The New England Journal of Statistics in Data Science Volume 4, 90–104 (2026)
DOI: https://doi.org/10.51387/26-NEJSDS102

Consistent and Scalable Variable Selection with Robust Link Functions

Eric ODOOM and Xia WANG*

Abstract

This study explores the application of the t-link model in high-dimensional variable selection for binary regression. The t-link model provides flexibility in binary modeling and offers robust inference in the presence of outliers, making it a preferable alternative to the commonly used probit and logit links. To address the computational challenges posed by a large number of covariates, the skinny Gibbs algorithm is employed, and the consistency of variable selection under this approximate algorithm is established. These advancements in both computational and theoretical perspectives enhance the practicality and ease of implementing the t-link model. The performance of the t-link model, with a specified degrees of freedom, is compared to logit link and the probit link through simulation studies and an application to PCR data. The results demonstrate the robustness and computational efficiency of the proposed method.

KEYWORDS AND PHRASES: Binary regression, Link functions, Robustness; skinny Gibbs, Spike-and-slab prior.

1. INTRODUCTION

Binary regression models are essential tools in statistical modeling, particularly for understanding relationships between binary outcomes and predictor variables. Traditional approaches often rely on the logit or probit link functions due to their theoretical appeal and interpretability. However, these standard link functions may fail when the sigmoid function of the underlying data exhibits heavy tails or extreme responses [10, 21, 29], as often encountered in genetics and biomedical research. For example, binary outcomes in genetic studies, such as the presence or absence of a specific disease, are influenced by complex biological processes that standard links may inadequately capture. Moreover, inference results from these models may become unreliable in the presence of outlying covariates or incoherent responses [22, 24, 40]. Addressing these limitations requires exploring alternative approaches that better account for the complexities of such data.

In this context, the t-distribution emerges as an attractive alternative due to its flexibility in modeling heavy-tailed data, controlled by the degrees of freedom parameter v. This study aims to further investigate the t-link model's potential to enhance robustness and mitigate sensitivity to outliers. Specifically, we explore its application to high-dimensional variable selection, a scenario that has received little, if any, attention in the current literature.

Bayesian variable selection is a theoretically intricate and computationally demanding problem. These challenges stem from the balance required between numerical computation, prior specification, and analytical evaluation, as highlighted in earlier works [18, 12, 35]. High-dimensional variable selection introduces additional computational burdens due to the increasing number of covariates. To address this, [32] proposed the skinny Gibbs sampler using the continuous spike-and-slab prior introduced by [31], demonstrating its consistency in selecting the true model in logistic regression. [38] extended this algorithm to the probit model, further proving its theoretical properties.

Building on these earlier work, we propose the hierarchical skinny Gibbs sampler with t-link (HSGT). The HSGT method extends the skinny Gibbs sampler framework by incorporating the t-distribution in the link function. Furthermore, we enhance the algorithms of [32] and [38] by introducing a hierarchical prior on the variance component of active covariates. This update eliminates the time-consuming and problematic tuning process required in previous approaches. The advantages of our method are twofold: (1) the t-link provides a flexible and robust family of link models, with probit and logit as special cases, enabling a more accurate representation of underlying biological and medical phenomena. Theoretically, the probit link can be recovered as $v \to \infty$, while the logit link can be approximated by a t-link with a specific value of v [1, 32]. And its robustness in estimation and inference under outliers - both in covariates and responses - is well established in the literature [40]. and (2) the skinny Gibbs sampler ensures computational efficiency and scalability to large datasets, a critical requirement in genomics and other high-dimensional fields.

The paper is organized as follows. In Section 2, we introduce the model for variable selection in t-link binary regression and discuss the hierarchical Gibbs algorithms, including the exact Gibbs (HEGT) and the skinny Gibbs (HSGT)

*Corresponding author.

methods. The theoretical foundations of the HSGT method, along with consistency results, are detailed in Section 3, with comprehensive proofs provided in Appendix A. Simulation studies evaluating model selection under various scenarios are presented in Section 4, with a particular focus on comparing the performance of the t-link and probit link models, both with and without outliers. In Section 5, we apply the proposed method to a genomic dataset, providing a thorough comparison in prediction accuracy among models. Section 6 concludes with a discussion.

2. METHODOLOGY

We consider a scenario of binary classification in high-dimensional analysis, specifically focusing on a t-link function to model binary response in medicine or genomics studies where outliers and heavy-tailed distributions are observed. For instance, identifying key genetic variants (SNPs) associated with the presence or absence of a disease (e.g., cancer). Let $E_i \in \{0,1\}$ represent the binary response variable, where $E_i = 1$ indicates that the ith subject has the disease, and $E_i = 0$ otherwise, $i = 1,\ldots,n$. The covariate vector $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})' \in \mathcal{R}^p$ contains the expression levels of p genes for the ith subject. The probability of $E_i = 1$ is modeled through the generalized liner regression model framework as

$$P(E_i = 1 \mid x_i, \boldsymbol{\beta}, v) = \Psi(\boldsymbol{x}_i'\boldsymbol{\beta}, v), \tag{2.1}$$

where $\Psi(\cdot)$ is the cumulative distribution function of a distribution, such as the standard normal distribution for the probit link and the standard t-distribution for the t-link. The parameter v may be a scalar or a vector of parameters that is fixed or estimated in the regression model. In the t-link regression, v represents the degrees of freedom. The regression coefficients $\boldsymbol{\beta}$ is a $p \times 1$ vector, which satisfies $\|\boldsymbol{\beta}\|_0 << p$, meaning that only a small subset of the components of $\boldsymbol{\beta}$ are non-zero. This sparsity assumption is especially relevant in high-dimensional contexts where the number of predictors p is much larger than the sample size n.

The observed-data likelihood function of all n observations is given as

$$L_n\left(\boldsymbol{\beta} \mid \boldsymbol{E}, \boldsymbol{X}, v\right) = \prod_{i=1}^{n} \left[\Psi(\boldsymbol{x}_i'\boldsymbol{\beta},\ v)\right]^{E_i} \left[1 - \Psi(\boldsymbol{x}_i'\boldsymbol{\beta},\ v)\right]^{(1-E_i)}. \tag{2.2}$$

Using the latent variable approach [1] and representing the t-distribution as a scale mixture of Gaussian, the complete-data likelihood function is expressed as:

$$\begin{aligned} &L(\boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{\omega} \mid \boldsymbol{E}, \boldsymbol{X}, v) \\ &\quad = \prod_{i=1}^{n} \mathcal{N}(Z_i \mid \boldsymbol{x}_i'\boldsymbol{\beta}, \omega_i^2)\mathcal{IG}(\omega_i^2 \mid v/2, v/2) \times \\ &\qquad \Big(E_i \mathcal{I}(Z_i \geq 0) + (1 - E_i)\mathcal{I}(Z_i < 0)\Big), \end{aligned} \tag{2.3}$$

where $Z_i|\omega_i^2$ is a normal with mean $\boldsymbol{x}_i'\boldsymbol{\beta}$ and variance ω_i^2 and ω_i^2 has an inverse-gamma distribution with the scale and the shape as $v/2$. The aim is to efficiently infer the sparsity pattern encoded in the regression coefficients $\boldsymbol{\beta}$ through a hierarchical Bayesian model with consistency guarantees. Throughout this methodological exposition, we assume that the columns of $\boldsymbol{X}$ are standardized to have zero mean and unit variance.

2.1 Variable Selection with Spike-and-Slab Priors

The spike-and-slab prior is used to induce sparsity on $\boldsymbol{\beta}$ [17, 25]. This approach employs a binary latent vector $\boldsymbol{\gamma}$ of dimension p, corresponding to the number of covariates, where each element γ_j indicates whether the jth covariate is active ($\gamma_j = 1$) or inactive ($\gamma_j = 0$), $j = 1,\ldots,p$. The prior distribution for the regression coefficients reflects this status: a "spike" component centered near zero when $\gamma_j = 0$, or a "slab" component with a diffuse probability density when $\gamma_j = 1$. Building on this framework, [33] demonstrated that allowing the variances of the spike-and-slab components to depend on the sample size ensures that, as the sample size grows, the posterior probability of selecting the true model converges to one.

Let $\boldsymbol{\beta}_k$ be the set of active covariates and $\boldsymbol{\beta}_{k^c}$ the set of inactive covariates such that $k = \{j : \mathcal{I}(\gamma_j \neq 0)\ \}$. Then $\sum_{j=1}^{p} \gamma_j = |k|$ is the number of non-zero entries in $\boldsymbol{\gamma}$. Similarly, let $\boldsymbol{X}_k$ represent a sub-matrix of $\boldsymbol{X}$ with columns corresponding to the nonzero indices of $\boldsymbol{\gamma}$ while $\boldsymbol{X}_{k^c}$ includes the remaining columns associated with $\gamma_j = 0$. The priors for the binary latent variables γ_j and the corresponding regression coefficients β_j, $j = 1,\ldots,p$, are specified as follows:

$$\beta_j \mid \gamma_j = 0 \sim N\left(0, \tau_0^2\right), \tag{2.4}$$

$$\beta_j \mid \gamma_j = 1 \sim N\left(0, \tau_1^2\right), \tag{2.5}$$

$$\pi\left(\gamma_j = 1\right) = 1 - \pi\left(\gamma_j = 0\right) = q, \tag{2.6}$$

$$\tau_1^2 \mid r, s \sim \mathcal{IG}(r, s), \tag{2.7}$$

for some positive constants τ_0^2, r, s and $q \in (0,1)$. The hyperparameters r and s are the shape and scale parameters of the inverse gamma prior on τ_1^2. The parameter τ_0^2 is the variance of the spike part and τ_1^2 controls the variance of the slab part. The prior probability q is the probability that $\beta_j \neq 0$. The rates determining τ_0^2 and q are as specified in [33, 32] and for asymptotic considerations (i.e. $n\tau_0^2 = o(1)$, and $q \sim p^{-1}$).

2.2 Hierarchical Exact Gibbs Sampler (HEGT)

The full joint posterior distributions of $\boldsymbol{\beta}$, $\boldsymbol{Z}$, $\boldsymbol{\gamma}$, $\boldsymbol{\omega}$ and τ_1^2 conditioned on $\mathbf{X}$, $\mathbf{E}$ and v using the complete-data likelihood in (2.3) and the spike-and-slab prior specified in (2.4)-(2.7) is given as

$$\begin{aligned}
&\pi(\boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \tau_1^2 \mid \boldsymbol{E}, \boldsymbol{X}, v) \\
&\quad \propto L(\boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{\omega} \mid \boldsymbol{E}, \boldsymbol{X}, v) \cdot \pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \tau_1^2) \cdot \pi(\boldsymbol{\gamma}) \cdot \pi(\tau_1^2) \\
&\quad \propto \exp\left\{-\frac{1}{2}\Big[(\mathbf{Z}-\boldsymbol{X}\boldsymbol{\beta})'\mathbf{W}(\mathbf{Z}-\boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\beta}'\boldsymbol{D}\boldsymbol{\beta}\Big]\right\} \times \\
&\qquad \prod_{i=1}^{n}\Big(E_i\mathcal{I}(Z_i \geq 0) + (1-E_i)\mathcal{I}(Z_i < 0)\Big) \times \\
&\qquad \prod_{i=1}^{n}\left(\left(\frac{1}{\omega_i^2}\right)^{\frac{v+3}{2}} \exp\left(-\frac{v}{2\omega_i^2}\right)\right)\left[\frac{1-q}{\tau_0\sqrt{2\pi}}\right]^p \times \\
&\qquad \left[\frac{\tau_0 q}{\tau_1(1-q)}\right]^{|\boldsymbol{k}|}\left(\left[\frac{1}{\tau_1^2}\right]^{r+1}\exp\left(-\frac{s}{\tau_1^2}\right)\right),
\end{aligned} \tag{2.8}$$

where $\mathbf{W} = \text{Diag}\left(1/\omega_1^2, \ldots, 1/\omega_n^2\right)$, $|\boldsymbol{k}| = \sum_{j=1}^{p}\gamma_j$ and $\mathbf{D} = \text{Diag}\left(\tau_0^{-2}(\mathbf{1}-\boldsymbol{\gamma}) + \tau_1^{-2}\boldsymbol{\gamma}\right)$.

(i) The conditional posterior distribution of γ_j is

$$\gamma_j|\cdot \sim \text{Bernoulli}\left(\frac{d_j}{1+d_j}\right), \quad j = 1, \ldots, p,$$

where

$$d_j = \frac{\pi(\gamma_j = 1 \mid \boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{E})}{\pi(\gamma_j = 0 \mid \boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{E})} = \frac{q\phi\left(\beta_j;\ 0, \tau_1^2\right)}{(1-q)\phi\left(\beta_j;\ 0, \tau_0^2\right)}.$$

(ii) The conditional posterior distribution of ω_i^2 is

$$\omega_i^2 \mid \cdot \sim \mathcal{IG}\left(\frac{1+v}{2}, \quad \frac{v + (Z_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2}{2}\right).$$

(iii) The conditional posterior distribution of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \mid \cdot \sim \mathcal{N}_p\left(\boldsymbol{V^{-1}X'WZ},\ \boldsymbol{V}^{-1}\right),$$

where $\mathbf{V} = \mathbf{X}'W\mathbf{X} + \mathbf{D}$.

(iv) The conditional posterior of Z_i is

$$\pi(Z_i \mid \cdot) \propto \begin{cases} \mathcal{N}(Z_i;\ \boldsymbol{x}_i'\boldsymbol{\beta},\ \omega_i^2)\mathcal{I}(Z_i \geq 0), & \text{if } E_i = 1, \\ \mathcal{N}(Z_i;\ \boldsymbol{x}_i'\boldsymbol{\beta},\ \omega_i^2)\mathcal{I}(Z_i < 0), & \text{if } E_i = 0. \end{cases}$$

(v) The conditional posterior distribution of τ_1^2 is

$$\tau_1^2 \mid \cdot \sim \mathcal{IG}\left(r + \frac{|\boldsymbol{k}|}{2},\ s + \frac{\sum_{j=1}^{p}\gamma_j\beta_j^2}{2}\right).$$

2.3 Hierarchical Skinny Gibbs Sampler(HSGT)

The computational complexity of the above algorithm is primarily driven by sampling $\boldsymbol{\beta}$ in step (iii), which involves handling a $p \times p$ precision matrix $\boldsymbol{V}$. Calculating its inverse $(\boldsymbol{V}^{-1})$ has a complexity of $O(p^3)$, while computing the product $\boldsymbol{V}^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Z}$ introduces an additional complexity of $O(p^2 n)$. As a result, the total computational cost of HEGT is $O(p^2(p \vee n))$, which increases rapidly with p. This high complexity makes the algorithm computationally expensive in terms of both time and memory, particularly within the context of an Markov chain Monte Carlo (MCMC) framework. [32] introduces the skinny Gibbs algorithm to handle this computational bottleneck. The skinny Gibbs algorithm is a simple yet effective modification of the Gibbs sampler, designed for high-dimensional settings with many predictors. At each MCMC iteration, $\boldsymbol{\beta}$ is divided into two components: the active component $(\boldsymbol{\beta}_k)$ and the inactive component $(\boldsymbol{\beta}_{k^c})$. The sparsity is imposed by structuring the precision matrix $(\boldsymbol{V})$ of $\boldsymbol{\beta}$ as a two-block diagonal matrix, assuming full independence between the active and inactive components of $\boldsymbol{\beta}$. That is,

$$\begin{aligned}
\boldsymbol{V} &= \begin{pmatrix} \boldsymbol{X}_k'\boldsymbol{W}\boldsymbol{X}_k + \tau_1^{-2}\boldsymbol{I}_{|\boldsymbol{k}|} & \boldsymbol{X}_k'\boldsymbol{W}\boldsymbol{X}_{k^c} \\ \boldsymbol{X}_{k^c}'\boldsymbol{W}\boldsymbol{X}_k & \mathbf{X}_{k^c}'\boldsymbol{X}_{k^c} + \tau_o^{-2}\boldsymbol{I}_{(p-|\boldsymbol{k}|)} \end{pmatrix} \\
&\approx \begin{pmatrix} \boldsymbol{X}_k'\boldsymbol{W}\boldsymbol{X}_k + \tau_1^{-2}\boldsymbol{I}_{|\boldsymbol{k}|} & \mathbf{0} \\ \mathbf{0} & (n-1+\tau_o^{-2})\boldsymbol{I}_{(p-|\boldsymbol{k}|)} \end{pmatrix}.
\end{aligned}$$

The joint skinny posterior of the skinny $\boldsymbol{\beta}$, $\boldsymbol{Z}$ $\boldsymbol{\omega}$, $\boldsymbol{\gamma}$ and τ_1^2 conditioned on $\boldsymbol{X}$, $\boldsymbol{E}$ and v is given as

$$\begin{aligned}
&\pi(\boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \tau_1^2 \mid \boldsymbol{X}, \boldsymbol{E}, v) \\
&\quad \propto L(\boldsymbol{\beta}_k,\ \boldsymbol{Z}, \boldsymbol{\omega} \mid \boldsymbol{E},\ \boldsymbol{X}_k, v) \cdot \pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \tau_1^2) \cdot \pi(\boldsymbol{\gamma})\pi(\tau_1^2) \\
&\quad \propto \exp\left\{-\frac{1}{2}\left((\boldsymbol{Z}-\boldsymbol{X_k}\boldsymbol{\beta_k})'\boldsymbol{W}(\boldsymbol{Z}-\boldsymbol{X_k}\boldsymbol{\beta_k}) + \frac{\boldsymbol{\beta}_k'\boldsymbol{\beta}_k}{\tau_1^2}\right)\right\} \\
&\qquad \times \exp\left\{-\frac{1}{2}\left(n-1+\tau_0^{-2}\right)\boldsymbol{\beta}_{k^c}'\boldsymbol{\beta}_{k^c}\right\}\left[\frac{(1-q)\tau_1}{q\tau_0}\right]^{-|\boldsymbol{k}|} \times \\
&\qquad \prod_{i=1}^{n}\Big(E_i\mathcal{I}(Z_i \geq 0) + (1-E_i)\mathcal{I}(Z_i < 0)\Big) \times \\
&\qquad \prod_{i=1}^{n}\left(\left(\frac{1}{\omega_i^2}\right)^{\frac{v+3}{2}}\exp\left(-\frac{v}{2\omega_i^2}\right)\right)\left[\frac{1}{\tau_1^2}\right]^{r+1}\exp\left\{-\frac{s}{\tau_1^2}\right\}.
\end{aligned} \tag{2.9}$$

(i) The conditional posterior distribution of $\boldsymbol{\beta}$ is

$$\begin{aligned}
\boldsymbol{\beta}_k \mid \cdot &\sim \mathcal{N}_{|\boldsymbol{k}|}\left(\boldsymbol{V}_k^{-1}\boldsymbol{X}_k'\boldsymbol{W}\boldsymbol{Z},\ \mathbf{V}_k^{-1}\right), \\
\boldsymbol{\beta}_{k^c} \mid \cdot &\sim \mathcal{N}_{p-|\boldsymbol{k}|}\left(0,\ \boldsymbol{V}_{k^c}^{-1}\right),
\end{aligned}$$

where $\boldsymbol{V}_k = \boldsymbol{X}_k'\boldsymbol{W}\boldsymbol{X}_k + \tau_1^{-2}\boldsymbol{I}_{|\boldsymbol{k}|}$ and $\boldsymbol{V}_{k^c} = (n-1+\tau_0^{-2})\boldsymbol{I}_{p-|\boldsymbol{k}|}$.

Comparing to Step (iii) in Section 2.2, the computation here has a much lower complexity of $O(n(p \vee |\boldsymbol{k}|^2))$. This reduction not only minimizes memory usage, but also

improves computational efficiency, making the MCMC algorithm more practical and desirable for high-dimensional data settings.

(ii) The conditional posterior distribution of ω_i^2 is

$$\omega_i^2 \mid \cdot \sim \mathcal{IG}\left(\frac{1+v}{2}, \quad \frac{v+(Z_i - \boldsymbol{x}'_{ik}\boldsymbol{\beta}_k)^2}{2}\right),$$

for $i = 1, \ldots, n$.

(iii) The conditional posterior distribution of Z_i is

$$\pi(Z_i \mid \cdot) \propto \begin{cases} \mathcal{N}(Z_i;\, \boldsymbol{x}'_{i,k}\boldsymbol{\beta}_k,\, \omega_i^2)\mathcal{I}(Z_i \geq 0), & \text{if } E_i = 1, \\ \mathcal{N}(Z_i;\, \boldsymbol{x}'_{i,k}\boldsymbol{\beta}_k,\, \omega_i^2)\mathcal{I}(Z_i < 0), & \text{if } E_i = 0, \end{cases}$$

for $i = 1, \ldots, n$.

(iv) The conditional posterior distribution of γ_j is

$$\gamma_j | \cdot \sim \text{Bernoulli}\left(\frac{\tilde{d}_j}{1+\tilde{d}_j}\right), \quad j = 1, \ldots, p,$$

where

$$\begin{aligned} \tilde{d}_j &= \frac{\pi(\gamma_j = 1 \mid \boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{E})}{\pi(\gamma_j = 0 \mid \boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{E})} \\ &= d_j \cdot \exp\left\{\frac{1}{2} X_j'(\boldsymbol{I} - \boldsymbol{W})\boldsymbol{X}_j\beta_j^2 + \right. \\ &\quad \left. (\boldsymbol{Z} - \boldsymbol{\beta}'_{k_{-j}}\boldsymbol{X}'_{k_{-j}})\boldsymbol{W}\boldsymbol{X}_j\beta_j\right\}. \end{aligned}$$

(v) The conditional posterior distribution of τ_1^2 is

$$\tau_1^2 \mid \cdot \sim \mathcal{IG}\left(r + \frac{|\boldsymbol{k}|}{2},\; s + \frac{\sum_{j=1}^p \gamma_j \beta_j^2}{2}\right).$$

3. THEORETICAL PROPERTIES

In this section, we present the asymptotic properties of the hierarchical prior setup of the algorithm on binary regression with the student t-link. Let $\boldsymbol{t} = \{t_1, t_2, \ldots, t_t\} \subseteq [p]$ represent the true model, indicating that the non-zero locations of the true coefficient vector correspond to $\boldsymbol{t} = (j,\ j \in \boldsymbol{t})$. We assume $|\boldsymbol{t}|$ is fixed. Let $\boldsymbol{\beta}_0 \in \mathcal{R}^p$ be the true coefficient vector, and $\boldsymbol{\beta}_{0,t} \in \mathcal{R}^{|\boldsymbol{t}|}$ represent the vector of true non-zero coefficients.

For a given model k with active components indexed by $\boldsymbol{k} \subseteq [p]$ with size $|\boldsymbol{k}|$ at most m_n, we define $\ell_n(\boldsymbol{\beta}_k)$ as the log-likelihood and $S_n(\boldsymbol{\beta}_k) = \partial\ell_n(\boldsymbol{\beta}_k)/\partial\boldsymbol{\beta}_k$ as the score function. The restriction on the largest model size is a common practice in Bayesian high-dimensional asymptotic theory to effectively control the posterior ratio and achieve desirable consistency results. The variable selection consistency by the skinny Gibbs has been shown under the logit link [32] and the probit link [38], respectively. Here, we extend the consistency results to a t-link model with a known degrees of freedom.

The Hessian matrix of $\ell_n(\boldsymbol{\beta}_k)$ is defined as

$$\begin{aligned} H_n(\boldsymbol{\beta}_k) = \; -\tfrac{\partial^2\ell_n}{\partial\boldsymbol{\beta}_k\partial\boldsymbol{\beta}_k'} \; &= \sum_{i=1}^n \boldsymbol{x}_{i,k}\boldsymbol{x}'_{i,k} w_i(\boldsymbol{\beta}_k) \\ &= \boldsymbol{X}'_k \boldsymbol{W}(\boldsymbol{\beta}_k)\boldsymbol{X}_k, \end{aligned}$$

where $\boldsymbol{W}(\boldsymbol{\beta}_k) = \text{Diag}\{w_i(\boldsymbol{\beta}_k)\}$ for $i = 1, \ldots, n$ with $w_i(\boldsymbol{\beta}_k)$ as

$$E_i\left\{\frac{(v+1)}{v}\left(1 + \frac{\mu_{i,k}^2}{v}\right)^{-1} \mu_{i,k}\; r_{i,k} + r_{i,k}^2\right\} +$$
$$(1 - E_i) \cdot \left\{-\frac{(v+1)}{v}\left(1 + \frac{\mu_{i,k}^2}{v}\right)^{-1} \mu_{i,k}\; R_{i,k} + R_{i,k}^2\right\}$$

with $r_{i,k} = \psi(\mu_{i,k},\ v)/\Psi(\mu_{i,k},\ v)$ and $R_{i,k} = \psi(\mu_{i,k},\ v)/[1 - \Psi(\mu_{i,k},\ v)]$, $\psi(\mu_{i,k},\ v)$ and $\Psi(\mu_{i,k},\ v)$ as the density and the cumulative probability functions evaluated at $\mu_{i,k}$ of a standard t distribution with v degrees of freedom, and $\mu_{i,k} = \boldsymbol{x}_{i,k}\boldsymbol{\beta}_{i,k}$.

Before presenting the main results, we introduce the following notation: (i) for any $x, y \in \mathbb{R}$, $x \vee y$ and $x \wedge y$ represent the maximum and minimum of x and y, respectively. For any positive real sequences x_n and y_n, (ii) $x_n \lesssim y_n$ or $x_n = O(y_n)$ implies there exists $C > 0$ a constant such that $|x_n| \leq C|y_n|$ for all large n. (iii) $x_n \ll y_n$ or $x_n = o(y_n)$ if $x_n/y_n \to 0$ as $n \to \infty$. (iv) If $x_n \sim y_n$, there exists $K_1 > K_2 > 0$ such that $K_2 < y_n/x_n \leq x_n/y_n < K_1$.

Let $\boldsymbol{a} = (a_1, a_2, \ldots, a_p)' \in \mathbb{R}^p$ then $\|\boldsymbol{a}\|_2$ and $\|\boldsymbol{a}\|_1$ is the ℓ_2 and ℓ_1 norm of a respectively and $\|\boldsymbol{a}\|_{\max} = \max_{1\leq j\leq p}|a_j|$. For any real symmetric matrix $\boldsymbol{M}$, $\lambda_{\max}(\boldsymbol{M})$ and $\lambda_{\min}(\boldsymbol{M})$ are the maximum and minimum eigenvalues of $\boldsymbol{M}$, respectively. The following conditions are assumed to obtain the asymptotic results:

Condition (C1) (Condition on dimension (n and p)) For some constant $0 < d' < 1$, let $p \gg n$, $\log\ p = o(n)$ and $m_n = O\Big((n/\log p)^{\frac{1-d'}{2}} \wedge p\Big)$, as $n \to \infty$.

Condition (C1) guarantees that our proposed method is capable of handling high-dimensional setting analysis where the number of predictors increases at a sub-exponential rate relative to n. The parameter m_n restricts the analysis to a set of sufficiently large models. Also, to avoid over-fitting, it is required that $m_n \ll n$. Similar assumptions regarding model size limitations are frequently encountered in the sparse estimation literature [14, 34, 28].

Condition (C2) (Condition on Design Matrix $\boldsymbol{X}$) For $\boldsymbol{x_i} \in \mathcal{R}^p, \quad i = 1, 2, \ldots, n$.

i. (Boundedness) $P(||\boldsymbol{x_i}||_{max} \leq C_0) = 1$ for some $C_0 > 0$ and $P(\cdot)$ is the probability measure.
ii (Sub-Gaussianity) for every $\alpha \in \mathcal{R}^p$, there exists a constant $C_\alpha > 0$ such that $\bar{E}\exp(\boldsymbol{x}_i'\boldsymbol{\alpha}) \leq \exp(C_\alpha||\boldsymbol{\alpha}||_2^2)$

The condition (C2) [i] requires each component of $\boldsymbol{x}_i$ is bounded with probability 1 as adopted by [32] for a deterministic design matrix. The condition [ii] of (C2) implies that a linear combination of $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ has a sufficiently light tail satisfying sub-Gaussianity. See similar conditions under the logistic regression in [32, 8].

Condition (C3) (Conditions on True Model) For some constant $c_0 > 0$,

$$\min_{j \in t} \|\beta_{0,j}\|_2^2 \geq c_0 |\boldsymbol{t}| \Lambda_{|\boldsymbol{t}|} \Big(\frac{\log p}{n} \vee \frac{1}{\log p} \Big),$$

and $\Lambda_{|\boldsymbol{t}|} = \max_{\boldsymbol{t}:\, |\boldsymbol{t}| \leq c_0} \lambda_{\max}(n^{-1} \boldsymbol{X}_k' \boldsymbol{X}_k)$ for any integer $c_0 > 0$.
This condition (also known as *beta-min* condition) guarantees true regression coefficients $\boldsymbol{\beta}_0$ have a finite number of nonzero entries and a bounded ℓ_1-norm. It holds if we assume that the true inactive beta components ($\boldsymbol{\beta}_{t^c}$) can be nonzero but $\|\boldsymbol{X}_{t^c} \boldsymbol{\beta}_{t^c}\|_2 = O(1)$ in line with similar assumption in [33].

Condition (C4) By adapting the line of reasoning(Conditions on prior hyperparameters) $n\tau_0^2 = O(1)$ and for some constants $a_1, a_2 > 0$, the hyperparameters satisfy $a_1 < r, s < a_2$ and $q^2 \sim p^{-1}$, where r and s are the shape and scale parameters of the inverse gamma prior on τ_1^2 defined in Section 2.1.
Condition (C4) suggests appropriate conditions for the hyperparameters. Similar assumptions have also been considered in [43], [26] and [9]. We impose a prior on τ_1^2 with its value estimated within the MCMC iterations. This eliminates the need for the tuning procedure required in [32].

Lemma 1 (Pseudoposterior of $\boldsymbol{\gamma}$)**.** *The marginal posterior for any model $\boldsymbol{\gamma}$, $m_k(\boldsymbol{E})$, under the skinny Gibbs is given by*

$$\begin{aligned} m_k(\boldsymbol{E}) = \pi(\boldsymbol{k} \mid \boldsymbol{X}, \boldsymbol{E}) \propto \quad & \left\{2\pi v_n^{-2} (n - 1 + \tau_0^{-2})\right\}^{\frac{|\boldsymbol{k}|}{2}} \times \\ & \int\int (\tau_1^2)^{-\frac{|\boldsymbol{k}|}{2}} \pi(\tau_1^2) \exp\left\{\ell_n(\boldsymbol{\beta}_k)\right\} \\ & \exp\Big(-\tfrac{1}{2\tau_1^2} \|\boldsymbol{\beta}_k\|_2^2 \Big) d\boldsymbol{\beta}_k d\tau_1^2, \end{aligned}$$

where $v_n = (1 - q)/(\tau_0 q)$.

Proof of Lemma 1.

$$\begin{aligned} m_k(\boldsymbol{E}) =& \pi(\boldsymbol{k} \mid \boldsymbol{X}, \boldsymbol{E}) \\ \propto & v_n^{-|\boldsymbol{k}|} \int\int \exp\left\{\ell_n(\boldsymbol{\beta}_k)\right\} \exp\left\{ -\frac{1}{2\tau_1^2} \boldsymbol{\beta}_k' \boldsymbol{\beta}_k \right\} \times \\ & \exp\left\{ -\frac{1}{2} \left(n - 1 + \tau_0^{-2}\right) \boldsymbol{\beta}_{k^c}' \boldsymbol{\beta}_{k^c} \right\} d\boldsymbol{\beta}_{k^c} d\boldsymbol{\beta}_k d\tau_1^2 \\ \propto & v_n^{-|\boldsymbol{k}|} \left\{2\pi \left(n - 1 + \tau_0^{-2}\right)\right\}^{-\frac{p - |\boldsymbol{k}|}{2}} \times \\ & \int\int \exp\left\{\ell_n(\boldsymbol{\beta}_k)\right\} \exp\left(-\frac{1}{2\tau_1^2} \boldsymbol{\beta}_k' \boldsymbol{\beta}_k \right) d\boldsymbol{\beta}_k d\tau_1^2 \\ \propto & \left\{2\pi v_n^{-2} \left(n - 1 + \tau_0^{-2}\right)\right\}^{\frac{|\boldsymbol{k}|}{2}} \times \\ & \int\int \exp\left\{\ell_n(\boldsymbol{\beta}_k)\right\} \exp\left(-\frac{1}{2\tau_1^2} \boldsymbol{\beta}_k' \boldsymbol{\beta}_k \right) d\boldsymbol{\beta}_k d\tau_1^2. \quad \square \end{aligned}$$

Lemma 1 is used to prove Theorem 1 which demonstrates the proposed method guarantees posterior ratio consistency.

3.1 Model Selection Consistency

Theorem 1 (Posterior ratio consistency)**.** *Assume Conditions (C1) - (C4) hold, for any $\boldsymbol{k} \neq \boldsymbol{t}$, $PR(\boldsymbol{k}, \boldsymbol{t}) = \frac{\pi(\boldsymbol{k}|\boldsymbol{X},\boldsymbol{E})}{\pi(\boldsymbol{t}|\boldsymbol{X},\boldsymbol{E})}$,*

$$PR(\boldsymbol{k}, \boldsymbol{t}) \xrightarrow{P} 0 \quad \text{as } n \to \infty, \tag{3.1}$$

where $\boldsymbol{t} \subseteq [p]$ is the true model.

Theorem 1 shows that, asymptotically, our posterior avoids over-fitting by not including an excessive number of unnecessary variables. However, this does not guarantee that the posterior will concentrate on the true model. To capture all significant variables, the nonzero entries in $\beta_{0,t}$ must have sufficiently large magnitudes. In the following theorem, we establish that our posterior achieves *strong selection consistency*, meaning that the posterior probability assigned to the true model t converges to 1. This requires a suitable condition on the lower bound for the magnitudes of the nonzero entries in $\boldsymbol{\beta}_{0,t}$. It is important to note that Theorem 1 does not require the smallest nonzero entries to be bounded away from 0. Furthermore, Theorem 2 guarantees *posterior ratio consistency*, which implies that the true model t will become the mode of the posterior as the sample size grows.

Theorem 2 (Strong selection consistency)**.** *Under Conditions (C1)–(C4), the following holds:*

$$\pi\big(\boldsymbol{t} \mid \boldsymbol{E}, \boldsymbol{X}\big) \xrightarrow{P} 1, \quad \text{as } n \to \infty.$$

The proofs for Theorems 1 and 2 are provided in the appendix. The authors in [8] established consistency under hierarchical nonlocal priors in logistic regression, and we adapt their proof strategy to demonstrate that the asymptotic results also hold under spike and slab priors within the skinny Gibbs framework.

The current consistency results are established for variable selection under a specified regression model. That is, under Conditions (C1)–(C4), posterior concentration for $\boldsymbol{\beta}$ holds with all the other components of the regression model fixed, including the random distribution of the response variable and the link function. Allowing the degrees of freedom parameter ν to be unknown incorporates the link function determination into the model selection process. Because of the interplay between the regression coefficients and the link function, additional conditions beyond (C1)–(C4) are required to ensure model identifiability and consistency.

4. SIMULATION

We assess the performance of the t-links, the logit link, and the probit link in variable selection, particularly in the presence of outlying observations. In addition, we evaluate the performance under both the exact and the skinny Gibbs algorithms. For each data replicate, ten models are fitted and compared, including the hierarchal skinny Gibbs t-link model(HSGT), the exact Gibbs t-link model (HEGT), the skinny Gibbs logit model (SLogit)[37], the exact Gibbs logit model (ELogit), the skinny Gibbs probit model (SProbit) and the exact Gibbs probit model (EProbit)[38]. For the t-link models, we consider $\nu = 1$(Cauchy), 3, and 7, respectively, where $\nu = 1$ represents the heaviest tail case among integer values of ν, and $\nu = 7$ is sometimes used as an approximation to the logit link. In addition, three LASSO models with different links are fitted to provide a frequentist benchmark and to further illustrate the robustness of the link functions under alternative estimation methods.

Here we assume ν as known and fix it at a specific value. The proposed algorithm is flexible and can be applied to any value of ν. The "best" ν for a given dataset can be determined ad hoc using standard model selection criteria, such as cross validation, DIC or WAIC. All the theoretical properties are proved for any fixed value of ν.

In the simulation, the first $|\boldsymbol{k}|$ columns of the design matrix $\boldsymbol{X}$ correspond to active covariates with non-zero coefficients, while the remaining columns represent inactive covariates with zero regression coefficients. Each row $\boldsymbol{x}_i$ of $\boldsymbol{X}$ is independently generated from a normal distribution with a p-dimensional covariance matrix $\boldsymbol{\Sigma}$. The binary response $\boldsymbol{E}$ is generated from a Bernoulli distribution with the probability of success $\Psi(\boldsymbol{X\beta}, v)$, where $\Psi(\cdot)$ is the cumulative distribution function of a standard normal for the probit link ($v \to \infty$), a standard logistic distribution for the logit link, and a standard t distribution with the degrees of freedom $v = 1, 3, 7$. We set $|\boldsymbol{k}| = 4$, $n = 100$, $\boldsymbol{\beta} = (3, 1.5, 1.0, 0.5)'$. We vary the true link models Ψ^{-1} (the probit link, the logit link or the Cauchy link ($v = 1$)), the dimension in $\boldsymbol{X}$ ($p = 100$ or 500) and its correlation structure (independent or dependent). For independent $\boldsymbol{X}$, the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{I}$; in the case of correlated covariates, $\boldsymbol{\Sigma}$ is structured into a two-block matrix corresponding to $\boldsymbol{X}_k$ and $\boldsymbol{X}_{k^c}$: with a constant correlation of 0.1 between blocks, and within each block, an AR(1) correlation structure with autocorrelation ρ values of 0.5 across the covariates. Thus, there are a total of 12 scenarios to simulate. A total of 50 data replicates are generated under each scenario.

For each simulated dataset, we also generate two modified datasets identical to the original except for one observation The first dataset is altered to create a "bad leverage point" as described in [22]. This observation is randomly selected from among those with $E_i = 1$ and the value of the first covariate is changed to -10 (note the regression coefficient associated with the first covariate is 3). It is referred to as a "bad leverage point" because it exerts leverage while contradicting the association between the response and explanatory variable observed in the rest of the data. The second dataset creates a "non-leverage outlier" by selecting the observation with the largest magnitude of $E_i - \Psi(\mathbf{x}_i\boldsymbol{\beta})$ and set the value of binary observation as $1 - E_i$. A comparison of the fits from the 10 Bayesian models and three LASSO models, both with and without the outlier, highlights the robustness of the t-link model compared to the probit link.

For all Bayesian models, we set τ_0^2 to $1/n$ and select $q = P(\gamma_i = 1)$ such that $P\left(\sum_{j=1}^{p} \gamma_j = 1 > K_\gamma\right) = 0.1$, with $K_\gamma = \max(10, \log(n))$. As discussed, unlike [32] and [38], the slab variance parameter τ_1^2 is imposed an inverse gamma prior with the hyperparameters s and r set to 2 and 1, respectively, providing a prior mean of 1 and an infinite prior variance. A total of 6, 000 iterations are run for the posterior sampling with a burn-in period of 2,000 iterations.

The metrics used for evaluating the methods are divided into two categories: variable selection performance and prediction performance. The variable selection metrics include sensitivity (SEN), specificity (SPE), and Matthew's correlation coefficient (MCC), defined as follows:

$$\text{SEN} = \text{Sensitivity} = \frac{TP}{TP + FN},$$
$$\text{SPE} = \text{Specificity} = \frac{TN}{TN + FP},$$
$$\text{MCC} = \frac{TP \times TN - FN \times FP}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. The MCC score provides a more comprehensive measure of performance than other metrics, as it incorporates all elements of the confusion matrix into a single summary statistic [13]. In addition, we present the out-of-sample mean square prediction error (MSPE) defined as:

$$\text{MSPE} = \frac{||E - \hat{p}||_2^2}{n},$$

where $\hat{p}$ is the estimated Prob($E = 1$).

Table 1 presents the comparison results, reporting average metrics from simulations based on data generated from the probit link model under two covariate correlation structures with $p = 500$. Results on the remaining 11 scenarios are provided in Tables S1–S5 of the Supplementary Materials.

Across all scenarios without data contamination ("No outlier"), the different link models exhibit similar performance. As expected, the true model generally performs slightly better than the others in terms of variable selection (higher MCC) and prediction accuracy (lower MSPE). Notably, the skinny Gibbs algorithm often outperforms the exact Gibbs algorithm for both the probit and t-link models, consistent with the findings of [32] and [38].

Table 1. Simulation Results: Probit as the True Model ($p = 500$).

	Bayesian										LASSO		
	Cauchy ($v = 1$)		t ($v = 3$)		t ($v = 7$)		logit		probit		Cauchy	logit	probit
	SG	EG	SG	EG	SG	EG	SG	EG	SG	EG			
$\Sigma = I$													
						No Outlier							
SEN	0.6100	0.5850	0.6000	0.6100	0.6400	0.5900	0.6300	0.6050	0.6400	0.6000	0.7450	0.6850	0.6800
SPE	0.9995	0.9995	0.9996	0.9997	0.9995	0.9996	0.9994	0.9998	1.0000	0.9998	0.9729	0.9898	0.9878
MCC	0.7435	0.7316	0.7432	0.7574	0.7673	0.7386	0.7562	0.7621	0.7772	0.7574	0.4188	0.5590	0.5510
FP	0.2600	0.2600	0.2000	0.1600	0.2600	0.1800	0.2800	0.0800	0.1600	0.1200	13.4200	5.0600	5.9000
FN	1.6600	1.6600	1.6000	1.5600	1.4400	1.6400	1.4800	1.5800	1.4400	1.6000	1.0200	1.2600	1.2800
MSPE	0.0994	0.1034	0.0923	0.0994	0.0930	0.1027	0.0916	0.0947	0.0902	0.1022	0.1233	0.1222	0.1194
						Bad Leverage Outlier							
SEN	0.6050	0.5750	0.5900	0.5600	0.5600	0.3950	0.5400	0.4200	0.4750	0.3550	0.7100	0.0400	0.0000
SPE	0.9999	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9982	1.0000	0.9864	1.0000	1.0000
MCC	0.7647	0.7427	0.7564	0.7374	0.7358	0.5946	0.7228	0.6098	0.5236	0.5260	0.5381	0.0626	0.0000
FP	0.0400	0.0800	0.0400	0.0400	0.0400	0.0400	0.0400	0.0600	1.4800	0.2200	6.7600	0.0000	0.0000
FN	1.5800	1.7000	1.6400	1.7600	1.7600	2.4200	1.8400	2.3200	2.1000	2.5800	1.1600	3.8400	4.0000
MSPE	0.0943	0.0956	0.0918	0.0984	0.0976	0.1480	0.1043	0.1377	0.1940	0.1838	0.1383	0.2472	0.2500
						Non-leverage Outlier							
SEN	0.5450	0.5650	0.5400	0.4850	0.4950	0.4700	0.5000	0.4600	0.4850	0.4500	0.6600	0.4900	0.4450
SPE	0.9998	0.9998	0.9995	0.9998	0.9994	0.9994	0.9996	0.9995	0.9992	1.0000	0.9849	0.9978	0.9988
MCC	0.7189	0.7291	0.7122	0.6680	0.6583	0.6414	0.6774	0.6356	0.6150	0.6354	0.4916	0.5978	0.5974
FP	0.1000	0.1200	0.2400	0.1200	0.2800	0.3000	0.1800	0.2600	0.7000	0.2200	7.4800	1.2800	0.7800
FN	1.8200	1.7400	1.8400	2.0600	2.0200	2.1200	2.0000	2.1600	2.0600	2.2000	1.3600	2.0400	2.2200
MSPE	0.1001	0.0998	0.1014	0.1146	0.1138	0.1229	0.1073	0.1176	0.1236	0.1250	0.1415	0.1562	0.1652
$\Sigma = \Sigma_{ar1(\rho=0.50)}$													
						No Outlier							
SEN	0.5600	0.5400	0.5500	0.5350	0.6100	0.5650	0.5850	0.5700	0.5950	0.5450	0.8700	0.8450	0.8200
SPE	0.9997	0.9997	0.9997	0.9997	0.9996	0.9996	0.9997	0.9997	1.0000	1.0000	0.9690	0.9828	0.9846
MCC	0.7238	0.7116	0.7225	0.7088	0.7530	0.7245	0.7400	0.7355	0.7548	0.7322	0.4192	0.5624	0.5790
FP	0.1600	0.1600	0.1600	0.1400	0.2000	0.1800	0.1600	0.1400	0.1200	0.0400	15.3800	8.4200	7.7200
FN	1.7600	1.8400	1.8000	1.8600	1.5600	1.7400	1.6600	1.7200	1.6200	1.8200	0.5200	0.6200	0.7200
MSPE	0.0662	0.0689	0.0653	0.0719	0.0638	0.0719	0.0650	0.0672	0.0616	0.0702	0.0785	0.0806	0.0816
						Bad Leverage Outlier							
SEN	0.5600	0.5300	0.5450	0.4650	0.5100	0.4000	0.5150	0.4300	0.5000	0.4200	0.8400	0.6200	0.1500
SPE	0.9997	0.9998	0.9998	0.9995	0.9988	0.9994	0.9995	0.9996	0.9996	1.0000	0.9888	0.9958	1.0000
MCC	0.7257	0.7115	0.7187	0.6447	0.6384	0.5788	0.6762	0.6211	0.6108	0.6104	0.6381	0.6844	0.2308
FP	0.1600	0.0800	0.1200	0.2400	0.6200	0.3200	0.2400	0.1800	0.7600	0.2000	5.5800	2.0832	0.0200
FN	1.7600	1.8800	1.8200	2.1400	1.9600	2.4000	1.9400	2.2800	2.0000	2.3200	0.6400	1.4400	3.4000
MSPE	0.0715	0.0783	0.0765	0.1133	0.1255	0.1491	0.1169	0.1365	0.1394	0.1454	0.1090	0.1694	0.2344
						Non-leverage Outlier							
SEN	0.5700	0.5300	0.5550	0.4750	0.5200	0.3700	0.5100	0.4200	0.4450	0.3750	0.8500	0.7050	0.6300
SPE	0.9994	0.9998	0.9994	0.9999	0.9996	0.9998	0.9997	0.9996	0.9994	1.0000	0.9822	0.9984	0.9974
MCC	0.7197	0.7124	0.7133	0.6745	0.6890	0.5808	0.6925	0.6139	0.5472	0.5588	0.5287	0.7171	0.7176
FP	0.2800	0.0800	0.2800	0.0600	0.2200	0.1200	0.1400	0.2000	1.0200	0.3400	8.8400	2.0832	1.2600
FN	1.7200	1.8800	1.7800	2.1000	1.9200	2.5200	1.9600	2.3200	2.2200	2.5000	0.6000	1.1800	1.4800
MSPE	0.0697	0.0698	0.0697	0.0768	0.0775	0.0962	0.0784	0.0905	0.0996	0.1020	0.0944	0.1242	0.1460

A comparison between the "No outlier" and "Bad Leverage Outlier" scenarios shows that the performance of the t-link models with $\nu = 1$ or 3 experiences only minimal changes. In contrast, the results for the probit link show noticeable deterioration, both in variable selection (evidenced by lower MCC and higher FP and FN) and prediction (reflected in higher MSPE). Due to this one outlying observation, the probit model performs much worse than the t-link even when the datasets are generated under the probit link. These findings highlight the sensitivity of the probit link and the robustness of the t-link model. The t-link with $\nu = 7$ and the logit link show generally comparable performance, both more robust than the probit link but less robust than the t-links with $\nu = 1$ or $\nu = 3$.

When comparing the "Non-leverage outlier" scenario with the other two cases, the t-links with $\nu = 1$ and $\nu = 3$ again demonstrate robustness in both variable selection and prediction. The other link functions show some deterioration, though to a lesser degree than under the "Bad Leverage Outlier" scenario, with the probit link remaining the most vulnerable.

Overall, the LASSO models perform worse than their Bayesian counterparts fitted with either the skinny Gibbs or exact Gibbs algorithms using the same link functions.

Regarding the link functions, similar patterns are observed for the Cauchy, logit, and probit links under the LASSO method.

Figure 1 displays the runtime (in seconds) per iteration for the four methods, highlighting the computational efficiency of the skinny methods compared to the exact Gibbs methods. The computation time for HEGT and EProbit appears to increase quadratically with p, whereas HSGT and SProbit show a linear increase, consistent with their respective computational complexities. The t-link requires a slightly more time than the probit under either the exact or the skinny algorithms (by a margin of microseconds). This difference stems from the additional time spent sampling extra inverse-gamma latent variables, which is necessary when employing the t-link method.

The Fast Sampling Algorithm in [3] has a typical computational complexity of n^2p, which is slightly higher than the $n(p \vee |\boldsymbol{k}|^2)$ complexity of Skinny Gibbs, where $|\boldsymbol{k}|$ denotes the active model size and typically satisfies $|\boldsymbol{k}| \ll p$. The Skinny Gibbs method is specifically designed to ensure scalability under sparsity, making it a more suitable choice for sparse Bayesian variable selection.

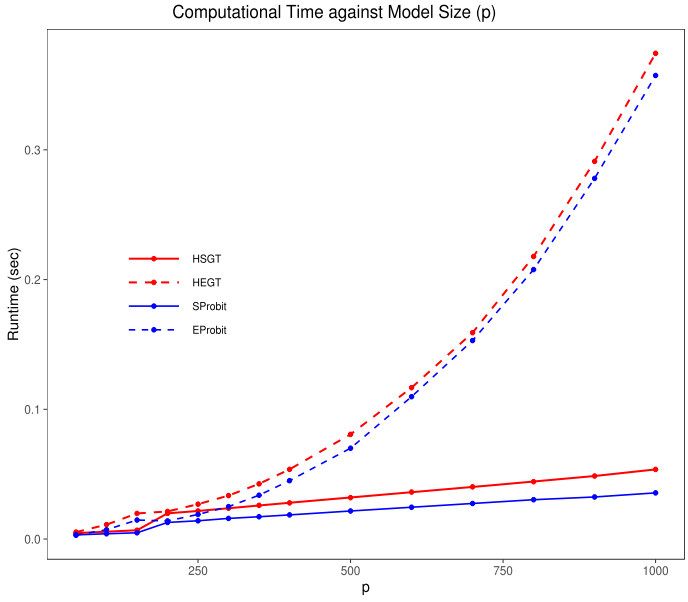


Figure 1: The average wall-clock time per iteration (in seconds) is compared across different covariate dimensions for the four methods: HSGT (solid red), SProbit (solid blue), HEGT (dashed red), and EProbit (dashed blue).

5. APPLICATION: PCR DATA

The PCR dataset originates from a study conducted by [27] to explore the genetics of two inbred mouse populations, B6 and BTBR. It includes gene expression data for 22,575 genes measured across 60 F2 mice, comprising 31 females and 29 males. The physiological phenotype, glycerol-3-phosphate acyltransferase (GPAT), was quantified using quantitative real-time Polymerase Chain Reaction (PCR). Both the phenotypic data (including GPAT measurements) and the gene expression data are publicly available on the GEO website (http://www.ncbi.nlm.nih.gov/geo) under accession number GSE3330.

GPAT is known to influence Hepatic Steatosis, a condition associated with obesity, with lower GPAT levels reducing its prevalence. Following the approach in [32], we define a binary response variable (E) based on GPAT values as $E = \mathbf{I}(\text{GPAT} < Q(0.4))$, where $Q(0.4)$ represents the 40th percentile of GPAT.

Given the high dimensionality of the gene expression data, an initial screening step was performed using the simple binary regression. This process identified the 99 most significant genes based on p-values, which were combined with the sex variable, resulting in a total of 100 predictors for further analysis. Figure 2 displays the boxplots of the standardized covariates in both datasets. The plot highlights the presence of numerous covariates with outlying values, which is a common occurrence in real-world data applications.

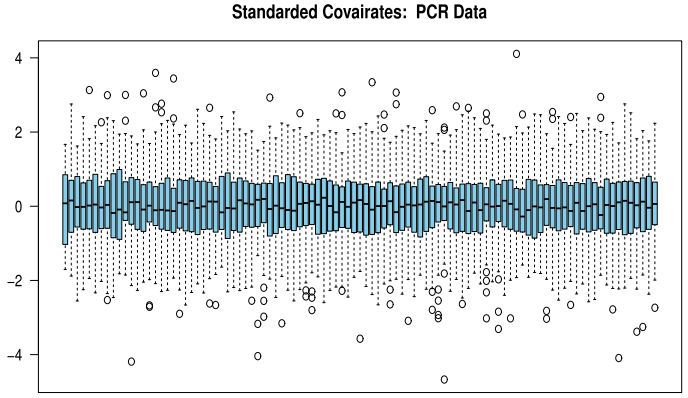


Figure 2: Boxplots of all the standard covariates in the PCR dataset.

The six models, incorporating three link functions and two computational algorithms, are then fitted to the data. The initial number of active genes influencing GPAT is set to 10 ($|k| = 10$), and the $\boldsymbol{\beta}$ vector is initialized to zero. Bayesian inference is performed using 4,000 MCMC samplings, following a burn-in period of 2,000 iterations. To assess the performance of these variable selection methods, we use 10-fold cross-validation and evaluate prediction accuracy based on mean squared prediction errors (MSPE), prediction accuracy and area under the curve (AUC). The dataset T is partitioned into 10 subsets, denoted $T_1, T_2, \ldots, T_{10}$. For each subset T_r (where $r = 1, \ldots, 10$), variable selection is performed on the remaining data ($T \setminus T_r$), and the model is used to compute predicted probabilities for the responses in T_r. The MSPE for subset r is calculated as

$\text{MSPE}_{CV(r)} = \sum_{i \in S_r}(E_i - \hat{p}_i)^2/n_r$ and $\text{Accuracy}_{CV(r)} = (TP_r + TN_r)/(TP_r + TN_r + FP_r + FN_r)$ where $n_r = |S_r|$, $\hat{p}_i$ is the predicted probability for the ith observation, and E_i is the observed response. The process is repeated for $r = 1, \ldots, 10$ and the overall cross-validated MSPE, accuracy and AUC are then obtained by averaging across all 10 subsets.

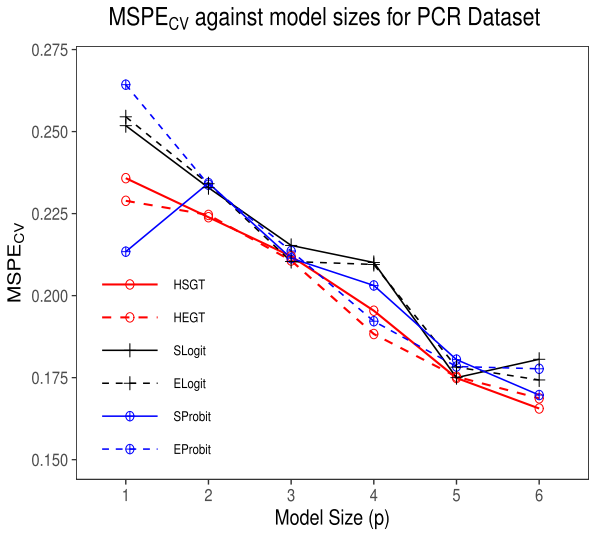


Figure 3: PCR Data: A 10-fold cross-validated mean square prediction error versus model size.

Figure 3 depicts the 10-fold cross-validation errors based on the selected number of covariates (model size). The MSPE_{CV} is plotted against the different model sizes ranging from $|k| = 1$ to 6 with lower values indicating better predictive performance. It is evident that the HSGT and HEGT yield the smallest MSPE_{CV} although the SProbit has the smallest MSPE when the number of active covariates is assumed to be one. The overall out-of-sample prediction accuracy in Figure 4 shows that the two t-link models demonstrate superior classification accuracy compared to the logit and probit methods except at model size 1.

Figure 5 illustrates the performance measured by AUC (Area Under the Curve) across varying model sizes. The plot shows that HSGT consistently achieves the highest AUC values as the model size increases, and has the highest AUC value from model size 3 to 6. The second best performed model is the HEGT. These results demonstrate the robustness of the t-link model and the practical performance of the scalable skinny Gibbs algorithm.

6. DISCUSSION

In this study, we establish the theoretical consistency results for the scalable algorithm to implement t-link (HSGT) in high-dimensional variable selection. The performance of different links and different algorithms is illustrated through simulation studies and an application to the real dataset.

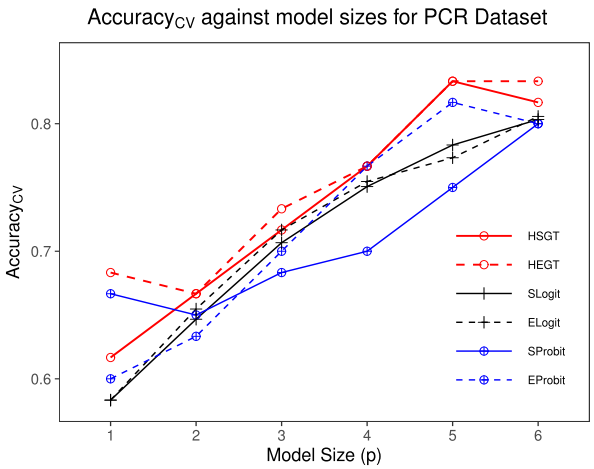


Figure 4: PCR Data: 10-fold cross-validated prediction accuracy versus model size.

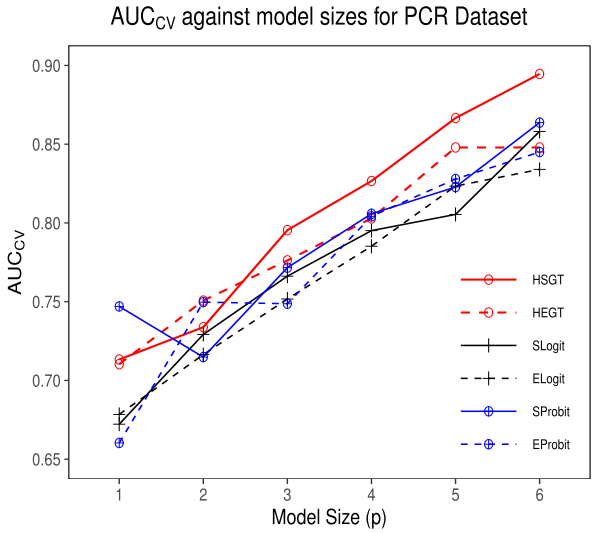


Figure 5: PCR Data: 10-fold cross-validated AUC performance evaluation versus model size.

The t-link model is preferable, compared to the probit link, given its robustness in variable selection and prediction in the presence of outlying cases. HSGT methods offers competitive computational efficiency. As observed in [37], HSGT has a lower variability of the MCMC chains for active variables than HEGT and thus better performance in variable selection. Our application to real datasets further supports the effectiveness of HSGT. This outcome underscores its utility in handling high-dimensional sparse data, a common characteristic of modern datasets.

Our work opens the door to several promising research directions. First, it would be valuable to further investigate the interplay between link function choice and variable selection in high-dimensional data settings. While [35] explored this interplay using reversible jump Markov chain Monte

Carlo (RJMCMC) to study the simultaneous selection of link functions and variables, we aim to address this question in high-dimensional settings by removing restrictions on the degrees of freedom v. Such an investigation would enhance the theoretical foundations and practical applications of Bayesian methods in modern data science.

Additionally, we are interested in establishing consistency results for more general link function models. We also plan to explore the use of Beta priors on q, as suggested by [41] and [11], and implement scalable spike-and-slab algorithms introduced by [4]. Another promising avenue is extending the methods to mixed-effects models, leveraging the normal scale mixture prior, as discussed by [45].

Robustness in generalized linear models can be pursued through the choice of link function or through robust estimation procedures such as M-estimators with adaptive weighting. In this study, we focus on the former. The weighting-based methods [23, 20] achieve robustness by downweighting observations with large residuals or leverage, thereby directly limiting the impact of contamination. Robustness by link modification stabilizes inference by the latent structures associated with the corresponding links. For example, [19] showed that the t-link's influence function is bounded and its weight function down-weights extreme observations, thereby providing robust estimation under contamination and outperforming probit or logit links in such settings. While the weighting-based methods provides more flexibility, alternative links provide robustness while retaining standard estimation procedures.

SUPPLEMENTARY MATERIAL

The supplementary materials include the R code used in this study.

APPENDIX A. PROOFS

A.1 Background

We begin by introducing the notations that will be utilized throughout the proofs. For the remainder of the paper, $\boldsymbol{E} = (E_1, \ldots, E_n)$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)' \in \mathbb{R}^{n\times p}$. For any k, let $\mu_{i,k} = \boldsymbol{x}_{i,k}\boldsymbol{\beta}_k$, where $\boldsymbol{x}_{i,k}$ is the i^{th} row vector in the active covariate matrix $\boldsymbol{X}_k \in \mathbb{R}^{n\times k}$. The log-likelihood function is

$$\ell\left(\boldsymbol{\beta}_k, v \mid \boldsymbol{E}, \boldsymbol{X}\right) = \sum_{i=1}^{n} \{E_i \ \log\ \Psi(\mu_{i,k},\ v) + (1-E_i)\ \log\left[1-\Psi(\mu_{i,k},\ v)\right]\}.$$

Then the score function and the Hessian matrix are given as:

$$S_n(\boldsymbol{\beta}_k) = \frac{\partial \ell}{\partial \boldsymbol{\beta}_k}$$

$$= \sum_{i=1}^{n}\left[E_i \frac{\psi(\mu_{i,k},\ v)}{\Psi(\mu_{i,k},\ v)} - (1-E_i)\frac{\psi(\mu_{i,k},\ v)}{1-\Psi(\mu_{i,k},\ v)}\right]\boldsymbol{x}_{i,k}$$

$$\equiv \boldsymbol{X}_k' D(\boldsymbol{\beta}_k, v)\left[\Sigma(\boldsymbol{\beta}_k, v)\right]^{-\frac{1}{2}}\left(\boldsymbol{E} - \Psi(\boldsymbol{\mu}_k, v)\right)$$

and the Hessian matrix

$$H_n(\boldsymbol{\beta}_k) = -\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\beta}_k'} = \sum_{i=1}^{n} \boldsymbol{x}_{i,k}\boldsymbol{x}_{i,k}' w_i(\boldsymbol{\beta}_k) = \boldsymbol{X}_k'\boldsymbol{W}(\boldsymbol{\beta}_k)\boldsymbol{X}_k,$$

where $D(\boldsymbol{\beta}_k, v) = \text{Diag}\left(s_1(\boldsymbol{\beta}_k, v), \ldots, s_n(\boldsymbol{\beta}_k, v)\right)$,
$\Sigma(\boldsymbol{\beta}_k, v) = \text{Diag}(\sigma_1^2(\boldsymbol{\beta}_k, v), \ldots, \sigma_n^2(\boldsymbol{\beta}_k, v))$,
$\sigma_i^2(\boldsymbol{\beta}_k, v) = \Psi(\mu_{i,k},\ v)[1-\Psi(\mu_{i,k},\ v)]$,
$\Psi(\boldsymbol{\mu}_k, v) = (\Psi(\mu_{1,k}, v), \ldots, \Psi(\mu_{n,k}, v))$, and
$\boldsymbol{W}(\boldsymbol{\beta}_k) = \text{Diag}(w_i(\boldsymbol{\beta}_k))$ for $i = 1, \ldots, n$.

Now, we restate several lemmas from [7], which are required for proving the consistency of model selection. The first lemma bounds $|H_n(\boldsymbol{\beta}_k)|$ below and above.

Lemma 2 (Lemma A1 in [7]). *Under (C1)- (C3) and $R_2 = (n/\log p)^{\frac{1-d}{2}}$ with $1/2 < d < 1$, for $\epsilon_n = CR_2\sqrt{\log p/n}$, we have*

$$(1-\epsilon_n)H_n(\boldsymbol{\beta}_{0,k}) \le H_n(\boldsymbol{\beta}_k) \le (1+\epsilon_n)H_n(\boldsymbol{\beta}_{0,k})$$

for any $\boldsymbol{k} \in S_1^* = \{\boldsymbol{k} : \boldsymbol{k} \supset \boldsymbol{t},\ \boldsymbol{k} \le R_2 + |\boldsymbol{t}|\}$ and $\boldsymbol{\beta}_k$ such that $||\boldsymbol{\beta}_k - \boldsymbol{\beta}_{0,k}||_2 \le (C'|\boldsymbol{k}|\log\ p/n)^{\frac{1}{2}}$, with probability at least $1 - 2\exp(-cn)$ for some positive constant c, C and C' NB: The probability $1-2\exp(-cn)$ is based on the fact that, there exist positive constants T_1 and T_2 such that

$$T_1 \le \min_{\boldsymbol{k}:|\boldsymbol{k}|\le R_2+|\boldsymbol{t}|} \lambda_{min}\left(\frac{\boldsymbol{X}_k'\boldsymbol{X}_k}{n}\right) \le \max_{\boldsymbol{k}:|\boldsymbol{k}|\le R_2+|\boldsymbol{t}|} \lambda_{max}\left(\frac{\boldsymbol{X}_k'\boldsymbol{X}_k}{n}\right) \le T_2$$

with probability at least $1 - 2\exp(-cn)$ for some constant $c > 0$ by Theorem 5.39 and Remark 5.40 in [15].

Lemma 3 provides an inequality for quadratic forms involving the projection matrices onto the column space of the design matrix.

Lemma 3 (Lemma A2 in [7]). *Let $\tilde{\boldsymbol{U}} = \Sigma^{-1/2}(\boldsymbol{E} - \boldsymbol{\mu})$ and $\boldsymbol{P}_k$ be the projection matrix onto the column space of $\boldsymbol{D}_{0,k}\boldsymbol{X}_k$, where $|\boldsymbol{k}| \le R_2$, and $R_2 = (n/logp)^{\frac{1-d}{2}}$ with $1/2 < d < 1$. Then, for some constant $\delta^* > 0$,*

$$P\left[\tilde{\boldsymbol{U}}'\boldsymbol{P}_k\tilde{\boldsymbol{U}} > (1+\delta^*)\{tr(\boldsymbol{P}_k + 2\sqrt{tr(\boldsymbol{P}_k)s} + 2s\}\right] \le e^{-s},$$

for all $s > 0$.

Lemma 4 (Lemma A3 in [7]). *Under conditions (C1)- (C3) and* $R_2 = (n/\log p)^{\frac{1-d}{2}}$ *with* $1/2 < d < 1$, *we have*

$$\sup_{\boldsymbol{k}:\boldsymbol{k}\supset\boldsymbol{t}} ||\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{0,k}||_2 = O\left(\sqrt{\frac{|\boldsymbol{k}|\log p}{n}}\right)$$

uniformly for all $|\boldsymbol{k}| \le R_2 + |\boldsymbol{t}|$ *with probability at least* $1 - 2\exp(-cn)$ *for some constant* $c > 0$. *Furthermore, under the same conditions and* $|\boldsymbol{t}| \ge 1$, *we have*

$$\sup_{\boldsymbol{k}:\boldsymbol{k}\not\supseteq\boldsymbol{t}} ||\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{0,k}||_2 = O\left(\sqrt{\frac{|\boldsymbol{k}|\log p}{n}}\right)$$

uniformly for all $|\boldsymbol{k}| \le R_2 + |\boldsymbol{t}|$ *with probability at least* $1 - 2p^{-|\boldsymbol{k}|} - 2\exp(-cn)$ *for some constant* $c > 0$.

Lemma 4 bounds the Euclidean norm of the different between the MLE and the true regression coefficient.

Lemma 5 (Lemma A4 in [7]). *Under conditions (A1)–(A3) and* $R_2 = (n/\log p)^{\frac{1-d}{2}}$ *with* $1/2 < d < 1$, *there exist a constant* $\lambda > 0$ *such that*

$$\lambda \le \min_{\boldsymbol{k}:|\boldsymbol{k}|\le R+|\boldsymbol{t}|} \lambda_{\min}\left(\frac{1}{n}H_n(\boldsymbol{\beta}_{0,k})\right) \le \boldsymbol{\Lambda}_{m_n} \le C^2\left(\frac{n}{\log p}\wedge \log p\right)^d,$$

and $\boldsymbol{\Lambda}_{m_n} = \max_{\boldsymbol{k}:\,|\boldsymbol{k}|\le\zeta} \lambda_{\max}(n^{-1}\boldsymbol{X}_k'\boldsymbol{X}_k)$ *for any integer* $\zeta > 0$. *Also, for any model* $\boldsymbol{k} \in \{\boldsymbol{k} \subseteq [p] : |\boldsymbol{k}| \le m_n\}$ *and any* $u \in \{u \in \mathbb{R}^n : u$ *is in the space spanned by the columns of* $\boldsymbol{\Sigma}^{1/2}\boldsymbol{X}_k\}$, *with probability at least* $1 - 2\exp(-cn)$ *for some constant* $c > 0$.

Lemma 5 indicates that the minimum eigenvalues of all the $\frac{1}{n}H_n(\boldsymbol{\beta}_{0,k})$ will be uniformly bounded below by a constant under the certain conditions. It also provides an upper bounds for max eigenvalues of all the $\lambda_{\max}\left(n^{-1}\boldsymbol{X}_k'\boldsymbol{X}_k\right)$, respectively, where k belongs to the set of reasonably large models. The lower bound can be viewed as a restricted eigenvalue condition for k-sparse vectors, which is typically satisfied with high probability for sub-Gaussian design matrices [32]. Similar idea have been employed in the linear regression literature [25, 46, 44].

Equipped with these lemmas, we can now proceed to establish the main results, as detailed in the follow-up subsections.

A.2 Proof of Theorem 1

Proof. We want to show that:

$$PR(\boldsymbol{k}, \boldsymbol{t}) \xrightarrow{P} 0 \quad \text{as } n \to \infty. \tag{E1}$$

Consider $S_1 = \{\boldsymbol{k} : \boldsymbol{k} \supsetneq \boldsymbol{t}, |\boldsymbol{k}| \le m_n\}$ (overfitted models).

By Taylor's expansion of $\ell_n(\boldsymbol{\beta}_k)$ around the MLE $(\widehat{\boldsymbol{\beta}}_k)$ of $\boldsymbol{\beta}_k$ under the model $\boldsymbol{k}$,

$$\ell_n(\boldsymbol{\beta}_k) - \ell_n\left(\widehat{\boldsymbol{\beta}}_k\right) = -\frac{1}{2}\left(\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_k\right)' H_n\left(\widetilde{\boldsymbol{\beta}}_k\right)\left(\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_k\right)$$

for some $\widetilde{\boldsymbol{\beta}}_k$ such that $\left\|\widetilde{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k\right\|_2 \le \left\|\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_k\right\|_2$. For any $\boldsymbol{\beta}_k$ such that $\left\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{0,k}\right\|_2 \le C\sqrt{|\boldsymbol{k}|\log p/n} \equiv Cw_n$ for some constant $C > 0$, by Lemma 4,

$$\begin{aligned}\left\|\widetilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{0,k}\right\|_2 &\le \left\|\widetilde{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k\right\|_2 + \left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{0,k}\right\|_2 \\ &\le \left\|\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_k\right\|_2 + \left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{0,k}\right\|_2 \\ &\le \left\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{0,k}\right\|_2 + 2\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{0,k}\right\|_2 \le 3Cw_n\end{aligned}$$

uniformly for all $k \in S_1$ with probability at least $1 - 2\exp(-cn)$ for some constant $c > 0$. Thus by Lemma 2,

$$\ell_n(\boldsymbol{\beta}_k) - \ell_n\left(\widehat{\boldsymbol{\beta}}_k\right) \le -\frac{1-\epsilon}{2}\left(\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_k\right)' H_n\left(\boldsymbol{\beta}_{0,k}\right)\left(\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_k\right)$$

for some small constant $\epsilon > 0$. For any $\boldsymbol{\beta}_k$ such that $\left\|\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_k\right\|_2 = Cw_n/2$, we have

$$\begin{aligned}&\ell_n\left(\boldsymbol{\beta}_k\right) - \ell_n\left(\widehat{\boldsymbol{\beta}}_k\right) \\ &\le -\frac{1-\epsilon}{2}\left\|\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_k\right\|_2^2 \lambda_{\min}\left(H_n\left(\boldsymbol{\beta}_{0,k}\right)\right) \\ &\le -\frac{1-\epsilon}{8}C^2\lambda|\boldsymbol{k}|\log p \longrightarrow -\infty \text{ as } n \to \infty\end{aligned} \tag{E1}$$

with probability at least $1 - 2\exp(-cn)$, by Lemma 5. Note that it also holds for any $\boldsymbol{\beta}_k$ such that $\left\|\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_k\right\|_2 > Cw_n/2$ due to the concavity of $\ell_n(\cdot)$ and the fact that $\widehat{\boldsymbol{\beta}}_k$ maximizes $\ell_n\left(\boldsymbol{\beta}_k\right)$.

Define $B_k = \left\{\boldsymbol{\beta}_k : \left\|\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_k\right\|_2 \le Cw_n/2\right\}$, then $B_k \subset \left\{\boldsymbol{\beta}_k : \left\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{0,k}\right\|_2 < Cw_n\right\}$ with probability at least $1 - 2\exp(-cn)$ uniformly in $\boldsymbol{k} \in S_1$. Therefore, for any $\boldsymbol{k} \in S_1$, with probability at least $1 - 2\exp(-cn)$, The proof for showing consistency under overfitted models will start by obtaining an upper bound for the marginal density $\pi(\boldsymbol{k} \mid \boldsymbol{X}, \boldsymbol{E})$ under any model $k \in S_1$, follow by obtaining a lower bound for $\pi(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{E})$ under the true model $\boldsymbol{t}$. To do that, we will deal with integrals over set B and set B^c separately.

Note that, we have shown that for any model $\boldsymbol{k}$, $\ell_n(\boldsymbol{\beta}_k) - \ell_n(\widehat{\boldsymbol{\beta}}_k) \le -\frac{1-\epsilon}{2}(\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_k)' H_n(\boldsymbol{\beta}_{0,k})(\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_k)$,for $\boldsymbol{\beta} \in B$ and $\ell_n(\boldsymbol{\beta}_k) - \ell_n(\widehat{\boldsymbol{\beta}}_k) \le -\frac{1-\epsilon}{8}c^2\lambda|k|\log p$, for $\boldsymbol{\beta} \in B^c$ with probability tending to 1. The proof below proceed with the above two inequalities to obtain the upper bound for $\pi(\boldsymbol{k} \mid \boldsymbol{X}, \boldsymbol{E})$. It follows that

$$\begin{aligned}&\pi(\boldsymbol{k} \mid \boldsymbol{X}, \boldsymbol{E}) \\ &\propto \left\{2\pi v_n^{-2}\left(n - 1 + \tau_0^{-2}\right)\right\}^{\frac{|\boldsymbol{k}|}{2}} s^r/\Gamma(r) \times \\ &\int\int (\tau_1^2)^{-\frac{|\boldsymbol{k}|+2r+2}{2}} \exp\left(\ell_n\left(\boldsymbol{\beta}_k\right) - \frac{\|\boldsymbol{\beta}_k\|_2^2 + 2s}{2\tau_1^2}\right) d\boldsymbol{\beta}_k d\tau_1^2 \\ &= \left\{2\pi q^2\left(n\tau_0^2 - \tau_0^2 + 1\right)\tau_1^{-2}/(1-q)^2\right\}^{|\boldsymbol{k}|/2} s^r/\Gamma(r) \times\end{aligned}$$

$$\int\int(\tau_1^2)^{-\frac{|\boldsymbol{k}|+2r+2}{2}}\exp\left(\ell_n(\boldsymbol{\beta}_k)-\frac{\|\boldsymbol{\beta}_k\|_2^2+2s}{2\tau_1^2}\right)d\boldsymbol{\beta}_k d\tau_1^2$$
$$\le\left(C_1 q^{2c_q}\tau_1^2\right)^{-|\boldsymbol{k}|/2}\times$$
$$\int\int(\tau_1^2)^{-\frac{|\boldsymbol{k}|+2r+2}{2}}\exp\left(\ell_n(\boldsymbol{\beta}_k)-\frac{\|\boldsymbol{\beta}_k\|_2^2+2s}{2\tau_1^2}\right)d\boldsymbol{\beta}_k d\tau_1^2$$

for some constant $C_1>0$. It proceed as follows. follows. follows. follows. follows. follows. follows.

$$\begin{aligned}
&\pi(\boldsymbol{k}\mid\boldsymbol{X},\boldsymbol{E})\\
&\propto\left\{2\pi v_n^{-2}\left(n-1+\tau_0^{-2}\right)\right\}^{\frac{|\boldsymbol{k}|}{2}}s^r/\Gamma(r)\times\\
&\int\int(\tau_1^2)^{-\frac{\boldsymbol{k}}{2}-r-1}\exp\left(\ell_n(\boldsymbol{\beta}_k)-\frac{\|\boldsymbol{\beta}_k\|_2^2+2s}{2\tau_1^2}\right)d\boldsymbol{\beta}_k d\tau_1^2\\
&\le\left(C_1p^2\right)^{-|\boldsymbol{k}|/2}\exp\left\{\ell_n\left(\hat{\boldsymbol{\beta}}_k\right)\right\}\times\\
&\Bigg[\int\int_{B_k}(\tau_1^2)^{-|\boldsymbol{k}|/2-r-1}e^{-\frac{s}{\tau_1^2}}\times\\
&\exp\left\{-\frac{1-\epsilon}{2}(\boldsymbol{\beta}_k-\hat{\boldsymbol{\beta}}_k)'H_n(\boldsymbol{\beta}_{0,k})(\boldsymbol{\beta}_k-\hat{\boldsymbol{\beta}}_k)-\frac{\|\boldsymbol{\beta}_{\boldsymbol{k}}\|_2^2}{2\tau_1^2}\right\}d\boldsymbol{\beta}_k d\tau_1^2\\
&+\exp\left(-\frac{1-\epsilon}{8}C^2\lambda|\boldsymbol{k}|\log p\right)\\
&\int\int_{B_k^c}(\tau_1^2)^{-|\boldsymbol{k}|/2-r-1}e^{-\frac{s}{\tau_1^2}}\exp\left(-\frac{\|\boldsymbol{\beta}_k\|_2^2}{2\tau_1^2}\right)d\boldsymbol{\beta}_k d\tau_1^2\Bigg].
\end{aligned}\tag{E2}$$

Note that for $A_k=(1-\epsilon)H_n(\boldsymbol{\beta}_{0,k})$ and ${\boldsymbol{\beta}_k}^*=(A_k+I_{|\boldsymbol{k}|}/\tau_1^2)^{-1}A_k\widehat{\boldsymbol{\beta}}_k$, we have

$$\begin{aligned}
&\int\int_{B_k}(\tau_1^2)^{-|\boldsymbol{k}|/2-r-1}e^{-\frac{s}{\tau_1^2}}\times\\
&\exp\left\{-\frac{1-\epsilon}{2}(\boldsymbol{\beta}_k-\widehat{\boldsymbol{\beta}}_k)'H_n(\boldsymbol{\beta}_{0,k})(\boldsymbol{\beta}_k-\widehat{\boldsymbol{\beta}}_k)-\frac{\|\boldsymbol{\beta}_k\|_2^2}{2\tau_1^2}\right\}d\boldsymbol{\beta}_k d\tau_1^2\\
&\le\int\int(\tau_1^2)^{-|\boldsymbol{k}|/2-r-1}\exp\left(-\frac{s}{\tau_1^2}\right)\times\\
&\quad\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_k-{\boldsymbol{\beta}_k}^*)'(A_k+I_{|\boldsymbol{k}|}/\tau_1^2)(\boldsymbol{\beta}_k-{\boldsymbol{\beta}_k}^*)\right\}\times\\
&\quad\exp\left\{-\frac{1}{2}\widehat{\boldsymbol{\beta}}_k'\left(A_k-A_k(A_k+I_{|\boldsymbol{k}|}/\tau_1^2)^{-1}A_k\right)\widehat{\boldsymbol{\beta}}_k\right\}d\boldsymbol{\beta}_k d\tau_1^2\\
&=\int(2\pi)^{|\boldsymbol{k}|/2}\det\left(A_k+I_{|\boldsymbol{k}|}/\tau_1^2\right)^{-1/2}(\tau_1^2)^{-|\boldsymbol{k}|/2-r-1}\exp\left(-\frac{s}{\tau_1^2}\right)\times\\
&\quad\exp\left\{-\frac{1}{2}\widehat{\boldsymbol{\beta}}_k'\left(A_k-A_k(A_k+I_{|\boldsymbol{k}|}/\tau_1^2)^{-1}A_k\right)\widehat{\boldsymbol{\beta}}_k\right\}d\tau_1^2,
\end{aligned}$$

where ${\boldsymbol{\beta}_k}^*$ denotes the mean with respect to a multivariate normal distribution and covariance matrix $A_k+I_{|\boldsymbol{k}|}/\tau_1^2$. Therefore, by noting that $\exp\{-1/2\widehat{\boldsymbol{\beta}}_k'\left(A_k-A_k(A_k+I_{|\boldsymbol{k}|}/\tau_1^2)^{-1}A_k\right)\widehat{\boldsymbol{\beta}}_k\}\ge 1$, it follows from (E2) that, for some constant $C_1>0$,

$$\begin{aligned}
&\int\int_{B_k}(\tau_1^2)^{-|\boldsymbol{k}|/2-r-1}e^{-\frac{s}{\tau_1^2}}\exp\left\{-\frac{1}{2\tau_1^2}\|\boldsymbol{\beta}_{\boldsymbol{k}}\|_2^2\right\}\times\\
&\exp\left\{-\frac{1-\epsilon}{2}(\boldsymbol{\beta}_k-\widehat{\boldsymbol{\beta}}_k)'H_n(\beta_{0,k})(\boldsymbol{\beta}_k-\widehat{\boldsymbol{\beta}}_k)\right\}d\boldsymbol{\beta}_k d\tau_1^2\\
&\le C_\pi\int\det\left(A_k+I_{|\boldsymbol{k}|}/\tau_1^2\right)^{-\frac{1}{2}}(\tau_1^2)^{-\frac{|\boldsymbol{k}|+2r+2}{2}}\exp\left(-\frac{s}{\tau_1^2}\right)d\tau_1^2\\
&\le\ C_\pi\int\det\left(A_k\right)^{-1/2}(\tau_1^2)^{-|\boldsymbol{k}|/2-r-1}\exp\left(-\frac{s}{\tau_1^2}\right)d\tau_1^2\\
&\le C_\pi\det(n^{-1}A_k)^{-1/2}s^{-(|\boldsymbol{k}|/2+r)}\Gamma\left(|\boldsymbol{k}|/2+r\right),
\end{aligned}\tag{E3}$$

where $C_\pi=(2\pi)^{|\boldsymbol{k}|/2}$. Next, note that

$$\begin{aligned}
&\int\int_{B_k^c}(\tau_1^2)^{-|\boldsymbol{k}|/2-r-1}e^{-\frac{s}{\tau_1^2}}\exp\left(-\frac{\|\boldsymbol{\beta}_k\|_2^2}{2\tau_1^2}\right)d\boldsymbol{\beta}_k d\tau_1^2\\
&\le\int\int(\tau_1^2)^{-|\boldsymbol{k}|/2-r-1}\exp\left(-\frac{s}{\tau_1^2}\right)\exp\left(-\frac{{\boldsymbol{\beta}_k}'\boldsymbol{\beta}_k}{2\tau_1^2}\right)d\beta_k d\tau_1^2\\
&\le\int(\tau_1^2)^{-|\boldsymbol{k}|/2-r-1}\exp\left(-\frac{s}{\tau_1^2}\right)\left(2\pi\tau_1^2\right)^{|\boldsymbol{k}|/2}d\tau_1^2\\
&\le(2\pi)^{|\boldsymbol{k}|}s^{-r}\Gamma(r).
\end{aligned}\tag{E4}$$

Combining (E3) and (E4) and using the Stirling approximation for the gamma function, we obtain the following upper bound for $\pi(k\mid X,\boldsymbol{E})$ given as

$$\begin{aligned}
&\pi(\boldsymbol{k}\mid\boldsymbol{X},\boldsymbol{E})\lesssim\\
&\quad(C_3p^2)^{-|\boldsymbol{k}|/2}\exp\left\{\ell_n(\widehat{\boldsymbol{\beta}}_k)\right\}n^{-|\boldsymbol{k}|/2}\det(n^{-1}A_k)^{-1/2},
\end{aligned}\tag{E5}$$

for any $\boldsymbol{k}\in S_1$ and some constant $C_3>0$.

Note that

$$\det(A_k)^{1/2}s^{-(|\boldsymbol{k}|/2+r)}\Gamma\left(|\boldsymbol{k}|/2+r\right)\ll\exp\left\{\frac{1-\epsilon}{8}c^2\lambda|\boldsymbol{k}|\log p\right\}$$

by conditions (C1)–(C3).

On the other hand, for the true model t and some constant $C_4>0$,

$$\begin{aligned}
&\pi\left(\boldsymbol{t}\mid\boldsymbol{X},\boldsymbol{E}\right)\propto\\
&\left\{2\pi v_n^{-2}\left(n-1+\tau_0^{-2}\right)\right\}^{\frac{|\boldsymbol{t}|}{2}}s^r/\Gamma(r)\times\\
&\int\int(\tau_1^2)^{-\frac{|\boldsymbol{t}|}{2}-r-1}\exp\left\{\ell_n\left(\boldsymbol{\beta}_t\right)-\frac{1}{2\tau_1^2}\|\boldsymbol{\beta}_t\|_2^2-\frac{s}{\tau_1^2}\right\}d\boldsymbol{\beta}_t d\tau_1^2\\
&\propto\left\{2\pi v_n^{-2}\left(n-1+\tau_0^{-2}\right)\right\}^{\frac{|\boldsymbol{t}|}{2}}\exp\left\{\ell_n\left(\boldsymbol{\beta}_t\right)\right\}\times\\
&\int\int(\tau_1^2)^{-|\boldsymbol{t}|/2-r-1}\exp\left(-\frac{s}{\tau_1^2}\right)\\
&\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_t-{\boldsymbol{\beta}_t}^*)'(A_t+I_{|\boldsymbol{t}|}/\tau_1^2)(\boldsymbol{\beta}_t-{\boldsymbol{\beta}_t}^*)\right\}\times\\
&\exp\left\{-\frac{1}{2}\widehat{\boldsymbol{\beta}}_t'\left(A_t-A_t(A_t+I_{|\boldsymbol{t}|}/\tau_1^2)^{-1}A_t\right)\widehat{\boldsymbol{\beta}}_t\right\}d\boldsymbol{\beta}_t d\tau_1^2\\
&\propto\left\{2\pi v_n^{-2}\left(n-1+\tau_0^{-2}\right)\right\}^{\frac{|\boldsymbol{t}|}{2}}\exp\left\{\ell_n\left(\boldsymbol{\beta}_t\right)\right\}\times
\end{aligned}$$

$$\int (2\pi)^{|\boldsymbol{t}|/2} \det\left(A_t + I_{|\boldsymbol{t}|}/\tau_1^2\right)^{-1/2} (\tau_1^2)^{-|\boldsymbol{t}|/2-r-1} \exp\left(-\frac{s}{\tau_1^2}\right) \times \exp\left\{-\frac{1}{2}\widehat{\boldsymbol{\beta}}_t'(A_t - A_t(A_t + I_{|\boldsymbol{t}|}/\tau_1^2)^{-1}A_t)\widehat{\boldsymbol{\beta}}_t\right\} d\tau_1^2,$$
$$\geq \left(C_4 p^2\right)^{-|\boldsymbol{t}|/2} \exp\left\{\ell_n\left(\widehat{\boldsymbol{\beta}_t}\right)\right\} \times n^{-|\boldsymbol{t}|/2} \det(n^{-1}A_t')^{-1/2} (\log p)^{-2d}.$$

Now, the lower bound is obtained using both the arguments and those in [38]. In particular, for some constant $C_4 > 0$,

$$\pi(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{E}) \geq (C_4 q^2)^{-|\boldsymbol{k}|/2} \times \exp\{\ell_n(\widehat{\beta}_t)\} n^{-|\boldsymbol{t}|/2} \det(n^{-1}A_t')^{-1/2} (\log p)^{-2d},$$

where $A_t' = (1+\epsilon)H_n(\beta_{0,t})$. Therefore, with probability tending to 1, combining with (E5), for some constant $C^* > 0$,

$$\frac{\pi(\boldsymbol{k} \mid \boldsymbol{X}, \boldsymbol{E})}{\pi(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{E})} \lesssim \{C^* n q^2\}^{-(|\boldsymbol{k}|-|\boldsymbol{t}|)/2} \exp\left\{\ell_n(\widehat{\boldsymbol{\beta}}_k) - \ell_n(\widehat{\boldsymbol{\beta}}_t)\right\} \times \frac{\det(n^{-1}A_t')^{1/2}}{\det(n^{-1}A_k)^{1/2}} (\log p)^{2d} \lesssim (C^* p)^{-(2+\delta/2)(|\boldsymbol{k}|-|\boldsymbol{t}|)} p^{(1+\delta^*)(1+2w)(|\boldsymbol{k}|-|\boldsymbol{t}|)}, \tag{A.1}$$

for any $k \in S_1$, where $A_t' = (1+\epsilon)H_n(\boldsymbol{\beta}_{0,t})$ and the second inequality holds by Lemma 3.

$$\ell_n(\widehat{\boldsymbol{\beta}}_k) - \ell_n(\widehat{\boldsymbol{\beta}}_t) \leq b_n(|\boldsymbol{k}| - |\boldsymbol{t}|) \tag{A.2}$$

for any $\boldsymbol{k} \in S_1$ with probability tending to 1, where $b_n = (1+\delta^*)(1+2w)\log p$ such that $1 + \delta/2 > (1+\delta^*)(1+2w)$. Note that since we will focus on sufficiently large n, δ^* can be considered as an arbitrarily small constant.

Hence, with probability tending to 1, it follows from (A.1) and (A.2) that

$$PR(\boldsymbol{k}, \boldsymbol{t}) = \frac{\pi(\boldsymbol{k} \mid X, E)}{\pi(\boldsymbol{t} \mid X, E)} \lesssim (C^* p)^{-(1+\delta^*)(|\boldsymbol{k}|-|\boldsymbol{t}|)} = o(1),$$

for some fixed constant $\delta' > 0$. By adapting the line of reasoning presented in [38], one can show that the above expression is $o(1)$ uniformly over all $\boldsymbol{k} \in S_2$, where $S_2 = \{\boldsymbol{k} : \boldsymbol{k} \not\supseteq \boldsymbol{t},\ |\boldsymbol{k}| \leq R_n\}$. Thus, we have proved the desired result (3.1). □

Proof of Theorem 2. It suffices to show that

$$\sum_{\boldsymbol{k}:\boldsymbol{k}\neq\boldsymbol{t}} \frac{\pi(\boldsymbol{k} \mid X, E)}{\pi(\boldsymbol{t} \mid X, E)} \xrightarrow{P} 0 \quad \text{as } n \to \infty \tag{A9}$$

Note that

$$\sum_{\boldsymbol{k}:\boldsymbol{k}\neq\boldsymbol{t}} \frac{\pi(\boldsymbol{k} \mid X, E)}{\pi(\boldsymbol{t} \mid X, E)} = \sum_{\boldsymbol{k}\in S_1} \frac{\pi(\boldsymbol{k} \mid X, E)}{\pi(\boldsymbol{t} \mid X, E)} + \sum_{\boldsymbol{k}\in S_2} \frac{\pi(\boldsymbol{k} \mid X, E)}{\pi(\boldsymbol{t} \mid X, E)}$$

where $S_1 = \{\boldsymbol{k} : \boldsymbol{k} \supseteq t, |\boldsymbol{k}| \leq m_n\}$ and $S_2 = \{\boldsymbol{k} : \boldsymbol{k} \not\supseteq t, |\boldsymbol{k}| \leq R_n\}$. By the proof of Theorem: 1, we have

$$\sum_{\boldsymbol{k}\in S_1} \frac{\pi(\boldsymbol{k} \mid X, Y)}{\pi(\boldsymbol{t} \mid X, Y)} \lesssim \sum_{|\boldsymbol{k}|-|\boldsymbol{t}|=1}^{m_n-|\boldsymbol{t}|} \sum_{\boldsymbol{k}\in S_1} \left\{(C^* p)^{-(1+\delta^*)}\right\}^{|\boldsymbol{k}|-|\boldsymbol{t}|} \leq \sum_{|\boldsymbol{k}|-|\boldsymbol{t}|=1}^{m_n-|\boldsymbol{t}|} \binom{p-|\boldsymbol{t}|}{|\boldsymbol{k}|-|\boldsymbol{t}|} (C^* p)^{-(1+\delta^*)(|\boldsymbol{k}|-|\boldsymbol{t}|)}, \tag{E5}$$

for some constant C^*, $\delta^* > 0$. Using $\binom{p}{|\boldsymbol{k}|} \leq p^{|\boldsymbol{k}|}$ and (E5), we have

$$\sum_{\boldsymbol{k}\in S_1} PR(\boldsymbol{k}, \boldsymbol{t}) = \sum_{|\boldsymbol{k}|=|\boldsymbol{t}|+1}^{m_n-|\boldsymbol{t}|} \left\{(C^* p)^{-(1+\delta)}\right\}^{|\boldsymbol{k}|-|\boldsymbol{t}|} = o(1).$$

Similar arguments from [8] can be used to show that S_2 is of order o(1). This completes the proof. □

ACKNOWLEDGEMENTS

This work was supported in part by an allocation of computing time from the Ohio Supercomputer Center. The datasets used and/or analyzed in this study are available in the Gene Expression Omnibus (GEO) database maintained by the National Center for Biotechnology Information.

FUNDING

There is no funding required for this project.

CONFLICTS OF INTEREST

The authors declare that they have no conflict of interest.

Accepted 3 February 2026

REFERENCES

[1] Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association* **88**(422) 669–679. MR1224394

[2] Anderson, D. and Kurtz, T. Continuous time Markov chain models for chemical reaction networks. http://www.math.wisc.edu/~kurtz/papers/AndKurJuly10.pdf. Accessed 27 July 2010.

[3] Bhattacharya, A., Chakraborty, A. and Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika* 042. https://doi.org/10.1093/biomet/asw042. MR3620452

[4] Biswas, N., Mackey, L. and Meng, X. -L. (2022). Scalable spike-and-slab. In *International Conference on Machine Learning* 2021–2040. PMLR.

[5] Blanchet, J., Leder, K. and Glynn, P. (2009). Efficient Simulation of Light-Tailed Sums: an Old-Folk Song Sung to a Faster New Tune... In *Monte Carlo and Quasi-Monte Carlo Methods* (P. L' Ecuyer and A. B. Owen, eds.) Springer, Berlin. MR274389713,

[6] Blanchet, J., Leder, K. and Shi, Y. (2011). Analysis of a splitting estimator for rare event probabilities in Jackson networks. *Stochastic Systems* **1** 306–339. https://doi.org/10.1214/11-SSY026. MR2949543

[7] Cao, X. and Lee, K. (2023). Consistent and scalable Bayesian joint variable and graph selection for disease diagnosis leveraging functional brain network. *Bayesian Analysis* **1**(1) 1–29. https://doi.org/10.1214/23-ba1376. MR4770325

[8] Cao, X. and Lee, K. (2024). Bayesian inference on hierarchical nonlocal priors in generalized linear models. *Bayesian Analysis* **19**(1) 99–122. https://doi.org/10.1214/22-ba1350. MR4692544

[9] Cao, X., Khare, K. and Ghosh, M. (2020). High-dimensional posterior consistency for hierarchical non-Local priors in regression. *Bayesian Analysis* **15**(1) 241–262. https://doi.org/10.1214/19-BA1154. MR4050884

[10] Cardie, C. and Howe, N. (1997). Improving minority class prediction using case-specific feature weights. In *Computer Science: Faculty Publications, Smith College, Northampton, MA*.

[11] Castillo, I. and Van Der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics* **40** 2069–2101. https://doi.org/10.1214/12-AOS1029. MR3059077

[12] Chen, M. -H., Ibrahim, J. G. and Yiannoutsos, C. (1999). Prior elicitation, variable selection and Bayesian computation for logistic regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(1) 223–242. https://doi.org/10.1111/1467-9868.00173. MR1664057

[13] Chicco, D. and Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining* **16**(1) 4.

[14] Dunn, P. K., Smyth, G. K. et al. (2018) *Generalized linear models with examples in R* **53**. Springer. https://doi.org/10.1007/978-1-4419-0118-7. MR3887706

[15] Eldar, Y. C. and Kutyniok, G. (2012) *Compressed sensing: theory and applications*. Cambridge university press. https://doi.org/10.1515/dmvm-2014-0014. MR3254193

[16] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). https://doi.org/10.1214/06-BA117A. MR2221284

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

[18] George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **35** 339–373.

[19] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association* **69**(346) 383–393. MR0362657

[20] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986) *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York. MR0829458

[21] He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* **21**(9) 1263–1284.

[22] Hosseinian, S. and Morgenthaler, S. (2011). Robust binary regression. *Journal of Statistical Planning and Inference* **141**(4) 1497–1509. https://doi.org/10.1016/j.jspi.2010.11.015. MR2747918

[23] Huber, P. J. (1981) *Robust Statistics*. Wiley, New York. MR0606374

[24] Iannario, M., Monti, A. C., Piccolo, D. and Ronchetti, E. (2017). Robust inference for ordinal response models. *Electronic Journal of Statistics* **11**(2) 3407–3445. https://doi.org/10.1214/17-EJS1314. https://doi.org/10.1214/17-EJS1314. MR3709859

[25] Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics* **33**(2) 730–773. https://doi.org/10.1214/009053604000001147. MR2163158

[26] Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association* **107**(498) 649–660. https://doi.org/10.1080/01621459.2012.682536. MR2980074

[27] Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E. T. q. K., Flowers, M. T., Schueler, K. L., Manly, K. F. et al. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics* **2**(1) 6. MR2709393

[28] Lee, K. and Cao, X. (2021). Bayesian group selection in logistic regression with application to MRI data analysis. *Biometrics* **77**(2) 391–400. https://doi.org/10.1111/biom.13290. MR4307642

[29] Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A. and Kumar, S. (2020). Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

[30] Mirfarah, E., Naderi, M., Lin, T. -I. and Wang, W. -L. (2024). Robust Bayesian inference for the censored mixture of experts model using heavy-tailed distributions. *Advances in Data Analysis and Classification* 1–29. https://doi.org/10.1007/s11634-024-00609-2. MR4993902

[31] Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association* **83**(404) 1023–1032. MR0997578

[32] Narisetty, N. N., Shen, J. and He, X. (2018). Skinny Gibbs: A consistent and scalable Gibbs sampler for model selection. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2018.1482754. MR4011773

[33] Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics* **42**(2) 789–817. https://doi.org/10.1214/14-AOS1207. MR3210987

[34] Nikooienejad, A., Wang, W. and Johnson, V. E. (2016). Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors. *Bioinformatics* **32**(9) 1338–1345.

[35] Ntzoufras, I., Dellaportas, P. and Forster, J. J. (2003). Bayesian variable and link determination for generalised linear models. *Journal of Statistical Planning and Inference* **111**(1-2) 165–180. https://doi.org/10.1016/S0378-3758(02)00298-7. MR1955879

[36] O'Brien, S. M. and Dunson, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics* **60**(3) 739–746. https://doi.org/10.1111/j.0006-341X.2004.00224.x. MR2089450

[37] Odoom, E., Ouyang, J., Cao, X. and Wang, X. (2026). Hierarchical skinny Gibbs sampler in logistic regression using Pólya-Gamma latent variables. *Statistics and Its Interface* **19**(2) 179–196. https://doi.org/10.4310/sii.260108020451. MR5012359

[38] Ouyang, J. and Cao, X. (2024). Consistent skinny Gibbs in probit regression. *Computational Statistics & Data Analysis* **198** 107993. https://doi.org/10.1016/j.csda.2024.107993. https://doi.org/10.1016/j.csda.2024.107993. MR4752161

[39] Rockova, V., Lesaffre, E., Luime, J. and Löwenberg, B. (2012). Hierarchical Bayesian formulations for selecting variables in regression models. *Statistics in medicine* **31**(11-12) 1221–1237. https://doi.org/10.1002/sim.4439. MR2925691

[40] Scalera, V., Iannario, M. and Monti, A. C. (2021). Robust link functions. *Statistics* **55**(4) 963–977. https://doi.org/10.1080/02331888.2021.1987436. MR4364403

[41] Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* **38** 2587–2619. https://doi.org/10.1214/10-AOS792. MR2722450

[42] Seumois, G., Ramírez-Suástegui, C., Schmiedel, B. J., Liang, S., Peters, B., Sette, A. and Vijayanand, P. (2020). Single-cell transcriptomic analysis of allergen-specific T cells in allergy and asthma. *Science Immunology* **5**(48) 6087.

[43] Shin, M., Bhattacharya, A. and Johnson, V. E. (2018). Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statistica Sinica* **28** 1053–1078. MR3791100

[44] Song, Q. and Liang, F. (2017). Nearly optimal Bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964*. https://doi.org/10.1007/s11425-020-1912-6. MR4535982

[45] Yang, M., Wang, M. and Dong, G. (2020). Bayesian vari-

able selection for mixed effects model with shrinkage prior. *Computational Statistics* **35** 227–243. https://doi.org/10.1007/s00180-019-00895-x. MR4066283

[46] Yang, Y., Wainwright, M. J., Jordan, M. I. et al. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics* **44**(6) 2497–2532. https://doi.org/10.1214/15-AOS1417. MR3576552

Eric Odoom.
Division of Statistics and Data Science
Department of Mathematical Sciences
University of Cincinnati
Cincinnat, OH, 45221, USA.
E-mail address: odoomec@mail.uc.edu

Xia Wang.
Division of Statistics and Data Science
Department of Mathematical Sciences
University of Cincinnati
Cincinnat, OH, 45221, USA.
E-mail address: xia.wang@uc.edu