A Mixture of $g$-priors for Variable Selection when the Number of Regressors Grows with the Sample Size

Minerva Mukhopadhyay
Indian Statistical Institute & Bethune College
minervamukherjee@gmail.com

Abstract: We consider variable selection problem in linear regression using mixture of $g$-priors. A number of mixtures are proposed in the literature which work well, especially when the number of regressors $p$ is fixed. In this paper, we propose a mixture of $g$-priors suitable for the case when $p$ grows with the sample size $n$. We study the performance of the method based on the proposed prior when $p = O(n^b)$, $0 < b < 1$. Along with model selection consistency, we also investigate the performance of the proposed prior when the true model does not belong to the model space considered. We find conditions under which the proposed prior is consistent in appropriate sense when normal linear models are considered. Further, we consider the case with non-normal errors in the regression model and study the performance of the model selection procedure. We also compare the performance of the proposed prior with that of several other mixtures available in the literature, both theoretically and using simulated data sets.

Keywords and phrases: Model false case, model selection consistency, model true case, normal linear models, scaled inverse chi-square prior.

1. Introduction

We consider the regression problem with a response variable $y$ and a set of $p$ potential regressors $x_1, x_2, \ldots, x_p$. Let $y_n = (y_1, y_2, \ldots, y_n)$ be a set of $n$ observations on $y$, and $X_n = (x_1, x_2, \ldots, x_p)$ be the $n \times p$ design matrix, where $x_i$ is the vector of $n$ observations on the $i^{th}$ regressor $x_i, i = 1, 2, \ldots, p$. We write

$$y_n = \mu_n + e_n,$$  \hspace{1cm} (1.1)
where $\mu_n = E(y_n|X_n)$ is the regression of $y_n$ on $X_n$ and $e_n$ is the vector of random errors. If we assume the normal linear regression model, then $\mu_n = \beta_0 1 + X_n \beta$ and $e_n \sim N_n(0, \sigma^2 I)$. Here $\beta_0$ is the intercept, 1 and 0 are the $n \times 1$ vectors of ones and zeros, respectively, and $\beta' = (\beta_1, \beta_2, \ldots, \beta_p)$ is the vector of regression coefficients.

In this article, we study the variable selection problem. Given a set of $p$ available regressor variables, there are $2^p$ possible linear regression models depending on which regressors are included in the model. The space of all these models is denoted by $A$ and indexed by $\alpha$, where each $\alpha$ consists of a subset of size $p(\alpha)$ ($0 \leq p(\alpha) \leq p$) of the set $\{1, 2, \ldots, p\}$, indicating which regressors are selected in the model. The $\alpha^{th}$ model $M_\alpha$ is stated as

$$M_\alpha : \mu_n = \beta_0 1 + X_\alpha \beta_\alpha,$$

where $X_\alpha$ is a sub-matrix of $X_n$ consisting of the $p(\alpha)$ columns specified by $\alpha$ and $\beta_\alpha$ is the corresponding vector of regression coefficients. We assume that all the components of $\beta_\alpha$ are non-zero. This ensures there is at most one true model in the model space. Our purpose is to choose the model $\alpha \in A$, which best explains the data.

In a Bayesian approach, each model $M_\alpha$ is associated with a prior probability $p(M_\alpha)$ and the corresponding set of parameters $\theta_\alpha = (\beta_0, \beta_\alpha, \sigma^2)$ involved in the model, is also associated with a prior distribution $p(\theta_\alpha|M_\alpha)$. Given the priors, one computes the posterior probability of $M_\alpha$ as

$$p(M_\alpha|y_n) = \frac{p(M_\alpha)m_\alpha(y_n)}{\sum_{\alpha \in A} p(M_\alpha)m_\alpha(y_n)},$$

where

$$m_\alpha(y_n) = \int p(y_n|\theta_\alpha, M_\alpha)p(\theta_\alpha|M_\alpha)d\theta_\alpha$$

is the marginal density of $y_n$ under $M_\alpha$ and $p(y_n|\theta_\alpha, M_\alpha)$ is the density of $y_n$ given $\theta_\alpha$ under $M_\alpha$. In our search for a model, $p(y_n|\theta_\alpha, M_\alpha)$ will be taken to be normal. We consider the model selection procedure that selects the model in $A$ with the highest posterior probability.

We consider a prevalent conventional prior on $\beta_\alpha$, the $g$-prior due to Zellner (1986). Properties of $g$-prior are studied extensively in the literature (see, e.g., George and Foster (2000), Berger and Pericchi (2001), Fernández, Ley and Steel (2001)). The prior specification induced by the $g$-prior method crucially
depends on the choice of the hyperparameter $g$ (see, e.g., Berger and Pericchi (2001), Liang et al. (2008)). It has also been argued in the literature that this method is subject to inconsistencies like Bartlett paradox (see Bartlett (1957), Jeffreys (1961)) and information paradox (see Zellner (1986), Berger and Pericchi (2001)). Liang et al. (2008) considered a mixture on $g$ instead of considering a fixed $g$ to overcome these inconsistencies. Subsequently, a number of mixtures on $g$ are proposed in the literature. In this paper, we propose a mixture $\pi(g)$ on $g$, suitable for the case when $p$ grows with $n$.

We assume without loss of generality that the columns of $X_\alpha$ are centered, so that $1^T x_i = 0$ for all $i$. The intercept $\beta_0$ and the scale parameter $\sigma^2$ are common to all models and assumed to be independent of other parameters.

We use standard non-informative priors for $\beta_0$ and $\sigma^2$. For justification of using such priors, see Bayarri et al. (2012, Sec 3.3). The vector $\beta_\alpha$ is assumed to follow a normal distribution with location parameter $0$ and scale $g\sigma^2 (X'_\alpha X_\alpha)^{-1}$, where the hyperparameter $g$ has density $\pi(g)$. The complete prior specification is given by

$$p(\beta_0, \sigma^2|M_\alpha) = \frac{1}{\sigma^2}, \quad \beta_\alpha| (\beta_0, \sigma^2, g, M_\alpha) \sim N_{p(\alpha)}(0, g\sigma^2 (X'_\alpha X_\alpha)(1,\beta))$$

and $g \sim \pi(g)$.

We do not consider any specific prior probability on the model space. We only impose some conditions on model prior probabilities under which our results hold. Similar setup has been considered by many authors, see, e.g., Liang et al. (2008), Bayarri et al. (2012).

Among the existing mixtures of $g$-priors, the earliest one, to the best of our knowledge, is due to Zellner and Siow (1980), who proposed a Cauchy prior on $\beta_\alpha$. Since Cauchy is an inverse gamma scale mixture of normal distributions, their prior proposition is considered as a mixture of $g$-priors. Other priors include hyper-$g$ and hyper-$g/n$ priors proposed by Liang et al. (2008), generalized $g$-prior of Maruyama and George (2011) and robust prior proposed by Bayarri et al. (2012). Henceforth, we will refer to these priors as Zellner-Siow prior, hyper-$g$ or hyper-$g/n$ prior, generalized $g$-prior and robust prior, respectively.

Inspite of existence of the above mixtures of $g$-priors there is not much discussion in the literature on which mixture one should use in a given situation. Bayarri et al. (2012) described some desirable properties a prior should satisfy in the context of Bayesian model selection. Ley and Steel (2012) made an extensive simulation study to compare several priors. However, none of them
considered the case when $p$ increases with $n$. Maruyama and George (2011) proposed a prior which is applicable when $p > n$, but proved consistency of their method for the case when $p$ is fixed. Shang and Clayton (2011) proved consistency for mixture of $g$-priors when $p$ increases with $n$ but their setup differs from the usual $g$-prior setup with respect to the covariance structure of the prior distribution of $\beta_n$. Wang and Sun (2014) investigated properties of different mixtures for the case with growing number of regressors. However, they only established results for Bayes factor consistency.

In this paper, we consider a scaled inverse chi-square prior on $g$ with appropriate parameters. The Zellner-Siow prior belongs to this family. In a sense we consider a modified form of Zellner-Siow prior by choosing an appropriate scale parameter. An advantage of this prior is that it provides an approximation to the marginal density in (1.4) with a closed-form expression which facilitates easy implementation and theoretical studies. Further, it satisfies many attractive consistency properties when $p$ increases with $n$. Most of the existing mixtures fail to be consistent for such rates of increase of $p$. The good properties of this prior are not restricted to the case where the error distribution is normal. For a general class of error distributions with minimal assumptions, the proposed prior performs reasonably well in terms of consistency. In Section 2, we explicitly describe the form of the prior and the motivation for considering the same.

In our investigation, we assume that the number of regressors $p$ increases with $n$ at a rate $p = O(n^b)$, $0 < b < 1$ and $p < n$. This is the so called ‘large $p$ large $n$ regime’ and is of theoretical interest in contemporary research (see, e.g., Fan and Peng (2004), Moreno, Girón and Casella (2010), Sparks, Khare and Ghosh (2012), Johnson and Rossell (2012)). In practice, one can conveniently use methods applicable in ‘large $p$ large $n$ regime’ when there is a sizable number of regressors compared to $n$ and the number of competing models in the model space is significantly large compared to $n$.

We show that our proposed prior is consistent in appropriate sense for a large class of models under reasonable assumptions. We consider the following two cases separately. First, we consider the case when the true model belongs to the model space $\mathcal{A}$, i.e., the true regression $\mu_n$ is as in (1.2) for some $\alpha$. This case is referred to as the ‘model true’ case. A well known notion of consistency in this regard is model selection consistency which requires the posterior probability of the true model to go to one as $n \to \infty$. We examine model selection consistency of the proposed mixture of $g$, along with that of some other existing mixtures. We then consider the case when $\mu_n$ can be any
unknown vector, not necessarily in the span of \( \{1, x_1, x_2, \ldots, x_p\} \). This case is referred to as the ‘model false’ case. ‘Model false’ case has been previously studied in Shao (1997), Chakrabarti and Ghosh (2006), Chakrabarti and Samanta (2008) and Mukhopadhyay, Samanta and Chakrabarti (2014). We investigate consistency of the proposed prior in this case using an appropriate notion of consistency.

The presence of information paradox in Zellner’s \( g \)-prior remains one of the key motivations for considering mixture of \( g \)-priors. So, it is important to verify whether the proposed mixture can resolve the information paradox, i.e., is information consistent in the sense of Bayarri et al. (2012). Along with the above notions of consistency we study information consistency as well.

Sections 3 and 4 of the paper deal with model selection consistency. In Section 3.1, we first consider the case when the error distribution is normal. In Section 3.2 we relax the condition of normality on \( e_n \). Here \( e_i s \) are assumed to be \( i.i.d. \) with mean 0 and finite fourth moment. In Section 4, we consider the performance of some other mixtures on \( g \) with respect to model selection consistency in normal linear model setup. Section 5 deals with the ‘model false’ case. We find sufficient conditions under which the proposed prior is consistent in an appropriate sense under general error distributions. In Section 6, we consider information consistency. In Section 7, we validate the performance of the proposed prior with extensive simulation studies. We study the performance of the proposed prior in comparison with several other existing priors in the literature. Finally, we make some concluding remarks in Section 8. Proofs of all the main results are presented in the Appendix (Section 9).

2. Scaled inverse chi-square mixture of \( g \)-priors

We first motivate our proposal of a mixture on \( g \). Most of the mixtures on \( g \) in the existing literature are highly positively skewed having a unique modal point close to zero and a very flat decay. For example, hyper-\( g \) and hyper-\( g/n \) priors are \( J \)-shaped with modal point at 0. Again, if we consider popular recommendations of \( g \) in Zellner’s \( g \)-prior, choices include the unit information prior \( (g = n, \text{Kass and Raftery (1995)}) \), the choice of \( g \) related to the risk inflation criterion \( (g = p^2, \text{see Foster and George (1994), George and Foster (2000)}) \), and the benchmark prior \( (g = \max\{n, p^2\}, \text{Fernández, Ley and Steel (2001)}) \). Recently, Mukhopadhyay, Samanta and Chakrabarti (2014) presented some theoretical results to explain why a relatively larger value
of $g$ yields better results, especially when $p$ grows with $n$ and recommended using $g = n^2$ for practical purposes. From such recommendations, it seems reasonable to put relatively higher probability masses to higher values of $g$ for a mixture. Thus, there persists a gap in the domain of $g$ getting relatively higher mass when a fixed $g$ is considered compared to that of a mixture. We now propose a mixture which gives more probability mass to a range of relatively higher values of $g$ compared to the existing mixtures. We consider the \textit{scaled inverse chi-square} mixture $\pi(g)$ on $g$ with scale parameter $\tau^2 = n^2$ and degrees of freedom $\nu$, given by

$$
\pi(g) = \frac{(\tau^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \frac{\exp \left[ -\tau^2\nu/(2g) \right]}{g^{1+\nu/2}}, \; g > 0, \; \nu > 0, \; \tau^2 > 0. \quad (2.1)
$$

Although, the hyperparameter $\nu$ can take any positive value, we recommend using values between 1 and $p$. In this paper, we will consider two extreme choices of $\nu$, namely, $\nu = 1$ and $\nu = p$. Note that such choices of hyperparameters ensure that the prior has a unique mode at $n^2\nu/(\nu + 2)$ and a very flat decay.

The use of inverse gamma distribution in a scale mixture of normal priors for $\beta_\alpha$ is a common practice in Bayesian model selection. It has already been stated that the Zellner-Siow prior for $\beta_\alpha$ is an inverse gamma scale mixture of normals with shape parameter $1/2$ and scale parameter $n/2$. In the context of linear regression models with shrinkage priors, \textit{Park and Casella (2008)} and \textit{Hans (2009)} , while introducing Bayesian version of lasso, used inverse-gamma priors for similar normal scale mixtures for $\beta_\alpha$. The proposed prior in (2.1) is same as the inverse gamma prior with shape parameter $\nu/2$ and scale parameter $n^2\nu/2$.

An advantage of considering this prior is that it yields a closed form approximation to the marginal density, which is similar to the form of the marginal for a $g$-prior with some fixed choice of $g$. Availability of closed form marginals (posterior probabilities) is not necessary for good inference but it serves as a desirable property for easy implementation. This approximation makes the application of this prior simple and theoretically tractable.
2.1. Posterior probability

For the linear model setup, the vector of parameters in model $M_\alpha$ is given by $\theta_\alpha = (\beta_0, \beta_\alpha, \sigma^2, g)$. The marginal density $m_\alpha(y_n)$ in (1.4) is given by

$$m_\alpha(y_n) = \int p(y_n|\beta_0, \beta_\alpha, \sigma^2, M_\alpha)p(\beta_\alpha|\beta_0, \sigma^2, g, M_\alpha)p(\beta_0, \sigma^2)\pi(g)d(\beta_0, \beta_\alpha, \sigma^2, g),$$

where $p(y_n|\beta_0, \beta_\alpha, \sigma^2, M_\alpha)$ is p.d.f. of the $n$-variate normal distribution with mean $\beta_01 + X_\alpha\beta_\alpha$, dispersion matrix $\sigma^2 I$ and $p(\beta_\alpha|\beta_0, \sigma^2, g, M_\alpha)$, $p(\beta_0, \sigma^2)$, $\pi(g)$ are as in equations (1.5) and (2.1).

Integrating the integrand above with respect to $\beta_0$, $\beta_\alpha$ and $\sigma^2$, we get a closed form expression which leads to

$$m_\alpha(y_n) = \frac{\Gamma(n-1)/2}{\pi^{(n-1)/2}}(S^2_y)^{-(n-1)/2}\int_0^\infty \frac{(1+g)^{(n-1-p(\alpha))/2}}{[1+g(1-R^2_\alpha)]^{(n-1)/2}}\pi(g)dg,$$

(2.2)

where $S^2_y = \|y_n - \bar{y}_n1\|^2/n$, $(1 - R^2_\alpha) = y'_n(I - P_n(\alpha))y_n/(nS^2_y)$, $P_n(\alpha) = Z_{na}(Z'_nZ_n)^{-1}Z'_na$, and $Z_{na} = (1\ X_\alpha)$.

Note that the marginal density of the intercept only model $M_N : y_n = \beta_01 + e_n$, which will be referred to as the null model, does not involve the hyperparameter $g$. It can be obtained as a special case of the marginal in expression (2.2) by putting $R^2_\alpha = 0$ and $p(\alpha) = 0$.

For models $\alpha \in A \setminus \{N\}$, the marginal density given by the proposed prior (2.1) does not have a closed form. But, we can make an approximation of the marginal density in a closed form expression when the proposed mixture (2.1) is used in (2.2). When $p$ is fixed, this approximation can obtain an accuracy of order $n^{-1}$ with probability tending to 1, which is the same as the accuracy of the Laplace approximation for fixed $p$ (see Kass and Raftery (1995)). When $p$ is increasing, the Laplace approximation may not be valid for the integral in (2.2) for commonly used priors on $g$, since the integrand may not be Laplace regular (see Kass, Tierney and Kadane (1990)). When $p = O(n^b)$, $0 < b < 1$ and $\nu = 1$ (or, when $\nu$ is free of $n$), the approximation is less accurate with an error of the order $n^{-(1-b)}$. But, if $\nu = p$ (or, if $\nu$ is of same order of $n$ as $p$), the approximation still attains an accuracy of the order $n^{-1}$.
We first state the assumptions under which the approximation holds. Throughout this paper $y_n$ is modeled as (1.1) and we assume the following:

(A.1) $\mu'_n \mu_n / n < M$ for some constant $M > 0$ as $n \to \infty$.

We make a mild assumption on distribution of $e_n$.

(A.2) The errors $e_1, e_2, \ldots, e_n$ are i.i.d. with a common density having mean 0 and finite fourth order moment.

We now state the result.

**Result 2.1.** Consider the set of priors (1.5) and (2.1) with $\nu$ varying from 1 to $p$. Under assumptions (A.1) and (A.2), the marginal density in (2.2) satisfies the following:

\[
m_\alpha(y_n) \leq \tilde{m}_\alpha(y_n) \left( 1 + \frac{p}{\nu n} O(1) \right)
\]

and

\[
m_\alpha(y_n) \geq \tilde{m}_\alpha(y_n) \left( 1 + \frac{p}{\nu n} O_p(1) \right),
\]

where

\[
\tilde{m}_\alpha(y_n) = \Gamma((n-1)/2) \Gamma((\nu + p(\alpha))/2) \sqrt{n \Gamma(\nu/2)} \left( \pi S^2_y (1 - R^2_{\alpha}) \right)^{-(n-1)/2} \left( \frac{n^2 \nu}{2} \right)^{-p(\alpha)/2},
\]

uniformly in $\alpha$, for any $\alpha \in \mathcal{A} \setminus \{N\}$ as $n \to \infty$.

From Result 2.1, we find an approximation to the marginal density as

\[
m_\alpha(y_n) \approx \tilde{m}_\alpha(y_n),
\]

in the sense that the ratio of $m_\alpha(y_n)$ and $\tilde{m}_\alpha(y_n)$ goes to 1 in probability. This approximation holds uniformly in $\alpha$ since the $O(1)$ and $O_p(1)$ terms can be made free of $\alpha$. It is easy to check from Result 2.1 that if $\nu = n^r$ for some $0 \leq r \leq b$, then the approximation is accurate up to an order $1/n^{1+r-b}$. A simulation study on the performance of this approximation is added in the supplementary file.

3. Model selection consistency of scaled inverse chi-square mixture

In this section, we assume that the true mean $\mu_n$ can be expressed as a linear combination of a subset of the $p$ regressors. Let $M_{\alpha_c}, \alpha_c \in \mathcal{A}$ be the true
model. An ideal model selection procedure should identify the true model in this framework. Therefore, a natural criterion in this case is model selection consistency, which is achieved if posterior probability of the true model, given by (1.3), converges to one in probability, i.e.,

\[ p(M_{\alpha_c}| y_n) \xrightarrow{P} 1 \quad \text{as} \quad n \to \infty. \]  

(3.1)

In Section 3.1, we provide sufficient conditions under which (3.1) holds when the error distribution in (1.1) is normal. In Section 3.2, we relax the assumption of normality and assume that \( e_n \) follows any distribution satisfying assumption (A.2). In this wider class of distributions, we investigate how the set of sufficient conditions for achieving (3.1) modify under the nested model setup.

### 3.1. Model selection consistency with normality assumption

Throughout this subsection, we will assume that \( e_n \sim \mathcal{N}(0, \sigma^2 I) \). We show that when \( p \) grows with \( n \) then under appropriate conditions, model selection consistency is achieved by the proposed prior considering all \( 2^p \) models in model space.

We split the model space into three mutually exclusive and exhaustive parts as follows:
\[ A_1 = \{ \alpha \in A : M_\alpha \supset M_{\alpha_c}, \alpha \neq \alpha_c \}, \quad A_2 = \{ \alpha \in A : \alpha \notin A_1, \alpha \neq \alpha_c \} \] and \( \{ \alpha_c \} \) where \( M_{\alpha_c} \) is the true model. We assume that

\[ \lim_{n \to \infty} n^s \min_{\alpha \in A_2} \mu_n'(I - P_n(\alpha)) \mu_n/n > \delta \quad \text{for some constants } \delta > 0 \quad \text{and} \quad 0 \leq s < 1. \]

We impose a general restriction on model prior probability as

\[ \max_{\alpha, \alpha' \in A} p(M_\alpha)/p(M_{\alpha'}) \leq C \quad \text{for some constant } C > 0. \]

A remark on each of the assumptions is made below (see Remark 3.2 and Remark 3.3).

**Theorem 3.1.** Let \( y_n \) be as in (1.1) with \( \mu_n \) satisfying (A.1) and \( e_n \sim \mathcal{N}(0, \sigma^2 I) \). If \( p = O(n^b) \), then the prior specification given by (1.5) and (2.1) is model selection consistent for \( 0 < b < 2/5 \) when \( \nu = 1 \), and \( 0 < b < 1/2 \) when \( \nu = p \), provided (A.3) holds with \( s < (1 - b)/2 \) and (A.4) holds.

**Remark 3.1.** This result is different from those obtained by Wang and Sun (2014) who have shown Bayes factor consistency or pairwise consistency for...
growing number of regressors. Our result deals with the asymptotic behavior of the posterior probability of the true model considering all the $2^p$ models in the model space, and it is indeed a much stronger result than pairwise consistency. See in this context Johnson and Rossell (2012, p.652).

**Remark 3.2.** Assumption (A.3) for $s = 0$ was assumed by many authors (see, e.g., Fernández, Ley and Steel (2001), Liang et al. (2008), Bayarri et al. (2012)). It is a key assumption for model selection consistency which ensures that the models can be differentiated. Here we relax this assumption by allowing $s > 0$, which is a natural extension for the situation when $p$ grows with $n$.

**Remark 3.3.** Assumption (A.4) may seem a bit restrictive when $p$ grows with $n$. For the results in this paper to hold we do not actually need assumption (A.4); we can use much weaker versions of the assumption. For each of these results we mention the weaker version of the assumption needed for the proof. As a whole we work with assumption (A.4) for simplicity in presentation.

For Theorem 3.1 to hold, we only need $\max_{\alpha \in A} p(M_\alpha)/p(M_{\alpha_c}) \leq C$ for some constant $C > 0$. This is a reasonable assumption, since this only indicates that the true model may not have a prior probability arbitrarily close to zero, which is necessary to achieve consistency.

### 3.2. Model selection consistency in general settings

In this subsection, we extend our results to situations where the error distributions belong to a larger class satisfying assumption (A.2). We investigate the strength of the model selection algorithm when the distribution of the errors is non-normal and the same model selection rule (based on the normal likelihood) is used. In other words we investigate robustness of our model selection rule for non-normal errors. Unlike the case for normal errors, here we do not consider all the $2^p$ models. Rather, we restrict our search within a class of nested models. By nested models, we mean a set of models where for every pair of models, say, $M_1$ and $M_2$, either $M_2 \subseteq M_1$, or $M_1 \subseteq M_2$. So, the index set $A^*$ of all nested models can be expressed as

$$A^* = \{\emptyset, \{1\}, \{1, 2\}, \ldots, \{1, 2, \ldots, p\}\}, \text{ with } A^* \subset A.$$  

Note that $A^*$ has $p + 1$ different models. When $p = O(n^b)$ with $0 < b < 1$, the number of models in $A^*$ also increases with $n$. While the cardinality of $A$ is
exponential in $p$ (i.e., $2^p$), for $\mathcal{A}^*$ it is linear in $p$. On one hand generalization to non-normal errors broadens the scope of the model selection algorithm, on the other, restriction to the class of nested models abridges the model space.

Comparison of the set of nested models is of great practical interest. The situation with a model space like $\mathcal{A}^*$ may occur, for example, when we have information on relative importance of the regressors and the regressors can be ordered accordingly. Model selection in nested models has been widely studied in the Bayesian paradigm when the error distribution is normal (see, e.g., Dawid (1992), Moreno (1997), Cui and George (2008), Wang and Sun (2014)). Unlike these authors, who study Bayes factor consistency, we consider model selection consistency restricted to the space of nested models when $p$ is of the said order. We summarize our findings in the following theorem.

**Theorem 3.2.** Let $y_n$ be as in (1.1) with $\mu_n$ satisfying (A.1) and $e_n$ satisfying (A.2). If $p = O(n^b)$, then the prior specification given by (1.5) and (2.1) with $\nu = 1$ or $\nu = p$ is model selection consistent in $\mathcal{A}^*$ for any $0 < b < 1$ provided (A.3) holds with $s < (1 - b)/2$ and (A.4) holds.

**Remark 3.4.** Here also we do not need assumption (A.4) to hold strictly. For Theorem 3.2 to hold we only need $\max_{\alpha \in \mathcal{A}^*} p(M_\alpha)/p(M_{\alpha_c}) \leq C\sqrt{n}$ for some constant $C > 0$. This assumption is quite general and includes many of the popular class of model prior probabilities.

4. Properties of Some Existing Mixtures

We now investigate the performance of some other mixtures from the perspective of model selection consistency. The beta prime (beta of second kind) prior is the commonly used prior for $g$ (see Liang et al. (2008), Maruyama and George (2011), Bayarri et al. (2012)). Therefore, it is worth investigating the performance of this prior when $p$ grows with $n$. Let $g$ follow a beta prime distribution with parameters $\gamma_0$ and $\gamma_1$, then

$$
\pi(g) = \frac{\Gamma(\gamma_0 + \gamma_1)}{\Gamma(\gamma_0)\Gamma(\gamma_1)} g^{\gamma_0-1}(1+g)^{-(\gamma_0+\gamma_1)}, \quad 0 < g < \infty, \quad \gamma_0 > 0, \quad \gamma_1 > 0. \quad (4.1)
$$

We show that when $p = n^b$, $0 < b < 1$, then for some inappropriate specification of the hyperparameters $\gamma_0$ and $\gamma_1$, the model selection rule given by the set of priors (1.5) and (4.1) becomes inconsistent.
Theorem 4.1. Consider the setup of Theorem 3.1 and let the true model be the null model. If the number of regressors $p = n^b$, $0 < b < 1$, then the set of priors given by (1.5) and (4.1) with $\gamma_1 > \epsilon$ for some $\epsilon > 0$ free of $n$, is inconsistent if $(\gamma_0/\gamma_1) = O(n^{2b})$.

Remark 4.1. In hyper-$g$ prior $\gamma_0 = 1$ and $\gamma_1 = (a/2 - 1)$ for some $a > 2$ free of $n$. Hence, it follows from the above theorem that hyper-$g$ prior is inconsistent for any $b > 0$. It is already shown in Liang et al. (2008) that hyper-$g$ prior is not consistent under the null model even for fixed $p$. Hyper-$g/n$ prior remains consistent when $p$ is fixed, but it fails to be consistent if $p = n^b$, for any $b > 0$ (the proof is in the supplementary file).

Remark 4.2. Generalized $g$-prior has $\gamma_0 = A + 1$ and $\gamma_1 = B + 1$ where the authors recommend using $A = (n - p(\alpha) - 1)/2 - B$ and $B < 1/2$ for the case when $p < n$. Hence, it is easy to check that generalized $g$-prior is inconsistent for this recommended settings if $b \geq 1/2$.

Remark 4.3. The robust prior can also be expressed as a truncated scaled beta prime distribution as $(g + B)/(\rho_\alpha(n + B)) - 1 \sim$ beta prime$(1, A)$ where $A > 0$, $B > 0$, $\rho_\alpha > B/(B + n)$. The recommended choices of hyperparameters are $A = 1/2$, $B = 1$ and $\rho_\alpha = 1/(1 + p(\alpha))$. It has also been recommended that $\rho_\alpha$ should be free of $n$. This makes choice of the parameter $\rho_\alpha$ difficult when $p = n^b$, since in that case the choice of $\rho_\alpha$ involves $n$. We check with two choices of $\rho_\alpha$, a constant $\rho_\alpha$ and $\rho_\alpha = 1/(1 + p(\alpha))$. It has been shown in the supplementary file that when $p = n^b$ a necessary condition for consistency of the robust prior under the null model is $b < 1/2$, for both choices of $\rho_\alpha$.

There are some other beta shrinkage priors available in the literature. A list of which can be obtained in Ley and Steel (2012). Similar results can be derived for them.

Remark 4.4. As we have already mentioned, Zellner Siow prior is also an inverse gamma prior with scale $n$, whereas the proposed one has scale $n^2\nu/2$. It can be shown that a sufficient condition of Zellner-Siow prior to be consistent under null model is $b < 1/5$ (the proof is similar to Case II in proof of Theorem 3.1 in Section A.3). The increment of scale from the order of $n$ to $n^2$ makes the prior more reasonable for ‘large p large n regime’ which is reflected in the increment of $b$ from $1/5$ to $1/2$.

However, similar improvement is not expected from all the priors we mentioned before. For example, if we change the scale of hyper-$g/n$ prior from $n$ to $n^r$ for any $r > 1$ free of $n$, it still remain inconsistent when the null model
is true, and the number of regressors is \( p = n^b \), for any \( 0 < b < 1 \). The proof is similar in idea to the proof of the result stated in Remark 4.1.

**Remark 4.5.** As mentioned in Remark 3.3, here also we do not need a strong assumption like (A.4). For all the above results in this section to hold, we only need to assume

\( (A.4^*) \) for all \( \alpha \in \mathcal{A} \), \( p(M_\alpha)/p(M_N) \geq \delta^{p(\alpha)} \) for some constant \( \delta > 0 \).

In situations when \( M_N \) is a candidate for the true model, one may like to put additional penalty to more complex models. Therefore, the prior probabilities for the high-dimensional models may be quite small compared to that of low dimensional models. Assumption (A.4*) gives us the scope to consider those set of prior probabilities also. Consider, for example, the case where the prior inclusion probability of each regressor is \( q \), which leads to the Bernoulli prior probability \( q^{p(\alpha)}(1-q)^{p-p(\alpha)} \) to the model \( M_\alpha \). The Bernoulli prior has been used by many authors (see, e.g., George and McCulloch (1993)). It is easy to see that this prior setting satisfies assumption (A.4*).

Note that the conditions on \( b \) mentioned in Theorem 3.1 are sufficient to achieve consistency for the proposed prior. The condition in Remark 4.4 is also sufficient for Zellner-Siow prior, whereas the above conditions on beta shrinkage priors (i.e., hyper-\( g/n \) prior, generalized \( g \)-prior and robust prior) are only necessary to achieve consistency. The range of \( b \) sufficient to achieve consistency is essentially a subset of the range of \( b \) necessary to achieve consistency. Also, achieving consistency under the null model is only a part of achieving full posterior consistency. Sufficient conditions for full posterior consistency may even be stronger.

### 5. Consistency of Scaled Inverse Chi-square Prior in ‘Model False’ Case

In Sections 3 and 4, we have considered situations when the true mean \( \mu_n \) in (1.1) belongs to the span of \( \{1, x_1, \ldots, x_n\} \). We will now consider a more general scenario where \( \mu_n \) is any \( n \)-dimensional vector, i.e., the true model does not necessarily belong to the model space \( \mathcal{A} \). Several authors have studied related problems of linear model selection under this framework (see, e.g., Li (1987), Shao (1997), Chakrabarti and Ghosh (2006), Chakrabarti and Samanta (2008), Mukhopadhyay, Samanta and Chakrabarti (2014)). Of course, one cannot compute the posterior probability of the true model here, and the usual notion of model selection consistency cannot be used in this
scenario. To validate a model selection rule we therefore adopt an alternative notion of consistency suited for the model false case as used in Mukhopadhyay, Samanta and Chakrabarti (2014).

Here consistency of a model selection procedure refers to the property of choosing the model which is closest to the unknown true model among all candidate models in the model space \( \mathcal{A} \) (in an asymptotic sense). We consider the Kullback-Leibler divergence as the measure of distance between two probability distributions. We define the distance \( \Delta_n(\alpha) \) between the true distribution (of \( y_n \)) and the model \( M_\alpha \) as the minimum of the Kullback-Leibler distance with respect to the underlying parameters \((\beta_0, \beta_\alpha)\) of the distribution under \( M_\alpha \). One would naturally like to choose a model \( M_{\alpha^*} \) which is as close as possible to the true distribution (i.e., for which \( \Delta_n(\alpha^*) = \min_{\alpha \in \mathcal{A}} \Delta_n(\alpha) \)). One can find the model \( M_{\alpha^*} \) only if the true distribution were known, which is not the case here. We show that our model selection procedure chooses a model which is closest to the unknown true model in an asymptotic sense.

We consider a general class of error distributions satisfying assumption (A.2) and the whole set of \( 2^p \) models \( \mathcal{A} \) for comparison. We make the following assumption which is analogous to assumption (A.3), by replacing \( \mathcal{A}_2 \) in (A.3) by \( \mathcal{A} \) as

\[
(A.3^*) \lim_{n \to \infty} n^s \min_{\alpha \in \mathcal{A}} \mu_n(I - P_n(\alpha)) \mu_n/n > \delta \quad \text{for some constant} \quad \delta > 0 \quad \text{and} \quad 0 \leq s < 1.
\]

Note that in the model false case \( \mathcal{A}_2 = \mathcal{A} \).

Let the true distribution of \( y_n \) has a density function \( f \). It can be easily verified that the Kullback-Leibler distance between the true distribution given by the density function \( f \), and the distribution \( N(1 \beta_0 + X_\alpha \beta_\alpha, \sigma^2 I) \) under \( M_\alpha \) equals

\[
\int f(y_n) \log f(y_n) \, dy_n + \frac{n}{2} \left( 1 + \log \sigma^2 \right) + \frac{1}{\sigma^2} (\mu_n - 1 \beta_0 - X_\alpha \beta_\alpha)' (\mu_n - 1 \beta_0 - X_\alpha \beta_\alpha).
\]

The distance \( \Delta_n(\alpha) \) between the true model \( f \), and the model \( M_\alpha \) is obtained by minimizing the above with respect to \((\beta_0, \beta_\alpha)\), as follows

\[
\Delta_n(\alpha) = \int f(y_n) \log f(y_n) \, dy_n + \frac{n}{2} \left( 1 + \log \sigma^2 \right) + D_n(\alpha), \quad (5.1)
\]
where
\[ D_n(\alpha) = \frac{1}{2\sigma^2}(\mu_n'(I - P_n(\alpha))\mu_n). \] (5.2)

Note that the first two terms of \( \Delta_n(\alpha) \) in (5.1) do not involve \( \alpha \). Therefore, \( \arg\min_\alpha \Delta_n(\alpha) = \arg\min_\alpha D_n(\alpha) \) and we show that
\[ \frac{D_n(\hat{\alpha})}{\min_{\alpha \in A} D_n(\alpha)} \xrightarrow{p} 1 \quad \text{as} \quad n \to \infty, \] (5.3)
where \( M_{\hat{\alpha}} \) is the model chosen by the model selection rule based on the proposed prior.

**Theorem 5.1.** Let \( y_n \) be as in (1.1) with \( \mu_n \) being any real vector in \( \mathbb{R}^n \) satisfying (A.1) and \( e_n \) satisfying (A.2). Suppose the assumption (A.4) holds and (A.3*) holds with \( s < (1 - b)/2 \). If the number of regressors \( p = O(n^b) \), \( 0 < b < 1 \), then the set of prior (1.5) and (2.1) is consistent in the sense that (5.3) holds for any \( 0 < b < 1 \) when \( \nu = 1 \) or \( \nu = p \).

In Theorem 5.1, it is only assumed that \( y_n \) is the sum of two components, namely, the true mean \( \mu_n \) and the random error \( e_n \). Here, \( \mu_n \) is allowed to be arbitrary and \( e_n \) can follow any distribution satisfying assumption (A.2); even symmetry or continuity is not required. Thus, given the additive structure of \( y_n \) as in (1.1), consistency is obtained in a much general setting.

Lastly, assumption (A.4) can be relaxed to a great extent here. For Theorem 5.1 to hold, we only need the following
\[ (A.4^*) \max_{\alpha, \alpha' \in A} p(M_\alpha)/p(M_{\alpha'}) \leq Cn^r \] for some \( C > 0 \) and \( r > 0 \).
This assumption is satisfied for a very large class of prior probabilities on the model space.

### 6. Information Consistency

The criterion of information consistency is considered by several authors (see, e.g., Jeffreys (1961), Berger and Pericchi (2001), Bayarri and García-Donato (2008), Liang et al. (2008), Bayarri et al. (2012)). While comparing the null model with any model \( M_\alpha \), suppose that \( ||\hat{\beta}_\alpha||^2 \to \infty \) (or equivalently, the usual F statistics goes to \( \infty \)) with both \( n \) and \( p(\alpha) \) are fixed, \( \hat{\beta}_\alpha \) being the least squares estimator of \( \beta_\alpha \). This is considered as a very strong evidence supporting the model \( M_\alpha \), and it is expected that the Bayes factor for comparing model \( M_\alpha \) to the null model would go to \( \infty \). The property that...
the Bayes factor goes to $\infty$ whenever $\|\hat{\beta}_0\|^2 \to \infty$ with fixed $n$ and $p(\alpha)$ is termed as information consistency in Bayarri et al. (2012). However, this does not hold in the case of Zellner’s $g$-prior. For mixture of $g$-priors, Liang et al. (2008, Theorem 2) give a sufficient condition which ensures information consistency. The following result gives conditions under which the proposed mixture in (2.1) is information consistent.

**Result 6.1.** Consider the set of prior probabilities (1.5). Then the mixture on $g$ given by (2.1) is information consistent if $n \geq p + 1$ when $\nu = 1$ and if $n \geq 2p$ when $\nu = p$.

The proof of this result is in the supplementary file.

Note that for $\nu = 1$ the proposed prior is information consistent with minimal sample size, i.e., information consistency holds for any $n > p$ (see Liang et al. (2008) in this context). But for $\nu = p$, the proposed prior fails to be information consistent with minimal sample size.

### 7. Performance of The Proposed Prior on Simulated Datasets

In this section we validate the performance of the proposed prior using simulated datasets. We present simulation results for model selection consistency under different simulation schemes. In each case, we consider our proposed prior with two choices of the hyperparameter $\nu$, viz., $\nu = 1$ (proposed prior I) and $\nu = p$ (proposed prior II). Along with the proposed prior we also consider four other priors on $g$, namely, Zellner-Siow prior, hyper-$g/n$ prior, generalized $g$-prior and robust prior.

Our results are designed for the case when $p$ increases with $n$, and therefore, we consider moderately large $p$ compared to $n$. Three choices of $n$ ($n = 50, 100, 150$) and two choices of $p$ ($p + 1 = 30, 50$) for each $n$ have been considered.

The theoretical results are not confined to the case with normal errors; any error distribution satisfying assumption (A.2) can be considered. We consider three different error distributions, namely, normal, Laplace and $t$ with degrees of freedom 3 ($t_{(3)}$). Note that $t_{(3)}$ does not satisfy the fourth order moment condition of (A.2). Moreover, we consider all the $2^p$ models in case of Laplace and $t_{(3)}$ distributions also, although our theoretical results for general case allow only nested model setup. We consider these settings to check the performance of the proposed mixture when some of the assumptions of Theorem 3.2 do not hold. The simulation scheme is described as follows.
For each combination of \((n, p)\), we generate \(n\) values of each of the \(p\) regressors \(x_1, x_2, \ldots, x_p\) and this gives the full design matrix \(X_n\). We choose \(p\) numbers \(\xi_i, i = 1, \ldots, p\) and generate the \(n\) values of the \(i^{th}\) regressor \(x_i\) from an \(N(\xi_i, 1)\) distribution, \(i = 1, \ldots, p\). We assume that the \(n\) values of the \(i^{th}\) regressor are coming from a homogeneous population. In order to fix a “true” model, we choose its dimension \(p(\alpha_c)\) and then choose the \(p(\alpha_c)\) non-zero regression coefficients \(\beta_i\)’s, the intercept \(\beta_0\) in the true model and also a value for the error variance \(\sigma^2\). The \(p(\alpha_c)\) columns of the design matrix \(X_{\alpha_c}\) for the true model are chosen at random from the \(p\) columns of \(X_n\). Here, \((\xi_1, \ldots, \xi_p)\) is chosen as a random permutation of \((0.2, 0.4, \ldots, 0.2 \times p)\). The dimension of the true model \(p(\alpha_c)\) is chosen as \([p/2]\) and the \(p(\alpha_c)\) non-zero regression coefficients \(\beta_i\)’s and the intercept \(\beta_0\) in the true model are randomly chosen from the set \([-0.2, 0.4, \ldots, (-1)^p \times 0.2 \times p]\). Lastly, we choose \(\sigma = 1\).

After choosing the dimension \(p(\alpha_c)\), the coefficients \((\beta_0, \beta_{\alpha_c})\), the error variance \(\sigma^2\) of the true model and the design matrix \(X_n\), we generate \(e_n\) from normal, Laplace and \(t(3)\) distributions, each with location vector \(0\) and dispersion matrix \(\sigma^2 I\). The vector of observations \(y_n\) is obtained by adding \(\mu_n = \beta_0 + X_{n\alpha_c}\beta_{\alpha_c}\) to \(e_n\). Having obtained the data, we compute the posterior probability of the true model using the set of priors (1.5) for several mixtures on \(g\) as indicated above. There are two issues to be mentioned here. Firstly, for calculation of the marginal, one needs to calculate the integral in (2.2), which is not of closed form for all the mixtures. We use numerical integration (available in R software) to calculate this integral for all the mixtures. Secondly, since \(p\) is large, calculation of the posterior probability for the candidate models \(\alpha \in A\) becomes quite infeasible. This is because calculation of posterior probability of any model requires the marginal densities \((m_{\alpha}(y_n))\) for all the \(2^p\) candidate models in \(A\). Therefore, we use Markov Chain Monte Carlo simulation techniques to approximate the posterior probabilities, where computation of marginal densities can be restricted only to the models visited by the chain. We have used the Gibbs sampling algorithm, to simulate from the relevant Markov chain. The sampling scheme and the method of computation of posterior probabilities are completely described in Chipman, George and McCulloch (2001, Section 3.5). We have generated a Markov chain of length 10000 of which the first 5000 have been used as burn-in. Similar simulation is also used in Scheme 2 of Mukhopadhyay, Samanta and Chakrabarti (2014).

For each combination of \((n, p)\) and each of the mixtures of \(g\), we repeat the above for 100 times fixing the chosen values of \(\xi_i\)’s, \(p(\alpha_c)\), \(\beta_j\)’s and \(\sigma\). The
mean and mean square error of the posterior probabilities of the true model are presented in Table 1. In the table, the mean posterior probabilities have been shown, keeping the mean squared errors (m.s.e.) in brackets.

From Table 1, it is evident that performance of the proposed mixture is better than the other mixtures. The differences in performances of the proposed priors from others increase with \( n \). Among the two choices of \( \nu \), the choice \( \nu = p \) performs better than \( \nu = 1 \). Among the other priors, robust prior has the best performance. For example, when \( t(3) \) is considered with \( n = 150 \) and \( p = 50 \), for the proposed prior I (proposed prior II), the Markov chain visits the true model in around 37\% (46\%) cases, whereas for the robust prior it visits the true model only in 6\% cases. A similar phenomenon is observed in other cases as well.

Next, we consider two sparse situations. We first consider the case where the null model is true (Scheme 1). We take \( p = 30 \) and \( \beta_0 = 5 \) and assume \( e_n \) to follow a normal distribution. The rest of the simulation scheme is as described above. The mean and m.s.e. of 100 replicates of posterior probabilities of different mixtures on \( g \) are shown in Figure 1.

Lastly, we consider the interesting situation where sparsity is present in the simulation scheme (Scheme 2) in the sense that a set of regression coefficients in the true model is negligible, even if not exactly zero. Here, it is desirable to select the parsimonious model (say \( \alpha_s \)) that includes all regressors with significant regression parameters, rather than the true model (\( \alpha_c \)). We take \( p = 30 \), \( p(\alpha_c) = 15 \), \( \sigma^2 = 1 \). The error distribution considered is normal. The regression parameters of the true model is as follows: \( \beta_i = i + 1 \) for \( i = 0, 1, \ldots, 4 \), \( 0 < |\beta_i| < 0.008 \) for \( i = 5, 6, \ldots 15 \) and \( \beta_i = 0 \) for \( i > 15 \). For different \( n \), we plot the mean and m.s.e. of posterior probabilities of the sparse model, \( \alpha_s = \{1, 2, 3, 4\} \) in Figure 2.

From both the figures, it can be seen that the performances of the proposed priors are distinctly better than that of the other priors. When the null model is true (see Figure 1), the posterior probabilities of all other priors are less than 0.00002 (therefore, the corresponding lines are not visually distinguishable in the figure), whereas the proposed prior for \( \nu = p \) achieves an average posterior probability of 0.59, when \( n = 150 \). In Scheme 2 (see Figure 2), the other priors perform relatively better than Scheme 1. The highest average posterior probability among all the other priors is achieved by the generalized \( g \)-prior for \( n = 150 \) and is less than 0.075, whereas the average posterior probabilities achieved by the proposed priors for \( n = 150 \) are 0.358 for \( \nu = 1 \), and 0.599 for \( \nu = p \).
8. Concluding Remarks

In this paper, we propose a mixture of $g$-priors suitable for the case when $p$ grows with $n$. The resulting marginal has an approximation with a closed form expression which makes implementation simple. We investigate the performance of the proposed prior by deriving consistency properties under different setups. We also compare its performance with that of several other mixtures using simulation results under different simulation schemes which demonstrates its nobility. Theoretically as well as in simulations, superiority of the performance of the proposed prior has also been shown under sparse situations.

The prior for $\beta_0$ arising from this mixture has a very thick tail which is recommended by Jeffreys (Jeffreys (1961)). Further, this structure of priors (1.5) has the properties like predictive matching and group invariance as described in Bayarri et al. (2012) (see Results 2-4 of Bayarri et al. (2012) in this context). The authors have explicitly justified the adoption of the form (1.5) in a broader context.

Finally, it may be mentioned that we have studied the performance of the proposed mixture for $\nu = 1$ and $\nu = p$. The performance of the mixture with $\nu = p$ is better than the other in the light of all the properties considered in this paper except information consistency. The prior with $\nu = p$ fails to be information consistent when $n \leq 2p$. In practice, when $n > 2p$ one can conveniently use the prior with $\nu = p$.

Appendix A

In this section, we present the proofs of most of the main results stated in this paper. Many of the statements in the following proofs hold with probability tending to 1 as $n \to \infty$, although this will not be always mentioned. Throughout this section we will assume that $\text{Var}(e_n) = \sigma^2 I$, where $\sigma^2 > 0$ is unknown.

A.1. Auxiliary Results

We first state three lemmas which will help in proving our main results. The proofs of these lemmas are given in the supplementary file.
### Table 1
Average and mean squared error of posterior probability of the true model.

| error density | Priors | Zeller-Siow mean (m.s.e.) | Hyper-$g/n$ mean (m.s.e.) | Generalized $g$ mean (m.s.e.) | Robust mean (m.s.e.) | Proposed I mean (m.s.e.) | Proposed II mean (m.s.e.) |
|---------------|--------|---------------------------|---------------------------|-----------------------------|---------------------|------------------------|--------------------------|
| 30 $p + 1$ 50 | 0.0948 (0.8294) | 0.0949 (0.8293) | 0.0956 (0.8282) | 0.0981 (0.8242) | 0.1251 (0.7836) | 0.1598 (0.7356) |
| 100 | 0.2916 (0.5095) | 0.2919 (0.5091) | 0.2926 (0.5081) | 0.3056 (0.4903) | 0.4598 (0.3131) | 0.6163 (0.1761) |
| Normal | 150 | 0.4932 (0.2954) | 0.4962 (0.2922) | 0.3758 (0.4192) | 0.5358 (0.2515) | 0.6834 (0.1393) | 0.9055 (0.0824) |
| 50 | 0.0000 (0.9999) | 0.0000 (0.9999) | 0.0000 (0.9999) | 0.0000 (0.9999) | 0.0006 (0.9999) | 0.0008 (0.9998) |
| 100 | 0.0018 (0.9964) | 0.0017 (0.9965) | 0.0018 (0.9965) | 0.0019 (0.9963) | 0.0255 (0.9525) | 0.0414 (0.9255) |
| 150 | 0.0474 (0.9086) | 0.0475 (0.9083) | 0.0477 (0.2749) | 0.0525 (0.8991) | 0.2581 (0.4382) | 0.4820 (0.3022) |
| 30 | 0.1611 (0.7156) | 0.1603 (0.7167) | 0.1614 (0.7151) | 0.1669 (0.7067) | 0.2155 (0.6387) | 0.2767 (0.5637) |
| 50 | 0.1937 (0.6620) | 0.1933 (0.6626) | 0.1944 (0.6608) | 0.2028 (0.6484) | 0.3374 (0.4764) | 0.4369 (0.3769) |
| Laplace | 150 | 0.2573 (0.5640) | 0.2575 (0.5636) | 0.2578 (0.5632) | 0.2734 (0.5419) | 0.4822 (0.3073) | 0.6144 (0.2055) |
| 50 | 0.0011 (0.9978) | 0.0011 (0.9978) | 0.0011 (0.9978) | 0.0011 (0.9978) | 0.0013 (0.9974) | 0.0068 (0.9866) |
| 100 | 0.0052 (0.9896) | 0.0053 (0.9895) | 0.0053 (0.9895) | 0.0055 (0.9890) | 0.0585 (0.8941) | 0.0816 (0.8554) |
| 150 | 0.0069 (0.9863) | 0.0070 (0.9861) | 0.0070 (0.9862) | 0.0075 (0.9851) | 0.1166 (0.8024) | 0.1563 (0.7495) |
| 30 | 0.1059 (0.8103) | 0.1060 (0.8103) | 0.1065 (0.8094) | 0.1091 (0.8053) | 0.1673 (0.7252) | 0.1921 (0.6918) |
| 50 | 0.3191 (0.4700) | 0.3197 (0.4693) | 0.3204 (0.4683) | 0.3342 (0.4497) | 0.5003 (0.2674) | 0.6593 (0.1332) |
| $t_{(3)}$ | 150 | 0.5419 (0.2103) | 0.5460 (0.2488) | 0.4292 (0.3629) | 0.5834 (0.2141) | 0.7100 (0.1190) | 0.8965 (0.0274) |
| 50 | 0.0004 (0.9993) | 0.0004 (0.9993) | 0.0004 (0.9992) | 0.0004 (0.9992) | 0.0307 (0.9444) | 0.0254 (0.9522) |
| 100 | 0.0080 (0.9840) | 0.0081 (0.9840) | 0.0080 (0.9840) | 0.0084 (0.9834) | 0.0713 (0.8765) | 0.0911 (0.8468) |
| 150 | 0.0562 (0.8924) | 0.0562 (0.8924) | 0.0557 (0.8931) | 0.0615 (0.8826) | 0.3744 (0.4279) | 0.4653 (0.3316) |
Fig 1: Mean and M.S.E. of Posterior Probabilities of the True Model in Scheme 1

Lemma A.1. If \( y_n = \mu_n + e_n \) with \( \mu_n \) satisfying assumption (A.1) and \( e_n \) satisfying assumption (A.2), then the following results hold as \( n \to \infty \):

(i) \( \hat{e} = \frac{\sum_{i=1}^{n} e_i}{n} \xrightarrow{p} 0 \),
(ii) \( \frac{\sum_{i=1}^{n} \mu_i e_i}{n} \xrightarrow{p} 0 \),
(iii) \( \bar{S}_e^2 \xrightarrow{p} \sigma^2 \) where \( n \bar{S}_e^2 = \sum_{i=1}^{n} (e_i - \bar{e})^2 \),
(iv) \( \max_{\alpha \in A} e'_n p_n(\alpha) e_n/n = O_p \left( \frac{p}{n} \right) \),
(v) \( \max_{\alpha \in A_2} |\mu'_n(I - P_n(\alpha))e_n|/n = O_p \left( \sqrt{\frac{p}{n}} \right) \) and
(vi) \( e'_n(I - P_n(\alpha))e_n/n \xrightarrow{p} \sigma^2 \) uniformly in \( \alpha \in A \).

Lemma A.2. Let \( R^2_\alpha \) is as in (2.2). Under assumptions (A.1) and (A.2), \( (1 - R^2_\alpha) > \sigma^2/(2M + 4\sigma^2) \) with probability tending to 1 uniformly in \( \alpha \in A \), where \( M \) is as in assumption (A.1).

Lemma A.3. Under the setup of Theorem 3.2, for any fixed \( R > 0 \), with probability tending to one

\[
\max_{\alpha \in A_1} \frac{e'_n(p_n(\alpha) - p_n(\alpha_c))e_n}{\sigma^2(p(\alpha) - p(\alpha_c))} \leq R \log p.
\]
A.2. Proof of Result 2.1

Using (2.1) and (2.2), we write,

$$m_{\alpha}(y_n) = \mathcal{C}_{1,y,n} \mathcal{I}$$

where,

$$\mathcal{C}_{1,y,n} = \frac{\Gamma(n-1)/2}{\Gamma(\nu/2) \pi^{(n-1)/2} \sqrt{n}} \left( \frac{\tau^2 \nu}{2} \right)^{\nu/2} \left( S_y^2 \right)^{-(n-1)/2},$$

and

$$\mathcal{I} = \int_{0}^{\infty} e^{-\tau^2 \nu/(2g)} g^{-(1+\nu/2)} (1+g)^{(n-p(\alpha)-1)/2} \left\{ 1 + (1 - R_{\alpha}^2) g \right\}^{-(n-1)/2} dg.$$ (A.3)

We first evaluate $\mathcal{I}$. After making a transformation $w = \tau^2 \nu/(2g)$, we observe that,

$$\mathcal{I} = \mathcal{C}_{2,y,n} \int_{0}^{\infty} e^{-\frac{w^{\nu/2-1}}{2w}} \left( 1 + \frac{\tau^2 \nu}{2w} \right)^{(n-p(\alpha)-1)/2} \left\{ 1 + (1 - R_{\alpha}^2) \frac{\tau^2 \nu}{2w} \right\}^{-(n-1)/2} dw.$$ (A.4)
where $C_{2,y,n} = (\tau^2 \nu / 2)^{-\nu/2}$. Next we use the fact that, for any $w > 0$,

$$\{1 + (1 - R_\alpha^2)\tau^2 \nu / (2w)\}^{-(n-1)/2} < \{(1 - R_\alpha^2)\tau^2 \nu / (2w)\}^{-(n-1)/2}.$$

Use of this inequality along with multiplication and division by $(\tau^2 \nu / (2w))^{(n-p(\alpha) - 1)/2}$ in R.H.S. of (A.4) gives,

$$I \leq C_{3,y,n} \int_0^\infty e^{-w} w^{(p(\alpha) + \nu)/2 - 1} \left(1 + \frac{2w}{\tau^2 \nu}\right)^{(n-p(\alpha) - 1)/2} dw, \quad (A.5)$$

where $C_{3,y,n} = C_{2,y,n}(\tau^2 \nu / 2)^{-p(\alpha)/2}(1 - R_\alpha^2)^{-(n-1)/2}$. The quantity $(n - p(\alpha) - 1)/2$ can either be an integer or a mixed fraction. We first deal with the case when $(n - p(\alpha) - 1)/2$ is an integer. We expand last term in (A.5) in binomial expansion as follows,

$$\left(1 + \frac{2w}{\tau^2 \nu}\right)^{(n-p(\alpha) - 1)/2} = \left(1 + \frac{(n - p(\alpha) - 1)w}{\tau^2 \nu} + \ldots + \left(\frac{2w}{\tau^2 \nu}\right)^{(n-p(\alpha) - 1)/2}\right). \quad (A.6)$$

For a fixed $n$ the sum in (A.6) is finite and hence the integration in (A.5) and the summation can be interchanged. It can be easily seen that after interchange has taken place, each integration under the summation forms a gamma integral of appropriate order. It then follows that,

$$I \leq C_{3,y,n} \left\{ \Gamma\left(\frac{p(\alpha) + \nu}{2}\right) + \frac{(n - p(\alpha) - 1)\Gamma\left(\frac{p(\alpha) + \nu}{2} + 1\right)}{\tau^2 \nu} + \ldots + \left(\frac{2}{\tau^2 \nu}\right)^{(n-p(\alpha) - 1)/2}\Gamma\left(\frac{n + \nu - 1}{2}\right)\right\},$$

$$= C_{3,y,n} \Gamma\left(\frac{p(\alpha) + \nu}{2}\right)\left\{1 + \frac{(n - p(\alpha) - 1)(p(\alpha) + \nu)}{2\tau^2 \nu} + \ldots + \frac{(p(\alpha) + \nu)(p(\alpha) + \nu + 2) + \ldots + (n + \nu - 3)}{(\tau^2 \nu)^{(n-p(\alpha) - 1)/2}}\right\}.$$
For $\tau = n$, the above expression is,

$$< C_{3,y,n} \Gamma \left( \frac{p(\alpha) + \nu}{2} \right) \left\{ 1 + \frac{n(p(\alpha) + \nu)}{2n^2\nu} + \frac{n^2(p(\alpha) + \nu)(p(\alpha) + \nu + 2)}{2!(2n^2\nu)^2} + \ldots + \text{upto } (n - p(\alpha) - 1)/2^{th} \text{ term} \right\}$$

using the fact that if $a > b > 0$ then $a/b > (a + 1)/(b + 1)$. The bracketed portion of the last expression is a G.P. series with positive terms. We add the terms with higher power and make it an infinite G.P. series and also replace $p(\alpha)$ by $p$ ($p(\alpha) \leq p$) to make the series free of $\alpha$. The resultant term is as follows,

$$< C_{3,y,n} \Gamma \left( \frac{p(\alpha) + \nu}{2} \right) \left\{ 1 + \frac{p + \nu}{2n\nu} + \left( \frac{p + \nu}{2n\nu} \right)^2 + \ldots \right\} \quad (A.7)$$

From (A.7) it is clear that

$$\mathcal{I} \leq \begin{cases} 
C_{3,y,n} \Gamma \left( \frac{p(\alpha) + \nu}{2} \right) \left( 1 + O \left( \frac{1}{n^2} \right) \right) & \text{if } \nu = 1 \text{ and } p = n^b, \\
C_{3,y,n} \Gamma \left( \frac{p(\alpha) + \nu}{2} \right) \left( 1 + O \left( \frac{1}{n} \right) \right) & \text{if } \nu = p \text{ and } p = n^b.
\end{cases} \quad (A.8)$$

We now consider the case when $(n - p(\alpha) - 1)/2$ is not an integer. Certainly then $(n - p(\alpha))/2$ is an integer. For any $w > 0$, we write the last term in (A.5) as

$$(1 + (2w)/(\tau^2\nu))^{(n-p(\alpha)-1)/2} \leq (1 + (2w)/(\tau^2\nu))^{(n-p(\alpha))/2}.$$

Using the above inequality, we proceed as before in (A.6), (A.7) and get the same bound as in (A.8).

Next, we assign a bound on $\mathcal{I}$ from other direction and show that the difference between the two bound is small. For this, we move back to (A.4) and use the inequality $(1 + (\tau^2\nu)/(2w)) > (\tau^2\nu)/(2w)$ along with a multiplication and division by the factor $((1 - R^2_\alpha)(\tau^2\nu)/(2w))^{(n-1)/2}$ in the integrand.
of (A.4). The resultant integral is as follows,

$$ I \geq C_{\alpha,n} \int_0^{\infty} w^{(\nu + p(\alpha))/2 - 1} e^{-w} \left(1 - \frac{1}{1 + (1 - R_a^2)\tau^2 \nu / (2w)}\right)^{(n-1)/2} \, dw, $$

$$ I \geq C_{\alpha,n} \int_0^{\infty} w^{(\nu + p(\alpha))/2 - 1} e^{-w} \left(1 - \frac{2w}{1 - R_a^2} \right)^{(n-1)/2} \, dw, $$

where $C_{\alpha,n}$ is the same as in (A.8). As before, here also we deal separately two cases when $(n - 1)/2$ is an integer and when it is not. First consider the case when it is an integer. Expanding as before we get,

$$ \left(1 - \frac{2w}{1 - R_a^2} \right)^{(n-1)/2} = 1 - \frac{(n - 1)w}{\tau^2 \nu (1 - R_a^2)} + \frac{(n - 1)(n - 3)w^2}{2! \tau^4 \nu^2 (1 - R_a^2)^2} - \ldots + \left(\frac{-2w}{1 - R_a^2} \right)^{(n-1)/2} $$

$$ \geq 1 - \frac{w}{n\nu(1 - R_a^2)} - \frac{1}{2!} \left(\frac{w}{(1 - R_a^2)n\nu}\right)^2 - \ldots - \frac{1}{((n - 1)/2)!} \left(\frac{w}{(1 - R_a^2)n\nu}\right)^{(n-1)/2}, $$

putting $\tau = n$ and replacing $(n - 1)$, $(n - 3)$, etc. by $n$. As before, we interchange the summation and integration and resultant term is as follows

$$ I \geq C_{\alpha,n} \Gamma \left(\frac{p(\alpha) + \nu}{2}\right) \left\{ 1 - \frac{(p(\alpha) + \nu)}{2n\nu(1 - R_a^2)} - \frac{(p(\alpha) + \nu)(p(\alpha) + \nu + 2)}{2! (2n\nu(1 - R_a^2))^2} - \ldots \right\} $$

$$ \geq C_{\alpha,n} \Gamma \left(\frac{p(\alpha) + \nu}{2}\right) \left\{ 1 - \frac{(p(\alpha) + \nu)}{2n\nu(1 - R_a^2)} - \frac{(p(\alpha) + \nu)}{2n\nu(1 - R_a^2)}^2 - \ldots - \left(\frac{(p(\alpha) + \nu)}{2n\nu(1 - R_a^2)}\right)^{(n-1)/2} - \ldots \right\} $$

$$ \geq C_{\alpha,n} \Gamma \left(\frac{p(\alpha) + \nu}{2}\right) \left\{ 1 - \frac{(p + \nu)}{2n\nu(1 - R_a^2)} - \frac{(p + \nu)}{2n\nu(1 - R_a^2)}^2 - \ldots \right\}. $$

From Lemma A.2, we know that for all $\alpha \in A$, $(1 - R_a^2)$ has a fixed positive lower bound with probability tending to 1. Using the lemma we get the following result

$$ I \geq \begin{cases} C_{\alpha,n} \Gamma \left(\frac{p + \nu}{2}\right) \left(1 + O_p \left(\frac{1}{n^{1-b}}\right)\right) & \text{if } \nu = 1 \text{ and } p = n^b, \\
C_{\alpha,n} \Gamma \left(\frac{p + \nu}{2}\right) \left(1 + O_p \left(\frac{1}{n}\right)\right) & \text{if } \nu = p \text{ and } p = n^b. \end{cases} \quad (A.9) $$
Hence, from (A.1), (A.8) and (A.9), the result follows.

\textbf{A.3. Proof of Theorem 3.1}

We need to show

\[\sum_{\alpha \in A_i} \frac{p(M_\alpha)}{p(M_{\alpha_c})} \frac{m_\alpha(y_n)}{m_{\alpha_c}(y_n)} \to 0, \quad \text{for } i = 1, 2. \tag{A.10}\]

We proof separately for the cases \(M_{\alpha_c} \neq M_N\) and \(M_{\alpha_c} = M_N\).

\textbf{Case I.} \(M_{\alpha_c} \neq M_N\).

We first consider (A.10) for \(i = 2\).

From Result 2.1 we have

\[
m_\alpha(y_n) / m_{\alpha_c}(y_n) \leq \left( \frac{n^2 \nu}{2} \right)^{-(p(\alpha) - p(\alpha_c)) / 2} \left( \frac{1 - R_{\alpha_c}^2}{1 - R_{\alpha}^2} \right)^{(n-1)/2} \Gamma((\nu + p(\alpha))/2) \{1 + O(1)/(\nu n)\} \Gamma((\nu + p(\alpha_c))/2) \{1 + O_p(1)/(\nu n)\}, \tag{A.11}\]

where the terms \(O(1)\) and \(O_p(1)\) are free of \(\alpha\). We consider the terms of R.H.S. of (A.11) one by one. First we consider the second term. We have for \(\alpha \in A_2\),

\[
\frac{1 - R_{\alpha_c}^2}{1 - R_{\alpha}^2} = \frac{e_n'(I - P_n(\alpha)) \mu_n + 2 \mu_n'(I - P_n(\alpha)) e_n}{e_n'(I - P_n(\alpha_c)) e_n} - 2 \max_{\alpha \in A_2} \frac{|\mu_n'(I - P_n(\alpha)) e_n|}{n} + \frac{e_n'(P_n(\alpha) e_n)}{n},
\]

By assumption (A.3) and from parts (iv) and (v) of Lemma A.1, we have

\[
\frac{1}{n} \left\{ \min_{\alpha \in A_2} \mu_n'(I - P_n(\alpha)) \mu_n - \max_{\alpha \in A_2} (2 |\mu_n'(I - P_n(\alpha)) e_n| + e_n'(P_n(\alpha) e_n)) \right\} > \frac{\delta_0}{n^s}
\]

for some \(\delta_0 > 0\), with probability tending to 1 provided \(s < (1 - b)/2\). Since \(e_n' e_n/n \overset{p}{\to} \sigma^2\),

\[
\max_{\alpha \in A_2} \left( \frac{1 - R_{\alpha_c}^2}{1 - R_{\alpha}^2} \right)^{(n-1)/2} \leq \left( 1 + \frac{\delta_1}{n^s} \right)^{-(n-1)/2} \tag{A.12}\]
for some $\delta_1 > 0$.

To evaluate third term of R.H.S. of (A.11), we make use of the result $\{x/(x + s)\}^s \leq \Gamma(x + s)/(x^s \Gamma x) \leq 1$ for $0 < s < 1$ and $x > 0$ from Wendel (1948). It can be shown that,

\[
\frac{\Gamma\{(\nu + p(\alpha))/2\}}{\Gamma\{(\nu + p(\alpha_c))/2\}} \leq \left(\frac{\nu + p}{2}\right)^{|p(\alpha) - p(\alpha_c)|/2}. \tag{A.13}
\]

Hence from (A.11), (A.12) and (A.13),

\[
\max_{\alpha \in A_2} \frac{m_\alpha(y_n)}{m_{\alpha_c}(y_n)} \leq \left(\frac{n^2 \nu}{2}\right)^{(p - p(\alpha_c))/2} \left(1 + \frac{\delta_1}{n^s}\right)^{-(n - 1)/2} \left(\frac{\nu + p}{2}\right)^{p/2} \frac{\{1 + p O(1)/(\nu n)\}}{\{1 + p O(1)/(\nu n)\}}.
\]

Using assumption (A.4) we also get an upper bound of the ratio of prior probabilities of the models. Therefore,

\[
\sum_{\alpha \in A_2} \frac{p(M_\alpha)}{p(M_{\alpha_c})} \frac{m_\alpha(y_n)}{m_{\alpha_c}(y_n)} \leq C' 2^p \left(\frac{n^2 \nu}{2}\right)^{(p - p(\alpha_c))/2} \left(1 + \frac{\delta_1}{n^s}\right)^{-(n - 1)/2} \left(\frac{\nu + p}{2}\right)^{p/2}, \tag{A.14}
\]

for some constant $C'$. It is easy to check that the above quantity goes to 0 as $n \to \infty$ when $s < (1 - b)/2$.

Next we prove (A.10) for $i = 1$. We recall (A.11) and consider each term of R.H.S.. We have, for any $\alpha \in A_1$,

\[
\left(\frac{1 - R_{2\alpha}}{1 - R_\alpha}\right)^{(n - 1)/2} = \left\{\frac{e_n'(I - P_n(\alpha))e_n}{e_n'(I - P_n(\alpha_c))e_n}\right\}^{(n - 1)/2} = \left\{1 - \frac{e_n'(P_n(\alpha) - P_n(\alpha_c))e_n}{e_n'(I - P_n(\alpha_c))e_n}\right\}^{-(n - 1)/2}.
\]

We now use Lemma 2 of Mukhopadhyay, Samanta and Chakrabarti (2014) which states that for any $R > 2$, with probability tending to 1,

\[
\max_{\alpha \in A_1} \frac{e_n'(P_n(\alpha) - P_n(\alpha_c))e_n}{\sigma^2(p(\alpha) - p(\alpha_c))} \leq R \log p. \tag{A.15}
\]
Again from part \((vi)\) of Lemma \(A.1\), \(e_n^T (I - P_n(\alpha_c)) e_n > n\sigma^2(1 - \epsilon)\) with probability tending to 1, for any \(\epsilon > 0\). These facts will imply that for any \(R > 2\), with probability tending to 1 uniformly in \(\alpha \in A_1\),

\[
\left(\frac{1 - R^2}{1 - R^2_\alpha}\right)^{(n-1)/2} \leq \left\{1 - \frac{\sigma^2(p(\alpha) - p(\alpha_c)) R \log p}{n\sigma^2(1 - \epsilon)}\right\}^{-(n-1)/2}
\leq \exp\left\{\frac{(p(\alpha) - p(\alpha_c)) R \log p}{2(1 - \epsilon)^2}\right\}
\text{ (for any } 0 < z < \epsilon < 1, (1 - z) > \exp\{-(1 - \epsilon)\}\}
= \left(\frac{p}{R}\right)^{(p(\alpha - p(\alpha_c))/2(1 - \epsilon)^2)}.
\tag{A.16}
\]

Combining (A.11), (A.13), (A.16) and using assumption (A.4) we have,

\[
\sum_{\alpha \in A_1} \frac{p(M_\alpha)}{p(M_{\alpha_c})} \frac{m_\alpha(y_n)}{m_{\alpha_c}(y_n)} \leq C \sum_{\alpha \in A_1} \left(\frac{\nu + p}{n^2 \nu}\right)^{(p(\alpha) - p(\alpha_c))/2}
\leq C \sum_{q=1}^{p-p(\alpha_c)} \left(\frac{p - p(\alpha_c)}{q}\right) \left(\frac{\sqrt{\nu + p} p^R/(2(1 - \epsilon)^2)}{n \sqrt{\nu}}\right)^q
\leq C \left\{1 + \left(\frac{\sqrt{\nu + p} p^R/(2(1 - \epsilon)^2)}{n \sqrt{\nu}}\right)^{(p-p(\alpha_c))}\right\}^{-1}
\tag{A.17}
\]

The above expression converges to 0 as \(n \to \infty\), if the first term in the curly braces converges to 1. If \(\nu = 1\) and \(p = O(n^b)\) then the first term is less than \(\left(1 + C' n^{-[-1-(R+1)/2(1 - \epsilon)^2]}\right)^{kn^b}\) for some positive constants \(C'\) and \(k\), any \(R > 2\) and any \(\epsilon > 0\). Also if \(\nu = p\) then this term is less than \(\left(1 + C'' n^{-[-1-Rb/(2(1 - \epsilon)^2)]}\right)^{kn^b}\) for some positive constants \(C''\) and \(k\), any \(R > 2\) and any \(\epsilon > 0\). Letting \(R \downarrow 2\) and \(\epsilon \downarrow 0\), we conclude that the last expression in (A.17) converges to 0 if \(b < 2/5\) when \(\nu = 1\) and if \(b < 1/2\) when \(\nu = p\).

\(\text{Case II. } M_{\alpha_c} = M_N.\) When the null model is true, the Bayes factor of any model with respect to the null model is given by,

\[
\frac{m_\alpha(y_n)}{m_N(y_n)} = \int_0^\infty (1 + g)^{(n - p(\alpha) - 1)/2} \left\{1 + g(1 - R^2_\alpha)\right\}^{-(n-1)/2} \pi(g)dg, \tag{A.18}
\]
where $\pi(g)$ is as in (2.1). Now
\[
\left[ \frac{1 + g}{1 + (1 - R_\alpha^2)g} \right]^{(n-1)/2} = \exp \left[ \frac{n - 1}{2} \left\{ \ln(1 + g) - \ln(1 + (1 - R_\alpha^2)g) \right\} \right] \\
= \exp \left[ \frac{n - 1}{2} \left\{ \ln(1 + g) - \ln(1 + g) + \frac{R_\alpha^2 g}{1 + g^*} \right\} \right] \\
\leq \exp \left[ \frac{n - 1}{2} \left\{ \frac{R_\alpha^2}{(1 - R_\alpha^2) + 1/g} \right\} \right] \\
\leq \exp \left[ \frac{n - 1}{2} \left\{ \frac{R_\alpha^2}{1 - R_\alpha^2} \right\} \right] \\
= \exp \left[ \frac{n - 1}{2} \left\{ \frac{e_n'(P_n(\alpha) - P_n(\alpha_c))e_n}{e_n'(I - P_n(\alpha))e_n} \right\} \right]
\]

since $nS_y^2 = e_n' e_n - ne^2$ and $ne^2 = e_n' P_n(\alpha) e_n$ when the null model is true.

Next we use the facts that with probability tending to 1,
\[
e^{'n}(I - P_n(\alpha))e_n > (n - 1)\sigma^2(1 - \epsilon), \text{ for any } \epsilon > 0 \text{ and for any } R > 2, \max_{\alpha \in A} e_n'(P_n(\alpha)e_n)/\sigma^2 < Rp(\alpha) \ln(p).\]
Combining these, we have for any $\epsilon > 0$,
\[
\left[ \frac{1 + g}{1 + (1 - R_\alpha^2)g} \right]^{(n-1)/2} \leq p^{Rp(\alpha)/2(1-\epsilon)}
\]

with probability tending to 1.

Hence from (A.18) we have,
\[
m_{\alpha}(y_n) \leq p^{Rp(\alpha)/2(1-\epsilon)} \int_0^{\infty} (1 + g)^{-p(\alpha)/2}\pi(g)dg.
\]

Now with $\pi(g)$ as given in (2.1)
\[
I = \int_0^{\infty} (1 + g)^{-p(\alpha)/2}\pi(g)dg \leq \frac{(\tau^2 \nu/2)^{\nu/2}}{\Gamma(\nu/2)} \int_0^{\infty} e^{-\tau^2g^2/2}(p(\alpha) + \nu)/\Gamma(\nu/2)\, dg
\]

by the fact that $(1 + g)^{-1} > g^{-1}$. We then have
\[
I = \begin{cases} 
\frac{(\tau^2/2)^{-p(\alpha)/2}}{\Gamma((p(\alpha) + 1)/2)} \Gamma(1/2) & \text{for } \nu = 1, \\
\frac{(\tau^2 p/2)^{-p(\alpha)/2}}{\Gamma((p(\alpha) + p)/2)} \Gamma(p/2) & \text{for } \nu = p.
\end{cases}
\]
To evaluate these terms, we again use the results of Wendel (1948) stated above. After little algebra one can show

\[ I < \begin{cases} \frac{(\tau^2/2)^{-\alpha/2}}{2} \left( \frac{p}{2} \right)^{\alpha/2} & \text{for } \nu = 1, \\ \frac{(\tau^2 p/2)^{-\alpha/2}}{2} & \text{for } \nu = p. \end{cases} \]

By assumption (A.4) and putting \( \tau = n \), it follows from (A.19) that

\[
\sum_{\alpha \in (A - \{\alpha_c\})} \frac{p(M_{\alpha}) m_{\alpha}(y_n)}{p(M_{\alpha_c}) m_{\alpha_c}(y_n)} < \begin{cases} C \left( \frac{1}{2} \right)^{1+R/(1-\epsilon)/n^2} \left( \frac{p}{2} \right)^{\alpha/2} / \sqrt{n} & \text{for } \nu = 1, \\ C \left( \frac{1+\sqrt{2}^{p R/(2(1-\epsilon))/n^2} \right)^{\alpha/2} & \text{for } \nu = p. \end{cases}
\]

for any \( R > 2 \) and any \( \epsilon > 0 \).

As before, we let \( R \downarrow 2 \) and \( \epsilon \downarrow 0 \) and observe that the above quantity converges to 0 when \( p \) is of order \( n^b \) if \( b < 2/5 \) for \( \nu = 1 \) and \( b < 1/2 \) for \( \nu = p \).

\[ \square \]

**A.4. Proof of Theorem 3.2**

We proceed as in the proof of the Theorem 3.1, and prove (A.10) with \( A_i \) replaced by its analog for nested models, \( A_i^* \). As before we consider separately the cases when the true model is null and when it is non-null.

Case-I: \( M_{\alpha_c} \neq M_N \). First we consider \( i = 2 \). Recall the proof of Theorem 3.1. It can easily be seen that this part in Theorem 3.1 is proved for the model space \( A_2 \) and without using the assumption of normality. Since \( A_2^* \) is a proper subset of \( A_2 \), here also the same proof holds.

Next consider (A.10) with \( A_i \) replaced by \( A_i^* \) and \( i = 1 \). Here we use Lemma A.3 which is equivalent to (A.15) when nested models are considered. This implies that (A.16) also holds here. Combining (A.11), (A.13) and (A.16) we have, for any \( R > 0 \) and any \( \epsilon > 0 \),
\[
\sum_{\alpha \in A^*_i} \frac{p(M_\alpha)}{p(M_{c_\alpha})} \frac{m_\alpha(y_n)}{m_{c_\alpha}(y_n)} 
\leq C \sum_{\alpha \in A_i} \left( \frac{(\nu + p)p^{R/(1-\epsilon)^2}}{n^2 \nu} \right)^{p(\alpha) - p(c_\alpha)}/2 
\leq C \sum_{q=1}^{p-p(c_\alpha)} \left( \frac{\sqrt{\nu + p} \ p^{R/2(1-\epsilon)^2}}{n^{\sqrt{\nu}}} \right)^q 
\leq C \left( \frac{\sqrt{\nu + p} \ p^{R/2(1-\epsilon)^2}}{n^{\sqrt{\nu}}} \right) \frac{1}{(1 - \sqrt{\nu + p} \ p^{R/2(1-\epsilon)^2}/n^{\sqrt{\nu}})}. \tag{A.21} \]

For suitably chosen \( R \) and \( \epsilon \), it can be easily seen that (A.21) converges to 0 for any \( 0 < b < 1 \).

**Case II.** \( M_{c_\alpha} = M_N \). We proceed as in Theorem 3.1. Observe that using Lemma A.3 we obtain the inequality similar to (A.20) with \( A \) replaced by \( A^* \). Thus we have

\[
\sum_{\alpha \in (A^* - \{c_\alpha\})} \frac{p(M_\alpha)}{p(M_{c_\alpha})} \frac{m_\alpha(y_n)}{m_{c_\alpha}(y_n)} 
\geq \left\{ \begin{array}{ll}
C \sum_{\alpha \in (A^* - \{c_\alpha\})} (p^{1+R/(1-\epsilon)/n^2})^{p(\alpha)/2} / \sqrt{\pi} & \text{for } \nu = 1, \\
C \sum_{\alpha \in (A - \{c_\alpha\})} (2p^{R/(1-\epsilon)/n^2})^{p(\alpha)/2} & \text{for } \nu = p.
\end{array} \right.
\]

\[
\left\{ \begin{array}{ll}
C \left\{ p^{1/2+R/(2(1-\epsilon))} / \left( n - p^{1/2+R/(2(1-\epsilon))} \right) \right\} & \text{for } \nu = 1, \\
C \left\{ \sqrt{2p^{R/(2(1-\epsilon))} / \left( n - \sqrt{2p^{R/(2(1-\epsilon))}} \right) \right\} & \text{for } \nu = p.
\end{array} \right.
\]

One can choose \( R > 0 \) and \( \epsilon > 0 \) suitably and show that the above quantities go to 0 for any \( 0 < b < 1 \). \( \square \)

**A.5. Proof of Theorem 4.1**

Let \( M_{c_\alpha} = M_N \). From (A.18) and (A.4) we have

\[
\sum_{\alpha \in A - \{c_\alpha\}} \frac{p(M_\alpha)}{p(M_{c_\alpha})} \frac{m_\alpha(y_n)}{m_{c_\alpha}(y_n)} \geq \frac{1}{C} \sum_{\alpha} \int_0^{\infty} (1 + g)^{-p(\alpha)/2} \pi(g) dg \tag{A.22}
\]
where $\pi(g)$ is given by (4.1). Putting the prior we get the R.H.S. of the above expression as

$$\frac{1}{C} \sum_{\alpha \in A - \{\alpha_c\}} \frac{\Gamma(\gamma_0 + \gamma_1)\Gamma(\gamma_1 + p(\alpha)/2)}{\Gamma\gamma_1 \Gamma(\gamma_0 + \gamma_1 + p(\alpha)/2)}.$$

Using the inequality of Wendel (1948) stated above and the fact that $\gamma_1 > \epsilon$ for some $\epsilon > 0$ free of $n$, it can be shown that for some constant $C' > 0$, the above expression is bigger than

$$C' \sum_{\alpha \in A - \{\alpha_c\}} \left( \frac{\gamma_1}{\gamma_0 + \gamma_1} \right)^{p(\alpha)/2} = C' \left\{ \left( 1 + \sqrt{\frac{\gamma_1}{\gamma_0 + \gamma_1}} \right)^p - 1 \right\}.$$

Thus if $p = nb$, the R.H.S. of (A.22) does not go to 0 if $\gamma_0 / \gamma_1 = O(nb)$. \qed

### A.6. Proof of Theorem 5.1

Our model selection criterion is to choose a model $\hat{\alpha}$ in the model space $A$, which maximizes $p(M{\alpha}) m_{\alpha}(y_n)$ with respect to $\alpha$. Now, from Result 2.1, this is equivalent to maximizing

$$p(M_{\alpha}) \Gamma \left( \frac{\nu + p(\alpha)}{2} \right) \{n\mathcal{S}_y^2(1 - R_\alpha^2)\}^{- (n-1)/2} \left( \frac{n^2 \nu}{2} \right)^{-p(\alpha)/2} (1 + \varepsilon_n(\alpha)),$$

where $|\varepsilon_n(\alpha)| = O(n)/(n\nu)$ uniformly in $\alpha$. We omit the other terms involved in the approximation of Result 2.1, since those are free of $\alpha$. Maximizing the above is equivalent to minimizing

$$\left[ p(M_{\alpha}) \Gamma \left( \frac{\nu + p(\alpha)}{2} \right) (1 + \varepsilon_n(\alpha)) \right]^{-2/(n-1)} \left( \frac{n^2 \nu}{2} \right)^{p(\alpha)/(n-1)} n\mathcal{S}_y^2(1 - R_\alpha^2)$$

with respect to $\alpha$. From (5.2) we have $n\mathcal{S}_y^2(1 - R_\alpha^2) = C_n + 2\sigma^2 D_n(\alpha)(1 + \xi_n(\alpha))$, where $C_n = e_n^t e_n$ and $\xi_n(\alpha) = \{2\mu_n^t(I - P_n(\alpha))e_n - e_n^t P_n(\alpha)e_n\}/(2\sigma^2 D_n(\alpha)).$

Hence, if $M_{\hat{\alpha}}$ is the model for which (A.23) is minimized, then

$$\frac{D_n(\hat{\alpha})}{D_n(\alpha)} \leq \frac{C_n(b_n(\alpha) - 1)}{2\sigma^2 D_n(\alpha)(1 + \xi_n(\hat{\alpha}))} + \frac{b_n(\alpha)(1 + \xi_n(\alpha))}{(1 + \xi_n(\hat{\alpha}))},$$
where

\[ b_n(\alpha) = \left( \frac{P(M_\hat{\alpha})}{p(M_\alpha)} \right)^{2/(n-1)} \left( \frac{\Gamma\{(n + p(\hat{\alpha})/2)}{\Gamma\{(n + p(\alpha)/2}\} \right)^{2/(n-1)} \frac{\left(\frac{n^2 \nu}{2}\right)^{(p(\hat{\alpha})-p(\alpha))/(n-1)}}{\left(\frac{1 + \varepsilon_n(\hat{\alpha})}{1 + \varepsilon_n(\alpha)}\right)^{2/(n-1)}} \]

Therefore, if \( \xi_n = \max_\alpha |\xi_n(\alpha)| \), we have

\[ 1 \leq \frac{D_n(\hat{\alpha})}{\min_\alpha D_n(\alpha)} \leq \frac{C_n}{2n\sigma^2(1 - \xi_n)} \times \max_\alpha \frac{n(b_n(\alpha) - 1)}{D_n(\alpha)} + \frac{1 + \xi_n}{(1 - \xi_n)} \times \max_\alpha b_n(\alpha). \]

The rest of the proof will follow from the following facts

\[ C_n/n \overset{p}{\to} \sigma^2, \]

\[ \xi_n \overset{p}{\to} 0, \]

\[ \max_\alpha \frac{n(b_n(\alpha) - 1)}{D_n(\alpha)} \overset{p}{\to} 0, \]

and \[ \max_\alpha b_n(\alpha) \overset{p}{\to} 1. \]

The proof of (A.26) is straightforward. To prove (A.27) we note that

\[ \xi_n \leq \frac{2\max_\alpha \mu'_n(I - P_n(\alpha))\epsilon_n/n - \min_\alpha \epsilon_n'P_n(\alpha)e_n/n}{2\min_\alpha \sigma^2 D_n(\alpha)/n} \leq \frac{O_p(\sqrt{p/n})}{\delta/n^s}, \]

from (iv) and (v) of Lemma A.1 and assumption (A.3). Clearly if \( s < (1 - b)/2 \), (A.27) holds.

Next we prove (A.29). We show that \( \log(\max_\alpha b_n(\alpha)) = \max_\alpha \log(b_n(\alpha)) \overset{p}{\to} 0 \). From (A.24) we have

\[ \max_\alpha \log b_n(\alpha) \leq \frac{2}{n - 1} \left\{ \log \left( \max_\alpha \frac{P(M_\hat{\alpha})}{p(M_\alpha)} \right) + \log \left( \frac{\Gamma\{(n + p(\hat{\alpha})/2)}{\Gamma\{(n + p(\alpha)/2}\} \right) \right. \]

\[ + \log \left( \frac{1 + \varepsilon_n(\hat{\alpha})}{1 - \max_\alpha |\varepsilon_n(\alpha)|} \right) \right\} + \max_\alpha \frac{p(\hat{\alpha}) - p(\alpha)}{n - 1} \log \left( \frac{n^2 \nu}{2} \right) \]

\[ \leq \frac{2}{n - 1} \left\{ \log C + \frac{p}{2} \log \left( \frac{p + \nu}{2} \right) + \log \left( \frac{1 + pO_p(1)/(n\nu)}{1 - pO_p(1)/(n\nu)} \right) \right\} \]

\[ + \frac{p}{n - 1} \log \left( \frac{n^2 \nu}{2} \right), \]
form (A.13) and assumption (A.4). It is now easy to show that when \( p = O(n^b) \), \( 0 < b < 1 \), the above expression is \( O_p \left( n^{-(1-b)} \log(n) \right) \) and converges to 0 with probability tending to 1. Hence (A.29) holds.

Finally we prove (A.28). By mean value theorem, for some \( z > 0 \), \((e^z - 1) = ze^{z^*} < ze^z\), where \( z^* \in [0, z] \). Replacing \( z \) by \( \log b_n(\alpha) \) we get

\[
\max_{\alpha} (b_n(\alpha) - 1) \leq \max_{\alpha} \log b_n(\alpha) \exp\left\{ \max_{\alpha} \log b_n(\alpha) \right\}.
\]

Thus by assumption (A.3) we have

\[
\max_{\alpha} \frac{n(b_n(\alpha) - 1)}{D_n(\alpha)} \leq \frac{\max_{\alpha} (b_n(\alpha) - 1)}{\min_{\alpha} D_n(\alpha)/n} \leq n^* O_p \left( n^{-(1-b)} \log(n) \right) \exp \left\{ O_p \left( n^{-(1-b)} \log(n) \right) \right\},
\]

which is going to 0, with probability tending to 1.

Acknowledgements. I am thankful to my Ph.D. supervisor Professor Tapas Samanta for his guidance. I am also thankful to Professor Arijit Chakrabarti for his insightful comments. Their suggestions have helped to improve this paper substantially.

References

Bartlett, M. (1957). A comment on D. V. Lindleys statistical paradox. *Biometrika* **44** 533-534.

Bayarri, M. J. and García-Donato, G. (2008). Generalization of Jeffreys divergence-based priors for Bayesian hypothesis testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 981–1003. MR2530326

Bayarri, M. J., Berger, J. O., Forte, A. and García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *Ann. Statist.* **40** 1550–1577. MR3015035

Berger, J. O. and Pericchi, L. R. (2001). Objective Bayesian methods for model selection: introduction and comparison. In *Model selection*. IMS Lecture Notes Monogr. Ser. **38** 135–207. Inst. Math. Statist., Beachwood, OH. With discussion by J. K. Ghosh, Tapas Samanta, and Fulvio De Santis, and a rejoinder by the authors. MR2000753

Chakrabarti, A. and Ghosh, J. K. (2006). A generalization of BIC for the general exponential family. *J. Statist. Plann. Inference* **136** 2847–2872. MR2281234
CHAKRABARTI, A. and SAMANTA, T. (2008). Asymptotic optimality of a cross-validatory predictive approach to linear model selection. In Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh. Inst. Math. Stat. Collect. 3 138–154. Inst. Math. Statist., Beachwood, OH. MR2459222 (2010a:62211)

CHIPMAN, H., GEORGE, E. I. and MCCULLOCH, R. E. (2001). The practical implementation of Bayesian model selection. In Model selection. IMS Lecture Notes Monogr. Ser. 38 65–134. Inst. Math. Statist., Beachwood, OH With discussion by M. Clyde, Dean P. Foster, and Robert A. Stine, and a rejoinder by the authors. MR2000752

CUI, W. and GEORGE, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. J. Statist. Plann. Inference 138 888–900. MR2416869

DAVID, A. P. (1992). Prequential analysis, stochastic complexity and Bayesian inference. In Bayesian statistics, 4 (Peñíscola, 1991) 109–125. Oxford Univ. Press, New York. MR1380273 (96j:62005)

FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. Ann. Statist. 32 928–961. MR2065194 (2005g:62047)

FERNÁNDEZ, C., LEY, E. and STEEL, M. F. J. (2001). Benchmark priors for Bayesian model averaging. J. Econometrics 100 381–427. MR1820410

FOSTER, D. P. and GEORGE, E. I. (1994). The risk inflation criterion for multiple regression. Ann. Statist. 22 1947–1975. MR1329177 (96c:62119)

GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. Biometrika 87 731–747. MR1813972 (2001k:62007)

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. J. Amer. Statist. Assoc. 88 881–889.

HANS, C. (2009). Bayesian lasso regression. Biometrika 96 835–845. MR2564494 (2010m:62202)

JEFFREYS, H. (1961). Theory of probability. Third edition. Clarendon Press, Oxford. MR0187257 (32 #4710)

JOHNSON, V. E. and ROSELL, D. (2012). Bayesian model selection in high-dimensional settings. J. Amer. Statist. Assoc. 107 649–660. MR2980074

KASS, R. and RAFTERY, A. (1995). Bayes Factors. J. Amer. Statist. Assoc. 90 773–795.

KASS, R., TIERNEY, L. and KADANE, J. (1990). The Validity of Posterior Expansions Based on Laplace’s Method. Bayesian and Likelihood Methods in Statistics and Economics.

LEY, E. and STEEL, M. F. J. (2012). Mixtures of g-priors for Bayesian
model averaging with economic applications. *J. Econometrics* **171** 251–266. MR2991863

Li, K.-C. (1987). Asymptotic optimality for $C_p$, $C_L$, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.* **15** 958–975. MR902239 (89c:62112)

Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103** 410–423. MR2420243

Maruyama, Y. and George, E. I. (2011). Fully Bayes factors with a generalized g-prior. *Ann. Statist.* **39** 2740–2765. MR2906885 (2012m:62022)

Moreno, E. (1997). Bayes factors for intrinsic and fractional priors in nested models. Bayesian robustness. In *L1-statistical procedures and related topics (Neuchâtel, 1997)*. IMS Lecture Notes Monogr. Ser. **31** 257–270. Inst. Math. Statist., Hayward, CA. MR1833592

Moreno, E., Girón, F. J. and Casella, G. (2010). Consistency of objective Bayes factors as the model dimension grows. *Ann. Statist.* **38** 1937–1952. MR2676879 (2011f:62016)

Mukhopadhyay, M., Samanta, T. and Chakrabarti, A. (2014). On consistency and optimality of Bayesian variable selection based on g-priors in normal linear regression models. *Annals of Institute of Statistical Mathematics.*

Park, T. and Casella, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. MR2524001

Shang, Z. and Clayton, M. K. (2011). Consistency of Bayesian linear model selection with a growing number of parameters. *J. Statist. Plann. Inference* **141** 3463–3474. MR2817355

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7** 221–264. With comments and a rejoinder by the author. MR1466682 (99m:62104)

Sparks, D. K., Khare, K. and Ghosh, M. (2012). Necessary and sufficient conditions for high-dimensional posterior consistency under g-priors. Submitted to *Annales de l’Institut Henri Poincaré - Probabilités et Statistiques.*

Wang, M. and Sun, X. (2014). Bayes factor consistency for nested linear models with a growing number of parameters. *J. Statist. Plann. Inference* **147** 95–105. MR3151848

Wendel, J. G. (1948). Note on the gamma function. *Amer. Math. Monthly* **55** 563–564. MR0029448 (10,605a)
ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti (P. K. Goel and A. Zellner, eds.)* 233-243.

ZELLNER, A. and SIOW, A. (1980). Posterior odds ratios for selected regression hypotheses. *Bayesian Statistics Proceedings of the first international meeting held in Valencia* 585603.