Bayesian Analysis (2020) **15**, Number 1, pp. 241–262

High-Dimensional Posterior Consistency for Hierarchical Non-Local Priors in Regression

Xuan Cao[*], Kshitij Khare[†], and Malay Ghosh[‡]

Abstract. The choice of tuning parameters in Bayesian variable selection is a critical problem in modern statistics. In particular, for Bayesian linear regression with non-local priors, the scale parameter in the non-local prior density is an important tuning parameter which reflects the dispersion of the non-local prior density around zero, and implicitly determines the size of the regression coefficients that will be shrunk to zero. Current approaches treat the scale parameter as given, and suggest choices based on prior coverage/asymptotic considerations. In this paper, we consider the fully Bayesian approach introduced in (Wu, 2016) with the pMOM non-local prior and an appropriate Inverse-Gamma prior on the tuning parameter to analyze the underlying theoretical property. Under standard regularity assumptions, we establish strong model selection consistency in a high-dimensional setting, where p is allowed to increase at a polynomial rate with n or even at a sub-exponential rate with n. Through simulation studies, we demonstrate that our model selection procedure can outperform other Bayesian methods which treat the scale parameter as given, and commonly used penalized likelihood methods, in a range of simulation settings.

Keywords: posterior consistency, high-dimensional data, non-local prior, model selection, multivariate regression.

1 Introduction

The literature on Bayesian variable selection in linear regression is vast and rich. Many priors and methods have been proposed. George and McCulloch (1993) propose the stochastic search variable selection which uses the Gaussian distribution with a zero mean and a small but fixed variance as the spike prior, and another Gaussian distribution with a large variance as the slab prior. Ishwaran, Kogalur, and Rao (2005) also use Gaussian spike and slab priors, but with continuous bimodal priors for the variance of the regression coefficient to alleviate the difficulty of choosing specific prior parameters. Narisetty and He (2014) introduce shrinking and diffusing priors as spike and slab priors, and establish model selection consistency of the approach in a high-dimensional setting. g-prior is introduced in (Zellner, 1986), and Liang et al. (2008) further propose the mixture of g priors based variable selection method and establish selection consistency. In recent years, the use of non-local priors in this context has generated a lot of interest.

Non-local priors were first introduced by Johnson and Rossell (2010) as densities that are identically zero whenever a model parameter is equal to its null value in the context

[*] Department of Mathematical Sciences, University of Cincinnati, xuan.cao@uc.edu
[†] Department of Statistics, University of Florida, kdkhare@stat.ufl.edu
[‡] Department of Statistics, University of Florida, kdkhare@stat.ufl.edu

https://doi.org/10.1214/19-BA1154

of hypothesis testing. Compared to local priors, which still preserve positive values at null parameter values, non-local prior distributions have relatively appealing properties for Bayesian model selection. In particular, non-local priors discard spurious covariates faster as the sample size n grows, while preserving exponential learning rates to detect non-zero coefficients as indicated in (Johnson and Rossell, 2010). These priors were further extended to Bayesian model selection problems in (Johnson and Rossell, 2012) by imposing non-local prior densities on a vector of regression coefficients. Posterior distributions on the model space based on non-local priors were found to be more tightly concentrated around the maximum a posteriori (MAP) model than the posterior based on for example, g-priors, which tend to be more dispersed, implying that these non-local priors yield a faster rate of posterior concentration, as indicated in (Shin et al., 2018).

In particular, let $\boldsymbol{y}_n$ denote a random vector of responses, X_n an $n \times p$ design matrix of covariates, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ a $p \times 1$ vector of regression coefficients. Under the linear regression model,

$$\boldsymbol{y}_n \sim N\left(X_n\boldsymbol{\beta}, \sigma^2 I_n\right).$$

In (Johnson and Rossell, 2012), the authors introduce the product moment (pMOM) non-local prior with density

$$d_p(2\pi)^{-\frac{p}{2}}(\tau\sigma^2)^{-rp-\frac{p}{2}}|A_p|^{\frac{1}{2}} \exp\left\{-\frac{\boldsymbol{\beta}_p' A_p \boldsymbol{\beta}}{2\tau\sigma^2}\right\}\prod_{i=1}^{p}\beta_i^{2r}. \tag{1.1}$$

Here A_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 τ and σ^2. Variations of the density in (1.1), called the piMOM and peMOM density, have also been developed in (Johnson and Rossell, 2012; Rossell et al., 2013). Clearly, the density in (1.1) is zero when any component of $\boldsymbol{\beta}$ is zero. Under appropriate regularity conditions, the authors in (Johnson and Rossell, 2012; Shin et al., 2018) demonstrate that in high-dimensional settings, model selection procedures based on the pMOM and piMOM non-local prior densities can achieve strong model selection consistency, i.e., the posterior probability of the true model converges to 1 as the sample size n increases.

As noted in (Johnson and Rossell, 2012), the scale parameter τ is of particular importance, as it reflects the dispersion of the non-local prior density around zero, and implicitly determines the size of the regression coefficients that will be shrunk to zero. Johnson and Rossell (2010, 2012) treat τ as given and suggest a choice of τ which leads to a high prior probability for significant values of the regression coefficients. Shin et al. (2018) again treat τ as given, and consider a setting where p and τ vary with the sample size n. They show that high-dimensional model selection consistency is achieved under the peMOM prior (another variation of the priors above introduced in (Rossell et al., 2013)), as long as τ is of a larger order than $\log p$ and smaller order than n.

In the context of generalized linear model, similar to the development from g prior in (Zellner, 1986) to the mixture of g prior in (Liang et al., 2008), Wu (2016) further extends the work in (Johnson and Rossell, 2012; Shin et al., 2018) by proposing a fully Bayesian approach with the pMOM non-local prior and an appropriate Inverse-Gamma prior on the parameter τ referred to as the hyper-pMOM prior, following the nomenclature in (Wu, 2016). In particular, Wu (2016) discusses the potential advantages of using hyper-pMOM priors and establish Bayes factor rates.

The primary goal and innovation of this paper is to investigate the underlying model selection consistency for the hyper-pMOM priors in linear regression setting. The extra prior layer of prior, however, creates technical challenges for a high-dimensional theoretical consistency analysis. Under standard regularity assumptions, which include the prior over all models is restricted to ones with model size less than an appropriate function of the sample size n, we establish *posterior ratio consistency* (Theorem 3.1), i.e., the ratio of the maximum marginal posterior probability assigned to a "non-true" model to the posterior probability assigned to the "true" model converges to zero in probability. In particular, this implies that the true model will be the mode of the posterior distribution with probability tending to 1 as $n \to \infty$.

Next, under the additional assumption that p increases at a polynomial rate with n, we show *strong model selection consistency* (Theorem 3.2). Strong model selection consistency implies that the posterior probability of the true model converges in probability to 1 as $n \to \infty$. The assumption of restricting the prior over models with appropriately bounded parameter size, i.e., putting zero prior mass on unrealistically large models) has been used in both (Narisetty and He, 2014) and (Shin et al., 2018) for regression models. Based on reviewers' comments, we relax the polynomial rate restriction on p to a sub-exponential rate by replacing the uniform type prior with a complexity prior on the model space to penalize larger models and establish model selection consistency under the complexity prior in Theorem 5.2.

For the hyper-piMOM priors, Bian and Wu (2017) establish model selection consistency in the framework of generalized linear model. While there are some connections between our model and the one in (Bian and Wu, 2017), there are fundamental differences between the two models and the corresponding analyses. A detailed explanation of this is provided in Remark 1.

The rest of the paper is structured as follows. In Section 2 we provide our hierarchical fully Bayesian model. Model selection consistency results are stated in Section 3, and the proofs are provided in Section 4. Section 5 establishes the model selection consistency under the complexity prior. Details about how to approximate the posterior density for model selection are demonstrated in Section 6. In Section 7 and Section 8, via simulation studies and real data analysis, we illustrate the model selection consistency result, and demonstrate the benefits of model selection using the fully Bayesian approach as compared to approaches which treat τ as given, and existing penalized likelihood approaches. We end our paper with a discussion in Section 9.

2 Model specification

We start by considering the standard Gaussian linear regression model with p coefficients and by introducing some required notation. Let $\boldsymbol{y}_n$ denote a random vector of responses, X_n an $n \times p$ design matrix of covariates, and $\boldsymbol{\beta}$ a $p \times 1$ vector of regression coefficients. Our goal is variable selection, i.e., to correctly identify all the non-zero regression coefficients.

In light of that, we denote a model by $\boldsymbol{k} = \{k_1, k_2, \ldots, k_m\}$ if and only if all the non-zero elements of $\boldsymbol{\beta}$ are $\beta_{k_1}, \beta_{k_2}, \ldots, \beta_{k_m}$ and denote $\boldsymbol{\beta}_k = \left(\beta_{k_1}, \beta_{k_2}, \ldots, \beta_{k_m}\right)^T$. For any $p \times p$ matrix A, let A_k represent the submatrix formed from the columns of A corresponding to model $\boldsymbol{k}$. In particular, Let X_k denote the design matrix formed from the columns of X_n corresponding to model $\boldsymbol{k}$. For the rest of the paper, simply let $k = |\boldsymbol{k}|$ represent the cardinality of model $\boldsymbol{k}$ for notational convenience.

The class of pMOM densities (1.1) can be used for model selection through the following hierarchical model.

$$\boldsymbol{Y}_n \mid \boldsymbol{\beta}_k, \sigma^2, \boldsymbol{k} \sim N(X_k \boldsymbol{\beta}_k, \sigma^2 I_n), \tag{2.1}$$

$$\pi\left(\boldsymbol{\beta}_k \mid \tau, \sigma^2, \boldsymbol{k}\right) = d_k (2\pi)^{-\frac{k}{2}} (\tau\sigma^2)^{-rk-\frac{k}{2}} |A_k|^{\frac{1}{2}} \exp\left\{-\frac{\boldsymbol{\beta}_k' A_k \boldsymbol{\beta}_k}{2\tau\sigma^2}\right\} \prod_{i=1}^{k} \beta_{k_i}^{2r}, \tag{2.2}$$

$$\pi(\tau) = \frac{\left(\frac{n}{2}\right)^{\frac{1}{2}}}{\Gamma(\frac{1}{2})} \tau^{-\frac{3}{2}} e^{-\frac{n}{2\tau}}, \tag{2.3}$$

$$\pi\left(\sigma^2\right) = \frac{(\alpha_2)^{\alpha_1}}{\Gamma(\alpha_1)} \left(\sigma^2\right)^{-(\alpha_1+1)} e^{-\frac{\alpha_2}{\sigma^2}}. \tag{2.4}$$

Note that in the currently presented hierarchical model, no specific form/condition has yet been assigned to the prior over the space of models. Some standard regularity assumptions for this prior will be provided later in Section 3. Following the nomenclature in (Wu, 2016), we refer to the mixture of priors in (2.2) and (2.3) as the hyper-pMOM prior. In particular, one can show the implied marginal density of $\boldsymbol{\beta}_k$ after integrating out τ have the following expression

$$\pi\left(\boldsymbol{\beta}_k \mid \sigma^2, \boldsymbol{k}\right) = \frac{\left(\frac{n}{2}\right)^{\frac{1}{2}}}{\Gamma(\frac{1}{2})} \frac{\Gamma(rk + \frac{k}{2} + \frac{1}{2})}{(\frac{n}{2} + \frac{\boldsymbol{\beta}_k' A_k \boldsymbol{\beta}_k}{2\sigma^2})^{rk+\frac{k}{2}+\frac{1}{2}}} d_k (2\pi)^{-\frac{k}{2}} \sigma^{-2rk-k} |A_k|^{\frac{1}{2}} \prod_{i=1}^{k} \beta_{k_i}^{2r}. \tag{2.5}$$

Note that compared to the pMOM density in (1.1) with given τ, $\pi\left(\boldsymbol{\beta}_k \mid \sigma^2, \boldsymbol{k}\right)$ now possesses thicker tails, which induces prior dependence. See Figure 1 and Figure 2,

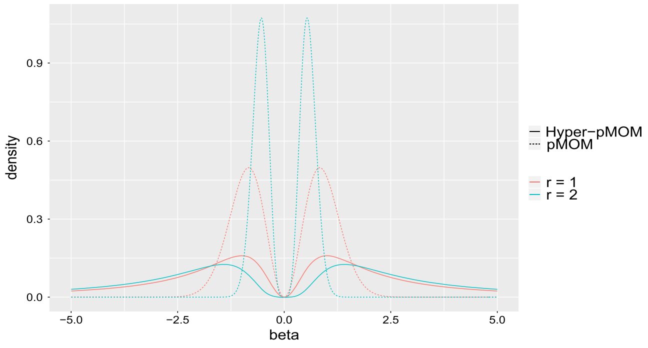
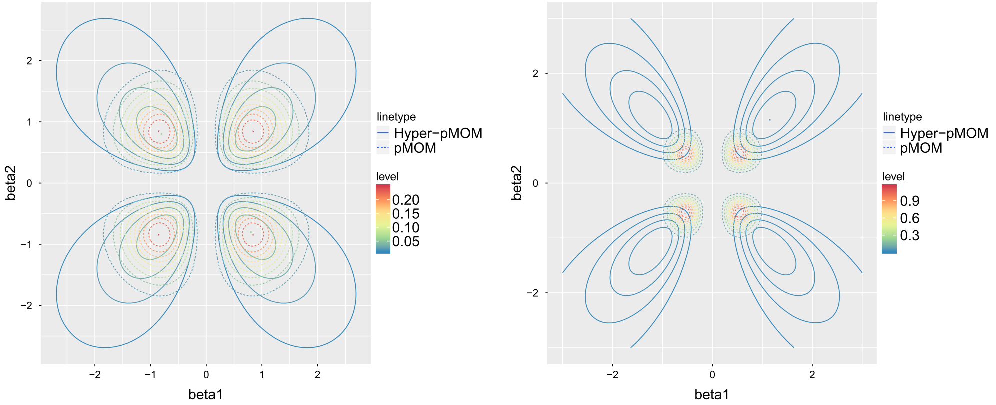


Figure 1: Comparison: Hyper-pMOM and pMOM when $p = 1$. $\tau = 0.072, 0.348$ for $r = 1, 2$ respectively for pMOM.



(a) $r = 1$ (b) $r = 2$

Figure 2: Comparison: Hyper-pMOM and pMOM when $p = 2$. $\tau = 0.072, 0.348$ for $r = 1, 2$ respectively for pMOM.

where we plot the marginal density $\pi\left(\boldsymbol{\beta}_k \mid \sigma^2, \boldsymbol{k}\right)$ when $A_p = 1$, $\sigma^2 = 1$ and $n = 1$ for the univariate and bivariate case, respectively. In addition, the hyper-pMOM prior could achieve better model selection performance especially for small samples. See for example (Liang et al., 2008) that investigates the finite sample performance for hyper-g priors.

By (2.1) and Bayes' rule, the resulting posterior probability for model $\boldsymbol{k}$ is denoted by,

$$\pi(\boldsymbol{k}|\boldsymbol{y}_n) = \frac{\pi(\boldsymbol{k})}{\pi(\boldsymbol{y}_n)} m_{\boldsymbol{k}}(\boldsymbol{y}_n), \tag{2.6}$$

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

$$\begin{aligned}
& m_{\boldsymbol{k}}(\boldsymbol{y}_n) \\
=& \int_0^\infty \int_0^\infty \pi\left(\boldsymbol{y}_n \mid \boldsymbol{\beta}_k, \sigma^2, \boldsymbol{k}\right) \pi\left(\boldsymbol{\beta}_k \mid \tau, \sigma^2, \boldsymbol{k}\right) \pi(\tau) \pi\left(\sigma^2\right) d\boldsymbol{\beta}_k d\sigma^2 d\tau \\
=& \frac{\left(\frac{n}{2}\right)^{\frac{1}{2}}}{\Gamma(\frac{1}{2})} \frac{(\alpha_2)^{\alpha_1}}{\Gamma(\alpha_1)} \int_0^\infty \int_0^\infty d_k (2\pi)^{-\frac{k}{2}} (\tau\sigma^2)^{-rk-\frac{k}{2}} |A_k|^{\frac{1}{2}} \exp\left[-\frac{\boldsymbol{\beta}_k' A_k \boldsymbol{\beta}_k}{2\tau\sigma^2}\right] \prod_{i=1}^{k} \beta_{k_i}^{2r} \\
& \times \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{(\boldsymbol{y}_n - X_k\boldsymbol{\beta}_k)^T(\boldsymbol{y}_n - X_k\boldsymbol{\beta}_k)}{2\sigma^2}\right\} \tau^{-\frac{3}{2}} e^{-\frac{n}{2\tau}} \left(\sigma^2\right)^{-(\alpha_1+1)} e^{-\frac{\alpha_2}{\sigma^2}} d\boldsymbol{\beta}_k d\sigma^2 d\tau \\
=& d_k \frac{\frac{\left(\frac{n}{2}\right)^{\frac{1}{2}}}{\Gamma(\frac{1}{2})}}{(\sqrt{2\pi})^n} \frac{(\alpha_2)^{\alpha_1}}{\Gamma(\alpha_1)} |A_k|^{\frac{1}{2}}
\end{aligned}$$

$$\times \int_0^\infty \int_0^\infty (\sigma^2)^{-\left(\frac{n}{2}+rk+\alpha_1+1\right)} \exp\left\{-\frac{R_k+2\alpha_2}{2\sigma^2}\right\} \tau^{-rk-\frac{k}{2}-\frac{3}{2}} e^{-\frac{n}{2\tau}} \frac{E_k(\prod_{i=1}^k \beta_{k_i}^{2r})}{|C_k|^{\frac{1}{2}}} d\sigma^2 d\tau, \tag{2.7}$$

where $C_k = X_k^T X_k + \frac{A_k}{\tau}, R_{\boldsymbol{k}} = \boldsymbol{y}_n^T(I_n - X_k C_k^{-1} X_k^T)\boldsymbol{y}_n$, and $E_k(.)$ denotes the expectation with respect to a multivariate normal distribution with mean $\tilde{\boldsymbol{\beta}}_k = C_k^{-1} X_k^T \boldsymbol{y}_n$, and covariance matrix $V = \sigma^2 C_k^{-1}$. In particular, these posterior probabilities can be used to select a model by computing the posterior mode defined by

$$\hat{\boldsymbol{k}} = \arg\max_{\boldsymbol{k}} \pi(\boldsymbol{k}|\boldsymbol{y}_n). \tag{2.8}$$

3 Model selection consistency: main results

In this section we will explore the high-dimensional asymptotic properties of the Bayesian model selection approach specified in Section 2. In particular, we consider a setting where the number of regression coefficients $p = p_n$ increases with the sample size n. The true data generating mechanism is given by $Y_n = X_n \boldsymbol{\beta}_0 + \boldsymbol{\epsilon}_n$. Here $\boldsymbol{\beta}_0$ is the true p_n dimensional vector of regression coefficients, whose dependence on n is suppressed for notational convenience, and the entries of $\boldsymbol{\epsilon}_n$ are i.i.d Gaussian with mean zero and variance σ_0^2. As in (Johnson and Rossell, 2012), we assume that the true vector of regression coefficients is sparse, i.e., all the entries of $\boldsymbol{\beta}_0$ are zero except those corresponding to a subset $\boldsymbol{t} \subseteq \{1, 2, \ldots, p_n\}$, and $\boldsymbol{t}, \boldsymbol{\beta}_{0,t}, \sigma_0^2$ do not vary with n. Our results can be easily extended to the case where $|\boldsymbol{t}|$, and entries of $\boldsymbol{\beta}_{0,t}$ and σ_0^2 vary with n but stay bounded. However, we assume these quantities stay fixed for ease of exposition.

For any $p \times p$ symmetric matrix A, let $eig_1(A) \leq eig_2(A) \ldots \leq eig_p(A)$ be the ordered eigenvalues of A and denote the j-th largest nonzero eigenvalue as $\nu_j(A)$. Let $\lambda_k^m = \min_{1 \leq j \leq \min(n,k)} \nu_j\left(\frac{X_k^T X_k}{n}\right)$ and $\lambda_k^M = \max_{1 \leq j \leq \min(n,k)} \nu_j\left(\frac{X_k^T X_k}{n}\right)$, respectively. In order to establish our asymptotic results, we need the following mild regularity assumptions.

Assumption 1. There exist $\epsilon_n < 1$, such that $0 < \epsilon_n \leq \lambda_k^m \leq \lambda_k^M \leq \epsilon_n^{-1}$, where $\epsilon_n^{-1} = O\left((\log n)^{\frac{1}{8}}\right)$.

Assumption 2. $p = O\left(n^\gamma\right)$, where $\gamma < r$.

Assumption 3. $\pi(\boldsymbol{k}) = 0$ for all $|\boldsymbol{k}| > q_n$, where $q_n = O\left(n^\xi\right)$ and $\xi < 1$.

Assumption 4. There exists a constant $\omega > 0$, such that $\frac{\pi(\boldsymbol{t})}{\pi(\boldsymbol{k})} > \omega$ for every $\boldsymbol{k}$ with $\pi(\boldsymbol{k}) > 0$.

Assumption 5. For every $n \geq 1$, the hyper-parameter for the non-local pMOM prior in 2.1 satisfy $0 < a_1 < eig_1(A_p) \leq eig_2(A_p) \leq \ldots \leq eig_p(A_p) < a_2 < \infty$. Here a_1, a_2 are constants not depending on n.

Johnson and Rossell (2012) assume all the eigenvalues of $\frac{X^T X}{n}$ to be bounded by a constant, which is unrealistic to achieve in the high-dimensional setting. In our work, Assumption 1 assumes that non-zero eigenvalues of any sub-matrices of the design matrix not to be bounded by a constant, but to be uniformly bounded over a function of n. Assumption 5 is standard which assumes the prior covariance matrix are uniformly bounded in n. Note that for the default value of $A_p = I_p$, Assumption 5 is immediately satisfied. Assumption 3 states that the prior on the space of the 2^{p_n} possible models, places zero mass on unrealistically large models (identical to Assumption in (Shin et al., 2018)). Assumption 4 states that the ratio of the prior probabilities assigned to the true model and any non-true model stays bounded below in n (identical to Assumption in (Johnson and Rossell, 2012)). This type of priors have also been considered in (Song and Liang, 2015) and (Shin et al., 2018). Assumption 2 states that p can grow at an appropriate polynomial rate with n. In Section 5, we also give the consistency results under the complexity priors on the model space, which penalize larger models, and consequently relax the assumption on the rate at which p can be growing.

We now state and prove the main model selection consistency results. Our first result establishes what we refer to as posterior ratio consistency. This notion of consistency implies that the true model will be the mode of the posterior distribution among all the models with probability tending to 1 as $n \to \infty$.

Theorem 3.1 (Posterior ratio consistency for hyper-pMOM priors). Under Assumptions 1, 3, 4 and 5, for the hierarchical model in (2.1) to (2.4) with hyper-pMOM priors, the following holds:

$$\max_{\boldsymbol{k} \neq \boldsymbol{t}} \frac{\pi(\boldsymbol{k}|\boldsymbol{y}_n)}{\pi(\boldsymbol{t}|\boldsymbol{y}_n)} \to 0, \quad \text{as } n \to \infty.$$

In particular, it implies that the probability that the posterior mode $\hat{\boldsymbol{k}}$ defined in (2.8) is equal to the true model $\boldsymbol{t}$ will converge to 1, i.e.,

$$P(\boldsymbol{t} = \arg\max_{\boldsymbol{k}} \pi(\boldsymbol{k}|\boldsymbol{y}_n)) \to 1, \quad \text{as } n \to \infty.$$

We would like to point out that posterior ratio consistency (Theorems 3.1) does not require any restriction on the number of predictors. This requirement is only needed for strong selection consistency (Theorem 3.2). Next, we establish a stronger result which implies that the posterior mass assigned to the true model $\boldsymbol{t}$ converges to 1 in probability. We refer to this notion of consistency as strong selection consistency.

Theorem 3.2 (Strong selection consistency for hyper-pMOM priors). Under Assumptions 1-5, with $\xi < 1 - \frac{4\gamma}{3r}$ in Assumption 3, for the hierarchical model in (2.1) to (2.4) with hyper-pMOM priors, the following holds:

$$\pi(\boldsymbol{t}|\boldsymbol{y}_n) \to 1, \quad \text{as } n \to \infty.$$

The above results have been established under the pMOM priors. Another class of non-local priors introduced in (Johnson and Rossell, 2012) are the piMOM priors on the regression coefficients, for which the density of the regression coefficients under the model $\boldsymbol{k} = \{k_1, k_2, \ldots, k_m\}$ is given by

$$\frac{(\tau\sigma^2)^{\frac{r|\boldsymbol{k}|}{2}}}{\Gamma(\frac{r}{2})^{|\boldsymbol{k}|}} \prod_{i=1}^{|\boldsymbol{k}|} |\beta_{k_i}|^{-(r+1)} \exp\left(-\frac{\tau\sigma^2}{\beta_{k_i}^2}\right), \tag{3.1}$$

where r is a positive integer and is refereed to as the order of the density. The corollary below establishes strong model selection consistency under the hyer-piMOM priors with piMOM priors on each linear regression coefficient (conditional on the hyper parameter τ) and an Inverse-Gamma prior on τ. The consistency can be obtained immediately by combining Theorem 3.2 with Eq. (59) and (60) in the supplementary material for (Johnson and Rossell, 2012).

Corollary 3.1 (Strong selection consistency for hyper-piMOM priors). Under the same conditions as in Theorem 3.2, when piMOM priors are imposed on $\boldsymbol{\beta}_k$ in model (2.2), the following holds:

$$\pi(\boldsymbol{t}|\boldsymbol{y}_n) \to 1, \text{ as } n \to \infty.$$

Remark 1. In the context of generalized linear regression, Bian and Wu (2017) consider the hierarchical Bayesian model with the following hyer-piMOM priors on regression coefficients.

$$\boldsymbol{\beta}_k \mid \tau_i \sim \frac{(\tau\sigma^2)^{\frac{r|\boldsymbol{k}|}{2}}}{\Gamma(\frac{r}{2})^{|\boldsymbol{k}|}} \prod_{i=1}^{|\boldsymbol{k}|} |\beta_{k_i}|^{-(r+1)} \exp\left(-\frac{\tau_i\sigma^2}{\beta_{k_i}^2}\right)$$
$$\tau_i \sim \text{Inverse-Gamma}\ (\frac{(r+1)}{2}, \lambda).$$

In particular, the authors put an independent piMOM prior on each linear regression coefficient (conditional on the hyper parameter τ_i), and an Inverse-Gamma prior on τ_i. In this setting, Bian and Wu (2017) establish strong selection consistency for the regression coefficients (assuming the prior is constrained to leave out unrealistically large models). There are similarities between the models and the consistency analysis in (Bian and Wu, 2017) and our work as in the usage of non-local priors and Inverse-Gamma distribution. However, despite these similarities, there are some fundamental differences in the two models and the 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 $\boldsymbol{\beta}_k$. In particular, pMOM distributions introduce correlations among the entries in $\boldsymbol{\beta}_k$ and creates more theoretical challenges. Because of the correlation introduced, some properties like detecting small coefficients are not apparent in our case. Also, the pMOM prior imposes exact sparsity in $\boldsymbol{\beta}_k$, which is not the case in (Bian and Wu, 2017). Hence it is structurally different than the prior in (Bian and Wu, 2017). Secondly, in order to prove consistency results, Bian and Wu (2017) assume the product of the response variables and the entries of design matrix are bounded by a constant. The former assumption is rarely seen in the literatures and the latter assumption can be problematic in practice. See Assumption C1 in (Bian and Wu, 2017). In addition, Assumption C2 in (Bian and Wu, 2017) states the lower bound for the true regression coefficients, which is not required in our analysis. Thirdly, in terms of proving posterior consistency, we bound the ratio of posterior probabilities for a non-true model and the true model by a 'prior term' which results from the Inverse-Gamma prior on τ, and a 'data term'. The consistency proof is then a careful exercise in balancing these two terms against each other on a case-by-case basis, while Bian and Wu (2017) directly follow the proof in (Shin et al., 2018) and requires additional assumptions on the Hessian matrix.

4 Proof of Theorems 3.1 and 3.2

The proof of Theorems 3.1 and 3.2 will be broken up into several steps. First we denote for any model $\boldsymbol{k}$, $R_k^* = \boldsymbol{y}_n^T \left(I - X_k(X_k^T X_k)^{-1} X_k^T\right) \boldsymbol{y}_n$, and $P_k = X_k(X_k^T X_k)^{-1} X_k^T$. Our method of proving consistency involves approximating R_t and R_k (in (2.7)) with R_t^* and R_k^* respectively. Fix a model $\boldsymbol{k} \neq \boldsymbol{t}$ arbitrarily, and let $\boldsymbol{u} = \boldsymbol{k} \cup \boldsymbol{t}$ and $u = |\boldsymbol{u}|$ be the cardinality of $\boldsymbol{u}$. Note that $\frac{R_t^*}{\sigma_0^2} \sim \chi^2_{n-t}$, $\frac{R_u^*}{\sigma_0^2} \sim \chi^2_{n-u}$, $R^*_{u \cap t^c} \sim \chi^2_{u-t}$, and $\frac{\boldsymbol{y}_n^T P_u \boldsymbol{y}_n}{\sigma_0^2} \sim \chi^2_u \left(\frac{\boldsymbol{\beta}_0^T X_t^T X_t \boldsymbol{\beta}_0}{\sigma_0^2}\right)$. Next, we establish two tail probability bounds for the χ^2 distribution and the non-central χ^2 distribution respectively, which will be useful in our analysis.

Lemma 4.1. For any $a > 0$, we have the following two inequalities,

$$P\left(|\chi_p^2 - p| > a\right) \leq 2\exp\left(-\frac{a^2}{4p}\right), \tag{4.1}$$

$$P\left(\chi_p^2(\lambda) - (p+\lambda) > a\right) \leq \exp\left(-\frac{p}{2}\left\{\frac{a}{p+\lambda} - \log\left(1 + \frac{a}{p+\lambda}\right)\right\}\right). \tag{4.2}$$

The proof for Lemma 4.1 is provided in the supplemental document (Cao et al., 2019). The following result is immediate from Lemma 4.1.

$$\begin{aligned} P\left[\left|\frac{R_t^*}{\sigma_0^2} - (n-t)\right| > \sqrt{n-t}\log n\right] &\leq P\left[\left|\frac{R_t^*}{\sigma_0^2} - (n-t)\right| > 4\sqrt{(n-t)\log n}\right] \\ &\leq 2n^{-1} \to 0, \end{aligned} \tag{4.3}$$

as $n \to \infty$. Similarly, we have

$$P\left[\left|\frac{R_u^*}{\sigma_0^2} - (n-u)\right| > \sqrt{n-u}\log n\right] \leq 2n^{-1} \to 0, \tag{4.4}$$

and

$$P\left[\left|\frac{R^*_{u\cap t^c}}{\sigma_0^2} - (u-t)\right| > \sqrt{u-t}\log n\right] \leq 2n^{-1} \to 0, \tag{4.5}$$

as $n \to \infty$. Next, by Lemma 4.1, since $\boldsymbol{u} \supset \boldsymbol{t}$, we have

$$\begin{aligned} &P\left[\frac{\boldsymbol{y}_n^T P_u \boldsymbol{y}_n}{\sigma_0^2} - \left(u + \frac{1}{\sigma_0^2}\boldsymbol{\beta}_0^T X_t^T X_t \boldsymbol{\beta}_0\right) > n\log n - u - \frac{1}{\sigma_0^2}\boldsymbol{\beta}_0^T X_t^T X_t \boldsymbol{\beta}_0\right] \\ &\leq \exp\left\{-\frac{u}{2}\left\{\frac{n\log n}{u + \frac{1}{\sigma_0^2}\boldsymbol{\beta}_0^T X_t^T X_t \boldsymbol{\beta}_0} - \log\left(1 + \frac{n\log n}{u + \frac{1}{\sigma_0^2}\boldsymbol{\beta}_0^T X_t^T X_t \boldsymbol{\beta}_0}\right)\right\}\right\} \\ &\leq \exp\left\{-\frac{u}{4}\left\{\frac{\log n}{1 + \frac{1}{\sigma_0^2 \epsilon_n}\boldsymbol{\beta}_0^T \boldsymbol{\beta}_0}\right\}\right\}. \\ &\preceq n^{-c'u} \to 0, \end{aligned} \tag{4.6}$$

as $n \to \infty$, for some constant $c' > 0$. Define the event C_n as

$$\begin{aligned} C_n = &\left\{ \left| \frac{R_t^*}{\sigma_0^2} - (n-t) \right| > \sqrt{n-t}\log n \right\} \cup \left\{ \left| \frac{R_u^*}{\sigma_0^2} - (n-u) \right| > \sqrt{n-u}\log n \right\} \\ &\cup \left\{ \left| \frac{R_{u\cap t^c}^*}{\sigma_0^2} - (u-t) \right| > \sqrt{u-t}\log n \right\} \cup \left\{ \frac{\boldsymbol{y}_n^T P_u \boldsymbol{y}_n}{\sigma_0^2} > n\log n \right\}, \end{aligned} \tag{4.7}$$

It follows from (4.3), (4.4), (4.5), and (4.6), that $P(C_n) \to 0$ as $n \to \infty$.

We now analyze the behavior of the posterior ratio under different scenarios in a sequence of lemmas. Recall that our goal is to find an upper bound for the posterior ratio, such that the upper bound converges to 0 as $n \to \infty$. **For the following lemmas, we will restrict ourselves to the event C_n^c.** The following lemma establishes the upper bound of the marginal posterior ratio for any "non-true" model compared to the true model.

Lemma 4.2. Under Assumption 1 and Assumption 5, for all $\boldsymbol{k} \neq \boldsymbol{t}$, there exists N (not depending on k), such that when $n > N$,

$$\begin{aligned} \frac{m_{\boldsymbol{k}}(\boldsymbol{y}_n)}{m_{\boldsymbol{t}}(\boldsymbol{y}_n)} < &BA^k \left(\frac{V}{\epsilon_n^2}\right)^{rk} k^k n^{-(k-t)} \frac{\{R_t^* + 2\alpha_2\}^{\frac{n}{2}+rt+\alpha_1}}{\{R_k^* + 2\alpha_2\}^{\frac{n}{2}+rk+\alpha_1}} \\ &+ BA^k k^{(r+1)k} n^{-(r+1)(k-t)-\frac{3}{4}rk-rt} \frac{\{R_t^* + 2\alpha_2\}^{\frac{n}{2}+rt+\alpha_1}}{\{R_k^* + 2\alpha_2\}^{\frac{n}{2}+\alpha_1}}, \end{aligned} \tag{4.8}$$

where A, B are constants and $V = \epsilon_n^{-4}\hat{\beta}_u^T\hat{\beta}_u$, in which $\hat{\beta}_u^T = (X_u^T X_u)^{-1} X_u \boldsymbol{y}_n$ with $\boldsymbol{u} = \boldsymbol{k} \cup \boldsymbol{t}$.

The proof for Lemma 4.2 is provided in supplemental document. The next two lemmas provide the upper bound of the marginal posterior ratio for $\boldsymbol{y}_n$ under different cases of the "non-true" model $\boldsymbol{k}$ with proof provided in the supplemental document.

Lemma 4.3. Under Assumptions 1, 3 and 5, for all $\boldsymbol{k} \not\supseteq \boldsymbol{t}$, there exists N, such that when $n > N'$ (not depending on k),

$$\frac{m_{\boldsymbol{k}}(\boldsymbol{y}_n)}{m_{\boldsymbol{t}}(\boldsymbol{y}_n)} < K'(L')^k n^{-\frac{3}{4}rk}, \tag{4.9}$$

where K' and L' are constants.

Lemma 4.4. Under Assumptions 1, 3 and 5, for all $\boldsymbol{k} \supset \boldsymbol{t}$, there exists N'' (not depending on k), such that when $n > N''$,

$$\frac{m_{\boldsymbol{k}}(\boldsymbol{y}_n)}{m_{\boldsymbol{t}}(\boldsymbol{y}_n)} < S'(T')^{k-t} n^{-\min\{\frac{3}{4},1-\xi\}r(k-t)}, \tag{4.10}$$

where S' and T' are constants.

Proof of Theorem 3.1 and 3.2. The proof of Theorem 3.1 will follow immediately from these two lemmas. By Lemma 4.3, if we restrict to C_n^c, for any $\boldsymbol{k} \neq \boldsymbol{t}$, if $\boldsymbol{k} \not\supseteq \boldsymbol{t}$,

$$\frac{m_{\boldsymbol{k}}(\boldsymbol{y}_n)}{m_{\boldsymbol{t}}(\boldsymbol{y}_n)} < K'(L')^k n^{-\frac{3}{4}rk} \to 0, \text{ as } n \to \infty.$$

Otherwise, when $\boldsymbol{k} \supset \boldsymbol{t}$,

$$\frac{m_{\boldsymbol{k}}(\boldsymbol{y}_n)}{m_{\boldsymbol{t}}(\boldsymbol{y}_n)} < S'(T')^{k-t} n^{--\min\left\{\frac{3}{4},1-\xi\right\}r(k-t)} \to 0, \text{ as } n \to \infty.$$

Note that $P\left(C_n^c\right) \to 1$ as $n \to \infty$. Following from (2.6) and Assumption 4, when $\boldsymbol{k} \neq \boldsymbol{t}$, we have

$$\frac{\pi(\boldsymbol{k}|\boldsymbol{y}_n)}{\pi(\boldsymbol{t}|\boldsymbol{y}_n)} \leq \frac{1}{\omega} \frac{m_{\boldsymbol{k}}(\boldsymbol{y}_n)}{m_{\boldsymbol{t}}(\boldsymbol{y}_n)} \to 0, \text{ as } n \to \infty. \tag{4.11}$$

We now move on to the proof of Theorem 3.2. First note that when $\xi < 1 - \frac{4\gamma}{3r}$, we have

$$\min\left\{\frac{3}{4}, 1-\xi\right\} r > \frac{3}{4}(1-\xi) r > \gamma. \tag{4.12}$$

It follows from (4.11) and Assumption 2 that if we restrict to C_n^c, then

$$\begin{aligned}
\frac{1-\pi(\boldsymbol{t}|\boldsymbol{y}_n)}{\pi(\boldsymbol{t}|\boldsymbol{y}_n)} =& \sum_{\boldsymbol{k} \neq \boldsymbol{t}} \frac{\pi(\boldsymbol{k}) m_{\boldsymbol{k}}(\boldsymbol{y}_n)}{\pi(\boldsymbol{t}) m_{\boldsymbol{t}}(\boldsymbol{y}_n)} \\
\leq& \frac{1}{\omega} \sum_{\boldsymbol{k} \not\supset \boldsymbol{t}} \frac{m_{\boldsymbol{k}}(\boldsymbol{y}_n)}{m_{\boldsymbol{t}}(\boldsymbol{y}_n)} + \frac{1}{\omega} \sum_{\boldsymbol{k} \supset \boldsymbol{t}} \frac{m_{\boldsymbol{k}}(\boldsymbol{y}_n)}{m_{\boldsymbol{t}}(\boldsymbol{y}_n)} \\
\leq& \frac{1}{\omega} \sum_{k=1}^{q_n} \binom{p}{k} K'(L')^k p^{-\frac{\frac{3}{4}r}{\gamma}k} \\
&+ \frac{1}{\omega} \sum_{k-t=1}^{q_n-t} \binom{p-t}{k-t} S'(T')^{k-t} p^{-\frac{\min\left\{\frac{3}{4},1-\xi\right\}r}{\gamma}(k-t)}.
\end{aligned}$$

Using $\binom{p}{k} \leq p^k$ and (4.12), we get

$$\frac{1-\pi(\boldsymbol{t}|\boldsymbol{y}_n)}{\pi(\boldsymbol{t}|\boldsymbol{y}_n)} \to 0, \text{ i.e. } \pi(\boldsymbol{t}|\boldsymbol{y}_n) \to 1,$$

as $n \to \infty$. □

5 Results for complexity priors

Note that under our model prior specified in Assumption 4, to achieve strong selection consistency, we are restricting p to grow at a polynomial rate of n (see Assumption 2). To address this limitation, based on reviewers' comments, we investigate the theoretical property under the complexity priors introduced in (Castillo et al., 2015). The hierarchical model with complexity priors placed on the model space, adapted to our notation and framework, can be described as follows:

$$\begin{aligned}
&\boldsymbol{Y}_n \mid \boldsymbol{\beta}_k, \sigma^2, \boldsymbol{k} \sim N(X_k\boldsymbol{\beta}_k, \sigma^2 I_n)\\
&\pi\left(\boldsymbol{\beta}_k \mid \tau, \sigma^2, \boldsymbol{k}\right) = d_k(2\pi)^{-\frac{k}{2}}(\tau\sigma^2)^{-rk-\frac{k}{2}}|A_k|^{\frac{1}{2}} \exp\left\{-\frac{\boldsymbol{\beta}_k' A_k \boldsymbol{\beta}_k}{2\tau\sigma^2}\right\}\prod_{i=1}^{k}\beta_{k_i}^{2r}\\
&\pi(k) \propto c_1^{-k} p^{-c_2 k},\\
&\pi(\tau) = \frac{\left(\frac{n}{2}\right)^{\frac{1}{2}}}{\Gamma(\frac{1}{2})}\tau^{-\frac{3}{2}}e^{-\frac{n}{2\tau}}.\\
&\pi\left(\sigma^2\right) = \frac{(\alpha_2)^{\alpha_1}}{\Gamma(\alpha_1)}\left(\sigma^2\right)^{-(\alpha_1+1)}e^{-\frac{\alpha_2}{\sigma^2}}.
\end{aligned} \tag{5.1}$$

where $c_1, c_2 > 0$ are fixed constants. As indicated in (Castillo et al., 2015), the rate of decrease for $\pi(k)$ reflects the number of models $\binom{p}{k}$ of given size k. Compared to the previous uniform-like prior, these complexity priors are explicitly penalizing larger models to accommodate larger dimensions. In particular, to achieve model selection consistency in this setup, the dimension p can be allowed to grow at a sub-exponential rate of n given in the following condition:

Condition A. There exists a constant $0 < \kappa < 1$, such that $\log p = O(n^{\kappa})$.

The next result establishes the posterior ratio consistency for the complexity prior based approach in (5.1).

Theorem 5.1 (Posterior ratio consistency for complexity priors). Consider the complexity prior based model described in (5.1). Under Assumptions 1, 3, 5 and Condition A, the following holds:

$$\max_{\boldsymbol{k}\neq\boldsymbol{t}} \frac{\pi(\boldsymbol{k}|\boldsymbol{y}_n)}{\pi(\boldsymbol{t}|\boldsymbol{y}_n)} \to 0, \quad \text{as } n \to \infty.$$

Next, we establish the strong selection consistency result which implies that the posterior mass assigned to the true model $\boldsymbol{t}$ converges to 1 in probability.

Theorem 5.2 (Strong selection consistency for complexity priors). Consider the complexity prior based model described in (5.1). Under Assumptions 1, 3, 5 and Condition A, if we future assume $c_2 > 1$, the following holds:

$$\pi(\boldsymbol{t}|\boldsymbol{y}_n) \to 1, \quad \text{as } n \to \infty.$$

The proof for Theorem 5.1 and 5.2 will also be broken into several steps. The following three lemmas establish the upper bound for the marginal posterior ratio between any “non-true” model and the true model.

Lemma 5.3. Under Assumptions 1, 3, 5 and Condition A, when $\boldsymbol{k} \subset \boldsymbol{t}$, for large enough $n > N_1''$ (not depending on k), the following holds:

$$\frac{\pi(\boldsymbol{k}|\boldsymbol{y}_n)}{\pi(\boldsymbol{t}|\boldsymbol{y}_n)} \leq 2M_1 p^{-2c_2 t}, \tag{5.2}$$

for some constants $M_1 > 0$.

Lemma 5.4. Under Assumptions 1, 3, 5 and Condition A, When $\boldsymbol{k} \supset \boldsymbol{t}$, for large enough $n > N''$ (not depending on k), the following holds:

$$\frac{\pi(\boldsymbol{k}|\boldsymbol{y}_n)}{\pi(\boldsymbol{t}|\boldsymbol{y}_n)} \leq c_1^{-(k-t)} p^{-c_2(k-t)}. \tag{5.3}$$

Lemma 5.5. Under Assumptions 1, 3, 5 and Condition A, when $\boldsymbol{k} \not\subseteq \boldsymbol{t}$, $\boldsymbol{k} \not\supseteq \boldsymbol{t}$ and $\boldsymbol{k} \neq \boldsymbol{t}$, denote $\boldsymbol{u} = \boldsymbol{k} \cup \boldsymbol{t}$. for large enough $n > N_3''$ (not depending on k), the following holds:

$$\frac{\pi(\boldsymbol{k}|\boldsymbol{y}_n)}{\pi(\boldsymbol{t}|\boldsymbol{y}_n)} \leq c_3^{-(k-t)} p^{-c_2 k}, \tag{5.4}$$

for some constant $c_3 > 0$.

Proof of Theorem 5.1 and 5.2. Theorem 5.1 immediately follows after Lemma 5.3 to 5.5. We now move on to the proof of Theorem 5.2. It follows from Lemma 5.3 to 5.5 that if we restrict to C_n^c, then

$$\begin{aligned}
\frac{1-\pi(\boldsymbol{t}|\boldsymbol{y}_n)}{\pi(\boldsymbol{t}|\boldsymbol{y}_n)} &= \sum_{\boldsymbol{k}\neq\boldsymbol{t}} \frac{\pi(\boldsymbol{k}|\boldsymbol{y}_n)}{\pi(\boldsymbol{t}|\boldsymbol{y}_n)} \\
&\leq \sum_{k\leq t} \frac{\pi(\boldsymbol{k}|\boldsymbol{y}_n)}{\pi(\boldsymbol{t}|\boldsymbol{y}_n)} + \sum_{k>t,\boldsymbol{k}\supset\boldsymbol{t}} \frac{\pi(\boldsymbol{k}|\boldsymbol{y}_n)}{\pi(\boldsymbol{t}|\boldsymbol{y}_n)} + \sum_{k>t,\boldsymbol{k}\not\supseteq\boldsymbol{t}} \frac{\pi(\boldsymbol{k}|\boldsymbol{y}_n)}{\pi(\boldsymbol{t}|\boldsymbol{y}_n)} \\
&\leq \sum_{k=1}^{t} \binom{p}{k} M_2 p^{-2c_2 t} + \sum_{k-t=1}^{q_n-t} \binom{p-t}{k-t} c_1^{-(k-t)} p^{-c_2(k-t)} \\
&\quad + \sum_{k=1}^{q_n} \binom{p}{k} c_3^{-(k-t)} p^{-c_2 k}.
\end{aligned}$$

By $\binom{p}{k} \leq p^k$ and $c_2 > 1$, we get

$$\frac{1-\pi(\boldsymbol{t}|\boldsymbol{y}_n)}{\pi(\boldsymbol{t}|\boldsymbol{y}_n)} \to 0, \text{ i.e. } \pi(\boldsymbol{t}|\boldsymbol{y}_n) \to 1, \quad \text{as } n \to \infty,$$

which completes our proof for Theorem 5.2. □

Remark 2. Note that though under the complexity priors, the restriction on p is relaxed in terms of proving strong selection consistency, we find that in our simulation studies, the model selection performance under uniform-like prior is much better than that under the complexity priors, hence from a practical point of view, one would still prefer the hyper-pMOM with uniform-like prior over the model space. As indicated in (Shin et al., 2018), since the pMOM priors already induce a strong penalty on the model size, it is no longer necessary to penalize larger models through priors on the model space.

6 Computation

The integral formulation in (2.6) is quite complicated, and hence the posterior probabilities can not be obtained in closed form. Hence, we use Laplace approximation to compute $m_{\boldsymbol{k}}(\boldsymbol{y}_n)$ and $\pi(\boldsymbol{k}|\boldsymbol{y}_n)$. A similar approach to compute posterior probabilities has been used in (Johnson and Rossell, 2012) and (Shin et al., 2018).

Note that for any model $\boldsymbol{k}$, when $A_k = I_k$, the normalization constant d_k in (2.1) is given by $d_k = ((2r-1)!!)^{-k}$. Let

$$\begin{aligned} f(\boldsymbol{\beta},\tau,\sigma^2) =& \log(m_{\boldsymbol{k}}(\boldsymbol{y}_n)) \\ =& \log\left(\pi(\boldsymbol{y}_n|\sigma^2)\pi(\boldsymbol{\beta}_k|\tau,\sigma^2)\pi(\tau)\pi(\sigma^2)\right) \\ =& -k\log\left((2r-1)!!\right) - \frac{n+k}{2}\log(2\pi) - \left(rk + \frac{n+k}{2} + \alpha_1 + 1\right)\log(\sigma^2) \\ & - \left(rk + \frac{k+3}{2}\right)\log\tau - \left(\frac{(\boldsymbol{y}_n - X_k\boldsymbol{\beta}_k)^T(\boldsymbol{y}_n - X_k\boldsymbol{\beta}_k)}{2\sigma^2}\right) \\ & - \left(\frac{\boldsymbol{\beta}_k^T\boldsymbol{\beta}_k}{2\tau\sigma^2} + \frac{\alpha_2}{\sigma^2} + \frac{n}{2\tau}\right) + \sum_{i=1}^{k} 2r\log(\beta_{k_i}) \end{aligned} \tag{6.1}$$

For any model $\boldsymbol{k}$, the Laplace approximation of $m_{\boldsymbol{k}}(\boldsymbol{y}_n)$ is given by

$$(2\pi)^{\frac{k}{2}+1}\exp\left\{f(\hat{\boldsymbol{\beta}}_k,\hat{\tau},\hat{\sigma^2})\right\}|V(\hat{\boldsymbol{\beta}}_k,\hat{\tau},\hat{\sigma^2})|^{-\frac{1}{2}}, \tag{6.2}$$

where $(\hat{\boldsymbol{\beta}}_k,\hat{\tau},\hat{\sigma^2}) = \arg\max_{(\boldsymbol{\beta},\tau,\sigma^2)} f(\boldsymbol{\beta},\tau,\sigma^2)$ obtained via the optimization function nlm in R using a Newton-type algorithm and $V(\hat{\boldsymbol{\beta}}_k,\hat{\tau},\hat{\sigma^2})$ is a $(k+2)\times(k+2)$ symmetric matrix with the following blocks:

$$\begin{aligned} & V_{11} = \frac{1}{\tau\sigma^2}I_k + \frac{1}{\sigma^2}X_k^TX_k + diag\left\{\frac{2r}{\beta_{k_1}^2},\ldots,\frac{2r}{\beta_{k_k}^2}\right\}, V_{12} = -\frac{\boldsymbol{\beta}_k}{\tau^2\sigma^2}, \\ & V_{13} = -\frac{\boldsymbol{\beta}_k}{\tau\sigma^4} - \frac{X_k^TX_k\boldsymbol{\beta}_k - X_k^T\boldsymbol{y}_n}{\sigma^4}, V_{22} = -\frac{rk + \frac{k}{2} + \frac{3}{2}}{\tau^2} + \frac{\boldsymbol{\beta}_k^T\boldsymbol{\beta}_k}{\tau^3\sigma^2} + \frac{n}{\tau^3}, V_{23} = \frac{\boldsymbol{\beta}_k^T\boldsymbol{\beta}_k}{2\tau^2\sigma^4}, \\ & V_{33} = -\frac{rk + \frac{k}{2} + \frac{n}{2} + \alpha_1 + 1}{\sigma^4} + \frac{\boldsymbol{\beta}_k^T\boldsymbol{\beta}_k}{\tau\sigma^6} + \frac{(\boldsymbol{y}_n - X_k\boldsymbol{\beta}_k)^T(\boldsymbol{y}_n - X_k\boldsymbol{\beta}_k)}{4\sigma^6} + \frac{2\alpha_2}{\sigma^6}. \end{aligned} \tag{6.3}$$

The above Laplace approximation can be used to compute the log of the posterior probability ratio between any given model $\boldsymbol{k}$ and true model $\boldsymbol{t}$, and select a model $\boldsymbol{k}$ with the highest probability. Based on a reviewer's comment, we would like to point out that Laplace approximation could have potential drawbacks. Firstly, as indicated in Rossell and Telesca (2017), for non-local priors, Laplace approximations fail to consistently estimate the marginal likelihood for overfitted models. Secondly, the Newton-type algorithm used for optimizing (6.1) could be quite time consuming, especially when the size of the model and the dimension p are large. For example, in Figure 5, the runtime for the hyper-pMOM approach increases as p grows. However, the computation cost could potentially be significantly improved by using other optimization algorithms in high dimensions. For example, the coordinate descent algorithm in (Friedman, Hastie, and Tibshirani, 2010), or other first-order based algorithms including gradient descent may reduce computational cost.

Despite these potential drawbacks of the Laplace approximation, we would like to point out that in these high-dimensional settings, full posterior sampling using Markov chain Monte Carlo algorithms is highly inefficient and often not feasible from a practical perspective. Hence, the usage of Laplace approximation is still much better than MCMC.

We then adopt the scalable stochastic search algorithm proposed by Shin et al. (2018) called Simplified Shotgun Stochastic Search with Screening (S5). Utilizing the Laplace approximations of the marginal probabilities in (6.2), the S5 method aims at rapidly identifying regions of high posterior probability and finding the maximum a posteriori (MAP) model. Detailed algorithm steps can be found in Shin et al. (2018) and the implementation can be found in the R package "BayesS5".

7 Experiments

7.1 Simulation I: illustration of posterior ratio consistency

In this section, we illustrate the model selection consistency results in Theorems 3.1 and 3.2 using a simulation experiment. The similar simulation setting was also considered in the literature (Cao et al., 2019) by Cao, Khare and Ghosh, in which the authors showed posterior consistency in graphical model setting. We generate our data according to a Gaussian linear model based on the following mechanism. First, we vary p from 500 to 3000 and let $n = p/5$. Then, for each fixed p, ten covariates are taken as active in the true model with coefficients $\boldsymbol{\beta}_0 = (1.1, 1.2, 1.3, \dots, 1.9, 2)^T$ and set $\sigma = 1$. Also, the signs of the true regression coefficients were randomly changed with probability 0.5. Next, we generate n i.i.d. observations from the $N(\mathbf{0}_p, \Sigma)$ distribution as rows of the covariate matrix X. We then examine posterior ratio consistency under three different cases of Σ by computing the log posterior ratio of a "non-true" model $\boldsymbol{k}$ and $\boldsymbol{t}$ as follows.

1. Case 1: Isotropic design, where $\Sigma = I_p$.
2. Case 2: Compound symmetry design, where $\Sigma_{ij} = 0.5$, if $i \neq j$ and $\Sigma_{ii} = 1$, for all $1 \leq i \leq j \leq p$.
3. Case 3: Autoregressive correlated design; where $\Sigma_{ij} = 0.5^{|i-j|}$, for all $1 \leq i \leq j \leq p$.

Throughout this simulation study, we set the hyperparameters $r = 2$ and $\alpha_1 = \alpha_2 = 0.01$. The log of the posterior probability ratio for various cases of Σ is provided in Figure 3. Note that for each of these cases, we compute the log ratio under four different scenarios of "non-true" model $\boldsymbol{k}$.

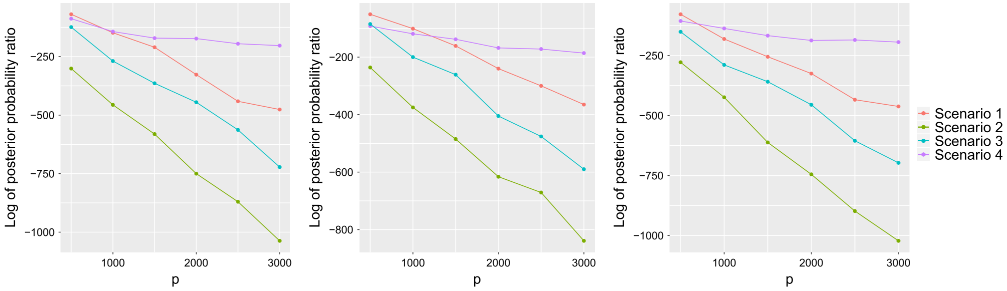


Figure 3: Log of posterior probability ratio for $\boldsymbol{k}$ and $\boldsymbol{t}$ for various choices of the "non-true" model $\boldsymbol{k}$. Left: case 1; middle: case 2; right: case 3.

1. Scenario 1: $\boldsymbol{k}$ is a subset of $\boldsymbol{t}$ and $|\boldsymbol{k}| = \frac{1}{2}|\boldsymbol{t}|$.
2. Scenario 2: $\boldsymbol{k}$ is a superset of $\boldsymbol{t}$ and $|\boldsymbol{k}| = 2|\boldsymbol{t}|$.
3. Scenario 3: $\boldsymbol{k}$ is not necessarily a subset of $\boldsymbol{t}$, but $|\boldsymbol{k}| = \frac{1}{2}|\boldsymbol{t}|$.
4. Scenario 4: $\boldsymbol{k}$ is not necessarily a superset of $\boldsymbol{t}$, but $|\boldsymbol{k}| = 2|\boldsymbol{t}|$.

As expected the log of the posterior probability ratio for any "non-true" model $\boldsymbol{k}$ compared to the true model $\boldsymbol{t}$ are all decreasing to large negative values as p increases, thereby providing a numerical illustration of Theorems 3.1 and 3.2.

7.2 Simulation II: illustration of model selection

In this section, we perform a simulation experiment to illustrate the potential advantages of using our Bayesian approach. Several different values of p ranging from 500 to 3000 are considered, while $n = p/5$. For each fixed p, we construct two sets of $\boldsymbol{\beta}_0$. The first set is generated by the same mechanism as in Section 7.1. The other set also considered is $(0.3, 0.35, 0.4, 0.45, 0.5, 1.1, 1.2, 1.3, 1.4, 1.5)^T$. Next, we generate n i.i.d. observations from the $N(\mathbf{0}_p, \Sigma)$ distribution as rows of covariate matrix X under the following three cases similar to Section 7.1.

1. Case 1: Isotropic design, where $\Sigma = I_p$.
2. Case 2: Compound symmetry design, where $\Sigma_{ij} = 0.5$, if $i \neq j$ and $\Sigma_{ii} = 1$, for all $1 \leq i \leq j \leq p$.
3. Case 3: Autoregressive correlated design; where $\Sigma_{ij} = 0.5^{|i-j|}$, for all $1 \leq i \leq j \leq p$.

Then, we perform model selection using our hierarchical Bayesian approach. This is done by computing the posterior probabilities using the Laplace approximation in (6.2),

		Lasso			SCAD			BayesS5			$r = 1, \tau = 0.348$			$r = 2, \tau = 0.072$			Hyper-pMOM	
p	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR
200	0.29	1	0.14	0.93	1	0.01	1	1	0	0.84	1	0.01	0.96	1	0	1	1	0
500	0.19	1	0.09	0.90	1	0	1	1	0	0.64	1	0.01	0.88	1	0	1	1	0
1000	0.18	1	0.04	0.87	1	0	0.98	1	0	0.51	1	0.01	0.69	1	0	1	1	0
1500	0.18	1	0.03	0.84	1	0	0.98	1	0	0.45	1	0.01	0.74	1	0	1	1	0
2000	0.17	1	0.02	0.82	1	0	0.98	1	0	0.30	1	0.01	0.59	1	0	1	1	0
2500	0.13	1	0.02	0.90	1	0	0.97	1	0	0.23	1	0.01	0.49	1	0	1	1	0

		Lasso			SCAD			BayesS5			$r = 1, \tau = 0.348$			$r = 2, \tau = 0.072$			Hyper-pMOM	
p	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR
200	0.27	1	0.15	0.96	1	0	0.94	1	0	0.83	1	0.01	0.81	1	0.01	1	1	0
500	0.21	1	0.09	0.94	1	0	0.95	1	0	0.57	1	0.03	0.59	1	0.02	1	1	0
1000	0.17	1	0.05	0.92	1	0	0.95	1	0	0.45	1	0.02	0.46	1	0.01	0.99	1	0
1500	0.19	1	0.03	0.90	1	0	0.94	1	0	0.27	1	0.01	0.42	1	0.01	1	1	0
2000	0.13	1	0.04	0.84	1	0	0.87	1	0	0.20	1	0.02	0.41	1	0.01	0.99	1	0
2500	0.12	1	0.03	0.92	1	0	0.88	1	0	0.18	1	0.02	0.36	1	0.01	0.99	1	0

		Lasso			SCAD			BayesS5			$r = 1, \tau = 0.348$			$r = 2, \tau = 0.072$			Hyper-pMOM	
p	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR
200	0.25	1	0.17	0.91	1	0	1	1	0	1	1	0	0.96	1	0	1	1	0
500	0.20	1	0.10	0.91	1	0	0.98	1	0	0.77	1	0.01	0.83	1	0	1	1	0
1000	0.18	1	0.05	0.85	1	0	0.97	1	0	0.59	1	0.01	0.73	1	0	1	1	0
1500	0.16	1	0.04	0.83	1	0	0.96	1	0	0.41	1	0.01	0.71	1	0	1	1	0
2000	0.17	1	0.04	0.83	1	0	0.96	1	0	0.36	1	0.01	0.57	1	0	0.99	1	0
2500	0.14	1	0.03	0.85	1	0	0.96	1	0	0.28	1	0.01	0.56	1	0	1	1	0

Table 1: Model selection performance comparison table when $|\boldsymbol{\beta}_0| = (1.1, 1.2, 1.3, \ldots, 1.9, 2)^T$. Top: case 1; middle: case 2; bottom: case 3.

		Lasso			SCAD			BayesS5			$r = 1, \tau = 0.348$			$r = 2, \tau = 0.072$			Hyper-pMOM	
p	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR
200	0.32	1	0.22	1	1	0	0.97	0.95	0	0.83	1	0.01	0.95	0.96	0	1	0.86	0
500	0.27	1	0.06	0.48	1	0.02	0.97	0.93	0	0.71	1	0.01	0.89	0.93	0	1	0.84	0
1000	0.13	1	0.07	0.59	1	0.01	0.95	0.95	0	0.41	1	0.01	0.88	0.89	0	1	0.88	0
1500	0.23	1	0.02	0.61	0.89	0	0.97	0.90	0	0.33	1	0.01	0.84	0.90	0	1	0.88	0
2000	0.19	1	0.03	0.63	1	0	0.97	0.89	0	0.25	1	0.01	0.74	0.89	0	1	0.84	0
2500	0.16	1	0.03	0.59	1	0.01	0.99	0.87	0	0.22	0.90	0.01	0.77	0.88	0	1	0.83	0

		Lasso			SCAD			BayesS5			$r = 1, \tau = 0.348$			$r = 2, \tau = 0.072$			Hyper-pMOM	
p	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR
200	0.32	1	0.12	0.88	0.7	0.01	0.99	0.77	0	0.82	0.90	0.01	0.79	0.84	0.01	1	0.81	0
500	0.26	1	0.06	1	0.83	0	1	0.72	0	0.47	0.91	0.02	0.74	0.82	0.01	1	0.83	0
1000	0.19	0.89	0.02	0.57	0.81	0.01	1	0.69	0	0.46	0.90	0.01	0.60	0.84	0.01	1	0.79	0
1500	0.19	0.91	0.03	0.57	0.80	0.05	0.99	0.65	0	0.26	0.85	0.02	0.70	0.80	0	1	0.79	0
2000	0.17	1	0.15	0.66	0.79	0.03	0.95	0.67	0	0.23	0.83	0.02	0.62	0.80	0	0.94	0.74	0
2500	0.18	1	0.19	0.51	0.72	0.03	0.95	0.64	0	0.21	0.82	0.01	0.57	0.78	0	0.95	0.70	0

		Lasso			SCAD			BayesS5			$r = 1, \tau = 0.348$			$r = 2, \tau = 0.072$			Hyper-pMOM	
p	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR
200	0.37	1	0.09	0.7	1	0.02	0.99	0.92	0	0.82	1	0.01	0.94	0.93	0	1	0.90	0
500	0.23	1	0.07	0.89	0.79	0	0.96	0.90	0	0.77	1	0.01	0.89	0.87	0	1	0.88	0
1000	0.13	1	0.07	0.48	0.95	0.01	0.96	0.88	0	0.67	1	0.01	0.77	0.86	0	1	0.84	0
1500	0.21	1	0.03	0.36	0.80	0.01	0.97	0.87	0	0.36	0.91	0.01	0.75	0.86	0	1	0.89	0
2000	0.16	0.9	0.03	0.35	0.71	0.01	0.95	0.88	0	0.25	1	0.01	0.84	0.82	0	1	0.86	0
2500	0.13	1	0.03	0.45	0.68	0	0.95	0.81	0	0.22	1	0.01	0.80	0.82	0	1	0.78	0

Table 2: Model selection performance comparison table when $\boldsymbol{\beta}_0 = (0.3, 0.35, 0.4, 0.45, 0.5, 1.1, 1.2, 1.3, 1.4, 1.5)^T$. Top: case 1; middle: case 2; bottom: case 3.

and exploring the model space using the simplified stochastic shotgun stochastic search algorithm in (Shin et al., 2018).

We would like to remind the readers that in our model, we don't need to specify a fixed value for τ, but rather put a prior on the parameter τ (as opposed to (Johnson and Rossell, 2012) and (Shin et al., 2018) when τ is treated as a fixed parameter). In Table 1 and Table 2, we also provide model selection performance results with fixed τ at $r = 2, \tau = 0.072$ and $r = 1, \tau = 0.348$ (the default value for the first and second-order pMOM prior suggested in Johnson and Rossell (2012)), and numerical values in R package

BayesS5 (a choice for fixed τ from the results in (Shin et al., 2018)). Additionally, we also provide model selection performance results for the Lasso (Tibshirani, 1996) and SCAD (Fan and Li, 2001) penalized likelihood methods.

The model selection performance of these five methods is then compared using several different measures of structure such as positive predictive value, true positive rate and false positive rate (average over 20 independent repetitions). Positive Predictive Value (PPV) represents the proportion of true model indexes among all the indexes detected by the given procedure. True Positive Rate (TPR) measures the proportion of true indexes detected by the given procedure among all the true indexes from the true model. False Positive Rate (FPR) represents the proportion of falsely identified indexes among all the non-true indexes from the true model. PPV, TPR and FPR are defined as

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}},$$

where TP, TN, FP and FN correspond to true positive, true negative, false positive and false negative, respectively. One would like the PPV and TPR values to be as close to 1 as possible, while FPR to be as close to 0 as possible. The results are summarized in Table 1 and Table 2.

To better visualized the results, in Figure 4, we provide the ROC curves when $|\boldsymbol{\beta}_0| = (1.1, 1.2, 1.3, \ldots, 1.9, 2)^T$ and Σ for generating $\boldsymbol{X}$ yields a compound symmetry design. We also include the complexity prior based approach illustrated in Section 5. As we can see, the complexity prior based approach captures fewer true indexes compared to other approaches.

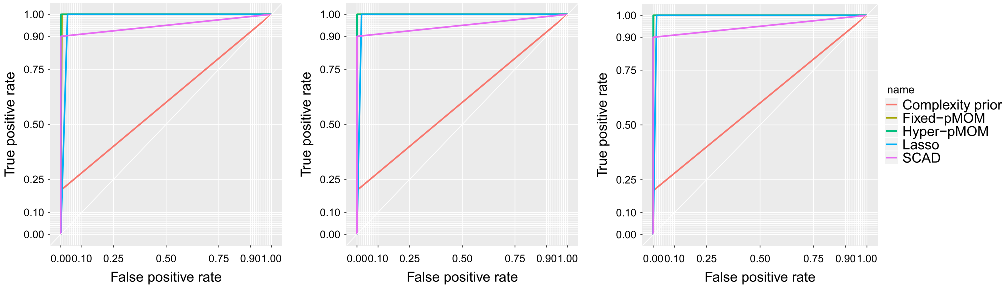


Figure 4: ROC curves when $p = 1500$ (left), $p = 2000$ (middle), $p = 2500$ (right).

Based on Table 1 and 2, it is clear that our Bayesian approach outperforms both the penalized likelihood approaches and the fixed τ settings based on almost all measures and under all cases. The PPV values for our hyper-pMOM approach are all higher than the other five methods, which means our method can identify the true model more precisely. In addition, The FPR values for the Bayesian approach are all significantly smaller than the FPR values for the penalized approaches. It is also worth noting that especially in lower dimensions, the numerical procedure for choosing τ implemented in BayesS5 needs additional run time as shown in Figure 5, while in our simulation studies, not only this step is omitted, we are still able to better simulation results. Overall,

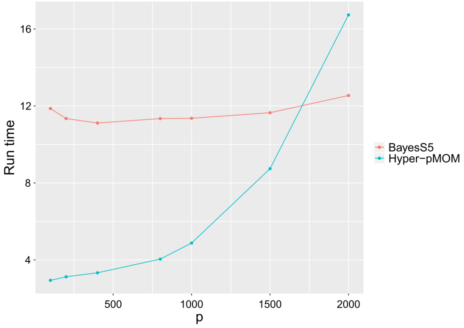


Figure 5: Run time comparison in seconds.

this experiment illustrates the fact that the Bayesian approach can lead to a significant improvement in model selection performance as compared to penalized likelihood methods. Also, the hierarchical Bayesian approach introduced in this paper can lead to a significant improvement in performance as compared to the fixed τ Bayesian approach when sample size is much smaller than the number of predictors.

8 Real data analysis

In this section, we carry out the real data analysis to examine the performance of proposed method based on the Boston housing dataset. The dataset contains the median value of owner-occupied homes in the Boston region as the response variable, together with several other possible predictor variables including the geographical characteristics. The total number of observations is $n = 506$ and 10 continuous variables: crim, indus, nox, rm, age, dis, tax, ptratio, b, and lstat are considered as the predictor variables. Several approaches for variable selection have been demonstrated via this housing dataset. See for example (Yuan and Lin, 2005; Shin et al., 2018).

We added 900 noise variables generated independently from a standard normal distribution, and additional 100 noise variables that obey the multivariate normal distribution with a autoregressive correlated Σ, where $\Sigma_{ij} = 0.5^{|i-j|}$, for all $1 \leq i \leq j \leq 100$, to perform the model selection in a $p > n$ regression setting. The design matrix is standardized and the dataset is divided into a training set of size 406 and a test set of size 100. We first obtain the model estimate based on the training set and then compare the proposed hyper-pMOM approach with the following four methods on the test set: pMOM with fixed $\tau = 0.072$, peMOM with simplified shotgun stochastic search, and two frequentist approaches, Lasso and SCAD.

The results are summarized in Table 3 averaged over 100 repetitions based on the following five measures also adopted in (Shin et al., 2018). MSPE represents the out-of-sample square prediction error calculated by

$$\text{MSPE} = \frac{1}{100} \sum_{i \in test} \left(y_i - X_i^T \hat{\boldsymbol{\beta}}_{\hat{k}}^{train} \right)^2,$$

	MSPE	MS-O	MS-N	FS-O	TS-O
Hyper-pMOM	25.41	3	0	3	3
pMOM	34.43	5.10	4	5	6
peMOM	27.05	5	1	5	5
Lasso	30.19	5.26	35.79	4	6
SCAD	26.11	5	11.87	5	5

Table 3: Model selection comparison based on the Boston housing data.

where $\hat{\boldsymbol{\beta}}_{\hat{k}}^{train}$ is the least squared estimator based on the model estimate obtained from the test set. MS-O and MS-N refer to the average original variables and falsely selected noise variables over 100 repetitions, respectively. FS-O is the number of original variables that are selected at least 95 out of 100 repetitions. TS-O refers to the number of original variables that are selected at least once from 100 repetitions.

As we see in Table 3, our hyper-pMOM approach consistently identifies the same model and had the lowest prediction error among all the five methods. In particular, the average number of the original variables that are selected at least 95 times is 3. Across all the 100 repetitions, our hyper-pMOM method successfully avoids selecting any noise variable, while all the other four methods falsely identify at least one noise variable. Overall, the real data application illustrates our hyper-pMOM approach yields the most stable and accurate model selection among all the five methods.

9 Discussion

This article describes and examines theoretical properties of hyper-pMOM priors proposed in (Wu, 2016) for variable selection in high-dimensional linear model settings. Under standard regularity assumptions, which include the prior over all models is restricted to ones with model size less than an appropriate function of the sample size n, we establish posterior ratio consistency (Theorem 3.1), i.e., the ratio of the maximum marginal posterior probability assigned to a "non-true" model to the posterior probability assigned to the "true" model converges to zero in probability. Next, under the additional assumption that p increases at a polynomial rate with n, we show strong model selection consistency (Theorem 3.2). Strong model selection consistency implies that the posterior probability of the true model converges in probability to 1 as $n \to \infty$.

Based on the reviewers' comments, we realize the polynomial rate restriction on p could be rather limited. By carefully examining our theoretical analysis, in Section 5, we add another result where we replace the uniform-like prior with the complexity prior on the model space to penalize larger models, and establish strong model selection consistency (Theorem 5.2) when p is allowed to grow at a sub-exponential rate of n. However, through simulation studies, we find out that the model selection performance under the uniform-like prior is much better than that under the complexity prior, hence from a practical point of view, one would still prefer the hyper-pMOM with uniform-like prior on the model space.

In Section 6, we provide details about the application of Laplace approximation to approximate the posterior density and illustrate the potential benefits for our hyper-pMOM based model selection procedure compared with other methods via simulation studies and real data analysis in Section 7 and Section 8, respectively.

Acknowledgment

The authors are grateful to the Editor and three anonymous referees for their encouraging and helpful comments which substantially improved the paper.

Supplementary Material

Supplementary Material for "High-Dimensional Posterior Consistency for Hierarchical Non-Local Priors in Regression"
(DOI: 10.1214/19-BA1154SUPP; .pdf).

References

Bian, Y. and Wu, H.-H. (2017). "A Note on Nonlocal Prior Method." *arXiv:1702.07778*. 243, 248

Cao, X., Khare, K., and Ghosh, M. (2019). "Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models." *Annals of Statistics*, 47(1): 319–348. MR3909935. doi: https://doi.org/10.1214/18-AOS1689. 255

Cao, X., Khare, K., and Ghosh, M. (2019). "Supplementary Material for "High-Dimensional Posterior Consistency for Hierarchical Non-Local Priors in Regression"." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1154. 249

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). "Bayesian linear regression with sparse priors." *Annals of Statistics*, 43: 1986–2018. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 251, 252

Fan, J. and Li, R. (2001). "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties." *Journal of the American Statistical Association*, 96: 1348–1360. MR1946581. doi: https://doi.org/10.1198/016214501753382273. 258

Friedman, J., Hastie, T., and Tibshirani, R. (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software, Articles*, 33(1): 1–22. MR1082147. 255

George, E. I. and McCulloch, R. E. (1993). "Variable Selection via Gibbs Sampling." *Journal of the American Statistical Association*, 88: 881–889. 241

Ishwaran, H., Kogalur, U. B., and Rao, J. S. (2005). "Spike and slab variable selection: Frequentist and Bayesian strategies." *Annals of Statistics*, 33: 730–773. MR2163158. doi: https://doi.org/10.1214/009053604000001147. 241

Johnson, V. and Rossell, D. (2010). "On the Use of Non-Local Prior Densities in

Bayesian Hypothesis Tests Hypothesis." *Journal of the Royal Statistical Society. Series B*, 72: 143–170. MR2830762. doi: https://doi.org/10.1111/j.1467-9868.2009.00730.x. 241, 242

Johnson, V. and Rossell, D. (2012). "Bayesian Model Selection in High-Dimensional Settings." *Journal of the American Statistical Association*, 107: 649–660. MR2980074. doi: https://doi.org/10.1080/01621459.2012.682536. 242, 246, 247, 248, 254, 257

Liang, F., Paulo, R., Molina, G., Clyde, A. M., and Berger, O. J. (2008). "Mixtures of g Priors for Bayesian Variable Selection." *Journal of the American Statistical Association*, 103: 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 241, 242, 245

Narisetty, N. and He, X. (2014). "Bayesian variable selection with shrinking and diffusing priors." *Annals of Statistics*, 42: 789–817. MR3210987. doi: https://doi.org/10.1214/14-AOS1207. 241, 243

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. 254

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. 242

Shin, M., Bhattacharya, A., and Johnson, V. (2018). "Scalable Bayesian Variable Selection Using Nonlocal Prior Densities in Ultrahigh-Dimensional Settings." *Statist. Sinica*, 28: 1053–1078. MR3791100. 242, 243, 247, 248, 253, 254, 255, 257, 258, 259

Song, Q. and Liang, F. (2015). "High-Dimensional Variable Selection With Reciprocal L_1-Regularization." *Journal of the American Statistical Association*, 110: 1607–1620. MR3449058. doi: https://doi.org/10.1080/01621459.2014.984812. 247

Tibshirani, R. (1996). "Regression Shrinkage and Selection Via the Lasso." *Journal of the Royal Statistical Society. Series B*, 58: 267–288. MR1379242. 258

Wu, H.-H. (2016). "Nonlocal Priors for Bayesian Variable Selection in Generalized Linear Models and Generalized Linear Mixed Models and Their Applications in Biology Data." Ph.D. thesis, University of Missouri. MR3698950. 241, 242, 243, 244, 260

Yuan, M. and Lin, Y. (2005). "Efficient Empirical Bayes Variable Selection and Estimation in Linear Models." *Journal of the American Statistical Association*, 100(472): 1215–1225. MR2236436. doi: https://doi.org/10.1198/016214505000000367. 259

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." *Bayesian Inference and Decision Techniques, Stud. Bayesian Econometrics Statist.*, 6: 233–243. MR0881437. 241, 242