsymbol{\theta}_I, \boldsymbol{\theta}_{IC}, I) d\boldsymbol{\theta}_{IC} d\boldsymbol{\theta}_I,$$

where Θ_I and Θ_{IC} are the spaces consisting of all $\boldsymbol{\theta}_I$ and $\boldsymbol{\theta}_{IC}$, respectively.

3.2. Main results

Notations: for any sequences a_n and b_n, we denote $a_n \sim b_n$ if $a_n/b_n \longrightarrow c$ for some $c > 0$. We denote $b_n \succeq a_n$, or equivalently $a_n \preceq b_n$, if $b_n = O(a_n)$. For any $a, b \in \mathbb{R}$, the maximum and minimum of a and b are denoted by $a \vee b$ and $a \wedge b$. For any real symmetric matrix A, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ are the maximum and minimum eigenvalues of A, respectively.

We now study the theoretical properties of the proposed method in terms of selection consistency for both predictive and prognostic parts of our subgroup model, as well as the posterior contraction behavior of the parameters with respect to the ℓ_2 error. We assume that both covariate spaces $\mathcal{X}$ and $\mathcal{Z}$ of x_i's and z_i's are bounded and consider the parameter space

$$\Theta(M) := \{\boldsymbol{\theta} : |\log \rho| \le M, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1 \le M\},$$

where $\boldsymbol{\theta}_0$ is the true parameter, $M > 0$ is a fixed constant, and $\|\cdot\|_1$ denotes the L_1 norm for any vector. We define $Z_n(\boldsymbol{\theta}_I) := l_n(\boldsymbol{\theta}_{0I}, I) - l_n(\boldsymbol{\theta}_I, I)$ and let $\lambda_0 = \sqrt{\log p/n}$. We further define

$$V_n = \sup_{I:|I|\le p} \sup_{\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_1 \le M} \frac{1}{n} \frac{|Z_n(\boldsymbol{\theta}_I) - \mathbb{E}Z_n(\boldsymbol{\theta}_I)|}{\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_1 \vee \lambda_0}.$$

We first state some necessary conditions and introduce two important lemmas.

Condition 1. The dimension satisfies $\log p_n = o(n)$ as $n \longrightarrow \infty$.

Condition 2. For all $x \in \mathcal{X}$ and $z \in \mathcal{Z}$, there exist some constants λ_1 and λ_2 such that

$$\begin{aligned} 0 < \lambda_1 &\le \min_{I \in \mathcal{I}(m_n)} \min\left(\lambda_{\min}\left(\frac{1}{n} X_I^T X_I\right), \lambda_{\min}\left(\frac{1}{n} \tilde{Z}_I^T \tilde{Z}_I\right)\right) \\ &\le \max_{I \in \mathcal{I}(m_n)} \max\left(\lambda_{\max}\left(\frac{1}{n} X_I^T X_I\right), \lambda_{\max}\left(\frac{1}{n} \tilde{Z}_I^T \tilde{Z}_I\right)\right) \le \lambda_2, \end{aligned}$$

where $\mathcal{I}(m_n) = \{I : |I| \le m_n\}$ with $m_n := ((n/\log p)^{1/2} \wedge p)$ and $\tilde{Z}_I = (Z_I, T)$. We also assume that $I_0 \in \mathcal{I}(m_n)$.

Condition 1 restricts the model dimension as a function of n which is satisfied if $p_n \le e^{nd_n}$ for some $d_n \longrightarrow 0$ as $n \longrightarrow \infty$. Such a condition is common in Bayesian variable selection literature [25, 31, 23].

Condition 2 gives lower and upper bounds on the eigenvalues of $n^{-1} X_I^T X_I$ and $n^{-1} \tilde{Z}_I^T \tilde{Z}_I$. The lower bound can be seen as a restricted eigenvalue condition common in the high-dimensional statistics literature and is satisfied by sub-Gaussian design matrices with high probability [32]. The upper bound is similar to the bounded maximum eigenvalue condition assumed in Zou [53] and Bondell and Reich [5]. We restrict the model size to be smaller than or equal to m_n in Condition 2, which means that we only consider models of reasonably large sizes. This can be achieved by restricting the prior distribution on I as commonly done by Liang et al. [25] and Narisetty et al. [32].

Lemma 3.1. *Under the logistic-normal mixture model with $\boldsymbol{\theta} \in \Theta(M)$ for some M and Condition 1, there exists some constant $\bar{C} > 0$, such that for any constant $R \geq \bar{C}$, as $n \to \infty$,*

$$P(V_n \leq R\lambda_0) \to 1.$$

Lemma 3.2. *Under Conditions 1 and 2, on $\{V_n \leq R\lambda_0\}$, it holds that for any model I and any $\boldsymbol{\theta} \in \Theta(M)$ there exist some constants $c, c_1, c_2, c_3 > 0$, only dependent on M, $\mathcal{X}$ and $\mathcal{Z}$, satisfying*

$$cn\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2 - c_2 nR\lambda_0^2(|I|+3) \leq Z_n(\boldsymbol{\theta}_I) \leq c_1 n\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2 + c_3 nR\lambda_0^2(|I|+3).$$

Lemma 3.1 constructs a useful set that holds with probability going to 1 as n goes to infinity. Within the set in Lemma 3.1, Lemma 3.2 shows that $Z_n(\boldsymbol{\theta}_I)$, the negative log-likelihood divergence for any model I, has upper and lower bounds in simple forms, which can be utilized to replace the non-convex log-likelihood with tractable parameter ℓ_2 distance and model size.

We now outline additional conditions necessary to ensure the strong selection consistency property of our proposed method:

Condition 3. For some constant $\tilde{C} > \bar{C}|I_0|$ in which $\bar{C}$ is the constant specified in Lemma 3.1, the prior parameters τ_1^2 and q satisfy the following orders:

$$n\tau_1^2 \sim (n \vee p^{2+2\tilde{C}}), \quad q \sim p^{-1}.$$

Condition 4. For some constant $C_0 > 0$,

$$\min_{j \in \{k: I_{0k}=1\}} |b_{0j}| \geq \sqrt{\frac{C_0|I_0|\log p}{n}}.$$

Condition 3 provides rates on the parameters of the spike and slab prior. The variance of the slab prior distributions τ_1^2 is assumed to grow with n. No assumption is made on the variance of the spike prior distributions τ_0^2 as the choice of τ_0^2 would not influence the asymptotic results, which was also stated by Wang et al. [46]. In addition, we assume that the prior inclusion probability q is proportional to the inverse of the number of covariates p, which will control the model size. Condition 4 is a beta-min condition that restricts the minimal signal strength of true nonzero coefficients. This is commonly assumed when considering model sparsity [6, 32, 23].

Theorem 3.3. *Under Conditions 1-4 and on the set $\{V_n \leq R\lambda_0\}$, it holds that on $\mathcal{I}(m_n)$, the marginal posterior distribution of the true model satisfies*

$$\Pi[I = I_0 \mid Y] \xrightarrow{\text{P}} 1, \quad \text{as } n \to \infty.$$

Moreover,

$$\sum_{I_1 \in \mathcal{I}(m_n)\setminus\{I_0\}} \frac{\Pi[I = I_1 \mid Y]}{\Pi[I = I_0 \mid Y]} \xrightarrow{\text{P}} 0.$$

Theorem 3.3 provides strong selection consistency for both the predictive and prognostic parts of the high-dimensional subgroup model. It is implied that with probability going to 1, the posterior probability of the true model I_0 grows to 1 as n goes to infinity, given that the considered model sizes are allowed to be reasonably large. Theorem 3.3 gives an even stronger result that the sum of the posterior probability ratios of all possible false models to the true model converges to 0 in probability, implying a larger gap between the posterior probabilities of the true model and the rest.

In the end, we present a theorem concerning the posterior contraction rates, for which we impose an additional assumption regarding the variance τ_0^2 in the prior distributions:
Condition 5. The prior parameter τ_0^2 satisfies $n\tau_0^2 \sim 1/p$.

Theorem 3.4. *Under Conditions 1-5 and on the set $\{V_n \leq R\lambda_0\}$, it holds that on $\mathcal{I}(m_n)$, the posterior distribution satisfies for some constant $C' > 0$,*

$$\Pi\left[(\boldsymbol{\theta}, I) \in \Theta(M) \times \mathcal{I}(m_n) : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \geq \sqrt{\frac{C'|I_0|\log p}{n}} \,\Big|\, Y\right] \xrightarrow{\mathrm{P}} 0.$$

Theorem 3.4 shows that the posterior allocates most of its mass around the true parameters at the optimal rates for both high-dimensional linear regression and logistic regression under sparsity assumptions. The result ensures that the posterior probability of any estimate deviating from $\boldsymbol{\theta}_0$ by the bound on the left-hand side converges to 0 asymptotically. Our proof of Theorem 3.4 is obtained based on Theorem 3.3, which follows an approach different from Ray and Szabó [35] and Ray et al. [36]. We defer all proofs to Appendix A.

4. Simulation studies

In this section, we investigate the performance of the proposed method for subgroup analysis. First, we focus on correctly specified model settings and examine variable selection and parameter estimation in finite sample situations in both $p < n$ and $p \geq n$ cases. We then consider misspecified settings where subgroup membership is determined by splitting rules, common in traditional subgroup analysis. We also compare our proposed method with other subgroup identification methods.

4.1. Selection and estimation under structured logistic-normal mixture settings

We first consider data from the structured logistic-normal mixture model (1). Each row of Z and X is generated independently from normal distributions where the means are 0 and the correlations between any pair of covariates are equal to ρ. An intercept column is added to both Z and X. The noises are independently drawn from the standard normal distribution $N(0, 1)$.

We set the dimension $p = 2p_1 = 2p_2$ with $p \in \{100, 500, 2000\}$, the correlation $\rho \in \{0, 0.25\}$, and the sample size $n \in \{200, 300\}$. The values of $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$ are assigned to be $(1, -1.5, 2, -2.5, 3)^T$ with the rest being 0, and the treatment effects in two different subgroups are set to be $\alpha_{10} = 40$ and $\alpha_{20} = 0$. To examine the impact of higher correlation, numerical studies with larger ρ values are presented in Appendix B.1.

For our proposed method, referred to as **BVSA**, we consider the median probability model when we select active variables, i.e., variables with posterior inclusion probability at least 0.5. We initialize the Gibbs chain with random samples from the priors and obtain the results based on a chain of length 20000 with a burn-in of length 5000. The maximum model size is restricted to $\max(d, \sqrt{n})$ for some constant d. Here we choose $d = 30$ as suggested by

Narisetty et al. [32]. The choice of hyperparameters in continuous spike-and-slab priors can be sensitive to the scale difference between the linear and logistic components of the model. Unless otherwise specified, we adopt the following default setting for the spike-and-slab prior parameters based on theoretical considerations:

$$\tau_{\beta 0} = \tau_{\gamma 0} = 5/n, \;\; \tau_{\beta 1} = \tau_{\gamma 1} = \max(\sqrt{p^2/400n}, 1), \;\; q_\beta = q_\gamma = \min(1/5, 20/p). \qquad (5)$$

As discussed in Iqbal et al. [19], prior calibration on the prior variances can improve the empirical performance of variable selection in finite-sample settings. To assess this, we conduct a sensitivity analysis on the variance hyperparameters in Section D.1. Based on our experience on simulations, we adjust $\tau_{\gamma 1} = \max(\sqrt{p_2^2/800n}, 1)$ when $p = 2000$ to mitigate excessive false positives in high-dimensional settings. Since the priors on α_1, α_2, and σ_α^2 are weakly informative, the results are insensitive to a range of choices, confirmed by an illustration in Appendix D.3. Throughout the numerical studies we set $a_0 = 2$, $b_0 = 1$, and $\sigma_\alpha^2 = 1$.

We adopt variable selection performance measures used in Narisetty et al. [32]: TP, TP$_s$, FP, "$I = I_0$", "$I \supset I_0$", and "$I_s = I_0$", where TP (true positive) is the number of active covariates chosen; TP$_s$ is the number of active covariates selected if the size of the chosen model is restricted to be $|I_0|$; FP (false positive) is the number of inactive covariates chosen; "$I = I_0$" is the proportion of choosing the true model exactly; "$I \supset I_0$" is the proportion of times the true model is included in the chosen model; and "$I_s = I_0$" is the proportion of choosing the true model exactly when the model chosen is restricted to size $|I_0|$. Note that the measures TP$_s$ and "$I_s = I_0$" indicate how well a method can rank variable importance and do not depend on the specific choice of the threshold on posterior inclusion probability. The results are averaged based on 100 randomly generated datasets.

From the left columns of Table 1, we can observe that when $n = 200$, BVSA correctly identifies all prognostic variables, with true positives 4 and false positives 0 across all settings. For the predictive variables, when $p = 100$, our method finds most of the active variables with the probability of including all active covariates exceeding 0.8. When p increases to 500, which is greater than the sample size n, our method still performs well. The high probabilities of $I_s = I_0$ indicate that our method can correctly rank the posterior inclusion probabilities of all the variables. When $p = 2000$, the performance deteriorates but still yields reasonable results. When the sample size increases to $n = 300$, as shown in the right columns of Table 1, the performance on predictive variable selection improves significantly, even when $p = 2000$. These results support our theoretical findings on variable selection consistency.

To evaluate estimation accuracy, we examine the ℓ_2 errors and report the results in Table 2. We consider $p = 100$ and $\rho = 0$ with the same values of $\boldsymbol{\beta}_0$, $\boldsymbol{\gamma}_0$, α_{10} and α_{20} as before. The means and standard errors of the parameters $\boldsymbol{\beta}$, α, and $\boldsymbol{\gamma}$ at various sample sizes are summarized from 100 random trials. For all parameters, the ℓ_2 errors from BVSA shrink towards 0 as the sample size grows, and the standard errors also decrease towards 0.

4.2. Comparison under traditional subgroup settings

In this subsection, we consider several misspecified settings where subgroup membership is not determined by a logistic model but by splitting rules. We consider two cases with $p_2 = 10$ and $p_2 = 100$. The first ten predictors are generated as follows: (1) X_1 is standard normal; (2)

Table 1

Variable selection results in structured logistic-normal mixture settings.

		$n = 200$						$n = 300$					
		TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
$p = 100$	I^β	4	4	0	1	1	1	4	4	0	1	1	1
$\rho = 0$	I^γ	3.89	3.89	0.36	0.63	0.89	0.89	3.99	3.98	0.43	0.64	0.99	0.98
$p = 500$	I^β	4	4	0	1	1	1	4	4	0	1	1	1
$\rho = 0$	I^γ	3.68	3.74	0.34	0.55	0.69	0.74	3.94	3.94	0.26	0.71	0.94	0.94
$p = 2000$	I^β	4	4	0	1	1	1	4	4	0	1	1	1
$\rho = 0$	I^γ	3.01	3.12	0.64	0.24	0.30	0.36	3.74	3.73	0.37	0.49	0.76	0.74
$p = 100$	I^β	4	4	0	1	1	1	4	4	0	1	1	1
$\rho = 0.25$	I^γ	3.80	3.81	0.32	0.60	0.80	0.81	3.98	3.98	0.46	0.66	0.98	0.98
$p = 500$	I^β	4	4	0	1	1	1	4	4	0	1	1	1
$\rho = 0.25$	I^γ	3.65	3.65	0.42	0.47	0.65	0.66	3.90	3.89	0.21	0.73	0.90	0.89
$p = 2000$	I^β	4	4	0	1	1	1	4	4	0	1	1	1
$\rho = 0.25$	I^γ	2.81	2.87	0.89	0.16	0.20	0.25	3.58	3.57	0.64	0.42	0.64	0.63

Table 2

The ℓ_2 errors of parameter estimation with growing sample sizes when $p = 100$ and $\rho = 0$. The true parameter values are set to be $\boldsymbol{\beta}_{0I_0} = (1, -1.5, 2, -2.5, 3)^T$, $\boldsymbol{\gamma}_{0I_0} = (1, -1.5, 2, -2.5, 3)^T$, $\alpha_{10} = 40$, and $\alpha_{20} = 0$.

n	$\boldsymbol{\beta}$	α	$\boldsymbol{\gamma}$
200	0.423 (0.095)	1.139 (0.175)	1.527 (0.272)
300	0.279 (0.076)	0.760 (0.130)	1.204 (0.272)
400	0.214 (0.058)	0.570 (0.117)	1.002 (0.243)
500	0.184 (0.054)	0.479 (0.109)	0.893 (0.205)
1000	0.100 (0.037)	0.228 (0.088)	0.625 (0.173)

X_2 and X_3 are correlated normal variables with mean 0 and covariance 0.5; (3) X_4 comes from an exponential distribution with mean 1; (4) X_5 is Bernoulli with success probability equal to 0.5; (5) X_6 is multinomial with 3 equal-probability cells; and (6) X_7 to X_{10} are correlated normal variables with mean 0 and pairwise covariance 0.5. In high-dimensional settings, the remaining 90 predictors are generated from independent standard normal distributions. We take Z to be the same as X and thus $p = 20$ in the low-dimensional case and $p = 200$ in the high-dimensional case, respectively.

We consider the following six settings similar to those in Loh et al. [28] but with more generality:

S01: $Y = 1 + X_1 + X_2 + I(X_6 = 2) + X_7 + X_{10} + \epsilon,$
S02: $Y = 1 + X_1 + X_2 + 40t + \epsilon,$
S1: $Y = 1 + X_1 + X_2 + X_4 + I(X_6 = 2) + X_7 + 40t \times I(X_1 > 0) + \epsilon,$
S2: $Y = 1 + X_2 + 40t \times I(X_1 > 0, X_4 < 1, X_6 = 2) + \epsilon,$
S3: $Y = 1 + X_1 + X_2 + X_4 + I(X_6 = 2) + X_7 + 40t \times I(X_1 > 0, X_4 < 1, X_6 = 2) + \epsilon,$
S4: $Y = 1 + X_1 + X_2 + 40t \times I(\mathrm{logit}(X_1 + I(X_6 = 2)) \geq 0.5) + \epsilon,$

where t is the treatment indicator and ϵ is standard normal noise. Setting S01 has neither a treatment effect nor subgroups, and Setting S02 has a treatment effect but no subgroups. In these two settings, no meaningful subgroups exist. The remaining settings S1 to S4 have both treatment effects and subgroups. The two-component mixture model assumption used in the

Table 3

Component-wise predictive variable selection probabilities and FPR under S01 and S02 with no subgroup structure when $p = 20$ and $n = 200$.

(a) S01: $Y = 1 + X_1 + X_2 + I(X_6 = 2) + X_7 + X_{10} + \epsilon$

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	FPR
BVSA	0.01	0.03	0.03	0.02	0.01	0.02	0.03	0.01	0.01	0.01	0.15
MOB	0.75	0.67	0.01	0	0	0.01	0.90	0.04	0.02	0.90	1
FindIt	0.02	0.01	0.01	0	0	0.01	0	0.01	0	0	0.04
PRIM	0.36	0.40	0.42	0.46	0	0.20	0.45	0.39	0.44	0.43	1
SeqBT	0.23	0.34	0.17	0.10	0.01	0.12	0.26	0.21	0.22	0.18	1
GUIDE	0.19	0.32	0.23	0.18	0.14	0.22	0.31	0.23	0.24	0.30	0.88

(b) S02: $Y = 1 + X_1 + X_2 + 40t + \epsilon$

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	FPR
BVSA	0	0	0	0	0	0	0	0	0	0	0
MOB	1	1	0	0	0	0	0.01	0	0	0	1
FindIt	0.08	0.01	0	0.02	0.03	0.05	0	0	0	0.02	0.14
PRIM	0.53	0.55	0.07	0.08	0	0	0.06	0.03	0.04	0.02	1
SeqBT	0.18	0.21	0.19	0.12	0.06	0.08	0.14	0.09	0.09	0.13	1
GUIDE	0.27	0.28	0.22	0.07	0.05	0.10	0.11	0.10	0.08	0.11	0.62

paper does not hold under the settings S1, S2, and S3, so our simulation studies examine the performance of the proposed method under model misspecification.

The results of several other methods for subgroup identification in the literature are also reported for comparison, including:

- MOB: model-based recursive partitioning [39];
- SeqBT: sequential bootstrapping and aggregating of threshold from trees [17];
- GUIDE: generalized unbiased interaction detection and estimation [27];
- FindIt: support vector machine model with Lasso penalties [18];
- PRIM: patient rule induction method [8].

The parameters for all the comparison methods are set at their suggested default values. Our method is carried out in the same manner as in Section 4.1, except that we set $\tau_{\gamma 1} = 5$ across all settings to capture weaker predictive signals in the misspecified traditional subgroup settings. The simulation results are summarized from 100 randomly generated data sets with sample size $n = 200$ for each setting.

We focus on the performance of predictive variable selection since our method can accurately identify all prognostic variables in different settings. For settings without treatment effect, we provide the variable selection frequencies for predictive covariates and the false positive rate, which is defined by the frequency of falsely selecting any covariate. For settings with treatment effect, in addition to variable selection probabilities, we also report the same variable selection performance measures used in Section 4.1: TP, TP$_s$, FP, "$I = I_0$", "$I \supset I_0$", and "$I_s = I_0$". Subgroup prediction errors are reported in Appendix B.2, which are estimated from an independent testing data with $n = 5000$.

For low-dimensional settings with $p = 20$, the results for settings S01 and S02 in Table 3 show that the posterior inclusion probabilities of BVSA are close to 0 as they should be, and the false positive rates are small. The penalty-based method FindIt also performs well, while other methods always mistakenly assign subgroups. We can conclude that BVSA is not likely to select any predictive covariate when there is no treatment effect, and thus has a low

Table 4

Component-wise selection probabilities in traditional subgroup settings when $p = 20$ *and* $n = 200$.

(a) S1: $Y = 1 + X_1 + X_2 + X_4 + I(X_6 = 2) + X_7 + 40t \times I(X_1 > 0) + \epsilon$

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
BVSA	1	0	0	0	0	0	0	0	0	0
MOB	1	0.88	0.03	0.63	0	0.07	0.79	0.05	0.01	0.02
FindIt	1	0.88	0.85	0.80	0.87	0.97	0.86	0.84	0.87	0.85
PRIM	1	0.05	0.02	0.05	0	0	0.03	0.01	0	0
SeqBT	1	0.03	0.03	0.03	0	0.01	0.02	0.01	0.03	0
GUIDE	0.96	0.16	0.13	0.18	0.06	0.12	0.18	0.14	0.09	0.14

(b) S2: $Y = 1 + X_2 + 40t \times I(X_1 > 0, X_4 < 1, X_6 = 2) + \epsilon$

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
BVSA	0.92	0.01	0	0.85	0.02	1	0	0	0	0
MOB	0.19	1	0.02	0.05	0	0.51	0	0.01	0	0
FindIt	0.99	0.84	0.81	1	0.86	1	0.93	0.86	0.80	0.89
PRIM	0.53	0.16	0.15	0.50	0	0.12	0.12	0.16	0.16	0.12
SeqBT	0.08	0	0.01	0.02	0	0.93	0	0	0	0
GUIDE	0.86	0.18	0.07	0.13	0.05	0.98	0.03	0.06	0.04	0.07

(c) S3: $Y = 1 + X_1 + X_2 + X_4 + I(X_6 = 2) + X_7 + 40t \times I(X_1 > 0, X_4 < 1, X_6 = 2) + \epsilon$

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
BVSA	0.90	0.01	0	0.82	0.02	1	0	0.01	0	0
MOB	0.90	0.74	0.01	0.52	0	0.86	0.75	0.03	0	0.03
FindIt	1	0.84	0.84	0.99	0.87	1	0.92	0.82	0.80	0.86
PRIM	0.51	0.19	0.23	0.43	0	0.12	0.19	0.22	0.14	0.19
SeqBT	0.09	0	0	0.01	0	0.94	0	0	0	0
GUIDE	0.87	0.22	0.14	0.26	0.04	0.98	0.21	0.08	0.13	0.09

(d) S4: $Y = 1 + X_1 + X_2 + 40t \times I(\text{logit}(X_1 + I(X_6 = 2)) \geq 0.5) + \epsilon$

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
BVSA	1	0	0	0	0	1	0	0	0	0
MOB	1	1	0	0.01	0	0.82	0	0	0.01	0
FindIt	1	0.89	0.95	0.84	0.95	1	0.85	0.90	0.84	0.90
PRIM	1	0.03	0.02	0.01	0	0	0.01	0.01	0.01	0.03
SeqBT	1	0.08	0.10	0	0.01	0.05	0	0.03	0.02	0.03
GUIDE	1	0.03	0.02	0.02	0	0.02	0	0.01	0.01	0.02

probability of falsely identifying any subgroup. We also notice that those tree-based methods assign high inclusion probabilities to the active prognostic covariates, indicating that they are less capable of distinguishing prognostic and predictive covariates.

For the settings S1 to S4 with treatment effects, we summarize the posterior predictive inclusion probabilities in Table 4 and variable selection performance measures in the left columns of Table 5. In all settings, BVSA outperforms other methods, especially when the setting is complicated, e.g., S2 or S3. The true positives are close to the true model sizes, while the false positives are much smaller than those of other methods, indicating that BVSA has a high probability of finding the exact set of predictive covariates. One possible reason for the failure of the tree-based methods in some settings is that those methods are more ambitious in being overly flexible compared to model-based methods and are sensitive to tuning parameters involved.

When $p = 200$, FindIt adds the interactions between all covariates into the model, making it intractable for high-dimensional settings. Thus we exclude FindIt in the comparison. The findings are similar to those of low-dimensional settings. We mainly focus on predictive

Table 5
Predictive variable selection results in traditional subgroup settings with $n = 200$.

	$p = 20$						$p = 200$					
(a) S1: $Y = 1 + X_1 + X_2 + X_4 + I(X_6 = 2) + X_7 + 40t \times I(X_1 > 0) + \epsilon$												
	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
BVSA	1	1	0	1	1	1	1	1	0.03	0.97	1	1
MOB	1	1	2.48	0	1	1	1	1	1.56	0.01	1	1
FindIt	1	1	7.79	0	1	1	-	-	-	-	-	-
PRIM	1	0.97	0.16	0.88	1	0.97	0.99	0.99	0.12	0.97	0.99	0.99
SeqBT	1	1	0.16	0.85	1	1	1	1	0.12	0.89	1	1
GUIDE	0.96	0.96	1.20	0.34	0.96	0.96	0.87	0.87	0.26	0.74	0.87	0.87
(b) S2: $Y = 1 + X_2 + 40t \times I(X_1 > 0, X_4 < 1, X_6 = 2) + \epsilon$												
	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
BVSA	2.77	2.92	0.03	0.81	0.81	0.92	2.28	2.53	0.23	0.41	0.42	0.63
MOB	0.75	0.74	1.03	0	0.01	0	0.44	0.44	1.03	0	0	0
FindIt	2.99	2.23	5.99	0	0.99	0.61	-	-	-	-	-	-
PRIM	1.15	1.12	0.87	0	0	0	0.29	0.29	2.44	0	0	0
SeqBT	1.03	1.03	0.01	0	0	0	1	1	0.06	0	0	0
GUIDE	1.97	2.11	0.50	0.03	0.03	0.14	1.55	1.55	0.34	0	0	0.01
(c) S3: $Y = 1 + X_1 + X_2 + X_4 + I(X_6 = 2) + X_7 + 40t \times I(X_1 > 0, X_4 < 1, X_6 = 2) + \epsilon$												
	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
BVSA	2.72	2.92	0.04	0.75	0.76	0.92	2.33	2.53	0.18	0.46	0.48	0.63
MOB	2.28	1.67	1.56	0.04	0.39	0.08	1.86	1.72	0.97	0.10	0.20	0.10
FindIt	2.99	1.71	5.95	0	0.99	0.25	-	-	-	-	-	-
PRIM	1.06	1.07	1.16	0	0	0	0.23	0.24	3.31	0	0	0
SeqBT	1.04	1.04	0	0	0	0	0.98	0.98	0.07	0	0	0
GUIDE	2.11	2.27	0.91	0.05	0.17	0.29	1.52	1.54	0.30	0	0	0.01
(d) S4: $Y = 1 + X_1 + X_2 + 40t \times I(\text{logit}(X_1 + I(X_6 = 2)) \geq 0.5) + \epsilon$												
	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
BVSA	2	2	0	1	1	1	2	2	0.04	0.96	1	1
MOB	1.82	1.78	1.02	0	0.82	0.78	1.76	1.74	0.97	0.01	0.76	0.74
FindIt	2	1.94	7.12	0	1	0.97	-	-	-	-	-	-
PRIM	1	1.01	0.12	0	0	0	0.99	0.99	0.05	0	0	0
SeqBT	1.02	1.04	0.11	0.02	0.02	0.02	1	1	0.12	0	0	0
GUIDE	1.85	1.86	0.62	0.46	0.85	0.86	1.81	1.82	0.43	0.60	0.81	0.82

variable selection performance and conclude from the right columns of Table 5 that in nearly all settings, the performance of all the methods deteriorates when the dimensions of covariates grow, but BVSA suffers less severely and outperforms other methods significantly.

With the above results demonstrating that BVSA remains robust in rule-based settings, we further discuss its limitations under high correlations among predictive covariates when compared with tree-based methods.

Since all covariates enter the model simultaneously, BVSA can be sensitive to high collinearity, an issue that is further exacerbated by model misspecification. As a result, BVSA may select redundant variables. In contrast, tree-based methods partition the data hierarchically based on individual variable thresholds. Even if two variables are highly correlated, tree-based methods typically select only one for a given split, making them less sensitive to collinearity.

To empirically assess these limitations, we conduct additional simulation studies under rule-based subgroup settings, examining varying levels of correlation among predictive covariates. The details are reported in Appendix E. The results exhibit the limitations of BVSA in rule-based settings with highly correlated predictive covariates.

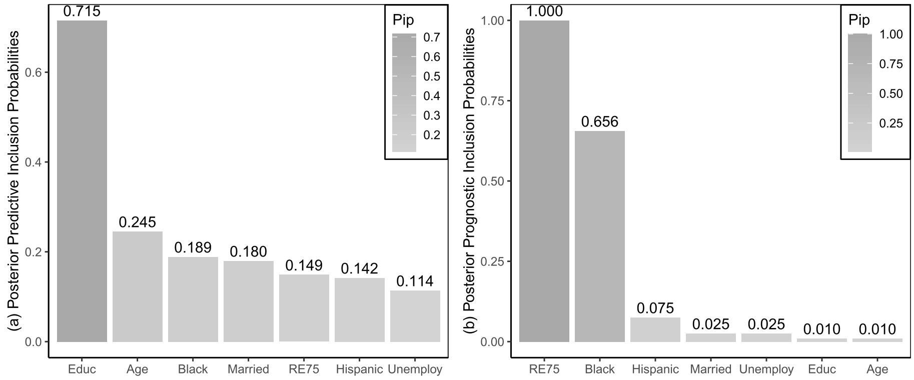


Fig 1. *Posterior inclusion probabilities for Lalonde data with posterior predictive inclusion probabilities in the left figure and posterior prognostic inclusion probabilities in the right figure.*

5. Real data applications

In this section, we apply the proposed method to two empirical studies, the Lalonde data from the National Supported Work program and the AIDS Clinical Trials Group 320 study.

5.1. Application to Lalonde data from NSW program

We first apply our method to the Lalonde data from the National Supported Work (NSW) program, a federally and privately funded program implemented from 1975 to 1978 in the United States to provide work experience to disadvantaged workers who had faced economic and social problems before enrollment. The data consists of 722 observations with 297 workers assigned to the training program and 425 workers in the control group. We focus on the earning increase in thousands of dollars after the job training program, which is equal to the difference between 1978 earnings and 1975 earnings, and aim to identify the subgroup where workers will benefit from the training program. The pre-treatment covariates include age, years of education (Educ), race (White, Black, Hispanic), marriage status (Married), the 1975 earnings in thousand dollars (RE75), and whether the worker was unemployed in 1975 before the program (Unemploy), and thereby $p = p_1 + p_2 = 14$. The covariates are standardized if they are continuous.

We first identify the important prognostic and predictive variables with the proposed method. We initialize the Gibbs chain based on the prior distributions, and the results are averaged from 5 random chains with a burn-in of 5000 and a subsequent length of 5000. We adopt the hyperparameter settings as specified in Section 4.1, with $\tau_{\beta 0} = \tau_{\gamma 0} = 0.007$, $\tau_{\beta 1} = \tau_{\gamma 1} = 1$, and $q_\beta = q_\gamma = 0.2$. The averaged posterior inclusion probabilities of all covariates for both prognostic and predictive consideration are shown in Figure 1.

The posterior prognostic inclusion probabilities of RE75 and Black are greater than 0.5, while the others are much smaller, with the largest among them below 0.1. The largest posterior predictive inclusion probability is 0.715 for Educ, followed by 0.245 for Age, while the rest are close to each other. We select important variables according to both the absolute values of the posterior inclusion probabilities and their gaps. As a result, we choose Black and RE75 to be prognostic and Educ to be predictive. The same active predictive variable was used for the group construction in Imai and Ratkovic [18]. As shown in Loh et al. [28], PRIM chooses Black, Educ, and Age as predictive variables, while GUIDE and SeqBT choose Married and Black as predictive variables, and FindIt uses a linear combination of all variables. However, our method identifies both predictive and prognostic variables and provides an explicit model with the estimated treatment effects in the subgroups.

Based on the selected variables, we obtain the estimated model as follows:

$$\begin{aligned} Increase &= \underset{(0.063)}{2.925} - \underset{(0.082)}{1.041}Black - \underset{(0.003)}{4.032}RE75 + \underset{(0.012)}{20.300}t\delta - \underset{(0.103)}{0.104}t(1-\delta) + \epsilon, \\ \operatorname{logit}(P[\delta = 1]) &= \underset{(0.032)}{-3.120} + \underset{(0.060)}{0.503}Educ, \end{aligned} \tag{6}$$

where $\hat{\sigma}_y = 5.501$ and the standard errors are provided under the estimated coefficients in brackets. Model (6) shows that the treatment effects on earning increase differ a lot in the two subgroups: in the first subgroup with $\delta = 1$, the treatment effect is over \$20, 000, while in the other subgroup, the treatment effect is close to 0.

Our method provides strong evidence for the selection of RE75 as a prognostic variable with posterior prognostic inclusion probability 1, but it is not selected as an active predictive variable due to its small posterior predictive inclusion probability. To understand the results, we examine the earning increase for different levels of RE75, with or without controlling the treatment. We divide all workers into two groups corresponding to high RE75 and low RE75 by its third quartile 3.993 as the threshold. The box plots of the earning increase of workers divided by high or low RE75 only overlap slightly, and the estimated density curves have two different peaks as shown in (a) and (b) of Figure 2, where the grey box and black solid line correspond to the high RE75 group, and the white box and grey dashed line correspond to the low RE75 group. In contrast, when we compare the earning increase differences between workers divided by the treatment in the high RE75 and the low RE75 groups, which are shown in (c) and (d) of Figure 2, the differences are similar in both groups. This indicates that the interaction between RE75 and treatment is negligible, which is consistent with our finding that RE75 is not predictive of the subgroup membership.

To demonstrate the effectiveness of our method in high-dimensional settings, we introduce additional noise features into the NSW data. We randomly assign 80% of the data as the training set and the remaining 20% as the testing set, repeating this process 100 times. In each trial, we increase the dimension of possible prognostic covariates from 7 to 571, resulting in $p = n = 578$, with all noise features drawn independently from the standard normal distribution. We evaluate the variable selection performance as well as the prediction errors. We perform the proposed method on the training set in the same manner as the analysis done earlier without the noise features added, to select active prognostic and predictive covariates, and obtain estimations of the corresponding parameters. Covariates are selected based on the median probability, and the selection frequency of each covariate is summarized over 100 trials. For the predictive covariates, the selection frequencies of Educ, Age, and Unemploy are 0.63, 0.01, and 0.01, respectively, with all others being 0. For the prognostic covariates, the selection frequencies of RE75 and Black are 1 and 0.58, while other covariates are not selected and the largest selection frequency of the noise features is 0.02. The variable selection results are consistent with those obtained without the noise features added to the data, showing the capability of our method for large p. In contrast, MOB selects RE75 as the only predictive covariate in all trials, which is a prognostic variable as discussed earlier.

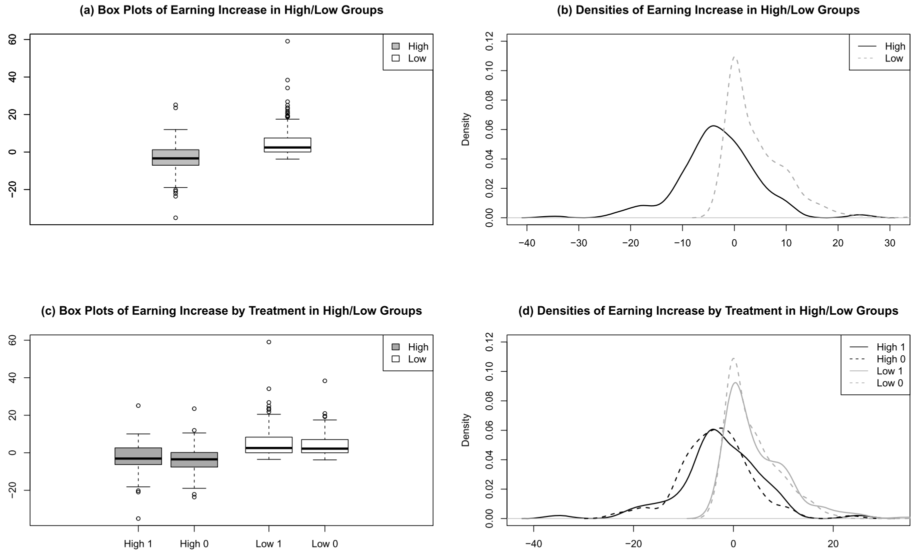


Fig 2. *Prognostic or predictive effects of RE75 (baseline earning in 1975) on the earning increase for Lalonde data. Prognostic effect: box plots of earning increase in high RE75 and low RE75 groups and density curves are provided in (a) and (b). Predictive effect: box plots of earning increase under treatment or control in high RE75 or low RE75 group and the corresponding density curves are provided in (c) and (d).*

Based on the estimated model from the training set with additional noise covariates, we obtain predictions on the testing set. The predictive root mean square error (PRMSE) of the earning increase of our method is 6.044 while the PRMSE based on MOB is 6.257. Our method exhibits lower prediction error and interpretable variable selection results, further supporting the superiority of the proposed method.

5.2. Application to ACTG 320 study

In this subsection, we apply our proposed method to the AIDS Clinical Trials Group (ACTG) 320 study. Following Hammer et al. [16], Zhao et al. [52], and Shen and He [40], we use the CD4 count change at week 24 as the response and aim to find the patient subgroup benefiting more from the three-drug combination. The dataset consists of 852 observations with 423 patients receiving the three-drug combination regimen and 429 patients receiving only the two-drug combination regimen, referred to as the control group. Our pre-treatment covariates include sex, injection-drug use (Ivdr), hemophilia (Hemo), weight (Weig), Karnofsky score (Karn), months of prior zidovudine therapy (PrZ), age, logarithm of baseline CD4 counts (Lcd40), logarithm of baseline HIV-1 RNA concentration with base 10 (Lrna0), and race (White, African, or Hispanic). We also include the interaction terms, and thus $p = p_1 + p_2 = 122$.

We perform our method in a similar manner to that in Section 5.1, except that we adjust the spike variances to $\tau_{\beta 0} = \tau_{\gamma 0} = 0.02$ because of higher dimensionality and weaker signal strength. We summarize the results from 5 random chains, each with a burn-in period of 10000 iterations followed by an additional 10000 iterations. The posterior prognostic and predictive inclusion probabilities of all covariates are represented in Figure 3.

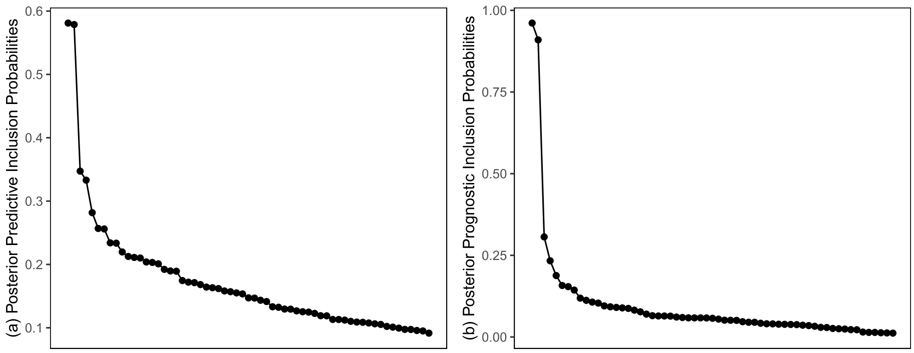


Fig 3. *Posterior inclusion probabilities for ACTG320 data with posterior predictive inclusion probabilities in the left figure and posterior prognostic inclusion probabilities in the right figure.*

For prognostic variables, the posterior inclusion probabilities of the interaction of Lrna0 and Lcd40 (Lrna0·Lcd40) and Lcd40 are 0.961 and 0.910, respectively, while the probabilities of other variables are no more than 0.3. For predictive variables, Lcd40 and Lrna0 have the largest posterior inclusion probability of 0.581 and 0.579, respectively, while the probabilities of other variables are less than 0.35. Based on posterior inclusion probabilities, we select Lrna0·Lcd40 and Lcd40 as the active prognostic variables, and Lcd40 and Lrna0 as the active predictive variables. Both Lcd40 and Lrna0 have been identified as predictive in previous studies [7, 52]. Based on the variable selection results, the estimated model of our method is given as follows:

$$\begin{aligned} Cd4\ change &= \underset{(0.313)}{-61.013} + \underset{(1.132)}{51.21} Lrna0 \cdot Lcd40 - \underset{(3.296)}{85.118} Lcd40 \\ &\quad + \underset{(0.391)}{161.403} t\delta + \underset{(1.268)}{10.974} t(1-\delta) + \epsilon, \\ \operatorname{logit}(P[\delta = 1]) &= -\underset{(0.047)}{0.24} - \underset{(0.06)}{1.049} Lcd40 + \underset{(0.074)}{0.622} Lrna0, \end{aligned} \tag{7}$$

with $\hat{\sigma}_y = 71.707$. We can observe from Model (7) that, although the new three-drug combination regimen has a positive effect on both subgroups, the first subgroup will benefit much more than the other.

The posterior inclusion probabilities of our method suggest strongly that Lrna0 is predictive but not prognostic, while Lcd40 is selected as both prognostic and predictive. To better interpret the roles of Lcd40 and Lrna0 in the prognostic and predictive models, we present additional graphical illustrations in Appendix C.2.

6. Discussion

Variable selection is crucial in subgroup analysis to identify subgroups with differential treatment effects defined by predictive variables, especially in a study with many possible covariates. In this paper, we consider the structured logistic-normal mixture model and propose a Bayesian method for finding the prognostic and/or predictive covariates. The strong selection consistency of this method is established under mild conditions, which guarantees that the posterior probability of the true model goes to 1 and separates from those of false models, and the posterior contraction rate is derived. The posterior computation can be implemented efficiently using a carefully designed Gibbs sampler. Simulation studies and application to real data show that our proposed method enjoys highly competitive performance for variable selection in subgroup analysis. Our methodology provides a good selection of predictive and prognostic variables and a satisfactory estimation of the treatment effects in the selected subgroups simultaneously.

Future work can explore strategies to mitigate the limitations discussed in Section 4.2, particularly in high-dimensional settings with highly correlated predictors. One promising approach is to perform variable selection in two stages. Since our variable ranking remains stable in lower-dimensional settings, we can first apply variable screening to filter out weakly associated variables, after which BVSA can be applied more effectively. Another possible improvement is to incorporate correlation-aware structure in the prior distribution. The standard spike-and-slab prior treats the variables independently, which can lead to redundant selection when they are highly correlated. Instead, we can modify the prior inclusion probability to depend on the covariate structure, allowing the model to suppress the inclusion of redundant correlated variables while still selecting relevant ones.

We can also extend our method to more flexible models to broaden its applicability. For example, generalizing BVSA to handle binary or survival outcomes would enhance its relevance in clinical studies. Additionally, extending BVSA to a multinomial logit framework could improve adaptive subgroup identification.

Appendix A: Proof of main results

In Appendix A.1 to A.4, we provide the proofs of the two lemmas and two theorems from Section 3.2, and then we give the proofs of the technical lemmas in Appendix A.5 to A.7. Given that the theoretical results are straightforward for finite p_n, we assume that $p_n \to \infty$ as $n \to \infty$ in the proof. We briefly discuss the case of finite p_n in Appendix A.8.

A.1. Proof of Lemma 3.1

For any model I, we define $\boldsymbol{\theta}_{\bar{I}} \in \mathbb{R}^{p+3}$ with $\boldsymbol{\theta}_I$ for I and $\mathbf{0} \in \mathbb{R}^{|I^c|}$ for I^c. For vector $\boldsymbol{v}$, $\boldsymbol{v}_I$ is used to denote the vector containing the components corresponding to model I. For any $I \supset I_0$, $\boldsymbol{\beta}_{0I}$ or $\boldsymbol{\gamma}_{0I}$ denotes the vector having $\boldsymbol{\beta}_{0I_0}$ or $\boldsymbol{\gamma}_{0I_0}$ for I_0 and zeroes for $I \cap I_0^c$.

As a common practice for the finite mixture of regressions, we consider a set of parameters:

$$\boldsymbol{\psi}(x, z, I) = (\pi(x^T \boldsymbol{\gamma}_{\bar{I}}), z^T \boldsymbol{\phi}_{\bar{I}} + ts_1, z^T \boldsymbol{\phi}_{\bar{I}} + ts_2, \rho).$$

Note that $\boldsymbol{\psi}(x, z, I)$ has a fixed dimension of 4, which is independent of n and p. We denote the density of Y by $f_{\boldsymbol{\psi}(x,z,I)}$ and $\ell_{\boldsymbol{\psi}(x,z,I)} = \log f_{\boldsymbol{\psi}(x,z,I)}$. Furthermore, we define the score function as

$$s_{\boldsymbol{\psi}(x,z,I)} = \frac{\partial \ell_{\boldsymbol{\psi}(x,z,I)}}{\partial \boldsymbol{\psi}(x, z, I)},$$

and the Fisher information as

$$I(\boldsymbol{\psi}(x, z, I)) = \int s_{\boldsymbol{\psi}(x,z,I)} s^T_{\boldsymbol{\psi}(x,z,I)} f_{\boldsymbol{\psi}(x,z,I)} \, d\mu,$$

where μ is the dominating measure of $f_{\psi(x,z,I)}(Y)$. By direct calculations, there exists a function $G_1(\cdot)$ for any I such that

$$\sup_{x\in\mathcal{X},z\in\mathcal{Z},\boldsymbol{\theta}_{\bar{I}}\in\Theta(M)} \|s_{\psi(x,z,I)}\|_\infty \leq G_1(Y) := C(Y^2+|Y|+1),$$

where C is a finite constant only depending on $M, \mathcal{X}, \mathcal{Z}$. Based on the inequalities in (B.2) of [48], we have for any positive number $\bar{M}$ and $\boldsymbol{\theta}_{\bar{I}} \in \Theta(\bar{M})$,

$$|\ell_{\psi(x,z,I)}(Y_i,x_i,z_i) - \ell_{\psi_0(x,z,I)}(Y_i,x_i,z_i)| \leq CG_1(Y_i)\|\boldsymbol{\theta}_{\bar{I}} - \boldsymbol{\theta}_{0\bar{I}}\|_1 \leq CG_1(Y_i)\bar{M},$$

and

$$\mathbb{E}[(\ell_{\psi(x,z,I)}(Y_i,x_i,z_i) - \ell_{\psi_0(x,z,I)}(Y_i,x_i,z_i))^2] \leq C^2\bar{M}^2\mathbb{E}[G_1^2(Y_i)] \leq C^*\bar{M}^2,$$

by Taylor expansions and the condition of boundedness of $\mathcal{X}$ and $\mathcal{Z}$. Due to Equation (B.14) in [48], for some $\bar{C} > 0$,

$$P\left[\sup_{\boldsymbol{\theta}_{\bar{I}}\in\Theta(\bar{M})} \frac{1}{n}|Z_n(\boldsymbol{\theta}_{\bar{I}}) - \mathbb{E}Z_n(\boldsymbol{\theta}_{\bar{I}})| > \bar{C}\bar{M}\lambda_0\right] \leq \bar{C}\left(\frac{1}{n}+\frac{1}{p}\right) \leq \frac{2\bar{C}}{p\wedge n}.$$

We apply the peeling device. For any given $M > 0$, divide $\Theta(M)$ to $\Theta(\lambda_0) \cup \left\{\Theta(M_j)\right\}_{j=1,2,\dots}$, where

$$\Theta(M_j) = \{\boldsymbol{\theta}_{\bar{I}} : |\log\rho| \leq M, e^{-j}M \leq \|\boldsymbol{\theta}_{\bar{I}} - \boldsymbol{\theta}_{0\bar{I}}\|_1 \leq e^{1-j}M\},$$

and

$$\Theta(\lambda_0) = \{\boldsymbol{\theta}_{\bar{I}} : |\log\rho| \leq M, \|\boldsymbol{\theta}_{\bar{I}} - \boldsymbol{\theta}_{0\bar{I}}\|_1 \leq \lambda_0\}.$$

It can be seen that the number of these sets is $\log(M/\lambda_0)+1$. Then we have for any constant $R \geq \bar{C}$ and any I,

$$\begin{aligned}
&P\left(\sup_{\|\boldsymbol{\theta}_{\bar{I}}-\boldsymbol{\theta}_{0\bar{I}}\|_1\leq M} \frac{1}{n}\frac{|Z_n(\boldsymbol{\theta}_{\bar{I}}) - \mathbb{E}Z_n(\boldsymbol{\theta}_{\bar{I}})|}{\|\boldsymbol{\theta}_{\bar{I}} - \boldsymbol{\theta}_{0\bar{I}}\|_1 \vee \lambda_0} > R\lambda_0\right)\\
&\leq \sum_j P\left(\sup_{\boldsymbol{\theta}_{\bar{I}}\in\Theta(M_j)} \frac{1}{n}\frac{|Z_n(\boldsymbol{\theta}_{\bar{I}}) - \mathbb{E}Z_n(\boldsymbol{\theta}_{\bar{I}})|}{\|\boldsymbol{\theta}_{\bar{I}} - \boldsymbol{\theta}_{0\bar{I}}\|_1} > R\lambda_0\right)\\
&+ P\left(\sup_{\boldsymbol{\theta}_{\bar{I}}\in\Theta(\lambda_0)} \frac{1}{n}\frac{|Z_n(\boldsymbol{\theta}_{\bar{I}}) - \mathbb{E}Z_n(\boldsymbol{\theta}_{\bar{I}})|}{\lambda_0} > R\lambda_0\right)\\
&\leq \sum_j P\left(\sup_{\boldsymbol{\theta}_{\bar{I}}\in\Theta(M_j)} \frac{1}{n}|Z_n(\boldsymbol{\theta}_{\bar{I}}) - \mathbb{E}Z_n(\boldsymbol{\theta}_{\bar{I}})| > \bar{C}e^{-j}M\lambda_0\right)\\
&+ P\left(\sup_{\boldsymbol{\theta}_{\bar{I}}\in\Theta(\lambda_0)} \frac{1}{n}|Z_n(\boldsymbol{\theta}_{\bar{I}}) - \mathbb{E}Z_n(\boldsymbol{\theta}_{\bar{I}})| > \bar{C}\lambda_0^2\right)\\
&\leq \frac{2\bar{C}(\log(M/\lambda_0)+1)}{p\wedge n} \longrightarrow 0,
\end{aligned} \tag{8}$$

as $n \longrightarrow \infty$ and $p \longrightarrow \infty$. Given the definition of $l_n(\theta_I, I)$ and $\boldsymbol{\theta}_{\bar{I}}$, for $\boldsymbol{\theta}_{\bar{I}} \in \Theta(M)$, we have $Z_n(\boldsymbol{\theta}_I) = Z_n(\boldsymbol{\theta}_{\bar{I}})$ and $\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_1 = \|\boldsymbol{\theta}_{\bar{I}} - \boldsymbol{\theta}_{0\bar{I}}\|_1 \leq M$. Thus,

$$P\left(\sup_{\|\boldsymbol{\theta}_I-\boldsymbol{\theta}_{0I}\|_1\leq M} \frac{1}{n}\frac{|Z_n(\boldsymbol{\theta}_I) - \mathbb{E}Z_n(\boldsymbol{\theta}_I)|}{\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_1 \vee \lambda_0} > R\lambda_0\right) \longrightarrow 0.$$

Due to the arbitrariness of I, we have that with probability going to one, $V_n \leq R\lambda_0$. □

A.2. *Proof of Lemma 3.2*

We denote the Kullback-Leibler information as

$$\varepsilon\left(\boldsymbol{\psi}(x, z, I) \mid \boldsymbol{\psi}_0(x, z, I)\right) = -\int \log\left[\frac{f_{\boldsymbol{\psi}(x,z,I)}}{f_{\boldsymbol{\psi}_0(x,z,I)}}\right] f_{\boldsymbol{\psi}_0(x,z,I)} d\mu,$$

where $\boldsymbol{\psi}_0(x, z, I) = (\pi(x^T\boldsymbol{\gamma}_{0I}), z^T\boldsymbol{\phi}_{0I} + ts_{10}, z^T\boldsymbol{\phi}_{0I} + ts_{20}, \rho_0)$. Further we define the average excess risk for fixed covariates $(x_1, z_1), \ldots, (x_n, z_n)$ to be

$$\bar{\varepsilon}\left(\boldsymbol{\psi}_I \mid \boldsymbol{\psi}_{0I}\right) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon\left(\boldsymbol{\psi}\left(x_i, z_i, I\right) \mid \boldsymbol{\psi}_0\left(x_i, z_i, I\right)\right).$$

Before the proof of Lemma 3.2, we first claim that the average excess risk is bounded lower and upper by the ℓ_2 distance of $\boldsymbol{\theta}_I$ and $\boldsymbol{\theta}_{0I}$.

Lemma A.1. *Under Condition 2, for some constants c and c_1 depending on M, $\mathcal{X}$ and $\mathcal{Z}$, we have for any I and any $\boldsymbol{\theta} \in \Theta(M)$*

$$c\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2 \leq \bar{\varepsilon}(\boldsymbol{\theta}_I \mid \boldsymbol{\theta}_{0I}) \leq c_1\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2.$$

Proof. We defer the proof to Appendix A.5.

Note that

$$\begin{aligned}
\bar{\varepsilon}(\boldsymbol{\theta}_I \mid \boldsymbol{\theta}_{0I}) &= -\frac{1}{n}\sum_{i=1}^{n}\int \log\frac{f_{\boldsymbol{\psi}(x_i,z_i,I)}(Y_i)}{f_{\boldsymbol{\psi}_0(x_i,z_i,I)}(Y_i)} f_{\boldsymbol{\psi}_0(x_i,z_i,I)}(Y_i) d\mu \\
&= -\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\log\frac{f_{\boldsymbol{\psi}(x_i,z_i,I)}(Y_i)}{f_{\boldsymbol{\psi}_0(x_i,z_i,I)}(Y_i)}\right] \\
&= \frac{1}{n}\mathbb{E}[l_n(\boldsymbol{\theta}_{0I}, I) - l_n(\boldsymbol{\theta}_I, I)] = \frac{1}{n}\mathbb{E}Z_n(\boldsymbol{\theta}_I).
\end{aligned}$$

Based on Lemma A.1, on the set $\{V_n \leq R\lambda_0\}$, we have

$$\begin{aligned}
Z_n(\boldsymbol{\theta}_I) &\geq \mathbb{E}Z_n(\boldsymbol{\theta}_I) - nR\lambda_0(\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_1 \vee \lambda_0) \\
&\geq cn\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2 - nR\lambda_0(\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_1 \vee \lambda_0),
\end{aligned}$$

and

$$\begin{aligned}
Z_n(\boldsymbol{\theta}_I) &\leq \mathbb{E}Z_n(\boldsymbol{\theta}_I) + nR\lambda_0(\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_1 \vee \lambda_0) \\
&\leq c_1 n\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2 + nR\lambda_0(\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_1 \vee \lambda_0).
\end{aligned}$$

Case 1. If $\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_1 \leq \lambda_0$, then we have

$$cn\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2 - nR\lambda_0^2 \leq Z_n(\boldsymbol{\theta}_I) \leq c_1 n\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2 + nR\lambda_0^2.$$

Case 2. If $\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_1 \geq \lambda_0$, then we have

$$\begin{aligned}
Z_n(\boldsymbol{\theta}_I) &\geq cn\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2 - nR\lambda_0\sqrt{|I|+3}\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2 \\
&= cn\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2 - 2\sqrt{n}R\lambda_0\sqrt{(|I|+3)/2c}\sqrt{cn\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2/2} \\
&\geq cn\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2 - \frac{n}{2c}R^2\lambda_0^2(|I|+3) - \frac{cn}{2}\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2 \\
&= \frac{cn}{2}\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2 - \frac{R}{2c}nR\lambda_0^2(|I|+3).
\end{aligned}$$

Similarly,

$$\begin{aligned}
Z_n(\boldsymbol{\theta}_I) &\leq c_1 n\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2 + \frac{n}{4c_1}R^2\lambda_0^2(|I|+3) + c_1 n\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2 \\
&= 2c_1 n\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2 + \frac{R}{4c_1}nR\lambda_0^2(|I|+3).
\end{aligned}$$

Combining the above two cases, we obtain the results we want. □

A.3. Proof of Theorem 3.3

The posterior of model I can be written as

$$\begin{aligned}
\Pi(I \mid Y) &= C\int_{\Theta} \exp\{l_n(\boldsymbol{\theta}_I, I)\}\exp\left\{-\frac{1}{2}(\tau_1^{-2}\boldsymbol{b}_I^T\boldsymbol{b}_I + \tau_0^{-2}\boldsymbol{b}_{I^c}^T\boldsymbol{b}_{I^c})\right\} \\
&\quad \times (\tau_1/q)^{-|I|}\,(\tau_0/1-q)^{|I|-p}\,\pi_N\,(s_1/\sigma_\alpha)\,\pi_N\,(s_2/\sigma_\alpha)\,\pi_\rho(\rho; a_0, b_0)d\boldsymbol{\theta} \\
&= C\,(\tau_1(1-q)/q)^{-|I|}\,(1-q)^p\exp\{l_n(\boldsymbol{\theta}_{0I}, I)\}(2\pi)^{-|I|/2} \\
&\quad \times \int_{\Theta_I}\exp\{-Z_n(\boldsymbol{\theta}_I)\}\exp\left\{-\frac{1}{2}\tau_1^{-2}\boldsymbol{b}_I^T\boldsymbol{b}_I\right\}\pi(s_1, s_2, \rho)d\boldsymbol{\theta}_I.
\end{aligned}$$

Therefore, with the notation $\nu_n = \tau_1(1-q)/q$, we can write the posterior as

$$\Pi(I \mid Y) = C\nu_n^{-|I|}\,(1-q)^p\exp\{l_n(\boldsymbol{\theta}_{0I}, I)\}R_I,$$

where, in the set $\{V_n \leq R\lambda_0\}$,

$$\begin{aligned}
R_I &= (2\pi)^{-\frac{|I|}{2}}\int_{\Theta_I}\exp\{-Z_n(\boldsymbol{\theta}_I)\}\exp\left\{-\frac{1}{2}\tau_1^{-2}\boldsymbol{b}_I^T\boldsymbol{b}_I\right\}\pi(s_1, s_2, \rho)d\boldsymbol{\theta}_I \\
&\leq (2\pi)^{-\frac{|I|}{2}}\int_{\boldsymbol{b}_I}\exp\left\{-cn\|\boldsymbol{b}_I - \boldsymbol{b}_{0I}\|_2^2 + c_2 nR\lambda_0^2\,(|I|+3)\right\}\exp\left\{-\frac{\boldsymbol{b}_I^T\boldsymbol{b}_I}{2\tau_1^2}\right\}d\boldsymbol{b}_I \\
&\quad \times \prod_{j=1}^{2}\int_{s_j}\frac{1}{\sqrt{2\pi}\sigma_\alpha}\exp\left\{-\frac{s_j^2}{2\sigma_\alpha^2} - cn(s_j - s_{j0})^2\right\}ds_j\int_\rho \pi_\rho(\rho)e^{-cn(\rho-\rho_0)^2}d\rho \\
&\leq \exp\left\{c_2 nR\lambda_0^2\,(|I|+3)\right\}\frac{1}{2cn\sigma_\alpha^2+1}\sqrt{\frac{2\pi}{2cn}}\mathbb{E}\pi_\rho(\rho)\prod_{j=1}^{2}\exp\left\{-\frac{cns_{j0}^2}{2cn\sigma_\alpha^2+1}\right\} \qquad (9) \\
&\quad \times (2\pi)^{-\frac{|I|}{2}}\exp\left\{-\frac{1}{2}\left(\frac{cn}{2cn\tau_1^2+1}\right)\boldsymbol{b}_{0I}^T\boldsymbol{b}_{0I}\right\}
\end{aligned}$$

$$\times \int_{\boldsymbol{b}_I} \exp\left\{-\frac{2cn\tau_1^2+1}{4\tau_1^2}\left(\boldsymbol{b}_I - \frac{2cn\tau_1^2}{2cn\tau_1^2+1}\boldsymbol{b}_{0I}\right)^T\left(\boldsymbol{b}_I - \frac{2cn\tau_1^2}{2cn\tau_1^2+1}\boldsymbol{b}_{0I}\right)\right\} d\boldsymbol{b}_I$$
$$\preceq \left(\frac{2\tau_1^2}{2cn\tau_1^2+1}\right)^{|I|/2}\frac{1}{2cn\sigma_\alpha^2+1}\sqrt{\frac{1}{2cn}}\exp\left\{c_2nR\lambda_0^2\left(|I|+3\right)\right\},$$

where the last approximation follows since $\pi_\rho(\rho) \leq C$ for some constant $C > 0$.

For the true model I_0, we have

$$\Pi[I = I_0 \mid Y] = C\nu_n^{-|I_0|}(1-q)^p \exp\{l_n(\boldsymbol{\theta}_{0I_0}, I_0)\}R_{I_0},$$

where

$$R_{I_0} = (2\pi)^{-\frac{|I_0|}{2}}\int_{\Theta_{I_0}} \exp\{-Z_n(\boldsymbol{\theta}_{I_0})\}\exp\left\{-\frac{1}{2}\tau_1^{-2}\boldsymbol{b}_{I_0}^T\boldsymbol{b}_{I_0}\right\}\pi(s_1, s_2, \rho)d\boldsymbol{\theta}_{I_0}. \tag{10}$$

We now derive a lower bound on R_{I_0}. Similarly, we have

$$R_{I_0} \geq \int_{s_1,s_2,\rho}\pi(s_1,s_2,\rho)ds_1ds_2d\rho(2\pi)^{-\frac{|I_0|}{2}}$$
$$\times\int_{\boldsymbol{b}_{I_0}}\exp\left\{-c_1n\|\boldsymbol{\theta}_{I_0} - \boldsymbol{\theta}_{0I_0}\|_2^2 - c_3nR\lambda_0\left(|I_0|+3\right)\right\}\exp\left\{-\frac{\boldsymbol{b}_{I_0}^T\boldsymbol{b}_{I_0}}{2\tau_1^2}\right\}d\boldsymbol{b}_{I_0} \tag{11}$$
$$\succeq \left(\frac{2\tau_1^2}{2c_1n\tau_1^2+1}\right)^{|I_0|/2}\frac{1}{2c_1n\sigma_\alpha^2+1}\sqrt{\frac{1}{2c_1n}}\exp\{-c_3nR\lambda_0^2\left(|I_0|+3\right)\}.$$

Here in the integral of ρ we have $\mathbb{E}(\pi_\rho(\rho))$ bounded below by some constant using a Gaussian distribution. In fact, it is larger than the integral near ρ_0 where $\pi_\rho(\rho) \geq C$ for some constant $C > 0$ in the interval.

Now we discuss the posterior ratios.

$$\frac{\Pi[I = I_1 \mid Y]}{\Pi[I = I_0 \mid Y]} \preceq \frac{\nu_n^{-|I_1|}\exp\{l_n(\boldsymbol{\theta}_{0I_1}, I_1)\}\left(\frac{2\tau_1^2}{2cn\tau_1^2+1}\right)^{|I_1|/2}}{\nu_n^{-|I_0|}\exp\{l_n(\boldsymbol{\theta}_{0I_0}, I_0)\}\left(\frac{2\tau_1^2}{2c_1n\tau_1^2+1}\right)^{|I_0|/2}}$$
$$\times\frac{\frac{1}{2cn\sigma_\alpha^2+1}\sqrt{\frac{1}{2cn}}\exp\{c_2nR\lambda_0^2\left(|I_1|+3\right)\}}{\frac{1}{2c_1n\sigma_\alpha^2+1}\sqrt{\frac{1}{2c_1n}}\exp\{-c_3nR\lambda_0^2\left(|I_0|+3\right)\}}$$
$$\preceq\left(\frac{1-q}{q}\right)^{-(|I_1|-|I_0|)}\frac{(cn\tau_1^2+1/2)^{-|I_1|/2}}{(c_1n\tau_1^2+1/2)^{-|I_0|/2}}$$
$$\times\exp\{c_2nR\lambda_0^2\left(|I_1|+3\right)+c_3nR\lambda_0^2\left(|I_0|+3\right)\}\exp\{l_n(\boldsymbol{\theta}_{0I_1}, I_1) - l_n(\boldsymbol{\theta}_{0I_0}, I_0)\}.$$

Given the orders of prior parameters, we have

$$\frac{\Pi[I = I_1 \mid Y]}{\Pi[I = I_0 \mid Y]}$$
$$\preceq p^{-(\tilde{C}+2)(|I_1|-|I_0|)+c_2R(|I_1|+3)+c_3R(|I_0|+3)}\exp\{l_n(\boldsymbol{\theta}_{0I_1}, I_1) - l_n(\boldsymbol{\theta}_{0I_0}, I_0)\}$$
$$= p^{-(\tilde{C}+2-c_2R)(|I_1|-|I_0|)+(c_2+c_3)R(|I_0|+3)}\exp\{l_n(\boldsymbol{\theta}_{0I_1}, I_1) - l_n(\boldsymbol{\theta}_{0I_0}, I_0)\}.$$

Next we consider the following three cases one by one:

1. Over-fitted models: $M_1 = \{I_1 : I_1 \supset I_0, I_1 \neq I_0, |I_1| \leq m_n\}$,
2. Large models: $M_2 = \{I_1 : |I_0| < |I_1| \leq m_n\}$,
3. Under-fitted models: $M_3 = \{I_1 : I_1 \not\supset I_0, |I_1| \leq |I_0|\}$.

Over-fitted models: if $I_1 \in M_1$, we have $l_n(\boldsymbol{\theta}_{0I_1}, I_1) = l_n(\boldsymbol{\theta}_{0I_0}, I_0)$, thus

$$\frac{\Pi[I = I_1 \mid Y]}{\Pi[I = I_0 \mid Y]} \preceq p^{-(\tilde{C}+2-c_2R)(|I_1|-|I_0|)+(c_2+c_3)R(|I_0|+3)}.$$

Since R is an arbitrary positive number no less than $\bar{C}$, we can set it to be $\bar{C}$. Define $\tilde{C}$ to be $(c_2 + (c_2 + c_3)(|I_0| + 3))\bar{C}$. Then for all models in M_1

$$\begin{aligned}
\sum_{I_1 \in M_1} \frac{\Pi[I = I_1 \mid Y]}{\Pi[I = I_0 \mid Y]} &\preceq \sum_{d=|I_0|+1}^{m_n} \binom{p-|I_0|}{d-|I_0|} p^{-(\tilde{C}+2-c_2\bar{C})(d-|I_0|)+(c_2+c_3)\bar{C}(|I_0|+3)} \\
&\preceq \sum_{d=|I_0|+1}^{m_n} p^{d-|I_0|} p^{-(\tilde{C}+2-c_2\bar{C})(d-|I_0|)+(c_2+c_3)\bar{C}(|I_0|+3)} \\
&\preceq \sum_{d=|I_0|+1}^{m_n} p^{-(c_2+c_3)(|I_0|+3)\bar{C}(d-|I_0|-1)} p^{-(d-|I_0|)} \\
&\preceq p^{-1} \longrightarrow 0.
\end{aligned}$$

Large models: if $|I_1| > |I_0|$, we use $I_1^* = I_1 \cup I_0$. Thus $\boldsymbol{\theta}_{0I_1^*}$ denotes the $|I_1^*| \times 1$ vector including $\boldsymbol{\theta}_{0I_0}$ for I_0 and zeros for $I_1 \cap I_0^c$. We use $\boldsymbol{\theta}_{1I_1^*} \in \Theta_{I_1^*}(M)$ to denote the vector with $\boldsymbol{\theta}_{0I_1}$ for I_1 and zeros for $I_1^c \cap I_0$. Then we have

$$\begin{aligned}
\frac{\Pi[I = I_1 \mid Y]}{\Pi[I = I_0 \mid Y]} &\preceq p^{-(\tilde{C}+2-c_2R)(|I_1|-|I_0|)+(c_2+c_3)R(|I_0|+3)} \exp\{-Z_n(\boldsymbol{\theta}_{1I_1^*})\} \\
&\preceq p^{-(\tilde{C}+2-2c_2R)(|I_1|-|I_0|)+(3c_2+c_3)R(|I_0|+3)} \exp\left\{-cn\|\boldsymbol{\theta}_{0,I_1^c\cap I_0}\|_2^2\right\} \\
&\preceq p^{-(\tilde{C}+2-2c_2R)(|I_1|-|I_0|)+(3c_2+c_3)R(|I_0|+3)} \exp\left\{-cn \min_j b_{j0}^2\right\}.
\end{aligned}$$

Thus, for all models in M_2,

$$\begin{aligned}
&\sum_{I_1 \in M_2} \frac{\Pi[I = I_1 \mid Y]}{\Pi[I = I_0 \mid Y]} \preceq \exp\left\{-cn \min_j b_{j0}^2\right\} \\
&\times \sum_{d=|I_0|+1}^{m_n} \sum_{h=0}^{|I_0|-1} \binom{p-|I_0|}{d-h} \binom{|I_0|}{h} p^{-(\tilde{C}+2-2c_2R)(d-|I_0|)+(3c_2+c_3)R(|I_0|+3)} \\
&\preceq \sum_{d=|I_0|+1}^{m_n} \sum_{h=0}^{|I_0|-1} p^{d-h}|I_0|^h p^{-(\tilde{C}+2-2c_2R)(d-|I_0|)+(3c_2+c_3)R(|I_0|+3)} \exp\left\{-cn \min_j b_{j0}^2\right\} \\
&\preceq \sum_{d=|I_0|+1}^{m_n} m_n p^{d-|I_0|} p^{-(\tilde{C}+2-2c_2R)(d-|I_0|)+(3c_2+c_3)R(|I_0|+3)+|I_0|} \exp\left\{-cn \min_j b_{j0}^2\right\} \\
&\preceq p^{-\tilde{C}+2c_2R+(3c_2+c_3)R(|I_0|+3)+|I_0|} \exp\left\{-cn \min_j b_{j0}^2\right\} \\
&\leq p^{-\tilde{C}+2c_2R+(3c_2+c_3)R(|I_0|+3)+|I_0|} p^{-cC_0|I_0|} \longrightarrow 0,
\end{aligned}$$

for some $C_0 \geq ((3c_2 + c_3)R + 1)/c + (2c_2R + 3(3c_2 + c_3)R - \tilde{C})/(c|I_0|)$ in Condition 4.
Under-fitted models: if $I_1 \not\supset I_0$ and $|I_1| \leq |I_0|$, similarly, we have

$$\frac{\Pi[I = I_1 \mid Y]}{\Pi[I = I_0 \mid Y]} \preceq p^{-(\tilde{C}+2-c_2R)(|I_1|-|I_0|)+(2c_2+c_3)R(|I_0|+3)} \exp\left\{-cn \min_j b_{j0}^2\right\}.$$

Then we have for all models in M_3

$$\begin{aligned}
&\sum_{I_1 \in M_3} \frac{\Pi[I = I_1 \mid Y]}{\Pi[I = I_0 \mid Y]} \preceq \exp\left\{-cn \min_j b_{j0}^2\right\} \\
&\times \sum_{d=0}^{|I_0|} \sum_{h=0}^{d} \binom{p - |I_0|}{d - h} \binom{|I_0|}{h} p^{-(\tilde{C}+2-c_2R)(d-|I_0|)+(2c_2+c_3)R(|I_0|+3)} \\
&\preceq \sum_{d=0}^{|I_0|} \sum_{h=0}^{d} p^{d-h} |I_0|^h p^{-(\tilde{C}+2-c_2R)(d-|I_0|)+(2c_2+c_3)R(|I_0|+3)} \exp\left\{-cn \min_j b_{j0}^2\right\} \\
&\preceq p^{(\tilde{C}+2+(c_2+c_3)R)|I_0|+3(2c_2+c_3)R} p^{-cC_0|I_0|} \longrightarrow 0,
\end{aligned}$$

for some $C_0 \geq (\tilde{C} + 2 + (c_2 + c_3)R)/c + 3(2c_2 + c_3)R/(c|I_0|)$.

Combing the results, we have $\sum_{I_1 \in \mathcal{I}(m_n)\setminus\{I_0\}} \frac{\Pi[I=I_1|Y]}{\Pi[I=I_0|Y]} \xrightarrow{\text{P}} 0$, which in turn implies that $\Pi[I = I_0 \mid Y] \xrightarrow{\text{P}} 1$ on $\mathcal{I}(m_n)$. □

A.4. *Proof of Theorem 3.4*

Define the set $\mathcal{D}_\varepsilon = \{(\boldsymbol{\theta}, I) \in \Theta(M) \times \mathcal{I}(m_n) : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \geq \varepsilon, I = I_0\}$. We have

$$\Pi\left[(\boldsymbol{\theta}, I) \in \Theta(M) \times \mathcal{I}(m_n) : \|\theta - \theta_0\|_2 \geq \varepsilon \mid Y\right] \leq \Pi\left[I \neq I_0 \mid Y\right] + \Pi\left[\mathcal{D}_\varepsilon \mid Y\right].$$

On $\{V_n \leq R\lambda_0\}$, by Theorem 3.3, we have on $\mathcal{I}(m_n)$, $\Pi[I \neq I_0 \mid Y] \xrightarrow{\text{P}} 0$, so we only need to study $\Pi[\mathcal{D}_\varepsilon \mid Y]$, which can be rewritten as

$$\Pi\left[\mathcal{D}_\varepsilon \mid Y\right] = \frac{\int_{\mathcal{D}_\varepsilon} \sum_I \exp\{l_n(\boldsymbol{\theta}_I, I) - l_n(\boldsymbol{\theta}_{0I}, I)\}\pi(\boldsymbol{\theta}_I, \boldsymbol{\theta}_{I^c}, I) d\boldsymbol{\theta}_I d\boldsymbol{\theta}_{I^c}}{\int_{\Theta(M)} \sum_I \exp\{l_n(\boldsymbol{\theta}_I, I) - l_n(\boldsymbol{\theta}_{0I}, I)\}\pi(\boldsymbol{\theta}_I, \boldsymbol{\theta}_{I^c}, I) d\boldsymbol{\theta}_I d\boldsymbol{\theta}_{I^c}}. \tag{12}$$

For the denominator in (12), we have

$$\begin{aligned}
&\int_{\Theta(M)} \sum_I \exp\{l_n(\boldsymbol{\theta}_I, I) - l_n(\boldsymbol{\theta}_{0I}, I)\}\pi(\boldsymbol{\theta}_I, \boldsymbol{\theta}_{I^c}, I) d\boldsymbol{\theta}_I d\boldsymbol{\theta}_{I^c} \\
&\geq \int_{\Theta(M)} \exp\{l_n(\boldsymbol{\theta}_{I_0}, I_0) - l_n(\boldsymbol{\theta}_{0I_0}, I_0)\}\pi(\boldsymbol{\theta}_I, \boldsymbol{\theta}_{I^c}, I = I_0) d\boldsymbol{\theta}_{I_0} d\boldsymbol{\theta}_{I_0^c} \\
&\geq \int_{\Theta(M)} \exp\{-Z_n(\boldsymbol{\theta}_{I_0})\}\pi(\boldsymbol{\theta}_I, \boldsymbol{\theta}_{I^c}, I = I_0) d\boldsymbol{\theta}_{I_0} d\boldsymbol{\theta}_{I_0^c} \\
&= C\nu_n^{-|I_0|}(1 - q)^p R_{I_0} \\
&\succeq \nu_n^{-|I_0|}(1 - q)^p (\tau_1^{-2}/2 + c_1 n)^{-|I_0|/2} \frac{\sqrt{1/2c_1 n}}{2c_1 n \sigma_\alpha^2 + 1} \exp\{-c_3 n R \lambda_0^2(|I_0| + 3)\},
\end{aligned}$$

in which R_{I_0} is defined in (10) and the last inequality follows directly from the lower bound on R_{I_0} in (11).

For the numerator in (12), we have

$$\begin{aligned}
&\int_{\mathcal{D}_\varepsilon}\sum_I \exp\{l_n(\boldsymbol{\theta}_I, I) - l_n(\boldsymbol{\theta}_{0I}, I)\}\pi(\boldsymbol{\theta}_I, \boldsymbol{\theta}_{I^c}, I)d\boldsymbol{\theta}_I d\boldsymbol{\theta}_{I^c} \\
&\leq \int_{\mathcal{D}_\varepsilon}\sum_I \exp\left\{-cn\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2 + c_2 nR\lambda_0^2(|I|+3)\right\}\pi(\boldsymbol{\theta}_I, \boldsymbol{\theta}_{I^c}, I)d\boldsymbol{\theta}_I d\boldsymbol{\theta}_{I^c} \\
&= \left(\frac{1-q}{\sqrt{2\pi}\tau_0}\right)^p \left(\frac{\tau_0 q}{\tau_1(1-q)}\right)^{|I_0|} \int_{\{\boldsymbol{\theta}\in\Theta(M):\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2\geq\varepsilon\}} \exp\left\{-cn\|\boldsymbol{\theta}_{I_0} - \boldsymbol{\theta}_{0I_0}\|_2^2\right\} \\
&\times \exp\left\{-\frac{\boldsymbol{b}_{I_0}^T\boldsymbol{b}_{I_0}}{2\tau_1^2} - \frac{\boldsymbol{b}_{I_0^c}^T\boldsymbol{b}_{I_0^c}}{2\tau_0^2}\right\}\pi(s_1, s_2, \rho)d\boldsymbol{\theta}_{I_0}d\boldsymbol{\theta}_{I_0^c}\exp\left\{c_2 nR\lambda_0^2(|I_0|+3)\right\}.
\end{aligned}$$

The integral set can be rewritten as

$$\begin{aligned}
&\{\boldsymbol{\theta} \in \Theta(M) : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2 \geq \varepsilon^2\} = \{\boldsymbol{\theta} \in \Theta(M) : \|\boldsymbol{\theta}_{I_0} - \boldsymbol{\theta}_{0I_0}\|_2^2 + \|\boldsymbol{\theta}_{I_0^c}\|_2^2 \geq \varepsilon^2\} \\
&= \{\boldsymbol{\theta} \in \Theta(M) : \|\boldsymbol{\theta}_{I_0} - \boldsymbol{\theta}_{0I_0}\|_2^2 + (1+p)\|\boldsymbol{\theta}_{I_0^c}\|_2^2 \geq \varepsilon^2 + p\|\boldsymbol{\theta}_{I_0^c}\|_2^2\}.
\end{aligned}$$

Thus the integral is bounded above by

$$\begin{aligned}
&\int_{\Theta_{I_0}} \exp\left\{-cn\|\boldsymbol{\theta}_{I_0} - \boldsymbol{\theta}_{0I_0}\|_2^2 - \frac{\boldsymbol{b}_{I_0}^T\boldsymbol{b}_{I_0}}{2\tau_1^2} + \frac{\|\boldsymbol{\theta}_{I_0} - \boldsymbol{\theta}_{0I_0}\|_2^2}{2(1+p)\tau_0^2}\right\}\pi(s_1, s_2, \rho)d\boldsymbol{\theta}_{I_0} \\
&\times \int_{\Theta_{I_0^c}} \exp\left\{-\frac{p}{2(1+p)\tau_0^2}\|\boldsymbol{b}_{I_0^c}\|_2^2\right\}db_{I_0^c}\exp\left\{-\frac{1}{2(1+p)\tau_0^2}\varepsilon^2\right\},
\end{aligned}$$

where the second integral is equal to

$$\left(\sqrt{2\pi}\sqrt{\frac{1+p}{p}}\tau_0\right)^{p-|I_0|} \preceq \left(\sqrt{2\pi}\tau_0\right)^{p-|I_0|}e^{1/2}.$$

Thus similar to steps for deriving the upper bound on R_I in (9), we can obtain an upper bound on the numerator as

$$\begin{aligned}
&\int_{\mathcal{D}_\varepsilon}\sum_I \exp\{l_n(\boldsymbol{\theta}_I, I) - l_n(\boldsymbol{\theta}_{0I}, I)\}\pi(\boldsymbol{\theta}_I, \boldsymbol{\theta}_{I^c}, I)d\boldsymbol{\theta}_I d\boldsymbol{\theta}_{I^c} \\
&\preceq \nu_n^{-|I_0|}(1-q)^p\left(\frac{1}{2\tau_1^2} + cn - \frac{1}{2(1+p)\tau_0^2}\right)^{-|I_0|/2}\left(\frac{2cn(1+p)\tau_0^2 - 1}{(1+p)\tau_0^2}\sigma_\alpha^2 + 1\right)^{-1} \\
&\times \sqrt{\frac{1}{2cn - 1/[(1+p)\tau_0^2]}}\exp\{c_2 nR\lambda_0^2(|I_0|+3)\}\exp\left\{-\frac{1}{2(1+p)\tau_0^2}\varepsilon^2\right\}.
\end{aligned}$$

Thus given conditions 3 and 5,

$$\begin{aligned}
&\Pi\left[\mathcal{D}_\varepsilon \mid Y\right] \preceq \left(\frac{\tau_1^{-2}/2 + c_1 n}{\tau_1^{-2}/2 + cn - 1/[2(1+p)\tau_0^2]}\right)^{|I_0|/2}\frac{2c_1 n\sigma_\alpha^2 + 1}{(2cn - 1/[(1+p)\tau_0^2])\sigma_\alpha^2 + 1} \\
&\times \sqrt{\frac{2c_1 n}{2cn - 1/[(1+p)\tau_0^2]}}\exp\{(c_2+c_3)nR\lambda_0^2(|I_0|+3)\}\exp\left\{-\frac{1}{2(1+p)\tau_0^2}\varepsilon^2\right\} \\
&\preceq \exp\{(c_2+c_3)R(|I_0|+3)\log p\}\exp\left\{-\frac{n}{2}\varepsilon^2\right\}.
\end{aligned}$$

If we choose $\varepsilon = \sqrt{C'|I_0|\log p/n}$ for $C' \geq 8(c_2 + c_3)R$, the result follows. □

A.5. Proof of Lemma A.1

Before the proof of Lemma A.1, we first claim two necessary lemmas as follows.

Lemma A.2. *Under Condition 2, we have*

$$\varepsilon(\boldsymbol{\psi}(x,z,I) \mid \boldsymbol{\psi}_0(x,z,I)) \geq c_0\|\boldsymbol{\psi}(x,z,I) - \boldsymbol{\psi}_0(x,z,I)\|_2^2,$$
$$\varepsilon(\boldsymbol{\psi}(x,z,I) \mid \boldsymbol{\psi}_0(x,z,I)) \leq c_1 \left\|\boldsymbol{\psi}(x,z,I) - \boldsymbol{\psi}_0(x,z,I)\right\|_2^2.$$

Proof: The proof can be found in Appendix A.6.

Lemma A.3. *For $\boldsymbol{\theta} \in \Theta(M)$ and $x \in \mathcal{X}$, there exist some constants c and C, such that*

$$c|x^T\boldsymbol{\gamma} - x^T\boldsymbol{\gamma}_0|^2 \leq |\pi(x^T\boldsymbol{\gamma}) - \pi(x^T\boldsymbol{\gamma}_0)|^2 \leq C|x^T\boldsymbol{\gamma} - x^T\boldsymbol{\gamma}_0|^2.$$

Proof: The proof can be found in Appendix A.7.

With the notation $\tilde{\boldsymbol{\phi}}_j = (\boldsymbol{\phi}^T, s_j)^T$ for $j = 1, 2$, by Lemma A.2, we have

$$\begin{aligned}\mathbb{E}Z_n(\boldsymbol{\theta}_I) &\geq c_0 \sum_{i=1}^n \|\boldsymbol{\psi}(x_i, z_i, I) - \boldsymbol{\psi}_0(x_i, z_i, I)\|_2^2 \\ &= c_0 \sum_{j=1}^2 (\tilde{\boldsymbol{\phi}}_{jI} - \tilde{\boldsymbol{\phi}}_{j0I})^T \tilde{Z}_I^T \tilde{Z}_I (\tilde{\boldsymbol{\phi}}_{jI} - \tilde{\boldsymbol{\phi}}_{j0I}) \\ &\quad + c_0 \left[\sum_{i=1}^n |\pi(x_{iI}^T\boldsymbol{\gamma}_I) - \pi(x_{iI}^T\boldsymbol{\gamma}_{0I})|^2 + n|\rho - \rho_0|^2\right].\end{aligned}$$

By the lower bound in Lemma A.3, we have

$$\begin{aligned}&\mathbb{E}Z_n(\boldsymbol{\theta}_I) \geq c_0 \sum_{j=1}^2 (\tilde{\boldsymbol{\phi}}_{jI} - \tilde{\boldsymbol{\phi}}_{j0I})^T \tilde{Z}_I^T \tilde{Z}_I (\tilde{\boldsymbol{\phi}}_{jI} - \tilde{\boldsymbol{\phi}}_{j0I}) \\ &+ c_0 c(\boldsymbol{\gamma}_I - \boldsymbol{\gamma}_{0I})^T X_I^T X_I (\boldsymbol{\gamma}_I - \boldsymbol{\gamma}_{0I}) + c_0 n|\rho - \rho_0|^2 \\ &\geq c_0 n\lambda_1 \left(2\|\boldsymbol{\phi} - \boldsymbol{\phi}_0\|_2^2 + |s_1 - s_{10}|^2 + |s_2 - s_{20}|^2 + c_2\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|_2^2 + |\rho - \rho_0|^2\right) \\ &\geq cn\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2.\end{aligned}$$

Similarly we can upper bound $\mathbb{E}Z_n(\boldsymbol{\theta}_I)$. We have

$$\begin{aligned}\mathbb{E}Z_n(\boldsymbol{\theta}_I) &\leq c_1 \sum_{j=1}^2 (\tilde{\boldsymbol{\phi}}_{jI} - \tilde{\boldsymbol{\phi}}_{j0I})^T \tilde{Z}_I^T \tilde{Z}_I (\tilde{\boldsymbol{\phi}}_{jI} - \tilde{\boldsymbol{\phi}}_{j0I}) \\ &\quad + c_1 c_2 (\boldsymbol{\gamma}_I - \boldsymbol{\gamma}_{0I})^T X_I^T X_I (\boldsymbol{\gamma}_I - \boldsymbol{\gamma}_{0I}) + c_1 n|\rho - \rho_0|^2 \\ &\leq c_1 n\|\boldsymbol{\theta}_I - \boldsymbol{\theta}_{0I}\|_2^2.\end{aligned}$$

The results then follow. □

A.6. *Proof of Lemma A.2*

By Lemma 1 in [42], with some slight modification, we have

$$c_0\|\boldsymbol{\psi}(x,z,I)-\boldsymbol{\psi}_0(x,z,I)\|_2^2 \le \varepsilon(\boldsymbol{\psi}(x,z,I)\mid\boldsymbol{\psi}_0(x,z,I)),$$

for some constant c_0. For the other side of the inequality, we adopt a similar proof procedure to [42]. For fixed design finite mixture regression models, their Conditions 1,2 and 3 are automatically met with appropriate C_3, $\Lambda_{\min}$ and $\{\alpha_\varepsilon\}$, only depending on M. Thus by Taylor expansion, we have

$$\varepsilon(\boldsymbol{\psi}(x,z,I)\mid\boldsymbol{\psi}_0(x,z,I)) = (\boldsymbol{\psi}_I-\boldsymbol{\psi}_{0I})^T I(\boldsymbol{\psi}_{0I})(\boldsymbol{\psi}_k-\boldsymbol{\psi}_{0I})/2 + r_{\boldsymbol{\psi}_I},$$

where

$$\begin{aligned}\left|r_{\boldsymbol{\psi}_I}\right| &\le \frac{\left\|\boldsymbol{\psi}_I-\boldsymbol{\psi}_{0I}\right\|_1^3}{6}\int \sup_{\boldsymbol{\psi}_I\in\Psi_I}\max_{j_1,j_2,j_3}\left|\frac{\partial^3 l_{\boldsymbol{\psi}_I}}{\partial\boldsymbol{\psi}_{j_1I}\partial\boldsymbol{\psi}_{j_2I}\partial\boldsymbol{\psi}_{j_3I}}\right| f_{\boldsymbol{\psi}_{0I}}d\mu \\ &\le \frac{4C_3}{3}\left\|\boldsymbol{\psi}_I-\boldsymbol{\psi}_{0I}\right\|_2^3.\end{aligned}$$

By direct calculations, for $\boldsymbol{\theta}\in\Theta(M)$, $x\in\mathcal{X}$ and $z\in\mathcal{Z}$, the largest eigenvalue of information matrix $I\left(\boldsymbol{\psi}_0(x,z,I)\right)$ is bounded above, i.e.,

$$\sup_{x\in\mathcal{X},z\in\mathcal{Z}}\lambda_{\max}\left(I\left(\boldsymbol{\psi}_0(x,z,I)\right)\right)\le\Lambda_{\max},$$

where $\Lambda_{\max}$ is some finite constant. Hence

$$\varepsilon(\boldsymbol{\psi}(x,z,I)\mid\boldsymbol{\psi}_0(x,z,I)) \le \frac{\Lambda_{\max}}{2}\left\|\boldsymbol{\psi}_I-\boldsymbol{\psi}_{0I}\right\|_2^2+\frac{4C_3}{3}\left\|\boldsymbol{\psi}_I-\boldsymbol{\psi}_{0I}\right\|_2^3.$$

Thus we have

$$\begin{aligned}\varepsilon(\boldsymbol{\psi}(x,z,I)\mid\boldsymbol{\psi}_0(x,z,I)) &\le \frac{\Lambda_{\max}}{2}\left\|\boldsymbol{\psi}_I-\boldsymbol{\psi}_{0I}\right\|_2^2+\frac{4C_3}{3}\left\|\boldsymbol{\psi}_I-\boldsymbol{\psi}_{0I}\right\|_2^2\left\|\boldsymbol{\psi}_I-\boldsymbol{\psi}_{0I}\right\|_1 \\ &\le (\Lambda_{\max}/2+16C_3M/3)\left\|\boldsymbol{\psi}_I-\boldsymbol{\psi}_{0I}\right\|_2^2 \\ &\equiv c_1\left\|\boldsymbol{\psi}_I-\boldsymbol{\psi}_{0I}\right\|_2^2.\end{aligned}$$

□

A.7. *Proof of Lemma A.3*

The upper bound inequality comes from Lemma 9 in [48]. For the lower bound inequality, let $a=x^T\boldsymbol{\gamma}$, $b=x^T\boldsymbol{\gamma}_0$. Since $\boldsymbol{\theta}\in\Theta(M)$ and $x\in\mathcal{X}$, we have $a\le Q$ and $b\le Q$ for some finite constant Q. Let $\pi(x)=e^x/(1+e^x)$, so we have

$$\pi'(x)=\frac{e^x}{(1+e^x)^2}=\frac{1}{e^x+e^{-x}+2}\le\frac{1}{4}.$$

Thus, by the mean value theorem,

$$\left|\frac{\pi(a)-\pi(b)}{a-b}\right| = |\pi'(c)|\le\frac{1}{4},$$

where c is some constant between a and b. Therefore,

$$|\pi(x^T\boldsymbol{\gamma})-\pi(x^T\boldsymbol{\gamma}_0)|^2\le\frac{1}{16}|x^T\boldsymbol{\gamma}-x^T\boldsymbol{\gamma}_0|^2.$$

□

A.8. Case of finite dimension

For the finite-dimensional case $p = \mathcal{O}(1)$, one can use Taylor expansion around the true parameter $\boldsymbol{\theta}_0$ to achieve results analogous to Lemma 3.2, under common regularity conditions [22, 42]. The remaining steps are similar to those in the high-dimensional case. Since the proof of the finite-dimensional case is straightforward, throughout the paper we focus on elaborating the proof when p is infinite.

Additionally, we can also see that when $p = \mathcal{O}(1)$, one can extend the parameters by including infinitely many zeros, allowing the same procedure of the high-dimensional case to be applicable.

Appendix B: Additional simulation results

B.1. High correlation settings

In this subsection, we conduct additional experiments to assess the impact of higher correlations under both pairwise correlation settings and autoregressive correlation settings.

B.1.1. Pairwise correlation setting

We first consider the same settings as in Section 4.1 with $n = 200$. We increase the pairwise correlation ρ among the active prognostic or predictive covariates to $\{0.7, 0.8\}$, while keeping the pairwise correlations among inactive covariates and between active and inactive covariates fixed at 0.25. The results of variable selection performance are summarized in Table 6.

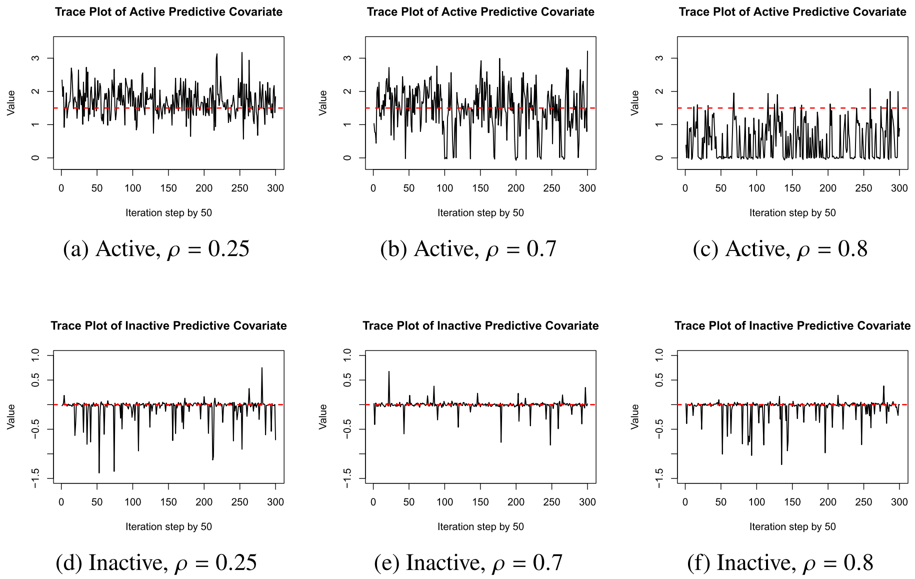
Table 6

Variable selection results in structured logistic-normal mixture settings with $n = 200$ and different values of pairwise covariate correlation ρ.

p	ρ	I^β						I^γ					
		TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
100	0.25	4	4	0	1	1	1	3.80	3.81	0.32	0.60	0.80	0.81
	0.7	4	4	0	1	1	1	3.39	3.44	0.57	0.33	0.48	0.49
	0.8	3.98	4	0	0.98	0.98	1	2.93	3.06	0.81	0.13	0.24	0.27
500	0.25	4	4	0	1	1	1	3.65	3.65	0.42	0.47	0.65	0.66
	0.7	3.99	4	0	0.99	0.99	1	2.83	3.02	0.54	0.14	0.20	0.29
	0.8	3.60	4	0	0.64	0.64	1	2.30	2.65	0.65	0.02	0.05	0.11

We observe that prognostic variable selection shows robustness to increasing correlations. When $\rho = 0.7$, the impact on the results is minimal, even in the high-dimensional setting with $p = 500$. When ρ increases to 0.8, our method still identifies more than half of the active covariates without false discoveries in the case of $p = 500$ and achieves perfect ranking in variable importance with TP$_s = 4$.

Predictive variable selection exhibits greater sensitivity to increasing correlation compared to prognostic selection, particularly in high-dimensional settings. The performance remains robust at $\rho = 0.7$, but deteriorates as correlation increases further. At the higher level of $\rho = 0.8$, TP declines and FP increases, suggesting that in the logistic model, some highly



(a) Active, $\rho = 0.25$ (b) Active, $\rho = 0.7$ (c) Active, $\rho = 0.8$

(d) Inactive, $\rho = 0.25$ (e) Inactive, $\rho = 0.7$ (f) Inactive, $\rho = 0.8$

Fig 4. *Post-burnin trace plots for an active predictive covariate and an inactive predictive covariate under different values of pairwise correlation ρ between active covariates.*

correlated active covariates may be overlooked, making it more difficult to accurately identify the true predictive variables.

To assess the impact of high correlation on MCMC convergence, we evaluate the mixing behavior of the Gibbs sampler under different pairwise correlation levels among active covariates, $\rho \in 0.25, 0.7, 0.8$, by examining trace plots and the effective sample size (ESS).

Since correlation primarily affects predictive covariates rather than prognostic covariates, our analysis focuses on the trace plots of the regression coefficients associated with the predictive covariates. We illustrate the MCMC mixing behavior using a representative trial under the same settings as in Section 4.1 with $n = 200$ and $p = 100$. Figure 4(a)-(c) presents the trace plots using every 50th iteration from a total of 15000 samples after the burn-in period for one active predictive covariate, while Figure 4(d)-(f) shows the trace plots for one inactive predictive covariate across different correlation levels.

The trace plots reveal a deterioration in convergence as ρ increases. For the active predictive covariate, the sampler mixes well when $\rho = 0.25$. However, as ρ increases to 0.8, the trace shows more frequent and prolonged visits to 0, suggesting that the sampler becomes more prone to switching the variable's inclusion status. For the inactive predictive covariate, the trace plots remain concentrated around zero across all levels of ρ as expected.

We further examine the impact of covariate correlation on MCMC convergence by computing the ESS across all prognostic and predictive covariates under different values of ρ, as summarized in Table 7. For prognostic covariates, the ESS remains notably high. For predictive covariates, the ESS is lower with more variation due to the additional complexity introduced by subgroup modeling. Overall, the ESS values for both components are satisfactory, indicating good mixing behavior of the Gibbs sampler regardless of the correlation.

Table 7

Averaged effective sample sizes with standard errors across all prognostic and predictive covariates under different values of pairwise correlation ρ with $n = 200$, $p = 100$, and post-burnin length 15000.

ρ	Prognostic	Predictive
0.25	13591 (1909.96)	924.1 (589.66)
0.7	13090 (3501.48)	1151.1 (624.92)
0.8	12887 (3792.69)	1328.0 (917.22)

Table 8

Variable selection results in structured logistic-normal mixture settings with $n = 200$ and different values of autoregressive correlation ρ.

p	ρ	I^β						I^γ					
		TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
100	0.25	4	4	0	1	1	1	3.84	3.81	0.50	0.51	0.85	0.81
	0.5	4	4	0	1	1	1	3.70	3.70	0.56	0.49	0.78	0.74
	0.6	4	4	0	1	1	1	3.31	3.44	0.51	0.38	0.56	0.59
	0.7	4	4	0	1	1	1	2.92	3.18	0.62	0.22	0.38	0.42
500	0.25	4	4	0	1	1	1	3.55	3.68	0.32	0.54	0.67	0.74
	0.5	4	4	0	1	1	1	3.23	3.34	0.57	0.39	0.54	0.58
	0.6	4	4	0	1	1	1	2.86	3.06	0.36	0.30	0.38	0.42
	0.7	3.98	4	0	0.99	0.99	1	2.35	2.71	0.47	0.21	0.23	0.26

B.1.2. Autoregressive correlation setting

We also consider the case of autoregressive (AR) correlation structure where prognostic and predictive covariates follow $N(0_p, \Sigma)$ with $\Sigma_{ij} = \rho^{|i-j|}$, exhibiting local dependencies rather than uniform pairwise correlation. The results are summarized in Table 8.

Similar conclusions can be drawn from the results in Table 8. Prognostic variable selection remains highly robust under the AR correlation setting, achieving perfect recovery across all ρ values. In contrast, predictive variable selection is more affected by the AR correlation structure, showing a gradual decline in true positives and an increase in false positives as ρ increases. This indicates that higher correlation among neighboring covariates makes it more challenging to distinguish true predictive variables from correlated noise, leading to reduced exact recovery rates.

B.2. Prediction errors for traditional subgroup settings

In this section, we provide the subgroup prediction errors of the traditional subgroup settings in addition to the results provided in Section 4.2. We only study settings S1 to S4 that have latent subgroups. We estimate the subgroup prediction errors using independent testing data with $n = 5000$. The generating procedure of the testing data is the same as that for the training data. The subgroup prediction error is calculated as the rate of observations that are misclassified into the wrong subgroup in our testing data. For our method, the classification for the subgroup membership is based on the estimated model from training data with the cutoff for the probability of the logistic model being 0.5. For tree-based methods, the classification depends on the split of the first node trained from the training data. The results are averaged based on the estimated models from 100 training trials.

Table 9

Subgroup prediction error of traditional subgroup settings in Section 4.2. In (a), the settings are low-dimensional with $p = 20$, while in (b) the settings are high-dimensional with $p = 200$.

(a) $p = 20$	S1	S2	S3	S4
BVSA	0.0258	0.1055	0.1055	0.1313
MOB	0.1543	0.3115	0.2874	0.1323
FindIt	0.4904	0.2876	0.1505	0.3931
PRIM	0.0494	0.3282	0.3016	0.1429
SeqBT	0.2731	0.2364	0.2354	0.3149
GUIDE	0.0443	0.2265	0.2266	0.1327

(b) $p = 200$	S1	S2	S3	S4
BVSA	0.0242	0.1058	0.1059	0.1279
MOB	0.0238	0.3015	0.2824	0.1295
PRIM	0.0321	0.2451	0.2205	0.1358
SeqBT	0.2570	0.2345	0.2370	0.2953
GUIDE	0.0837	0.2149	0.2134	0.1302

The results for low-dimensional settings with $p = 20$ are shown in (a) of Table 9. BVSA outperforms all other methods in all the settings, demonstrating its capability in subgroup prediction. The performance on subgroup identification in high-dimensional settings with $p = 200$ is shown in (b) of Table 9. BVSA has the smallest prediction errors in settings S2, S3, and S4, while in S1, its prediction error is also quite low compared to most other candidates, again indicating that BVSA is effective and stable.

Appendix C: Additional real data results

C.1. Predictive performance evaluation

In this subsection, we assess the predictive performance of our method. We compute the log predictive scores (LPS) on an independent test dataset following the formulation in [11]. Specifically, given a training dataset $\{(y_i, z_i, x_i, t_i)\}_{i=1}^{n_{\text{train}}}$, we adopt g-priors on the linear coefficients $\boldsymbol{\beta}$ and the logistic coefficients $\boldsymbol{\gamma}$ based on the selected model, and estimate the posterior distributions using Gibbs sampling. Let $\boldsymbol{\zeta}^{(s)}$ denote the posterior sample at the sth iteration. The LPS on the testing dataset $\{((y_i, z_i, x_i, t_i)\}_{i=1}^{n_{\text{test}}}$ is computed as

$$\widehat{\text{LPS}} = -\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i \mid z_i, x_i, t_i, \boldsymbol{\zeta}^{(s)})\right),$$

where S is the number of posterior samples and set as 3000 in our analysis.

In addition to LPS, we assess the effectiveness of the subgroup identification by evaluating treatment effect heterogeneity. For each test sample, we compute a predicted subgroup score using the estimated logistic component $\hat{\eta}_i = \mathbf{x}_i^T \hat{\boldsymbol{\gamma}}$. Based on the median $\hat{\eta}_{\text{med}}$, we divide the testing dataset into the "low" subgroup with $\hat{\eta}_i \leq \hat{\eta}_{\text{med}}$ and the "high" subgroup with $\hat{\eta}_i > \hat{\eta}_{\text{med}}$. Within each subgroup, we estimate the average treatment effect (ATE) by computing the difference in adjusted outcomes between treated and untreated individuals:

$$\widehat{\text{ATE}} = \frac{1}{n_1} \sum_{i:t_i=1} (y_i - z_i^T \hat{\boldsymbol{\beta}}) - \frac{1}{n_0} \sum_{i:t_i=0} (y_i - z_i^T \hat{\boldsymbol{\beta}}),$$

Table 10

Comparison of predictive performance and subgroup treatment effect.

Model	LPS ↓	Low Group ATE	High Group ATE	ATE Group Difference ↑
Baseline	3.172	0.352	1.447	1.095
+Age (pred)	3.171	0.574	1.141	0.567
+Hispanic (prog)	3.174	0.349	1.436	1.087

where n_1 and n_0 are the numbers of treated and untreated subjects in the subgroup, respectively. A large contrast between the estimated ATEs of the "low" and "high" subgroups suggests that the model successfully identifies meaningful heterogeneity in treatment response, thereby validating the subgroup discovery.

We conduct a comparison on the NSW dataset using the following three models: (1) the baseline model that includes the set of variables selected by BVSA; (2) a modified model that additionally includes *Age* in the predictive component; and (3) a modified model that additionally includes *Hispanic* in the prognostic component.

For each model, we compute the LPS on a held-out testing dataset comprising 20% of the observations, and estimate the subgroup treatment effects. The results are averaged over 100 independent trials and summarized in Table 10.

The results show that all three models achieve comparable LPS values, indicating similar predictive accuracy on the test data. However, when *Age* is included in the predictive component, the contrast between subgroup treatment effects becomes notably smaller. In contrast, the baseline model identified by BVSA yields a clearer separation between the low and high subgroups, suggesting more meaningful treatment effect heterogeneity. On the other hand, including *Hispanic* in the prognostic component does not lead to improvement in either LPS or subgroup contrast. The subgroup treatment effects remain nearly identical to the baseline model, indicating that *Hispanic* may not provide additional explanatory power beyond the selected covariates.

These findings support the effectiveness of BVSA in selecting variables. Including variables with seemingly small effects does not necessarily improve model performance or subgroup identification.

C.2. Prognostic and predictive illustration for ACTG320

In this section, we illustrate the roles of Lcd40 and Lrna0 in the structured models in Figure 5 and Figure 6, respectively. Patients are divided into high and low groups based on median cd40 and rna0, and we investigate the CD4 count change in different groups with or without the interaction of treatment.

Our method chooses Lcd40 as an active variable in the prognostic model with a posterior inclusion probability of 1, while Lrna0 is not included as a prognostic variable. We can observe in (a) and (b) of Figure 5 that the CD4 count changes in high and low Lcd40 levels show differential patterns. Their box plots do not overlap, and the estimated density curves have two different peaks, which supports the choice of Lcd40 as prognostic. However, in (a) and (b) of Figure 6, the CD4 count changes in high and low Lrna0 groups are similar, indicating that Lrna0 does not have a direct effect on the CD4 count change. For the predictive part, our method chooses both Lcd40 and Lrna0. From (c) and (d) of both Figure 5 and Figure 6, the

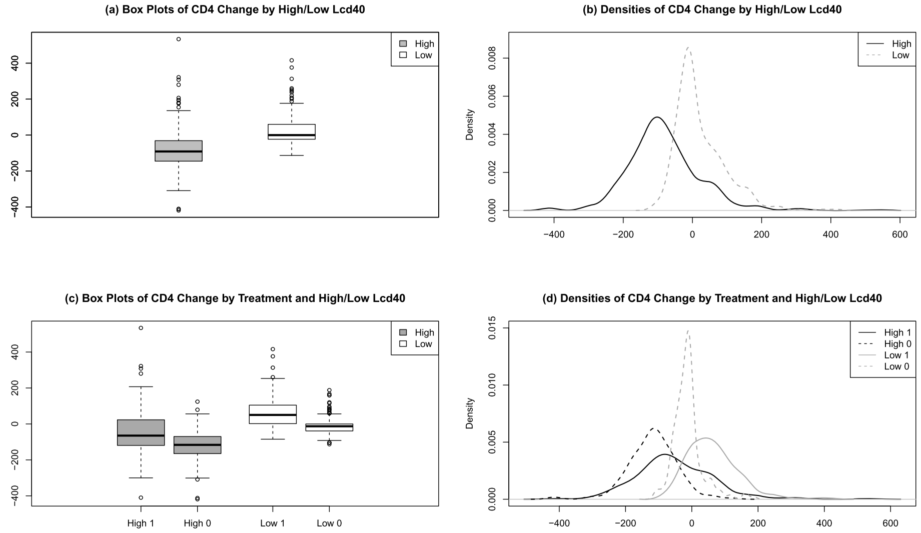


Fig 5. *Prognostic or predictive effects of the natural logarithm of baseline CD4 counts on the CD4 count change at week 24 for ACTG320 study with the prognostic effect shown in (a) and (b), and the predictive effect shown in (c) and (d).*

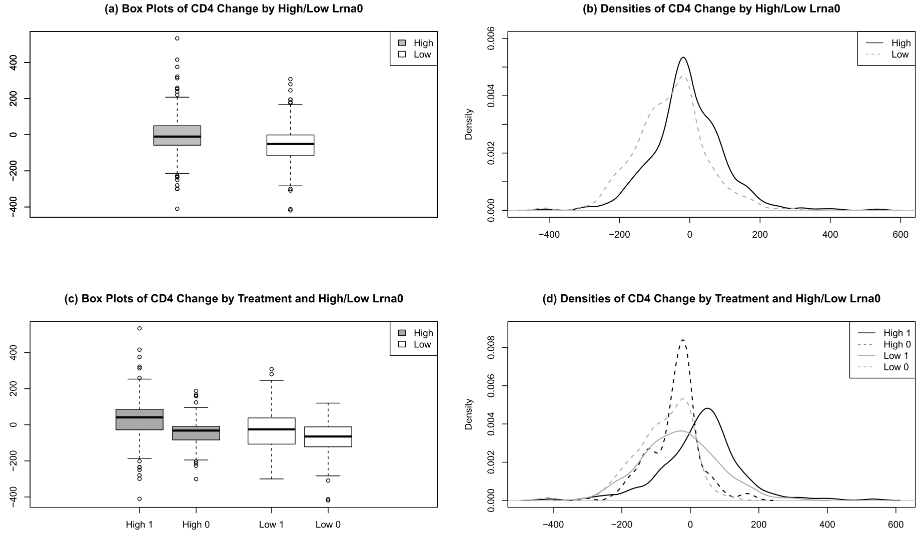


Fig 6. *Prognostic or predictive effects of the logarithm of baseline HIV-1 RNA concentration with base 10 on the CD4 count change at week 24 for ACTG320 study with the prognostic effect shown in (a) and (b), and the predictive effect shown in (c) and (d).*

treatment effects for high and low groups of patients are different. The treatment will show more influence in the low Lcd40 group and the high Lrna0 group. These results validate that the prognostic and predictive models are reasonably selected.

C.3. Sensitivity analysis of hyperparameters

In this subsection, we assess the sensitivity of our method to the choice of the hyperparameters on the NSW dataset in Section 5.1. We vary key hyperparameters, including slab variances

Table 11

Posterior inclusion probabilities for prognostic and predictive covariates under different values of hyperparameters: $(\tau_{\beta 1}, \tau_{\gamma 1})$, $(\tau_{\beta 0}, \tau_{\gamma 0})$, and (q_β, q_γ). When varying one group of hyperparameters, the others are fixed at the values specified in (5).

(a)	$\tau_{\beta 1}$	$\tau_{\gamma 1}$		Age	Educ	Black	Hisp	Marr	RE75	Unemploy
	0.8	0.8	Prog	0.015	0.014	0.722	0.081	0.036	1	0.031
			Pred	0.274	0.525	0.258	0.190	0.187	0.149	0.154
	1	**1**	Prog	0.010	0.010	0.656	0.075	0.025	1	0.025
			Pred	0.245	0.715	0.189	0.142	0.180	0.149	0.114
	2	2	Prog	0.006	0.005	0.463	0.046	0.015	1	0.008
			Pred	0.188	0.733	0.074	0.147	0.097	0.096	0.112
(b)	$\tau_{\beta 0}$	$\tau_{\gamma 0}$		Age	Educ	Black	Hisp	Marr	RE75	Unemploy
	0.007	**0.007**	Prog	0.010	0.010	0.656	0.075	0.025	1	0.025
			Pred	0.245	0.715	0.189	0.142	0.180	0.149	0.114
	0.01	0.01	Prog	0.012	0.012	0.682	0.065	0.026	1	0.024
			Pred	0.212	0.592	0.217	0.211	0.155	0.205	0.150
	0.02	0.02	Prog	0.009	0.009	0.586	0.055	0.017	1	0.018
			Pred	0.153	0.477	0.120	0.106	0.087	0.107	0.092
(c)	q_β	q_γ		Age	Educ	Black	Hisp	Marr	RE75	Unemploy
	0.15	0.15	Prog	0.008	0.008	0.590	0.055	0.021	1	0.015
			Pred	0.170	0.524	0.098	0.122	0.099	0.111	0.097
	0.20	**0.20**	Prog	0.010	0.010	0.656	0.075	0.025	1	0.025
			Pred	0.245	0.715	0.189	0.142	0.180	0.149	0.114
	0.25	0.25	Prog	0.014	0.014	0.758	0.085	0.038	1	0.031
			Pred	0.315	0.656	0.235	0.214	0.197	0.200	0.192

$(\tau_{\beta 1}, \tau_{\gamma 1})$, spike variances $(\tau_{\beta 0}, \tau_{\gamma 0})$, and prior inclusion probabilities (q_β, q_γ), while keeping the remaining hyperparameters fixed at the values specified in (5) to ensure a controlled comparison. For each setting, we run five independent Gibbs sampling chains and report the posterior inclusion probabilities of both prognostic and predictive covariates, averaged across these runs.

For $(\tau_{\beta 1}, \tau_{\gamma 1})$, the results are presented in Table 11(a). For prognostic variable selection, different choices of $(\tau_{\beta 1}, \tau_{\gamma 1})$ do not affect the selection of RE75, which is consistently identified with a posterior inclusion probability of 1. While the posterior inclusion probabilities of other variables vary across different hyperparameter values, Black remains the second highest probability with a substantial gap from the remaining covariates, indicating a stable variable ranking despite changes in hyperparameter settings. For predictive variable selection, Educ consistently has the highest inclusion probability across all settings. While the absolute posterior inclusion probabilities of all covariates vary with different choices of $(\tau_{\beta 1}, \tau_{\gamma 1})$, the relative rankings of variable importance remain largely stable, further demonstrating the robustness of our method.

The results for $(\tau_{\beta 0}, \tau_{\gamma 0})$ and (q_β, q_γ) are presented in Tables 11(b) and 11(c), respectively. We observe similar patterns in these analyses, further supporting the conclusion that our method is stable with respect to hyperparameter choices.

Table 12

Variable selection results in structured logistic-normal mixture settings with $\mathbf{n} = \mathbf{200}$ *and* $\rho = 0$. *(a) Results under the same hyperparameter strategies on* $\boldsymbol{\beta}$ *and* $\boldsymbol{\gamma}$*, where spike hyperparameters* $(\tau_{\beta 0}, \tau_{\gamma 0})$ *or slab hyperparameters* $(\tau_{\beta 1}, \tau_{\gamma 1})$ *are modified with multipliers* 0.5 *or* 2. *(b) Results under the separate hyperparameter strategies on* $\boldsymbol{\beta}$ *and* $\boldsymbol{\gamma}$*, where slab hyperparameters* $\tau_{\beta 1}$ *or* $\tau_{\gamma 1}$ *are modified with multipliers* 0.5 *or* 2.

(a)	Hyper		I^{β}						I^{γ}					
p	Spike	Slab	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
500	×1	×1	4	4	0	1	1	1	3.68	3.74	0.34	0.55	0.69	0.74
	×2	×1	4	4	0	1	1	1	3.69	3.72	0.31	0.55	0.70	0.73
	×0.5	×1	4	4	0	1	1	1	3.65	3.64	0.39	0.54	0.68	0.67
	×1	×2	4	4	0	1	1	1	3.70	3.61	1.11	0.31	0.72	0.65
	×1	×0.5	4	4	0	1	1	1	3.52	3.68	0.10	0.51	0.56	0.68

(b)	Slab		I^{β}						I^{γ}					
p	$\tau_{\beta 1}$	$\tau_{\gamma 1}$	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
500	×1	×1	4	4	0	1	1	1	3.68	3.74	0.34	0.55	0.69	0.74
	×0.5	×1	4	4	0	1	1	1	3.69	3.71	0.36	0.54	0.71	0.73
	×1	×0.5	4	4	0	1	1	1	3.54	3.68	0.13	0.53	0.57	0.68
	×0.5	×0.5	4	4	0	1	1	1	3.52	3.68	0.10	0.51	0.56	0.68
	×2	×1	4	4	0	1	1	1	3.72	3.76	0.30	0.61	0.73	0.76
	×1	×2	4	4	0	1	1	1	3.73	3.67	0.91	0.36	0.75	0.69
	×2	×2	4	4	0	1	1	1	3.70	3.61	1.11	0.31	0.72	0.65

Appendix D: Sensitivity analysis

D.1. Sensitivity analysis of spike-and-slab variances

As noted in [19], continuous spike-and-slab priors are sensitive to the choice of variances, making prior calibration an essential consideration in Bayesian variable selection. To assess their sensitivity, we first conduct an analysis where we simultaneously adjust the variance hyperparameters in both the linear and logistic components. Specifically, we scale the spike or slab variances by multipliers of 0.5 or 2, and examine the impact on variable selection. The results under the same settings in Section 4.1 with $n = 200$ from 100 random trials are summarized in Table 12(a). We present the results with $\rho = 0$, while the results with $\rho = 0.25$ show similar patterns.

We have the following three observations. First, for the linear part, the results on $\boldsymbol{\beta}$ are perfect across all settings, indicating that the linear part is not sensitive to the choice of $(\tau_{\beta 1}, \tau_{\gamma 1})$ and $(\tau_{\beta 0}, \tau_{\gamma 0})$ within a reasonable range. Second, for the logistic part, the variation in spike variances $(\tau_{\beta 0}, \tau_{\gamma 0})$ does not significantly affect the variable selection results for predictive covariates. Third, for the logistic part, the choice of slab variance $(\tau_{\beta 1}, \tau_{\gamma 1})$ exhibits a trade-off between true positives and false positives. Smaller slab variances lead to a more conservative selection on $\boldsymbol{\gamma}$, while larger values may increase the false positives.

Based on these observations, we find that when applying the same hyperparameter settings for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, the overall variable selection performance remains stable within reasonable variations. However, we also observe that the selection on $\boldsymbol{\gamma}$ is more sensitive to the scales of slab variances $(\tau_{\beta 1}, \tau_{\gamma 1})$ than that on $\boldsymbol{\beta}$. This suggests that the linear and logistic components may operate on different scales of $(\tau_{\beta 1}, \tau_{\gamma 1})$ and could benefit from separate calibration

Table 13

Variable selection results in structured logistic-normal mixture settings with $\mathbf{n} = \mathbf{200}$ *and* $\rho = 0$. *Prior inclusion probabilities* q_β *and* q_γ *are modified with multipliers* 0.5 *or* 2.

p	Prior		I^β						I^γ					
	q_β	q_γ	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
100	×1	×1	4	4	0	1	1	1	3.89	3.89	0.36	0.63	0.89	0.89
	×0.5	×1	4	4	0	1	1	1	3.89	3.88	0.33	0.63	0.89	0.88
	×1	×0.5	4	4	0	1	1	1	3.82	3.87	0.05	0.81	0.82	0.87
	×0.5	×0.5	4	4	0	1	1	1	3.81	3.89	0.07	0.79	0.82	0.89
	×2	×1	4	4	0	1	1	1	3.89	3.89	0.32	0.66	0.89	0.89
	×1	×2	4	4	0	1	1	1	3.99	3.88	2.34	0.09	0.99	0.88
	×2	×2	4	4	0	1	1	1	4.00	3.86	2.42	0.09	1.00	0.86
500	×1	×1	4	4	0	1	1	1	3.68	3.74	0.34	0.55	0.69	0.74
	×0.5	×1	4	4	0	1	1	1	3.68	3.70	0.27	0.55	0.69	0.70
	×1	×0.5	4	4	0	1	1	1	3.64	3.73	0.18	0.58	0.68	0.73
	×0.5	×0.5	4	4	0	1	1	1	3.60	3.70	0.20	0.56	0.65	0.71
	×2	×1	4	4	0	1	1	1	3.70	3.69	0.33	0.57	0.70	0.69
	×1	×2	4	4	0	1	1	1	3.69	3.62	0.63	0.43	0.70	0.64
	×2	×2	4	4	0	1	1	1	3.74	3.68	0.64	0.40	0.76	0.69

strategies. Therefore, we further investigate the individual impact of $\tau_{\beta 1}$ and $\tau_{\gamma 1}$ by modifying them separately, and the corresponding results are presented in Table 12(b).

From Table 12(b), we observe that the results are not sensitive to the change of $\tau_{\beta 1}$, while the trade-off between true positives and false positives is primarily influenced by the choice of $\tau_{\gamma 1}$. This observation underscores the necessity of distinct prior calibration for the linear and logistic components in practical applications. Proper calibration of $\tau_{\gamma 1}$ is particularly crucial for guaranteeing robust variable selection on predictive covariates, especially in challenging scenarios involving high dimensionality and potential model misspecification.

D.2. Sensitivity analysis of prior inclusion probabilities

In this subsection, we examine the choice of prior inclusion probabilities q_β and q_γ by applying a similar sensitivity analysis under the same settings in Section 4.1 with $n = 200$. Specifically, we scale each of these probabilities by multipliers of 0.5 or 2 separately to assess their impact on variable selection performance. The results from 100 random trials are summarized in Table 13.

Based on the results in Table 13, we observe the following key findings regarding the sensitivity of prior inclusion probabilities q_β and q_γ. First, the results for $\boldsymbol{\beta}$ remain highly stable across different values of q_β. Second, the results for $\boldsymbol{\gamma}$ exhibit an expected trade-off between true positives and false positives. A larger q_γ increases the number of selected variables, leading to higher true positives and false positives, particularly when $p = 500$. This highlights the need for calibration of q_γ in predictive variable selection. Third, despite variations in q_γ, TP$_s$ and $I_s = I_0$ for I^γ remain stable, suggesting that the ranking ability of the model remains largely unaffected within a reasonable range of hyperparameter choices.

We conclude from these results that, while theoretically the same order of q_β and q_γ guarantees model selection and parameter estimation consistency, empirically they can be

Table 14
Sensitivity to σ_α^2, a_0, and b_0.

(a) $a_0 = 2$ and $b_0 = 1$							
		TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
$\sigma_\alpha^2 = 1$	I^β	4	4	0	1	1	1
	I^γ	3.86	3.87	0.25	0.7	0.87	0.87
$\sigma_\alpha^2 = 2$	I^β	4	4	0	1	1	1
	I^γ	3.86	3.86	0.29	0.67	0.87	0.86
$\sigma_\alpha^2 = 3$	I^β	4	4	0	1	1	1
	I^γ	3.87	3.86	0.32	0.66	0.88	0.86
(b) $\sigma_\alpha^2 = 1$ and $b_0 = 1$							
		TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
$a_0 = 1$	I^β	4	4	0	1	1	1
	I^γ	3.87	3.87	0.3	0.66	0.88	0.87
$a_0 = 2$	I^β	4	4	0	1	1	1
	I^γ	3.84	3.87	0.3	0.65	0.85	0.87
$a_0 = 3$	I^β	4	4	0	1	1	1
	I^γ	3.86	3.87	0.3	0.64	0.87	0.87
(c) $\sigma_\alpha^2 = 1$ and $a_0 = 2$							
		TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
$b_0 = 1$	I^β	4	4	0	1	1	1
	I^γ	3.84	3.87	0.3	0.65	0.85	0.87
$b_0 = 2$	I^β	4	4	0	1	1	1
	I^γ	3.85	3.87	0.3	0.67	0.86	0.87
$b_0 = 3$	I^β	4	4	0	1	1	1
	I^γ	3.85	3.85	0.29	0.66	0.86	0.86

calibrated differently to achieve improved finite-sample performance. Specifically, while q_β remains robust across different values, q_γ requires careful tuning to prevent excessive false positives.

D.3. Sensitivity analysis of other hyperparameters

In this section, we explore the sensitivity of our methods to other hyperparameters in the assigned weak informative prior distributions. To be specific, we investigate the influence of σ_α^2 in the priors of α_1 and α_2 as well as a_0 and b_0 in the prior of σ_y. We consider the setting with $p = 100$ and $\rho = 0$ in Section 4.1 with $n = 200$.

For σ_α^2, we fix $a_0 = 2$ and $b_0 = 1$ and consider a range of $\sigma_\alpha^2 \in \{1, 2, 3\}$. The variable selection results are presented in Table 14(a). All measures do not vary much, especially for variable importance ranking, indicating that BVSA is not sensitive to the choice of σ_α^2.

Similarly, we consider $a_0 \in \{1, 2, 3\}$ with fixed $\sigma_\alpha^2 = 1$ and $b_0 = 1$ and $b_0 \in \{1, 2, 3\}$ with fixed $\sigma_\alpha^2 = 1$ and $a_0 = 2$. The results are reported in Table 14(b) and Table 14(c), respectively, indicating our method is robust to the choices of the hyperparameters.

D.4. Sensitivity analysis of initialization

In this subsection, we examine the sensitivity of our method to initialization. In our approach, the Gibbs sampler is randomly initialized by drawing samples from the prior distributions. To

Table 15

Variable selection results in structured logistic-normal mixture settings with $n = 200$ and $\rho = 0$. "Prior": initialization by random sampling according prior distributions; "EM-r": initialization via random active variable selection and EM algorithm; "EM-s": initialization via active variable selection by lasso and GUIDE [27] and EM algorithm.

p	Init	I^β						I^γ					
		TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
100	Prior	4	4	0	1	1	1	3.89	3.89	0.36	0.63	0.89	0.89
	EM-r	4	4	0	1	1	1	3.88	3.90	0.33	0.64	0.88	0.90
	EM-s	4	4	0	1	1	1	3.87	3.90	0.34	0.62	0.87	0.90
500	Prior	4	4	0	1	1	1	3.68	3.74	0.34	0.55	0.69	0.74
	EM-r	4	4	0	1	1	1	3.67	3.67	0.35	0.50	0.69	0.68
	EM-s	4	4	0	1	1	1	3.67	3.70	0.33	0.54	0.68	0.70

further assess the impact of initialization, we consider two alternative EM-based initialization strategies:

- **EM-r**: We randomly select active prognostic and predictive covariates, with the size determined as $\min(30, 0.2p_Z)$ and $\min(30, 0.2p_X)$, respectively. Using this subset of covariates, we estimate the initial model parameters via the EM algorithm on this selected low-dimensional model.
- **EM-s**: We apply a tree-based method, GUIDE [27], to select predictive covariates, and use LASSO to select prognostic covariates. The EM algorithm is then used to estimate initial model parameters based on the selected variables.

We compare the performance of these three initialization strategies under the same settings as in Section 4.1 with $n = 200$. Specifically, we evaluate their impact on variable selection and convergence behavior of the Gibbs sampler.

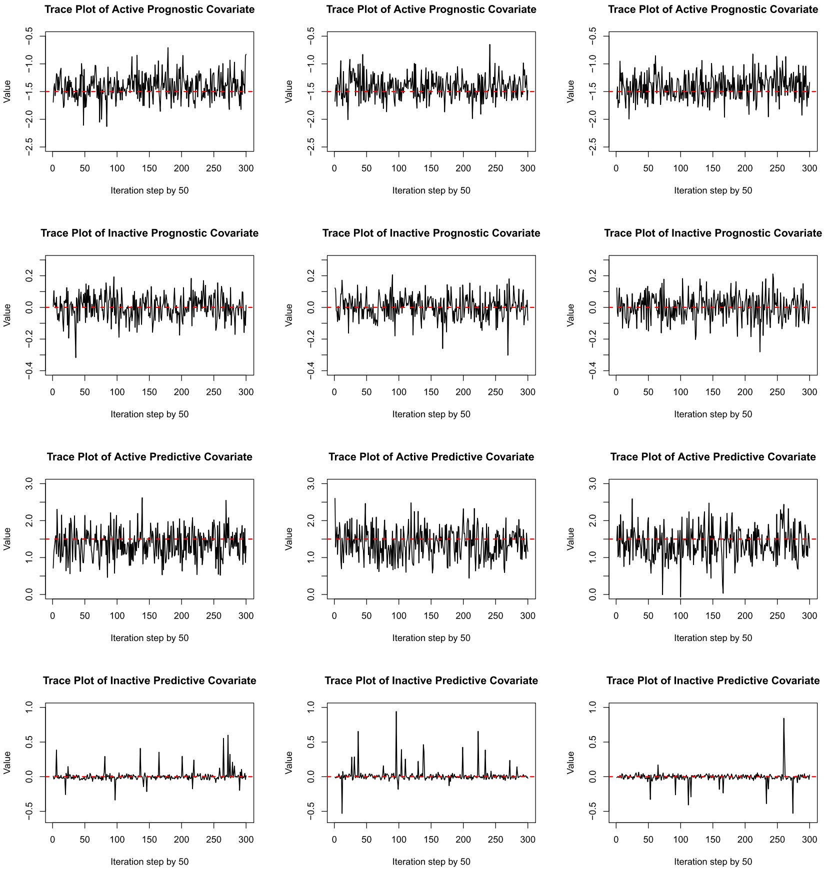
For variable selection, as indicated in Table 15, the choice of initialization has a small impact on the selection performance.

We investigate the impact on mixing of the sampler using trace plots and effective sample size (ESS). We take one representative trial under $p = 100$ as an example. We present the trace plots using every 50th iteration from a total of 15000 samples after the burn-in period for different initialization methods (Prior, EM-r, EM-s) across different variable categories in Figure 7. The trace plots exhibit a high degree of consistency across all settings, indicating that our approach achieves stable mixing regardless of the initialization strategy.

We further present the averaged ESS across all prognostic and predictive covariates for different initialization methods in Table 16. The results indicate that the prognostic covariates consistently achieve high ESS values across all initialization methods, suggesting stable and efficient sampling. For the predictive covariates, while the ESS values are lower due to the challenge from subgroup modeling, they remain comparable across different initializations. These results reinforce the robustness of our approach to initialization choices.

D.5. Sensitivity analysis of model misspecification

In this subsection, we investigate the robustness of our method to two types of model misspecification: (1) misspecification in the subgroup membership structure; (2) misspecification in the noise distribution of the linear model.



(a) Prior (b) EM-r (c) EM-s

Fig 7. *Trace plots for different initialization methods (Prior, EM-r, EM-s) across different variable categories.*

Table 16

Averaged effective sample sizes with standard errors across all prognostic and predictive covariates for different initialization methods (Prior, EM-r, EM-s) with $n = 200$, $p = 100$, $\rho = 0$, *and post-burnin length* 15000.

Init	Prognostic	Predictive
Prior	14016 (1752.04)	1412.7 (606.62)
EM-r	13831 (1682.09)	1392.8 (561.98)
EM-s	13893 (1602.60)	1350.1 (685.41)

For the first type of misspecification, we have conducted extensive experiments to evaluate its impact under traditional rule-based subgroup settings in Section 4.2, where our method exhibits strong robustness. We further consider the scenario where the subgroup membership structure follows a probit link function and summarize the results in Table 17.

The results demonstrate that prognostic variable selection remains highly stable, achieving perfect recovery of the true set under the probit link case. Moreover, predictive variable selection is also robust to link function misspecification, with only minimal variations. These findings indicate that our method effectively adapts to different link functions, demonstrating strong robustness to subgroup membership structure misspecification.

For the second type of misspecification, we consider scenarios where the assumed normal

Table 17

Variable selection results in structured logistic-normal mixture settings with $n = 200$ *and* $\rho = 0$. *Probit link function is considered as an example of subgroup membership structure misspecification.*

p	Link	I^β						I^γ					
		TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
100	Logit	4	4	0	1	1	1	3.89	3.89	0.36	0.63	0.89	0.89
	Probit	4	4	0	1	1	1	3.88	3.92	0.26	0.70	0.88	0.92
500	Logit	4	4	0	1	1	1	3.68	3.74	0.34	0.55	0.69	0.74
	Probit	4	4	0	1	1	1	3.68	3.73	0.25	0.63	0.68	0.73

Table 18

Variable selection results in structured logistic-normal mixture settings with $n = 200$ *and* $\rho = 0$. *Different noise distributions are considered, including the standard normal distribution and Student's t distribution with degrees of freedom of 2 or 3.*

p	Noise	I^β						I^γ					
		TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
100	$N(0,1)$	4	4	0	1	1	1	3.89	3.89	0.36	0.63	0.89	0.89
	$t(3)$	4	4	0	1	1	1	3.88	3.89	0.36	0.61	0.88	0.89
	$t(2)$	4	4	0	1	1	1	3.88	3.88	0.36	0.61	0.88	0.88
500	$N(0,1)$	4	4	0	1	1	1	3.68	3.74	0.34	0.55	0.69	0.74
	$t(3)$	4	4	0	1	1	1	3.66	3.71	0.30	0.56	0.68	0.72
	$t(2)$	4	4	0	1	1	1	3.73	3.74	0.29	0.57	0.73	0.74

distribution of the noise term in the response model differs from the true data-generating process. Specifically, we allow the true noises to follow a heavy-tailed distribution, including Student's t-distributions with degrees of freedom 2 or 3. To evaluate the impact of such misspecification, we conducted experiments under the same settings as Section 4.1 with $n = 200$. The results are summarized in Table 18.

We observe that prognostic variable selection is highly stable across different noise distributions. The prognostic selection performance remains perfect even in the $t(2)$ noise case, where the noise distribution has heavier tails. Similarly, predictive variable selection is insensitive to the noise distribution misspecification. These results confirm the robustness of our method under deviations from the normality assumption.

Appendix E: Critical evaluation

In our simulation studies in Section 4.2, we primarily compare BVSA with tree-based methods, such as GUIDE [27] and MOB [39]. These methods are particularly well-suited for traditional rule-based subgroup settings, where the subgroup membership follows a predefined splitting structure. Unlike BVSA, they directly model subgroup boundaries using recursive partitioning, rather than relying on a global parametric model.

Despite our results in Section 4.2 demonstrating that BVSA remains robust in such misspecified settings, we did not examine the impact of high correlation among predictive covariates in rule-based settings. Since all covariates enter the model simultaneously, BVSA is sensitive to high correlation. Additionally, high correlation can create difficulties with convergence. As a result, BVSA may select redundant or irrelevant variables, leading to decreased true positives and increased false positives. In contrast, tree-based methods divide the data into

Table 19
Predictive variable selection results in rule-based subgroup settings with $n = 200$ and various pairwise covariate correlations.

p	ρ	Method	TP	TP$_s$	FP	$I = I_0$	$I \supset I_0$	$I_s = I_0$
20	0.5	BVSA	1.98	2.53	0.09	0.19	0.20	0.54
		MOB	2.69	2.69	0.07	0.67	0.69	0.67
		PRIM	2.71	2.58	0.62	0.51	0.74	0.52
		SeqBT	2.49	2.49	0.02	0.54	0.56	0.54
		GUIDE	2.67	2.67	0.33	0.48	0.67	0.46
	0.6	BVSA	1.76	2.41	0.10	0.10	0.12	0.43
		MOB	2.60	2.60	0.09	0.58	0.60	0.58
		PRIM	2.71	2.52	0.94	0.37	0.75	0.38
		SeqBT	2.41	2.41	0.04	0.46	0.50	0.46
		GUIDE	2.57	2.56	0.35	0.44	0.57	0.40
	0.7	BVSA	1.42	2.14	0.08	0.03	0.03	0.26
		MOB	2.39	2.39	0.11	0.37	0.39	0.37
		PRIM	2.55	2.35	0.96	0.27	0.59	0.27
		SeqBT	2.14	2.14	0.05	0.26	0.28	0.26
		GUIDE	2.40	2.44	0.46	0.23	0.41	0.17
200	0.5	BVSA	1.77	1.99	1.15	0.07	0.16	0.18
		MOB	2.52	2.52	0.04	0.52	0.52	0.52
		PRIM	1.82	1.77	0.92	0.12	0.15	0.12
		SeqBT	2.55	2.55	0.06	0.56	0.60	0.56
		GUIDE	2.40	2.37	0.31	0.31	0.41	0.30
	0.6	BVSA	1.58	1.69	1.17	0.02	0.03	0.03
		MOB	2.44	2.44	0.10	0.44	0.44	0.44
		PRIM	1.70	1.55	1.84	0.08	0.11	0.08
		SeqBT	2.36	2.36	0.12	0.37	0.41	0.37
		GUIDE	2.24	2.21	0.49	0.17	0.25	0.17
	0.7	BVSA	1.17	1.56	0.95	0	0.03	0.01
		MOB	2.13	2.13	0.22	0.14	0.14	0.14
		PRIM	1.55	1.35	2.24	0	0	0
		SeqBT	2.21	2.21	0.13	0.24	0.30	0.24
		GUIDE	2.07	2.06	0.63	0.06	0.08	0.05

subgroups hierarchically based on individual variable thresholds. Even if two variables are highly correlated, a tree-based approach will typically only use one of them in a given split. This structure makes tree-based methods less sensitive to correlation, while BVSA's global logistic model is more susceptible to the selection of highly correlated inactive variables.

To assess the limitations of BVSA, we conduct simulations under the rule-based subgroup settings with high pairwise correlations among predictive covariates. Specifically, we generate data based on the following model:

$$Y = 1 - 1.5Z_1 + 2Z_2 - 2.5Z_3 + 3Z_4 + 40t \times I(X_1 \geq -1, X_3 < 1, X_5 < 0.5) + \epsilon,$$

where both Z and X follow multivariate normal distributions with varying pairwise correlation levels $\rho \in \{0.5, 0.6, 0.7\}$. We consider $p = 2p_1 = 2p_2$ with $p \in \{20, 200\}$ and compare BVSA with tree-based methods, including MOB [39], PRIM [8], SeqBT [17], and GUIDE [27]. The results from 100 random trials are summarized in Table 19.

When $p = 20$, our method exhibits a conservative behavior in variable selection, with both low TP and low FP, whereas tree-based methods identify most of the active covariates. As ρ increases, the performance of BVSA deteriorates, reflected in a decline in TP and reduced exact recovery of the true predictive set. Despite these limitations, BVSA maintains TP_s values that are comparable to those of tree-based methods, demonstrating its robustness in ranking variable importance even under high correlation. When $p = 200$, BVSA's performance further deteriorates in the misspecified setting under high dimensionality and strong correlation, highlighting the challenge of identifying true predictive covariates when all variables are jointly handled in the regression model. Our method becomes more susceptible to selecting redundant covariates, leading to an increase in FP. In contrast, most tree-based methods demonstrate greater robustness. Notably, MOB and SeqBT maintain stable TP values, showcasing their ability to select relevant covariates despite strong correlation. These findings validate the key limitation of BVSA in rule-based settings with highly correlated predictive covariates.

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor, and the Editor for their constructive comments that improved the quality of this paper.

Funding

This work was partially supported by the National Nature and Science Foundation of China (11871165, 12331009), and the National Science Foundation of US (CAREER-1943500).

References

[1] Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association* **88** 669–679. MR1224394

[2] Berkhof, J., van Mechelen, I. and Gelman, A. (2003). A Bayesian approach to the selection and testing of mixture models. *Statistica Sinica* **13** 423–442. MR1977735

[3] Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons, New York. MR1274699

[4] Bhattacharya, A., Pati, D., Pillai, N. S. and Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association* **110** 1479–1490. MR3449048

[5] Bondell, H. D. and Reich, B. J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association* **107** 1610–1624. MR3036420

[6] Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549

[7] Cai, T., Tian, L., Wong, P. H. and Wei, L. J. (2010). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* **12** 270–282.

[8] Chen, G., Zhong, H., Belousov, A. and Devanarayan, V. (2015). A PRIM approach to predictive-signature development for patient stratification. *Statistics in Medicine* **34** 317–342. MR3293151

[9] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581

[10] Finocchio, G. and Schmidt-Hieber, J. (2023). Posterior contraction for deep Gaussian process priors. *Journal of Machine Learning Research* **24** 1–49. MR4582488

[11] Gelman, A., Hwang, J. and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing* **24** 997–1016. MR3253850

[12] George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88** 881–889.

[13] Ghosh, J., Herring, A. H. and Siega-Riz, A. M. (2011). Bayesian variable selection for latent class models. *Biometrics* **67** 917–925. MR2829266

[14] Guo, X. and He, X. (2021). Inference on selected subgroups in clinical trials. *Journal of the American Statistical Association* **116** 1498–1506. MR4309288

[15] Guo, X., Wei, W., Liu, M., Cai, T., Wu, C. and Wang, J. (2023). Assessing the most vulnerable subgroup to type II diabetes associated with statin usage: Evidence from electronic health record data. *Journal of the American Statistical Association* **118** 1488–1499. MR4646578

[16] Hammer, S. M., Squires, K. E., Hughes, M. D., Grimes, J. M., Demeter, L. M., Currier, J. S., Eron, J. J., Feinberg, J. E., Balfour, H. H., Deyton, L. R., Chodakewitz, J. A., Fischl, M. A., Phair, J. P., Pedneault, L., Nguyen, B.-Y. and Cook, J. C. (1997). A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine* **337** 725–733.

[17] Huang, X., Sun, Y., Trow, P., Chatterjee, S., Chakravartty, A., Tian, L. and Devanarayan, V. (2017). Patient subgroup identification for clinical drug development. *Statistics in Medicine* **36** 1414–1428. MR3631969

[18] Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* **7** 443–470. MR3086426

[19] Iqbal, A., Ogundimu, E. O. and Rubio, F. J. (2025). Bayesian variable selection in sample selection models using spike-and-slab priors.

[20] Italiano, A. (2011). Prognostic or predictive? It's time to get back to definitions! *Journal of Clinical Oncology* **29** 4718–4718.

[21] Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association* **107** 649–660. MR2980074

[22] Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association* **102** 1025–1038. MR2411662

[23] Lee, K. and Cao, X. (2021). Bayesian group selection in logistic regression with application to MRI data analysis. *Biometrics* **77** 391–400. MR4307642

[24] Li, J., Li, Y., Jin, B. and Kosorok, M. R. (2021). Multi-threshold change plane model: estimation theory and applications in subgroup identification. *Statistics in Medicine* **40** 3440–3459. MR4269063

[25] LIANG, F., SONG, Q. and YU, K. (2013). Bayesian subset modeling for high-dimensional generalized linear models. *Journal of the American Statistical Association* **108** 589–606. MR3174644

[26] LIU, Y., MA, X., ZHANG, D., GENG, L., WANG, X., ZHENG, W. and CHEN, M.-H. (2019). Look before you leap: Systematic evaluation of tree-based statistical methods in subgroup identification. *Journal of Biopharmaceutical Statistics* **29** 1082–1102.

[27] LOH, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica* **12** 361–386. MR1902715

[28] LOH, W.-Y., CAO, L. and ZHOU, P. (2019). Subgroup identification for precision medicine: A comparative review of 13 methods. *WIREs Data Mining and Knowledge Discovery* **9** e1326.

[29] LU, Z. and LOU, W. (2023). Bayesian approaches to variable selection in mixture models with application to disease clustering. *Journal of Applied Statistics* **50** 387–407. MR4536600

[30] MA, J., STINGO, F. C. and HOBBS, B. P. (2019). Bayesian personalized treatment selection strategies that integrate predictive with prognostic determinants. *Biometrical Journal* **61** 902–917. MR3982424

[31] NARISETTY, N. N. and HE, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics* **42** 789–817. MR3210987

[32] NARISETTY, N. N., SHEN, J. and HE, X. (2019). Skinny Gibbs: A consistent and scalable Gibbs sampler for model selection. *Journal of the American Statistical Association* **114** 1205–1217. MR4011773

[33] PEDONE, M., ARGIENTO, R. and STINGO, F. C. (2024). Personalized treatment selection via product partition models with covariates. *Biometrics* **80** ujad003. MR4867257

[34] POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association* **108** 1339–1349. MR3174712

[35] RAY, K. and SZABÓ, B. (2022). Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association* **117** 1270–1281. MR4480711

[36] RAY, K., SZABÓ, B. and CLARA, G. (2020). Spike and slab variational Bayes for high dimensional logistic regression. In *Advances in Neural Information Processing Systems* (H. LAROCHELLE, M. RANZATO, R. HADSELL, M. F. BALCAN and H. LIN, eds.) **33** 14423–14434. Curran Associates, Inc.

[37] ROSSELL, D. and RUBIO, F. J. (2023). Additive Bayesian variable selection under censoring and misspecification. *Statistical Science* **38** 13–29. MR4534642

[38] ROČKOVÁ, V. and GEORGE, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association* **109** 828–846. MR3223753

[39] SEIBOLD, H., ZEILEIS, A. and HOTHORN, T. (2016). Model-based recursive partitioning for subgroup analyses. *The International Journal of Biostatistics* **12** 45–63. MR3505686

[40] SHEN, J. and HE, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *Journal of the American Statistical Association* **110** 303–312. MR3338504

[41] SHEN, J. and QU, A. (2019). Subgroup analysis based on structured mixed-effects models for longitudinal data. *Journal of Biopharmaceutical Statistics* **30** 607–622.

[42] Städler, N., Bühlmann, P. and Van De Geer, S. (2010). $\ell 1$-penalization for mixture regression models. *Test* **19** 209–256. MR2677722

[43] Tadesse, M. G. and Vannucci, M. (2021). *Handbook of Bayesian variable selection.* CRC Press.

[44] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **58** 267–288. MR1379242

[45] Wang, B., Li, J. and Wang, X. (2022). Multi-threshold proportional hazards model and subgroup identification. *Statistics in Medicine* **41** 5715–5737. MR4515038

[46] Wang, J., Cai, X. and Li, R. (2021). Variable selection for partially linear models via Bayesian subset modeling with diffusing prior. *Journal of Multivariate Analysis* **183** 104733. MR4222385

[47] Wang, K. and Ghosal, S. (2023). Posterior contraction and testing for multivariate isotonic regression. *Electronic Journal of Statistics* **17** 798–822. MR4554662

[48] Wang, Y. (2016). Logistic-normal mixtures with heterogeneous components and high dimensional covariates., PhD thesis, University of Michigan. MR3641123

[49] West, M. (1993). Mixture models, Monte Carlo, Bayesian updating and dynamic models. *Computing Science and Statistics* **24** 325–333.

[50] Yang, Y., Wainwright, M. J. and Jordan, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics* **44** 2497–2532. MR3576552

[51] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894–942. MR2604701

[52] Zhao, L., Tian, L., Cai, T., Claggett, B. and Wei, L. J. (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association* **108** 527–539. MR3174639

[53] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101** 1418–1429. MR2279469