Bayesian Analysis (2024) **19**, Number 1, pp. 99–122

Bayesian Inference on Hierarchical Nonlocal Priors in Generalized Linear Models*

Xuan Cao† and Kyoungjae Lee‡,§

Abstract. Variable selection methods with nonlocal priors have been widely studied in linear regression models, and their theoretical and empirical performances have been reported. However, the crucial model selection properties for hierarchical nonlocal priors in high-dimensional generalized linear regression have rarely been investigated. In this paper, we consider a hierarchical nonlocal prior for high-dimensional logistic regression models and investigate theoretical properties of the posterior distribution. Specifically, a product moment (pMOM) nonlocal prior is imposed over the regression coefficients with an Inverse-Gamma prior on the tuning parameter. Under standard regularity assumptions, we establish strong model selection consistency in a high-dimensional setting, where the number of covariates is allowed to increase at a sub-exponential rate with the sample size. We implement the Laplace approximation for computing the posterior probabilities, and a modified shotgun stochastic search procedure is suggested for efficiently exploring the model space. We demonstrate the validity of the proposed method through simulation studies and an RNA-sequencing dataset for stratifying disease risk.

Keywords: high-dimensional, nonlocal prior, strong selection consistency.

MSC2020 subject classifications: Primary 62F15, 62J12; secondary 62F12.

1 Introduction

The advance in modern technology has led to an increased ability to collect and store data on a large scale. This brings opportunities and, at the same time, tremendous challenges in analyzing data with a large number of covariates per observation, the so-called high-dimensional problem. In high-dimensional analysis, variable selection is one of the very important tasks and commonly used techniques, especially in radiological and genetic research, with the high-dimensional data naturally extracted from imaging scans and gene sequencing. There is an extensive frequentist literature on variable selection, especially ones that are based on regularization techniques enforcing sparsity through penalization functions that share the common property of shrinkage toward sparse models (Tibshirani, 1996; Fan and Li, 2001; Zhang, 2010). On the other hand, Bayesian model selection expresses the sparsity through a sparse prior, such as the popular spike and slab prior (Ishwaran and Rao, 2005; George and McCulloch, 1993;

*This work was supported in part by the Simons Foundation's collaboration grant (No.635213) and the Taft Summer Research Fellowship at University of Cincinnati. This paper was supported by Samsung Research Fund, Sungkyunkwan University, 2022.

†Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio, USA, caox4@ucmail.uc.edu

‡Department of Statistics, Sungkyunkwan University, Seoul, South Korea, leekjstat@gmail.com

§Corresponding author.

© 2024 International Society for Bayesian Analysis https://doi.org/10.1214/22-BA1350

Narisetty and He, 2014) and continuous shrinkage prior (Liang et al., 2008; Johnson and Rossell, 2012; Liang et al., 2013), i.e., a distribution that supports on the sparse model or model with sparse parameters, and inference is carried out through posterior inference.

In this paper, we are interested in nonlocal priors (Johnson and Rossell, 2010) that are identically zero whenever a model parameter is equal to its null value. Compared to local priors, nonlocal prior distributions have relatively appealing properties for Bayesian model selection. Specifically, nonlocal priors discard spurious covariates faster as the sample size grows, while preserving exponential learning rates to detect nontrivial coefficients (Johnson and Rossell, 2010). Under the setup of linear regression with p predictors, Johnson and Rossell (2012) introduced the product moment (pMOM) nonlocal prior with density

$$d_p(2\pi)^{-\frac{p}{2}}(\tau\sigma^2)^{-rp-\frac{p}{2}}|U_p|^{\frac{1}{2}}\exp\Big(-\frac{\beta_p^\top U_p\beta_p}{2\tau\sigma^2}\Big)\prod_{i=1}^{p}\beta_i^{2r}. \tag{1.1}$$

Here U_p is a $p\times p$ nonsingular matrix, r is a positive integer referred to as the order of the density and d_p is the normalizing constant independent of the scale parameter τ and the error variance σ^2. Variations of the density in (1.1), called the product inverse-moment (piMOM) and product exponential moment (peMOM) density, have also been developed in Johnson and Rossell (2012) and Rossell et al. (2013). Under regularity conditions, Johnson and Rossell (2012); Shin et al. (2018) and Cao and Lee (2020) demonstrated that the posterior distributions based on the pMOM and piMOM nonlocal prior densities can achieve strong model selection consistency in high-dimensional settings. It implies that the posterior probability assigned to the true model converges to one as the sample size grows. When the number of covariates is much smaller than the sample size, Shi et al. (2019) established the posterior convergence rate of the probability regarding the Hellinger distance between the posterior model and the true model under pMOM priors in a logistic regression model.

In the pMOM prior (1.1), the hyperparameter τ controls the dispersion of the density around the origin, and thus implicitly determines the magnitude of the regression coefficients that will be shrunk to zero (Johnson and Rossell, 2012). Wu et al. (2020) and Cao et al. (2020) extended the work in Johnson and Rossell (2012) and Shin et al. (2018) by proposing a fully Bayesian approach with the pMOM nonlocal prior and an appropriate Inverse-Gamma prior on the hyperparameter τ referred to as the hyper-pMOM prior. Compared with pMOM priors, the hyer-pMOM density possesses a thicker tail and is able to accommodate large magnitudes of regression coefficients to carry out robust inference (Wu et al., 2020). In particular, Wu et al. (2020) investigated model selection properties of hyper-pMOM priors in a generalized linear model (GLM) under a fixed dimension p, and Cao et al. (2020) established strong model selection consistency of hyper-pMOM priors in linear regression when p is allowed to grow at a polynomial rate of n. For the hyper-piMOM priors composed of the mixture of piMOM and Inverse-Gamma densities, Bian and Wu (2017) established model selection consistency in generalized linear models under rather restrictive assumptions.

Despite recent developments in model selection using nonlocal priors, a rigorous Bayesian inference of hyper-pMOM priors in GLMs has not been undertaken to the

best of our knowledge. Motivated by this gap, we establish model selection consistency of the hyper-pMOM prior on regression coefficients in a GLM, in particular, logistic regression when the number of covariates grows at a sub-exponential rate of the sample size (Theorems 3.1 to 3.4). Our theory reveals that, under a uniform model prior, increasingly diffuse priors on the scale parameter τ are needed for model selection consistency, where Shin et al. (2018) reported similar conditions for the piMOM and peMOM priors. Furthermore, it is known that the computation problem can arise for Bayesian approaches due to the non-conjugate nature of priors in GLMs. To address this issue, we obtain posterior probabilities via Laplace approximation and then implement a slightly modified shotgun stochastic search algorithm for exploring the sparsity pattern of the regression coefficients. We demonstrate that the proposed method can outperform existing state-of-the-art methods including both penalized likelihood and Bayesian approaches in various settings. Finally, the proposed method is applied to an RNA-sequencing dataset consisting of gene expression levels to identify differentially expressed genes for disease risk stratification.

The rest of paper is organized as follows. Section 2 provides background material regarding GLMs and revisits the hyper-pMOM distribution. We detail strong selection consistency results in Section 3, and proofs are provided in the Supplementary Material (Cao and Lee, 2022). The posterior computation algorithm is described in Section 4, and we show the performance of the proposed method and compare it with other competitors through simulation studies in Section 5. In Section 6, we conduct a data analysis for predicting asthma and show that the hyper-pMOM prior yields better prediction performance compared with other contenders. We conclude with a discussion in Section 7.

2 Methodology

2.1 Variable Selection in Logistic Regression

We first describe the framework and introduce some notations for Bayesian variable selection in logistic regression. Let $y \in \{0,1\}^n$ be the binary response vector and $X \in \mathbb{R}^{n\times p}$ be the design matrix. Without loss of generality, we assume that the columns of X are standardized to have zero mean and unit variance. Let $x_i \in \mathbb{R}^p$ denote the ith row vector of X that contains the covariates for the ith subject. Let β be the $p \times 1$ vector of regression coefficients. We first consider the standard logistic regression model:

$$P\big(y_i = 1 \mid x_i, \beta\big) = \frac{\exp\big(x_i^\top \beta\big)}{1 + \exp\big(x_i^\top \beta\big)}, \qquad \text{for } i = 1, 2, \ldots, n. \tag{2.1}$$

We present a scenario where the dimension of predictors, p, grows with the sample size n. Thus, the number of predictors is a function of n, that is, $p = p_n$, but we denote it as p for notational simplicity. The goal of this paper is variable selection, i.e., to correctly identify all the locations of nonzero regression coefficients.

We denote a model by $k = \big\{k_1, k_2, \ldots, k_{|k|}\big\} \subseteq [p] =: \{1, 2, \ldots, p\}$ if and only if all the nonzero elements of β are $\beta_{k_1}, \beta_{k_2}, \ldots, \beta_{k_{|k|}}$, where $|k|$ is the cardinality of k.

For any $\beta \in \mathbb{R}^p$ and $k \subseteq [p]$, let $\beta_k = \left(\beta_{k_1}, \beta_{k_2}, \ldots, \beta_{k_{|k|}}\right)^\top \in \mathbb{R}^{|k|}$. Similarly, for any $m \times p$ matrix A and $k \subseteq [p]$, let $A_k \in \mathbb{R}^{m \times |k|}$ denote the submatrix of A containing the columns of A indexed by model k. In particular, for any $1 \leq i \leq n$ and $k \subseteq [p]$, we denote $x_{ik} \in \mathbb{R}^{|k|}$ as the subvector of $x_i \in \mathbb{R}^p$ containing the entries of x_i corresponding to model k.

2.2 Hierarchical Nonlocal Priors

The class of the following hierarchical nonlocal priors can be used for variable selection:

$$\pi\left(\beta_k \mid \tau, k\right) = d_k (2\pi)^{-\frac{|k|}{2}} (\tau)^{-r|k| - \frac{|k|}{2}} |U_k|^{\frac{1}{2}} \exp\Big(- \frac{{\beta_k}^\top U_k \beta_k}{2\tau} \Big) \prod_{i=1}^{|k|} \beta_{k_i}^{2r},\ \beta_k \in \mathbb{R}^{|k|}, \quad (2.2)$$

$$\pi(\tau) = \frac{\psi_2^{\psi_1}}{\Gamma(\psi_1)} \tau^{-\psi_1 - 1} \exp\Big(- \frac{\psi_2}{\tau} \Big), \quad \tau > 0, \quad (2.3)$$

where U is a $p \times p$ nonsingular matrix, r is a positive integer and ψ_1, ψ_2 are positive constants. we refer to the mixture of densities of pMOM and Inverse-Gamma in (2.2) and (2.3) as the hyper-pMOM prior (Wu et al., 2020; Cao et al., 2020). It is easy to see that the marginal density of β_k, after integrating out τ, has the following form:

$$\begin{aligned} \pi\left(\beta_k \mid k\right) &= \int \pi(\beta_k \mid \tau, k) \pi(\tau) d\tau \\ &= \frac{\psi_2^{\psi_1}}{\Gamma(\psi_1)} \frac{\Gamma(r|k| + \frac{|k|}{2} + \psi_1)}{(\psi_2 + \frac{\beta_k^\top U_k \beta_k}{2})^{r|k| + \frac{|k|}{2} + \psi_1}} d_k (2\pi)^{-\frac{|k|}{2}} |U_k|^{\frac{1}{2}} \prod_{i=1}^{|k|} \beta_{k_i}^{2r}. \end{aligned}$$

Compared to the pMOM density in (2.2) with a given τ, $\pi\left(\beta_k \mid k\right)$ possesses heavier tails as shown in Figure 1 and Figure 2. To be more specific, we consider the univariate and bivariate cases corresponding to $|k| = 1$ and $|k| = 2$ respectively. For a fair comparison, we set $\psi_1 = 1$ and let the fixed hyperparameter τ in the pMOM density equal to the expectation of the Inverse-Gamma prior (2.3), leading to $\psi_2 = \tau$. It is also important to note that the hyper-pMOM priors become increasingly diffuse as ψ_2 grows, suggesting a data-dependent value of ψ_2 should be adopted for asymptotic considerations in high dimensions.

There are several effects resulting from the mixture of priors as noted in Wu et al. (2020). First, it is clearly reflected in Figures 1 and 2 that when approaching the tail, the hyper-pMOM prior vanishes more gently than the pMOM prior, making hyper-pMOM better suited for detecting nonzero regression coefficients with relatively larger magnitudes. Second, the scale mixture of priors could achieve better empirical model selection performance especially for smaller dimensions. See for example Liang et al. (2008) and Wu et al. (2020) that investigate the finite sample performance of hyper-g and hyper-pMOM priors.

For the prior over the model space, we suggest using the following uniform prior and restricting the analysis to models with a size of less than or equal to m_n:

$$\pi(k) \quad \propto \quad \mathbb{1}(|k| \leq m_n). \quad (2.4)$$

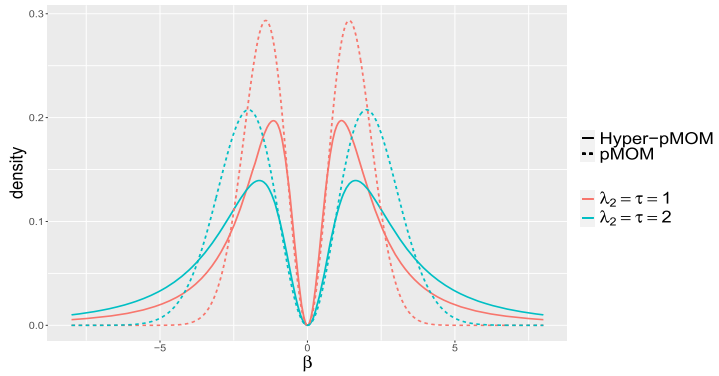
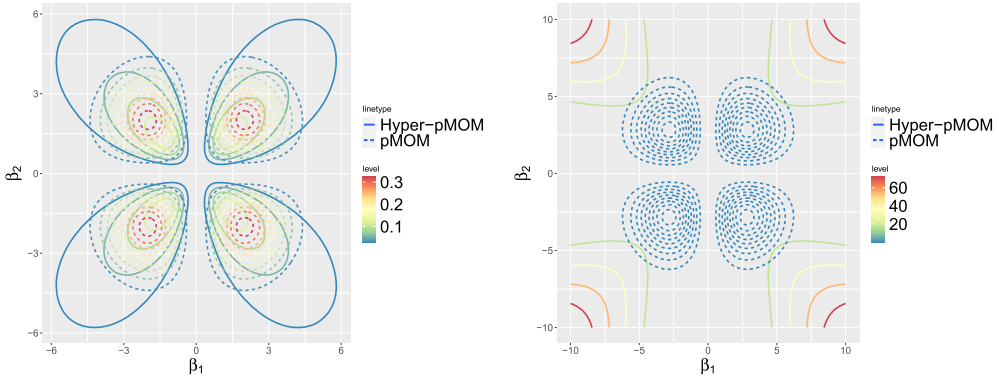


Figure 1: The univariate comparison of hyper-pMOM and pMOM densities.



(a) $\psi_2 = \tau = 1$ (b) $\psi_2 = \tau = 2$

Figure 2: The bivariate comparison of hyper-pMOM and pMOM densities.

A similar setup has also been considered in Narisetty et al. (2019), Shin et al. (2018) and Cao et al. (2020). One may view the uniform prior (2.4) constructed in the following way by assuming all models with the same size receive equal prior probability, i.e.,

$$\pi(k) = \frac{\pi(|k|)}{\binom{p}{|k|}} \text{ and } \pi(|k|) \propto \binom{p}{|k|}.$$

As alternatives to the uniform prior, one may also consider the Beta-Binomial(1,1) prior (Rossell, 2022; Scott and Berger, 2010) taking the form of

$$\pi(|k|) \propto \frac{1}{p+1}\binom{p}{|k|}^{-1}, \tag{2.5}$$

or the complexity prior (Castillo et al., 2015) with the expression of,

$$\pi(|k|) \propto c_1^{-|k|} p^{-c_2|k|},$$

for some positive constants c_1 and c_2. Note that the Beta-Binomial(1,1) yields a slightly more sparse formulation of prior on the model space compared with the uniform prior under the constraint of only considering realistically large models. One can obtain similar model selection consistency under the Beta-Binomial prior as shown in Theorem 3.4.

It is worthwhile to mention that the penalty over large models can be derived directly from the nonlocal densities themselves with large τ values, and there is no need of the extra penalization through the prior over the model space (Shin et al., 2018). In particular, Cao et al. (2020) conducted simulation studies to compare the model selection results under a uniform prior and a complexity prior using the same hyper-pMOM density with $\tau = n/2$, and they showed the superior performance of model selection under a uniform prior. If one is using nonlocal priors with finite τ or hierarchical nonlocal priors with fixed ψ_2, model priors like complexity prior or independent Bernoulli priors with small variable indicator probability (Narisetty and He, 2014) should be adopted for establishing high-dimensional model selection consistency. However, as shown in Rossell (2022), although the complexity prior attains better asymptotic results, even with a small τ value, the combined pMOM and Beta-Binomial priors are still able to discard spurious parameters and achieve better power/sparsity tradeoffs in finite samples.

Note that in the hierarchical nonlocal prior (2.1) to (2.4), no specific conditions have yet been assigned to the hyperparameters. Some standard regularity assumptions on the hyperparameters will be provided in Section 3.

By the hierarchical model (2.1) to (2.4) and Bayes' rule, the resulting posterior probability for model k is denoted by

$$\pi(k \mid y) = \frac{\pi(k)}{m(y)} m_k(y),$$

where $m(y)$ is the marginal density of y, and $m_k(y)$ is the marginal density of y under model k given by

$$\begin{aligned} m_k(y) &= \iint \exp\left\{L_n(\beta_k)\right\} \pi\left(\beta_k \mid \tau, k\right) \pi(\tau)\, d\tau d\beta_k \\ &= \int \exp\left\{L_n(\beta_k)\right\} \pi\left(\beta_k \mid k\right)\, d\beta_k, \end{aligned} \tag{2.6}$$

where

$$L_n(\beta_k) = \log\left(\prod_{i=1}^{n} \left\{\frac{\exp\left(x_{ik}^{\top}\beta_k\right)}{1+\exp\left(x_{ik}^{\top}\beta_k\right)}\right\}^{y_i} \left\{\frac{1}{1+\exp\left(x_{ik}^{\top}\beta_k\right)}\right\}^{1-y_i}\right) \tag{2.7}$$

is the log-likelihood function. The above marginal posterior probabilities for model k can be used to find the posterior mode, $\hat{k} = \arg\max_k \pi(k \mid y)$. The closed form of these posterior probabilities cannot be obtained due to the non-conjugate nature of nonlocal

densities. Therefore, special efforts need to be devoted for both consistency results and computational strategy as we shall see in the following sections. In Section 4, we will adopt a (modified) stochastic search algorithm that utilizes posterior probabilities to target the mode in a more efficient way compared with Markov chain Monte Carlo (MCMC).

2.3 Extension to Generalized Linear Model

In this section, we extend our previous discussion on logistic regression to a GLM. Given predictors x_i and an outcome y_i for $1 \leq i \leq n$, a GLM has a probability density function or probability mass function of the form

$$p(y_i \mid \theta_i) = \exp\big\{a(\theta_i)y_i + b(\theta_i) + c(y_i)\big\},$$

in which $a(\cdot)$ is a continuously differentiable function with respect to θ with nonzero derivative, $b(\cdot)$ is also a continuously differentiable function of θ, $c(\cdot)$ is some constant function of y, and $\theta_i = \theta_i(\beta) = x_i^\top \beta$ is the natural parameter.

The class of hierarchical pMOM densities specified in (2.2) and (2.3) can still be used for model selection in the generalized setting by noting that the log-likelihood function in (2.6) and (2.7) now takes the general form of

$$L_n(\beta_k) = \sum_{i=1}^{n} \big\{a(\theta_i(\beta_k))y_i + b(\theta_i(\beta_k)) + c(y_i)\big\}.$$

Using similar techniques in Section 4, one can also develop efficient search algorithms based on different log-likelihood functions to navigate the posterior mode through the model space.

3 Main Results

In this section, we show that the hyper-pMOM prior enjoys desirable model selection properties in a GLM. The results in this section are based on the assumption that p grows to infinity as n increases and do not apply to the case where p is fixed. Let $t = \{t_1, t_2, \ldots, t_{|t|}\} \subseteq [p]$ be the true model, which means that the nonzero locations of the true coefficient vector are $t = (j, j \in t)$. We consider $|t|$ to be a fixed value. Let $\beta_0 \in \mathbb{R}^p$ be the true coefficient vector and $\beta_{0,t} \in \mathbb{R}^{|t|}$ be the vector of the true nonzero coefficients. For a given model $k \subseteq [p]$, we denote $L_n(\beta_k)$ and $s_n(\beta_k) = \partial L_n(\beta_k)/(\partial \beta_k)$ as the log-likelihood and score function, respectively. In the following analysis, we will focus on logistic regression, but our argument can be extended to any other GLMs such as a probit regression model by imposing certain conditions on the design matrix to effectively bound the Hessian matrix. Let

$$H_n(\beta_k) \;=\; -\frac{\partial^2 L_n(\beta_k)}{\partial \beta_k \partial \beta_k^\top} \;=\; \sum_{i=1}^{n} \sigma_i^2(\beta_k) x_{ik} x_{ik}^\top \;=\; {X_k}^\top \Sigma(\beta_k) X_k$$

be the negative Hessian of $L_n(\beta_k)$, where $\Sigma(\beta_k) \equiv \Sigma_k = \text{diag}(\sigma_1^2(\beta_k), \ldots, \sigma_n^2(\beta_k))$, $\sigma_i^2(\beta_k) = \mu_i(\beta_k)\{1-\mu_i(\beta_k)\}$ and $\mu_i(\beta_k) = \exp\left(x_{ik}^\top \beta_k\right)/\{1+\exp\left(x_{ik}^\top \beta_k\right)\}$. In the rest of the paper, we denote $\Sigma = \Sigma(\beta_{0,t})$, $\mu = (\mu_i(\beta_{0,t}))$ and $\sigma_i^2 = \sigma_i^2(\beta_{0,t})$ for simplicity.

Before establishing the main results, we introduce the following notation. For any $a, b \in \mathbb{R}$, $a \vee b$ and $a \wedge b$ mean the maximum and minimum of a and b, respectively. For any positive real sequences a_n and b_n, we denote $a_n \lesssim b_n$, or equivalently $a_n = O(b_n)$, if there exists a constant $C > 0$ such that $|a_n| \leq C|b_n|$ for all large n. We denote $a_n \ll b_n$, or equivalently $a_n = o(b_n)$, if $a_n/b_n \longrightarrow 0$ as $n \to \infty$. We denote $a_n \sim b_n$, if there exist constants $C_1 > C_2 > 0$ such that $C_2 < b_n/a_n \leq a_n/b_n < C_1$. The ℓ_2-norm for a given vector $v = (v_1, v_2, \ldots, v_p)^\top \in \mathbb{R}^p$ is defined as $\|v\|_2 = (\sum_{j=1}^p v_j^2)^{1/2}$. For any real symmetric matrix A, let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ be maximum and minimum eigenvalue of A, respectively. We assume the following standard conditions for obtaining the asymptotic results:

Condition (A1) For some constant $0 < d' < 1$, $\log n \lesssim \log p = o(n)$ and $m_n = O\Big((n/\log p)^{\frac{1-d'}{2}} \wedge p\Big)$, as $n \to \infty$.

Condition (A2) For some constants $C > 0$, $\lambda > 0$ and $0 \leq d < (1+d)/2 \leq d' \leq 1$, $\|\beta_{0,t}\|_2^2 = O\big((\log p)^d\big)$ and

$$\begin{aligned} \max_{i,j} |x_{ij}| &\leq C, \\ 0 < \lambda \leq \min_{k:|k|\leq m_n} \lambda_{\min}\Big(n^{-1}H_n(\beta_{0,k})\Big) &\leq \Lambda_{m_n} \leq C^2\Big(\frac{n}{\log p} \wedge \log p\Big)^d, \end{aligned}$$

and $\Lambda_\zeta = \max_{k:|k|\leq\zeta} \lambda_{\max}(n^{-1}X_k^\top X_k)$ for any integer $\zeta > 0$. Furthermore, for any model $k \in \{k \subseteq [p] : |k| \leq m_n\}$ and any

$$u \in \{u \in \mathbb{R}^n : u \text{ is in the space spanned by the columns of } \Sigma^{1/2}X_k\},$$

there exists a small constant $\delta^* > 0$ such that for any $n \geq N(\delta^*)$,

$$\mathbb{E}\Big[\exp\big\{u^\top \Sigma^{-1/2}(y-\mu)\big\}\Big] \leq \exp\Big\{\frac{(1+\delta^*)u^\top u}{2}\Big\}.$$

Condition (A3) For some constant $c_0 > 0$,

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

Condition (A4) For some constants $\delta, a_1, a_2 > 0$, the hyperparameters satisfy

$$a_1 < \psi_1 < a_2, \;\; \psi_2^{r+1/2} \sim n^{-1/2}p^{2+\delta} \;\text{ and }\; a_1 < \lambda_{\min}(U) \leq \lambda_{\max}(U) < a_2,$$

where r is a positive integer in (2.2).

Condition (A1) ensures our proposed method can accommodate high dimensions where the number of predictors grows at a sub-exponential rate of n. Condition (A1) also specifies the parameter m_n in the uniform prior (2.4) that restricts our analysis on

a set of *reasonably large* models. We essentially require $m_n \ll n$ to avoid over-fitting problems. Similar assumptions restricting the model size have been commonly assumed in the sparse estimation literature (Liang et al., 2013; Narisetty et al., 2019; Shin et al., 2018; Lee et al., 2019).

Condition (A2) gives lower and upper bounds of $\lambda_{\min}\big(n^{-1}H_n(\beta_{0,k})\big)$ and $\lambda_{\max}\big(n^{-1}X_k^\top X_k\big)$, respectively, where k belongs to the set of reasonably large models. The lower bound condition can be seen as a restricted eigenvalue condition for k-sparse vectors and is satisfied with high probability for sub-Gaussian design matrices (Narisetty et al., 2019). Similar conditions have been used in the linear regression literature (Ishwaran and Rao, 2005; Yang et al., 2016; Song and Liang, 2017). For the last assumption in Condition (A2), as stated in Narisetty et al. (2019), due to the sub-Gaussianity of $\Sigma^{-1/2}(y-\mu)$, for typical random designs, the variable $u^\top\Sigma^{-1/2}(y-\mu)/\|u\|$ is asymptotically distributed as the standard normal, so this assumption is expected to hold for some small constant $\delta^* > 0$.

Condition (A2) also allows the magnitude of true signals to increase to infinity but stay bounded above by $(\log p)^d$ up to some constant, while Condition (A3), the well-known *beta-min* condition, gives a lower bound for nonzero signals. In general, this type of condition is necessary to not neglect any small signals.

Condition (A4) suggests appropriate conditions for the hyperparameters in (2.2) and (2.3). Similar assumption has also been considered in Shin et al. (2018), Johnson and Rossell (2012) and Cao et al. (2020). In particular, we extend the previous polynomial rate of the dimension in Cao et al. (2020) by considering a larger order of the hyperparameter ψ_2. Note that the prior expectation and variance of τ are $\psi_2/(\psi_1-1)$ and $\psi_2^2/\{(\psi_1-1)^2(\psi_1-2)\}$, respectively, if $\psi_1 > 2$. Therefore, the priors on τ satisfying Condition (A4) are increasingly diffuse as n grows.

3.1 Model Selection Consistency

Theorem 3.1 (No super set)**.** *Under Conditions* (A1)*,* (A2) *and* (A4)*,*

$$\pi\big(k \supsetneq t \mid y\big) \xrightarrow{P} 0, \quad \text{as } n \to \infty.$$

Theorem 3.1 says that, asymptotically, our posterior does not overfit the model, i.e., it does not include unnecessarily many variables. Of course, the result does not guarantee that the posterior will concentrate on the true model. To capture every significant variable, we require the magnitudes of nonzero entries in $\beta_{0,t}$ not to be too small. Theorem 3.2 shows that with an appropriate lower bound specified in Condition (A3), the true model t will be the mode of the posterior.

Theorem 3.2 (Posterior ratio consistency)**.** *Under Conditions* (A1)–(A4) *with* $c_0 = \{(1-\epsilon_0)\lambda\}^{-1}\big[2(4+2\delta)+5\{(1-\epsilon_0)\lambda\}^{-1}\big]$ *for some small constant* $\epsilon_0 > 0$*,*

$$\max_{k\neq t}\frac{\pi\big(k \mid y\big)}{\pi\big(t \mid y\big)} \xrightarrow{P} 0, \quad \text{as } n \to \infty.$$

Posterior ratio consistency is a useful property especially when we are interested in the point estimation with the posterior mode, but does not provide how large is the probability that the posterior puts on the true model. In the following theorem, we state that our posterior achieves *strong selection consistency*. By strong selection consistency, we mean that the posterior probability assigned to the true model t converges to 1, which requires a slightly stronger condition on the lower bound for the magnitudes of nonzero entries in $\beta_{0,t}$ compared to that in Theorem 3.2.

Theorem 3.3 (Strong selection consistency)**.** *Under Conditions (A1)–(A4) with* $c_0 = \{(1-\epsilon_0)\lambda\}^{-1}\big[2(9+2\delta)+5\{(1-\epsilon_0)\lambda\}^{-1}\big]$ *for some small constant* $\epsilon_0 > 0$*, the following holds:*

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

As discussed in Section 2, one may also choose other types of priors over the model space. Indeed, next theorem establishes that the strong selection consistency can likewise be achieved under the Beta-Binomial prior.

Theorem 3.4 (Consistency under Beta-Binomial (1,1))**.** *If instead of the uniform prior, the Beta-Binomial (1,1) distribution in (2.5) is imposed over the model space, then under the exact same conditions stated in Theorem 3.3, the following holds:*

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

3.2 Comparison with Existing Work

We compare our results and assumptions with those of existing methods using nonlocal priors in generalized linear regression. Shi et al. (2019) established the posterior convergence rate for nonlocal priors under the assumption of $p\log(1/\epsilon_n^2) \ll n\epsilon_n^2$ for some $\epsilon_n \in (0,1]$ satisfying $n\epsilon_n^2 \gg 1$, which indicates that p can increase with the sample size but slower than n. Wu et al. (2020) investigated the model selection performance of hyper-nonlocal priors that combine the Fisher information matrix with the pMOM density and established asymptotic properties under a fixed dimension of predictors. Both works considered the setting of low to moderate dimensions, while we allow p to grow at a sub-exponential rate of n, the so-called "ultra high-dimensional" setting (Shin et al., 2018).

Bian and Wu (2017) considered the following hyper-piMOM priors for regression coefficients in GLMs and established the high-dimensional model selection consistency:

$$\begin{aligned} \beta_k \mid \tilde{\tau}_k \quad &\sim \quad \prod_{i=1}^{|k|} \frac{(\tau_i\sigma^2)^{\frac{r}{2}}}{\Gamma(\frac{r}{2})}|\beta_{k_i}|^{-(r+1)} \exp\Big(-\frac{\tau_i\sigma^2}{\beta_{k_i}^2}\Big), \\ \tau_i \quad &\overset{i.i.d.}{\sim} \quad \text{Inverse-Gamma}\,\Big(\frac{r+1}{2}, \psi\Big), \qquad \text{for } i = 1, \ldots, |k|, \end{aligned}$$

where $\tilde{\tau}_k = (\tau_1, \tau_2, \ldots, \tau_{|k|})^\top$. In particular, the authors put an independent piMOM prior on each linear regression coefficient (conditional on the hyperparameter τ_i) and an Inverse-Gamma prior on τ_i.

There are some fundamental differences between Bian and Wu (2017) and our work in terms of the models considered and corresponding analysis. Firstly, unlike the piMOM prior, the pMOM prior in our model does not in general correspond to assigning an independent prior to each entry of β_k. In particular, pMOM distributions introduce correlations among the entries in β_k through U_k and create more theoretical challenges. Furthermore, as $\beta \to 0$, piMOM converges to 0 faster at a quasi-exponential shrinkage rate, while pMOM densities yield a polynomial shrinkage rate Rossell and Telesca (2017). These different prior dispersions near the origin lead to the piMOM densities imposing a stronger penalty on coefficients with magnitudes close to 0 and being more conservative in identifying nonzero signals compared with the pMOM densities. This means by exploiting piMOM, one may avoid more false positives at the cost of a potential decrease in statistical power. Finally, from a practical standpoint, as shown in Rossell et al. (2021), pMOM can facilitate the approximate Laplace approximation to attain a faster and at the same time better inference compared with the Laplace approximation. In terms of technical assumptions, Bian and Wu (2017) assumed the eigenvalues of the Hessian matrix to be bounded below and above by some constants, while we allow the upper bound to grow with n (Condition (A2)). In addition, to prove the model selection consistency, Bian and Wu (2017) required the spectral norm of the difference between the Hessian matrices corresponding to any two models to be bounded above by a function of the ℓ_2-norm difference between the respective regression coefficients, and they assumed that the product of the response variables and the entries of design matrix are bounded by a constant, while these constraints are not imposed in our study. See assumptions B1, B2 and C1 in Bian and Wu (2017) for details. Thirdly, no simulation studies were conducted in Bian and Wu (2017), leaving the empirical validity of the proposed method in question, while we include the computational strategy in the following section and examine the practical utility of the hyper-pMOM prior through gene expression analysis.

4 Posterior Computation

In this section, we describe how to approximate the marginal density of data and conduct the model selection procedure. The integral formulation in (2.6) cannot be calculated in a closed form. Hence, we use Laplace approximation to compute $m_k(y)$ and $\pi(k \mid y)$. Similar approaches to compute posterior probabilities have been used in Johnson and Rossell (2012), Shi et al. (2019) and Shin et al. (2018). We note here that the model selection results in Section 3 assume the exact marginal likelihoods and do not apply to the Laplace-approximated marginal likelihoods.

For any model k, when $U_k = I_k$, the normalization constant d_k in (2.2) is given by $d_k = \left\{(2r-1)!!\right\}^{-|k|}$. Let

$$
\begin{aligned}
f(\beta_k) &= \log\left(\exp\left\{L_n(\beta_k)\right\}\pi(\beta_k \mid k)\right) \\
&= \sum_{i=1}^{n}\left\{y_i x_{ik}^{\top}\beta_k - \log\left(1+\exp(x_{ik}^{\top}\beta_k)\right)\right\} - |k|\log\left((2r-1)!!\right) - \frac{|k|}{2}\log(2\pi)
\end{aligned}
$$

Algorithm 1 Shotgun Stochastic Search (SSS).

Set an initial model $k^{(1)}$
for $i = 1$ to $i = N - 1$ **do**
(a) Compute $\pi(k \mid y)$ using (4.1) for all $k \in \text{nbd}(k^{(i)})$
(b) Sample k^+, k^- and k^0 from Γ_k^+, Γ_k^- and Γ_k^0 with probabilities proportional to $\pi(k \mid y)$
(c) Sample the next model $k^{(i+1)}$ from $\{k^+, k^-, k^0\}$ with probability proportional to $\{\pi(k^+ \mid y), \pi(k^- \mid y), \pi(k^0 \mid y)\}$
end for

$$+ \log\Big(\frac{\psi_2^{\psi_1}}{\Gamma(\psi_1)}\Big) - \Big(r|k| + \frac{|k|}{2} + \psi_1\Big) \log\Big(\psi_2 + \frac{1}{2}\|\beta_k\|_2^2\Big)$$
$$+ 2r \sum_{i=1}^{|k|} \log\big(|\beta_{k_i}|\big) + \log\Gamma\Big(r|k| + \frac{|k|}{2} + \psi_1\Big).$$

For any model k, the Laplace approximation of $m_k(y)$ is given by

$$(2\pi)^{\frac{|k|}{2}} \exp\big\{f(\widehat{\beta}_k)\big\} |V(\widehat{\beta}_k)|^{-\frac{1}{2}}, \tag{4.1}$$

where $\widehat{\beta}_k = \arg\max_{\beta_k} f(\beta_k)$ is obtained via the optimization function `optim` in R using a quasi-Newton method, and $V(\beta_k)$ is a $|k| \times |k|$ symmetric matrix defined as

$$\begin{aligned} V(\beta_k) \;=\; & -\sum_{i=1}^{n} \frac{x_{ik} x_{ik}^\top \exp(x_{ik}^\top \beta_k)}{\big\{1 + \exp(x_{ik}^\top \beta_k)\big\}^2} - \text{diag}\Big(\frac{2r}{\beta_{k_1}^2}, \ldots, \frac{2r}{\beta_{k_{|k|}}^2}\Big) \\ & - \Big(r|k| + \frac{|k|}{2} + \psi_1\Big)\Big\{\frac{1}{\psi_2 + \|\beta_k\|_2^2/2} I_{|k|} - \frac{1}{(\psi_2 + \|\beta_k\|_2^2/2)^2} \beta_k \beta_k^\top\Big\}. \end{aligned}$$

The above Laplace approximation can be used to compute the posterior probability ratio between two models.

The shotgun stochastic search (SSS) algorithm (Hans et al., 2007; Shin et al., 2018) is inspired by MCMC but enables much more efficient identification of probable models by swiftly moving around in the model space as the dimension escalates. The SSS algorithm explores high-dimensional model spaces and quickly identifies "interesting" regions of high posterior probability over models. The SSS evaluates numerous models guided by the unnormalized posterior probabilities that can be approximated using the Laplace approximations of the marginal probabilities in (4.1). Let $\text{nbd}(k) = \{\Gamma_k^+, \Gamma_k^-, \Gamma_k^0\}$ containing all the neighbors of model k, in which $\Gamma_k^+ = \big\{k \cup \{j\} : j \notin k\big\}$, $\Gamma_k^- = \big\{k \setminus \{j\} : j \in k\big\}$ and $\Gamma_k^0 = \big\{k \setminus \{j\} \cup \{l\} : j \in k, l \notin k\big\}$. Algorithm 1 describes the SSS procedure.

However, as pointed out by Shin et al. (2018), the SSS algorithm can be computationally expensive in high-dimensional settings. The computational bottleneck is calculating the Laplace approximations of the marginal probabilities for the models in Γ_k^+ and Γ_k^0, whose cardinalities are $p - |k|$ and $(p - |k|)|k|$, respectively. To alleviate computational burden, we slightly modify the SSS algorithm by reducing the number of entries in Γ_k^+ and Γ_k^0. Specifically, we reduce the number of models in Γ_k^+ by selecting only (1) the top K_1 variables having large absolute sample correlation with y and (2) K_2 randomly

Algorithm 2 Reduced Shotgun Stochastic Search (RSSS).

Set an initial model $k^{(1)}$
for $i = 1$ to $i = N - 1$ **do**
(a) Compute $\pi(k \mid y)$ using (4.1) for all $k \in \text{nbd}_R(k^{(i)}) = \{\Gamma^+_{R,k}, \Gamma^-_k, \Gamma^0_{R,k}\}$
(b) Sample k^+, k^- and k^0 from $\Gamma^+_{R,k}$, Γ^-_k and $\Gamma^0_{R,k}$ with probabilities proportional to $\pi(k \mid y)$
(c) Sample the next model $k^{(i+1)}$ from $\{k^+, k^-, k^0\}$ with probability proportional to $\{\pi(k^+ \mid y), \pi(k^- \mid y), \pi(k^0 \mid y)\}$
end for

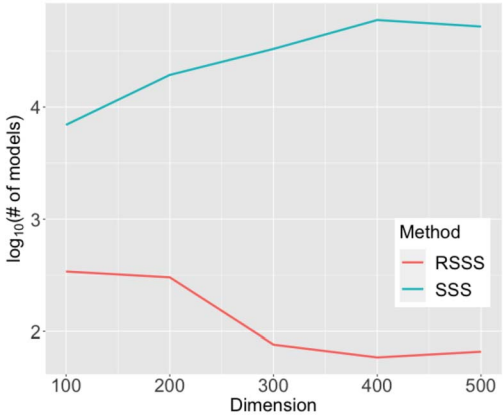


Figure 3: The average number of models searched before visiting the posterior mode.

selected variables, and we define the resulting set as $\Gamma^+_{R,k}$. Similarly, we define a reduced set $\Gamma^0_{R,k} = \{k \setminus \{j\} \cup \{l\} : j \in k, l \in \Gamma^+_{R,k}\}$ and replace Γ^0_k with $\Gamma^0_{R,k}$ in the algorithm. By doing so, we can efficiently reduce the computational complexity of the algorithm. Note that the cardinalities of $\Gamma^+_{R,k}$ and $\Gamma^0_{R,k}$ are $K_1 + K_2$ and $(K_1 + K_2)|k|$, respectively. We call this modified algorithm the reduced SSS (RSSS) algorithm and describe it in Algorithm 2. In the subsequent simulation study and real data analysis, the RSSS algorithm with $K_1 = K_2 = 10$ is adopted for the posterior inference of the hyper-pMOM prior.

Note that the RSSS algorithm is different from the simplified shotgun stochastic search with screening (S5) algorithm (Shin et al., 2018). The three main differences are that the RSSS algorithm (i) does not completely ignore the set Γ^0_k, (ii) does not introduce temperature parameters and (iii) uses the marginal correlation between X and y. Note that Shin et al. (2018) used the correlation between X and residuals of the current model to construct the set $\Gamma^+_{R,k}$. In the Supplementary Material, we conduct additional simulation studies to compare the performances of the two RSSS algorithms using (a) the marginal correlation between X and y and (b) the correlation between X and residuals of the current model.

To demonstrate the computational efficiency of the RSSS algorithm and compare it with the SSS algorithm, we conduct a simulation study. We generate the data from the

model (2.1) with the true coefficient $\beta_0 = (1, 1, 1, 0, \ldots, 0)^\top \in \mathbb{R}^p$ and design matrix $X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n \times p}$, where $x_i \overset{i.i.d.}{\sim} N_p(0, \Sigma)$ for $i = 1, \ldots, n$. The covariance matrix $\Sigma = (\Sigma_{ij})$ is chosen as $\Sigma_{ij} = 0.8^{|i-j|}$ for any $1 \leq i \leq j \leq p$ to consider highly correlated covariates. The number of samples is fixed at $n = 200$, while the number of variables varies over $p \in \{100, \ldots, 500\}$. Figure 3 shows the average number of models searched before visiting the posterior mode for each p, where the averages are calculated based on 100 repetitions. When compared with the SSS algorithm, the RSSS algorithm investigates a far less number of models before hitting the posterior mode. Although the two algorithms did not always find the same posterior modes, they found exactly the same models in approximately 64 simulations among 100 repetitions. We also found that the RSSS algorithm produced fewer false positives and false negatives than the SSS algorithm. Specifically, when $p \in \{100, \ldots, 500\}$, the average numbers of false positives/negatives of the RSSS algorithm are 1.80, 1.87, 1.59, 1.65 and 1.73, respectively, while those of the SSS algorithm are 3.16, 4.08, 4.59, 5.40 and 4.55, respectively. Therefore, the RSSS algorithm can achieve reasonable performance compared to the SSS algorithm while boosting computing efficiency.

We note here that the proposed RSSS algorithm can also be used to obtain estimated posterior probabilities. As discussed in Shin et al. (2018), the normalizing constant of the posterior model probability can be approximated by adding the unnormalized posterior probabilities of all models visited by the RSSS algorithm. Furthermore, the Associate Editor suggested an alternative method, which considers the RSSS as a proposal kernel of a Metropolis-Hastings algorithm. Then, after obtaining MCMC samples of models, we can estimate posterior inclusion probabilities of each variable and posterior model probabilities of each model by approximating the normalizing constant as described above.

The RSSS algorithm can be further modified by selecting variables based on other criteria instead of considering randomly selected variables in the set Γ_k^+. For example, similar to the suggestion of Griffin et al. (2021), we can choose K_2 variables having the largest estimated posterior inclusion probabilities.

5 Simulation Studies

In this section, we investigate the performance of the hyper-pMOM prior for logistic regression models. For given $n = 200$ and $p \in \{100, 300, 500\}$, simulated data sets are generated from (2.1) with the true coefficient vector β_0 and design matrix $X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n \times p}$. We set the index set for nonzero values in β_0 at $t = \{1, 2, 3\}$, where nonzero coefficients $\beta_{0,t}$ are generated under the following two different settings:

- Setting 1 (Weak signals): All the entries of $\beta_{0,t}$ are set to 0.5.
- Setting 2 (Moderate signals): All the entries of $\beta_{0,t}$ are set to 1.
- Setting 3 (Large signals): All the entries of $\beta_{0,t}$ are set to 2.

We generate covariate vectors as $x_i \overset{i.i.d.}{\sim} N_p(0, \Sigma)$ for $i = 1, \ldots, n$, under the following cases of Σ:

- Case 1 (Isotropic design): $\Sigma = I_p$

- Case 2 (Correlated design): $\Sigma = (\Sigma_{ij})$, where $\Sigma_{ij} = 0.3^{|i-j|}$ for any $1 \le i \le j \le p$.

We also generate test samples $\{(y_{\text{test},1}, x_{\text{test},1}), \ldots, (y_{\text{test},n_{\text{test}}}, x_{\text{test},n_{\text{test}}})\}$ with $n_{\text{test}} = 50$ to evaluate the prediction performance.

In the various scenarios mentioned above, we compare the performance of the hyper-pMOM prior (H-pMOM) with existing variable selection methods. As Bayesian contenders, we consider the nonlocal pMOM prior (Cao and Lee, 2020), piMOM prior (Johnson and Rossell, 2012), spike and slab (SS) prior (Tüchler, 2008) and empirical Bayesian Lasso (EBLasso) (Cai et al., 2011), while we consider Lasso (Friedman et al., 2010), smoothly clipped absolute deviations (SCAD) (Breheny and Huang, 2011) and minimax concave penalty (MCP) (Zhang, 2010) as frequentist competitors.

The R codes for implementing the hyper-pMOM prior are publicly available at `https://github.com/leekjstat/Hierarchical-nonlocal`. Among the hyperparameters in (2.2) and (2.3), we set $U = I_p$, $r = 1$ and $\psi_1 = 1$. Furthermore, we consider $\psi_2 \in \{10^l n^{-1/3} p^{(2+0.001)2/3} : l = -1, 0, 1, 2\}$ and choose the value of ψ_2 that gives the minimum mean squared prediction error based on 5-fold cross-validation. The hyperparameters of the pMOM prior are set at $U = I_p$ and $r = 1$, and we find a set of τs that makes the univariate marginal prior variance of β_j of the pMOM prior and that of the hyper-pMOM prior with each $\psi_2 \in \{10^l n^{-1/3} p^{(2+0.001)2/3} : l = -1, 0, 1, 2\}$ the same. After that, we choose the value of τ giving the minimum mean squared prediction error through 5-fold cross-validation. For a fair comparison between the hyper-pMOM and pMOM priors, we use the same RSSS algorithm to find a posterior mode for both priors. For the RSSS algorithm, initial models are set by randomly taking three nonzero entries. The piMOM prior is implemented by the `BVSNLP` package in R, where an adaptive hyperparameter selection method (Nikooienejad et al., 2016) is used. For the regularization approaches, the tuning parameters are chosen by 5-fold cross-validation.

To examine the performance of each method, the values of the precision, sensitivity, specificity, Matthews correlation coefficient (MCC) (Matthews, 1975) and mean squared prediction error (MSPE) are used. These criteria are defined as

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Sensitivitiy} = \frac{TP}{TP+FN}, \quad \text{Specificity} = \frac{TN}{TN+FP},$$
$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

and $\text{MSPE} = (n_{\text{test}})^{-1} \sum_{i=1}^{n_{\text{test}}} \left(\hat{y}_i - y_{\text{test},i}\right)^2$, where TP, TN, FP and FN are true positive, true negative, false positive and false negative, respectively. Here, $\hat{y}_i = \exp(x_{\text{test},i}^\top \hat{\beta})/\{1+\exp(x_{\text{test},i}^\top \hat{\beta})\}$, where $\hat{\beta}$ is the estimated coefficient vector. For the hyper-pMOM and pMOM priors, the nonzero part in $\hat{\beta}$ is chosen as the posterior mode with the estimated model $\hat{k}$, i.e., $\hat{\beta}_{\hat{k}} = \arg\max_{\hat{k}} f(\beta_{\hat{k}})$. For the spike and slab prior, the posterior mean based on 2,000 posterior samples is used as $\hat{\beta}$. The averages of each criterion based on 100 repetitions are summarized in Tables 1–6. Furthermore, we also report the proportion of simulations where the true model was selected, which we denote as $P(\hat{k} = t)$.

Based on the simulation results, Bayesian methods tend to achieve high precision, specificity, MCC and $P(\hat{k} = t)$, which means that they produce low false positives.

		Precision	Sensitivity	Specificity	MCC	$P(\hat{k}=t)$	MSPE
Setting 1	H-pMOM	0.933	0.420	0.998	0.599	0.03	0.242
	pMOM	0.958	0.380	**0.999**	0.588	0.02	0.240
	piMOM	**0.962**	0.423	**0.999**	0.524	**0.10**	0.238
	SS	0.946	0.460	**0.999**	0.582	0.08	0.243
	EBLasso	0.704	0.683	0.977	**0.635**	0.09	0.246
	Lasso	0.367	**0.890**	0.912	0.493	0.03	0.234
	SCAD	0.377	0.887	0.924	0.506	0.02	**0.233**
	MCP	0.570	0.800	0.968	0.598	0.09	**0.233**
Setting 2	H-pMOM	0.965	0.960	0.998	0.957	0.79	0.174
	pMOM	**0.988**	0.937	**0.999**	0.955	0.80	0.173
	piMOM	0.956	0.997	0.998	**0.974**	**0.82**	**0.167**
	SS	0.958	0.990	0.998	0.971	**0.82**	0.173
	EBLasso	0.835	0.957	0.989	0.874	0.49	0.234
	Lasso	0.287	**1.000**	0.888	0.491	0.01	0.179
	SCAD	0.325	**1.000**	0.918	0.536	0.01	0.172
	MCP	0.623	**1.000**	0.973	0.767	0.20	0.171
Setting 3	H-pMOM	0.981	1.000	0.999	0.989	0.95	0.107
	pMOM	**0.998**	1.000	**1.000**	**0.999**	**0.99**	**0.106**
	piMOM	0.954	1.000	0.998	0.974	0.84	0.108
	SS	0.899	1.000	0.996	0.943	0.63	**0.106**
	EBLasso	0.681	1.000	0.981	0.810	0.16	0.258
	Lasso	0.189	1.000	0.833	0.391	0.00	0.119
	SCAD	0.471	1.000	0.955	0.661	0.02	0.108
	MCP	0.750	1.000	0.984	0.850	0.38	0.108

Table 1: The summary statistics for Case 1 (isotropic design) when $p = 100$.

		Precision	Sensitivity	Specificity	MCC	$P(\hat{k}=t)$	MSPE
Setting 1	H-pMOM	0.962	0.443	0.999	0.630	0.06	0.227
	pMOM	**0.980**	0.360	**1.000**	0.583	0.01	0.229
	piMOM	0.938	0.523	0.998	0.669	0.06	0.223
	SS	0.945	0.713	0.998	**0.790**	**0.31**	0.226
	EBLasso	0.769	0.593	0.983	0.620	0.04	0.238
	Lasso	0.425	**0.920**	0.936	0.568	0.03	**0.220**
	SCAD	0.479	0.907	0.953	0.609	0.05	0.221
	MCP	0.665	0.817	0.978	0.693	0.15	**0.220**
Setting 2	H-pMOM	0.933	0.890	0.996	0.896	0.56	0.159
	pMOM	**0.982**	0.863	**0.999**	0.911	0.60	0.158
	piMOM	0.944	0.960	0.998	**0.947**	**0.72**	**0.153**
	SS	0.929	0.970	0.997	0.944	0.71	0.161
	EBLasso	0.845	0.923	0.989	0.862	0.48	0.222
	Lasso	0.317	**1.000**	0.901	0.519	0.03	0.159
	SCAD	0.363	**1.000**	0.927	0.568	0.03	0.157
	MCP	0.646	0.997	0.975	0.781	0.19	0.156
Setting 3	H-pMOM	0.988	**1.000**	0.999	**0.994**	**0.96**	**0.089**
	pMOM	**0.990**	0.997	**1.000**	0.993	0.95	0.090
	piMOM	0.960	**1.000**	0.998	0.978	0.84	0.090
	SS	0.879	**1.000**	0.995	0.932	0.56	**0.089**
	EBLasso	0.774	**1.000**	0.988	0.869	0.32	0.227
	Lasso	0.213	**1.000**	0.849	0.417	0.00	0.101
	SCAD	0.347	**1.000**	0.930	0.561	0.00	0.094
	MCP	0.630	**1.000**	0.976	0.774	0.16	0.093

Table 2: The summary statistics for Case 2 (correlated design) when $p = 100$.

On the other hand, frequentist methods show high sensitivity, which means that they produce low false negatives. Compared with the pMOM prior, the hyper-pMOM prior gives better MCC, $P(\hat{k}=t)$ and MSPE for weak and moderate signal settings (Settings

		Precision	Sensitivity	Specificity	MCC	$P(\hat{k}=t)$	MSPE
Setting 1	H-pMOM	0.818	0.323	0.999	0.501	0.02	0.245
	pMOM	0.827	0.283	0.999	0.480	0.00	0.245
	piMOM	**0.971**	0.227	**1.000**	0.310	**0.03**	0.244
	SS	0.919	0.277	**1.000**	0.393	0.01	0.247
	EBLasso	0.585	0.557	0.988	**0.502**	0.02	0.249
	Lasso	0.441	**0.687**	0.970	0.416	0.01	0.241
	SCAD	0.444	**0.687**	0.974	0.422	0.01	0.240
	MCP	0.570	0.600	0.988	0.461	**0.03**	**0.238**
Setting 2	H-pMOM	0.973	0.880	0.999	0.909	0.72	0.182
	pMOM	**1.000**	0.760	**1.000**	0.852	0.56	0.189
	piMOM	0.973	0.980	**1.000**	**0.973**	**0.85**	**0.172**
	SS	0.967	0.940	**1.000**	0.948	0.74	0.186
	EBLasso	0.828	0.867	0.995	0.811	0.41	0.235
	Lasso	0.267	**1.000**	0.957	0.486	0.01	0.190
	SCAD	0.225	**1.000**	0.956	0.453	0.00	0.184
	MCP	0.520	**1.000**	0.987	0.700	0.10	0.181
Setting 3	H-pMOM	0.991	1.000	**1.000**	0.995	0.97	0.110
	pMOM	**0.998**	1.000	**1.000**	**0.999**	**0.99**	0.110
	piMOM	0.969	1.000	**1.000**	0.983	0.89	0.110
	SS	0.902	1.000	0.999	0.946	0.66	**0.109**
	EBLasso	0.562	1.000	0.990	0.738	0.06	0.269
	Lasso	0.129	1.000	0.912	0.335	0.00	0.128
	SCAD	0.268	1.000	0.967	0.501	0.00	0.115
	MCP	0.601	1.000	0.991	0.759	0.14	0.113

Table 3: The summary statistics for Case 1 (isotropic design) when $p = 300$.

		Precision	Sensitivity	Specificity	MCC	$P(\hat{k}=t)$	MSPE
Setting 1	H-pMOM	0.967	0.340	**1.000**	0.568	0.00	0.232
	pMOM	**0.980**	0.327	**1.000**	0.564	0.00	0.232
	piMOM	0.953	0.383	**1.000**	0.566	0.00	0.231
	SS	0.954	0.520	**1.000**	**0.670**	**0.09**	0.233
	EBLasso	0.720	0.520	0.994	0.564	0.01	0.242
	Lasso	0.378	**0.883**	0.970	0.526	0.03	**0.226**
	SCAD	0.382	0.873	0.972	0.530	0.03	**0.226**
	MCP	0.588	0.727	0.989	0.606	0.05	0.227
Setting 2	H-pMOM	0.963	0.753	0.999	0.830	0.39	0.171
	pMOM	**0.998**	0.630	**1.000**	0.776	0.20	0.178
	piMOM	0.953	0.903	0.999	0.919	0.60	**0.158**
	SS	0.948	0.913	0.999	**0.923**	**0.62**	0.169
	EBLasso	0.755	0.900	0.994	0.792	0.26	0.228
	Lasso	0.276	**1.000**	0.956	0.493	0.00	0.164
	SCAD	0.285	**1.000**	0.965	0.510	0.00	0.162
	MCP	0.557	**1.000**	0.988	0.725	0.14	0.160
Setting 3	H-pMOM	0.985	**1.000**	0.999	0.991	0.97	0.093
	pMOM	**0.995**	**1.000**	**1.000**	**0.997**	**0.98**	**0.091**
	piMOM	0.966	**1.000**	0.999	0.981	0.88	0.092
	SS	0.854	**1.000**	0.998	0.919	0.51	0.091
	EBLasso	0.692	0.993	0.994	0.816	0.23	0.240
	Lasso	0.151	**1.000**	0.923	0.361	0.00	0.107
	SCAD	0.205	**1.000**	0.955	0.437	0.00	0.099
	MCP	0.478	**1.000**	0.985	0.673	0.06	0.097

Table 4: The summary statistics for Case 2 (correlated design) when $p = 300$.

1 and 2), while the pMOM tends to show slightly better MCC, $P(\hat{k}=t)$ and MSPE for strong signal setting (Setting 3).

Although we only report the results of hyper-pMOM and pMOM priors with adap-

		Precision	Sensitivity	Specificity	MCC	$P(\hat{k}=t)$	MSPE
Setting 1	H-pMOM	0.807	0.287	**1.000**	**0.477**	0.00	0.246
	pMOM	0.810	0.280	**1.000**	0.473	0.00	0.246
	piMOM	**0.971**	0.203	**1.000**	0.289	**0.01**	0.245
	SS	0.938	0.250	**1.000**	0.369	0.00	0.247
	EBLasso	0.533	0.477	0.992	0.441	**0.01**	0.250
	Lasso	0.433	**0.667**	0.982	0.389	**0.01**	0.240
	SCAD	0.424	**0.667**	0.981	0.383	**0.01**	0.240
	MCP	0.571	0.573	0.992	0.416	**0.01**	**0.239**
Setting 2	H-pMOM	0.945	0.903	0.999	0.912	0.63	0.183
	pMOM	**0.994**	0.750	**1.000**	0.845	0.49	0.193
	piMOM	0.956	0.980	**1.000**	**0.960**	**0.80**	**0.174**
	SS	0.944	0.960	**1.000**	0.947	0.72	0.186
	EBLasso	0.780	0.870	0.996	0.787	0.29	0.240
	Lasso	0.187	**1.000**	0.960	0.407	0.00	0.191
	SCAD	0.160	**1.000**	0.961	0.383	0.00	0.185
	MCP	0.394	**1.000**	0.988	0.611	0.02	0.181
Setting 3	H-pMOM	0.983	1.00	**1.000**	0.990	0.96	0.109
	pMOM	**0.998**	1.00	**1.000**	**0.999**	**0.99**	**0.108**
	piMOM	0.961	1.00	**1.000**	0.979	0.86	0.109
	SS	0.884	1.00	0.999	0.936	0.59	**0.108**
	EBLasso	0.529	1.00	0.993	0.716	0.05	0.271
	Lasso	0.125	1.00	0.941	0.329	0.00	0.127
	SCAD	0.221	1.00	0.975	0.457	0.00	0.113
	MCP	0.579	1.00	0.994	0.744	0.14	0.112

Table 5: The summary statistics for Case 1 (isotropic design) when $p = 500$.

		Precision	Sensitivity	Specificity	MCC	$P(\hat{k}=t)$	MSPE
Setting 1	H-pMOM	0.990	0.347	**1.000**	0.582	0.01	0.227
	pMOM	0.990	0.333	**1.000**	0.573	0.00	0.226
	piMOM	**0.992**	0.407	**1.000**	0.611	0.02	0.222
	SS	0.962	0.573	**1.000**	**0.714**	**0.15**	0.228
	EBLasso	0.704	0.517	0.996	0.559	0.01	0.239
	Lasso	0.356	**0.873**	0.979	0.503	0.01	**0.221**
	SCAD	0.353	0.863	0.982	0.500	0.01	**0.221**
	MCP	0.555	0.730	0.994	0.592	0.03	**0.221**
Setting 2	H-pMOM	0.956	0.733	0.999	0.814	0.33	0.170
	pMOM	**1.000**	0.563	**1.000**	0.734	0.14	0.175
	piMOM	0.965	0.893	**1.000**	**0.921**	**0.61**	**0.154**
	SS	0.948	0.893	**1.000**	0.910	0.56	0.166
	EBLasso	0.796	0.830	0.997	0.777	0.25	0.222
	Lasso	0.231	**1.000**	0.965	0.446	0.02	0.162
	SCAD	0.236	**1.000**	0.970	0.458	0.02	0.159
	MCP	0.477	0.987	0.990	0.664	0.07	0.156
Setting 3	H-pMOM	0.979	0.993	**1.000**	0.984	0.92	0.090
	pMOM	**0.993**	0.993	**1.000**	**0.992**	**0.95**	0.088
	piMOM	0.968	**1.000**	**1.000**	0.982	0.89	0.088
	SS	0.890	**1.000**	0.999	0.940	0.59	**0.087**
	EBLasso	0.668	**1.000**	0.996	0.805	0.21	0.242
	Lasso	0.126	**1.000**	0.945	0.334	0.00	0.103
	SCAD	0.171	**1.000**	0.966	0.399	0.00	0.093
	MCP	0.447	**1.000**	0.990	0.649	0.04	0.092

Table 6: The summary statistics for Case 2 (correlated design) when $p = 500$.

tively chosen hyperparameters in Tables 1–6, the whole simulation results for each hyperparameter value are presented in the Supplementary Material.

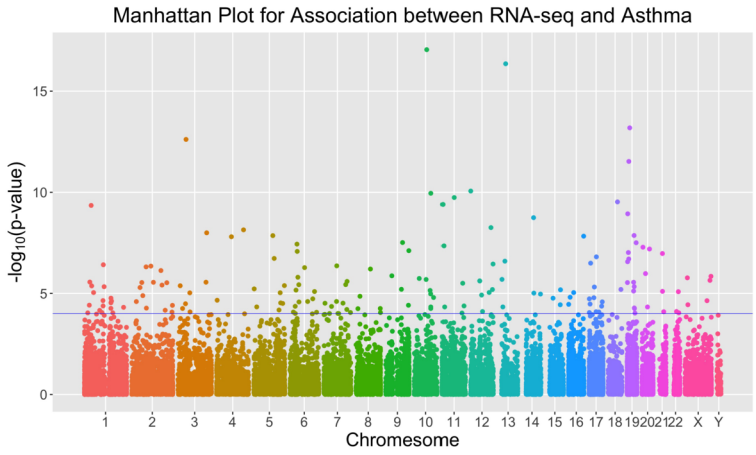


Figure 4: Manhattan plot for association between RNA-seq and asthma. The points beyond the blue line represents the 180 selected genes with $p < 10^{-4}$ based on the DEG analysis.

6 Application to the Analysis of Differentially Expressed Genes

Asthma has been recognized as a systemic disease consisting of networks of genes showing inflammatory changes involving a broad spectrum of adaptive and innate immune systems. Utilizing measurable characteristics of asthmatic patients, including biologic gene expression markers, can help to identify phenotypic categories in asthma. Identification of these phenotypes may help develop strategies for preventing progression of disease severity (Carr and Bleecker, 2016). We aim to apply the proposed variable selection method to develop an RNA-seq-based risk score for asthma stratification.

To construct the risk score, gene expression analysis is performed using an asthma RNA-seq dataset GSE146046 in the Gene Expression Omnibus (GEO) database (Seumois et al., 2020). There are 95 individuals in the GSE146046 dataset including 51 asthmatic subjects and 44 non-asthmatic subjects. The gene expression levels of all the 95 individuals are first randomly split into 2/3 as training and 1/3 as test data while maintaining the same ratio between asthma and control groups. Next we conduct the analysis of differentially expressed genes (DEG) based on the training set and construct data tables containing raw count values for approximately 20,000 unique genes, with genes in rows and sample GEO accession numbers in columns. `DESeq2` R package is used to store the read counts and the intermediate estimated quantities during statistical analysis (Love et al., 2014). We extract summary statistics including p-values for all genes and retain a total of 180 DEGs with p-values less than 10^{-4} visualized in a Manhattan plot (Figure 4). The proposed method and other contenders are applied to the resulting dataset with $p = 180$. The hyperparameters for all the methods are set as in the simulation studies.

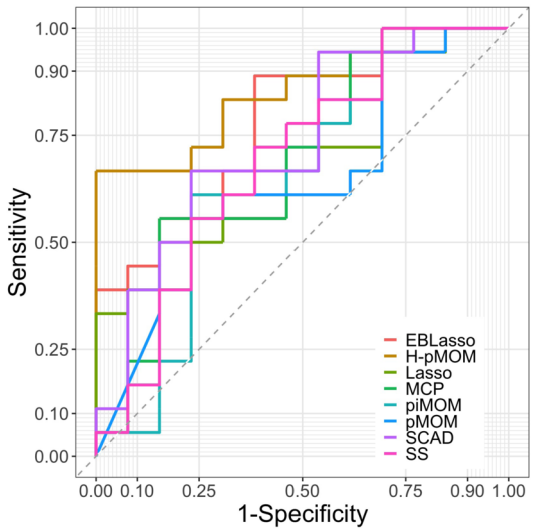


Figure 5: Receiver operating characteristic (ROC) curves comparison between different methods. X-axis: false positive rate (1 - specificity). Y-axis: true positive rate (sensitivity).

	Precision	**Sensitivity**	**Specificity**	**MCC**	**MSPE**
H-pMOM	0.765	**0.722**	0.692	**0.411**	0.277
pMOM	0.769	0.556	0.769	0.325	0.357
piMOM	0.667	0.333	0.769	0.111	0.374
SS	0.733	0.611	0.692	0.300	0.294
EBLasso	**1**	0	**1**	0	0.405
Lasso	0.786	0.611	0.769	0.377	**0.211**
SCAD	0.769	0.556	0.769	0.325	0.260
MCP	0.769	0.556	0.769	0.325	0.321

Table 7: The summary statistics for prediction performance in the testing set.

In Figure 5, we draw the ROC curves for all the methods. The results are further summarized in Table 7 where a common cutoff value 0.5 is adopted for thresholding prediction. From Table 7 and Figure 5, we can tell that the hyper-pMOM prior has overall better prediction performance compared with other methods. Of the retained DEGs, eight genes, namely, TRIM26, MTRNR2L6, DCLRE1B, MRPL45, PSMB8, CBLN3, RPP21 and CSNK2B, are selected by the proposed method. These identified genes seem plausible and have been established in the asthma GWAS catalog (Schoettler et al., 2019; Fodil et al., 2016), which may help better understand the omics architecture that drives complex diseases.

7 Discussion

In this paper, we consider the hyper-pMOM prior and investigate asymptotic properties of the resulting posterior distribution. The hyper-pMOM prior still has the hyperparameters, ψ_1 and ψ_2, so a cross-validation-based selection approach for ψ_2 with a fixed $\psi_1 = 1$ is proposed to alleviate the hyperparameter choice problem. Although it performed reasonably well in our numerical studies, studying the theoretical properties of the posterior based on the adaptively chosen hyperparameters will be a challenging but important task for the future work.

Furthermore, as mentioned in Section 3, deriving strong model selection consistency in a broader class of GLMs is an interesting future research direction. Note that, in this work, we focus on logistic regression models when proving strong model selection consistency of the posterior. An extension to general GLMs might require more conditions on the design matrix based on the current techniques used in the proof, due to more complicated structure of the Hessian matrix for other GLMs compared with that for the logistic regression model.

Another possible extension of our research is to adapt the approximate Laplace approximation (ALA) (Rossell et al., 2021) to estimate the marginal likelihood (2.6). Because the Laplace approximation is computationally expensive and does not consistently estimate the marginal likelihood for models that are supersets of the true model (Rossell and Telesca, 2017), replacing the Laplace approximation with the ALA may improve the performance of the proposed variable selection method.

Supplementary Material

Supplementary to "Bayesian inference on hierarchical nonlocal priors in generalized linear models" (DOI: 10.1214/22-BA1350SUPP; .pdf). We present the proofs for the main results and other auxiliary results.

References

Bian, Y. and Wu, H.-H. (2017). "A Note on Nonlocal Prior Method." arXiv:1702.07778. 100, 108, 109

Breheny, P. and Huang, J. (2011). "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection." *Ann. Appl. Stat.*, 5(1): 232–253. MR2810396. doi: https://doi.org/10.1214/10-AOAS388. 113

Cai, X., Huang, A., and Xu, S. (2011). "Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping." *BMC Bioinformatics*, 12(211). 113

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. MR4050884. doi: https://doi.org/10.1214/19-BA1154. 100, 102, 103, 104, 107

Cao, X. and Lee, K. (2020). "Variable Selection Using Nonlocal Priors in High-

Dimensional Generalized Linear Models With Application to fMRI Data Analysis." *Entropy*, 22(8). MR4220303. doi: https://doi.org/10.3390/e22080807. 100, 113

Cao, X. and Lee, K. (2022). "Supplementary to "Bayesian inference on hierarchical nonlocal priors in generalized linear models"." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1350SUPP. 101

Carr, T. F. and Bleecker, E. (2016). "Asthma heterogeneity and severity." *The World Allergy Organization journal*, 9(1): 41–41. 117

Castillo, I., Schmidt-Hieber, J., Van der Vaart, A., et al. (2015). "Bayesian linear regression with sparse priors." *The Annals of Statistics*, 43(5): 1986–2018. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 104

Fan, J. and Li, R. (2001). "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties." *Journal of the American Statistical Association*, 96(456): 1348–1360. MR1946581. doi: https://doi.org/10.1198/016214501753382273. 99

Fodil, N., Langlais, D., and Gros, P. (2016). "Primary Immunodeficiencies and Inflammatory Disease: A Growing Genetic Intersection." *Trends in Immunology*, 37(2): 126–140. 118

Friedman, J., Hastie, T., and Tibshirani, R. (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of statistical software*, 33(1): 1–22. 113

George, E. I. and McCulloch, R. E. (1993). "Variable Selection via Gibbs Sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 99

Griffin, J., Łatuszyński, K., and Steel, M. (2021). "In search of lost mixing time: adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large p." *Biometrika*, 108(1): 53–69. MR4226189. doi: https://doi.org/10.1093/biomet/asaa055. 112

Hans, C., Dobra, A., and West, M. (2007). "Shotgun Stochastic Search for "Large p" Regression." *Journal of the American Statistical Association*, 102(478): 507–516. MR2370849. doi: https://doi.org/10.1198/016214507000000121. 110

Ishwaran, H. and Rao, J. S. (2005). "Spike and slab variable selection: frequentist and Bayesian strategies." *The Annals of Statistics*, 33(2): 730–773. MR2163158. doi: https://doi.org/10.1214/009053604000001147. 99, 107

Johnson, V. and Rossell, D. (2010). "On the Use of Non-Local Prior Densities in Bayesian Hypothesis Tests Hypothesis." *J. R. Statist. Soc. B*, 72: 143–170. MR2830762. doi: https://doi.org/10.1111/j.1467-9868.2009.00730.x. 100

Johnson, V. E. and Rossell, D. (2012). "Bayesian model selection in high-dimensional settings." *Journal of the American Statistical Association*, 107(498): 649–660. MR2980074. doi: https://doi.org/10.1080/01621459.2012.682536. 100, 107, 109, 113

Lee, K., Lee, J. L., and Lin, L. (2019). "Minimax Posterior Convergence Rates and Model Selection Consistency in High-dimensional DAG Models based on Sparse Cholesky

Factors." *The Annals of Statistics*, 47(6): 3413–3437. MR4025747. doi: https://doi.org/10.1214/18-AOS1783. 107

Liang, F., Paulo, R., Molina, G., Clyde, A. M., and Berger, O. J. (2008). "Mixtures of g Priors for Bayesian Variable Selection." *J. Amer. Statist. Assoc*, 103: 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 100, 102

Liang, F., Song, Q., and Yu, K. (2013). "Bayesian subset modeling for high-dimensional generalized linear models." *Journal of the American Statistical Association*, 108(502): 589–606. MR3174644. doi: https://doi.org/10.1080/01621459.2012.761942. 100, 107

Love, M. I., Huber, W., and Anders, S. (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology*, 15(12): 550. 117

Matthews, B. W. (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme." *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2): 442–451. 113

Narisetty, N. N. and He, X. (2014). "Bayesian variable selection with shrinking and diffusing priors." *The Annals of Statistics*, 42(2): 789–817. MR3210987. doi: https://doi.org/10.1214/14-AOS1207. 99, 104

Narisetty, N. N., Shen, J., and He, X. (2019). "Skinny Gibbs: A Consistent and Scalable Gibbs Sampler for Model Selection." *Journal of the American Statistical Association*, 114(527): 1205–1217. MR4011773. doi: https://doi.org/10.1080/01621459.2018.1482754. 103, 107

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. 113

Rossell, D. (2022). "Concentration of Posterior Model Probabilities and Normalized L_0 Criteria." *Bayesian Analysis*, 17(2): 565 – 591. MR4483231. doi: https://doi.org/10.1214/21-ba1262. 103, 104

Rossell, D., Abril, O., and Bhattacharya, A. (2021). "Approximate Laplace approximations for scalable model selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4): 853–879. MR4320004. 109, 119

Rossell, D. and Telesca, D. (2017). "Nonlocal Priors for High-Dimensional Estimation." *Journal of the American Statistical Association*, 112(517): 254–265. MR3646569. doi: https://doi.org/10.1080/01621459.2015.1130634. 109, 119

Rossell, D., Telesca, D., and Johnson, V. E. (2013). "High-Dimensional Bayesian Classifiers Using Non-Local Priors." In *Statistical Models for Data Analysis*. Heidelberg: Springer International Publishing. 100

Schoettler, N., Rodríguez, E., Weidinger, S., and Ober, C. (2019). "Advances in asthma and allergic disease genetics: Is bigger always better?" *Journal of Allergy and Clinical Immunology*, 144(6): 1495–1506. 118

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 38(5): 2587 – 2619. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 103

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): eaba6087. 117

Shi, G., Lim, C. Y., and Maiti, T. (2019). "Bayesian model selection for generalized linear models using non-local priors." *Computational Statistics & Data Analysis*, 133: 285 – 296. MR3926481. doi: https://doi.org/10.1016/j.csda.2018.10.007. 100, 108, 109

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. 100, 101, 103, 104, 107, 108, 109, 110, 111, 112

Song, Q. and Liang, F. (2017). "Nearly optimal Bayesian shrinkage for high dimensional regression." *arXiv preprint* arXiv:1712.08964. 107

Tibshirani, R. (1996). "Regression Shrinkage and Selection Via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288. MR1379242. 99

Tüchler, R. (2008). "Bayesian Variable Selection for Logistic Models Using Auxiliary Mixture Sampling." *Journal of Computational and Graphical Statistics*, 17(1): 76–94. MR2424796. doi: https://doi.org/10.1198/106186008X289849. 113

Wu, H.-H., Ferreira, M. A., Elkhouly, M., and Ji, T. (2020). "Hyper nonlocal priors for variable selection in generalized linear models." *Sankhya A*, 82(1): 147–185. MR4155019. doi: https://doi.org/10.1007/s13171-018-0151-9. 100, 102, 108

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. MR3576552. doi: https://doi.org/10.1214/15-AOS1417. 107

Zhang, C.-H. (2010). "Nearly unbiased variable selection under minimax concave penalty." *The Annals of statistics*, 38(2): 894–942. MR2604701. doi: https://doi.org/10.1214/09-AOS729. 99, 113

Acknowledgments

We are grateful to the Associate Editor and reviewer for their valuable comments which have significantly improved the quality of presentation and technical content of our paper.