High-dimensional posterior consistency for hierarchical non-local priors in regression

XUAN CAO\textsuperscript{1}, KSHITIJ KHARE\textsuperscript{2} and MALAY GHOSH\textsuperscript{2}

\textsuperscript{1}Department of Mathematical Sciences, University of Cincinnati
4199 French Hall West
Cincinnati, OH 45221, USA

\textsuperscript{2}Department of Statistics, University of Florida
102 Griffin-Floyd Hall
Gainesville, FL, 32611 USA
E-mail: caox4@ucmail.uc.edu; kdkhare@stat.ufl.edu; ghoshm@ufl.edu

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 a constant, and suggest choices based on prior coverage/asymptotic considerations. In this paper, we introduce a fully Bayesian approach with the pMOM non-local prior where we place an appropriate Inverse-Gamma prior on the tuning parameter to analyze a more robust model that is comparatively immune to misspecification of the scale parameter. 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 \( n \). Through simulation studies, we demonstrate that our model selection procedure can outperform other Bayesian methods which treat the scale parameter as a constant, and commonly used penalized likelihood methods, in a range of simulation settings.

MSC 2010 subject classifications: Primary 62F15; secondary 62G20.
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, such as the stochastic search variable selection ([6]), empirical Bayes variable selection ([5]), spike and slab selection ([7], [13]), \( g \)-prior based variable selection ([11]), and penalized credible regions ([2]), to name a few. In recent years, the use of non-local priors in this context has generated a lot of interest. Non-local priors were first introduced in [8] by Johnson and Rossell as densities that are identically zero whenever a model parameter is equal to its null value in the context 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 [8]. These priors were further extended to Bayesian model selection problems in [9] by imposing non-local prior densities on a vector of regression coefficients.

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

\[
y_n \sim \mathcal{N}(X_n \beta, \sigma^2 I_n).
\]

In [9], the authors introduce the product moment (pMOM) non-local prior with density

\[
d_p(2\pi)^{-\frac{p}{2}}(\tau \sigma^2)^{-\frac{p-r}{2}}|A_p|^{\frac{1}{2}} \exp \left\{ \frac{-\beta' A_p \beta}{2\tau \sigma^2} \right\} \prod_{i=1}^{p} \beta_i^{2r}.
\]

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 \( \tau \) and \( \sigma^2 \). Variations of the density in (1.1), called the piMOM and peMOM density, have also been developed in [9, 14]. Clearly, the density in (1.1) is zero when any component of \( \beta \) is zero. Under appropriate regularity conditions, the authors in [9, 15] 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 [9], the scale parameter \( \tau \) 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. In [8, 9], the authors treat \( \tau \) as a constant and suggest a choice of \( \tau \) which leads to a high prior probability for significant values of the regression coefficients. In [15], the authors again treat \( \tau \) as a constant, and consider a setting where \( p \) and \( \tau \) 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 [14]), as long as \( \tau \) is of a larger order than \( \log p \) and smaller order than \( n \).

The primary goal and innovation of this paper is to propose and analyze a fully Bayesian approach with the pMOM non-local prior where we place an appropriate Inverse-Gamma prior on the parameter \( \tau \). A clear advantage of such an approach, compared to previous approaches, is that it is more robust and comparatively immune to misspecification of \( \tau \). The extra prior layer, 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 [13] and [15] for regression models.

The authors in [1] also discuss the potential advantages of using priors on \( \tau \) in the context of generalized linear model setting and establish model selection consistency. While there are some connections between our model and the one in [1], there are fundamental differences between the 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. Details about how to approximate the posterior density for model selection are demonstrated in Section 5. In Section 6 we use simulation experiments to illustrate the posterior ratio consistency result, and demonstrate the benefits of model selection using our fully Bayesian approach as compared to approaches which treat \( \tau \) as a constant, and also compared to existing penalized likelihood approaches.

2. Model specification

We start by considering the standard Gaussian linear regression model with \( p \) coefficients and by introducing some required notation. Let \( y_n \) denote a random vector of responses, \( X_n \) an \( n \times p \) design matrix of covariates, and \( \beta \) 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 \( k = \{k_1, k_2, \ldots, k_m\} \) if and only if all the non-zero elements of \( \beta \) are \( \beta_{k_1}, \beta_{k_2}, \ldots, \beta_{k_m} \) and denote \( \beta_k = (\beta_{k_1}, \beta_{k_2}, \ldots, \beta_{k_m})^T \). For any \( p \times p \) matrix \( A \), let \( A_k \) represent the submatrix formed from the columns of \( A \) corresponding to model \( k \). In particular, Let \( X_k \) denote the design matrix formed from the columns of \( X_n \) corresponding to model \( k \). For the rest of the paper, simply let \( k = |k| \) represent the cardinality of model \( k \) for notational convenience.

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

\[
Y_n \mid \beta_k, \sigma^2, k \sim N(X_k \beta_k, \sigma^2 I_n)
\]

\[
\pi(\beta_k \mid \tau, \sigma^2, k) = d_k(2\pi)^{-\frac{1}{2}(\tau \sigma^2)^{-\frac{1}{2}}\frac{1}{2}}A_k^\frac{1}{2} \exp \left\{ -\frac{\beta_k' A_k \beta_k}{2\tau \sigma^2} \right\} \prod_{i=1}^{k} \beta_{k_i}^{2r_i}
\]

\[
\pi(\tau) = \frac{1}{\Gamma(\frac{1}{2})} \tau^{-\frac{1}{2}} e^{-n \tau},
\]

\[
\pi(\sigma^2) = \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} (\sigma^2)^{-\alpha_1 - 1} e^{-\frac{\alpha_2}{\sigma^2}}.
\]

(2.1)
Note that in (2.1), 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.

By (2.1) and Bayes’ rule, the resulting posterior probability for model \( k \) is denoted by,

\[
\pi(k | y_n) = \frac{\pi(k)}{\pi(y_n)} m_k(y_n), \tag{2.2}
\]

where \( \pi(y_n) \) is the marginal density of \( y_n \), and \( m_k(y_n) \) is the marginal density of \( y_n \) under model \( k \) given by,

\[
m_k(y_n) = \int_0^\infty \int_0^\infty \pi(y_n | \beta_k, \sigma^2, k) \pi(\beta_k | \tau, \sigma^2, k) \pi(\tau) \pi(\sigma^2) \, d\beta_k \, d\sigma^2 \, d\tau
\]

\[
= \frac{(\frac{n}{2})^{\frac{3}{2}} (\alpha_2)^{\alpha_1}}{\Gamma(\frac{3}{2}) \Gamma(\alpha_1)} \int_0^\infty \int_0^\infty d_k(2\pi)^{-\frac{3}{2}} (\sigma^2)^{-\frac{k}{2} - \frac{3}{2}} |A_k|^\frac{3}{2} \exp \left\{ \frac{\beta_k^T A_k \beta_k}{2\tau \sigma^2} \right\} \prod_{i=1}^k \beta_{ki}^{2\tau} \times \frac{1}{(2\pi \sigma^2)^\frac{k}{2}} \exp \left\{ -\frac{(y_n - X_k \beta_k)^T (y_n - X_k \beta_k)}{2\sigma^2} \right\} \tau^{-\frac{3}{2} \tau - \frac{3}{2} \tau} e^{-\frac{\tau}{2\sigma^2}} \pi((\sigma^2)^{-\alpha_1 + 1}) e^{-\frac{\sigma^2}{2\tau}} d\beta_k \, d\sigma^2 \, d\tau
\]

\[
= d_k \left( \frac{(\frac{n}{2})^{\frac{3}{2}} (\alpha_2)^{\alpha_1}}{\Gamma(\frac{3}{2}) \Gamma(\alpha_1)} |A_k|^\frac{3}{2} \right) \int_0^\infty \int_0^\infty (\sigma^2)^{-\frac{k}{2} + 1 + \alpha_1} \exp \left\{ \frac{-R_k + 2\alpha_2 \sigma^2}{2\sigma^2} \right\} \tau^{-\frac{k}{2} + \alpha_1 + 1} e^{-\frac{\tau}{2\sigma^2}} E_k \left( \prod_{i=1}^k \beta_{ki}^{2\tau} \right) \, d\sigma^2 \, d\tau, \tag{2.3}
\]

where \( C_k = X_k^T X_k + \frac{1}{\tau} \), \( R_k = y_n^T (I_n - X_k C_k^{-1} X_k^T) y_n \), and \( E_k(\cdot) \) denotes the expectation with respect to a multivariate normal distribution with mean \( \hat{\beta}_k = C_k^{-1} X_k^T y_n \), 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

\[
\hat{k} = \arg \max_k \pi(k | y_n). \tag{2.4}
\]

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 \( Y_n = X_n \beta_0 + \epsilon_n \). Here \( \beta_0 \) is the true \( p_n \) dimensional vector of regression coefficients, whose dependence on \( n \) is suppressed for notational convenience, and the entries of \( \epsilon_n \) are i.i.d Gaussian with mean zero and variance \( \sigma_0^2 \). As
in [9], we assume that the true vector of regression coefficients is sparse, i.e., all the entries of \( \beta_0 \) are zero except those corresponding to a subset \( t \subseteq \{1, 2, \ldots, p_n\} \), and \( t, \beta_{0,t}, \sigma_0^2 \) do not vary with \( n \). Our results can be easily extended to the case where \( |t| \), and entries of \( \beta_{0,t} \) and \( \sigma_0^2 \) vary with \( n \) but stay bounded. However, we assume these quantities stay fixed for ease of exposition. For any matrix \( A \), denote the \( j \)-th largest nonzero eigenvalue as \( \nu_j(A) \) and 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 \( \epsilon_n^{-1} = O \left( (\log n)^{\frac{1}{18}} \right) \).

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

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

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

**Assumption 5.** For every \( n \geq 1 \), the hyper-parameter for the non-local pMOM prior in 2.1 satisfy \( 0 < a_1 < \text{eig}_1(A_p) \leq \text{eig}_2(A_p) \leq \ldots \leq \text{eig}_p(A_p) < a_2 < \infty \). Here \( a_1, a_2 \) are constants not depending on \( n \).

The authors in [9] 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 \( n \). 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 [15]). 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 [9]). This type of priors have also been considered in [16] and [15]. As indicated in [15], since the pMOM and piMOM priors already induce a strong penalty on the model size, it is no longer necessary to put other priors to penalize larger models, for example, in [18] and [3]. Assumption 2 states that \( p \) can grow at an appropriate polynomial rate with \( n \).

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). Under Assumptions 1, 3, 4 and 5, the following holds:
\[
\max_{k \neq t} \frac{\pi(k|y_n)}{\pi(t|y_n)} \to 0, \quad n \to \infty.
\]
In particular, it implies that the probability that the posterior mode \(\hat{k}\) defined in (2.4) is equal to the true model \(t\) will converge to 1, i.e.,
\[
P(t = \arg \max_k \pi(k|y_n)) \to 1, \quad n \to \infty.
\]

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 \(t\) converges to 1 in probability. We refer to this notion of consistency as strong selection consistency.

Theorem 3.2 (Strong selection consistency). Under Assumptions 1-5, with \(\xi < 1 - \frac{4\gamma}{3\nu}\) in Assumption 3, the following holds:
\[
\pi(t|y_n) \to 1, \quad n \to \infty.
\]

The above results have been established under the pMOM priors. Another class of non-local priors introduced in [9] are the piMOM priors on the regression coefficients, for which the density of the regression coefficients under the model \(k = \{k_1, k_2, \ldots, k_m\}\) is given by
\[
\frac{(\tau \sigma^2)^{\frac{1}{4k}} |k| \prod_i |\beta_{k_i}|^{-(r+1)} \exp \left(-\frac{\tau \sigma^2}{\beta_{k_i}^2} \right)}{\Gamma\left(\frac{r}{2}\right) |k|^{r/2}}.
\]
where \(r\) is a positive integer and is refered to as the order of the density. The corollary below establishes strong model selection consistency under the piMOM priors, and can be obtained immediately by combining Theorem 3.2 with [10, Eq. (59),(60)].

Corollary 3.1 (Strong selection consistency for piMOM priors). Under the same conditions as in Theorem 3.2, when piMOM priors are imposed on \(\beta_k\) in model (2.1), the following holds:
\[
\pi(t|y_n) \to 1, \quad n \to \infty.
\]

Remark 1. In the context of generalized linear regression, Bian and Wu [1] consider the following hierarchical Bayesian model on regression coefficients.
\[
\beta_k | \tau_i \sim \frac{(\tau \sigma^2)^{\frac{1}{4k}} |k| \prod_i |\beta_{k_i}|^{-(r+1)} \exp \left(-\frac{\tau \sigma^2}{\beta_{k_i}^2} \right)}{\Gamma\left(\frac{r}{2}\right) |k|^{r/2}},
\]
\[
\sigma^2 \sim \text{Inverse-Gamma} \left(\frac{(r+1)}{2}, \lambda\right).
\]
Posterior consistency for non-local priors in regression

In particular, they put an independent piMOM prior on each linear regression coefficient (conditional on the hyper parameter $\tau_i$), and an inverse Gamma prior on $\tau_i$. In this setting, the authors in [1] 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 [1] 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 $\beta_k$. In particular, pMOM distributions introduce correlations among the entries in $\beta_k$ 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 $\beta_k$, which is not the case in [1]. Hence it is structurally different than the prior in [1]. Secondly, in order to prove consistency results, the authors in [1] 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 [1]. In addition, Assumption C2 in [1] 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 $\tau$, 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 the authors in [1] directly follow the proof in [15] 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 $k$, $R_{k} = y_n^T (I - X_k (X_k^T X_k)^{-1} X_k^T) y_n$, 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 (2.3)) with $R^*_t$ and $R^*_k$ respectively. Fix a model $k \neq t$ arbitrarily, and let $u = k \cup t$ and $u = |u|$ be the cardinality of $u$. Note that $R^*_t \sim \chi^2_{n-t}$, $R^*_u \sim \chi^2_{n-u}$, $R^*_u \sim \chi^2_{u-t}$, and $\frac{y_n^T P_k y_n}{\sigma_0^2} \sim \chi^2(u \bigg( \beta_0^T X_k X_k \beta_0 \bigg)).$ Next, we establish two tail probability bounds for the $\chi^2$ distribution and the non-central $\chi^2$ distribution respectively, which will be useful in our analysis.

**Lemma 4.1.** For any $a > 0$, we have,

$$P \left( |\chi^2_p - p| > a \right) \leq 2 \exp \left( -\frac{a^2}{4p} \right).$$

**Proof of Lemma 4.1.** First note that

$$P(\chi^2_p - p > a) \leq \inf_{0 < t < \frac{a}{\sqrt{2}}} \left[ \exp \{-t(p+a)\} E \left( t \chi^2_p \right) \right]$$
Let 

\[ g(t) = -t(p + a) - \frac{p}{2} \log(1 - 2t) \]

then \( g'(t) = -(p + a) + \frac{tp}{2t - 2} \) and \( g''(t) = \frac{2p}{(1 - 2t)^2} > 0 \). Hence, \( g(t) \) is minimized at 

\[ t = t_0 = \frac{a}{2(p + a)} < \frac{1}{2}. \]

Hence, 

\[ \Pr\left( \chi^2_{p} - p > a \right) \leq \exp \left( -\frac{a^2}{4p} \right). \]

Similarly, we get 

\[ \Pr\left( \chi^2_{p} - p < -a \right) \leq \exp \left( -\frac{a^2}{4p} \right). \]

Therefore, 

\[ \Pr\left( |\chi^2_{p} - p| > a \right) \leq 2 \exp \left( -\frac{a^2}{4p} \right). \]

**Lemma 4.2.** For any \( c > 0 \), we have,

\[ \Pr\left( \chi^2_{p}(\lambda) - (p + \lambda) > c \right) \leq \exp \left( -\frac{p}{2} \left\{ \frac{c}{p + \lambda} - \log \left( 1 + \frac{c}{p + \lambda} \right) \right\} \right). \]

**Proof of Lemma 4.2.** Note that \( X \sim \chi^2_{p}(\lambda) \) is equivalent to \( X|K = k \sim \chi^2_{p+2k} \) and \( K \sim \text{Poisson}\left(\frac{\lambda}{2}\right) \). Thus, \( E(X) = EE(X|K) = p + \lambda \) and the moment-generating function of \( X \) is given by

\[
E[\exp(tX)] = EE[\exp(tX)|K] = E \left[ (1 - 2t)^{-\frac{p+2K}{2}} \right] = (1 - 2t)^{-\frac{p}{2}} \exp \left[ \frac{\lambda t}{2(1 - 2t)} \right], \quad 0 < t < \frac{1}{2}.
\]

Hence, for \( c > 0 \),

\[
\Pr\left( \chi^2_{p}(\lambda) - (p + \lambda) > c \right) \leq \inf_{0 < t < \frac{1}{2}} \left[ \exp \left\{ -t(p + \lambda + c) + \frac{\lambda t}{1 - 2t} \right\} (1 - 2t)^{-\frac{p}{2}} \right]. \tag{4.1}
\]
Let \( g(t) = -t(p + \lambda + c) + \frac{\lambda}{2}t^2 \) \( - \frac{c}{2} \log(1 - 2t) \). Since \( g''(t) > 0 \), \( g(t) \) is minimized at \( t = t_0 = \frac{c}{2(p + \lambda + c)} < \frac{1}{2} \) where \( g'(t_0) = 0 \).

Now, for \( c > 0 \), by (4.1),

\[
P \left( \chi_p^2(\lambda) - (p + \lambda) > c \right) 
\leq \exp \left[ -t_0(p + \lambda + c) + \frac{\lambda t_0}{1 - 2t_0} (1 - 2t_0)^{-\frac{c}{2}} \right] 
\leq \exp \left( -\frac{p}{2} \left\{ \frac{c}{p + \lambda} - \log \left( 1 + \frac{c}{p + \lambda} \right) \right\} \right). 
\]

The following result is immediate from Lemma 4.1.

\[
P \left[ \frac{R^n_y}{\sigma_0} - (n - t) > \sqrt{n - 1} \log n \right] \leq P \left[ \frac{R^n_y}{\sigma_0} - (n - t) > 2\sqrt{(n - t) \log n} \right] \leq 2n^{-1} \to 0, 
\]

as \( n \to \infty \). Similarly, we have

\[
P \left[ \frac{R^n_y}{\sigma_0} - (n - u) > \sqrt{n - u} \log n \right] \leq 2n^{-1} \to 0, 
\]

as \( n \to \infty \). Next, by Lemma 4.2, since \( u \supset t \), we have

\[
P \left[ y^n_T P_u y_n \sigma_0^2 - \left( u + \frac{1}{\sigma_0^2} \beta_0^T X^T_i X_i \beta_0 \right) > n \log n - u - \frac{1}{\sigma_0^2} \beta_0^T X^T_i X_i \beta_0 \right]
\leq \exp \left( -\frac{u}{2} \left\{ \frac{n \log n}{u + \frac{1}{\sigma_0^2} \beta_0^T X^T_i X_i \beta_0} - \log \left( 1 + \frac{n \log n}{u + \frac{1}{\sigma_0^2} \beta_0^T X^T_i X_i \beta_0} \right) \right\} \right) \leq \exp \left( -\frac{u}{4} \left\{ \frac{\log n}{1 + \frac{1}{\sigma_0^2} \beta_0^T X^T_i X_i \beta_0} \right\} \right), 
\]

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

\[
C_n = \left\{ \left| \frac{R^n_y}{\sigma_0} - (n - t) \right| > \sqrt{n - 1} \log n \right\} \cup \left\{ \left| \frac{R^n_y}{\sigma_0} - (n - u) \right| > \sqrt{n - u} \log n \right\} 
\cup \left\{ \left| \frac{R^n_y}{\sigma_0} - (n - t) \right| > \sqrt{n - u} \log n \right\} \cup \left\{ \frac{y^n_T P_u y_n}{\sigma_0^2} > n \log n \right\}.
\]
It follows from (4.2), (4.3), (4.4), and (4.5), 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 $C_n^c$.**

**Lemma 4.3.** Under Assumption 1 and Assumption 5, for all $k \neq t$, there exists $N$, such that when $n > N$,

$$
m_k(y_n) < B A^k \left( \frac{V}{\epsilon_n^2} \right)^{rk} k^{k_n-(k-\ell)} \frac{R_1^* + 2\alpha_2}{R_1^* + 2\alpha_2} \frac{\frac{2}{r} + r + \omega_1}{r + \omega_1},
$$

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

**Proof.** It follows by the form of the non-local priors in (2.1), and by Assumptions 1 and 5 that,

$$
\left( \frac{a_1}{a_2} \right)^{\frac{1}{r}} \left( \frac{a_1^r}{(2r-1)!} \right)^k < d_k < \left( \frac{a_2}{a_1} \right)^{\frac{1}{r}} \left( \frac{a_2^r}{(2r-1)!} \right)^k
$$

and

$$
(n\epsilon_n^{-1} + \frac{a_2}{\epsilon_n})^k > |C_k|^k \frac{k!}{n\epsilon_n}.
$$

Hence, by (2.3) and (4.9),

$$
m_k(y_n)
$$

$$
< \left( \frac{\epsilon_n}{\sqrt{2\pi}} \right)^{\frac{1}{2}} \frac{(\alpha_2)^{\alpha_1}}{G(\alpha_1)} \left( \frac{a_2^{1+r}}{\sqrt{\alpha_1(2r-1)!}} \right)^k (n\epsilon_n)^{-\frac{r}{2}}
$$

$$
\times \int_0^\infty \int_0^\infty (\sigma^2)^{-(\frac{\epsilon_n}{\sqrt{2\pi}})^{2r-1+\alpha_1+1}} \exp \left\{ -\frac{R_1^* + 2\alpha_2}{2\sigma^2} \right\} \tau^{-r_k - \frac{\epsilon_n}{\sqrt{2\pi}} - \frac{\epsilon_n}{\sqrt{2\pi}} E_k(\prod_{i=1}^k \beta_{k_i}^2) d\sigma^2 d\tau.
$$

(4.10)

When $k \neq t$, by [10, Lemma 6], we get

$$
Q_k = E_k(\prod_{i=1}^k \beta_{k_i}^2) < \left( \frac{\epsilon_n^{-1} + \frac{a_2}{\epsilon_n}}{\epsilon_n} \right)^{\frac{1}{r}} \left( \frac{4V}{k} + [(2r-1)!]^\frac{1}{r} \frac{4\sigma^2}{\epsilon_n} \right)^{rk}
$$

$$
\leq \left( \frac{1 + \frac{a_2}{\epsilon_n}}{\epsilon_n^{2r-1}} \right)^{\frac{1}{r}} 2^{rk-1} \left( \frac{4V}{k} \right)^{rk} + \left( \frac{4\sigma^2(2r-1)!}{\epsilon_n} \right)^{rk}.
$$
Note that \( V \leq \frac{1}{n} y_n^T P_u y_n \) under Assumption 1, where \( P_u = X_u (X_u^T X_u)^{-1} X_u^T \).

Therefore, from (4.10), for large enough constants \( M_1, M_2 \), we get

\[
m_k(y_n) < \frac{\sqrt{n}}{(2\pi)^{\frac{k}{2}}} M_1 M_2 \left( \frac{V}{k} \right)^{rk} n^{-\frac{k}{2}} \epsilon_n^{-\frac{k}{2}}
\]

\[
\times \int_0^\infty \int_0^\infty (\sigma^2)^{-(\frac{d}{2}+rk+\alpha_1+1)} \exp \left\{ -\frac{R_k + 2\alpha_2}{2\sigma^2} \right\} \tau^{-\frac{k}{2} - \frac{d}{2} - \frac{\alpha_1}{2}} (1 + \frac{a_\epsilon n}{n\tau})^{\frac{k}{2}} d\sigma^2 d\tau
\]

\[
+ \frac{\sqrt{n}}{(2\pi)^{\frac{k}{2}}} M_1 M_2 \left( \frac{V}{k} \right)^{rk} n^{-\frac{k}{2}} \epsilon_n^{-\frac{k}{2}}
\]

\[
\times \int_0^\infty \int_0^\infty (\sigma^2)^{-(\frac{d}{2}+\alpha_1+1)} \exp \left\{ -\frac{R_k + 2\alpha_2}{2\sigma^2} \right\} \tau^{-\frac{k}{2} - \frac{d}{2} - \frac{\alpha_1}{2}} (1 + \frac{a_\epsilon n}{n\tau})^{\frac{k}{2}} d\sigma^2 d\tau.
\]

Since \( 1 + x \leq e^x \), integrating out both \( \tau \) and \( \sigma^2 \) gives us

\[
m_k(y_n) \leq \frac{\sqrt{n}}{(2\pi)^{\frac{k}{2}}} M_1 M_2 \left( \frac{V}{k} \right)^{rk} n^{-\frac{k}{2}} \epsilon_n^{-\frac{k}{2}}
\]

\[
\times \frac{\Gamma \left( r_k + \frac{k}{2} + \frac{1}{2} \right) \Gamma \left( \frac{n}{2} + r_k + \alpha_1 \right)}{\left( \frac{n}{2} - \frac{a_\epsilon k}{2n\alpha_1} \right)^{r_k+\frac{n}{2}+\frac{1}{2}}} \left\{ R_k + 2\alpha_2 \right\}^{\frac{1}{2}+r_k+\alpha_1}
\]

\[
+ \frac{\sqrt{n}}{(2\pi)^{\frac{k}{2}}} M_1 M_2 \left( \frac{V}{k} \right)^{rk} n^{-\frac{k}{2}} \epsilon_n^{-\frac{k}{2}}
\]

\[
\times \frac{\Gamma \left( r_k + \frac{k}{2} + \frac{1}{2} \right) \Gamma \left( \frac{n}{2} + \alpha_1 \right)}{\left( \frac{n}{2} - \frac{a_\epsilon k}{2n\alpha_1} \right)^{r_k+\frac{n}{2}+\frac{1}{2}}} \left\{ R_k + 2\alpha_2 \right\}^{\frac{1}{2}+\alpha_1}.
\]

Similarly for the true model \( t \), using (2.3) and (4.9), \( 1 + x \leq e^x \), and first integrating out \( \sigma^2 \), we get

\[
m_t(y_n) > \frac{\sqrt{n}}{(2\pi)^{\frac{k}{2}}} M_1 M_2 \left( n\epsilon_n \right)^{-\frac{d}{2}}
\]

\[
\times \int_0^\infty \int_0^\infty (\sigma^2)^{-(\frac{d}{2}+rt+\alpha_1+1)} \exp \left\{ -\frac{R_t + 2\alpha_2}{2\sigma^2} \right\} \tau^{-\frac{k}{2} - \frac{d}{2} - \frac{\alpha_1}{2}} \left( \frac{a_\epsilon n}{2n\tau} \right)^{\frac{k}{2}} d\sigma^2 d\tau
\]

\[
\geq \frac{\sqrt{n}}{(2\pi)^{\frac{k}{2}}} M_1 M_2 \left( n\epsilon_n \right)^{-\frac{d}{2}} \Gamma \left( \frac{n}{2} + rt + \alpha_1 \right) \left\{ R_t + 2\alpha_2 \right\}^{\frac{1}{2}+rt+\alpha_1+1}
\]

\[
\times \int_0^\infty \frac{1}{\left\{ R_t + 2\alpha_2 \right\}^{\frac{1}{2}+rt+\alpha_1+1}} \tau^{-\frac{k}{2} - \frac{d}{2} - \frac{\alpha_1}{2}} \left( \frac{a_\epsilon n}{2n\tau} \right)^{\frac{k}{2}} d\tau.
\]
Note that
\[
\{R_t + 2\alpha_2\}^{\frac{1}{2} + rt + \alpha_1} \leq \{R_t^* + 2\alpha_2\}^{\frac{1}{2} + rt + \alpha_1} \left\{1 + \frac{R_t - R_t^*}{R_t^*}\right\}^{\frac{1}{2} + rt + \alpha_1},
\]
and, by Assumptions 1 and 5,
\[
\frac{R_t - R_t^*}{R_t^*} \leq \frac{1}{R_t^*} y_n^T X_t (X_t^T X_t)^{-\frac{1}{2}} \left(I - \left(I + \frac{(X_t^T X_t) - \frac{1}{2} A_t (X_t^T X_t)^{-\frac{1}{2}}}{\tau}\right)^{-1}\right) (X_t^T X_t)^{-\frac{1}{2}} X_t^T y_n.
\]

From (4.12), we have
\[
m_t(y_n) > \frac{\sqrt{n}}{(\sqrt{2\pi})^n} M_1 M_2^*(n\epsilon_n)^{-\frac{1}{2} + rt + \alpha_1} \frac{2^{\frac{1}{2} + rt + \alpha_1}}{\{R_t^* + 2\alpha_2\}^{\frac{1}{2} + rt + \alpha_1}}
\times \int_0^\infty \exp \left[ - (n + rt + \alpha_1) \frac{\sigma_\alpha^2 \log n}{\epsilon_n + \tau \sigma_\alpha^2(n - t - \sqrt{n - t \log n})} \right] \tau^{rt - \frac{1}{2} - \frac{1}{2} \log \tau} \left(\frac{\sqrt{\pi} + \frac{\alpha_2 \log n}{\epsilon_n}}{2n^{-1/2}}\right)^{\frac{1}{2}} d\tau. \tag{4.13}
\]

Note that \(P_t \leq P_u\). By (4.2) and (4.5), there exists \(N_1\), such that for all \(n > N_1\),
\[
\left(\frac{n}{2} + rt + \alpha_1\right) \frac{\sigma_\alpha^2 \log n}{\epsilon_n + \tau} \frac{y_n^T P_t y_n}{R_t^*} \leq \frac{\sigma_\alpha^2 \log n}{\epsilon_n + \tau} \frac{n + rt + \alpha_1}{\sigma_\alpha^2(n - t - \sqrt{n - t \log n})} < \frac{a_2 \log n}{\epsilon_n + \tau}. \tag{4.14}
\]

Using (4.13) and (4.14), integrating out \(\tau\), we get
\[
m_t(y_n) > \frac{\sqrt{n}}{(\sqrt{2\pi})^n} M_1 M_2^*(n\epsilon_n)^{-\frac{1}{2}} \left(\frac{n}{2} + rt + \alpha_1\right) \frac{1}{\{R_t^* + 2\alpha_2\}^{\frac{1}{2} + rt + \alpha_1}}
\times \int_0^\infty \tau^{rt - \frac{1}{2} - \frac{1}{2} \log \tau} \left(\frac{\sqrt{\pi} + \frac{\alpha_2 \log n}{\epsilon_n}}{2n^{-1/2}}\right)^{\frac{1}{2}} d\tau
\geq \frac{\sqrt{n}}{(\sqrt{2\pi})^n} M_1 M_2^*(n\epsilon_n)^{-\frac{1}{2}}
\times \frac{\Gamma\left(n + rt + \frac{1}{2}\right) \Gamma\left(n + rt + \alpha_1\right)}{(n + \frac{a_2 \log n}{\epsilon_n} + \frac{a_2 \alpha_1}{2n\epsilon_n})} \frac{1}{\{R_t^* + 2\alpha_2\}^{\frac{1}{2} + rt + \alpha_1}} \frac{1}{2^{\frac{1}{2} + rt + \alpha_1}}. \tag{4.15}
\]
It follows from (4.11) and Assumption 1 that, the upper bound for \( m_k(y_n) \) will be given by,

\[
m_k(y_n) < B A^k \left( \frac{V}{\epsilon_n} \right)^{rk} n^{-\frac{k}{2}(k-1)} \frac{\Gamma \left( rk + \frac{k}{2} + \frac{1}{2} \right) \Gamma \left( \frac{\alpha}{2} + rk + \alpha_1 \right)}{\Gamma \left( \frac{\alpha}{2} + rt + \alpha_1 \right)}
\times \frac{\left( \frac{n}{2} + \frac{a_2 \log n}{c} + \frac{a_2 \sigma t}{2m} \right)^{rt+\frac{k}{2}+\frac{r}{2}}}{\left( \frac{n}{2} - \frac{a_2 \sigma k}{2m} \right)^{rk+\frac{k}{2}+\frac{r}{2}}}
\times \left\{ R_t^* + 2\alpha_2 \right\}^{\frac{1}{2}+rt+\alpha_1}
\times \left\{ R_k^* + 2\alpha_2 \right\}^{\frac{1}{2}+rk+\alpha_1},
\]

(4.16)

where \( A_1 \) and \( B_1 \) are large enough constants.

Using Stirling approximation, there exists \( N_2 \), such that for all \( n > N_2 \) and a large enough constant \( M' \), we have

\[
\frac{\Gamma \left( rk + \frac{k}{2} + \frac{1}{2} \right) \Gamma \left( \frac{k}{2} + rk + \alpha_1 \right)}{\Gamma \left( \frac{k}{2} + rt + \alpha_1 \right)} \leq M' k^{(r+1)k} n^{r(k-1)},
\]

and

\[
\frac{\Gamma \left( rk + \frac{k}{2} + \frac{1}{2} \right) \Gamma \left( \frac{k}{2} + rt + \alpha_1 \right)}{\Gamma \left( \frac{k}{2} + rt + \alpha_1 \right)} \leq M' k^{(r+1)k} n^{-rt}.
\]

By (4.16), there exists \( N = \max\{N_1, N_2\} \), such that for \( n > N \), we get

\[
m_k(y_n) < B A^k \left( \frac{V}{\epsilon_n} \right)^{rk} n^{-\frac{k}{2}(k-1)} \frac{\left\{ R_t^* + 2\alpha_2 \right\}^{\frac{1}{2}+rt+\alpha_1}}{\left\{ R_k^* + 2\alpha_2 \right\}^{\frac{1}{2}+rk+\alpha_1}}
\times \left\{ R_t^* + 2\alpha_2 \right\}^{\frac{1}{2}+rt+\alpha_1}
\times \left\{ R_k^* + 2\alpha_2 \right\}^{\frac{1}{2}+rk+\alpha_1},
\]

(4.17)

where \( A = A_1 \times M' \) and \( B = B_1 \times M' \).

In order to show the upper bound converges to 0, we need the following lemma in our proof. Let \( \delta = \frac{1}{2} \min_{j \in t} |\beta_0,j| \).

Lemma 4.4. When \( k \not\subset t \) and \( u = k \cup t \), \( \frac{R_t^* - R_k^*}{\sigma_0} \sim \chi^2_{u-k}(\lambda) \), where \( \lambda = \frac{\beta_0^T (X_u^T P_u X_u) \beta_0}{\sigma_0^2} > \frac{nt \delta^2}{\sigma_0^2} \) and \( P_* \) is idempotent.

Proof. First write \( X_u = [X_k, X_{u \setminus k}] \) and \( X_k = X_u W \), where \( W = (I_k, 0_{k \times (u-k)})^T \). We let

\[
P_* = P_u - P_k = X_u (X_u^T X_u)^{-1} X_u^T - X_k (X_k^T X_k)^{-1} X_k^T,
\]

where
From the assumptions of our theorem, we have that the non-centrality parameter $k$ under different cases of the “non-true” model $P$. Hence, $X_u (X^T u X_u)^{-1} X^T u W (W^T X^T u X_u W)^{-1} W^T X^T u = P_\ast$. Now further note that

$$P_\ast^2 = [X_u (X^T u X_u)^{-1} X^T u - X_u W (W^T X^T u X_u W)^{-1} W^T X^T u]^2$$

$$= X_u (X^T u X_u)^{-1} X^T u - X_u W (W^T X^T u X_u W)^{-1} W^T X^T u - X_u W (W^T X^T u X_u W)^{-1} W^T X^T u + X_u W (W^T X^T u X_u W)^{-1} W^T X^T u = P_\ast.$$ 

Hence, $P_\ast$ is idempotent and $\frac{R_\ast^2 - R_\ast^*}{\sigma_\ast^2} = \frac{\log P_u - \log P_k}{\sigma_\ast^2} \sim \chi_2^2(\lambda)$, where $d = \text{rank}(P_\ast) = \text{rank}(P_u) - \text{rank}(P_k) = u - k$ and $\lambda = \frac{\beta_0^2 (X^T u X_u) \beta_0}{\sigma_\ast^2}$. Next, note that

$$P_\ast X_u = P_u X_u - P_k X_u = X_u - [X_k, P_k X_{i \cap k^c}] = [0_{n \times k}, (I_n - P_k) X_{i \cap k^c}].$$

Therefore,

$$X^T u P_\ast X_u = (P_\ast X_u)^T P_\ast X_u = \begin{bmatrix} 0_{k \times k} & 0_{k \times k^c} \\ 0_{k^c \times k} & X^T u (I_n - P_k) X_{i \cap k^c} \end{bmatrix}.$$

Since $X^T u (I_n - P_k) X_{i \cap k^c}$ is the Schur complement of $X^T u X_k$, we have

$$\text{eig} (X^T u P_\ast X_u) > \text{eig} (X^T u X_k) > n \epsilon_n.$$ 

From the assumptions of our theorem, we have that the non-centrality parameter

$$\lambda = \frac{\beta_0^2 (X^T u P_\ast X_u) \beta_0}{\sigma_\ast^2} > \frac{n \epsilon_n \delta^2}{\sigma_\ast^2}.$$ 

Let $Z$ be a standard normal distribution. By Lemma 4.4, it follows from the relation between noncentral chi-squared and normal distribution that,

$$P \left( \exp \left[ \frac{(R_\ast^2 - R_\ast^*)}{6 \sigma_\ast^2} \right] > n^{-2rt} \right) \leq P \left( \frac{R_\ast^2 - R_\ast^*}{\sigma_\ast^2} < 12rt \log n \right) \leq P \left( (Z - \sqrt{\lambda})^2 < 12rt \log n \right) \leq P \left( Z > \sqrt{\lambda} - \sqrt{12rt \log n} \right) < e^{-\frac{n \epsilon_n \delta^2}{4 \sigma_\ast^2}} \to 0, \text{ as } n \to \infty.$$ 

The next two lemmas provide the upper bound of the posterior ratio a function of $n$ under different cases of the “non-true” model $k$. 


Lemma 4.5. Under Assumptions 1, 3 and 5, for all $k \not\supset t$, there exists $N$, such that when $n > N$,

$$\frac{m_k(y_n)}{m_t(y_n)} < K'(L')^k n^{-\frac{4}{3}rk},$$

(4.19)

where $K'$ and $L'$ are constants.

Proof. Recall that $u = k \cup t$ is a supset of $t$. Therefore, $R_u^* \leq R_t^*$. By (4.5) and Assumption 1, for large enough $n$, say $n > N_1$, we obtain,

$$\frac{V}{\epsilon_n^2} = 1 \beta_u^T \hat{\beta}_u \leq \frac{1}{n\epsilon_n^2} y^T P_u y < \frac{1}{\epsilon_n^2} \log n < n^{\frac{1}{\xi}}.$$

(4.20)

Consider the first term of (4.17). When $\max \left\{ \frac{4(r+1)t}{r}, \frac{4}{(1-\xi)^2} \right\} \leq k < n^\xi$, we have $(r + 1)t \leq \frac{4}{\xi}rk$ and $\sqrt{u - t} \leq \sqrt{k} \leq \frac{1}{\xi}k$. For large enough $n > N_1$, using (4.3), (4.4) and Assumption 1, we have

$$BA^k \left( \frac{V}{\epsilon_n^2} \right)^{rk} k^{k_n} n^{-(k-t)} \left\{ \frac{R_t^* + 2\sigma_2}{R_t^* + 2\sigma_1} \right\}^{\frac{4}{4} + rt + \alpha_1}$$

$$\leq BA^k \left( \frac{V}{\epsilon_n^2} \right)^{rk} k^{k_n} n^{-(k-t)} \left\{ \frac{R_t^* + 2\sigma_2}{R_t^* + 2\sigma_1} \right\}^{\frac{4}{4} + rt + \alpha_1}$$

$$\leq BA^k n^{\frac{1}{2} - \xi} n^\xi k^{n} n^{-(k-t)} \left( 1 + \frac{R_t^* - R_t^*}{R_t^* - R_t^*} \right)^{\frac{4}{4} + rt + \alpha_1} \frac{1}{(R_t^* + 2\sigma_2)^{rt(k-t)}}.$$

(4.21)

It follows from $1 + x \leq e^x$ and Assumption 1 that

$$BA^k \left( \frac{V}{\epsilon_n^2} \right)^{rk} k^{k_n} n^{-(k-t)} \left\{ \frac{R_t^* + 2\sigma_2}{R_t^* + 2\sigma_1} \right\}^{\frac{4}{4} + rt + \alpha_1}$$

$$\leq K_1 L_1 n^{\frac{1}{2} - \xi} n^\xi k^{n} n^{-(k-t)} \left\{ \frac{u - t}{n - u} \log n \left( \frac{n}{2} + rt + \alpha_1 \right) \right\}$$

$$\times \left( \frac{\sigma_0^2}{n - u - \sqrt{n - u} \log n + 2\sigma_2} \right)^{r(k-t)}$$

$$\leq K_1 L_1 n^{\frac{1}{2} - \xi} n^\xi k^{n} n^{-(k-t)} \left\{ \frac{u - t}{n - u} \log n - rt(k-t) \right\}$$

$$\leq K_1 L_1 n^{-\frac{4}{3}rk},$$

where $K_1$ and $L_1$ are constants.

When $k < \max \left\{ \frac{4(r+1)t}{r}, \frac{4}{(1-\xi)^2} \right\}$, it follows by the boundedness of $k$, (4.2), (4.3), (4.4)
and (4.20) that the first part of (4.17) will be bounded by

\[
BA^{k} \left( \frac{V}{c_{n}^{2}} \right)^{r_{k}} k^{k_{n}-(k-t)} \left\{ \frac{R_{t}^{*} + 2\alpha_{2}}{R_{k}^{*} + 2\alpha_{2}} \right\}^{-\frac{t}{2} + r_{t} + \alpha_{1}} ^{2r_{t} + k + \alpha_{1}}
\]

\[
\leq BA^{k} \left( \frac{V}{c_{n}^{2}} \right)^{r_{k}} k^{k_{n}-(k-t)} \left\{ \frac{R_{t}^{*} + 2\alpha_{2}}{R_{k}^{*} + 2\alpha_{2}} \right\}^{-\frac{t}{2} + r_{t} + \alpha_{1}} ^{2r_{t} + k + \alpha_{1}}
\]

\[
\times \left( n - t - \sqrt{n} \log n + 2\alpha_{2} \right)^{r_{t}-(k-t)} \left( 1 - \frac{R_{k}^{*} - R_{n}^{*}}{\sigma_{n}^{2}} \right)^{r_{t} + k + \alpha_{1}} \right).
\]

(4.22)

By \(1 + x \leq e^{x}\) and \(1 - x \leq e^{-x}\), we get

\[
BA^{k} \left( \frac{V}{c_{n}^{2}} \right)^{r_{k}} k^{k_{n}-(k-t)} \left\{ \frac{R_{t}^{*} + 2\alpha_{2}}{R_{k}^{*} + 2\alpha_{2}} \right\}^{-\frac{t}{2} + r_{t} + \alpha_{1}} ^{2r_{t} + k + \alpha_{1}}
\]

\[
\leq BA^{k}(\log n)^{2r_{k}} k^{k_{n}-(k-t)} \left\{ \frac{R_{t}^{*} + 2\alpha_{2}}{\sigma_{n}^{2}} \right\}^{-\frac{t}{2} + r_{t} + \alpha_{1}}
\]

\[
\times \exp \left\{ - \frac{R_{k}^{*} - R_{n}^{*}}{\sigma_{n}^{2}} \left( n - k + \sqrt{n} \log n + 2\alpha_{2} \right)^{r_{t} + k + \alpha_{1}} \right\}
\]

\[
\leq K_{2} L_{2}^{k}(\log n)^{2r_{k}} n^{k_{n}-(r+1)(k-t)} \exp \left\{ - \frac{R_{k}^{*} - R_{n}^{*}}{6\sigma_{n}^{2}} \right\}
\]

\[
< K_{2} L_{2}^{k}(\log n)^{2r_{k}} n^{k_{n}-(r+1)(k-t)} \exp \left\{ - \frac{R_{k}^{*} - R_{n}^{*}}{6\sigma_{n}^{2}} \right\}
\]

for large enough constants \(K_{2}, L_{2}\), since \((\log n)^{2r_{k}} < n^{r_{k}}\) for large enough \(n\), say \(n > N_{3}\), and \(\sqrt{u-t} \leq \sqrt{k} \leq k\).

Now consider the second term of (4.17). When \(\max \left\{ \frac{u-t}{\xi}, \frac{1}{(1-\xi)^{2}} \right\} \leq k \leq n^{\xi}\), we have \(\sqrt{u-t} \leq \sqrt{k} \leq (1-\xi)k\) and \((\frac{t}{2} + 1)t \leq (1-\xi)rt\). By (4.3), (4.4) and Assumption 3,
for large enough $n$, say $n > N_4$, we have

\[
BA^k n^{-r(k-t) - \frac{4}{3} r k - rt} \left\{ \frac{R_i^* + 2\alpha_2}{R_k^* + 2\alpha_2} \right\}^{\frac{3}{2} + rt + \alpha_1} \geq \frac{3}{2} + rt + \alpha_1
\]

\[
\leq B A^k n^{-r(k-t) - \frac{4}{3} r k - rt} \left\{ \frac{R_i^* + 2\alpha_2}{R_k^* + 2\alpha_2} \right\}^{\frac{3}{2} + rt + \alpha_1} (1 + \frac{\sigma_0^2}{\sigma_0^2}) \geq (R_i^* + 2\alpha_2)^{rt}. \tag{4.23}
\]

Similar to (4.21), we obtain,

\[
BA^k n^{-r(k-t) - \frac{4}{3} r k - rt} \left\{ \frac{R_i^* + 2\alpha_2}{R_k^* + 2\alpha_2} \right\}^{\frac{3}{2} + rt + \alpha_1} \leq B A^k n^{-r(k-t) - \frac{4}{3} r k - rt} \left\{ \frac{R_i^* + 2\alpha_2}{R_k^* + 2\alpha_2} \right\}^{\frac{3}{2} + rt + \alpha_1} \times \exp \left\{ \frac{(u-t) + \sqrt{u-t} \log n}{n - u - \sqrt{n-u} \log n} \right\}
\]

\[
\leq K_3 L_3 n^{-\frac{3}{2} r k},
\]

where $K_3$ and $L_3$ are large enough constants.

When $k < \max \left\{ \frac{\frac{2}{3} + 1}{1 + \xi}, \frac{1}{1 - \xi} \right\}$, using (4.18) and argument similar to (4.22), we have

\[
BA^k n^{-r(k-t) - \frac{4}{3} r k - rt} \left\{ \frac{R_i^* + 2\alpha_2}{R_k^* + 2\alpha_2} \right\}^{\frac{3}{2} + rt + \alpha_1} \leq B A^k n^{-r(k-t) - \frac{4}{3} r k - rt} \left\{ \frac{R_i^* + 2\alpha_2}{R_k^* + 2\alpha_2} \right\}^{\frac{3}{2} + rt + \alpha_1} \times \exp \left\{ \frac{(u-t) + \sqrt{u-t} \log n}{n - u - \sqrt{n-u} \log n} \right\}
\]

\[
\leq B A^k n^{-r(k-t) - \frac{4}{3} r k - rt} \left\{ \frac{R_i^* + 2\alpha_2}{R_k^* + 2\alpha_2} \right\}^{\frac{3}{2} + rt + \alpha_1} \times \left( 1 + \frac{\sigma_0^2}{\sigma_0^2} \right) \geq (R_i^* + 2\alpha_2)^{rt}. \tag{4.23}
\]

By $1 + x \leq e^x$, we get,

\[
BA^k n^{-r(k-t) - \frac{4}{3} r k - rt} \left\{ \frac{R_i^* + 2\alpha_2}{R_k^* + 2\alpha_2} \right\}^{\frac{3}{2} + rt + \alpha_1} \leq B A^k n^{-r(k-t) - \frac{4}{3} r k - rt} \exp \left\{ \frac{(u-t) + \sqrt{u-t} \log n}{n - u - \sqrt{n-u} \log n} \right\}
\]

\[
\leq B A^k n^{-r(k-t) - \frac{4}{3} r k - rt} \exp \left\{ \frac{(u-t) + \sqrt{u-t} \log n}{n - u - \sqrt{n-u} \log n} \right\} \left\{ \frac{\sigma_0^2}{\sigma_0^2} \right\}. \tag{4.23}
\]
\[ \times (n - t - \sqrt{n} \log n - 2 \alpha_2) \left( \frac{R_k^* - R_*}{\sigma_0^*} \right) \exp \left\{ \frac{R_k^* - R_*}{\sigma_0^*} \left( \frac{n}{2} + \alpha_1 \right) \right\} \]

\[ \leq K_4 L_4^k n^{-(r+1)(k-t) - \frac{2}{1-\xi} r k - r t} n^2 \left( \frac{n}{2} \log n - 2 \alpha_2 \right) \]

\[ \leq K_4 L_4^k n^{\frac{2}{1-\xi} r k} n^2 \left( \frac{n}{2} \log n - 2 \alpha_2 \right) \]

where \( K_4 \) and \( L_4 \) are large enough constants, for large enough \( n \), say \( n > N_6 \).

Combining all cases and letting \( N = \max \{ N_1, N_2, N_3, N_4, N_5, N_6 \} \), we get that

\[ \frac{m_k(y_n)}{m_t(y_n)} < K'(L')^k n^{\frac{-2}{1-\xi} r k}, \]

where \( K' = \max \{ K_1, K_2, K_3, K_4 \} \) and \( L' = \max \{ L_1, L_2, L_3, L_4 \} \) for all \( n > N \).

**Lemma 4.6.** Under Assumptions 1, 3 and 5, for all \( k \geq t \), there exists \( N \), such that when \( n > N \),

\[ \frac{m_k(y_n)}{m_t(y_n)} < S'(T')^{k-t} n^{-\min \{ \frac{-2}{1-\xi}, r(k-t) \}}, \]

where \( S' \) and \( T' \) are constants.

**Proof.** Similar to (4.20), for large enough \( n > N_1 \), we have \( V < n^\frac{2}{1-\xi} r t \). Next, consider the first term in (4.17). When \( \max \left\{ \frac{2((1+\xi) t)^4 r t}{1 - (1-\xi) r}, \max \{ r k, r k \} \right\} \leq k \leq n^\frac{2}{1-\xi} r t \), we obtain \( \frac{4}{1-\xi} r k \geq \frac{1}{2} (k - t) + r t \) and \( \sqrt{k - t} \leq \frac{1}{2} (k - t) \). For large enough \( n \), say \( n > N'_2 \), when \( k \geq t \), using (4.3), (4.4) and Assumption 3, we get

\[ BA^k \left( \frac{V}{\epsilon_n^2} \right)^{\frac{1}{2} k} n^{-(k-t)} \left\{ \frac{R_k^* + 2 \alpha_2}{R_k^*} \right\}^{\frac{2}{1-\xi} r t + \alpha_1} \left\{ \frac{R_k^* + 2 \alpha_2}{\sigma_0^*} \right\}^{\frac{2}{1-\xi} r k + \alpha_1} \]

\[ \leq BA^k n^\frac{1}{2} k n^{-\frac{1}{2} k} n^{-(k-t)} \left( 1 + \frac{R_k^* - R_*}{\sigma_0^*} \right)^{\frac{2}{1-\xi} r t + \alpha_1} \left( \frac{R_k^* + 2 \alpha_2}{\sigma_0^*} \right)^{\frac{2}{1-\xi} r k + \alpha_1} \]

\[ \leq S_1 T_1^k n^\frac{1}{2} k n^{-(k-t)} n^{r k} \exp \left\{ \left( k - t \right) + \sqrt{k - t} \log n \left( \frac{n}{2} + r t + \alpha_1 \right) \right\} \]

\[ \times \left( \frac{n - k - \sqrt{n - u \log n + 2 \alpha_2}}{\sigma_0^*} \right) \left( \frac{n - k - \sqrt{n - u \log n + 2 \alpha_2}}{\sigma_0^*} \right)^{r(k-t)} \]

\[ \leq S_1 T_1^k n^\frac{1}{2} k n^{-(k-t)} n^{\left( k - t \right) r(k-t)} \]

\[ \leq S_1 T_1^k n^{\frac{-2}{1-\xi} r k}, \]

where \( S_1 \) and \( T_1 \) are large enough constants.

When \( t < k < \max \left\{ \frac{2((1+\xi) t)^4 r t}{1 - (1-\xi) r}, \max \{ r k, r k \} \right\} \), by the boundedness of \( k \), (4.2), (4.3) and
for large enough constants $S_2, T_2$, since $\sqrt{k-t} \leq k-t$ and $(\log n)^{2rk} < n^{1/r} \leq n^{4r(k-t)}$ for large enough $n$, say $n > N_3'$. Now for the second part of (4.17), when $\max \left\{ \frac{16}{t^2} + t, \frac{1+4r}{r} t \right\}$ is bounded by $\frac{k-t}{4}(k-t)$ and $(1+\frac{4}{r}) t \leq \frac{1}{r} k$. Again using (4.3), (4.4) and Assumption 3, for large enough $n$, say $n > N_4'$, we get

\[
BA^k k^{(r+1)k} n^{-(r+1)(k-t)-\frac{4}{r} r-k rt} \left\{ \frac{R_t^* + 2\alpha_2}{R_k^* + 2\alpha_2} \right\}^{\frac{1}{2} + rt + \alpha_1}
\leq BA^k n^{\xi(r+1)k} n^{-(r+1)(k-t)-\frac{4}{r} r-k rt} \left( 1 + \frac{\sigma_n^2}{\sigma_0^2} \right)^{\frac{1}{2} + rt + \alpha_1} (R_k^* + 2\alpha_2)^rt
\leq BA^k n^{\xi(r+1)k} n^{-(r+1)(k-t)-\frac{4}{r} r-k rt} \exp \left\{ \left( \frac{k-t} {n-k} - \frac{n-k-\log n}{n-k} \right)^{\frac{n}{2} + rt + \alpha_1} \right\}
\times (n-k) r \leq \sqrt{n-k} \log n + 2\alpha_2 \right\}^{rt}
\leq S_3 T_3^k n^{-r(k-t)-\frac{4}{r} r-k rt} n^{r t n^t}
\leq S_3 T_3^k n^{-(1-\xi)(r+1)k},
\]

for large enough constants $S_3$ and $T_3$. When $t < k < \max \left\{ \frac{16}{t^2} + t, \frac{1+4r}{r} t \right\}$, following the exact arguments in (4.27), using the boundedness of $k$ and $\sqrt{k-t} \leq k-t$, it follows that for large enough $n$, say $n > N_5'$, the second part of (4.17) is bounded by $S_4 T_4^k n^{-(k-t)}$, for large enough constants $S_4$ and $T_4$. Combining all cases and letting $N = \max\{N_1, N_2', N_3', N_4', N_5'\}$,

\[
\frac{m_k(y_n)}{m_k(y_n)} < S'(T^r k-t) n^{-\min \left\{ \frac{1}{2}, 1-\xi \right\} r(k-t)},
\]
for \( n > N \), where \( S' = \max\{S_1, S_2, S_3, S_4\} \) and \( T' = \max\{T_1, T_2, T_3, T_4\} \).

The proof of Theorem 3.1 will follow immediately from these two lemmas. By Lemma 4.5, if we restrict to \( C_n^c \), for any \( k \neq t \), if \( k \nsubseteq t \),

\[
\frac{m_k(y_n)}{m_t(y_n)} < K'(L')^kn^{-\frac{2}{r}r} \to 0, \text{ as } n \to \infty.
\]

Otherwise, when \( k \supset t \),

\[
\frac{m_k(y_n)}{m_t(y_n)} < S'(T')^{k-t}n^{-\min\{\frac{3}{4}, 1-\xi\}}r^{(k-t)} \to 0, \text{ as } n \to \infty.
\]

Note that \( P(C_n^c) \to 1 \) as \( n \to \infty \). Following from (2.2) and Assumption 4, when \( k \neq t \), we have

\[
\frac{\pi(k|y_n)}{\pi(t|y_n)} \leq \frac{1}{\omega} \frac{m_k(y_n)}{m_t(y_n)} \to 0, \text{ as } n \to \infty.
\]

(4.30)

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

\[
\min\left\{\frac{3}{4}, 1-\xi\right\}r > \frac{3}{4}(1-\xi)r > \gamma.
\]

(4.31)

It follows from (4.30) and Assumption 2 that if we restrict to \( C_n^c \), then

\[
\frac{1-\pi(t|y_n)}{\pi(t|y_n)} = \sum_{k \neq t} \frac{\pi(k)m_k(y_n)}{\pi(t)m_t(y_n)} \\
\leq \frac{1}{\omega} \sum_{k \nsubseteq t} \frac{m_k(y_n)}{m_t(y_n)} + \frac{1}{\omega} \sum_{k \supset t} \frac{m_k(y_n)}{m_t(y_n)} \\
\leq \frac{1}{\omega} \sum_{k=1}^{q_n} \binom{p}{k} K'(L')^kn^{-\frac{2}{r}r} \\
+ \frac{1}{\omega} \sum_{k-t=1}^{q_n-t} \binom{p-t}{k-t} S'(T')^{k-t}n^{-\min\{\frac{3}{4}, 1-\xi\}}r^{(k-t)}.
\]

Using \( \binom{p}{k} \leq p^k \) and (4.31), we get

\[
\frac{1-\pi(t|y_n)}{\pi(t|y_n)} \to 0, \text{ i.e. } \pi(t|y_n) \to 1,
\]

as \( n \to \infty \).
5. Computation

The integral formulation in (2.2) is quite complicated, and hence the posterior probabilities can not be obtained in closed form. Hence, we use Laplace approximation to compute \( m_k(y_n) \) and \( \pi(k | y_n) \). A similar approach to compute posterior probabilities has been used in [9] and [15].

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

\[
\log \frac{m_k(y_n)}{\pi(y_n | \sigma^2)} = \log \left( \pi(y_n | \sigma^2) \pi(\beta | \tau, \sigma^2) \right) - \left( \frac{rk + \frac{k}{2}}{2} \right) \log \tau - \left( \frac{rk + \frac{k}{2} + \alpha_1 + 1}{2} \right) \log(\sigma^2) - \left( \frac{(y_n - X_k \beta_k)^T (y_n - X_k \beta_k)}{2\sigma^2} \right) - \left( \frac{\beta_k^T \beta_k}{2\tau \sigma^2} + \frac{\alpha_2}{\sigma^2} + \frac{r}{2\tau} \right) + \sum_{i=1}^{k} 2r \log(\beta_k_i)
\]

(5.1)

For any model \( k \), the Laplace approximation of \( m_k(y_n) \) is given by

\[
(2\pi)^{\frac{k}{2} + 1} \exp \left\{ \log(\hat{\beta}_k, \hat{\tau}, \hat{\sigma}^2) \right\} |V(\hat{\beta}_k, \hat{\tau}, \hat{\sigma}^2)|^{-\frac{k}{2}},
\]

(5.2)

where \((\hat{\beta}_k, \hat{\tau}, \hat{\sigma}^2) = \arg\max_{(\beta, \tau, \sigma^2)} f(\beta, \tau, \sigma^2)\), and \( V(\hat{\beta}_k, \hat{\tau}, \hat{\sigma}^2) \) is a \((k + 2) \times (k + 2)\) symmetric matrix with the following blocks:

\[
V_{11} = \frac{1}{\tau \sigma^2} I_k + \frac{1}{\sigma^2} X_k^T X_k + \text{diag} \left\{ \frac{2r}{\beta_k^2}, \ldots, \frac{2r}{\beta_k^2} \right\},
\]

\[
V_{12} = -\frac{\beta_k^T \beta_k}{2\tau \sigma^2},
\]

\[
V_{13} = -\frac{\beta_k X_k \beta_k - X_k^T y_n}{\sigma^4},
\]

\[
V_{22} = -\frac{r k + \frac{k}{2} + \frac{3}{2}}{\tau^2} + \frac{\beta_k^T \beta_k}{\tau^3 \sigma^2} + \frac{n}{\tau^3},
\]

\[
V_{23} = \frac{\beta_k^T \beta_k}{2\tau^2 \sigma^4},
\]

\[
V_{33} = -\frac{r k + \frac{k}{2} + \frac{3}{2} + \alpha_1 + 1}{\sigma^4} + \frac{\beta_k^T \beta_k}{\tau \sigma^6} + \frac{(y_n - X_k \beta_k)^T (y_n - X_k \beta_k)}{4\sigma^6} + \frac{2\alpha_2}{\sigma^6},
\]

(5.3)

The above Laplace approximation can be used to compute the log of the posterior probability ratio between any given model \( k \) and true model \( t \), and select a model \( k \) with the highest probability.

6. Experiments
6.1. 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 [3] 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 \( \beta_0 = (1.1, 1.2, 1.3, \ldots, 1.9, 2) \)' 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 \( N(0, \Sigma) \) distribution as rows of the covariate matrix \( X \). We then examine posterior ratio consistency under three different cases of \( \Sigma \) by computing the log posterior ratio of a “non-true” model \( k \) and \( t \) 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 \( \Sigma_{ij} = 0.5^{\left| i-j \right|} \), 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 is provided in Table 6.1. Note that for each of these cases, we compute the log ratio under four different scenarios of “non-true” model \( k \).

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

In Table 6.1, columns 3, 4, 5, 6 correspond to Scenarios 1, 2, 3 and 4 respectively. As expected the log of the posterior probability ratio for any “non-true” model \( k \) compared to the true model \( t \) are all decreasing to large negative values as \( p \) increases, thereby providing a numerical illustration of Theorems 3.1 and 3.2.

6.2. Simulation II: Illustration of model selection

In this section, we perform a simulation experiment to illustrate the potential advantages of using our Bayesian approach. We also consider several different values of \( p \) ranging from 500 to 3000, while \( n = p/5 \). For each fixed \( p \), \( \beta_0 \) is generated by the same mechanism as in Section 6.1. Next, we generate \( n \) i.i.d. observations from the \( N(0_p, \Sigma) \) distribution as rows of covariate matrix \( X \) under the following three cases similar to Section 6.1.

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 \( \Sigma_{ij} = 0.5^{\left| i-j \right|} \), 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 (5.2), and exploring the model space using the simplified stochastic shotgun stochastic search algorithm in [15].

We would like to remind the readers that in our model, we don’t need to specify a fixed value for $\tau$, but rather put a prior on the parameter $\tau$ (as opposed to [9] and [15] when $\tau$ is treated as a fixed parameter). We also provide model selection performance results with fixed $\tau$ at $\tau = n$ (the reciprocal of the expected value of $1/\tau$ under our prior), and numerical values in R package BayesS5 (a choice for fixed $\tau$ from the results in [15]). Additionally, we also provide model selection performance results for the Lasso and SCAD penalized likelihood methods.

It is clear that our Bayesian approach outperforms both the penalized likelihood approaches and the fixed $\tau$ setting based on almost all measures and under all cases. The TPR values for the other three methods except the fixed $\tau$ setting are all very close to 1, which means that the Bayesian approach with fixed $\tau$ values capture much fewer true indexes than the other three methods. Also, the PPV values for our Bayesian approach are all higher than Lasso and SCAD methods, though just slightly lower than the fixed $\tau$ case. 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 the numerical procedure for choosing $\tau$ implemented in BayesS5 needs additional run time, while in

| $p$ | $n$ | $|k| = \frac{1}{2} |t|$ | $|k| = 2 |t|$ | $|k| = \frac{1}{4} |t|$ | $|k| = 2 |t|$ |
|-----|-----|----------------|----------------|----------------|----------------|
| 500 | 100 | -60.1710       | -88.4197       | -123.9463      | -300.8400      |
| 1000| 200 | -148.1302      | -142.7311      | -268.6699      | -455.6704      |
| 1500| 300 | -210.0730      | -170.5842      | -364.0275      | -580.5925      |
| 2000| 400 | -326.7857      | -172.8720      | -445.2006      | -749.7216      |
| 2500| 500 | -440.5580      | -194.7749      | -563.8925      | -870.3250      |
| 3000| 600 | -476.2002      | -203.3338      | -722.3714      | -1036.5460     |

Table 1. Log of posterior probability ratio for $k$ and $t$ for various choices of the “non-true” model $k$. 
Top: case 1; middle: case 2; bottom: case 3.
Table 2. Model selection performance table when $n=200$. Top: case 1; middle: case 2; bottom: case 3.

| Lasso | SCAD | BayesS5 | Bayesian |
|-------|------|---------|----------|
| p PPV TPR FPR | PPV TPR FPR | PPV TPR FPR | PPV TPR FPR |
| 100 0.3192 0.2600 0.8892 1 0.0189 1 1 0 1 0.4000 1 1 0 | 0.3278 0.2506 0.9568 1 0.0061 1 1 0 1 0.5200 0 1 1 0 | 0.3482 0.2506 0.9568 1 0.0061 1 1 0 1 0.3900 0 0.9955 1 0.0006 |
| 200 0.2865 0.1466 0.9271 1 0.0040 1 1 0 1 0.6800 0 1 1 0 | 0.2567 0.1716 0.9122 1 0.0066 1 1 0 1 0.6400 0 1 1 0 | 0.3482 0.2181 0.9242 1 0.0128 1 1 0 1 0.8400 0 1 1 0 |
| 500 0.1942 0.0942 0.9865 1 0.0040 1 1 0 1 0.6800 0 1 1 0 | 0.1938 0.0964 0.9466 1 0.0011 1 0.9250 0 1 0.4800 0 0.9818 1 0.0001 | 0.1938 0.0964 0.9466 1 0.0011 1 0.9250 0 1 0.4800 0 0.9818 1 0.0001 |
| 1000 0.1868 0.0497 0.8701 1 0.0021 1 1 0 1 0.6400 0 1 1 0 | 0.1877 0.0330 0.9000 1 0.0011 1 0.9600 0 1 0.7000 0 0.9742 1 0.0002 | 0.1877 0.0330 0.9000 1 0.0011 1 0.9600 0 1 0.7000 0 0.9742 1 0.0002 |
| 1500 0.1824 0.0344 0.8484 1 0.0019 1 1 0 1 0.5650 0 1 1 0 | 0.1799 0.0518 0.9199 1 0.0011 1 0.9555 0 1 0.5000 0 0.9756 1 0.0003 | 0.1958 0.1016 0.9897 1 0.0002 1 1 0 1 0.6800 0 1 1 0 |
| 2000 0.1729 0.0288 0.8155 1 0.0019 1 1 0 1 0.5500 0 1 1 0 | 0.1595 0.0385 0.9269 1 0.0005 1 0.9000 0 0.9750 0 0.4500 0 0.9750 1 0.0003 | 0.1595 0.0385 0.9269 1 0.0005 1 0.9000 0 0.9750 0 0.4500 0 0.9750 1 0.0003 |
| 2500 0.1395 0.0288 0.9010 1 0.0009 1 0.9756 0 1 0.5800 0 0.9909 1 0 | 0.1424 0.0135 0.5073 1 0.0004 1 0.9600 0 1 0.4500 0 0.9909 1 0 | 0.1424 0.0135 0.5073 1 0.0004 1 0.9600 0 1 0.4500 0 0.9909 1 0 |
| 5000 0.1428 0.0135 0.9073 1 0.0004 1 0.9909 1 1 0.4450 0 1 1 0 | 0.1428 0.0135 0.9073 1 0.0004 1 0.9909 1 1 0.4450 0 0.9909 1 0 | 0.1428 0.0135 0.9073 1 0.0004 1 0.9909 1 1 0.4450 0 0.9909 1 0 |

References

[1] Yuanyuan Bian and Ho-Hsiang Wu. A note on nonlocal prior method. https://arxiv.org/abs/1702.07778, 2017.
[2] H. D. Bondell and B. J. Reich. Consistent high-dimensional bayesian variable selection via penalized credible regions. J. Amer. Statist. Assoc, 107(500):1610–1624, 2012.
[3] X. Cao, K. Khare, and M. Ghosh. Posterior graph selection and estimation consistency for high-dimensional bayesian dag models. https://arxiv.org/abs/1611.01205, 2017.
[4] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc, 96(456):1348–1360, 2001.
Posterior consistency for non-local priors in regression

[5] E. I. George and D. P. Foster. Calibration and empirical bayes variable selection. *Biometrika*, 87(4):731–747, 2000.

[6] E. I. George and R. E. McCulloch. Variable selection via gibbs sampling. *J. Amer. Statist. Assoc.*, 88(423):881–889, 1993.

[7] H. Ishwaran, U. B. Kogalur, and J. S. Rao. Spike and slab variable selection: Frequentist and bayesian strategies. *Ann. Statist.*, 33(2):730–773, 04 2005.

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

[9] V. Johnson and D. Rossell. Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.*, 107(498):649–660, 2012.

[10] V. Johnson and D. Rossell. Supplementary material to “bayesian model selection in high-dimensional settings”. *J. Amer. Statist. Assoc.*, 2012.

[11] F. Liang, R. Paulo, G. Molina, A. M. Clyde, and O. J. Berger. Mixtures of g priors for bayesian variable selection. *J. Amer. Statist. Assoc.*, 103(481):410–423, 2008.

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

[13] N. Narisetty and X. He. Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.*, 42:789–817, 2014.

[14] David Rossell, Donatello Telesca, and Valen E. Johnson. *High-Dimensional Bayesian Classifiers Using Non-Local Priors*, pages 305–313. Springer International Publishing, Heidelberg, 2013.

[15] M. Shin, A. Bhattacharya, and V. Johnson. Scalable bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *arXiv*, 2015.

[16] Q. Song and F. Liang. High-dimensional variable selection with reciprocal l1-regularization. *J. Amer. Statist. Assoc.*, 110(512):1607–1620, 2015.

[17] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

[18] Yun Yang, Martin J. Wainwright, and Michael I. Jordan. On the computational complexity of high-dimensional bayesian variable selection. *Ann. Statist.*, 44(6):2497–2532, 12 2016.

[19] C-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 04 2010.

[20] H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.