Journal of Mathematics and Modeling in Finance (JMMF)
Vol. 2, No. 2, Summer & Autumn 2022
Research paper

Bayesian Inference Using Hyper Product Inverse Moment Prior in the Ultrahigh-Dimensional Generalized Linear Models

Robabeh Hosseinpour Samim Mamaghani[1], Farzad Eskandari[2]

[1] Department of Statistics, Faculty of Statistics, Mathematics and Computer, Allameh Tabatabai University, Tehran, Iran

r_hosseinpour@atu.ac.ir

[2] Department of Statistics, Faculty of Statistics, Mathematics and Computer, Allameh Tabatabai University, Tehran, Iran

askandari@atu.ac.ir

Abstract:
In this paper, we considered a Bayesian hierarchical method using the hyper product inverse moment prior in the ultrahigh-dimensional generalized linear model (UDGLM), that was useful in the Bayesian variable selection. We showed the posterior probabilities of the true model converge to 1 as the sample size increases. For computing the posterior probabilities, we implemented the Laplace approximation. The Simplified Shotgun Stochastic Search with Screening (S5) procedure for generalized linear model was suggested for exploring the posterior space. Simulation studies and real data analysis using the Bayesian ultrahigh-dimensional generalized linear model indicate that the proposed method had better performance than the previous models. *Keywords:* Ultrahigh dimensional; Nonlocal prior; Optimal properties; Bayesian Variable Selection; Generalized Linear Model.
JEL Classification: C68, G10, C45.

1 Introduction

Identifying a sparse subset from a large number of covariates per observation to balance parsimony and predictive power in high dimensional statistical problems, especially in generalized linear model (GLM), is a variety of scientific research and technological development.

When the number of covariates grows at a sub-exponential rate of n, variable selection is the first step for dimension reduction to estimate the parameters of the model. Our objective is to fit a GLM by efficiently estimating regression coefficients $\boldsymbol{\beta}$ and use it for subsequent inference. Recently, many common methods for selecting variables from both pluralistic and Bayesian perspectives have been developed. Most of the frequentist methods can be interpreted from a Bayesian perspective, because they share the basic desire of shrinkage toward sparse models. In the con-

[2]Corresponding author
Received: 31/08/2022 Accepted: 01/11/2022
https://doi.org/10.22054/jmmf.2022.69844.1072

text of Bayesian testing in regression models, [10] have shown that the local priors (LPs) put a positive probability on the null value of the parameter whereas nonlocal priors (NLPs) put zero probability on the null value. Thus, NLPs consider a clear separation between the null hypothesis that some regression coefficients are equal to zero and the alternative hypothesis that these coefficients are different from zero. Now, let $\mathbf{Y}_n = (y_1, y_2, \cdots, y_n)^{'}$ be n-dimensional response vector of Gaussian linear model $\mathbf{Y}_n \sim N(\mathbf{X}_n\boldsymbol{\beta}, \phi_k\mathbf{I_n})$ where $\mathbf{X}_n$ is a $n \times p$ design matrix with n sample size and p number of covariates, $\boldsymbol{\beta} \in R^p$ is vector of parameters of interest and $\phi_k \in R^+$ is a fixed dimension nuisance parameter. Since, we do not know which covariates truly predict $\mathbf{y}_n$, we consider $k = 2^p$ models by setting the elements in $\boldsymbol{\beta}$ to zero. Let $\mathbf{M}$ denote the model space that collects all the model indices $\mathbf{k}$; i.e., $\mathbf{M} = \{\mathbf{k} : \mathbf{k} \subseteq \{0, 1\}^p\}$. The nonlocal priors, moment (pMOM) and product inverse moment (piMOM) priors are introduced by [10] as follows

$$\pi_M\left(\boldsymbol{\beta} \mid \phi_k, \tau, r, M_k\right) = \prod_{j \in M_k} \left[(2r-1)!!\right]^{-1} \frac{\beta_j^{2r}}{(\tau\phi_k)^r} N\left(\beta_j; 0, \tau\phi_k\right) \tag{1}$$

$$\pi_I\left(\boldsymbol{\beta} \mid \phi_k, \tau, r, M_k\right) = \prod_{j \in M_k} \frac{(\tau\phi_k)^{\frac{r}{2}}}{\Gamma(\frac{r}{2})\,|\beta_j|^{(r+1)}} \exp\left\{-\frac{\tau\phi_k}{\beta_j^2}\right\} \tag{2}$$

Here τ and r are scale parameter and the order of the density, respectively, and (.)!! is double factorial. They find one important result in Bayesian estimation that NLPs discard spurious covariates faster than the sample size n grows, but maintains an exponential rate to detect non-zero coefficients.

In Bayesian model selection with the $p \leq n$ setting, [11] have presented model selection procedures based on NLPs with strong model selection property. As the sample size n increases, the posterior probabilities of the true model converge to 1. With the $p \gg n$ setting, [20] studied high-dimensional estimation problems and obtained the rate of convergence of the Johnson-Rossell moment and inverse moment of a model when meets Walker's condition. They have shown that for NLP in the linear models, based on the Bayesian model averaging (BMA), spurious parameters shrink either at fast polynomial or quasi-exponential rates as the sample size n increases, while non-spurious parameter estimates are not shrunk.
[22] have studied the behavior of nonlocal priors for variable selection and their consistency properties in linear regression and also proposed the scalable and efficient algorithm, Simplified Shotgun Stochastic Search with Screening (S5), to explore the massive model space. They bound the model space by placing a uniform prior on the space of the model to induce a penalty on the size of the model space in ultrahigh-dimensional setting. They need this bound to ensure that the least square estimator of a model is consistent when a model contains the true model. Similar priors have been considered in the literature by [9], [13] and [1].
For generalized linear regression with the $p \leq n$ setting, [21] established the posterior convergence rate for NLPs in a logistic regression model and propose the

Metropolis-Hastings algorithm for computation. [24] propose hyper nonlocal priors for variable selection in generalized linear models. They combine the Fisher information matrix with the Johnson-Rossell moment and inverse moment priors and assign hyper priors to the scale parameters.
They show that if $\text{diag}(\mathbf{I}_{k0}) = \mathbf{I}_{k0}$ is unit Fisher information matrix evaluated at $\boldsymbol{\beta} = \mathbf{0}$, piMOM priors will be

$$\pi_I\left(\boldsymbol{\beta}_k \mid M_k\right) = \pi_I\left(\boldsymbol{\beta}_k \mid k\right) = \frac{(\tau)^{\frac{r|k|}{2}}\left|I_{k0}\right|^{-\frac{r}{2}}}{\left(\Gamma\left(\frac{r}{2}\right)\right)^{|k|}} \times \exp\{-(\tau)\left(\boldsymbol{\beta}_k{}^{\top} I_{k0}\boldsymbol{\beta}_k\right)^{-1}\} \prod_{i=1}^{|k|}\left|\beta_{ki}\right|^{-(r+1)}. \tag{3}$$

According to [24], while these priors are useful for variable selection in GLM, but they require the specification of prior scale parameter at nonlocal prior. To overcome this difficulty, they assign the inverse gamma hyper prior to scale parameter in pmGLM and a gamma hyper prior to scale parameter in pimGLM that are able to learn about the prior scale parameter from data and provid robust inferential results. They call them, hyper nonlocal (HYN) priors.
Under certain regularity conditions, they have shown that HYN priors methods achieve variable selection consistency. That is, the posterior probabilities of the true model converge to 1 as the sample size increases. This means the prior probabilities of HYN identify the true model much faster.
With the $p > n$ settings, [16] develop Bayesian variable selection in logistic models for binary outcomes in genomic studies of the piMOM density class of NLPs. [17] propose Bayesian variable selection method using NLP (BVSNLP) for high and ultrahigh dimensional datasets with survival time as a result and using piMOM on nonzero regression coefficients.
[1] undertake high-dimensional posterior consistency properties for the class of pMOM of NLPs.
However, as far as we know, to date there have been no published manuscript in case of hyper nonlocal priors for variable selection in high-dimensional generalized linear models. Motivated by this gap, our first goal is to investigate the model selection properties of the hyper product inverse moment prior in a UDGLM.
We use piMOM priors because as [17] mention, the piMOM priors assign negligible probability to a wider range near zero than do pMOM priors. The pMOM priors decrease to zero at the polynomial rate while the piMOM priors decrease much faster with exponential rate. On the other hand, pMOM priors have tails that converge to zero at an exponential rate, while piMOM priors have heavier Cauchy-like tails. In the $p \gg n$ setting, [22] have studied consistency properties of piMOM priors for linear models, but this property does not apply to pMOM priors.
It is known that the computation problem can arise for Bayesian approaches due to the non-conjugate structure in generalized linear regression. Therefore, our second goal is to develop efficient algorithms to explore the massive posterior space. These are challenging goals of course, as the posterior distributions are not available in

closed form for this type of nonlocal priors.
First, we obtained the posteriors via Laplace approximation, and then implemented the efficient Simplified Shotgun Stochastic Search with Screening (S5) presented by [22] algorithm to explore the sparsity pattern of the regression coefficients in generalized linear regression. Finally, our proposed method was validated through simulation studies and illustrated by a real data. The Golub leukemia data ([6]) is publicly available and has good clinical annotations. We discriminated between two types of acute leukemia, myeloid (AML) and lymphoblastic (ALL). This data set contains 72 samples and 7,129 genes. The design matrix consisted of gene expression levels produced by cDNA microarrays from bone marrow samples, and was pre-processed by RMA ([8]).
The remainder of this article is organized as follows. Section 2 presents problem of modeling on the UDGLM and Methodology. In this section we compute the MAP estimator by selecting the best estimator with the $0-1$ loss function and use model selection method with a Laplace integration procedure to compute posterior model probabilities. Section 3 studies their optimal properties as theoretical properties. Section 4 studies Computational Strategy. Section 5 presents a simulation studies to investigate the performance of our proposed method under binomial models. Section 6 illustrates the application of our method with the analysis of the real data example on the Golub leukemia data to discriminate between two types of acute leukemia, myeloid (AML) and lymphoblastic (ALL). Section 7 concludes the document with discussions and possible directions for future research directions. For greater clarity, proofs are presented in the Appendix.

2 Problem modeling

2.1 Methodology

Let $\mathbf{y}_n = (y_1, \ldots, y_n)$ be an n-dimensional response vector and $\mathbf{X}_n$ be a $n \times p$ design matrix, where n is the sample size and p is the total number of covariates, $\mathbf{k} \subseteq \{0,1\}^p$ index a model consisting of a subset of columns of $\mathbf{X}_n$, $|k|$ is the cardinality of subset $\mathbf{k}$ and $\mathbf{M}$ denotes the model space that collects all the model indices $\mathbf{k}$; i.e., $\mathbf{M} = \{\mathbf{k} : \mathbf{k} \subseteq \{0,1\}^p\}$. Then, with a given link function $g(.)$, we consider model $\mathbf{k}$ of the form

$$\eta_{\mathbf{k}i} = g(E(y_i \mid x_k)) = \beta_0 + \mathbf{X}_{\mathbf{k}i}^{\top}\boldsymbol{\beta}_{\mathbf{k}}, \tag{4}$$

where β_0 denotes the intercept, $\mathbf{X}_{\mathbf{k}}$ is design matrix for model $\mathbf{k}$, $\mathbf{X}_{\mathbf{k}i}$ is the ith row of $\mathbf{X}_{\mathbf{k}}$, $\boldsymbol{\beta}_{\mathbf{k}}$ denotes the corresponding vector of nonzero regression coefficients. Thus, considering the intercept, model $\mathbf{k}$ has dimension $|\mathbf{k}| + 1$. We are interested in selecting the best subset of covariates to predict response variable, where y can follow any probability distribution in the exponential family including binomial, Poisson and negative binomial, and the $\mathbf{X}_{\mathbf{k}}$s can be continuous or discrete.

We assume that the true model exists, and is defined as the smallest model in the model space $\mathbf{M}$ that contains the true data-generating distribution. Consequently, the problem of selecting the best subset is now equivalent to the problem of identifying the true model in $\mathbf{M}$.
Then, for a given model $\mathbf{k}$, the likelihood function of $\mathbf{y}_n = (y_1, \ldots, y_n)$ given linear predictors $\boldsymbol{\eta}_\mathbf{k} = (\eta_{\mathbf{k}1}, \ldots, \eta_{\mathbf{k}n})^\top$ and dispersion parameter ϕ is

$$f(\mathbf{y}_n \mid \boldsymbol{\eta}_\mathbf{k}, \phi) = \prod_{i=1}^{n} a(y_i, \frac{\phi}{w_i}) \exp\left\{\frac{w_i(y_i\theta(\eta_{\mathbf{k}i}) - b(\theta(\eta_{\mathbf{k}i})))}{\phi}\right\} \tag{5}$$

For proposed Bayesian variable selection approach, we assume the dispersion parameter ϕ is known and assign an improper uniform prior for the intercept β_0.
According to [24] in the pimGLM prior, we assign for τ a gamma hyper prior with shape parameter a and rate parameter b. 1 induces prior dependence among the regression coefficients and difference of pimGLM and hpimGLM. The rate of decay at the null, indicates the rate of Bayes factor in favor of the null hypothesis. Meanwhile, the heaviness of the tail indicates the degree of robustness to large effect size. It shows that scale mixturing goes to zero faster than pimGLM, but preserves the same tail heaviness.
According to the equation (3), the class of hpimUGLM for unknown regression coefficients $\boldsymbol{\beta}_\mathbf{k}$ is given by

$$\pi\left(\boldsymbol{\beta}_\mathbf{k} \mid \phi, r, a, b\right) = \frac{b^a\Gamma(\frac{r|k|}{2} + a)}{\Gamma(a)\Gamma\left(\frac{r}{2}\right)^{|k|}}|I_{k0}|^{-\frac{r}{2}} \times \exp\left\{b + (\boldsymbol{\beta}_k{}^\top I_{k0}\boldsymbol{\beta}_k)^{-1}\right\}^{-\frac{r|k|}{2}-a} \times \prod_{i=1}^{|k|}|\beta_{ki}|^{-(r+1)}. \tag{6}$$

Equations (5) and (6) are valid for the $p < n$ settings. In order to use these equations in our proposed Bayesian variable selection approach in ultrahigh-dimensional settings ($p \gg n$), we need to restrict the size of largest model because of having nonsingular Gram matrix. According to [9], we consider a uniform prior on the model space and the model space prior is assumed as follows

$$\pi(\mathbf{k}) \propto I\left(|\mathbf{k}| \leq m_n\right) \tag{7}$$

where $m_n = \left(\frac{n}{\log p}\right)^\alpha$ for $0 < \alpha < 1$, is a positive integer restricting the size of the largest model, and a uniform prior is placed on the model space restricting our analysis to models having size less than or equal to m_n.
According to [19], suppose that prior distribution $\pi(\boldsymbol{\beta})$ is available then the posterior distribution $\pi\left(\boldsymbol{\beta}_k \mid \mathbf{y}_n\right)$ can be derived from the observation distribution $f\left(\mathbf{y}_n \mid \boldsymbol{\beta}_k\right)$. The posterior distribution can be used to describe the properties of $\boldsymbol{\beta}_k$. The way of selecting a best estimator is using a loss criterion. With the $0-1$ loss function, a possible estimator of $\boldsymbol{\beta}_k$ based on posterior is the maximum a posteriori (MAP) estimator which also maximizes $\pi\left(\boldsymbol{\beta}_k \mid \mathbf{y}_n\right)\pi\left(\boldsymbol{\beta}_k \mid k\right)$.
By the hierarchical Bayesian model (5) to (7) and Bayes rule, the resulting posterior

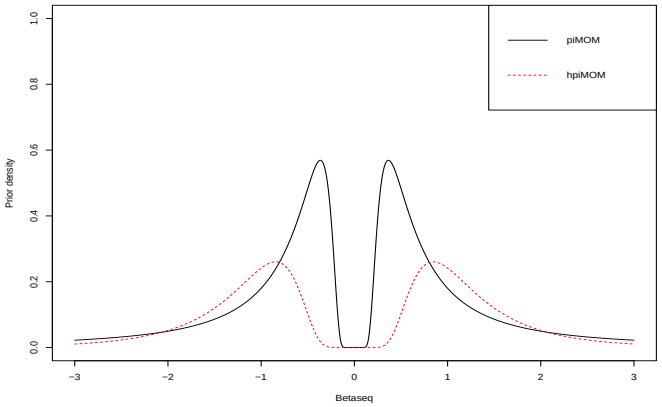

Figure 1: A depiction of Nonlocal prior density function pimGLM (solid line) and hyper product inverse moment hpimGLM (long-dash line), priors under a binomial GLM with $n = 30$, $w = 30$, $p = 1$, $r = 1$ and X is a vector with elements that are normal $N(0, 1)$.

probability for $\boldsymbol{\beta}_k$ is denoted by,

$$\pi(\boldsymbol{\beta}_k \mid \mathbf{y}_n) = \frac{\ell(\boldsymbol{\beta}_k)\pi(\boldsymbol{\beta}_k \mid k)\pi(k)}{\sum_{k\in M} \ell(\boldsymbol{\beta}_k)\pi(\boldsymbol{\beta}_k \mid k)\pi(k)} \propto \ell(\boldsymbol{\beta}_k)\pi(\boldsymbol{\beta}_k \mid k)\pi(k), \tag{8}$$

posterior mode is defined by

$$\hat{\boldsymbol{\beta}}_k = \arg\max_{\beta} \pi(\boldsymbol{\beta}_k \mid \mathbf{y}_n). \tag{9}$$

Depending on the complexity of the loss and the posterior distribution, the estimator will be determined analytically or numerically. Of course in our proposed method, the closed form of these posterior probabilities cannot be obtained due to not only the nature of GLMs but also the structure of hpimGLM prior. Therefore, special efforts need to be devoted to computational strategy. We use Laplace approximation ([23], among others; [12]; [18]) to maximize the logarithm of the unnormalized joint posterior density with one of several optimization algorithms and the goal is to estimate the posterior mode and variance of each parameter. We use the limited memory version of the BroydenFletcherGoldfarbShanno optimization algorithm (L-BFGS) which is particularly suited to problems with very large number of variables ([14])) to find the MAP, which requires only the initial values and the computation of the scoring function to obtain the posterior mode and numerical Hessian matrix. For implementing Laplace approximation, we use Maximum Likelihood Estimation (MLE) as initial values of parameters because the global posterior mode under the true model converges toward the MLE with probability one.

The first order derivatives vector $U = (U_0, ..., U_{|k|})^\top$ of the log likelihood function of GLM with respect to the coefficient vector $\boldsymbol{\beta}_k$ is $U = \frac{1}{\phi}\mathbf{X}^\top\mathbf{WM}(\mathbf{Y} - \boldsymbol{\mu})$, where $\mathbf{W}$ is the diagonal matrix of working weights and $\mathbf{M}$ is the diagonal matrix of link derivatives $\left(\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i}\right)$.
Thus, according to [24], the scoring function of posterior probability for hpimGLM can be written as follow:

$$\mathbf{S}_{hpimGLM} = \frac{1}{\phi}\mathbf{X}^\top\mathbf{WM}(\mathbf{Y}-\boldsymbol{\mu})+\frac{(r\,|k|+2a)\left[\mathrm{diag}(\hat{\boldsymbol{\beta}}^{-3}\boldsymbol{\beta}^\top)I_{k0}^{-1}\right]\boldsymbol{\beta}^{-1}}{(b+\mathrm{tr}((\mathrm{diag}(\boldsymbol{\beta}\boldsymbol{\beta}^\top)I_{k0})^{-1}))} -(r+1)1_{|k|}\otimes\hat{\boldsymbol{\beta}}^{-1}. \tag{10}$$

Where $\otimes$ denotes the Hadamard product that is element-wise multiplication, $\mathrm{diag}(\mathbf{I}_{k0}) = \mathbf{I}_{k0}$ shows unit Fisher information matrix evaluated at $\boldsymbol{\beta} = \mathbf{0}$ and $1_{|k|}$ denote the length-$|k|$ vector of ones.
In variable selection perspective, the essence is to force the estimated model to be sparse by penalizing dense models. By the hierarchical Bayesian model (5) to (7) and Bayes' rule, the resulting posterior probability for model $\mathbf{k}$ is denoted by,

$$\pi(\mathbf{k} \mid \mathbf{y}_n) = \frac{\pi(\mathbf{k})\, m_{\mathbf{k}}(\mathbf{y}_n)}{\sum_{\mathbf{j}\in M}\pi(\mathbf{j})\, m_{\mathbf{j}}(\mathbf{y}_n)}, \tag{11}$$

where $m_{\mathbf{k}}(\mathbf{y}_n)$ is the marginal density of $\mathbf{y}_n$ under model $\mathbf{k}$ given by

$$m_{\mathbf{k}}(\mathbf{y}_n) = \int \exp\left\{\ell(\boldsymbol{\beta}_{\mathbf{k}})\right\}\, \pi(\boldsymbol{\beta}_{\mathbf{k}} \mid \mathbf{k})\, \mathrm{d}\boldsymbol{\beta}_{\mathbf{k}}, \tag{12}$$

and log likelihood function is

$$\ell(\boldsymbol{\beta}_{\mathbf{k}}) = \log\left(f(\mathbf{y}_n \mid \boldsymbol{\eta}_{\mathbf{k}},\, \phi)\right) = \sum_{i=1}^{n}\left\{y_i\theta_i(\boldsymbol{\beta}_{\mathbf{k}}) - b(\theta_i(\boldsymbol{\beta}_{\mathbf{k}})) + a(y_i)\right\}. \tag{13}$$

In particular, these posterior probabilities can be used to select a model by computing the posterior mode which is defined by

$$\hat{\mathbf{k}} = \arg\max_{\mathbf{k}}\pi(\mathbf{k} \mid \mathbf{y}_n). \tag{14}$$

3 Theoretical properties

In this section, we present some theoretical results for Bayesian model selection based on the hpimUGLM. We consider a known dispersion parameter $\phi = 1$ and a canonical link function such that $\theta(\eta_i) = \eta_i$ for $i = 1, 2, ..., n$. We show that the proposed Bayesian model enjoys desirable theoretical properties.
Let $\mathbf{t} \subseteq [p] = \{1, 2, ..., p\}$ be the true model, which means that the nonzero locations of the true coefficient vector are $\mathbf{t} = \{j, j \in \mathbf{t}\}$. We consider $\mathbf{t}$ as a fixed vector.
As mentioned in [2], log likelihood function for a generalized linear model of the

exponential family is

$$\ell(\boldsymbol{\beta_k}) = \log\left(f(\mathbf{y}_n \mid \boldsymbol{\eta}_k,\, \phi)\right) = \sum_{i=1}^{n} \left\{ \frac{w_i}{\phi}(y_i\theta_i(\eta_{ki}) - b(\theta_i(\eta_{ki}))) + a(y_i, \frac{\phi}{w_i}) \right\}, \quad (15)$$

where linear predictor is $\eta_{\mathbf{k}i} = g(E(y_i \mid x_k)) = \mathbf{X}_{\mathbf{k}i}^{\top}\boldsymbol{\beta_k}$ and the Hessian (Second derivative of log likelihood function) is

$$\mathbf{H}(\boldsymbol{\beta}_k) = \left(\frac{\partial^2 \ell(\boldsymbol{\beta}_k)}{\partial\beta_k \partial\beta_k^{\top}}\right) = \sum_{i=1}^{n} \left(\frac{\partial\,(y_i - \mu_i)}{\partial\beta_k} (\frac{w_i}{\phi V(\mu_i)} \frac{x_k}{(\frac{\mathrm{d}\eta_{ki}}{\mathrm{d}\mu_i})}) + (y_i - \mu_i)\frac{\partial}{\partial\beta_k} (\frac{w_i}{\phi V(\mu_i)} \frac{x_k}{(\frac{\mathrm{d}\eta_{ki}}{\mathrm{d}\mu_i})}) \right) \quad (16)$$

The unit Fisher information matrix (standardized by sample size n) which in the prior leads to feasible and efficient computation, is expectation of the negative Hessian,

$$E(-\mathbf{H}(\boldsymbol{\beta}_k)) = \mathbf{I}(\beta_{0,k}) = \frac{1}{n}\sum_{i=1}^{n}(-\frac{w_i}{\phi V(\mu_i)} \frac{x_i x_i^{\top}}{(\frac{\mathrm{d}\eta_{ki}}{\mathrm{d}\mu_i})^2}) = \frac{1}{\phi\nu n}\mathbf{X}_k^{\top}\mathbf{W}\mathbf{X}_k \quad (17)$$

Here, $\mathbf{W}$ is an $n \times n$ diagonal matrix with diagonal elements w_i, which denotes known weight. For example, in a binomial regression model, it is the number of trials for observation i, and $\nu = V(\mu_i) = \frac{\mathrm{d}\,\mu_i}{\mathrm{d}\,\beta_k} = \frac{\mathrm{d}^2\, b(\beta_k)}{\mathrm{d}\,\beta_k^2}$ is variance function that for Binomial distribution is $\mu_i(1-\mu_i)$ where $\mu_i = \frac{\exp(x_{ik}^{\top}\beta_k)}{1+\exp(x_{ik}^{\top}\beta_k)}$. We use the notation $\mathrm{diag}(\mathbf{I}(\boldsymbol{\beta}_{0,k})) = \mathbf{I}(\boldsymbol{\beta}_{0,k})$ as unit Fisher information matrix evaluated at $\boldsymbol{\beta}$ and $\mathrm{diag}(\mathbf{I}_{k0}) = \mathbf{I}_{k0}$ as unit Fisher information matrix evaluated at $\boldsymbol{\beta} = \mathbf{0}$.

We consider the following regularity conditions for theoretical properties of our posterior:
Condition (A1): For some $0 < \alpha < 1$,

$$\log p = \mathrm{O}(n^{\alpha}) \quad (18)$$

$$m_n = (\frac{n}{\log p})^{\alpha} \quad (19)$$

$$0 < \tau \leq \log p \quad (20)$$

Condition (A1) ensures that our proposed method can accommodate high dimensions where the number of covariates grows at a sub-exponential rate of n. Also specifies the parameter m_n in the uniform model space prior that restricts our analysis on a set of reasonably large models. The scale parameter τ in the nonlocal prior density reflects the dispersion of the nonlocal prior density around zero, and implicitly determines the size of the regression coefficients that will be shrunk to zero.

Condition (A2): For some $0 < \alpha < 1$,

$$\max_{i,k} |x_{ik}| \ \leq 1 \tag{21}$$

$$\lim_{p\to\infty} \sum_{k=1}^{p} |\beta_k^*| \ < \infty \tag{22}$$

$$\|\beta_{0,t}\|_2^2 = \mathrm{O}((\log p)^{\alpha}) \tag{23}$$

In condition (A2), for simplicity we will assume that all covariates are bounded and standardized such that $\max_{i,k} |x_{ik}| \ \leq 1$ for all $k = 1, ..., m_n$. The regression parameter $\boldsymbol{\beta}^*$ corresponding to the true model are bounded, which satisfies some "sparseness" conditions, when most components of $\boldsymbol{\beta}^*$ are very small in magnitude. The last assumption in this Condition says that the magnitude of true regression parameter is bounded above $(\log p)^{\alpha}$ up to some constant, which allows the magnitude of regression parameter to increase to infinity.

Condition (A3): For some $\lambda > 0$, suppose the ordered eigenvalues $0 \leq \lambda_1 \leq \lambda_2 \leq \ ... \ \leq \lambda_{\max}$ of unit Fisher information matrix $\mathbf{I}(\boldsymbol{\beta}_{0,k})$ and the Gram matrix $\left(\frac{X_{\mathbf{k}}^{\top} X_{\mathbf{k}}}{n}\right)$ over model $\mathbf{k}$, then,

$$0 < \lambda \leq \min_{\mathbf{k}\,:\,|\mathbf{k}|\leq m_n} \lambda_1 \left(\mathbf{I}(\boldsymbol{\beta}_{0,k})\right) \ \leq \max_{\mathbf{k}\,:\,|\mathbf{k}|\leq m_n} \lambda_{\max} \left(\frac{X_{\mathbf{k}}^{\top} X_{\mathbf{k}}}{n}\right) \leq (\log p)^{\alpha} \,. \tag{24}$$

As [15] has noted, restricted eigenvalue conditions are routinely assumed in high-dimensional theory to guarantee some level of curvature of the objective function and are satisfied with high probability for sub-Gaussian design matrices.

Also the minimum of the minimum eigenvalue of $\mathbf{I}(\boldsymbol{\beta}_{k0})$ and the maximum of the maximum eigenvalue of the Gram matrix $\left(\frac{X_{\mathbf{k}}^{\top} X_{\mathbf{k}}}{n}\right)$ are bounded above and below overall sub models $\mathbf{k}$ with $|\mathbf{k}| \leq m_n$ are allowed to decrease with increasing n and p and causes none singularity and invertible in symmetric matrixes $\mathbf{I}_{k0}$ and $\mathbf{I}(\boldsymbol{\beta}_{0,k})$, and consistency for posterior probability.

Condition (A4)(Beta-min condition)**:** Let $\mathbf{t} \subseteq [p] = \{1, 2, ..., p\}$ be the true model, for some constant $C > 0$,

$$\min_{\mathbf{k}\in t} \beta_{\mathbf{k}}^2 \ \geq \ C \max \left\{ \frac{|t| \ \max\limits_{\mathbf{k}\,:\,|\mathbf{k}|\leq t} \lambda_{\max} \left(\frac{X_{\mathbf{k}}^{\top} X_{\mathbf{k}}}{n}\right) \ \log p}{n} \ , \ \frac{1}{\log p} \right\} \tag{25}$$

The beta-min condition is a lower bound for nonzero regression parameters. In general, this type of condition is necessary for catching every nonzero regression parameter. Due to Conditions (A1) and (A3), the right-hand side of Condition (A4) decreases to zero as $n \to \infty$ thus it allows the smallest nonzero coefficients to tend to zero as we observe more data.

Suppose conditions (A1) - (A4) hold. Let $\pi(\mathbf{t} \mid \mathbf{y}_n)$ denote the posterior probability

of the true model obtained under the hyper product inverse moment nonlocal prior (hpimUGLM) over coefficients. Also, assume a uniform prior on all models of size less than or equal to $\mathbf{m}_n$, i.e., $\pi(\mathbf{k}) \propto \mathrm{I}(|\mathbf{k}| \leq \mathbf{m}_n)$. Then, $\pi(\mathbf{t} \mid \mathbf{y}_n)$ converges to one in probability as n goes to ∞.

$$\pi(\mathbf{t} \mid \mathbf{y}_n) \xrightarrow{P} 1 \ , \ as \ n \to \infty. \tag{26}$$

We know that

$$\pi(\mathbf{t} \mid \mathbf{y}_n) = \pi(\mathbf{t} = \mathbf{k} \mid \mathbf{y}_n) + \pi(\mathbf{t} \underset{\neq}{\subset} \mathbf{k} \mid \mathbf{y}_n) + \pi(\mathbf{t} \neq \mathbf{k} \mid \mathbf{y}_n) \tag{27}$$

Let $\mathbf{K}_1 = \left\{ \mathbf{k} : \ |\mathbf{k}| \leq \mathbf{m}_n \ , \ \mathbf{t} \underset{\neq}{\subset} \mathbf{k} \right\}$ and $\mathbf{K}_2 = \{ \mathbf{k} : \ |\mathbf{k}| \leq \mathbf{m}_n \ , \ \mathbf{t} \neq \mathbf{k} \}$, by Bayes' rule and the resulting posterior probability for model $\mathbf{k}$, this equation is denoted by

$$\pi(\mathbf{t} \mid \mathbf{y}_n) = \frac{\pi(\mathbf{t})\, m_{\mathbf{t}}(\mathbf{y}_n)}{\sum_{\mathbf{k}} \pi(\mathbf{k})\, m_{\mathbf{k}}(\mathbf{y}_n)} = \left[1 + \sum_{\mathbf{k}:\, \mathbf{k} \in \mathbf{K}_1} \frac{\pi(\mathbf{k})\, m_{\mathbf{k}}(\mathbf{y}_n)}{\pi(\mathbf{t})\, m_{\mathbf{t}}(\mathbf{y}_n)} + \sum_{\mathbf{k}:\, \mathbf{k} \in \mathbf{K}_2} \frac{\pi(\mathbf{k})\, m_{\mathbf{k}}(\mathbf{y}_n)}{\pi(\mathbf{t})\, m_{\mathbf{t}}(\mathbf{y}_n)} \right]^{-1}. \tag{28}$$

Following [1] expression, in order to proof "Strong selection consistency Theorem", we have to show two more theorems, the first one is "No super set theorem":

$$\sum_{\mathbf{k}:\, \mathbf{k} \in \mathbf{K}_1} \frac{\pi(\mathbf{k})\, m_{\mathbf{k}}(\mathbf{y}_n)}{\pi(\mathbf{t})\, m_{\mathbf{t}}(\mathbf{y}_n)} = \sum_{\mathbf{k}:\, \mathbf{k} \in \mathbf{K}_1} \frac{\pi(\mathbf{k} \mid \mathbf{y}_n)}{\pi(\mathbf{t} \mid \mathbf{y}_n)} \xrightarrow{P} 0 \quad as \quad n \to \infty \tag{29}$$

It says that, asymptotically, posterior will not include unnecessarily many variables and not over fit the model.
The second theorem is "**Posterior ratio consistency theorem**":

$$\sum_{\mathbf{k}:\, \mathbf{k} \in \mathbf{K}_2} \frac{\pi(\mathbf{k})\, m_{\mathbf{k}}(\mathbf{y}_n)}{\pi(\mathbf{t})\, m_{\mathbf{t}}(\mathbf{y}_n)} = \sum_{\mathbf{k}:\, \mathbf{k} \in \mathbf{K}_2} \frac{\pi(\mathbf{k} \mid \mathbf{y}_n)}{\pi(\mathbf{t} \mid \mathbf{y}_n)} \xrightarrow{P} 0 \quad as \quad n \to \infty \tag{30}$$

It shows that, with an appropriate lower bound specified in Condition (A4), the true model $\mathbf{t}$ will be the mode of the posterior. Posterior ratio consistency is a useful property especially when we are interested in the point estimation with the posterior mode. The Proofs are available in Appendix.

4 Computational Strategy

4.1 Computing posterior probability of models

The following Algorithm 1 illustrates the procedure employed to compute the posterior probability of models.

[16] have employed a Birth-death scheme in Metropolis-Hastings to sample from the posterior distribution on the model space to obtain a sequence of sampled models and estimate the MAP model.

Algorithm 1 Computing posterior probability of models

[1] Let M is the model space. The following steps are repeated as k converges to the solution: Choose $k \in M$ as a given model. Then log likelihood function $\ell(\boldsymbol{\beta}_{\mathbf{k}})$, the priors $\pi(\boldsymbol{\beta}_{\mathbf{k}})$ and $\pi(\mathbf{k})$ are available. Define the posterior probability for $\boldsymbol{\beta}_{\mathbf{k}}$ by

$$\pi(\boldsymbol{\beta}_{\mathbf{k}} \mid \mathbf{y}_n) \propto \ell(\boldsymbol{\beta}_{\mathbf{k}})\pi(\boldsymbol{\beta}_{\mathbf{k}} \mid \mathbf{k})\pi(\mathbf{k}).$$

For model $\mathbf{k}$, obtain the MLEs as initial values of parameters in optimizing the posterior distribution. Estimate the posterior mode by $\tilde{\boldsymbol{\beta}}_{\mathbf{k}} = \arg\max_{\boldsymbol{\beta}} \pi(\boldsymbol{\beta}_{\mathbf{k}} \mid \mathbf{y}_n)$ with the optimization algorithm (L-BFGS). Obtain the numerical Hessian matrix of the logarithm of the posterior of $\boldsymbol{\beta}_{\mathbf{k}}$ evaluated at $\tilde{\boldsymbol{\beta}}_{\mathbf{k}}$. Compute the marginal likelihood of model $\mathbf{k}$ observing $\mathbf{y}_n$ by Laplace approximation

$$\tilde{m}_{\mathbf{k}}(\mathbf{y}_n) = (2\pi)^{\frac{|k|}{2}} \left|-\hat{\mathbf{H}}_k\right|^{-\frac{1}{2}} \exp[\ell(\tilde{\boldsymbol{\beta}}_{\mathbf{k}})]\pi(\tilde{\boldsymbol{\beta}}_{\mathbf{k}} \mid \mathbf{k}).$$

Compute the approximate of posterior probability for model $\mathbf{k}$ by

$$\tilde{\pi}(\mathbf{k} \mid \mathbf{y}_n) \propto \frac{\pi(\mathbf{k})\tilde{m}_{\mathbf{k}}(\mathbf{y}_n)}{\sum_{\mathbf{j}\in M} \pi(\mathbf{j})\tilde{m}_{\mathbf{j}}(\mathbf{y}_n)}.$$

For any $k \in M$, employ $\tilde{\pi}(\mathbf{k} \mid \mathbf{y}_n)$ as the criterion for Bayesian variable selection.

One way to avoid complicated computational schemes such as Metropolis-Hastings algorithm, is the [7] method which approximates the posterior density $\pi(\boldsymbol{\beta}_{\mathbf{k}} \mid \mathbf{y}_n)$, by normal distribution with the posterior mode $(\tilde{\boldsymbol{\beta}}_{\mathbf{k}})$ as mean and inverse of Hessian matrix $(\hat{\mathbf{H}}_k)^{-1}$ as covariance matrix denoted by $\pi(\boldsymbol{\beta}_{\mathbf{k}} \mid \mathbf{y}_n) \sim N(\tilde{\boldsymbol{\beta}}_{\mathbf{k}}, \hat{\mathbf{H}}_k^{-1})$. Since we obtained the approximated posterior density of the coefficient vector $(\boldsymbol{\beta}_{\mathbf{k}})$ in a conjugate family, we can simply perform Gibbs sampler for its evaluation.
With the $p \gg n$ setting, full posterior sampling using the existing Markov chain Monte Carlo (MCMC) algorithms is very inefficient and often impractical from a point of view. To achieve dimension reduction, we use prior density with some characteristic to shrink each coefficient. Therefor, we use another stochastic algorithm to search the model space by rapidly identifying regions with high posterior probability and finding the maximum a posteriori (MAP) model.

4.2 Simplified shotgun stochastic search algorithm with screening (S5) for GLMs

In ultrahigh-dimensional settings, to increase the efficiency of exploring the model space, we use the S5 algorithm. S5 is proposed by [22] for variable selection in linear regression problems. It is a stochastic search method that screens covariates at each step. Screening is the essential part of the S5 algorithm. According to [4] in linear regression, screening is based on the correlation between the excluded covariates and the residuals of the regression using the current model. The concept

of screening covariates for GLMs response data is proposed in [5] and is defined as a more general version of the independent learning with ranking the maximum marginal likelihood estimator (MMLE) or the maximum marginal likelihood itself. Let k be the current model and complement of k contains columns of the design matrix that are not present in the current model. The S5 algorithm for GLM data works as follows:
At each step the $d = 2[\log p]$ covariates with highest maximum likelihood (estimator) are candidates to be added to the current model and called the addition set, Γ^{+}_{scr}. The deletion set, Γ^{-}, contains the current model except that one variable is removed. From the current model, we consider moves to each of its neighbors in Γ^{+}_{scr} and Γ^{-} with a probability proportional to the marginal probabilities of these neighboring models. To avoid local maxima, the model probabilities used in S5 are raised to the power of $1/t^l$, where t^l is the lth temperature in an annealing schedule in which "temperatures"decrease. To increase the number of visited models, a specified number of iterations are performed at each temperature. Therefore, the model with the highest posterior probability in visited models is identified as the HPPM.
The following Algorithm 2 illustrates the procedure which has been employed for the Simplified Shotgun Stochastic Search with Screening (S5) for GLMs.

Algorithm 2 Simplified Shotgun Stochastic Search with Screening (S5) for GLMs

[1] Choose S as an initial value for screening size of variables. Set a temperature schedule $t_1 > t_2 > ... > t_S > 0$ to avoid local maxima. Choose $ITER$ as a specified number of iterations at each temperature to increase the number of visited models. Choose an initial model $k^{(1;1)}$ and a set of variables after screening $S_{k^{(1;1)}}$ based on $k^{(1;1)}$. Choose an initial number c_0 for repetition of the S5 algorithm. For $l = 1$ in $l = s$

For i in $1, ..., ITER - 1$

Compute all $\pi(k \mid y_n)$ for all $k \in \text{nbd}_{scr}(k^{(i,l)}) = \{\Gamma^{+}_{scr}\ , \Gamma^{-}\}$

Sample k^{+} and k^{-}, from Γ^{+}_{scr} and Γ^{-}, with probabilities proportional to $\pi(k \mid y_n)^{\frac{1}{t_l}}$

Sample $k^{(i+1,l)}$ from $\{k^{+}, k^{-}\}$, with probability proportional to $\left\{\pi(k^{+} \mid y_n)^{\frac{1}{t_l}}, \pi(k^{-} \mid y_n)^{\frac{1}{t_l}}\right\}$

Update the set of considered variables $S_{k^{(i+1,l)}}$ to be the union of variables in $k^{(i+1,l)}$ and top, $d = 2[\log p]$ variables according to highest maximum likelihood estimator.

end for

end for.

5 Simulation Studies

We applied our method to both simulated data and real data, to investigate the performance of the proposed method. Iterative Sure Independence Screening (ISIS) was introduced by [4] to reduce the computation in ultra-high dimensional variable selection. It refers to ranking covariates according to marginal utility, namely, each covariate is used independently as a predictor to decide its usefulness for predicting the response.
In the ISIS-SCAD/LASSO method, first the Iterative Sure Independence Screening for different variants implements, and then fits the final regression model using the SCAD/LASSO regularized log likelihood for the variables picked by ISIS. The ISIS-SCAD/LASSO has proven to be among the most successful model selection procedures used in practice. To run ISIS-SCAD/LASSO, we used the R package 'SIS'([3]) available from CRAN.
In simulation studies, Let $n = 100$, $p = 120$, X be the design matrix and for a true model K, the response vector represents a sequence of Bernoulli samples with probability of success $\pi_i = \frac{e^{X_{ik}^\top \beta_k}}{1+e^{X_{ik}^\top \beta_k}}$.
Elements of the design matrix X were sampled from a multivariate normal distribution with mean 0 and covariance matrix Σ, under the following three different cases of Σ:

- Case (1): Compound symmetry design, where $\Sigma_{ij} = 0.5$ if $i \neq j$ and $\Sigma_{ii} = 1$, for all $1 \leq i \leq j \leq p$.
- Case (2): Autoregressive correlated design, where $\Sigma_{ij} = 0.5^{|i-j|}$, for all $1 \leq i \leq j \leq p$.
- Case (3): Isotropic design, where $\Sigma = I_p$ i.e., no correlation imposed between different covariates.

With the fixed true model $\mathbf{t} = (1, 2, 3, 4, 5)$ and coefficient $\boldsymbol{\beta}_{\mathbf{t}}^0 = (0.5, 0.75, 1, 1.25, 1.5)$, the signs were randomly determined with probability one-half.
From a practical perspective, in order to have efficient and feasible algorithms and average computation time to first hit the MAP model, the variable selection procedure in all algorithms was run 50 times and each time with different random seed numbers in order to generate different datasets. In each trial, true and false positive values for UDGLM and ISIS-SCAD/LASSO were counted by comparing the selected model with the true one. TP and FP rates were defined as the average true and false positive values over 50 trials. A true positive, TP, was defined to be the number of variables that were correctly selected, while false positives, FP, were the number of variables that were mistakenly selected.
To evaluate the performance of variable selection, the precision, sensitivity, specificity, Matthew's correlation coefficient (MCC), mean-squared prediction error (MSPE) and mean squared error (MSE) of success probability were computed. The criteria

are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{31}$$

$$\text{Sensitivity} = \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{32}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{33}$$

$$\text{MCC} = \frac{\text{TP} \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{34}$$

$$\text{MSPE} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\hat{y}_i - y_{test,i})^2 \tag{35}$$

In order to calculate the MSPE, If we consider $\hat{\boldsymbol{\beta}}$ as the estimated regression coefficients based on the training samples for each method, $\hat{y}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ will be estimated response based on that method. In Simulation Studies for Bayesian methods, the usual GLM estimates based on the selected support are used as $\hat{\boldsymbol{\beta}}$. We generated test samples y_test with $n$test $= 50$ to calculate the MSPE.
We compared the mean squared error in estimating the probability of success for each binary observation. The point estimates of the regression coefficients were estimated as the posterior mode under the highest posterior probability model. Note that the prediction of the response vector involves both coefficient estimation and variable selection. The MSE of success probability was defined as follows:

$$MSE(\hat{\pi}) = \frac{1}{n} \sum_{i=1}^{n} (\hat{\pi}_i - \pi_i)^2 \tag{36}$$

Following tables summarize the results of applying UDGLM, ISIS-SCAD and ISIS-LASSO approaches to the simulated data.

Table 1: The summary statistics to evaluate the performance of variable selection in three methods UDGLM ,ISIS-SCAD and ISIS-LASSO in Case (1) (*Compound symmetry* covariance of the design matrix).

	Precision	Sensitivity	Specificity	MCC	MSPE	Selected variables
UDGLM	1	1	1	1	0.2338	X1 X5
ISIS-SCAD	0.7246377	0.8333333	0.525	0.3795661	1.0621	X5
ISIS-LASSO	0.7323944	0.8666667	0.525	0.4228575	1.673	X1 X5

Table 2: The summary statistics to evaluate the performance of variable selection in three methods UDGLM ,ISIS-SCAD and ISIS-LASSO in Case (2) (*Autoregressive* covariance of the design matrix).

	Precision	Sensitivity	Specificity	MCC	MSPE	Selected variables
UDGLM	1	1	**1**	1	0.2286	X2 X5
ISIS-SCAD	0.7727273	0.85	0.625	0.4912333	1.9065	X1 X5
ISIS-LASSO	0.7833333	0.7833333	0.675	0.4583333	2.4506	X1 X5 X115

Table 3: The summary statistics to evaluate the performance of variable selection in three methods UDGLM ,ISIS-SCAD and ISIS-LASSO in Case (3) (*Isotropic* covariance of the design matrix).

	Precision	Sensitivity	Specificity	MCC	MSPE	Selected variables
UDGLM	1	1	1	1	0.2289	X1 X5
ISIS-SCAD	0.7246377	0.8333333	0.525	0.3795661	1.0796	X5
ISIS-LASSO	0.7323944	0.8666667	0.525	0.4228575	1.6909	X1 X5

- Based on the simulation results, we could see that under the *Compound symmetry* covariance of the design matrix, the UDGLM method worked better than the frequentist methods. They had higher precision, Sensitivity, Specificity, MCC, and lower MSPE than others. Generally ISIS-SCAD method suffered from lower precision, sensitivity, and MCC but still had lower mean-squared prediction error (MSPE) compared with the frequentist approach ISIS-LASSO.

- When the covariance of the design matrix is *autoregressive*, ISIS-LASSO method had higher precision and specificity and mean-squared prediction error (MSPE) but had lower Sensitivity and MCC than the ISIS-SCAD method. Again the UDGLM strategy worked better than the frequentist procedures.

- In case of *Isotropic* covariance of the design matrix, the UDGLM strategy worked superior than the others. ISIS-LASSO method had higher precision, Sensitivity, specificity, MCC and mean-squared prediction error (MSPE) compared with the ISIS-SCAD method.

6 Real data analysis

[6] described a generic approach to cancer classification based on gene expression monitoring by DNA microarrays and applied it to human acute leukemia as a test case. We applied our method in hpimUGLM to the Golub leukemia data. The goal of our analysis for these data was to discriminate between two types of acute leukemia, myeloid (AML) and lymphoblastic (ALL). The design matrix consisted of gene expression levels produced by cDNA microarrays from bone marrow samples, and was pre-processed by RMA ([8]). There are 72 samples and 7129 genes in the data set.
Following [6], we split the data into training and test sets. The testing data is from 34 patients with acute leukemia (20 in class ALL and 14 in class AML) and the training data is from 38 patients with acute leukemia (27 in class ALL and 11 in class AML). The hpimUGLM method restricts size of the largest model by $m_n = (\frac{34}{\log(7129)})^{\frac{1}{2}} = 2.97$, then maximum number of variables which could be selected in the model is two variables.

Table 4 summarizes the results of applying the UDGLM, ISIS-SCAD and ISIS-LASSO approaches to the real data. Because of having higher precision, Sensitivity, Specificity, MCC, and lower MSPE and MSE, the UDGLM method worked better than the frequentist methods. In the first steps of variable selection, it behaved as similar as the ISIS-SCAD method but following up, it changed one of the selected variables. Among of the frequentist methods, ISIS-LASSO method had higher sensitivity and MCC but ISIS-SCAD had lower mean-squared prediction error (MSPE) and mean squared error of success probability (MSE).

Table 4: The summary statistics to evaluate the performance of variable selection on the testing set of Golub leukemia data based in three methods UDGLM ,ISIS-SCAD and ISIS-LASSO.

	Precision	Sensitivity	Specificity	MCC	MSPE	Selected variables	MSE($\hat{\pi}$)
UDGLM	1	1	1	1	0.3529	X461 X1744	0.06718414
ISIS-SCAD	1	0.6428571	1	0.7171372	134.5048	X1834 X2020	0.1733124
ISIS-LASSO	1	0.7142857	1	0.7715167	31139.71	X3320 X4847	0.1344605

The comparison between residuals of binary observations and predicted values in methods is shown in figure 2. As in the comparisons of MSPE and MSE, this figure shows UDGLM is preferred to ISIS-SCAD and ISIS-LASSO in estimating of binary observations because of having lower residuals.

7 Conclusion

The purpose of the current study is Bayesian inference with the UDGLM prior specification over regression coefficients to find MAP estimator and perform variable selection in ultrahigh-dimensional generalized linear models settings. The model selection consistency of UDGLM prior is established under mild conditions. The efficient model search strategy for the increasingly large model space, the Simplified shotgun stochastic search algorithm with screening (S5), can be used for the implementation of this approach.

Because the explicit form of the marginal likelihood of the nonlocal priors is not available, we have used the Laplace approximation throughout the paper. The Bayesian procedures require numerical optimization to obtain the maximum a posteriori estimate used in the evaluation of the Laplace approximation to the marginal density of each model visited. However, the procedures used to search the model space, given the value of a marginal density or objective function, are approximately equally complex for procedure.

The main conclusion to be drawn from our simulation and real data studies indicate that the proposed method, UDGLM, because of lower MSPE and MSE compared to the frequentist methods, has better performance under different configurations with different data generation mechanisms for variable selection.

A promising avenue for future research is the extension of our proposed method to the UDGLM prior to perform variable selection in ultrahigh-dimensional generalized linear models settings. We used the uniform prior on the model space, It is

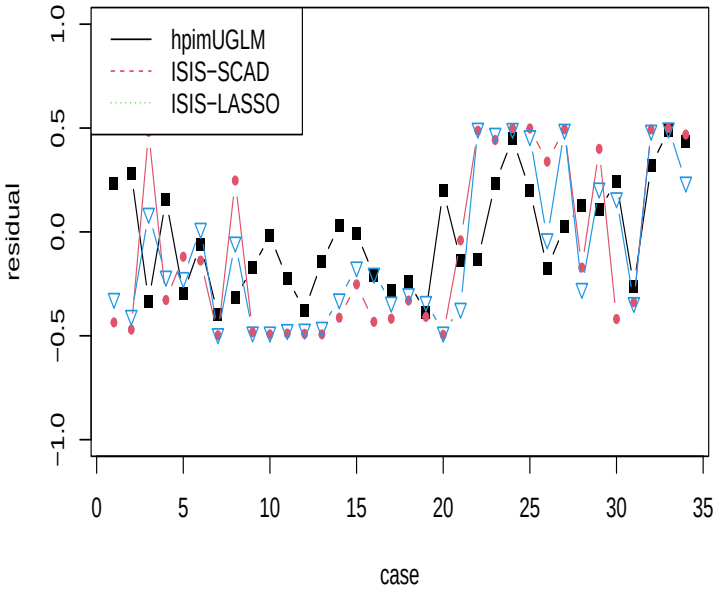

Figure 2: Residual of binary observations and predicted values in three methods UDGLM, ISIS-SCAD and ISIS-LASSO.

suggested additional directions for future research that the beta-binomial prior can be used on the model space.

Compliance with Ethical Standards

• **Funding:** The authors did not receive support from any organization for the submitted paper, and this research received no specific grant from any funding agency in the public, commercial, or not-forprofit sectors.

• **Conflicts of interest/Competing interests :** The authors have no Conflicts of interest/Competing interests to declare that are relevant to the content of this article.

• **Ethical conduct:** Not applicable because the data available to the public at
$(http://www.affymetrix.com/analysis/download\_center2.affx)$
where 14 human genes were spikedin at concentrations ranging from 0 to 1024 pM and one from GeneLogic
$(http://qolotus02.genelogic.com/datasets.nsf/)$
where 11 control cRNA fragments were spiked-in at concentrations ranging from 0 to 100 pM.

8 Appendix

Suppose conditions (A1), (A2) hold, then

$$\sum_{\mathbf{k}:\,\mathbf{k}\in\mathbf{K}_1} \frac{\pi(\mathbf{k})\, m_{\mathbf{k}}(\mathbf{y}_n)}{\pi(\mathbf{t})\, m_{\mathbf{t}}(\mathbf{y}_n)} \xrightarrow{P} 0 \quad as \quad n\to\infty.$$

Proof. Using Taylor's expansion on the log-likelihood $\ell(\boldsymbol{\beta}_{\mathbf{k}})$ around $\hat{\boldsymbol{\beta}}_{\mathbf{k}}$, which is the MLE of $\boldsymbol{\beta}_{\mathbf{k}}$, and the Fisher information matrix $\mathbf{I}(\boldsymbol{\beta}_{0,k})$ is expectation of the negative Hessian under the model $\mathbf{k}$, we have

$$\ell(\boldsymbol{\beta}_{\mathbf{k}}) = \ell(\hat{\boldsymbol{\beta}}_{\mathbf{k}}) - \frac{1}{2}(\boldsymbol{\beta}_{\mathbf{k}}-\hat{\boldsymbol{\beta}}_{\mathbf{k}})^{\top}\mathbf{I}(\boldsymbol{\beta}_{0,k})(\boldsymbol{\beta}_{\mathbf{k}}-\hat{\boldsymbol{\beta}}_{\mathbf{k}}) + \mathrm{O}(n^{-1})$$

For $\tilde{\boldsymbol{\beta}}_{\mathbf{k}}$, such that $\left\|\tilde{\boldsymbol{\beta}}_{\mathbf{k}}-\hat{\boldsymbol{\beta}}_{\mathbf{k}}\right\|_2 \le \left\|\boldsymbol{\beta}_{\mathbf{k}}-\hat{\boldsymbol{\beta}}_{\mathbf{k}}\right\|_2$

$$\ell(\boldsymbol{\beta}_{\mathbf{k}}) = \ell(\hat{\boldsymbol{\beta}}_{\mathbf{k}}) - \frac{1}{2}(\boldsymbol{\beta}_{\mathbf{k}}-\hat{\boldsymbol{\beta}}_{\mathbf{k}})^{\top}\mathbf{I}(\tilde{\boldsymbol{\beta}}_{0,k})(\boldsymbol{\beta}_{\mathbf{k}}-\hat{\boldsymbol{\beta}}_{\mathbf{k}}) + \mathrm{O}(n^{-1})$$

, by Condition (A4), for any $\mathbf{k}\in\mathbf{K}_1$, $C_1, C_2 > 0$ and any $\boldsymbol{\beta}_{\mathbf{k}}$ such that

$$\|\boldsymbol{\beta}_{\mathbf{k}} - \boldsymbol{\beta}_{0,k}\|_2 < C_1\sqrt{\frac{|\mathbf{k}|\ \max\limits_{\mathbf{k}\,:\,|\mathbf{k}|\le m_n} \lambda_{\max}\left(\frac{X_{\mathbf{k}}^{\top}X_{\mathbf{k}}}{n}\right)\ \log p}{n}} = C_1 a_n$$

, and $\varepsilon = C_2\sqrt{\frac{|m_n|\ \max\limits_{\mathbf{k}\,:\,|\mathbf{k}|\le m_n} \lambda_{\max}\left(\frac{X_{\mathbf{k}}^{\top}X_{\mathbf{k}}}{n}\right)\ \log p}{n}} = \mathrm{o}(1)$ with probability 1,

$$\ell(\boldsymbol{\beta}_{\mathbf{k}}) - \ell(\hat{\boldsymbol{\beta}}_{\mathbf{k}}) \ \le\ -\frac{1-\varepsilon}{2}(\boldsymbol{\beta}_{\mathbf{k}} - \hat{\boldsymbol{\beta}}_{\mathbf{k}})^{\top}\,\mathbf{I}(\boldsymbol{\beta}_{0,k})(\boldsymbol{\beta}_{\mathbf{k}} - \hat{\boldsymbol{\beta}}_{\mathbf{k}}) \tag{A1}$$

For $\boldsymbol{\beta}_{\mathbf{k}}$ such that $\left\|\boldsymbol{\beta}_{\mathbf{k}} - \hat{\boldsymbol{\beta}}_{\mathbf{k}}\right\|_2 = \frac{C_1 a_n}{2}$ (because of concavity of the log-likelihood of GLMs also for $\boldsymbol{\beta}_{\mathbf{k}}$ that $\left\|\boldsymbol{\beta}_{\mathbf{k}} - \hat{\boldsymbol{\beta}}_{\mathbf{k}}\right\|_2 > \frac{C_1 a_n}{2}$) and Condition (A3),

$$\begin{aligned}
\ell(\boldsymbol{\beta}_{\mathbf{k}}) - \ell(\hat{\boldsymbol{\beta}}_{\mathbf{k}}) \ &\le\ -\frac{1-\varepsilon}{2}\left\|\boldsymbol{\beta}_{\mathbf{k}} - \hat{\boldsymbol{\beta}}_{\mathbf{k}}\right\|_2^2\ \lambda_{\min}(\mathbf{I}(\boldsymbol{\beta}_{0,k})) \\
&\le\ -\frac{1-\varepsilon}{2}\ \frac{C_1^2 a_n^2}{4}\ n\lambda \\
&=\ -\frac{1-\varepsilon}{8}\, C_1^2\,\lambda\,|\mathbf{k}|\ \max_{\substack{\mathbf{k}\,:\,|\mathbf{k}|\le m_n \\ n\to\infty}} \lambda_{\max}\left(\frac{X_{\mathbf{k}}^{\top}X_{\mathbf{k}}}{n}\right)\ \log p \to\ -\infty
\end{aligned}$$

Define, $I = \left\{\boldsymbol{\beta}_{\mathbf{k}} :\ \left\|\boldsymbol{\beta}_{\mathbf{k}} - \hat{\boldsymbol{\beta}}_{\mathbf{k}}\right\|_2 < \frac{C_1 a_n}{2}\right\}$ and $I^c = \left\{\boldsymbol{\beta}_{\mathbf{k}} :\ \left\|\boldsymbol{\beta}_{\mathbf{k}} - \hat{\boldsymbol{\beta}}_{\mathbf{k}}\right\|_2 \ge \frac{C_1 a_n}{2}\right\}$.
Using the hyper product inverse moment nonlocal prior (hpimUGLM) over coefficients,

$$\pi(\boldsymbol{\beta}_k \mid k) = \frac{(\tau)^{\frac{r|k|}{2}}\,|I_{k0}|^{-\frac{r}{2}}}{(\Gamma(\frac{r}{2}))^{|k|}} \times \exp\left\{-(\tau)(\boldsymbol{\beta}_k^{\top} I_{k0}\boldsymbol{\beta}_k)^{-1}\right\}\prod_{i=1}^{|k|}|\boldsymbol{\beta}_{ki}|^{-(r+1)}$$

The marginal density of $\mathbf{y}_n$ under model $\mathbf{k}$, in $\mathbf{K}_1 = \left\{\mathbf{k} :\ |\mathbf{k}| \le \mathbf{m}_n, \mathbf{t} \underset{\neq}{\subset} \mathbf{k}\right\}$ is

$$\begin{aligned}
m_{\mathbf{k}}(\mathbf{y}_n) &= \int_{\boldsymbol{\beta}_{\mathbf{k}}} \exp\{\ell(\boldsymbol{\beta}_{\mathbf{k}})\}\ \pi(\boldsymbol{\beta}_{\mathbf{k}} \mid \mathbf{k})\,\mathrm{d}\boldsymbol{\beta}_{\mathbf{k}} \\
&= \int_{\boldsymbol{\beta}_{\mathbf{k}}} \exp\{\ell(\boldsymbol{\beta}_{\mathbf{k}})\}\ \frac{(\tau)^{\frac{r|k|}{2}}\,|I_{k0}|^{-\frac{r}{2}}}{(\Gamma(\frac{r}{2}))^{|k|}} \times \exp\left\{-(\tau)(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})^{-1}\right\}\prod_{i=1}^{|\mathbf{k}|}|\boldsymbol{\beta}_{\mathbf{k}i}|^{-(r+1)}\ \mathrm{d}\boldsymbol{\beta}_{\mathbf{k}}
\end{aligned}$$

$$
\begin{aligned}
&\leq_{(A1)} \int_{\boldsymbol{\beta}_{\mathbf{k}}} \exp\left\{\ell(\hat{\boldsymbol{\beta}}_{\mathbf{k}}) - \frac{1-\varepsilon}{2}\,(\boldsymbol{\beta}_{\mathbf{k}} - \hat{\boldsymbol{\beta}}_{\mathbf{k}})^{\top}\mathbf{I}(\boldsymbol{\beta}_{0,\mathbf{k}})(\boldsymbol{\beta}_{\mathbf{k}} - \hat{\boldsymbol{\beta}}_{\mathbf{k}})\right\} \frac{(\tau)^{\frac{r|k|}{2}}|I_{k0}|^{-\frac{r}{2}}}{(\Gamma(\frac{r}{2}))^{|k|}} \\
&\times \exp\left\{-(\tau)(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})^{-1}\right\} \prod_{i=1}^{|\mathbf{k}|} |\boldsymbol{\beta}_{\mathbf{k}i}|^{-(r+1)}\ \mathrm{d}\boldsymbol{\beta}_{\mathbf{k}} \\
&= \exp\left\{\ell(\hat{\boldsymbol{\beta}}_{\mathbf{k}})\right\} \\
&\times \left\{\left[\int_{\boldsymbol{\beta}_{\mathbf{k}}:\boldsymbol{\beta}_{\mathbf{k}}\in\mathbf{I}} \exp\left\{-\frac{1-\varepsilon}{2}\,(\boldsymbol{\beta}_{\mathbf{k}} - \hat{\boldsymbol{\beta}}_{\mathbf{k}})^{\top}\mathbf{I}(\boldsymbol{\beta}_{0,\mathbf{k}})(\boldsymbol{\beta}_{\mathbf{k}} - \hat{\boldsymbol{\beta}}_{\mathbf{k}})\right\}\right.\right. \\
&\times \left.\left. \frac{(\tau)^{\frac{r|k|}{2}}|I_{k0}|^{-\frac{r}{2}}}{(\Gamma(\frac{r}{2}))^{|k|}} \exp\left\{-(\tau)(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})^{-1}\right\} \prod_{i=1}^{|\mathbf{k}|} |\boldsymbol{\beta}_{\mathbf{k}i}|^{-(r+1)}\ \mathrm{d}\boldsymbol{\beta}_{\mathbf{k}}\right] \right. \\
&+ \left.\left[\int_{\boldsymbol{\beta}_{\mathbf{k}}:\boldsymbol{\beta}_{\mathbf{k}}\in\mathbf{I}^{c}} \exp\left\{-\frac{1-\varepsilon}{2}\,(\boldsymbol{\beta}_{\mathbf{k}} - \hat{\boldsymbol{\beta}}_{\mathbf{k}})^{\top}\mathbf{I}(\boldsymbol{\beta}_{0,\mathbf{k}})(\boldsymbol{\beta}_{\mathbf{k}} - \hat{\boldsymbol{\beta}}_{\mathbf{k}})\right\}\right.\right. \\
&\times \left.\left. \frac{(\tau)^{\frac{r|k|}{2}}|I_{k0}|^{-\frac{r}{2}}}{(\Gamma(\frac{r}{2}))^{|k|}} \exp\left\{-(\tau)(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})^{-1}\right\} \prod_{i=1}^{|\mathbf{k}|} |\boldsymbol{\beta}_{\mathbf{k}i}|^{-(r+1)}\ \mathrm{d}\boldsymbol{\beta}_{\mathbf{k}}\right]\right\} \\
&= \exp\left\{\ell(\hat{\boldsymbol{\beta}}_{\mathbf{k}})\right\} \\
&\times \left\{\left[\int_{\boldsymbol{\beta}_{\mathbf{k}}:\boldsymbol{\beta}_{\mathbf{k}}\in\mathbf{I}} \exp\left\{-\frac{1-\varepsilon}{2}\,(\boldsymbol{\beta}_{\mathbf{k}} - \hat{\boldsymbol{\beta}}_{\mathbf{k}})^{\top}\mathbf{I}(\boldsymbol{\beta}_{0,\mathbf{k}})(\boldsymbol{\beta}_{\mathbf{k}} - \hat{\boldsymbol{\beta}}_{\mathbf{k}})\right\}\right.\right. \\
&\times \left.\left. \frac{(\tau)^{\frac{r|k|}{2}}|I_{k0}|^{-\frac{r}{2}}}{(\Gamma(\frac{r}{2}))^{|k|}} \exp\left\{-(\tau)(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})^{-1}\right\} \prod_{i=1}^{|\mathbf{k}|} |\boldsymbol{\beta}_{\mathbf{k}i}|^{-(r+1)}\ \mathrm{d}\boldsymbol{\beta}_{\mathbf{k}}\right] \right. \\
&+ \left.\left[\exp\left\{-\frac{1-\varepsilon}{8}\,C_1^2\,\lambda\,|\mathbf{k}|_{\substack{\max \\ \mathbf{k}\,:\,|\mathbf{k}|\leq m_n}} \lambda_{\max}\left(\frac{X_{\mathbf{k}}^{\top}X_{\mathbf{k}}}{n}\right)\log p\right\}\right.\right. \\
&\times \left.\left. \int_{\boldsymbol{\beta}_{\mathbf{k}}:\boldsymbol{\beta}_{\mathbf{k}}\in\mathbf{I}^{c}} \frac{(\tau)^{\frac{r|k|}{2}}|I_{k0}|^{-\frac{r}{2}}}{(\Gamma(\frac{r}{2}))^{|k|}} \exp\left\{-(\tau)(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})^{-1}\right\} \prod_{i=1}^{|\mathbf{k}|} |\boldsymbol{\beta}_{\mathbf{k}i}|^{-(r+1)}\ \mathrm{d}\boldsymbol{\beta}_{\mathbf{k}}\right]\right\}
\end{aligned}
$$

In other hand, from Proof of Corollary 2 in the supplementary material of [11], for any m, $m > 4 + \frac{(r+1)}{2}$, from the inequality

$$
\begin{aligned}
\left(\frac{\boldsymbol{\beta}_{\mathbf{k}}^{2}}{\tau\ |\mathbf{I}_{k0}|^{-1}}\right)^{-\frac{(r+1)}{2}} \exp\left(-\frac{\tau}{(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})}\right) &= \left(\frac{(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})}{\tau}\right)^{-\frac{(r+1)}{2}} \frac{1}{\sum_{j=0}^{\infty}\frac{1}{j!}\left(\frac{\tau}{(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})}\right)^{j}} \\
&< r!\left(\frac{(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})}{\tau}\right)^{m-(\frac{r+1}{2})} \qquad m \in \mathrm{Z}^{+}
\end{aligned}
$$

Conversely, for some constant c and $\left|\boldsymbol{\beta}_{k_i}\right| > \varepsilon$ for any $\varepsilon > 0$,

$$
\left(\frac{\boldsymbol{\beta}_{\mathbf{k}_i}^{2}}{\tau\ |\mathbf{I}_{k0}|^{-1}}\right)^{-\frac{(r+1)}{2}} \exp\left(-\frac{\tau}{(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})}\right) > c\left(\frac{(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})}{\tau}\right)^{m-(\frac{r+1}{2})} \exp\left(-\frac{(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})}{2\tau}\right)
$$

Then we have,

$$
\begin{aligned}
m_{\mathbf{k}}(\mathbf{y}_n) < \exp\left\{\ell(\hat{\boldsymbol{\beta}}_{\mathbf{k}})\right\} & \\
\times \Bigg\{ \Bigg[\int_{\boldsymbol{\beta}_{\mathbf{k}}:\boldsymbol{\beta}_{\mathbf{k}}\in\mathbf{I}} & \exp\left\{-\frac{1-\varepsilon}{2}(\boldsymbol{\beta}_{\mathbf{k}}-\hat{\boldsymbol{\beta}}_{\mathbf{k}})^{\top}\mathbf{I}(\boldsymbol{\beta}_{0,\mathbf{k}})(\boldsymbol{\beta}_{\mathbf{k}}-\hat{\boldsymbol{\beta}}_{\mathbf{k}})\right\} \\
\times \prod_{i=1}^{|\mathbf{k}|} c\left(\frac{(\boldsymbol{\beta}_{\mathbf{k}_i}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}_i})}{\tau}\right)^{m-(\frac{r+1}{2})} & \exp\left(-\frac{(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})}{2\tau}\right)\mathrm{d}\boldsymbol{\beta}_{\mathbf{k}}\Bigg] \\
+ \Bigg[\exp\left\{-\frac{1-\varepsilon}{8}\,C_1^2\,\lambda\,|\mathbf{k}|\max_{\mathbf{k}\,:\,|\mathbf{k}|\le m_n}\lambda_{\max}\left(\frac{X_{\mathbf{k}}^{\top}X_{\mathbf{k}}}{n}\right)\log p\right\} & \\
\times \int_{\boldsymbol{\beta}_{\mathbf{k}}:\boldsymbol{\beta}_{\mathbf{k}}\in\mathbf{I}^c} \frac{(\tau)^{\frac{r|k|}{2}}|I_{k0}|^{-\frac{r}{2}}}{(\Gamma(\frac{r}{2}))^{|k|}}\,\exp\left\{-(\tau)(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})^{-1}\right\}\prod_{i=1}^{|\mathbf{k}|}|\boldsymbol{\beta}_{\mathbf{k}i}|^{-(r+1)}\,\mathrm{d}\boldsymbol{\beta}_{\mathbf{k}}\Bigg]\Bigg\} &
\end{aligned}
$$

$$
\begin{aligned}
m_{\mathbf{k}}(\mathbf{y}_n) < \exp\left\{\ell(\hat{\boldsymbol{\beta}}_{\mathbf{k}})\right\} & \\
\times \Bigg\{ \Bigg[\int_{\boldsymbol{\beta}_{\mathbf{k}}:\boldsymbol{\beta}_{\mathbf{k}}\in\mathbf{I}} & \exp\left\{-\frac{1-\varepsilon}{2}(\boldsymbol{\beta}_{\mathbf{k}}-\hat{\boldsymbol{\beta}}_{\mathbf{k}})^{\top}\mathbf{I}(\boldsymbol{\beta}_{0,\mathbf{k}})(\boldsymbol{\beta}_{\mathbf{k}}-\hat{\boldsymbol{\beta}}_{\mathbf{k}}) - \frac{(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})}{2\tau}\right\} \\
\times \prod_{i=1}^{|\mathbf{k}|} c\left(\frac{(\boldsymbol{\beta}_{\mathbf{k}_i}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}_i})}{\tau}\right)^{m-(\frac{r+1}{2})} & \mathrm{d}\boldsymbol{\beta}_{\mathbf{k}}\Bigg] \qquad \text{(A2)} \\
+ \Bigg[\exp\left\{-\frac{1-\varepsilon}{8}\,C_1^2\,\lambda\,|\mathbf{k}|\max_{\mathbf{k}\,:\,|\mathbf{k}|\le m_n}\lambda_{\max}\left(\frac{X_{\mathbf{k}}^{\top}X_{\mathbf{k}}}{n}\right)\log p\right\} & \\
\times \int_{\boldsymbol{\beta}_{\mathbf{k}}:\boldsymbol{\beta}_{\mathbf{k}}\in\mathbf{I}^c} \frac{(\tau)^{\frac{r|k|}{2}}|I_{k0}|^{-\frac{r}{2}}}{(\Gamma(\frac{r}{2}))^{|k|}}\,\exp\left\{-(\tau)(\boldsymbol{\beta}_{\mathbf{k}}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{\mathbf{k}})^{-1}\right\}\prod_{i=1}^{|\mathbf{k}|}|\boldsymbol{\beta}_{\mathbf{k}i}|^{-(r+1)}\,\mathrm{d}\boldsymbol{\beta}_{\mathbf{k}}\Bigg]\Bigg\} & \qquad \text{(A3)}
\end{aligned}
$$

Now, we separately calculate equations (A2) and (A3). Let $\mathbf{H}_k = (1-\varepsilon)\mathbf{I}(\boldsymbol{\beta}_{0,k})$ and $\boldsymbol{\beta}_k^* = (\mathbf{H}_k + \frac{\mathbf{I}_{k0}}{\tau})^{-1}\mathbf{H}_k\hat{\boldsymbol{\beta}}_k$, then

$$
\begin{aligned}
\text{(A2)} \le \int_{\boldsymbol{\beta}_k:\boldsymbol{\beta}_k\in\mathbf{I}} & \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_k-\boldsymbol{\beta}_k^*)^{\top}(\mathbf{H}_k+\frac{\mathbf{I}_{k0}}{\tau})(\boldsymbol{\beta}_k-\boldsymbol{\beta}_k^*)\right\}\prod_{i=1}^{|k|} c\left(\frac{(\boldsymbol{\beta}_{k_i}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{k_i})}{\tau}\right)^{(m-\frac{(r+1)}{2})}\mathrm{d}\boldsymbol{\beta} \\
\times \exp & \left\{-\frac{1}{2}\hat{\boldsymbol{\beta}}_k^{\top}\left(\mathbf{H}_k-\mathbf{H}_k(\mathbf{H}_k+\frac{\mathbf{I}_{k0}}{\tau})^{-1}\mathbf{H}_k\right)\hat{\boldsymbol{\beta}}_k\right\} \\
= (2\pi)^{\frac{|k|}{2}} & \det(\mathbf{H}_k+\frac{\mathbf{I}_{k0}}{\tau})^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\hat{\boldsymbol{\beta}}_k^{\top}\left(\mathbf{H}_k-\mathbf{H}_k(\mathbf{H}_k+\frac{\mathbf{I}_{k0}}{\tau})^{-1}\mathbf{H}_k\right)\hat{\boldsymbol{\beta}}_k\right\} \\
\times E_k & \left(\prod_{i=1}^{|k|} c\left(\frac{(\boldsymbol{\beta}_{k_i}^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_{k_i})}{\tau}\right)^{(m-\frac{(r+1)}{2})}\right)
\end{aligned}
$$

, where E_k denotes the expectation with respect to a multivariate normal distribution with mean $\boldsymbol{\beta}_k^*$ and covariance matrix $V_k = (\mathbf{H}_k + \frac{\mathbf{I}_{k0}}{\tau})^{-1}$. It follows from Lemma 6 in the supplementary material for [11] that, if $t \not\subset k$ and conditions (A1)-(A4) apply, then

$$
P\left[E\left(\prod_{i=1}^{k}\boldsymbol{\beta}_{k_i}^{2r}\right) > q\right] < P\left[T > q^*\right]
$$

, then T follows a chi-squared distribution with non-centrality parameter $\lambda = (\boldsymbol{\beta}_k^{\top}\mathbf{I}_{k0}\boldsymbol{\beta}_k)$ and $t \cup k$ degrees of freedom. $T \sim \chi^2_{(t\cup k,\lambda)}$,

For any m, $m > 4 + \frac{(r+1)}{2}$,

$$E_k\left(\prod_{i=1}^{|k|} c\left(\frac{(\boldsymbol{\beta}_{k_i}^\top \mathbf{I}_{k0}\boldsymbol{\beta}_{k_i})}{\tau}\right)^{(m-\frac{(r+1)}{2})}\right) = c^{|k|}\tau^{-(m-\frac{(r+1)}{2})} E_k\left(\prod_{i=1}^{|k|}\left((\boldsymbol{\beta}_{k_i}^\top \mathbf{I}_{k0}\boldsymbol{\beta}_{k_i})\right)^{(m-\frac{(r+1)}{2})}\right)$$

$$\leq c^{|k|}\tau^{-(m-\frac{(r+1)}{2})}\left(\frac{n \max\limits_{k\,:\,|k|\leq m_n}\lambda_{\max}\left(\frac{X_k^\top X_k}{n}\right)+\tau^{-1}}{n\lambda+\tau^{-1}}\right)^{\frac{|k|}{2}}\left\{\frac{4\left\|\boldsymbol{\beta}_k^*\right\|_2^2}{|k|}+\frac{4\left[(2r-1)!!\right]^{\frac{1}{r}}}{n(\lambda+\tau^{-2})}\right\}^{r|k|}$$

$$\leq c^{|k|}\tau^{-(m-\frac{(r+1)}{2})}\left(\frac{n \max\limits_{k\,:\,|k|\leq m_n}\lambda_{\max}\left(\frac{X_k^\top X_k}{n}\right)+\tau^{-1}}{n\lambda+\tau^{-1}}\right)^{\frac{|k|}{2}} 2^{r|k|-1}\left\{\left(\frac{4\left\|\boldsymbol{\beta}_k^*\right\|_2^2}{|k|}\right)^{r|k|}+\left(\frac{4\left[(2r-1)!!\right]^{\frac{1}{r}}}{n(\lambda+\tau^{-2})}\right)^{r|k|}\right\}$$

, and for some constant $C > 0$

$$\int_{\boldsymbol{\beta}_k:\boldsymbol{\beta}_k\in\mathbf{I}^c}\exp\left\{-(\tau)(\boldsymbol{\beta}_k^\top \mathbf{I}_{k0}\boldsymbol{\beta}_k)^{-1}\right\}\prod_{i=1}^{|k|}|\boldsymbol{\beta}_{ki}|^{-(r+1)}[d\boldsymbol{\beta}_k \leq \int_{\boldsymbol{\beta}_k:\boldsymbol{\beta}_k\in\mathbf{I}^c} r!\left(\frac{(\boldsymbol{\beta}_k^\top \mathbf{I}_{k0}\boldsymbol{\beta}_k)}{\tau}\right)^{m-(\frac{r+1}{2})}\mathrm{d}\boldsymbol{\beta}_k \leq (C\tau^{-1})^{\frac{|k|}{2}}$$

In other hand, by condition (A1), (A3) and (A4)

$$\begin{aligned}\det(\mathbf{H}_k+\frac{\mathbf{I}_{k0}}{\tau})^{\frac{1}{2}} &= \det\left\{(1-\varepsilon)\mathbf{I}(\boldsymbol{\beta}_{0,k})+\frac{\mathbf{I}_{k0}}{\tau}\right\}^{\frac{1}{2}}\\ &\leq (n\times\max_{k\,:\,|k|\leq m_n}\lambda_{m_n}\left(\frac{X_k^\top X_k}{n}\right)+\frac{1}{\tau})^{\frac{|k|}{2}}\\ &\leq \exp\left\{C\,|k|\log n\right\}\\ &\ll \exp\left\{\frac{1-\varepsilon}{8}c^2\lambda\,|k|\max_{k\,:\,|k|\leq m_n}\lambda_{m_n}\left(\frac{X_k^\top X_k}{n}\right)\log p\right\}\end{aligned}$$

Therefore, for some constant $c > 0$

$$\begin{aligned}m_k(\mathbf{y}_n) \leq \exp\left\{\ell(\hat{\boldsymbol{\beta}}_k)\right\}\frac{(\tau)^{\frac{r|k|}{2}}}{\Gamma(\frac{r}{2})^{|k|}}|\mathbf{I}_{k0}|^{-\frac{r}{2}}\times(2\pi)^{\frac{|k|}{2}}\det\left\{(1-\varepsilon)\mathbf{I}(\boldsymbol{\beta}_{0,k})+\frac{\mathbf{I}_{k0}}{\tau}\right\}^{-\frac{1}{2}}\\ \times E_k\left(c\left(\frac{(\boldsymbol{\beta}_k^\top \mathbf{I}_{k0}\boldsymbol{\beta}_k)}{\tau}\right)^{(m-\frac{(r+1)}{2})}\right)\end{aligned} \tag{A4}$$

Next, note that it follows from Lemma A.3 in the supplementary material of [15], that

$$\begin{aligned}V &= \left\|\boldsymbol{\beta}_k^*\right\|_2^2\\ &\leq \left\|\hat{\boldsymbol{\beta}}_k\right\|_2^2\\ &\leq \left(\left\|\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_{0,k}\right\|_2+\left\|\boldsymbol{\beta}_{0,k}\right\|_2\right)^2\\ &\leq \left(\sqrt{\frac{|k|\max\limits_{k\,:\,|k|\leq m_n}\lambda_{\max}\left(\frac{X_k^\top X_k}{n}\right)\log p}{n}}+\sqrt{\log p}\right)^2\\ &\leq 2\left(\frac{|k|\max\limits_{k\,:\,|k|\leq m_n}\lambda_{\max}\left(\frac{X_k^\top X_k}{n}\right)\log p}{n}+\log p\right).\end{aligned}$$

Suppose conditions (A1) - (A4) hold, for any $\mathbf{k} \in \mathbf{K}_2 = \{\mathbf{k} : |\mathbf{k}| \leq \mathbf{m}_n, \mathbf{t} \neq \mathbf{k}\}$, and small $c > 0$ then,

$$\sum_{\mathbf{k}:\, \mathbf{k}\in\mathbf{K}_2} \frac{\pi(\mathbf{k})\, m_{\mathbf{k}}(\mathbf{y}_n)}{\pi(\mathbf{t})\, m_{\mathbf{t}}(\mathbf{y}_n)} \xrightarrow{P} 0 \quad as \quad n \to \infty.$$

Proof. Let $\mathbf{k}^* = \mathbf{k} \cup \mathbf{t}$, then $\mathbf{k}^* \in \mathbf{K}_1$, and suppose $\boldsymbol{\beta}_{\mathbf{k}^*}$ is the $|\mathbf{k}^*| - \dim$ensional vector of $\boldsymbol{\beta}_{\mathbf{k}}$ for $\mathbf{k}$ and zero for $\mathbf{t}$. Then by Taylor's expansion and Lemmas A.1 and A.3 in [15], for any $\boldsymbol{\beta}_{\mathbf{k}^*}$, such that for $c > 0$

$$\left\|\boldsymbol{\beta}_{k^*} - \boldsymbol{\beta}_{0,k^*}\right\|_2 \leq c\sqrt{\frac{|k^*| \max\limits_{k^*:|k^*|\leq m_n} \lambda_{\max}\left(\frac{X_{k^*}^\top X_{k^*}}{n}\right) \log p}{n}} = cw$$

, then with probability tending to 1,

$$\begin{aligned}
\ell(\boldsymbol{\beta}_{k^*}) &= \ell(\hat{\boldsymbol{\beta}}_{k^*}) - \frac{1}{2}(\boldsymbol{\beta}_{k^*} - \hat{\boldsymbol{\beta}}_{k^*})^\top \mathbf{I}(\tilde{\boldsymbol{\beta}}_{k^*})(\boldsymbol{\beta}_{k^*} - \hat{\boldsymbol{\beta}}_{k^*}) + \mathrm{O}(n^{-1}) \\
&\leq \ell(\hat{\boldsymbol{\beta}}_{k^*}) - \frac{1-\varepsilon}{2}(\boldsymbol{\beta}_{k^*} - \hat{\boldsymbol{\beta}}_{k^*})^\top \mathbf{I}(\boldsymbol{\beta}_{0,k^*})(\boldsymbol{\beta}_{k^*} - \hat{\boldsymbol{\beta}}_{k^*}) \\
&\leq \ell(\hat{\boldsymbol{\beta}}_{k^*}) - \frac{n(1-\varepsilon)\lambda}{2}\left\|\boldsymbol{\beta}_{k^*} - \hat{\boldsymbol{\beta}}_{k^*}\right\|_2^2
\end{aligned}$$

Let $\mathbf{H}_k = n(1-\varepsilon)\lambda \mathbf{I}(\boldsymbol{\beta}_{0,k})$ and $\boldsymbol{\beta}_k^* = (\mathbf{H}_k + \frac{\mathbf{I}_{k0}}{\tau})^{-1}\mathbf{H}_k\hat{\boldsymbol{\beta}}_k$, then

$$\begin{aligned}
&\int \exp\{-\frac{n(1-\varepsilon)\lambda}{2}\left\|\boldsymbol{\beta}_{k^*} - \hat{\boldsymbol{\beta}}_{k^*}\right\|_2^2\}\frac{(\tau)^{\frac{r|k|}{2}}}{\Gamma(\frac{r}{2})^{|k|}}\,|\mathbf{I}_{k0}|^{-\frac{r}{2}}\prod_{i=1}^{|k|}|\boldsymbol{\beta}_{ki}\,|^{-(r+1)} \times \exp\left\{-(\tau)(\boldsymbol{\beta}_k^\top \mathbf{I}_{k0}\boldsymbol{\beta}_k)^{-1}\right\}\mathrm{d}\boldsymbol{\beta}_k \\
&= \int \exp\{-\frac{n(1-\varepsilon)\lambda}{2}\left\|\boldsymbol{\beta}_k - \hat{\boldsymbol{\beta}}_k\right\|_2^2 - (\tau)(\boldsymbol{\beta}_k^\top \mathbf{I}_{k0}\,\boldsymbol{\beta}_k)^{-1}\}\prod_{i=1}^{|k|}|\boldsymbol{\beta}_{ki}|^{-(r+1)}\mathrm{d}\boldsymbol{\beta}_k \times \exp\{-\frac{n(1-\varepsilon)\lambda}{2}\left\|\hat{\boldsymbol{\beta}}_t\right\|_2^2\} \\
&= (2\pi)^{\frac{|k|}{2}}\ \det(\mathbf{H}_k + \frac{\mathbf{I}_{k0}}{\tau})^{-\frac{1}{2}}\ \exp\left\{-\frac{1}{2}\hat{\boldsymbol{\beta}}_k^\top\left(\mathbf{H}_k - \mathbf{H}_k(\mathbf{H}_k + \frac{\mathbf{I}_{k0}}{\tau})^{-1}\mathbf{H}_k\right)\hat{\boldsymbol{\beta}}_k\right\} \times |\mathbf{I}_{k0}|^{-\frac{r}{2}} \\
&\times E_k\left(c\left(\frac{(\boldsymbol{\beta}_k^\top \mathbf{I}_{k0}\boldsymbol{\beta}_k)}{\tau}\right)^{(m-\frac{(r+1)}{2})}\right) \times \exp\{-\frac{n(1-\varepsilon)\lambda}{2}\left\|\hat{\boldsymbol{\beta}}_t\right\|_2^2\}
\end{aligned}$$

, where E_k denotes the expectation with respect to a multivariate normal distribution with mean $\boldsymbol{\beta}_k^*$ and covariance matrix $V_k = (\mathbf{H}_k + \frac{\mathbf{I}_{k0}}{\tau})^{-1}$. It follows from Lemma 6 in the supplementary material of [11], that if $t \not\subset k$ and conditions (A1) - (A4) apply, then

$$P\left[E\left((\boldsymbol{\beta}_k^\top \mathbf{I}_{k0}\boldsymbol{\beta}_k)\right) > q\right] < P\left[T > q^*\right]$$

,then T follows a chi-squared distribution with non-centrality parameter $\lambda = (\boldsymbol{\beta}_k^\top \mathbf{I}_{k0}\boldsymbol{\beta}_k)$ and $t \cup k$ degrees of freedom. $T \sim \chi^2_{(t\cup k,\lambda)}$,

$$\begin{aligned}
&E_k\left(\prod_{i=1}^{|k|} c\left(\frac{(\boldsymbol{\beta}_{k_i}^\top \mathbf{I}_{k0}\boldsymbol{\beta}_{k_i})}{\tau}\right)^{(m-\frac{(r+1)}{2})}\right) = c^{|k|}\tau^{-(m-\frac{(r+1)}{2})}E_k\left(\prod_{i=1}^{|k|}\left((\boldsymbol{\beta}_{k_i}^\top \mathbf{I}_{k0}\boldsymbol{\beta}_{k_i})\right)^{(m-\frac{(r+1)}{2})}\right) \\
&\leq c^{|k|}\tau^{-(m-\frac{(r+1)}{2})}\left(\frac{n\max\limits_{k:|k|\leq m_n}\lambda_{\max}\left(\frac{X_k^\top X_k}{n}\right) + \tau^{-1}}{n\lambda + \tau^{-1}}\right)^{\frac{|k|}{2}}\left\{\frac{4\left\|\beta_k^*\right\|_2^2}{|k|} + \frac{4\left[(2r-1)!!\right]^{\frac{1}{r}}}{n(\lambda + \tau^{-2})}\right\}^{r|k|} \\
&\leq c^{|k|}\tau^{-(m-\frac{(r+1)}{2})}\left(\frac{n\max\limits_{k:|k|\leq m_n}\lambda_{\max}\left(\frac{X_k^\top X_k}{n}\right) + \tau^{-1}}{n\lambda + \tau^{-1}}\right)^{\frac{|k|}{2}} 2^{r|k|-1}\left\{\left(\frac{4\left\|\boldsymbol{\beta}k^*\right\|_2^2}{|k|}\right)^{r|k|} + \left(\frac{4\left[(2r-1)!!\right]^{\frac{1}{r}}}{n(\lambda + \tau^{-2})}\right)^{r|k|}\right\}
\end{aligned}$$

Define the set $H_* = \{\boldsymbol{\beta}_k : \left\|\boldsymbol{\beta}_{k^*} - \hat{\boldsymbol{\beta}}_{k^*}\right\|_2 \leq \frac{cw}{2}\}$, then with probability tending to 1, for any $\mathbf{k} \in \mathbf{K}_2 = \{\mathbf{k} : |\mathbf{k}| \leq \mathbf{m}_n \ , \ \mathbf{t} \neq \mathbf{k}\}$

$$
\begin{aligned}
\pi(\mathbf{k} \mid y) &= \int_{H_* \cup H_*^C} \exp\{\ell(\boldsymbol{\beta}_k)\}\, \pi(\boldsymbol{\beta}_k \mid \mathbf{k}) \mathrm{d}\boldsymbol{\beta}_k \\
&= \int_{H_* \cup H_*^C} \exp\{\ell(\boldsymbol{\beta}_{k^*})\}\ \frac{(\tau)^{\frac{r|k|}{2}}}{\Gamma(\frac{r}{2})^{|k|}} |\mathbf{I}_{k0}|^{-\frac{r}{2}} \prod_{i=1}^{|k|} |\boldsymbol{\beta}_{ki}|^{-(r+1)} \times \exp\left\{-(\tau)(\boldsymbol{\beta}_k^\top \mathbf{I}_{k0} \boldsymbol{\beta}_k)^{-1}\right\} \mathrm{d}\boldsymbol{\beta}_k \\
&\leq (C\tau)^{\frac{r|k|}{2}} \exp\left\{\ell(\hat{\boldsymbol{\beta}}_{k^*})\right\} \det(\mathbf{H}_k + \frac{\mathbf{I}_{k0}}{\tau})^{-\frac{1}{2}} \\
&\times \left[\exp\{-\frac{n(1-\varepsilon)\lambda}{2} \left\|\hat{\boldsymbol{\beta}}_t\right\|_2^2\} (\frac{\log p}{|k|})^{r|k|} + \exp\{-c\, C \left|k^*\right| \max_{k^* : |k^*| \leq m_n} \lambda_{\max}\left(\frac{X_{k^*}^\top X_{k^*}}{n}\right) \log p\}\right]
\end{aligned}
$$

There for,

$$
\begin{aligned}
\frac{m_k(\mathbf{y}_n)}{m_t(\mathbf{y}_n)} \leq\ & \left(C\tau^{\frac{r}{2}}\right)^{|k|-|t|} \frac{\det\left\{(1+\varepsilon)n^{-1}\mathbf{I}(\boldsymbol{\beta}_{0,t}) + (n^{-1}\tau^{-1})\mathbf{I}_{k0}\right\}^{\frac{1}{2}}}{\det\left\{(1-\varepsilon)\lambda\mathbf{I}(\boldsymbol{\beta}_{0,k}) + (n^{-1}\tau^{-1})\mathbf{I}_{k0}\right\}^{\frac{1}{2}}} \\
&\times \exp\left\{\ell(\hat{\boldsymbol{\beta}}_{k^*}) - \ell(\hat{\boldsymbol{\beta}}_t)\right\} \exp\left\{-\frac{n(1-\varepsilon)\lambda}{2} \left\|\hat{\boldsymbol{\beta}}_{\frac{t}{k}}\right\|_2^2\right\} (\frac{\log p}{|k|})^{r|k|} (\log p)^{r|t|} \qquad \text{(A7)} \\
&+ \left(C\tau^{\frac{r}{2}}\right)^{|k|-|t|} \det\left\{(1+\varepsilon)n^{-1}\mathbf{I}(\boldsymbol{\beta}_{0,t}) + (n^{-1}\tau^{-1})\mathbf{I}_{k0}\right\}^{\frac{1}{2}} \\
&\times \exp\left\{\ell(\hat{\boldsymbol{\beta}}_{k^*}) - \ell(\hat{\boldsymbol{\beta}}_t)\right\} \exp\{-c\, C \left|k^*\right| \max_{k^* : |k^*| \leq m_n} \lambda_{\max}\left(\frac{X_{k^*}^\top X_{k^*}}{n}\right) \log p\} (\log p)^{r|t|} \qquad \text{(A8)}
\end{aligned}
$$

We separately calculate (A7) and (A8), first in (A7)

$$
\begin{aligned}
&\frac{\det\left\{(1+\varepsilon)n^{-1}\mathbf{I}(\boldsymbol{\beta}_{0,t}) + (n^{-1}\tau^{-1})\mathbf{I}_{k0}\right\}^{\frac{1}{2}}}{\det\left\{(1-\varepsilon)\lambda\mathbf{I}(\boldsymbol{\beta}_{0,k}) + (n^{-1}\tau^{-1})\mathbf{I}_{k0}\right\}^{\frac{1}{2}}} \leq \frac{\left\{(1+\varepsilon)\ |t| \max_{t\,:\,|t| \leq m_n} \lambda_{\max}\left(\frac{X_t^\top X_t}{n}\right) \log p + (n^{-1}\tau^{-1})\right\}^{\frac{|t|}{2}}}{\{(1-\varepsilon)\lambda + (n^{-1}\tau^{-1})\}^{\frac{|k|}{2}}} \\
&= \left\{\frac{(1+\varepsilon)\ |t| \max_{t\,:\,|t| \leq m_n} \lambda_{\max}\left(\frac{X_t^\top X_t}{n}\right) \log p + (n^{-1}\tau^{-1})}{(1-\varepsilon)\lambda + (n^{-1}\tau^{-1})}\right\}^{\frac{|t|}{2}} \left\{\frac{1}{(1-\varepsilon)\lambda + (n^{-1}\tau^{-1})}\right\}^{\frac{(|k|-|t|)}{2}} \\
&\leq \exp\left\{C\,|t| \log\left(|t| \max_{t\,:\,|t| \leq m_n} \lambda_{\max}\left(\frac{X_t^\top X_t}{n}\right) \log p\right)\right\} \left\{\frac{1}{(1-\varepsilon)\lambda + (n^{-1}\tau^{-1})}\right\}^{\frac{(|k|-|t|)}{2}}
\end{aligned}
$$

Also, for $\gamma_* = (1+\delta)(1+2w)\log p$, we have

$$
\ell(\hat{\boldsymbol{\beta}}_{k^*}) - \ell(\hat{\boldsymbol{\beta}}_t) \ \leq\ \gamma_*(\left|k^*\right| - |t|)\log p = \gamma_*\left(\left|\frac{t}{k}\right|\right)\log p + \gamma_*(|k| - |t|)\log p
$$

Then, with Condition (A1), in part of (A7),

$$
\begin{aligned}
&\left(C\tau^{\frac{r}{2}}\right)^{|k|-|t|} \frac{\det\left\{(1+\varepsilon)n^{-1}\mathbf{I}(\boldsymbol{\beta}_{0,t}) + (n^{-1}\tau^{-1})\mathbf{I}_{k0}\right\}^{\frac{1}{2}}}{\det\left\{(1-\varepsilon)\lambda\mathbf{I}(\boldsymbol{\beta}_{0,k}) + (n^{-1}\tau^{-1})\mathbf{I}_{k0}\right\}^{\frac{1}{2}}} (\frac{\log p}{|k|})^{r|k|} (\log p)^{r|t|} \\
&= \left(C\tau^{\frac{r}{2}}\right)^{|k|-|t|} \left\{\frac{1}{(1-\varepsilon)\lambda + (n^{-1}\tau^{-1})}\right\}^{\frac{(|k|-|t|)}{2}} p^{\gamma_*(|k|-|t|)} \times \exp\left\{r\,|k|\ \log\left(\frac{\log p}{|k|}\right) + r\,|t|\log(\log p)\right\} \\
&\leq \left(C^{(|k|-|t|)}(\log p)^{\frac{r}{2}(|k|-|t|)}\right) p^{\gamma_*(|k|-|t|)} = o(1).
\end{aligned}
$$

In other hand, with Condition (A4), in other part of (A7)

$$\begin{aligned}
\exp\left\{-\frac{n(1-\varepsilon)\lambda}{2}\left\|\hat{\boldsymbol{\beta}}_{\frac{t}{k}}\right\|_2^2\right\} &\leq \exp\left\{-\frac{n(1-\varepsilon)\lambda}{2}\left\{\left\|\hat{\boldsymbol{\beta}}_{\frac{t}{k}}+\boldsymbol{\beta}_{0,\frac{t}{k}}-\boldsymbol{\beta}_{0,\frac{t}{k}}\right\|_2^2\right\}\right\} \\
&\leq \exp\left\{-\frac{n(1-\varepsilon)\lambda}{2}\left\{\left\|\boldsymbol{\beta}_{0,\frac{t}{k}}\right\|_2^2-\left\|\hat{\boldsymbol{\beta}}_{\frac{t}{k}}-\boldsymbol{\beta}_{0,\frac{t}{k}}\right\|_2^2\right\}\right\} \\
&\leq \exp\left\{-\frac{(1-\varepsilon)\lambda}{2}\left\{c_0\left|\frac{t}{k}\right|^2\min_{j\in t}\boldsymbol{\beta}_{0,j}^2-c_1\left|\frac{t}{k}\right|\max_{t\,:\,|t|\leq m_n}\lambda_{\max}\left(\frac{X_t^\top X_t}{n}\right)\log p\right\}\right\} \\
&\leq \exp\left\{-\frac{(1-\varepsilon)\lambda}{2}(c_0-c_1)\left|\frac{t}{k}\right|^2|t|\max_{t\,:\,|t|\leq m_n}\lambda_{\max}\left(\frac{X_t^\top X_t}{n}\right)\log p\right\}
\end{aligned}$$

Then (A7) is bounded by,

$$\begin{aligned}
(A9) &= \left(C\tau^{\frac{r}{2}}\right)^{|k|-|t|}\frac{\det\left\{(1+\varepsilon)n^{-1}\mathbf{I}(\boldsymbol{\beta}_{0,t})+(n^{-1}\tau^{-1})\mathbf{I}_{k0}\right\}^{\frac{1}{2}}}{\det\left\{(1-\varepsilon)\lambda\mathbf{I}(\boldsymbol{\beta}_{0,k})+(n^{-1}\tau^{-1})\mathbf{I}_{k0}\right\}^{\frac{1}{2}}} \\
&\times\exp\left\{\ell(\hat{\boldsymbol{\beta}}_{k*})-\ell(\hat{\boldsymbol{\beta}}_t)\right\}\exp\left\{-\frac{n(1-\varepsilon)\lambda}{2}\left\|\hat{\boldsymbol{\beta}}_{\frac{t}{k}}\right\|_2^2\right\}\left(\frac{\log p}{|k|}\right)^{r|k|}(\log p)^{r|t|} \\
&\leq \mathrm{o}(1)\times\exp\left\{\gamma_*\left(\left|\frac{t}{k}\right|\right)\log p+\gamma_*(|k|-|t|)\log p\right\} \\
&\times\exp\left\{-\frac{(1-\varepsilon)\lambda}{2}(c_0-c_1)\left|\frac{t}{k}\right|^2|t|\max_{t\,:\,|t|\leq m_n}\lambda_{\max}\left(\frac{X_t^\top X_t}{n}\right)\log p\right\} \\
&\times\left(\frac{\log p}{|k|}\right)^{r|k|}(\log p)^{r|t|} \\
&\leq \exp\left\{-\left(\frac{(1-\varepsilon)\lambda}{2}(c_0-c_1)-\mathrm{o}(1)-\gamma_*\right)\left|\frac{t}{k}\right|^2|t|\max_{t\,:\,|t|\leq m_n}\lambda_{\max}\left(\frac{X_t^\top X_t}{n}\right)\log p\right\} \\
&\leq \exp\left\{-\left(\frac{(1-\varepsilon)\lambda}{2}(c_0-c_1)-\mathrm{o}(1)-\gamma_*\right)|t|\max_{t\,:\,|t|\leq m_n}\lambda_{\max}\left(\frac{X_t^\top X_t}{n}\right)\log p\right\}
\end{aligned}$$

It is easy to see that the maximum (A8) is also of order o(1) with probability tending to 1 by the similar arguments. Since we have (A2) in the proof of No super set theorem, it completes the proof. □

Proof of "**Strong selection consistency Theorem**". By putting two summations of "**No super set theorem**"and "**Posterior ratio consistency theorem**"together in

$$\pi(\mathbf{t}\mid\mathbf{y}_n)=\frac{\pi(\mathbf{t})\,m_{\mathbf{t}}(\mathbf{y}_n)}{\sum_{\mathbf{k}}\pi(\mathbf{k})\,m_{\mathbf{k}}(\mathbf{y}_n)}=\left[1+\sum_{\mathbf{k}:\mathbf{k}\in\mathbf{K}_1}\frac{\pi(\mathbf{k})\,m_{\mathbf{k}}(\mathbf{y}_n)}{\pi(\mathbf{t})\,m_{\mathbf{t}}(\mathbf{y}_n)}+\sum_{\mathbf{k}:\mathbf{k}\in\mathbf{K}_2}\frac{\pi(\mathbf{k})\,m_{\mathbf{k}}(\mathbf{y}_n)}{\pi(\mathbf{t})\,m_{\mathbf{t}}(\mathbf{y}_n)}\right]^{-1}$$

, and the use of Slutsky's theorem leads to

$$\tilde{\pi}(\mathbf{t}\mid\mathbf{y}_n)\quad\xrightarrow{P}\quad 1\ ,\ \ as\ \ n\to\infty$$

The claimed result follows because the Laplace method is valid with error rate $\mathrm{O}(n^{-1})$.

Bibliography

[1] Xuan Cao, Kshitij Khare, and Malay Ghosh. High-dimensional posterior consistency for hierarchical non-local priors in regression. *Bayesian Analysis*, 15(1):241–262, 2020.

[2] Peter K Dunn, Gordon K Smyth, et al. *Generalized linear models with examples in R*. Springer, 2018.

[3] Jianqing Fan, Yang Feng, Diego Franco Saldana, Richard Samworth, Yichao Wu, and Maintainer Diego Franco Saldana. Package sis. *CRAN, https://cran. r-project. org/web/packages/SIS/index. html*, 2015.

[4] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[5] Jianqing Fan and Rui Song. Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 2010.

[6] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

[7] Peter D Hoff. *A first course in Bayesian statistical methods*, volume 580. Springer, 2009.

[8] Rafael A Irizarry, Benjamin M Bolstad, Francois Collin, Leslie M Cope, Bridget Hobbs, and Terence P Speed. Summaries of affymetrix genechip probe level data. *Nucleic acids research*, 31(4):e15–e15, 2003.

[9] Wenxin Jiang. On the consistency of bayesian variable selection for high dimensional binary regression and classification. *Neural computation*, 18(11):2762–2776, 2006.

[10] Valen E Johnson and David Rossell. On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170, 2010.

[11] Valen E Johnson and David Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660, 2012.

[12] Robert E Kass and Larry Wasserman. A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association*, 90(431):928–934, 1995.

[13] Faming Liang, Qifan Song, and Kai Yu. Bayesian subset modeling for high-dimensional generalized linear models. *Journal of the American Statistical Association*, 108(502):589–606, 2013.

[14] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

[15] Naveen N Narisetty, Juan Shen, and Xuming He. Skinny gibbs: A consistent and scalable gibbs sampler for model selection. *Journal of the American Statistical Association*, 2018.

[16] Amir Nikooienejad, Wenyi Wang, and Valen E Johnson. Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors. *Bioinformatics*, 32(9):1338–1345, 2016.

[17] Amir Nikooienejad, Wenyi Wang, and Valen E Johnson. Bayesian variable selection for survival data using inverse moment priors. *The annals of applied statistics*, 14(2):809, 2020.

[18] Stephen W Raudenbush, Meng-Li Yang, and Matheos Yosef. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of computational and Graphical Statistics*, 9(1):141–157, 2000.

[19] Christian P Robert et al. *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer, 2007.

[20] David Rossell and Donatello Telesca. Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association*, 112(517):254–265, 2017.

[21] Guiling Shi, Chae Young Lim, and Tapabrata Maiti. Bayesian model selection for generalized linear models using non-local priors. *Computational Statistics & Data Analysis*, 133:285–296, 2019.

[22] Minsuk Shin, Anirban Bhattacharya, and Valen E Johnson. Scalable bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statistica Sinica*, 28(2):1053, 2018.

[23] Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86, 1986.

[24] Ho-Hsiang Wu, Marco AR Ferreira, Mohamed Elkhouly, and Tieming Ji. Hyper nonlocal priors for variable selection in generalized linear models. *Sankhya A*, 82(1):147–185, 2020.

How to Cite: Robabeh Hosseinpour Samim Mamaghani[1], Farzad Eskandari[2], *Bayesian Inference Using Hyper Product Inverse Moment Prior in the Ultrahigh-Dimensional Generalized Linear Models*, Journal of Mathematics and Modeling in Finance (JMMF), Vol. 2, No. 2, Pages:63–89, (2022).