Statistica Sinica Preprint No: SS-2025-0226	
Title	Variational Bayes for High-dimensional Structured Mixture Model
Manuscript ID	SS-2025-0226
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0226
Complete List of Authors	Ruqian Zhang and Juan Shen
Corresponding Authors	Juan Shen
E-mails	shenjuan@fudan.edu.cn
Notice: Accepted author version.	

Statistica Sinica

VARIATIONAL BAYES FOR HIGH-DIMENSIONAL STRUCTURED MIXTURE MODEL

Ruqian Zhang and Juan Shen

Fudan University

Abstract: Bayesian methods are widely employed for variable selection; however, the computational complexity associated with Markov Chain Monte Carlo (MCMC) techniques often limits their scalability in high-dimensional contexts. The computation becomes more challenging in mixture models with a substantial number of latent variables. We propose a variational Bayesian (VB) approach for high-dimensional structured mixture models to identify important variables for subgroup analysis. Our method enables efficient and simultaneous variable selection and parameter estimation by approximating the posterior distribution. We establish model selection consistency and derive contraction rates for estimation errors, advancing existing VB theoretical results. Additionally, a coordinate ascent variational inference algorithm with data augmentation is developed. Numerical studies illustrate that our method achieves accuracy comparable to MCMC while significantly improving computational efficiency. The effectiveness of our method is validated through real-world applications.

Key words and phrases: Model selection consistency, spike-and-slab prior, variational Bayes.

1. Introduction

Mixture models are widely used to capture heterogeneous subgroups within a population, as seen in applications such as precision medicine (van der Vliet et al., 2020) and recommendation systems (Van Dat et al., 2022). In such models, responses from distinct subgroups follow different distributions based on mixture proportions (McLachlan et al., 2019). Usually, mixture proportions are modeled as functions of observed variables, as in structured mixture models (Shen and He, 2015; Shen and Qu, 2020). The baseline variables associated with subgroup memberships are referred to as “predictive” variables (Loh, 2002), while those directly influencing the response are termed “prognostic” variables (Italiano, 2011).

When a large number of covariates are available, identifying the active ones is important for improving interpretability and obtaining a parsimonious model. This step becomes crucial in high-dimensional settings, where redundant covariates complicate model estimation and increase uncertainty in subgroup identification (Ghosh et al., 2011). For mixture models, variable selection is often achieved through penalization (Khalili and Chen, 2007; Städler et al., 2010). However, the non-convexity of the penalized objective poses challenges for high-dimensional structured mixture models in both computational implementation and theoretical justification (Wang, 2016). These issues motivate the adoption of Bayesian variable selection (Narisetty and He, 2014). Zhang et al. (2025) uses the spike-and-slab prior to select important covariates in high-dimensional structured mixture models. Bayesian methods avoid the complexities of non-convex optimization by using sampling strategies like Markov chain Monte Carlo (MCMC).

Although Bayesian approaches provide a flexible framework, they often incur high computational costs when MCMC is used, especially in high-dimensional settings. This difficulty is exacerbated in large-sample mixture models, where latent subgroup indicators are introduced. To overcome the computational difficulty, variational Bayes (VB) has emerged as a scalable alternative (Blei et al., 2017). VB approximates the exact posterior by finding the closest distribution within a tractable variational family, measured by the Kullback-Leibler (KL) divergence. This transforms the problem from sampling to optimization, substantially reducing computational burden while maintaining much of the accuracy of MCMC. VB has been widely applied to variable selection in high-dimensional linear regression Carbonetto and Stephens (2012), logistic regression (Zhang et al., 2019), and proportional hazards models (Komodromos et al., 2022). In this paper, we tackle a more challenging scenario of structured mixture models, in which the hierarchical layer of subgroup memberships brings additional theoretical and computational complexity.

Theoretical developments in variational inference have gained increasing attention. In low-dimensional settings, the consistency of VB estimation has been analyzed from a frequentist perspective (Westling and McCormick, 2019), and Bernstein-von Mises type theorems for variational approximation have been established (Wang and Blei, 2019). In high-dimensional and nonparametric settings, general results on variational posterior contraction have been derived through the prior mass and testing approach (Zhang and Gao, 2020; Yang et al., 2020). Recent advances in model selection priors have shown near-optimal contraction rates for parameter estimation in linear regression (Ray and Szabó, 2022). It is also shown that the variational posterior concentrates on models of sizes at most a multiple of the true model size. These results have been extended to logistic models (Ray et al., 2020) and group sparse regression (Ge et al., 2025).

Nevertheless, model selection consistency in high-dimensional settings remains relatively underexplored for variational Bayesian variable selection. In this paper, we develop a VB approach that performs variable selection of prognostic and predictive covariates and estimates model parameters without post hoc analysis. Our work advances previous theoretical results in two aspects. First, we establish model selection consistency under an adjusted beta-min condition, proving VB's ability to identify the true model in high-dimensional scenarios. Second, we show that the VB posterior achieves near-optimal contraction rates for parameter estimation in the challenging mixture model setting, where the concave log-likelihood condition in Atchadé (2017) and Ray et al. (2020) no longer applies.

We develop a scalable coordinate ascent variational inference (CAVI) algorithm to optimize the variational posterior distribution. Data augmentation techniques are used to ensure model conjugacy and efficient computation. Extensive simulations demonstrate that the VB method achieves accuracy comparable to MCMC in both variable selection and parameter estimation, while significantly reducing computational cost.

The remainder of this paper is organized as follows. Section 2 introduces the structured mixture model and the variational Bayesian method. Section 3 establishes the theoretical guarantees for the VB posterior. Section 4 details the computational algorithm and implementation specifics. Section 5 provides extensive simulation studies to evaluate the performance of the proposed method, followed by two real data applications in Section 6. Finally, Section 7 concludes the paper with a discussion.

2. Problem Setup

In this section, we introduce the structured mixture model and specify the prior distributions used for variable selection. We then propose a variational Bayesian method for simultaneous model selection and parameter estimation in high-dimensional structured mixture models.

2.1 Structured mixture model

Let Y be a continuous response variable, $\boldsymbol{z} \in \mathbb{R}^{pz}$ be the prognostic covariates directly influencing Y, and $t \in \{0, 1\}$ be the treatment indicator. We assume the presence of heterogeneous treatment effects across two subgroups. For example, in one subgroup, the treatment effect is negligible, whereas in the other subgroup, the treatment effect is significant. Within each subgroup, Y, conditional on $\boldsymbol{z}$ and t, follows a linear model with Gaussian noise from $N(0, \sigma_y^2)$. The density of Y is given by

$$f(y \mid \boldsymbol{z}, t, \pi_1, \pi_2) = \sum_{k=1}^{2} \frac{\pi_k}{\sqrt{2\pi}\sigma_y} \exp\left\{-\frac{1}{2\sigma_y^2}(y - \boldsymbol{z}^T\boldsymbol{\beta} - t\alpha_k)^2\right\}, \qquad (2.1)$$

where $\boldsymbol{\beta} \in \mathbb{R}^{pz}$ represents the shared prognostic effects, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ denotes the subgroup-specific treatment effects, and π_k denotes the mixture proportions, with $\pi_1 + \pi_2 = 1$ and $0 < \pi_k < 1$ for $k = 1, 2$. We assume that the subgroup identity follows a hierarchical structure determined by predictive covariates $\boldsymbol{x} \in \mathbb{R}^{p_X}$, modeled with logistic regression:

$$\pi_1 = 1 - \pi_2 = \frac{\exp(\boldsymbol{x}^T \boldsymbol{\gamma})}{1 + \exp(\boldsymbol{x}^T \boldsymbol{\gamma})}, \tag{2.2}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^{p_X}$ are the coefficients for subgroup identification. The unknown parameters are $\boldsymbol{\theta} := (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma_y) \in \mathbb{R}^{p+3}$ with $p = p_Z + p_X$.

2.2 Bayesian variable selection

We assume both $\boldsymbol{z}$ and $\boldsymbol{x}$ are high-dimensional with a sparse true model. To identify the true model, we apply spike-and-slab priors on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ to select active covariates with nonzero coefficients(George and McCulloch, 1993).

We introduce binary model indicators $I_j^{\boldsymbol{\beta}}$ for $j = 1, \ldots, p_Z$ where $I_j^{\boldsymbol{\beta}} = 1$ indicates that the jth prognostic covariate is active and $I_j^{\boldsymbol{\beta}} = 0$ otherwise. Similarly, for predictive covariates, we define $I_\ell^{\boldsymbol{\gamma}} \in \{0, 1\}$ for $\ell = 1, \ldots, p_X$. Let the model structure indicator be $I = (I_1^{\boldsymbol{\beta}}, \ldots, I_{p_Z}^{\boldsymbol{\beta}}, I_1^{\boldsymbol{\gamma}}, \ldots, I_{p_X}^{\boldsymbol{\gamma}}) \in \{0, 1\}^p$. Given I, the model selection prior follows a hierarchical form

$$I_j^{\boldsymbol{\beta}} \sim \text{Bern}(q_{\boldsymbol{\beta}n}), \quad \beta_j \mid I_j^{\boldsymbol{\beta}} \sim I_j^{\boldsymbol{\beta}} N(0, \sigma_y^2 \tau_{\boldsymbol{\beta}n}^2) + (1 - I_j^{\boldsymbol{\beta}})\delta_0,$$

$$I_\ell^{\boldsymbol{\gamma}} \sim \text{Bern}(q_{\boldsymbol{\gamma}n}), \quad \gamma_\ell \mid I_\ell^{\boldsymbol{\gamma}} \sim I_\ell^{\boldsymbol{\gamma}} N(0, \tau_{\boldsymbol{\gamma}n}^2) + (1 - I_\ell^{\boldsymbol{\gamma}})\delta_0,$$

where δ_0 denotes the Dirac mass at 0, $\tau_{\boldsymbol{\beta}n}^2$ and $\tau_{\boldsymbol{\gamma}n}^2$ are the variances of slab distributions, and $q_{\boldsymbol{\beta}n}$ and $q_{\boldsymbol{\gamma}n}$ are the prior inclusion probabilities. The subscript n indicates that their choices may depend on the sample size, which is omitted later for conciseness. We set Gaussian priors $N(0, \sigma_y^2\sigma_\alpha^2)$ on α_k and an inverse gamma prior $\mathrm{IG}(a_0, b_0)$ on σ_y^2, with σ_α^2, a_0, and b_0 being the hyperparameters.

Let $\{(y_i, \boldsymbol{z}_i, \boldsymbol{x}_i, t_i)\}_{i=1}^n$ be a sample of n independent observations, where $\boldsymbol{Y} = (y_1, \ldots, y_n)$ denotes the response and $L_n(\boldsymbol{\theta})$ the likelihood function. Given the prior $\pi(\boldsymbol{\theta})$, the joint posterior satisfies $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y}) \propto \pi(\boldsymbol{\theta})L_n(\boldsymbol{\theta})$, and MCMC methods are typically used to estimate $\boldsymbol{\theta}$. However, despite recent improvements in sampling efficiency, MCMC remains computationally expensive for large sample sizes and high-dimensional models. To overcome this scalability limitation, we adopt variational Bayes as an alternative.

2.3 Variational Bayesian approximation

Variational Bayes aims to approximate the exact posterior distribution with a tractable variational distribution. We consider a mean-field variational family, which assumes a factorized structure as

$$\mathcal{Q} = \left\{ q(\boldsymbol{\theta}) = \prod_{j=1}^{p_Z} q(\beta_j) \times \prod_{\ell=1}^{p_X} q(\gamma_\ell) \times \prod_{k=1}^{2} q(\alpha_k) \times q(\sigma_y^2) \right\}, \tag{2.3}$$

where $q(\cdot)$ denotes the variational density for each parameter. Variational distributions of $\{\beta_j\}_{j=1}^{p_Z}$ and $\{\gamma_\ell\}_{\ell=1}^{p_X}$ are assumed in a spike-and-slab form:

$$I_j^{\boldsymbol{\beta}} \sim \mathrm{Bern}(\eta_j^{\boldsymbol{\beta}}), \quad \beta_j \mid I_j^{\boldsymbol{\beta}} \sim I_j^{\boldsymbol{\beta}} N(\mu_j^{\boldsymbol{\beta}}, \sigma_j^{\boldsymbol{\beta}2}) + (1 - I_j^{\boldsymbol{\beta}})\delta_0,$$

$$I_\ell^{\gamma} \sim \mathrm{Bern}(\eta_\ell^{\gamma}), \quad \gamma_\ell \mid I_\ell^{\gamma} \sim I_\ell^{\gamma} N(\mu_\ell^{\gamma}, \sigma_\ell^{\gamma 2}) + (1 - I_\ell^{\gamma})\delta_0.$$

The variational distributions of α_1, α_2, and σ_y^2 are assumed as $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, and $\mathrm{IG}(a_1, b_1)$, respectively.

The desired VB posterior distribution $Q^*(\boldsymbol{\theta})$ within the family $\mathcal{Q}$ minimizes the KL divergence from the exact posterior distribution $\Pi(\boldsymbol{\theta} \mid \boldsymbol{Y})$,

$$Q^*(\boldsymbol{\theta}) = \underset{Q(\boldsymbol{\theta}) \in \mathcal{Q}}{\arg\min}\, \mathrm{KL}[Q(\boldsymbol{\theta}) \| \Pi(\boldsymbol{\theta} \mid \boldsymbol{Y})],$$

which replaces MCMC sampling with an optimization task. Since directly calculating the KL divergence involves the intractable marginal distribution of $\boldsymbol{Y}$, we instead optimize the evidence lower bound (ELBO), which is equivalent to the KL divergence up to a constant,

$$\mathcal{L}(\boldsymbol{\theta}) = \int q(\boldsymbol{\theta}) \log \frac{\pi(\boldsymbol{\theta}) L_n(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}. \tag{2.4}$$

Although $\mathcal{Q}$ is fully factorized for $\boldsymbol{\theta}$, the dependence of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ on I in the hierarchical structure prevents direct derivation of variational posterior distributions via conjugacy. To overcome this difficulty, we explicitly compute the ELBO $\mathcal{L}(\boldsymbol{\theta})$, allowing for efficient optimization using CAVI (Blei et al., 2017). Algorithmic details are provided in Section 4.

3. Variational Bayes Model Selection and Estimation

In this section, we establish theoretical guarantees for the proposed VB method concerning model selection and parameter estimation.

3. VARIATIONAL BAYES MODEL SELECTION AND ESTIMATION

3.1 Notations for theoretical analysis

For each model indicator, we define an active set $S \subseteq \{1, \ldots, p\}$ to include all indices j where $I_j = 1$, with its size denoted as $|S|$. The considered model space is denoted as $\mathcal{S}$, and let the true active set corresponding to I_0 be S_0 with size s_0. For any S, we define $\boldsymbol{\theta}_S = ((\beta_j)_{j\in S}, (\gamma_\ell)_{\ell\in S}, \boldsymbol{\alpha}, \sigma_y) \in \mathbb{R}^{|S|+3}$ as the coefficient vector for model S. We assume that $(\tau^2_{\boldsymbol{\beta}}, \tau^2_{\boldsymbol{\gamma}})$ and $(q_{\boldsymbol{\beta}}, q_{\boldsymbol{\gamma}})$ are of the same order, respectively, and without loss of generality, we omit the subscripts and denote them as τ^2 and q. We denote the treatment vector as $\boldsymbol{T} = (t_1, \ldots, t_n)$ and design matrices as $\boldsymbol{Z} \in \mathbb{R}^{n\times p_Z}$ and $\boldsymbol{X} \in \mathbb{R}^{n\times p_X}$. For any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p_A}$, the sub-matrix containing columns indexed by S is denoted as $\boldsymbol{A}_S$. We define the norm $\|\boldsymbol{A}\| = \max_{j\in p_A} (\boldsymbol{A}^T\boldsymbol{A})^{1/2}_{jj}$.

3.2 Asymptotic properties under known variance case

We now establish theoretical properties of the VB posterior Q^* with the proofs deferred to Section S1 in the supplementary materials. To simplify technicalities, we first study the case of known noise variance $\sigma_y^2 = 1$, which is commonly considered in VB literature (Ray and Szabó, 2022; Komodromos et al., 2025). The parameter is adjusted to $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) \in \mathbb{R}^{p+2}$ with its true value denoted as $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{\alpha}_0)$. We assume that the covariate spaces $\mathcal{Z}$ and $\mathcal{X}$ are bounded, and consider the ℓ_1-norm bounded parameter space $\Theta(M) := \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1 \leq M\}$, where M is a fixed constant. We require the following regularity conditions.

Condition 1. *(i) (Model dimension) The dimension satisfies* $\log p_n = o(n)$ *as* $n \to \infty$. *(ii) (True parameter) The true parameter satisfies* $\boldsymbol{\theta}_{0S_0^c} = \mathbf{0}$.

Condition 1(i) is common in high-dimensional literature (Lee and Cao, 2021). Condition 1(ii) assumes that inactive signals are negligible (Yang et al., 2016). Although $\boldsymbol{\theta}_{0S_0^c} = \mathbf{0}$ is required for simplicity, it can be relaxed to $\|\boldsymbol{Z}_{S_0^c}\boldsymbol{\beta}_{0S_0^c}\|_2^2 = o(\|\boldsymbol{Z}_{S_0}\boldsymbol{\beta}_{0S_0}\|_2^2)$ and $\|\boldsymbol{X}_{S_0^c}\boldsymbol{\gamma}_{0S_0^c}\|_2^2 = o(\|\boldsymbol{X}_{S_0}\boldsymbol{\gamma}_{0S_0}\|_2^2)$, under which we obtain $L_n(\boldsymbol{\theta}_{0S_0})/L_n(\boldsymbol{\theta}_0) = \mathcal{O}(1)$ as needed in the proof.

Condition 2 (Restricted eigenvalue). *For all* $\boldsymbol{X} \in \mathcal{X}$ *and* $\boldsymbol{Z} \in \mathcal{Z}$, *in the considered model space* $\mathcal{S}$, *there exist constants* λ_1 *and* λ_2 *such that*

$$
\begin{aligned}
0 < \lambda_1 &\leq \min_{S\in\mathcal{S}} \min\left(\lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}_S^T\boldsymbol{X}_S\right), \lambda_{\min}\left(\frac{1}{n}\tilde{\boldsymbol{Z}}_S^T\tilde{\boldsymbol{Z}}_S\right)\right) \\
&\leq \max_{S\in\mathcal{S}} \max\left(\lambda_{\max}\left(\frac{1}{n}\boldsymbol{X}_S^T\boldsymbol{X}_S\right), \lambda_{\max}\left(\frac{1}{n}\tilde{\boldsymbol{Z}}_S^T\tilde{\boldsymbol{Z}}_S\right)\right) \leq \lambda_2,
\end{aligned}
$$

where $\tilde{\boldsymbol{Z}}_S = (\boldsymbol{Z}_S, \boldsymbol{T})$ *combines prognostic variables and the treatment.*

Condition 2 ensures that the eigenvalues of the Gram matrix corresponding to model S are bounded, which is satisfied if, for any $S \in \mathcal{S}$, $|S| \leq m_n + s_0$ with $m_n := (\sqrt{n}/\log p \wedge p)$ (Narisetty et al., 2019).

Condition 3 (Prior distribution). *(i) For some constant* $c > 0$, τ^2 *satisfies* $n\tau^2 \sim (n \vee p^2)^{1+c}$. *(ii) The prior inclusion probability satisfies* $q \sim 1/p$.

Condition 3 specifies the rates of the hyperparameters in the spike-and-slab priors, ensuring posterior concentration on sparse models and enabling consistent variable selection (Narisetty and He, 2014).

Lemma 1. *Under Conditions 1-3, there exists some constant $L_0 > 2$ such that, for any sequence $L_n \geq L_0$, as $n \to \infty$, the VB posterior Q^* satisfies*

$$\mathbb{E}_{\boldsymbol{\theta}_0}\left[Q^*(\boldsymbol{\theta} \in \Theta(M) : |S| \geq L_n s_0)\right] \leq \mathcal{O}\left(\frac{C_L}{L_n}\right) + o(1),$$

with some constant $C_L > 0$.

Lemma 1 shows that the variational posterior distribution puts most of the mass on models of size at most a multiple of s_0, ensuring bounded false positives. If $L_n \to \infty$ at any arbitrarily slow rate, the VB posterior probability on the right-hand side converges to 0.

Theorem 1. *Under the conditions in Lemma 1, there exists some constant $M_0 > 0$ such that, for any sequence $M_n \geq M_0$ growing more slowly than L_n in Lemma 1, as $n \to \infty$, the VB posterior Q^* satisfies*

$$\mathbb{E}_{\boldsymbol{\theta}_0}\left[Q^*\left(\boldsymbol{\theta} \in \Theta(M) : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \geq \frac{\sqrt{M_n s_0 \log p}}{\|\boldsymbol{X}\| \vee \|\tilde{\boldsymbol{Z}}\|}\right)\right] \leq \mathcal{O}\left(\frac{C_M}{M_n}\right) + o(1),$$

with some constant $C_M > 0$.

Theorem 1 shows the VB posterior concentrates in an ℓ_2-ball around the true $\boldsymbol{\theta}_0$. If $M_n \to \infty$ at a slow rate, the VB posterior probability tends to 0 as $n \to \infty$. By combining Lemma 1 and Theorem 1, the VB posterior achieves estimation consistency without selecting excessively large models.

Remark 1. For the exact posterior, we can replace the sequences L_n and M_n with constants L_0 and M_0 and prove similar results without the terms $\mathcal{O}(C_L/L_n)$ and $\mathcal{O}(C_M/M_n)$. These terms quantify the approximation errors between the VB and exact posterior, unveiling a trade-off between computational efficiency and accuracy in the variational approach. As L_n and M_n grow with the sample size n, the approximation errors vanish asymptotically, which is supported by empirical evidence in Section S4.2.

In the following, we introduce additional conditions to strengthen the model selection guarantees in Lemma 1.

Condition 4. *There exists some constant $\kappa_0 > 0$ such that, for any sequence $\kappa_n \geq \kappa_0$, (i) (Refined prior specification) For the constant c in Condition 3, τ^2 satisfies $n\tau^2 \sim (n \vee p^2)^{1+c\kappa_n s_0}$. (ii) (Beta-min) For all $j \in S_0$, the true signals satisfy $|\theta_{0j}| \geq \kappa_n \sqrt{s_0 \log p/n}$.*

Condition 4(i) requires a flatter slab prior than Condition 3(i) to enhance signal capture accuracy. Condition 4(ii) ensures the minimal signal strength of true nonzero coefficients to be sufficiently large, as typically assumed in modeling sparsity (Bühlmann, 2013).

Theorem 2. *Under Conditions 1-4, for any κ_n growing more slowly than L_n defined in Lemma 1, as $n \to \infty$, the VB posterior Q^* satisfies*

$$\mathbb{E}_{\boldsymbol{\theta}_0}\left[Q^*\left(\boldsymbol{\theta} \in \Theta(M) : S \neq S_0\right)\right] \leq \mathcal{O}\left(\frac{C_\kappa}{\kappa_n}\right) + o(1),$$

with some constant $C_\kappa > 0$.

Theorem 2 establishes that, under certain conditions, the VB posterior of the true model converges to 1, achieving model selection consistency within the VB framework. Our results extend beyond existing VB literature on model selection (Ray et al., 2020; Ray and Szabó, 2022), and differ from Narisetty et al. (2019) by relaxing the restrictions on the model space.

3.3 Extension to unknown variance case

Recent studies on VB have relaxed the assumption of a known variance (Ge et al., 2025). In this subsection, we extend our analysis to the more general case with an unknown σ_y^2 and consider $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma_y) \in \mathbb{R}^{p+3}$ with true $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \sigma_{y0})$. Following Ge et al. (2025), we modify the priors of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ to be independent of σ_y to avoid coupling between $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and σ_y.

Obtaining the asymptotic results under an unknown σ_y^2 introduces additional technical challenges beyond the known-variance setting. First, existing theoretical results for mixture of regressions often rely on a reparameterization to establish posterior contraction rates (Zhang et al., 2025).

However, this transformation induces a non-equivalent variational family, leading to a mismatch between the parameters in exact and variational posteriors. To address this, we refine the theoretical arguments in Städler et al. (2010) to avoid reparameterization. Furthermore, σ_y^2 follows an inverse gamma distribution instead of a Gaussian, inducing a more complicated variational family than that in Ray and Szabó (2022).

Under the same conditions in Section 3.2, we extend the variational posterior contraction properties under the unknown variance case. The detailed theoretical results and their derivations are provided in Section S2 of the supplementary materials to avoid repetition.

4. Numerical Algorithm for Variational Inference

In this section, we introduce a coordinate ascent variational inference algorithm to optimize the evidence lower bound.

4. NUMERICAL ALGORITHM FOR VARIATIONAL INFERENCE

4.1 Data augmentation

To facilitate computation, we introduce latent subgroup indicators $\{\delta_i\}_{i=1}^n \in \{0,1\}^n$. The joint likelihood of $\{(y_i, \delta_i)\}_{i=1}^n$ can be rewritten as

$$L_n(\boldsymbol{\theta}, \boldsymbol{\Delta}) = (2\pi\sigma_y^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma_y^2}\left\|\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta} - \alpha_1\boldsymbol{\Delta}\boldsymbol{T} - \alpha_2(\mathbf{I} - \boldsymbol{\Delta})\boldsymbol{T}\right\|_2^2\right\}$$
$$\times \prod_{i=1}^n \frac{\exp(x_i^T\boldsymbol{\gamma})^{\delta_i}}{1+\exp(x_i^T\boldsymbol{\gamma})},$$

where $\boldsymbol{\Delta} = \mathrm{D}(\delta_1, \ldots, \delta_n)$, with $\mathrm{D}(\cdot)$ reshaping a vector into a diagonal matrix. Since the logistic model does not exhibit direct conjugacy, a straightforward CAVI approach is intractable (Durante and Rigon, 2019). To address this, we introduce Pólya-Gamma (PG) latent variables $\{\omega_i\}_{i=1}^n$ for data augmentation (Polson et al., 2013), which induces conjugacy with the Gaussian prior on $\boldsymbol{\gamma}$. The joint likelihood involving (δ_i, ω_i) is given by

$$L_n(\boldsymbol{\theta}, \boldsymbol{\phi}) = (2\pi\sigma_y^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma_y^2}\left\|\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta} - \alpha_1\boldsymbol{\Delta}\boldsymbol{T} - \alpha_2(\mathbf{I} - \boldsymbol{\Delta})\boldsymbol{T}\right\|_2^2\right\}$$
$$\times 2^{-n} \exp\left\{\mathbf{1}^T\left(\boldsymbol{\Delta} - \frac{1}{2}\mathbf{I}\right)\boldsymbol{X}\boldsymbol{\gamma}\right\} \exp\left\{-\frac{1}{2}\boldsymbol{\gamma}^T\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X}\boldsymbol{\gamma})\right\} \prod_{i=1}^n p(\omega_i), \tag{4.5}$$

where $\boldsymbol{\phi} = (\boldsymbol{\Delta}, \boldsymbol{\Omega})$, $\boldsymbol{\Omega} = \mathrm{D}(\omega_1, \ldots, \omega_n)$, and $p(\omega_i)$ denotes the density of $\mathrm{PG}(1,0)$ variable. The variational family with $\boldsymbol{\Delta}$ and $\boldsymbol{\Omega}$ is factorized as

$$\mathcal{Q} = \left\{q(\boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{j=1}^{p_Z} q(\beta_j) \prod_{\ell=1}^{p_X} q(\gamma_\ell) \prod_{k=1}^{2} q(\alpha_k) q(\sigma_y^2) \prod_{i=1}^n [q(\delta_i)q(\omega_i)]\right\}, \tag{4.6}$$

where $q(\delta_i)$ and $q(\omega_i)$ are the densities of $\mathrm{Bern}(\pi_i)$ and $\mathrm{PG}(1, c_i)$, respectively, with $\{\pi_i\}_{i=1}^n$ and $\{c_i\}_{i=1}^n$ being variational parameters.

4.2 Coordinate ascent variational inference

The CAVI updates are derived by optimizing the ELBO as

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \int q(\boldsymbol{\theta}, \boldsymbol{\phi}) \log \frac{\pi(\boldsymbol{\theta}) L_n(\boldsymbol{\theta}, \boldsymbol{\phi})}{q(\boldsymbol{\theta}, \boldsymbol{\phi})} d\boldsymbol{\theta} d\boldsymbol{\phi}. \tag{4.7}$$

For each factor, to obtain its variational posterior distribution, we fix the distributions of other factors and maximize the ELBO in (4.7). The updates for non-hierarchical factors are directly derived from

$$q(\theta_j) \propto \exp\left\{\mathbb{E}_{-q(\theta_j)} \log\left[\pi(\boldsymbol{\theta}) L_n(\boldsymbol{\theta}, \boldsymbol{\phi})\right]\right\}, \tag{4.8}$$

where the subscript $-q(\theta_j)$ indicates the expectation is taken over all other factors except θ_j. In details, updates for $q(\alpha_1)$ and $q(\alpha_2)$ are given by

$$\begin{aligned} \mu_1 &= \left(\boldsymbol{T}^T \mathbb{E}\boldsymbol{\Delta} \boldsymbol{T} + \sigma_\alpha^{-2}\right)^{-1} \left[(\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{\eta}^{\boldsymbol{\beta}} \odot \boldsymbol{\mu}^{\boldsymbol{\beta}}))^T \mathbb{E}\boldsymbol{\Delta}\right] \boldsymbol{T}, \\ \sigma_1^2 &= \left[a_1 \left(\boldsymbol{T}^T \mathbb{E}\boldsymbol{\Delta} \boldsymbol{T} + \sigma_\alpha^{-2}\right) / b_1\right]^{-1}, \end{aligned} \tag{4.9}$$

where $\mathbb{E}\boldsymbol{\Delta} = \mathrm{D}(\pi_1, \ldots, \pi_n)$ and $\odot$ denotes element-by-element product, and

$$\begin{aligned} \mu_2 &= \left(\boldsymbol{T}^T (\mathbf{I} - \mathbb{E}\boldsymbol{\Delta}) \boldsymbol{T} + \sigma_\alpha^{-2}\right)^{-1} \left[(\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{\eta}^{\boldsymbol{\beta}} \odot \boldsymbol{\mu}^{\boldsymbol{\beta}}))^T (\mathbf{I} - \mathbb{E}\boldsymbol{\Delta})\right] \boldsymbol{T}, \\ \sigma_2^2 &= \left[a_1 \left(\boldsymbol{T}^T (\mathbf{I} - \mathbb{E}\boldsymbol{\Delta}) \boldsymbol{T} + \sigma_\alpha^{-2}\right) / b_1\right]^{-1}. \end{aligned} \tag{4.10}$$

Updates for (a_1, b_1) in $q(\sigma_y^2)$, π_i in $q(\delta_i)$, and c_i in $q(\omega_i)$ for $i = 1, \ldots, n$ can be derived similarly, and for conciseness, are deferred to Section S3.2.

The updates for hierarchical factors including γ_ℓ, I_ℓ^γ, β_j, and $I_j^{\boldsymbol{\beta}}$ cannot be directly obtained from (4.8). However, we can leverage the hierarchical structure in the variational posterior distribution to calculate the ELBO. For example, the updates of γ_ℓ for $\ell = 1, \ldots, p_X$ can be divided into two cases. When $I_\ell^\gamma = 0$, the variational posterior of γ_ℓ is δ_0, while conditional on $I_\ell^\gamma = 1$, the maximizers of the ELBO are

$$\begin{aligned} \mu_\ell^\gamma &= \frac{\mathbf{1}^T(\mathbb{E}\boldsymbol{\Delta} - 1/2)x_\ell - (\boldsymbol{\eta}_{-\ell}^\gamma \odot \boldsymbol{\mu}_{-\ell}^\gamma)^T \boldsymbol{X}_{-\ell}^T \mathbb{E}\boldsymbol{\Omega} x_\ell}{x_\ell^T \mathbb{E}\boldsymbol{\Omega} x_\ell + \tau_\gamma^{-2}}, \\ \sigma_\ell^{\gamma 2} &= \frac{1}{x_\ell^T \mathbb{E}\boldsymbol{\Omega} x_\ell + \tau_\gamma^{-2}}, \end{aligned} \tag{4.11}$$

where the subscript $-\ell$ excludes the ℓth component or column in a vector or matrix, and $\mathbb{E}\boldsymbol{\Omega} = \mathrm{D}(\tanh(c_1/2)/2c_1, \ldots, \tanh(c_n/2)/2c_n)$. The updates of I_ℓ^γ for $\ell = 1, \ldots, p_X$ are given by solving

$$\log \frac{\eta_\ell^\gamma}{1 - \eta_\ell^\gamma} = \frac{\mu_\ell^{\gamma 2}}{2\sigma_\ell^{\gamma 2}} + \log \frac{q_\gamma \sigma_\ell^\gamma}{(1 - q_\gamma)\tau_\gamma}. \tag{4.12}$$

For the updates of β_j for $j = 1, \ldots, p_Z$, conditional on $I_j^{\boldsymbol{\beta}} = 0$, the variational posterior of β_j is δ_0, while conditional on $I_j^{\boldsymbol{\beta}} = 1$, the maximizers of the ELBO are given by

$$\begin{aligned} \mu_j^{\boldsymbol{\beta}} &= \frac{\left[\boldsymbol{Y} - \mu_1 \mathbb{E}\boldsymbol{\Delta}\boldsymbol{T} - \mu_2(\mathbf{I} - \mathbb{E}\boldsymbol{\Delta})\boldsymbol{T}\right]^T z_j - (\boldsymbol{\eta}_{-j}^{\boldsymbol{\beta}} \odot \boldsymbol{\mu}_{-j}^{\boldsymbol{\beta}})\boldsymbol{Z}_{-j}^T z_j}{z_j^T z_j + \tau_{\boldsymbol{\beta}}^{-2}}, \\ \sigma_j^{\boldsymbol{\beta}} &= \frac{1}{a_1(z_j^T z_j + \tau_{\boldsymbol{\beta}}^{-2})/b_1}, \end{aligned} \tag{4.13}$$

The updates of $I_j^{\boldsymbol{\beta}}$ can be obtained as

$$\log \frac{\eta_j^{\boldsymbol{\beta}}}{1 - \eta_j^{\boldsymbol{\beta}}} = \frac{\mu_j^{\boldsymbol{\beta}2}}{2\sigma_j^{\boldsymbol{\beta}2}} - \frac{1}{2}(\log b_1 - \psi(a_1)) + \log \frac{q_{\boldsymbol{\beta}} \sigma_j^{\boldsymbol{\beta}}}{(1 - q_{\boldsymbol{\beta}})\tau_{\boldsymbol{\beta}}}, \tag{4.14}$$

where $\psi(\cdot)$ is the digamma function. The detailed derivation and updates can be found in Section S3.2 of the supplementary materials.

4.3 Implementation details

Initialization. Since the VB method optimizes a non-convex objective, its performance can be sensitive to initialization. We initialize $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ in two steps. For $\boldsymbol{\gamma}$, we first use subgroup methods to identify an active predictive covariate set. Specifically, we adopt GUIDE (Loh, 2002) for efficient variable screening. As $\boldsymbol{\beta}$ is less sensitive, we randomly select prognostic covariates based on a predetermined size of p_Z. We then run an EM algorithm using the selected I^{γ} and $I^{\boldsymbol{\beta}}$ to obtain initial parameter values.

Hyperparameters. Hyperparameters are chosen based on prior assumptions. We set $q_{\boldsymbol{\beta}} = \min(0.2, 20/p_Z)$, $\tau_{\boldsymbol{\beta}} = \max(p_Z/(10\sqrt{n}), 1.3)$, and $\tau_{\boldsymbol{\gamma}} = \max(p_X/(10\sqrt{n}), 1.3)$. Since subgroup signal strength is often weak, we recommend a larger predictive inclusion probability of $q_{\boldsymbol{\gamma}} = 0.5$ for finite samples. For other hyperparameters, we use $a_0 = 2$, $b_0 = 1$, and $\sigma_\alpha^2 = 1$.

Updating process. We adopt the prioritized scheme from Ray and Szabó (2022), where updates for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ begin with the component having the largest absolute value and proceed to the smallest. The iteration stops when the maximum difference in the entropies of posterior inclusion probabilities, Δ_H, falls below a threshold ϵ, which is calculated as $\max_j |H(\eta_j^{\boldsymbol{\beta}}) - H(\eta_{j,old}^{\boldsymbol{\beta}})| \vee \max_\ell |H(\eta_\ell^{\gamma}) - H(\eta_{\ell,old}^{\gamma})|$ with $H(p) = -p\log p - (1-p)\log(1-p)$. The algorithm is summarized in Algorithm 1, where order$(|\boldsymbol{\mu}|)$ returns the indices of $|\boldsymbol{\mu}|$ in descending order.

Algorithm 1 Variational Bayes for structured mixture models

```
**Input:** observations $\{(y_i, z_i, x_i, t_i)\}_{i=1}^n$
**Output:** $\boldsymbol{\mu}^{\boldsymbol{\beta}}$, $\boldsymbol{\eta}^{\boldsymbol{\beta}}$, $\boldsymbol{\mu}^{\gamma}$, $\boldsymbol{\eta}^{\gamma}$, $\mu_1$, $\mu_2$, $\{\pi_i\}_{i=1}^n$
Initialize $\boldsymbol{\mu}^{\boldsymbol{\beta}}$, $\sigma^{\boldsymbol{\beta}2}$, $\boldsymbol{\eta}^{\boldsymbol{\beta}}$, $\boldsymbol{\mu}^{\gamma}$, $\sigma^{\gamma 2}$, $\boldsymbol{\eta}^{\gamma}$, $\mu_1$, $\mu_2$, $\sigma_1^2$, $\sigma_2^2$, $\{\pi_i\}_{i=1}^n$, $\Delta_H$.
**while** $\Delta_H \geq \epsilon$ **do**
  $R^{\boldsymbol{\beta}} :=$ order($|\boldsymbol{\mu}^{\boldsymbol{\beta}}|$), $R^{\gamma} :=$ order($|\boldsymbol{\mu}^{\gamma}|$), $\boldsymbol{\eta}^{\boldsymbol{\beta}}_{old} = \boldsymbol{\eta}^{\boldsymbol{\beta}}$, $\boldsymbol{\eta}^{\gamma}_{old} = \boldsymbol{\eta}^{\gamma}$
  **for** $j \in \{1, \ldots, p_Z\}$ **do**
    $m := R^{\boldsymbol{\beta}}_j$
    Update $\mu^{\boldsymbol{\beta}}_m$ and $\sigma^{\boldsymbol{\beta}2}_m$ for the prognostic coefficient $\beta_m$ via Eq.(4.13),
    and its variational posterior inclusion probability $\eta^{\boldsymbol{\beta}}_m$ via Eq.(4.14).
  **end for**
  **for** $\ell \in \{1, \ldots, p_X\}$ **do**
    $m := R^{\gamma}_{\ell}$
    Update $\mu^{\gamma}_m$ and $\sigma^{\gamma 2}_m$ for the predictive coefficient $\gamma_m$ via Eq.(4.11),
    and its variational posterior inclusion probability $\eta^{\gamma}_m$ via Eq.(4.12).
  **end for**
  Update $\mu_1$ and $\sigma_1^2$ for the treatment effect $\alpha_1$ via Eq.(4.9), and $\mu_2$ and
  $\sigma_2^2$ for the treatment effect $\alpha_2$ via Eq.(4.10).
  **for** $i \in \{1, \ldots, n\}$ **do**
    Update $\pi_i$ for the subgroup indicator $\delta_i$ via Eq.(S3.1) and $c_i$ for the
    latent Pólya-Gamma variable $\omega_i$ via Eq.(S3.2).
  **end for**
  Update $a_1 = n/2 + 1 + \sum_{j=1}^{p_Z} \eta^{\boldsymbol{\beta}}_j / 2 + a_0$ and $b_1$ for $\sigma_y^2$ via Eq.(S3.3).
  Compute $\Delta_H = \max_j |H(\eta^{\boldsymbol{\beta}}_j) - H(\eta^{\boldsymbol{\beta}}_{j,old})| \vee \max_{\ell} |H(\eta^{\gamma}_{\ell}) - H(\eta^{\gamma}_{\ell,old})|$.
**end while**
```

5. Simulation Studies

We evaluate the proposed Variational Structured Mixture models (VSM) using comprehensive simulations. Variable selection performance is assessed using true positive rate (**TPR**), false discovery rate (**FDR**), and the **F1** score calculated as $\mathrm{F1} = \frac{2\,\mathrm{TPR}(1-\mathrm{FDR})}{\mathrm{TPR}+(1-\mathrm{FDR})}$. The F1 score offers a trade-off between TPR and FDR. We also introduce **Ext** as the probability of selecting the exact true model when the size is restricted to $|I_0|$, which reflects the ability of ranking variable importance and is independent of thresholds on posterior inclusion probabilities.

5.1 Accuracy and time comparison with MCMC

We assess the finite sample performance of VSM for both $p < n$ and $p \geq n$, with $n \in \{200, 300\}$, $p \in \{100, 500, 2000\}$, and $p_Z = p_X = p/2$. Data are generated based on the structured mixture model in (2.1) and (2.2), with $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ independently drawn from the standard normal distribution. Intercept columns are included in $\boldsymbol{Z}$ and $\boldsymbol{X}$. The true values of $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$ are $(1, -1.5, 2, -2.5, 3, 0, \ldots, 0)$, and the true treatment effects are $\alpha_{10} = 40$ and $\alpha_{20} = 0$. The responses y_i's are independently sampled according to model (2.1) with variance equal to 1.

We compare VSM with a scalable MCMC method, BVSA (Zhang et al., 2025). For BVSA, we set the Gibbs chain length to 20000 with a burn-in of 5000, and select hyperparameters as recommended. Both methods employ a posterior inclusion probability threshold of 0.5. Results are summarized from 100 independent trials and presented in Table 1. We additionally examine scenarios with correlated covariates in Section S4.1 in the supplementary materials, and discuss estimation errors in Section S4.2.

Table 1: Performance on variable selection under structured mixture model settings with different p and n. All metrics are averaged over 100 trials.

p	n	Method	$\boldsymbol{\beta}$				$\boldsymbol{\gamma}$			
			TPR	FDR	F1	Ext	TPR	FDR	F1	Ext
100	200	VSM	1	0	1	100%	0.958	0.044	0.953	90%
		BVSA	1	0	1	100%	0.970	0.066	0.946	92%
	300	VSM	1	0	1	100%	0.995	0.041	0.974	98%
		BVSA	1	0	1	100%	0.998	0.071	0.958	97%
500	200	VSM	1	0	1	100%	0.903	0.140	0.873	64%
		BVSA	1	0	1	100%	0.885	0.045	0.911	73%
	300	VSM	1	0	1	100%	0.983	0.141	0.911	90%
		BVSA	1	0	1	100%	0.968	0.029	0.966	92%
2000	200	VSM	1	0	1	100%	0.758	0.200	0.755	40%
		BVSA	1	0	1	100%	0.670	0.104	0.744	29%
	300	VSM	1	0	1	100%	0.955	0.113	0.910	88%
		BVSA	1	0	1	100%	0.873	0.054	0.898	69%

From the results of $\boldsymbol{\beta}$ in Table 1, we observe that VSM achieves accurate prognostic variable selection, with TPR of 1 across all settings. For both $p = 100$ and $p = 500$, VSM and BVSA identify most of the active predictive covariates, with a small sample size of $n = 200$. The performance of VSM is comparable to BVSA, suggesting that VB approximation retains high accuracy. As n increases to 300, the performance of VSM improves, reducing the difference between VSM and BVSA to negligible levels.

As p increases to 2000, the performance of both methods declines, but they remain effective. When $n = 200$, VSM achieves higher F1 scores than BVSA, highlighting its capability when $p \gg n$. The threshold-free measure Ext results suggest that VSM ranks variable importance more effectively. When $n = 300$, the advantage of VSM in Ext becomes more pronounced. These findings confirm the consistency of variable selection and demonstrate the improved accuracy of VB approximation with larger sample sizes.

To showcase the scalability of VSM, we consider $n = 300$ with varying model dimensions $p \in \{100, 200, \ldots, 2000\}$, using the same true parameter values as before. We compare VSM with BVSA, which is specifically designed for high-dimensional models and offers computational improvements over traditional MCMC algorithms. All experiments are conducted on a single core of a MacBook Pro with an Apple M2 chip and 16 GB of memory, utilizing the Rcpp interface and the Armadillo library.

For each p, we record the average running time in seconds from 10 random trials. As shown in Figure 1, VSM computes substantially faster than BVSA, even though BVSA is designed for scalable MCMC inference.

The results highlight the computational advantage of the VB methods over MCMC methods when handling high-dimensional problems.

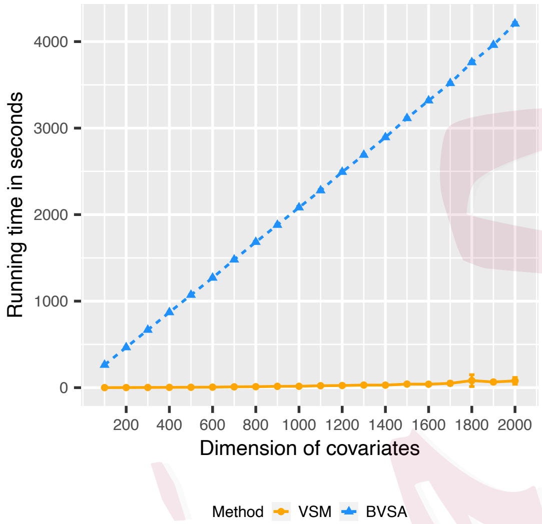

Figure 1: Running time in seconds with varying p when $n = 300$.

5.2 Results under traditional subgroup settings

We assess the robustness of VSM in a traditional subgroup setting, where subgroups are determined by splitting rules based on certain covariates. We consider two settings used in the subgroup literature (Loh et al., 2019):

$$\mathbf{S1}: Y = 1 + Z_2 + 40tI_{(X_1>0,X_4<1,X_6=2)} + \varepsilon,$$
$$\mathbf{S2}: Y = 1 + Z_1 + Z_2 + Z_4 + I_{(Z_6=2)} + Z_7 + 40tI_{(X_1>0,X_4<1,X_6=2)} + \varepsilon,$$

with $\varepsilon \sim N(0, 1)$. We consider two scenarios with $n = 200$ and $p_X = 10$ or $p_X = 100$. For $p_X = 10$, predictive covariates are generated according to $X_1 \sim N(0,1)$, $(X_2, X_3) \sim N(\mathbf{0}, \Sigma)$ with $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = \Sigma_{21} = 0.5$, $X_4 \sim \text{Exp}(1)$, $X_5 \sim \text{Bern}(0.5)$, $X_6 \sim \text{Multinomial}(3, 1/3)$, and $(X_7, X_8, X_9, X_{10}) \sim N(\mathbf{0}, \Sigma)$ with diagonal elements of 1 and nondiagonal elements of 0.5. For $p_X = 100$, covariates $X_{11}, \ldots, X_{100}$ are independently sampled from $N(0,1)$. The prognostic covariates $\boldsymbol{z}_i$ are set to be identical to $\boldsymbol{x}_i$, leading to a total dimension p of 20 and 200. We simulate independent testing datasets with $n = 5000$ to estimate subgroup prediction error (PE). We further include a setting without true subgroups to examine whether VSM may incorrectly identify subgroups, as discussed in Section S4.3.

Comparison methods include BVSA and subgroup identification approaches. We consider splitting-rule-based methods, including GUIDE (Loh, 2002), PRIM (Chen et al., 2015), MOB (Seibold et al., 2016), and SeqBT (Huang et al., 2017), as well as the penalized SVM-based FindIt (Imai and Ratkovic, 2013). Details on their implementations are provided in Section S4.3. All methods are evaluated across 100 random trials.

In the low-dimensional setting, the results of predictive variable selection are presented in the left panel of Table 2. VSM selects covariates more accurately than other subgroup methods and achieves lower prediction errors. The performance of VSM is comparable to BVSA, indicating that VB achieves computational efficiency without sacrificing much accuracy.

Table 2: Predictive variable selection results with different p when $n = 200$.

	$p = 20$					$p = 200$				
(a) S1: $Y = 1 + Z_2 + 40tI_{(X_1>0,X_4<1,X_6=2)} + \varepsilon$										
	TPR	FDR	F1	Ext	PE	TPR	FDR	F1	Ext	PE
VSM	0.883	0.021	0.904	77%	0.074	0.707	0.265	0.669	38%	0.091
BVSA	0.917	0.013	0.941	93%	0.105	0.783	0.198	0.781	50%	0.106
GUIDE	0.657	0.138	0.726	14%	0.226	0.517	0.119	0.610	1%	0.215
FindIt	0.997	0.656	0.507	61%	0.288	-	-	-	-	-
PRIM	0.383	0.324	0.455	0	0.328	0.097	0.771	0.129	0	0.245
MOB	0.250	0.689	0.274	0	0.312	0.147	0.810	0.164	0	0.302
SeqBT	0.343	0.010	0.507	0	0.236	0.333	0.050	0.490	0	0.234
(b) S2: $Y = 1 + Z_1 + Z_2 + Z_4 + I_{(Z_6=2)} + Z_7 + 40tI_{(X_1>0,X_4<1,X_6=2)} + \varepsilon$										
	TPR	FDR	F1	Ext	PE	TPR	FDR	F1	Ext	PE
VSM	0.873	0.027	0.896	77%	0.132	0.710	0.272	0.664	37%	0.090
BVSA	0.907	0.012	0.934	93%	0.106	0.810	0.182	0.802	51%	0.106
GUIDE	0.703	0.240	0.703	29%	0.227	0.507	0.124	0.602	1%	0.213
FindIt	0.997	0.657	0.507	25%	0.151	-	-	-	-	-
PRIM	0.353	0.409	0.406	0	0.302	0.077	0.845	0.094	0	0.220
MOB	0.760	0.399	0.662	8%	0.287	0.620	0.348	0.624	10%	0.282
SeqBT	0.347	0.000	0.512	0	0.235	0.327	0.060	0.482	0	0.237

In the high-dimensional setting, FindIt is excluded from the comparison because it includes all covariate interactions and becomes computationally infeasible. As shown in the right panel of Table 2, the performance of all methods declines as the dimensionality increases. However, VSM significantly outperforms other subgroup methods. Although VB approximation experiences some loss in accuracy, the results remain comparable to BVSA, highlighting the reliability of VSM under model misspecification.

6. Real Data Application

In this section, we apply our proposed method to two datasets: the International Warfarin Pharmacogenetics Consortium dataset and the AIDS Clinical Trials Group 320 study.

6.1 Application to IWPC dataset

The International Warfarin Pharmacogenetics Consortium (IWPC) dataset (International Warfarin Pharmacogenetics Consortium, 2009) includes clinical and genetic information from over 5700 warfarin-treated patients, covering demographic characteristics, therapeutic dose, and genotype variants of CYP2C9 and VKORC1, which are well-established factors influencing warfarin sensitivity and dose requirements (Sconce et al., 2005).

Although warfarin's effectiveness has been studied at the population level (Anderson et al., 2007; Pirmohamed et al., 2013), increasing focus has been placed on subgroup analysis to identify patients who benefit more from the therapy (Stack and Maurice, 2016; Liu et al., 2025). This motivates the investigation of treatment effect heterogeneity based on baseline covariates to improve dosing decisions across patient subpopulations.

In our study, the response variable is the post-treatment international normalized ratio (INR), a common measure of blood coagulation. Predictors include demographic variables (age, gender, weight, height, BMI, race), clinical covariates (use of aspirin, amiodarone, enzyme inducers), and genetic factors (CYP2C9 and VKORC1 genotypes). We follow the preprocessing steps in Liu et al. (2025) and further include pairwise interactions among predictors. The same set of 136 covariates is used for both prognostic and predictive components, resulting in $p = 272$. To adapt to our model, we define a binary warfarin treatment indicator by dichotomizing the dosage at its median value into high and low dose groups. After excluding subjects with missing records, the sample consists of $n = 2836$ patients.

Active prognostic and predictive covariates are selected based on the variational posterior inclusion probabilities, averaged over five independent runs with random initializations. The highest posterior prognostic inclusion probability is 0.821 for the interaction between amiodarone and CYP2C9*1/*2, while the probabilities of other covariates are below 0.5. For predictive covariates, the highest inclusion probability is 1 for the interaction between age and Asian race, followed by 0.614 for the interaction between age and weight, with other covariates having probabilities below 0.5. The identification of Asian race and age as predictive factors aligns with previous findings and can be explained by pharmacogenetic and metabolic differences across patients (Jensen et al., 2012; Gaikwad et al., 2014). Meanwhile, the discovery of novel predictive interactions offers new directions for investigating treatment effect heterogeneity and refining personalized dosing strategies in future studies. We present the probabilities of all covariates in Section S5.1. Using a hard threshold of 0.5 for inclusion probabilities, the estimated model, omitting inactive covariates, is given by

$$Y \sim \hat{\pi} N(-0.184 - 0.516\,\text{amiodarone} \cdot \text{CYP2C9}^{*}\,1/^{*}2 + 0.117t, 0.985^2)$$

$$+ (1-\hat{\pi}) N(-0.184 - 0.516\,\text{amiodarone} \cdot \text{CYP2C9}^{*}\,1/^{*}2 - 0.421t, 0.985^2),$$

$$\log[\hat{\pi}/(1-\hat{\pi})] = -8.091 - 2.522\,\text{age} \cdot \text{weight} + 0.187\,\text{age} \cdot \text{Asian},$$

where the opposite signs of treatment effects suggest that, for patients in one subgroup, high warfarin dosage may lead to adverse effects.

To validate the identified subgroups, we examine the treatment effects in two subgroups (Group U and Group L) with the highest and lowest 25% predicted subgroup proportions. Figure 2 illustrates the response under different treatments (1 and 0) in each subgroup. In Group U, the response is significantly higher under treatment 1, while Group L exhibits substantial overlap in the box plots, indicating a negligible treatment effect. This confirms that the identified subgroups reflect meaningful differences in treatment response.

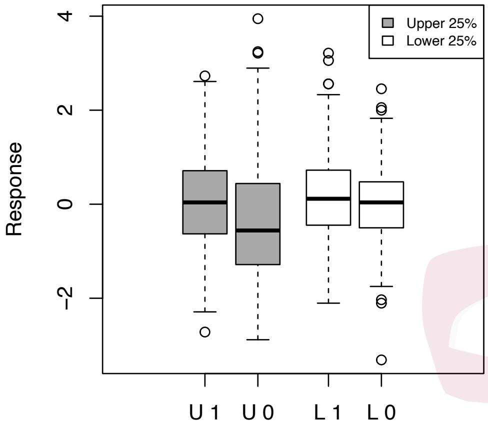

Figure 2: Response under treatment and no treatment in two groups with the highest and lowest 25% predicted subgroup proportions, respectively.

6.2 Application to ACTG 320 study

The AIDS Clinical Trials Group (ACTG) 320 study (Hammer et al., 1997) evaluated the efficacy of a three-drug regimen of indinavir, zidovudine, and lamivudine versus a two-drug regimen of zidovudine and lamivudine for HIV-infected patients. Following Zhao et al. (2013), we define the CD4 count change at week 24 as the response variable to identify patients benefiting from the three-drug regimen. The dataset consists of 852 observations with 11 pre-treatment covariates. To simulate high-dimensional scenarios, we add 415 noise covariates from $N(0, 1)$ to both the prognostic and predictive feature sets, resulting in a dimension of $p = n = 852$.

We identify active covariates based on posterior inclusion probabilities.

Among prognostic covariates, log baseline HIV-1 RNA concentration (L_r) and log baseline CD4 counts (L_c) exhibit the highest probabilities of 1 and 0.916, respectively, while all others remain below 0.3. For predictive covariates, both L_r and L_c attain probabilities of 1, with the remaining variables showing probabilities under 0.3. The identification of L_r and L_c as predictive variables is consistent with findings in previous studies (Cai et al., 2010; Zhao et al., 2013). The estimated treatment effects for the two subgroups are 139.25 and -7.63, respectively, indicating that the three-drug regimen may have adverse effects for a subset of patients.

We further compare the predictive variable selection results of VSM with existing subgroup identification methods. We report the selection frequencies of all covariates as well as the average number of selected noise variables from 100 random trials. As shown in Table 3, VSM consistently identifies L_r and L_c, while rarely selecting noise covariates. In contrast, GUIDE and SeqBT are highly sensitive to noise and fail to distinguish informative variables. Although PRIM and MOB select L_c, they fail to capture L_r, resulting in incomplete subgroup identification. These results highlight the robustness of VSM in high-dimensional settings.

Table 3: Selection frequencies of predictive covariates and the average number of selected noises from 100 trials

	sex	dr	hemo	wt	Ks	zido	age	Lr	Lc	Afri	Hisp	num of noises
VSM	0	0	0	0	0	0	0	0.94	0.84	0	0.06	0.06
GUIDE	0	0	0	0	0	0	0.01	0	0	0.01	0	0.07
PRIM	0	0	0.05	0	0	0	0	0	1.00	0	0	0.03
MOB	0	0	0	0	0	0	0	0	1.00	0	0	0.25
SeqBT	0	0	0	0	0	0	0	0	0	0	0	1.31

7. Discussion

In this paper, we propose VSM, a scalable method for high-dimensional structured mixture models. By approximating the exact posterior with a variational distribution, our method enables efficient and simultaneous inference for both variable selection and parameter estimation. We establish theoretical guarantees for model selection consistency of both prognostic and predictive variables, as well as for consistency in parameter estimation. A coordinate ascent variational inference algorithm is developed, and computational scalability is ensured via data augmentation strategies. Comprehensive simulation studies demonstrate that VSM achieves accuracy comparable to MCMC methods, while offering substantial improvements in computational efficiency. Applications to real-world datasets further highlight the practical utility of VSM in identifying meaningful subgroups and uncovering treatment heterogeneity.

Future extensions of our method may consider structured mixture models with more than two components. Recent advances in variational inference for multinomial logistic models have introduced sophisticated data augmentation techniques (Galy-Fajou et al., 2020; Wojnowicz et al., 2022), which could be utilized to construct surrogate ELBOs that accommodate latent subgroup memberships beyond the binary case. Furthermore, settings with heteroscedastic noise structures are for future study.

Supplementary Materials

The online supplementary materials contain (1) proofs of the theoretical results under a known noise variance; (2) extended theoretical results and proofs under an unknown noise variance; (3) detailed CAVI updates and their derivation; (4) additional results of simulation studies and sensitivity analyses; (5) additional information on the real applications.

Acknowledgments

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

References

Anderson, J. L., B. D. Horne, S. M. Stevens, A. S. Grove, S. Barton, Z. P. Nicholas, and et al. (2007). Randomized trial of genotype-guided versus standard warfarin dosing in patients initiating oral anticoagulation. *Circulation 116*(22), 2563–2570.

Atchadé, Y. A. (2017). On the contraction properties of some high-dimensional quasi-posterior distributions. *The Annals of Statistics 45*(5), 2248 – 2273.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association 112*(518), 859–877.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli 19*(4), 1212–1242.

Cai, T., L. Tian, P. H. Wong, and L. J. Wei (2010). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics 12*(2), 270–282.

Carbonetto, P. and M. Stephens (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis 7*(1), 73–108.

Chen, G., H. Zhong, A. Belousov, and V. Devanarayan (2015). A PRIM approach to predictive-signature development for patient stratification. *Statistics in Medicine 34*(2), 317–342.

Durante, D. and T. Rigon (2019). Conditionally conjugate mean-field variational Bayes for logistic models. *Statistical Science 34*(3), 472 – 485.

Gaikwad, T., K. Ghosh, and S. Shetty (2014). Vkorc1 and cyp2c9 genotype distribution in asian countries. *Thrombosis Research 134*(3), 537–544.

Galy-Fajou, T., F. Wenzel, C. Donner, and M. Opper (2020). Multi-class gaussian process classification made conjugate: Efficient inference via data augmentation. *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference 115*, 755–765.

Ge, C., B. Lin, and J. S. Liu (2025). A Variational Spike-and-Slab Approach for Group Variable Selection. *Bayesian Analysis*, 1 – 31.

George, E. I. and R. E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*(423), 881–889.

Ghosh, J., A. H. Herring, and A. M. Siega-Riz (2011). Bayesian variable selection for latent class models. *Biometrics 67*(3), 917–925.

Hammer, S. M., K. E. Squires, M. D. Hughes, J. M. Grimes, L. M. Demeter, J. S. Currier, and et al. (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*(11), 725–733.

Huang, X., Y. Sun, P. Trow, S. Chatterjee, A. Chakravartty, L. Tian, and et al. (2017). Patient subgroup identification for clinical drug development. *Statistics in Medicine 36*(9), 1414–1428.

Imai, K. and M. Ratkovic (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics 7*(1), 443–470.

International Warfarin Pharmacogenetics Consortium (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine 360*(8), 753–764.

Italiano, A. (2011). Prognostic or predictive? It's time to get back to definitions! *Journal of Clinical Oncology 29*(35), 4718–4718.

Jensen, B. P., P. K. L. Chin, R. L. Roberts, and E. J. Begg (2012). Influence of adult age on the total and free clearance and protein binding of (r)- and (s)-warfarin. *British Journal of Clinical Pharmacology 74*(5), 797–805.

Khalili, A. and J. Chen (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association 102*(479), 1025–1038.

Komodromos, M., E. O. Aboagye, M. Evangelou, S. Filippi, and K. Ray (2022). Variational Bayes for high-dimensional proportional hazards models with applications within gene expression. *Bioinformatics 38*(16), 3918–3926.

Komodromos, M., M. Evangelou, S. Filippi, and K. Ray (2025). Group Spike-and-Slab Variational Bayes. *Bayesian Analysis*, 1 – 29.

Lee, K. and X. Cao (2021). Bayesian group selection in logistic regression with application to MRI data analysis. *Biometrics 77*(2), 391–400.

Liu, P., Y. Li, and J. Li (2025). Change surface regression for nonlinear subgroup identification with application to warfarin pharmacogenomics data. *Biometrics 81*(1), ujae169.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica 12*(2), 361–386.

Loh, W.-Y., L. Cao, and P. Zhou (2019). Subgroup identification for precision medicine: A comparative review of 13 methods. *WIREs Data Mining and Knowledge Discovery 9*(5), e1326.

McLachlan, G. J., S. X. Lee, and S. I. Rathnayake (2019). Finite mixture models. *Annual Review of Statistics and Its Application 6*, 355–378.

Narisetty, N. N. and X. He (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics 42*(2), 789 – 817.

Narisetty, N. N., J. Shen, and X. He (2019). Skinny Gibbs: A consistent and scalable Gibbs sampler for model selection. *Journal of the American Statistical Association 114*(527), 1205–1217.

Pirmohamed, M., G. Burnside, N. Eriksson, A. L. Jorgensen, C. H. Toh, T. Nicholson, and et al. (2013). A randomized trial of genotype-guided dosing of warfarin. *New England Journal of Medicine 369*(24), 2294–2303.

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

Ray, K. and B. Szabó (2022). Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association 117*(539), 1270–1281.

Ray, K., B. Szabó, and G. Clara (2020). Spike and slab variational Bayes for high dimensional logistic regression. In *Advances in Neural Information Processing Systems*, Volume 33, pp. 14423–14434.

Sconce, E. A., T. I. Khan, H. A. Wynne, P. Avery, L. Monkhouse, B. P. King, and et al. (2005). The impact of CYP2C9 and VKORC1 genetic polymorphism and patient characteristics upon warfarin dose requirements: proposal for a new dosing regimen. *Blood 106*(7), 2329–2333.

Seibold, H., A. Zeileis, and T. Hothorn (2016). Model-based recursive partitioning for subgroup analyses. *The International Journal of Biostatistics 12*(1), 45–63.

Shen, J. and X. He (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *Journal of the American Statistical Association 110*(509), 303–312.

Shen, J. and A. Qu (2020). Subgroup analysis based on structured mixed-effects models for longitudinal data. *Journal of Biopharmaceutical Statistics 30*(4), 607–622.

Stack, G. and C. B. Maurice (2016). Warfarin pharmacogenetics reevaluated: Subgroup analysis reveals a likely underestimation of the maximum pharmacogenetic benefit by clinical trials. *American Journal of Clinical Pathology 145*(5), 671–686.

Städler, N., P. Bühlmann, and S. Van De Geer (2010). $\ell 1$-penalization for mixture regression models. *Test 19*(2), 209–256.

Van Dat, N., P. Van Toan, and T. M. Thanh (2022). Solving distribution problems in content-based recommendation system with gaussian mixture model. *Applied Intelligence 52*(2), 1602–1614.

van der Vliet, R., R. W. Selles, E.-R. Andrinopoulou, R. Nijland, G. M. Ribbers, M. A. Frens, and et al. (2020). Predicting upper limb motor impairment recovery after stroke: A mixture model. *Annals of Neurology 87*(3), 383–393.

Wang, Y. (2016). *Logistic-normal mixtures with heterogeneous components and high dimensional covariates.* Ph. D. thesis, University of Michigan.

Wang, Y. and D. M. Blei (2019). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association 114*(527), 1147–1161.

Westling, T. and T. H. McCormick (2019). Beyond prediction: A framework for inference with variational approximations in mixture models. *Journal of Computational and Graphical Statistics 28*(4), 778–789.

Wojnowicz, M. T., S. Aeron, E. L. Miller, and M. Hughes (2022). Easy variational inference for categorical models via an independent binary approximation. *Proceedings of the 39th International Conference on Machine Learning 162*, 23857–23896.

Yang, Y., D. Pati, and A. Bhattacharya (2020). α-variational inference with statistical guarantees. *The Annals of Statistics 48*(2), pp. 886–905.

Yang, Y., M. J. Wainwright, and M. I. Jordan (2016). On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics 44*(6), 2497 – 2532.

Zhang, C.-X., S. Xu, and J.-S. Zhang (2019). A novel variational Bayesian method for variable selection in logistic regression models. *Computational Statistics & Data Analysis 133*, 1–19.

Zhang, F. and C. Gao (2020). Convergence rates of variational posterior distributions. *The Annals of Statistics 48*(4), 2180 – 2207.

Zhang, R., N. N. Narisetty, X. He, and J. Shen (2025). Bayesian variable selection on structured logistic-normal mixture models for subgroup analysis. *Electronic Journal of Statistics 19*(1), 2876–2922.

Zhao, L., L. Tian, T. Cai, B. Claggett, and L. J. Wei (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association 108*(502), 527–539.

Ruqian Zhang

Department of Statistics and Data Science, Fudan University, Shanghai, China

E-mail: rqzhang20@fudan.edu.cn

Juan Shen

Department of Statistics and Data Science, Fudan University, Shanghai, China

E-mail: shenjuan@fudan.edu.cn