Ann Inst Stat Math (2015) 67:963–997
DOI 10.1007/s10463-014-0483-8

On consistency and optimality of Bayesian variable selection based on g-prior in normal linear regression models

Minerva Mukhopadhyay · Tapas Samanta · Arijit Chakrabarti

Received: 7 August 2013 / Revised: 22 May 2014 / Published online: 24 September 2014

Abstract Consider Bayesian variable selection in normal linear regression models based on Zellner's g-prior. We study theoretical properties of this method when the sample size n grows and consider the cases when the number of regressors, p is fixed and when it grows with n. We first consider the situation where the true model is not in the model space and prove under mild conditions that the method is consistent and "loss efficient" in appropriate sense. We then consider the case when the true model is in the model space and prove that the posterior probability of the true model goes to one as n goes to infinity. "Loss efficiency" is also proved in this situation. We give explicit conditions on the rate of growth of g, possibly depending on that of p as n grows, for our results to hold. This helps in making recommendations for the choice of g.

Keywords Model selection consistency · Loss efficiency · Variable selection · g-prior

Electronic supplementary material The online version of this article (doi:10.1007/s10463-014-0483-8) contains supplementary material, which is available to authorized users.

M. Mukhopadhayay (✉) · T. Samanta · A. Chakrabarti
Applied Statistics Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road,
Kolkata 700108, India
e-mail: minervamukherjee@gmail.com

T. Samanta
e-mail: tapas@isical.ac.in

A. Chakrabarti
e-mail: arc@isical.ac.in

1 Introduction

One of the very popular ways to model dependence between a *response* variable and *explanatory* or *predictor* variables in a given problem is through the linear regression model. In such a case, choosing from a set of candidate models is equivalent to the problem of variable selection. This problem has been widely studied in the literature and arises in myriad applications, see, for example, Shao (1997), George (2000) and Miller (2001) and the references therein.

We consider the regression problem with response variable y and a set of potential predictor variables $x_1, x_2, \dots, x_p$. Let $\mathbf{y}_n = (y_1, y_2, \dots, y_n)'$ be a vector of observations on the response variable and $\mathbf{X}_n = \big(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p\big)$ be an $n \times p$ design matrix. Here $\mathbf{x}_i$ is an $n \times 1$ vector of observations on the ith regressor x_i and the jth component of $\mathbf{x}_i$ is associated with y_j, $i = 1, \dots, p$, $j = 1, \dots, n$. We assume, without loss of generality, that the columns of $\mathbf{X}_n$ have been centered so that $\mathbf{1}_n'\mathbf{x}_i = 0$ for all i where $\mathbf{1}_n$ is a vector of 1's of length n. Let $\boldsymbol{\mu}_n$ denote $E(\mathbf{y}_n|\mathbf{X}_n)$ and assume

$$\mathbf{y}_n \sim N_n\left(\boldsymbol{\mu}_n, \sigma^2 I_n\right),$$

where σ^2 is unknown and I_n is the $n \times n$ identity matrix. We are interested in capturing the functional relationship, if any, between $\boldsymbol{\mu}_n$ and $\mathbf{X}_n$.

We restrict our search within the class of normal linear models under which $\boldsymbol{\mu}_n$ may be expressed as

$$\boldsymbol{\mu}_n = \mathbf{1}_n\beta_0 + \mathbf{X}_n\boldsymbol{\beta}, \tag{1}$$

where β_0 is an intercept and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)'$ is a vector of regression coefficients. Our problem is to select a subset of the potential predictor variables $x_1, x_2, \dots, x_p$. Thus, we have a model selection problem and our model space, denoted by $\boldsymbol{\mathcal{A}}$, may be indexed by α, where each α consists of a subset of size $p(\alpha)$ $(1 \le p(\alpha) \le p)$ of $\{1, 2, \dots, p\}$, indicating which regressors are included in the model. The model M_α corresponding to $\alpha \in \boldsymbol{\mathcal{A}}$ may be expressed as a sub-model of (1),

$$M_\alpha \ \ : \ \ \boldsymbol{\mu}_n = \mathbf{1}_n\beta_0 + \mathbf{X}_{n\alpha}\boldsymbol{\beta}_\alpha, \tag{2}$$

where the intercept β_0 is common to all models, $\mathbf{X}_{n\alpha}$ is a sub-matrix of $\mathbf{X}_n$ consisting of the $p(\alpha)$ columns specified by α and $\boldsymbol{\beta}_\alpha$ is the $p(\alpha)$-dimensional vector of regression coefficients.

Bayesian model selection requires specification of prior distribution of the parameters $\boldsymbol{\theta}_\alpha = (\beta_0, \boldsymbol{\beta}_\alpha, \sigma^2) \in \Theta_\alpha$ under each model M_α and prior probabilities $p(M_\alpha)$ of the models. Let $p(\mathbf{y}_n|\boldsymbol{\theta}_\alpha, M_\alpha)$ denote the density of $\mathbf{y}_n$ given $\boldsymbol{\theta}_\alpha$ under M_α and $p(\boldsymbol{\theta}_\alpha|M_\alpha)$ denote the prior density of $\boldsymbol{\theta}_\alpha$ under M_α. Then the posterior probability of the model M_α, $\alpha \in \boldsymbol{\mathcal{A}}$, is given by

$$p(M_\alpha|\mathbf{y}_n) = \frac{p(M_\alpha)m_\alpha(\mathbf{y}_n)}{\sum_{\alpha\in\boldsymbol{\mathcal{A}}} p(M_\alpha)m_\alpha(\mathbf{y}_n)}, \tag{3}$$

$$\text{where} \ \ m_\alpha(\mathbf{y}_n) = \int p(\mathbf{y}_n|\boldsymbol{\theta}_\alpha, M_\alpha)p(\boldsymbol{\theta}_\alpha|M_\alpha)d\boldsymbol{\theta}_\alpha \tag{4}$$

is the marginal density of $\mathbf{y}_n$ under M_α. In this paper, we consider the model selection procedure that selects the model with highest posterior probability.

A very popular conventional prior for the parameters $\boldsymbol{\beta}_\alpha$ is the conjugate g-prior due to Zellner (1986) given in (6). In the present scenario, β_0 and σ^2 may be regarded as parameters common to all the models and the suggested default priors are

$$p(\beta_0, \sigma^2 | M_\alpha) = \frac{1}{\sigma^2} \tag{5}$$

$$\boldsymbol{\beta}_\alpha | \beta_0, \sigma^2, M_\alpha \sim N_{p(\alpha)}(\mathbf{0}, g\sigma^2(\mathbf{X}'_{n\alpha}\mathbf{X}_{n\alpha})^{-1}) \tag{6}$$

for some $g > 0$ [see, for example, (Liang et al. 2008, Section 2.1)].

A major advantage of Zellner's g-prior is the availability of closed-form expressions of the marginal likelihoods $m_\alpha(\mathbf{y}_n)$ and the resulting computational efficiency. It may be noted that the prior covariance matrix is related to the Fisher information matrix in the linear model. This prior and its variants have been widely used in the literature in linear models; see, for example, Zellner (1986), Chaturvedi et al. (1997), Fernández et al. (2001), Consonni and Veronese (2008), Krishna et al. (2009) and Bornn et al. (2010).

It has been shown in George and Foster (2000) that g in the g-prior can be properly calibrated so that model selection using the g-prior is equivalent to that using the Akaike information criterion (AIC) or Bayesian information criterion (BIC) or the risk information criterion (RIC). There have been many suggestions regarding the proper choice of g based on various considerations, see, for example, the unit information prior of (Kass and Wasserman 1995) taking $g = n$, Benchmark prior of (Fernández et al. 2001) taking $g = \max(n, p^2)$ and choices of g coming out of local and global empirical Bayes estimation of g (see, e.g., Liang et al. 2008, Section 2.4).

In this paper, we study the performance of the Bayesian variable selection method based on g-prior when n is large. We consider a general setting which simultaneously allows the potential number of regressors p to remain fixed or grow with n. We also consider both the cases when the "true" model lies in the model space and when it does not. The main objective of the present work is to find conditions under which the model selection rule based on g-prior has some natural desirable properties. We have been able to come up with sufficient conditions under which such properties hold true. We use these results and simulations to make a recommendation for a suitable choice of g.

We first consider in Sect. 2 the case when the "true" model is not in the space of models from which we are selecting one. We refer to this as the "model false" case. This is the more realistic situation, the true model not being one of the entertained models. We are not aware of any work related to g-prior that considers the "model false" case. We first show in Sect. 2.1 that the method based on g-prior chooses the model which asymptotically minimizes the distance from the unknown true model, for some appropriate measure of distance between models [see (Chakrabarti and Ghosh 2006), for a related work]. We define this property as consistency in the "model false" case. This property is shown to hold for a wide range of choices of g and p. Following Li (1987) and Shao (1997), we next show in Sect. 2.2 that the method based on g-

prior has another optimality property which we call “Loss Efficiency” as in Shao (1997). As in Li (1987) and Shao (1997), we consider as our evaluation criterion the average squared error deviation between the true and estimated regression function, given by $L_n(\alpha) = n^{-1} \left\| \boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n(\alpha) \right\|^2$, where $\hat{\boldsymbol{\mu}}_n(\alpha)$ is the chosen estimator of the true regression function $\boldsymbol{\mu}_n$ when model M_α is selected. We show in Sect. 2.2 that for common choices of $\hat{\boldsymbol{\mu}}_n(\alpha)$, the ratio of $L_n(\hat{\alpha})$ and $\min_{\alpha \in \boldsymbol{\mathcal{A}}} L_n(\alpha)$ goes to 1 in probability as $n \to \infty$ when $M_{\hat{\alpha}}$ is the model chosen by the method under study. This property is referred to as “Loss Efficiency”.

In this paper, we are mainly concerned with normal linear models and also in the “model false” case, we assume that $\mathbf{y}_n$ is normally distributed. However, in Sects. 2.1.1 and 2.2.1 we make a modest attempt to study consistency and loss efficiency under a particular setup for the case when the true distribution is not normal.

In Sect. 3, we consider the “model true” case, that is, the case where the true model is one of the entertained models. We show in Sect. 3.1 that the posterior probability of the true model goes to *one* as $n \to \infty$ if one uses the g-prior. We call this property *model selection consistency* following Liang et al. (2008). We find explicit conditions on the rate of growth of g, depending on that of p, as n grows, under which the result holds. In Sect. 3.2, we also show “loss efficiency” of the model selection procedure based on g-prior in the “model true” case.

Model selection consistency for g-prior was studied, among others, by Fernández et al. (2001) for the case when p is fixed. Shang and Clayton (2011) considered the case when p grows with n and proved consistency for a prior that can be related to the g-prior in some particular cases. Liang et al. (2008) studied model selection consistency for mixtures of g-priors for the case when p is fixed.

In Sect. 4, we present simulation results which demonstrate that our theoretical results on loss efficiency and consistency can be greatly relied on even for moderate sample sizes compared to the model dimension p. Along with different possible choices of n and p, we also consider different choices of g to understand its role in the performance of the method.

Section 5 presents a discussion on the choice of g in the model selection procedure under study. The arguments based on our theoretical results and simulations lead to a recommendation for choice of g in this section.

In Sect. 6, we summarize our results and include some related discussions. Scope of future research is also explored.

Proofs of all the main results, except Theorems 1 and 6 are given in the Appendix (Sect. 7) and those of some other results are given in the supplementary file.

2 The “Model False” case

In this section, we consider the case when the true model is not in the model space $\{M_\alpha, \alpha \in \boldsymbol{\mathcal{A}}\}$ from which we are selecting one. This is referred to as the “model false” case and is indeed the case in almost all practical situations.

We consider the situation where $\mathbf{y}_n$ may be assumed to be normally distributed but the true regression function $\boldsymbol{\mu}_n$ is not expressible as a linear combination of some of the columns of $\mathbf{X}_n$ as stated in (2).

Given the priors (5) and (6), the marginal likelihood under the model $M_\alpha, \alpha \in \boldsymbol{\mathcal{A}}$, is given by

$$m_\alpha(\mathbf{y}_n) = \frac{\Gamma(n-1)/2}{\pi^{(n-1)/2}\sqrt{n}\,(1+g)^{p(\alpha)/2}} \times \left[(1-a)\sum_{i=1}^{n}(y_i-\overline{y})^2 + a\mathbf{y}_n'(I_n - P_n(\alpha))\mathbf{y}_n\right]^{-(n-1)/2} \tag{7}$$

where $a = g/(1+g)$ and $P_n(\alpha) = \mathbf{Z}_{n\alpha}\left[\mathbf{Z}_{n\alpha}'\mathbf{Z}_{n\alpha}\right]^{-1}\mathbf{Z}_{n\alpha}'$ is the projection matrix onto the span of $\mathbf{Z}_{n\alpha} = [\mathbf{1}_n, \mathbf{X}_{n\alpha}], \alpha \in \boldsymbol{\mathcal{A}}$. If $g = g_n$ varies with n, we write $a_n = g_n/(1+g_n)$. The model selection rule is to choose the model M_α with highest posterior probability, that is, we choose the model M_α for which $p(M_\alpha)m_\alpha(\mathbf{y}_n)$ is the largest among all $\alpha \in \boldsymbol{\mathcal{A}}$. We note that maximizing $p(M_\alpha)m_\alpha(\mathbf{y}_n)$ with respect to α is equivalent to minimizing

$$\Psi(\alpha) = [p(M_\alpha)]^{-2/(n-1)}\,(1+g)^{p(\alpha)/(n-1)} \times \left[(1-a)\sum_{i=1}^{n}(y_i-\overline{y})^2 + a\mathbf{y}_n'(I_n - P_n(\alpha))\mathbf{y}_n\right]. \tag{8}$$

In this section, we show that under certain conditions, the above model selection procedure asymptotically performs as well as an *Oracle*. By an Oracle we mean an imaginary model selection procedure which behaves optimally in some sense using the knowledge of the true regression function which is not known to us.

We now state the assumptions under which we prove the results. Throughout this paper we assume

(A.1) $\boldsymbol{\mu}_n'\boldsymbol{\mu}_n = O(n)$ as $n \to \infty$.

For the case when p is fixed, we assume

(A.2) $\varliminf_{n\to\infty} \min_{\alpha\in\boldsymbol{\mathcal{A}}} \frac{1}{n}\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\boldsymbol{\mu}_n > \Delta$ for some constant $\Delta > 0$.

For the case when $p = p_n$ grows with n, we replace assumption (A.2) by

(A.2)* $\varliminf_{n\to\infty} n^s \min_{\alpha\in\boldsymbol{\mathcal{A}}} \boldsymbol{\mu}_n'(I_n - P_n(\alpha))\boldsymbol{\mu}_n/n > \delta$ for some constants $\delta > 0$ and $0 \leq s < 1$.

In this case, we also assume

(A.3) The prior probabilities $p(M_\alpha)$'s satisfy $\max_{\alpha,\alpha'\in\boldsymbol{\mathcal{A}}} p(M_\alpha)/p(M_{\alpha'}) \leq \mathrm{e}^{n^\beta}$ for some $\beta < (1-s)$ where s is as in (A.2)*.

Remark 1 Assumption (A.1) holds if the μ_i's are of comparable magnitude and they do not grow too fast with i. This holds, for example, when the sequence $\{\mu_1, \mu_2, \ldots, \mu_n\}$ is bounded.

Bayarri et al. (2012) describe (A.2) as a key assumption for consistent model selection under which the models are differentiated in some sense; see also Shao (1997, p. 225), Fernández et al. (2001) and Liang et al. (2008, p. 416) who make this assumption when p is fixed. Assumption (A.2)* seems to us to be a natural extension of assumption (A.2) for the case when p grows with n. Assumption (A.3) is satisfied for a very large class of probabilities on the model space.

2.1 Consistency

We show that the model selection procedure under study chooses a model that is, in some asymptotic sense, closest to the unknown true model among all the models in the model space $\boldsymbol{\mathcal{A}}$. We define this properly as consistency in the "model false" case. Below, we will consider the Kullback–Leibler distance between two probability distributions. The Kullback–Leibler distance between the true distribution $N(\boldsymbol{\mu}_n, \sigma^2 I_n)$ and the distribution $N(\mathbf{1}_n\beta_0 + \mathbf{X}_{n\alpha}\boldsymbol{\beta}_\alpha, \sigma^2 I_n)$ under M_α is

$$\frac{1}{2\sigma^2}\left(\boldsymbol{\mu}_n - \mathbf{1}_n\beta_0 - \mathbf{X}_{n\alpha}\boldsymbol{\beta}_\alpha\right)'\left(\boldsymbol{\mu}_n - \mathbf{1}_n\beta_0 - \mathbf{X}_{n\alpha}\boldsymbol{\beta}_\alpha\right).$$

We define the distance $D_n(\alpha)$ between the true distribution and the model M_α as the minimum of the above distance with respect to $(\beta_0, \boldsymbol{\beta}_\alpha)$ which is given by

$$D_n(\alpha) = \frac{1}{2\sigma^2}\left\|\boldsymbol{\mu}_n - P_n(\alpha)\boldsymbol{\mu}_n\right\|^2 = \frac{1}{2\sigma^2}\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\boldsymbol{\mu}_n. \tag{9}$$

One would naturally like to choose a model M_α which is as close as possible to the true distribution, that is, for which $D_n(\alpha) = \min_{\alpha\in\boldsymbol{\mathcal{A}}} D_n(\alpha)$. Obviously, one could find the model M_α for which $D_n(\alpha)$ is minimized if the true distribution were known, which is never the case. We prove here that if $M_{\hat\alpha}$ is the model chosen by our model selection rule, then as $n \to \infty$,

$$\frac{D_n(\hat\alpha)}{\min_{\alpha\in\boldsymbol{\mathcal{A}}} D_n(\alpha)} \xrightarrow{p} 1. \tag{10}$$

In the "model false" case, we say that the model selection rule is consistent if (10) holds.

We now state our result.

Theorem 1 *Consider the model selection rule based on the priors* (5) *and* (6) *that chooses a model with the highest posterior probability and let* $M_{\hat\alpha}$ *be the model chosen by this rule. Let* $g = g_n = kn^r$ *for* $r \geq 0$ *and* $k > 0$. *Then we have the following results.*

- **(a)** *If p, the total number of predictors, is fixed, then under assumptions* (A.1) *and* (A.2), (10) *holds.*
- **(b)** *Suppose that* $p = p_n$ *grows with n. Assume that* $p_n = O(n^b)$ *for* $0 < b < 1$, (A.1) *and* (A.3) *holds and* (A.2)* *holds with* $s < (1-b)/2$. *Then* (10) *holds.*

Let $\mathbf{e}_n = \mathbf{y}_n - \boldsymbol{\mu}_n$. The following lemma will be used to prove the theorem.

Lemma 1 *Under assumption* (A.1), *we have as* $n \to \infty$,

- **(a)** $\max_{\alpha\in\boldsymbol{\mathcal{A}}}\left|\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\mathbf{e}_n\right|/n = O_p\left(\sqrt{p_n/n}\right)$, *and*
- **(b)** $\max_{\alpha\in\boldsymbol{\mathcal{A}}}\mathbf{e}_n' P_n(\alpha)\mathbf{e}_n/n = O_p\left(p_n/n\right)$.

The proof of Lemma 1 is given in the Appendix.

Proof of Theorem 1. We first derive results for the case when $p = p_n$, $p(\alpha) = p_n(\alpha)$, $\alpha \in \boldsymbol{\mathcal{A}}$ and $g = g_n$ may vary with n. This also includes the case with fixed p, $p(\alpha)$ and g.

We note that

$$\begin{aligned} &\mathbf{y}_n'(I_n - P_n(\alpha))\mathbf{y}_n \\ &\quad = \left(\boldsymbol{\mu}_n + \mathbf{e}_n\right)' (I_n - P_n(\alpha)) \left(\boldsymbol{\mu}_n + \mathbf{e}_n\right) \\ &\quad = 2\boldsymbol{\mu}_n'\mathbf{e}_n + \mathbf{e}_n'\mathbf{e}_n + 2\sigma^2 D_n(\alpha) - 2\boldsymbol{\mu}_n' P_n(\alpha)\mathbf{e}_n - \mathbf{e}_n' P_n(\alpha)\mathbf{e}_n \end{aligned} \tag{11}$$

and therefore, $\Psi(\alpha)$, given by (8), can be expressed as

$$\begin{aligned} \Psi(\alpha) &= [p(M_\alpha)]^{-2/(n-1)} (1 + g_n)^{p_n(\alpha)/(n-1)} \\ &\quad \times \left[C_n + 2a_n\sigma^2 D_n(\alpha) + 2a_n\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\mathbf{e}_n - a_n\mathbf{e}_n' P_n(\alpha)\mathbf{e}_n \right] \\ &= [p(M_\alpha)]^{-2/(n-1)} (1 + g_n)^{p_n(\alpha)/(n-1)} \left[\, C_n + 2a_n\sigma^2 D_n(\alpha)(1 + \xi_n(\alpha)) \, \right], \end{aligned} \tag{12}$$

where $C_n = (1 - a_n)\sum_{i=1}^n (y_i - \overline{y})^2 + a_n\mathbf{e}_n'\mathbf{e}_n$, $D_n(\alpha)$ is as defined in (9) and

$$\xi_n(\alpha) = \frac{2\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\mathbf{e}_n - \mathbf{e}_n' P_n(\alpha)\mathbf{e}_n}{2\sigma^2 D_n(\alpha)}. \tag{13}$$

We shall show below that

$$\max_\alpha |\xi_n(\alpha)| \xrightarrow{p} 0. \tag{14}$$

As $\Psi(\hat{\alpha}) \le \Psi(\alpha)$ for all $\alpha \in \boldsymbol{\mathcal{A}}$, from (12) we have, with probability tending to *one* uniformly in $\alpha \in \boldsymbol{\mathcal{A}}$,

$$\frac{D_n(\hat{\alpha})}{D_n(\alpha)} \le \frac{C_n(b_{n\alpha} - 1)}{2a_n\sigma^2 D_n(\alpha)(1 + \xi_n(\hat{\alpha}))} + \frac{(1 + \xi_n(\alpha))}{(1 + \xi_n(\hat{\alpha}))} b_{n\alpha}$$

and therefore,

$$1 \le \frac{D_n(\hat{\alpha})}{\min_\alpha D_n(\alpha)} \le \frac{C_n/n}{2a_n\sigma^2(1 - \xi_n)} \times \max_\alpha \frac{n(b_{n\alpha} - 1)}{D_n(\alpha)} + \frac{1 + \xi_n}{1 - \xi_n} \times \max_\alpha b_{n\alpha}, \tag{15}$$

where $\xi_n = \max_\alpha |\xi_n(\alpha)|$ and

$$b_{n\alpha} = \left(\frac{p(M_\alpha)}{p(M_{\hat{\alpha}})} \right)^{-2/(n-1)} (1 + g_n)^{(p_n(\alpha) - p_n(\hat{\alpha}))/(n-1)}. \tag{16}$$

We can now prove that

$$C_n = O_p(n) \tag{17}$$

$$\text{and} \quad \max_{\alpha} |n(b_{n\alpha} - 1)| \leq 2(p_n \log(1 + g_n) + 2 \log C_{0n}) \\ \times \exp\{(p_n \log(1 + g_n) + 2 \log C_{0n})/(n-1)\}, \tag{18}$$

where $C_{0n} = \max_{\alpha,\alpha' \in \mathcal{A}} p(M_\alpha)/p(M_{\alpha'})$. The calculations that lead to (17) and (18) are given in the Appendix. Also,

$$|\log(\max_{\alpha} b_{n\alpha})| \leq \max_{\alpha} |\log(b_{n\alpha})| \leq \frac{1}{(n-1)} \left[p_n \log(1 + g_n) + 2 \log(C_{0n})\right]. \tag{19}$$

We now prove part **(a)** of the theorem. If $p_n = p$ is fixed, by assumption (A.2),

$$\min_{\alpha} D_n(\alpha) > \frac{n\Delta}{2\sigma^2} \tag{20}$$

for all sufficiently large n, and C_{0n} is a fixed finite number. Therefore, from (18),

$$\max_{\alpha} \frac{n(b_{n\alpha} - 1)}{D_n(\alpha)} \xrightarrow{p} 0. \tag{21}$$

From (19),

$$\max_{\alpha} b_{n\alpha} \xrightarrow{p} 1. \tag{22}$$

Now, from (13) and (20),

$$\max_{\alpha} |\xi_n(\alpha)| \leq \frac{2}{n\Delta} \max_{\alpha} |\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\mathbf{e}_n| + \frac{1}{n\Delta} \max_{\alpha} \mathbf{e}_n' P_n(\alpha)\mathbf{e}_n$$

and hence by Lemma 1, (14) holds. Part **(a)** of the theorem now follows from (14), (15), (17), (21) and (22).

For part **(b)**, we note that by assumption (A.2)*,

$$\min_{\alpha} D_n(\alpha) > \frac{\delta n^{1-s}}{2\sigma^2} \tag{23}$$

for all sufficiently large n and by assumption (A.3), $C_{0n} \leq e^{n^\beta}$ for some $\beta < (1 - s)$. Therefore from (18), for $p = p_n = O(n^b)$ with $(1 - b)/2 > s$, the convergence (21) holds. Obviously, the convergence (22) also holds. From (13) and (23),

$$\max_{\alpha} |\xi_n(\alpha)| \leq \frac{2}{\delta n^{1-s}} \max_{\alpha} |\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\mathbf{e}_n| + \frac{1}{\delta n^{1-s}} \max_{\alpha} \mathbf{e}_n' P_n(\alpha)\mathbf{e}_n$$

and hence by Lemma 1, (14) holds. Part **(b)** of the theorem now follows from (14), (15), (17), (21) and (22). □

2.1.1 The case when the true model is not normal

In this paper, we are mainly concerned with normal linear models and also in the case when the true model is not in the model space considered, assume that $\mathbf{y}_n$ is normally distributed. This is a commonly used assumption for a wide variety of data. The consistency result proved above in this section is based on this assumption.

A natural question is what can be said regarding consistency if the unknown true distribution of $\mathbf{y}_n$ is not normal. We do not try to study this problem in its full generality in this paper. We make a modest attempt and give a proof of consistency for this case under the setup of Li (1987) and Shao (1997) for variable selection in linear regression models where the truth is not necessarily normal. We state below the result obtained by us. The proof is given in the supplementary file.

Let the true distribution of $\mathbf{y}_n$ has a density f. It can be easily seen that the Kullback–Leibler distance between the true distribution given by the density f and the distribution $N\left(\mathbf{1}_n\beta_0 + \mathbf{X}_{n\alpha}\boldsymbol{\beta}_\alpha, \sigma^2 I_n\right)$ under M_α is equal to

$$\int f\left(\mathbf{y}_n\right)\log f\left(\mathbf{y}_n\right)d\mathbf{y}_n + \frac{n}{2}\left(1+\log\sigma^2\right) \\ +\frac{1}{\sigma^2}\left(\boldsymbol{\mu}_n - \mathbf{1}_n\beta_0 - \mathbf{X}_{n\alpha}\boldsymbol{\beta}_\alpha\right)'\left(\boldsymbol{\mu}_n - \mathbf{1}_n\beta_0 - \mathbf{X}_{n\alpha}\boldsymbol{\beta}_\alpha\right).$$

Then, the distance $D_n^*(\alpha)$ between the true model f and the model M_α, obtained by minimizing the above with respect to $(\beta_0, \boldsymbol{\beta}_\alpha)$ is given by

$$D_n^*(\alpha) = \int f\left(\mathbf{y}_n\right)\log f\left(\mathbf{y}_n\right)d\mathbf{y}_n + \frac{n}{2}\left(1+\log\sigma^2\right) + D_n(\alpha), \tag{24}$$

where $D_n(\alpha)$ is as given in (9).

We note that the first two terms in the right hand side of (24) is free of α and therefore, minimizing $D_n^*(\alpha)$ with respect to α is equivalent to minimizing $D_n(\alpha)$ with respect to α. We prove that under certain assumptions

$$\frac{D_n(\hat{\alpha})}{\min_{\alpha\in\boldsymbol{\mathcal{A}}} D_n(\alpha)} \xrightarrow{p} 1 \tag{25}$$

as $n \to \infty$, where $M_{\hat{\alpha}}$ is the model chosen by the model selection rule based on g prior and treat this as our definition of consistency of the model selection rule.

We prove our results under the following assumption made by Li (1987) and Shao (1997) while proving asymptotic validity of various linear model selection procedures when the true model is not necessarily normal. It is assumed that

$$\sum_{\alpha\in\boldsymbol{\mathcal{A}}} \frac{1}{[nR_n(\alpha)]^m} \to 0 \tag{26}$$

as $n \rightarrow \infty$, for some positive integer m for which $E\left(\mathrm{e}_1^{4m}\right) < \infty$ where $\boldsymbol{\mathcal{A}}$ is the class of models considered and for $\alpha \in \boldsymbol{\mathcal{A}}$, $R_n(\alpha) = E\left[L_n(\alpha)\right]$ and $nL_n(\alpha) = \|\boldsymbol{\mu}_n - P_n(\alpha)\mathbf{y}_n\|^2$.

The result that is obtained can be stated as follows.

Theorem 2 *Consider the setup of Theorem 1 and consider a class of models not containing the true model for which (26) holds. Then under the conditions of Theorem 1, (25) holds.*

The proof of Theorem 2 is given in the supplementary file.

2.2 Loss efficiency

In this section, we prove "loss efficiency" of the model selection procedure using g-prior currently under study. The concept of "loss efficiency" [(Li 1987) and (Shao 1997)] has been briefly described in the Sect. 1. As in Li (1987) and Shao (1997), we consider as our evaluation criterion the average squared error deviation between the true and estimated regression function, given by

$$L_n(\alpha) = \frac{1}{n}\left\|\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n(\alpha)\right\|^2, \tag{27}$$

where $\hat{\boldsymbol{\mu}}_n(\alpha)$ is the chosen estimator of the true regression function $\boldsymbol{\mu}_n$ when model M_α is selected. We consider two estimators—the Bayes estimator for the g-prior and the least squares estimator which is also the Bayes estimator for the standard non-informative priors (considered, for example, in Chakrabarti and Samanta (2008)). If $\boldsymbol{\mu}_n$ were known, one could find the model α_n^L which minimizes $L_n(\alpha)$ for each $\mathbf{y}_n$. This model will be called the Oracle model since it is based on the unknown truth $\boldsymbol{\mu}_n$ and it cannot be achieved in practice. We show that for both the above choices of $\hat{\boldsymbol{\mu}}_n(\alpha)$,

$$\frac{L_n(\hat{\alpha})}{\min_{\alpha\in\boldsymbol{\mathcal{A}}} L_n(\alpha)} \xrightarrow{p} 1 \qquad \text{as } n \rightarrow \infty \tag{28}$$

if $M_{\hat{\alpha}}$ is the model chosen by the method based on g-prior. Thus the g-prior method is shown to perform equivalently to an Oracle asymptotically. We first consider in Theorem 3, the case when $\boldsymbol{\mu}_n$ is estimated by the Bayes estimator under the selected model $M_{\hat{\alpha}}$.

Theorem 3 *Consider model selection rule under study as in Theorem 1 and let $M_{\hat{\alpha}}$ be the model selected by this rule. Let $g = g_n = kn^r$ for $r > 0$ and $k > 0$ and $\hat{\boldsymbol{\mu}}_n(\alpha)$ be the Bayes estimator of $\boldsymbol{\mu}_n$ under model M_α. Then we have the following:*

- **(a)** *For p fixed, under the same assumptions as in Theorem 1* **(a)***, $\hat{\alpha}$ satisfies (28).*
- **(b)** *Suppose now $p = p_n$ grows with n. Then under the same assumptions as in Theorem 1* **(b)***, $\hat{\alpha}$ satisfies (28) provided $r > s$.*

The proof of Theorem 3 is given in the Appendix.

Remark 2 When one uses $\hat{\boldsymbol{\mu}}_n(\alpha)$ as the least squares estimator of $\boldsymbol{\mu}_n$ under model α, the proof of Theorem 3 can be suitably adapted to accommodate this case and (28) can be shown to hold under this situation as well. A proof is given in the Appendix after the proof of Theorem 3.

2.2.1 The case when the true model is not normal

As mentioned in Sect. 2.1.1, we do not try to study this problem in its full generality for this case in the present paper. We give a proof of loss efficiency under the setup of Li (1987) and Shao (1997) and under their assumption stated in (26). For simplicity in presentation, we consider only the case where $L_n(\alpha)$ is as defined in (27) with $\hat{\boldsymbol{\mu}}_n(\alpha) = P_n(\alpha)\mathbf{y}_n$, the least squares estimate of $\boldsymbol{\mu}_n$.

Our result can be stated as follows.

Theorem 4 *Consider the setup of Theorem 3 with $\hat{\boldsymbol{\mu}}_n(\alpha) = P_n(\alpha)\mathbf{y}_n$ and consider a class $\boldsymbol{\mathcal{A}}$ of models not containing the true model for which (26) holds. Then under the conditions of Theorem 3, (28) holds.*

The proof of Theorem 4 is given in the supplementary file.

3 The "Model True" case

In this section, we assume that the true model is in the model space and prove that the model selection procedure based on g-prior is consistent in appropriate sense. We also prove "loss efficiency" as described in the Introduction. We assume that under each model M_α, $\boldsymbol{\beta}_\alpha$ is a $p_n(\alpha)$-dimensional vector of *non-zero* regression coefficients. This ensures that there is exactly one true model in the model space.

3.1 Model selection consistency

Let M_{α_c}, $\alpha_c \in \boldsymbol{\mathcal{A}}$ be the true model. The posterior probability of M_{α_c}, given by (3) can be expressed as

$$p(M_{\alpha_c}|\mathbf{y}_n) = \left(1 + \sum_{\alpha \in \boldsymbol{\mathcal{A}}, \alpha \neq \alpha_c} \frac{p(M_\alpha)}{p(M_{\alpha_c})} \times \frac{m_\alpha(\mathbf{y}_n)}{m_{\alpha_c}(\mathbf{y}_n)}\right)^{-1}. \tag{29}$$

We will show, under suitable conditions, that for the priors (5) and (6),

$$p(M_{\alpha_c}|\mathbf{y}_n) \xrightarrow{p} 1 \tag{30}$$

under the model M_{α_c}. This is known as model selection consistency which also implies that the true model M_{α_c} is selected with probability tending to *one*. We divide the model space into two parts,

$$\boldsymbol{\mathcal{A}}_1 = \{\alpha \in \boldsymbol{\mathcal{A}} : M_\alpha \supset M_{\alpha_c}, \alpha \neq \alpha_c\}, \text{ and } \boldsymbol{\mathcal{A}}_2 = \{\alpha \in \boldsymbol{\mathcal{A}} : \alpha \notin \boldsymbol{\mathcal{A}}_1, \alpha \neq \alpha_c\}.$$

We make the following assumptions, which are analogous versions of assumptions (A2) and (A2)* of Sect. 2 in the "model true" case, replacing $\boldsymbol{\mathcal{A}}$ by $\boldsymbol{\mathcal{A}}_2$.

For the case when p is fixed, we assume

(B.2) $\underline{\lim}_{n\to\infty} \min_{\alpha\in\boldsymbol{\mathcal{A}}_2} \boldsymbol{\mu}_n'(I_n - P_n(\alpha))\boldsymbol{\mu}_n/n > \Delta$ for some constant $\Delta > 0$.

For the case when $p = p_n$ grows with n, we replace assumption (B.2) by

(B.2)* $\underline{\lim}_{n\to\infty} n^s \min_{\alpha\in\boldsymbol{\mathcal{A}}_2} \boldsymbol{\mu}_n'(I_n - P_n(\alpha))\boldsymbol{\mu}_n/n > \delta$ for some constants $\delta > 0$ and $0 \le s < 1$.

We will prove the following result.

Theorem 5 *Consider the priors* (5) *and* (6) *and let* $g = g_n = kn^r$ *for* $r > 0,\ k > 0$. *Then we have the following results.*

(a) *If p is fixed, then under assumptions* (A.1) *and* (B.2), (30) *holds under the true model* M_{α_c}.

(b) *Suppose that $p = p_n$ grows with n. Assume that $p_n = O(n^b)$ for* $0 < b < 1$, $r > 4b$, (A.1) *holds,* (B.2)* *holds with* $s < (1-b)/2$, *and*

$$\max_{\alpha,\alpha'\in\boldsymbol{\mathcal{A}}} p(M_\alpha)/p(M_{\alpha'}) \le k_0 n^{b_0} \quad \textit{for some } k_0 > 0 \textit{ and } 0 < b_0 < (r-4b)/2. \tag{31}$$

Then (30) *holds under the true model* M_{α_c}.

The proof of Theorem 5 is given in the Appendix.

Remark 3 It is interesting to note that a "very small" choice of $g = g_n$ may actually lead to inconsistency in the "model true" case. Consider for example, the situation when the true model is the null model, M_N, under which $\boldsymbol{\mu}_n = \mathbf{1}_n\beta_0$ and all the candidate models are given equal probability. It has been shown in the supplementary file that $\sum_{\alpha\in\boldsymbol{\mathcal{A}}_1} m_\alpha(\mathbf{y}_n)/m_{\alpha_c}(\mathbf{y}_n)$ cannot converge to zero in probability if one chooses $g_n = kn^r$ with $r \le 2b$ when $p_n = n^b$, $0 < b < 1$. This implies, vide (29), that for model selection consistency in the "model true" case it is necessary to choose $g_n = kn^r$ with $r > 2b$, $k > 0$ (if $p_n = n^b$, $0 < b < 1$).

Remark 4 This is worth noting that "too big" a $g = g_n$ may also lead to inconsistency in the "model true" case. Consider in this case, the situation where the true model is the full model, M_F. It has been proved in the supplementary file that if $p_n = n^b$, $0 < b < 1$ then by choosing $g_n > D^{n/p_n}$, for some appropriately selected D (> 1), one can conclude that $p(M_{\alpha_c}|\mathbf{y}_n)$ does not converge to 1 in probability.

3.2 Loss efficiency

We now prove "loss efficiency" of the model selection procedure based on g-prior in the "model true" case. We first consider the case when $\boldsymbol{\mu}_n$ is estimated by the Bayes estimator under the selected model. Indeed, we prove a result stronger than loss efficiency as stated below.

Theorem 6 *Consider the model selection rule under study as in Theorem 1 and let $M_{\hat{\alpha}}$ be the model selected by this rule. Let $\hat{\boldsymbol{\mu}}_n(\alpha)$ be the Bayes estimator of $\boldsymbol{\mu}_n$ for the g-prior under model M_α. Then under the conditions of Theorem 5, we have*

$$L_n(\hat{\alpha}) = \min_{\alpha \in \boldsymbol{\mathcal{A}}} L_n(\alpha)$$

with probability tending to one.

Proof As shown in the Appendix [see (36)], for all $\alpha \in \boldsymbol{\mathcal{A}}$,

$$\begin{aligned} nL_n(\alpha) = C_n + a_n(2-a_n)\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\boldsymbol{\mu}_n + a_n^2\mathbf{e}_n' P_n(\alpha)\mathbf{e}_n \\ -2a_n(1-a_n)\boldsymbol{\mu}_n' P_n(\alpha)\mathbf{e}_n, \end{aligned} \tag{32}$$

where

$$C_n = (1-a_n)^2\boldsymbol{\mu}_n'\boldsymbol{\mu}_n + (1-a_n^2)n\overline{y}^2 - 2(1-a_n)\overline{y}\sum_{i=1}^{n}\mu_i + 2a_n(1-a_n)\overline{y}\mathbf{1}_n'\mathbf{y}_n.$$

Let M_{α_c}, $\alpha_c \in \boldsymbol{\mathcal{A}}$, be the true model and $\boldsymbol{\mathcal{A}}_1$ and $\boldsymbol{\mathcal{A}}_2$ be the subspaces of the model space $\boldsymbol{\mathcal{A}}$, as defined at the beginning of Sect. 3.1. We shall prove that with probability tending to *one*,

$$\min_{\alpha \in \boldsymbol{\mathcal{A}}_i} L_n(\alpha) > L_n(\alpha_c) \qquad \text{for all } i = 1, 2. \tag{33}$$

As $P_n(\alpha)\boldsymbol{\mu}_n = \boldsymbol{\mu}_n$ for all $\alpha \in \boldsymbol{\mathcal{A}}_1 \cup \{\alpha_c\}$, from (32), we have for all $\alpha \in \boldsymbol{\mathcal{A}}_1 \cup \{\alpha_c\}$,

$$nL_n(\alpha) = C_n + a_n^2\mathbf{e}_n' P_n(\alpha)\mathbf{e}_n - 2a_n(1-a_n)\boldsymbol{\mu}_n'\mathbf{e}_n. \tag{34}$$

Therefore, for all $\alpha \in \boldsymbol{\mathcal{A}}_1$,

$$nL_n(\alpha) - nL_n(\alpha_c) = a_n^2\mathbf{e}_n'(P_n(\alpha) - P_n(\alpha_c))\mathbf{e}_n$$

which implies

$$\min_{\alpha \in \boldsymbol{\mathcal{A}}_1} nL_n(\alpha) - nL_n(\alpha_c) = a_n^2 \min_{\alpha \in \boldsymbol{\mathcal{A}}_1} \mathbf{e}_n'(P_n(\alpha) - P_n(\alpha_c))\mathbf{e}_n. \tag{35}$$

Since for all $\alpha \in \boldsymbol{\mathcal{A}}_1$, $(P_n(\alpha) - P_n(\alpha_c))$ is also a projection matrix, $\mathbf{e}_n'\,(P_n(\alpha) - P_n(\alpha_c))\,\mathbf{e}_n/\sigma^2$ follows a χ^2 distribution and therefore, by (35), the result (33) holds for $i = 1$ with probability *one*.

We now prove (33) for $i = 2$. From (32) and (34) we have for $\alpha \in \boldsymbol{\mathcal{A}}_2$,

$$\begin{aligned} nL_n(\alpha) - nL_n(\alpha_c) = a_n(2-a_n)\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\boldsymbol{\mu}_n + a_n^2\mathbf{e}_n' P_n(\alpha)\mathbf{e}_n \\ -a_n^2\mathbf{e}_n' P_n(\alpha_c)\mathbf{e}_n + 2a_n(1-a_n)\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\mathbf{e}_n \end{aligned}$$

and therefore,

$$\begin{aligned}&\min_{\alpha\in\boldsymbol{\mathcal{A}}_2} L_n(\alpha)-L_n(\alpha_c)\\&\quad\geq a_n(2-a_n)\min_{\alpha\in\boldsymbol{\mathcal{A}}_2}\frac{1}{n}\boldsymbol{\mu}_n'(I_n-P_n(\alpha))\boldsymbol{\mu}_n-a_n^2\frac{1}{n}\mathbf{e}_n'P_n(\alpha_c)\mathbf{e}_n\\&\qquad-2a_n(1-a_n)\max_{\alpha\in\boldsymbol{\mathcal{A}}_2}\frac{1}{n}\left|\boldsymbol{\mu}_n'(I_n-P_n(\alpha))\mathbf{e}_n\right|.\end{aligned}$$

The result (33) with $i=2$ now follows from Lemma 1 by assumption (B.2) (for the case with fixed p) or (B.2)* (for the case when $p=p_n$ grows with n).

As $\hat{\alpha}=\alpha_c$ with probability tending to *one* from Theorem 5, the result now follows from (33). □

When one uses $\hat{\boldsymbol{\mu}}_n(\alpha)$ as the least squares estimator of $\boldsymbol{\mu}_n$ under the model α, Theorem 3.2 can be proved using similar arguments. In this case, $\hat{\boldsymbol{\mu}}_n(\alpha)=P_n(\alpha)\mathbf{y}_n$ and

$$nL_n(\alpha)=\boldsymbol{\mu}_n'\left(I_n-P_n(\alpha)\right)\boldsymbol{\mu}_n+\mathbf{e}_n'P_n(\alpha)\mathbf{e}_n.$$

Therefore for all $\alpha\in\boldsymbol{\mathcal{A}}_1$,

$$\begin{aligned}&\min_{\alpha\in\boldsymbol{\mathcal{A}}_1} nL_n(\alpha)-nL_n(\alpha_c)=\min_{\alpha\in\boldsymbol{\mathcal{A}}_1}\mathbf{e}_n'\left(P_n(\alpha)-P_n(\alpha_c)\right)\mathbf{e}_n\\&\qquad\text{and for all }\alpha\in\boldsymbol{\mathcal{A}}_2,\\&\min_{\alpha\in\boldsymbol{\mathcal{A}}_2} L_n(\alpha)-L_n(\alpha_c)\geq\min_{\alpha\in\boldsymbol{\mathcal{A}}_2}\frac{1}{n}\boldsymbol{\mu}_n'\left(I_n-P_n(\alpha)\right)\boldsymbol{\mu}_n-\frac{1}{n}\mathbf{e}_n'P_n(\alpha_c)\mathbf{e}_n.\end{aligned}$$

The result now follows by arguments similar to those used in the proof of Theorem 3.2.

4 Simulation results

In this section, we present some simulation results to demonstrate the performance of the model selection procedure under study. We perform the simulation under both the scenarios when the true model is in the model space and when it is not. We consider sample sizes (n) varying from moderate to large, compared to the model dimension p and also allow p to grow from small to large. In each of these cases, we consider different choices of g to understand its role in the performance of the method. In the simulation results presented, the properties of loss efficiency and consistency are demonstrated satisfactorily most of the time even for moderate sample sizes. Below we describe our scheme of simulation and discuss the results obtained.

For both the "model false" and "model true" cases, we consider $n=50$, 100 and 150. We denote by p the total number of available regressors from which the selection is made so that the full design matrix $[\mathbf{1}_n,\mathbf{X}_n]$, with the column of 1's for the intercept, is of dimension $n\times(p+1)$. For the "model true" case, we take different choices of p

such that $p + 1 \leq n$. For "model false" case, we take different choices of p such that $p + 1$ is *strictly less than* n. For each combination of n and p, four different choices of g are considered, viz., $g = \sqrt{n}, n, p^2$, and n^2. Note that the choice of $g = n$ was recommended by Kass and Wasserman (1995) and the choice of $g = p^2$ was recommended by Foster and George (1994). Fernández et al. (2001) recommended use of $g = \max(n, p^2)$. We take $g = \sqrt{n}$ to see how the method performs for a relatively small g. The arguments for considering $g = n^2$ are given in Sect. 5.

We first describe the simulation scheme for the "model true" case. For each combination of $(n,\ \ p)$, we generate n values of each of the p regressor variables $x_1, x_2, \ldots, x_p$ and this gives the design matrix $\mathbf{X}_n$. We choose p numbers $\nu_i, i = 1, \ldots, p$ and generate the n values of the i^{th} regressor x_i from an $N(\nu_i, 1)$ distribution, $i = 1, \ldots, p$. We assume that the n values of the i^{th} regressor are coming from a homogeneous population. To fix a "true" model, we choose its dimension $p(\alpha_c)$ and then choose the $p(\alpha_c)$ non-zero regression coefficients β_j's, the intercept β_0 in the true model and also a value for the error variance σ^2. The $p(\alpha_c)$ columns of the design matrix $\mathbf{X}_{n\alpha_c}$ for the true model are chosen at random from the p columns of $\mathbf{X}_n$. We use two schemes to select the values of ν_i's, $p(\alpha_c)$, β_j's and σ. In Scheme 1, we select ν_i's at random from a normal distribution, $\nu_i \sim N(10, 10)$, and thus making the possible range of the x values very wide. In Scheme 2, $(\nu_1, \ldots, \nu_p)$ 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 $[2p/3]$ in Scheme 1 and $[p/2]$ in Scheme 2. The $p(\alpha_c)$ non-zero regression coefficients β_j's and the intercept β_0 in the true model are randomly chosen from a uniform distribution over $(-10, 10)$ in Scheme 1, and from the set $\{-0.2, 0.4, \ldots, (-1)^p 0.2p\}$ in Scheme 2. In Scheme 1, we take $\sigma = 3$ and in Scheme 2, we take $\sigma = 1$.

After choosing the dimension $p(\alpha_c)$, the coefficients $\left(\beta_0, \boldsymbol{\beta}_{\alpha_c}\right)$, the error variance σ^2 of the true model and the design matrix $\mathbf{X}_n$, we generate data $\mathbf{y}_n$ following $N_n\left(\mathbf{1}_n\beta_0 + \mathbf{X}_{n\alpha_c}\boldsymbol{\beta}_{\alpha_c}, \sigma^2 I_n\right)$. Having obtained the data, we compute the posterior probability of the true model using the g-prior for several choices of g as indicated in the second paragraph of this section. We also find the model $M_{\hat{\alpha}}$ selected by the method based on g-prior (by finding the model with highest posterior probability) and calculate the *loss ratio*, $L_n(\hat{\alpha})/\min_{\alpha \in \boldsymbol{\mathcal{A}}} L_n(\alpha)$, for each choice of g. It needs to be mentioned that for exact calculation of posterior probability of any model, the marginal densities of the data $(m_\alpha(\mathbf{y}_n))$ for all the candidate models $\alpha \in \boldsymbol{\mathcal{A}}$ are needed. It is possible to do this for small p $(p + 1 = 10)$ but for large p $(p + 1 = 30, 50)$, this becomes quite infeasible. Therefore, for such cases we take resort to Markov Chain Monte Carlo simulation techniques to approximate the posterior probabilities, whereby computation of marginal densities can be restricted only to the models visited by the chain. To simulate from the relevant Markov chain, we have used the Gibbs sampling algorithm. The sampling scheme and the method of computation of posterior probabilities are completely described in Chipman et al. (2001, Section 3.5). For the simulation, we have generated a Markov chain of length 20000 of which the first 10000 have been used as burn-in. For the cases $p + 1 = 30$ and 50, the quantity $\min_{\alpha \in \boldsymbol{\mathcal{A}}} L_n(\alpha)$ is approximated by taking the minimum of $L_n(\alpha)$ over the models visited by the chain.

For each combination of $(n,\ p)$, each of the choices of g and each of the two schemes, we repeat the above for 100 times fixing the chosen values of ν_i's, $p(\alpha_c)$, β_j's and σ. The mean and standard deviation of the posterior probabilities of the true model and those of the *loss ratios* are presented in the upper halves of Table 1 (for Scheme 1) and Table 2 (for Scheme 2). In the tables these measures are denoted by "Post. prob. mean", "Post. prob. s.d.", "Loss ratio mean" and "Loss ratio s.d.", respectively.

For "model false" case, the $n \times p$ design matrix $\mathbf{X}_n$ is generated as described above but the true regression $\boldsymbol{\mu}_n$ is generated in a different manner.We consider a basis of the orthogonal complement of the column space of the full design matrix $[\mathbf{1}_n, \mathbf{X}_n]$. Then, we choose k_1 columns of the design matrix and k_2 basis vectors of the orthogonal complement to generate the true regression $\boldsymbol{\mu}_n$ by taking linear combination of them with randomly chosen *non-zero* coefficients. The coefficients are selected in the same way as the regression coefficients and intercept are selected above in the "model true" case using two schemes. We choose k_1 and k_2 as $k_1 = [2p/3]$ and $k_2 = \min((n-p), [p/3])$ for Scheme 1 and as $k_1 = [p/2]$ and $k_2 = \min((n-p), [p/2])$ for Scheme 2. Once the truth $\boldsymbol{\mu}_n$ is fixed, we generate the data $\mathbf{y}_n$ following $N_n\left(\boldsymbol{\mu}_n, \sigma^2 I_n\right)$. We now select the model $M_{\hat{\alpha}}$ using the method based on g-prior and calculate the *distance ratio*, $D_n\left(\hat{\alpha}\right) / \min_{\alpha \in \boldsymbol{\mathcal{A}}} D_n(\alpha)$, and *loss ratio* $L_n(\hat{\alpha}) / \min_{\alpha \in \boldsymbol{\mathcal{A}}} L_n(\alpha)$, where $D_n(\alpha)$ and $L_n(\alpha)$ are as in (9) and (27), respectively. Again, as in the "model true" case, for each combination of $(n,\ p)$, each of the choices of g and each of the two schemes, we repeat the above for 100 times and obtain 100 *loss ratios* and *distance ratios*. The mean and standard deviation of these 100 values are presented in the lower halves of Table 1 (for Scheme 1) and Table 2 (for Scheme 2). In the tables these measures are denoted by "Dist. ratio mean", "Dist. ratio s.d.", "Loss ratio mean" and "Loss ratio s.d.", respectively. Note that the case $n = p+1$ does not fall in the "model false" case as any $\boldsymbol{\mu}_n$ can be generated by the $p+1 = n$ columns of $[\mathbf{1}_n, \mathbf{X}_n]$, and therefore, the corresponding blocks in the tables are left blank.

Figures in the upper halves of Tables 1 and 2 show how the posterior probabilities of true models increase and how the *loss ratios* (described above) decrease to one as the sample size n increases in the "model true" case. We observe that, in "model true" case, the choice of $g = n^2$ gives substantially better results than the other choices of g. The posterior probabilities for this choice of g are significantly larger than those for the other choices of g in all the cases, except for the case $p+1 = 30, n = 100$ in Scheme 2. However, for larger n ($n = 150$) with $p+1 = 30$ in Scheme 2, $g = n^2$ performs better than the other choices. Figures in the lower halves of the tables show how the *distance ratios* and the *loss ratios* (as described above) decrease to one as the sample size n increases in the "model false" case. For the "model false" case, the choice of $g = n^2$ seems to perform the best with little exception for Scheme 1. For Scheme 2, some of the smaller choices of g perform slightly better. However, as n increases, the differences in their performance become small.

5 On the choice of g

A crucial ingredient of the model selection mechanism studied in this paper is the choice of hyper-parameter g in the prior for the regression parameter vector $\boldsymbol{\beta}_\alpha$ given

Table 1 Simulation results for Scheme 1

$p+1$	Measures	$n=50$				$n=100$				$n=150$			
		$g=\sqrt{n}$	$g=n$	$g=p^2$	$g=n^2$	$g=\sqrt{n}$	$g=n$	$g=p^2$	$g=n^2$	$g=\sqrt{n}$	$g=n$	$g=p^2$	$g=n^2$
Model True Case													
10	Post. prob. mean	0.1902	0.5646	0.6069	0.8466	0.3981	0.5887	0.5887	0.8981	0.4361	0.6648	0.6399	0.9556
	Post. prob. s.d.	0.0616	0.1075	0.1354	0.1561	0.0477	0.1394	0.1394	0.1280	0.0423	0.1372	0.1047	0.0494
	Loss ratio mean	1.0979	1.0825	1.0993	1.0735	1.0052	1.1178	1.1178	1.0099	1.0024	1.0634	1.0499	1.0000
	Loss ratio s.d.	0.2360	0.4078	0.3580	0.3810	0.0367	0.3999	0.3999	0.0990	0.0241	0.3606	0.4485	0.0000
30	Post. prob. mean	0.0181	0.1149	0.3050	0.3349	0.0460	0.2287	0.4622	0.7214	0.0518	0.2640	0.4553	0.8081
	Post. prob. s.d.	0.0073	0.0370	0.1923	0.2217	0.0046	0.0502	0.1357	0.1534	0.0042	0.0777	0.1583	0.1258
	Loss ratio mean	2.0868	1.4175	1.1890	1.7878	1.2011	1.0088	1.0311	1.0240	1.0166	1.0352	1.0507	1.0046
	Loss ratio s.d.	0.7983	0.6683	0.3044	0.3097	0.2792	0.0881	0.1239	0.1145	0.0830	0.1322	0.1403	0.0457
50	Post . prob. mean	0.0001	0.0021	0.0020	0.0179	0.0097	0.0518	0.2721	0.4650	0.0151	0.0771	0.3600	0.6531
	Post. prob. s.d.	0.0001	0.0020	0.0054	0.0574	0.0025	0.0170	0.1750	0.2161	0.0025	0.0254	0.1571	0.2137
	Loss ratio mean	3.8688	3.2942	1.1768	1.1908	2.2354	1.1744	1.0867	1.0557	1.1547	1.0093	1.0352	1.0387
	Loss ratio s.d.	1.1500	1.6079	0.1146	0.1610	0.6079	0.3064	0.1564	0.1230	0.2675	0.0489	0.0922	0.1088
Model False Case													
10	Dist. ratio mean	1.0014	1.0166	1.0448	1.0000	1.0039	1.0726	1.0726	1.0000	1.0014	1.0636	1.0167	1.0000
	Dist. ratio s.d.	0.0140	0.0765	0.1405	0.0000	0.0276	0.2570	0.2570	0.0000	0.0140	0.2351	0.0827	0.0000
	Loss ratio mean	1.0036	1.0176	1.0463	1.0000	1.0079	1.0744	1.0744	1.0000	1.0032	1.0636	1.0173	1.0000
	Loss ratio s.d.	0.0366	0.0806	0.1444	0.0000	0.0564	0.2634	0.2634	0.0000	0.0318	0.2362	0.0853	0.0000
30	Dist. ratio mean	1.8176	1.2972	1.1644	1.5958	1.1578	1.0056	1.0083	1.0000	1.0002	1.0041	1.0044	1.0000
	Dist. ratio s.d.	0.4697	0.3532	1.1668	1.0733	0.1917	0.0326	0.0368	0.0000	0.0020	0.0300	0.0264	0.0000
	Loss ratio mean	2.5895	1.3010	1.1640	1.5958	1.3413	1.0058	1.0083	1.0000	1.0004	1.0041	1.0044	1.0000
	Loss ratio s.d.	1.0166	0.3551	1.1633	1.0732	0.4096	0.0342	0.0369	0.0000	0.0044	0.0295	0.0265	0.0000

Table 1 continued

$p+1$	Measures	$n=50$				$n=100$				$n=150$			
		$g=\sqrt{n}$	$g=n$	$g=p^2$	$g=n^2$	$g=\sqrt{n}$	$g=n$	$g=p^2$	$g=n^2$	$g=\sqrt{n}$	$g=n$	$g=p^2$	$g=n^2$
50	Dist. ratio mean					1.7617	1.0268	1.0242	1.0069	1.2132	1.0095	1.0000	1.0031
	Dist. ratio s.d.					0.3914	0.0559	0.0829	0.0508	0.2118	0.0256	0.0000	0.0223
	Loss ratio mean					2.3480	1.0274	1.0242	1.0069	1.4363	1.0097	1.0000	1.0031
	Loss ratio s.d.					0.6862	0.0565	0.0829	0.0508	0.4398	0.0263	0.0000	0.0223

Table 2 Simulation results for Scheme 2

$p+1$	Measures	$n=50$				$n=100$				$n=150$			
		$g=\sqrt{n}$	$g=n$	$g=p^2$	$g=n^2$	$g=\sqrt{n}$	$g=n$	$g=p^2$	$g=n^2$	$g=\sqrt{n}$	$g=n$	$g=p^2$	$g=n^2$
Model True Case													
10	Post. prob. mean	0.0400	0.0864	0.1183	0.3817	0.1536	0.2070	0.2070	0.8847	0.1221	0.4921	0.2804	0.8244
	Post. prob. s.d.	0.0283	0.0882	0.1061	0.2772	0.0590	0.1548	0.1548	0.0828	0.0628	0.1350	0.1504	0.2100
	Loss ratio mean	1.8058	2.4050	1.8145	2.3168	1.3145	1.8261	1.8261	1.0000	1.2504	1.1081	1.5997	1.3089
	Loss ratio s.d.	1.0629	2.6497	3.0123	2.7942	0.6327	1.4873	1.4873	0.0000	0.3828	0.4283	0.9239	1.1370
30	Post. prob. mean	0.0000	0.0003	0.0352	0.1474	0.0001	0.0200	0.3682	0.1400	0.0010	0.1404	0.1254	0.7325
	Post. prob. s.d.	0.0000	0.0002	0.0584	0.1632	0.0000	0.0142	0.1213	0.1947	0.0003	0.0497	0.1223	0.1597
	Loss ratio mean	1.6061	2.2431	1.1851	1.4828	1.2257	1.2022	1.0808	1.2016	1.0618	1.1009	1.3126	1.0459
	Loss ratio s.d.	0.4527	1.0202	0.3423	0.5106	0.1122	0.1467	0.2923	0.1612	0.1189	0.3423	0.3633	0.2214
50	Post. prob. mean	0.0000	0.0000	0.0417	0.0417	0.0000	0.0000	0.0398	0.2917	0.0000	0.0152	0.3231	0.5151
	Post. prob. s.d.	0.0000	0.0000	0.0696	0.0696	0.0000	0.0000	0.0698	0.1994	0.0000	0.0052	0.0906	0.2025
	Loss ratio mean	2.6489	4.6062	1.1911	1.1911	1.6119	2.5703	1.1553	1.1646	1.3730	1.3684	1.0260	1.1000
	Loss ratio s.d.	1.0091	2.2677	0.1871	0.1871	0.3265	1.1229	0.1905	0.2581	0.2038	0.5959	0.1260	0.2339
Model False Case													
10	Dist. ratio mean	1.1620	1.2393	1.2400	1.4864	1.1439	1.1640	1.1640	1.3527	1.1302	1.1617	1.1596	1.284
	Dist. ratio s.d.	0.1734	0.2470	0.2398	0.5191	0.1788	0.1970	0.1970	0.3447	0.2002	0.2029	0.1947	0.2974
	Loss ratio mean	1.1978	1.2425	1.2412	1.4864	1.1682	1.1652	1.1652	1.3527	1.1508	1.1622	1.1607	1.2840
	Loss ratio s.d.	0.2152	0.2527	0.2425	0.5191	0.2118	0.1985	0.1985	0.3447	0.2308	0.2025	0.1962	0.2974
30	Dist. ratio mean	1.1776	1.2417	1.6638	2.1367	1.0616	1.0633	1.0956	1.1546	1.0391	1.0279	1.0438	1.0814
	Dist. ratio s.d.	0.1074	0.1528	0.5222	0.7420	0.0512	0.0431	0.0544	0.0939	0.0387	0.0260	0.0312	0.0529
	Loss ratio mean	1.2021	1.2430	1.6638	2.1367	1.0696	1.0634	1.0956	1.1546	1.0446	1.0279	1.0438	1.0814
	Loss ratio s.d.	0.1251	0.1541	0.5223	0.7420	0.0583	0.0432	0.0544	0.0939	0.0440	0.0261	0.0312	0.0529

Table 2 continued

$p+1$	Measures	$n = 50$				$n = 100$				$n = 150$			
		$g = \sqrt{n}$	$g = n$	$g = p^2$	$g = n^2$	$g = \sqrt{n}$	$g = n$	$g = p^2$	$g = n^2$	$g = \sqrt{n}$	$g = n$	$g = p^2$	$g = n^2$
50	Dist. ratio mean					1.1356	1.1319	1.2416	1.3148	1.0778	1.0580	1.0950	1.1260
	Dist. ratio s.d.					0.0943	0.0771	0.1332	0.1392	0.0474	0.0453	0.0394	0.0523
	Loss ratio mean					1.1526	1.1321	1.2416	1.3148	1.0879	1.0581	1.0950	1.1260
	Loss ratio s.d.					0.1076	0.0773	0.1332	0.1392	0.0538	0.0454	0.0394	0.0522

in (6). The results we obtain through theoretical investigations and simulations help us understand which choices of g lead to good performance of the model selection procedure. The theoretical investigations are done in the asymptotic framework since it makes many things more transparent. We treat both the cases when the number of potential regressors (p) remains fixed and when it grows with sample size n such that $p = O(n^b)$ as $n \to \infty$ where $0 < b < 1$.

We first discuss the case when p is fixed. There have been several recommendations in the literature on the choice of g in this case. A thorough study is made by Fernández et al. (2001) in the "model true" case and they recommend use of $g = \max(n, p^2)$. The resulting prior is called "benchmark prior" and this recommendation is well accepted in the literature. Therefore, our main focus in this section lies in the choice of g when p grows with n. Before we discuss this in detail, we would like to mention that we have also studied the "model false" case in the fixed-p scenario. It may be observed (vide Theorems 1 and 3) that in this case the choice of g for ensuring good performance can be very flexible. In particular, the choice of $g = \max(n, p^2)$, as in Fernández et al. (2001), ensures both consistency and loss efficiency. Therefore, this choice is suitable for both the "model false" and "model true" cases when p is fixed.

We now discuss the case when p grows with n. We first summarize our findings from our theoretical investigations. It is very pleasing to note (vide Theorem 1) that this method is consistent for the "model false" case for any choice of g such that $g = g_n = kn^r$ for some $r \geq 0$ and $k > 0$. This also includes the situations when g does not vary with n. It follows from the proof of Theorem 3 that for loss efficiency in "model false" case, we need to make g grow to infinity with n. However, from the statement of the theorem it also follows that it is sufficient to choose $g = g_n = kn^r$, $r \geq 1/2$, $k > 0$ to achieve loss efficiency. This comes from the observation that we need $r > s$ and $s < (1 - b)/2$ (where $0 < b < 1$) for Theorem 3 to hold.

For the "model true" case, we show that choosing $g = g_n = kn^r$ with $k > 0$ and $r > 4b$ is sufficient to ensure both consistency and loss efficiency (vide Theorems 5 and 6). However, as observed in Remark 3, it is also necessary that we must have $r > 2b$ for consistency in the "model true" case as $r \leq 2b$ leads to inconsistency.

It is clear from the discussion above that it is advisable to choose a g which is not "too small". In particular, a choice of $g = n^r$ with some $r > \max(4b, 1/2)$ ensures good performance on the whole. This takes care of both the "model false" and "model true" cases and follows from the above discussions of these two cases. A natural question now is how large a g should be chosen. It may be noted that making g large arbitrarily makes the priors on $\boldsymbol{\beta}_\alpha$'s arbitrarily vague in the sense that most compact subsets of the parameter space gets nearly zero probability. Such priors are not usually recommended in the model selection literature. See in this context Bayarri et al. (2012). It may be recalled from Remark 4, that a choice of g_n of the form $g_n = D^{n/p_n}$ for some appropriately chosen constant $D > 1$, leads to inconsistency in the "model true" case. This supports the common wisdom of not having "too large" a g.

We now make some final remarks on the choice of g based on the above theoretical inputs and also the simulation results reported in Sect. 4. In practice, we have a data at hand with some fixed n and p, and we may apply our theoretical results (on choice of g) for the scenario "p grows with n" if the n and p for our given data are both large enough. If indeed this is the case, we will typically also have p at least as large as $n^{1/2-\delta}$

for some small positive δ and then p can be thought of as $O(n^b)$ for any $b \geq 1/2 - \delta$ for some small δ. For example, if n is 100 and p is between 30 and 50, it does not seem right to take $p = O(n^b)$ with $b = 1/2 - \Delta$ with Δ not small. Since our theoretical results show that by taking $g = n^r$ with $r > 4b$ one can achieve consistency, in view of the above observation it appears that $g = n^2$ might be a reasonable choice. In principle, one could try larger choices of g. But one would not want to use larger g if $g = n^2$ itself gives good enough results since too large a g is not necessarily a good idea as discussed above. We have simulation results taking $g = \sqrt{n}$, n and n^2. For the "model true" case the choice of $g = n^2$ gives clearly the best results. For the "model false" case, the performance with $g = n^2$ is very satisfactory for large n (compared to p) although it may not always be the best performing choice of g. Even when $g = n^2$ does not give the best results, the difference in its performance compared to the best one is quite negligible. Combining all these facts we feel comfortable in using $g = n^2$ as the preferred choice of g when p and n are both large.

6 Concluding remarks

In this paper, we have studied theoretical properties of the method of model selection based on g-prior when the sample size grows. The g-prior has been one of the most popular priors in use for the normal linear regression model and the motivation of the paper is to study under what conditions this popular method gives desirable results. We have studied the properties in the asymptotic framework since it makes many things more transparent. We have first shown that in a situation where the true model is not one of the candidate models, this model selection procedure selects a model that is in a sense closest to the true model. Also, the ratio of the loss incurred in estimating the unknown regression function under the selected model and the loss of an Oracle tends to one. These results have been proved under appropriate conditions on the rate of growth of g as n grows and for both the cases when the number p of potential predictors remains fixed with n and when $p = O(n^b)$ for some $0 < b < 1$. We do not know of any work related to g priors that considers the "model false" case and we have shown efficacy of the method based on g-prior with respect to two natural criteria. We think these two criteria, namely, the ability to choose the candidate model closest to the truth and the ability to estimate the unknown regression as well as an Oracle should be considered as desirable criteria that a model selection rule should satisfy in the "model false" case. In this context, it may be remarked that Bayarri et al. (2012) suggested several other desirable criteria for model selection procedures. In the "model true" case, we have been able to come up with precise conditions on the growth of $g = g_n \to \infty$ depending on that of $p = p_n \to \infty$ such that the posterior probability of the true model goes to 1 as $n \to \infty$. This, in particular, implies that the true model is chosen with probability tending to *one*. We have also derived conditions for attaining the Oracle loss asymptotically in this scenario. The specifications of the rate of growth of g in the "model false" and "model true" cases when p grows with n helps in making useful recommendations for appropriate choice of g. From our theoretical investigations it turns out that a choice of $g = kn^r$ with $r > \max\{4b, 1/2\}$ ensures desirable performance in both the "model false" and "model true" scenarios.

It has been argued in Section 5 that from practical considerations a choice of $g = n^2$ should be a reasonable one and this choice has been substantiated in the simulation studies presented in the paper. It needs special mention that we have been able to prove the fact that the true model is chosen with probability tending to one even in the case when the null model (i.e., model with only the intercept term) is the truth. A result of this nature is likely to be a useful addition to the literature.

This paper is an attempt to clarify the role of choice of g in variable selection using Zellner's g-prior when the sample size grows. It has been observed in Liang et al. (2008) that the use of g-prior in the fixed sample size scenario can lead to certain undesirable inference, which are referred to as paradoxes. The first of them, namely, the Bartlett paradox is observed when $g \to \infty$ while n and p remain fixed. The paradox lies in the fact that this method always favors the null model over other models even if it is not the truth when $g \to \infty$. It is to be noted that such a phenomenon occurs not only for a g-prior but also for other proper priors if we let the spread of the prior go to infinity. In our asymptotic framework, we have shown that the Bartlett paradox can be avoided in the sense that the true model will be selected with probability tending to *one* irrespective of whether the truth is the null model or a non-null model, provided $g \to \infty$ at some proper rate depending on the behavior of p as $n \to \infty$. It may thus be inferred that g should be allowed to go to infinity only when $n \to \infty$. The second paradox mentioned in Liang et al. (2008), namely, the Information paradox, loses much of its relevance when n is large since the Bayes factor in question goes to an extremely large number even for moderately large n, if $n - p$ is also moderately large and g is bounded away from $zero$.

We have considered the case when all the 2^{p_n} models are allowed in the model space. This is not to suggest that one should always consider this model space. In fact, when p_n is large, this may lead to serious computational issues and one probably would need something like a stochastic search and Monte Carlo simulation to find the model with the highest value of the criterion $p(M_\alpha)m_\alpha(\mathbf{y}_n)$. It would be interesting to see how the results obtained here would modify if one selects from a much smaller subset of the class of all possible models, for example, a specific nested sequence of models where α varies among $\{1\}, \{1, 2\}, \{1, 2, 3\}, \ldots, \{1, 2, \ldots, p\}$.

The method based on g-prior cannot be applied directly in situations where $p \geq n$. It is applicable when $p < n$ and we have simulation results even for the case $p = n-1$. For the case when $p \geq n$, there has been recent attempt by Maruyama and George (2011) to generalize the g-prior. However, they have not studied consistency in this setup and it remains an open problem worth studying.

It is worth investigating the issues of consistency and loss efficiency in the "model false" case for the local and global empirical Bayes choices of g and also for mixtures of g-priors (see, for example, Liang et al. 2008). It is also important to study model selection consistency for mixtures of g-priors for the case when p grows with n. Model selection using mixtures of g-priors necessitates calculation of marginal likelihood which may not have analytically tractable form. Standard approximations, like Laplace approximation, become very challenging when $p_n \to \infty$ since one needs to show that the error in approximation is uniformly small over the class of all candidate models. Another problem is the choice of the mixture from the wide class of mixtures available. All these are important issues which will be addressed in our future work.

7 Appendix

We present below proofs of some of the results presented in Sects. 2 and 3.

Proof of Lemma 1 **(a)** Let $U_{n\alpha} = \boldsymbol{\mu}_n'(I_n - P_n(\alpha))\mathbf{e}_n/\sqrt{\sigma^2\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\boldsymbol{\mu}_n}$. As $U_{n\alpha} \sim N(0,1)$, using the property of $N(0,1)$, we have for $t > 0$

$$P\left(\max_{\alpha\in\boldsymbol{\mathcal{A}}} |U_{n\alpha}| > t\right) \leq \sum_{\alpha\in\boldsymbol{\mathcal{A}}} P\left(|U_{n\alpha}| > t\right) \leq c_0 2^{p_n} e^{-\frac{t^2}{4}}$$

for some constant $c_0 > 0$. Then for $c > 0$

$$P\left(\max_{\alpha\in\boldsymbol{\mathcal{A}}} |U_{n\alpha}|/\sqrt{p_n} > c\right) \leq c_0 2^{p_n} e^{-c^2\frac{p_n}{4}}$$

which goes to zero as $n \to \infty$ for appropriately chosen c. Thus $\max\limits_{\alpha\in\boldsymbol{\mathcal{A}}} |U_{n\alpha}| = O_p(\sqrt{p_n})$. As $\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\boldsymbol{\mu}_n \leq \boldsymbol{\mu}_n'\boldsymbol{\mu}_n$, by assumption (A.1), the result follows.

(b) Let $V_{n\alpha} = \mathbf{e}_n' P_n(\alpha)\mathbf{e}_n/\sigma^2$. As $V_{n\alpha} \sim \chi^2_{p_n(\alpha)}$, using the Markov inequality and the moment generating function of χ^2 distribution we have for $0 < \lambda < \frac{1}{2}$ and $t > 0$,

$$\begin{aligned} P\left(\max_{\alpha\in\boldsymbol{\mathcal{A}}} \mathbf{e}_n' P_n(\alpha)\mathbf{e}_n > t\right) &\leq \sum_{\alpha\in\boldsymbol{\mathcal{A}}} P\left(V_{n\alpha} > \frac{t}{\sigma^2}\right) \\ &\leq \sum_{\alpha\in\boldsymbol{\mathcal{A}}} P(e^{\lambda V_{n\alpha}} > e^{\lambda t/\sigma^2}) \\ &\leq e^{-\lambda t/\sigma^2}\left(\frac{2}{\sqrt{1-2\lambda}}\right)^{p_n}. \end{aligned}$$

Thus for $c > 0$,

$$P\left(\max_{\alpha\in\boldsymbol{\mathcal{A}}} \frac{\mathbf{e}_n' P_n(\alpha)\mathbf{e}_n}{p_n} > c\right) \leq e^{-\lambda C p_n/\sigma^2}\left(\frac{2}{\sqrt{1-2\lambda}}\right)^{p_n}$$

which goes to zero as $n \to \infty$ for appropriately chosen λ and c and the result follows. □

Proof of (17) and (18) We have

$$C_n/n = (1-a_n)\sum_{i=1}^{n}(y_i - \overline{y})^2/n + a_n\sum_{i=1}^{n} e_i^2/n,$$

where $\sum_{i=1}^{n} e_i^2/n \xrightarrow{p} \sigma^2$.

Now $E(\sum_{i=1}^{n} y_i^2/n) = \sum_{i=1}^{n} \mu_i^2/n + \sigma^2$ and therefore,

$$P\left(\sum_{i=1}^{n} y_i^2/n > k\right) = \left(\sum_{i=1}^{n} \mu_i^2/n + \sigma^2\right)/k \to 0 \quad \text{as} \quad k \to \infty.$$

This implies $\sum_{i=1}^{n} y_i^2/n = O_p(1)$ and hence $\sum_{i=1}^{n}(y_i - \overline{y})^2/n = O_p(1)$. Thus (17) is proved.

Now from (16),

$$\begin{aligned} b_{n\alpha} &= e^{\log b_{n\alpha}} \\ &= \exp\left\{\frac{p_n(\alpha) - p_n(\hat{\alpha})}{n-1}\log(1+g_n) - \frac{2}{n-1}\log\left(\frac{p(M_\alpha)}{p(M_{\hat{\alpha}})}\right)\right\}. \end{aligned}$$

Therefore, by the mean value theorem,

$$b_{n\alpha} = 1 + \left\{\frac{p_n(\alpha) - p_n(\hat{\alpha})}{n-1}\log(1+g_n) - \frac{2}{n-1}\log\left(\frac{p(M_\alpha)}{p(M_{\hat{\alpha}})}\right)\right\} e^{U_n},$$

where U_n lies between 0 and $(p_n(\alpha) - p_n(\hat{\alpha}))\log(1+g_n)/(n-1) - 2\log(p(M_\alpha)/p(M_{\hat{\alpha}}))/(n-1)$. Then we have

$$\max_\alpha |n(b_{n\alpha} - 1)| \le 2\left[p_n \log(1+g_n) + 2\log C_{0n}\right]\exp\left\{\frac{p_n\log(1+g_n) + 2\log C_{0n}}{n-1}\right\}$$

and (18) is proved. □

Proof of Theorem 3. We first note that $nL_n(\alpha) = \|\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n(\alpha)\|^2$ for any fixed α, where $\hat{\boldsymbol{\mu}}_n(\alpha) = \mathbf{1}_n\hat{\beta}_0 + a_n\mathbf{X}_{n\alpha}\hat{\boldsymbol{\beta}}_\alpha = (1-a_n)\mathbf{1}_n\overline{y} + a_nP_n(\alpha)\mathbf{y}_n$. We also recall the definition of $\Psi(\alpha)$ as in (8). The proof hinges upon first establishing a relationship between $\Psi(\alpha)$ and $nL_n(\alpha)$. Towards that, we first observe that

$$\begin{aligned} nL_n(\alpha) &= \|\boldsymbol{\mu}_n - a_nP_n(\alpha)\mathbf{y}_n - (1-a_n)\mathbf{1}_n\overline{y}\|^2 \\ &= \left(\boldsymbol{\mu}_n - a_nP_n(\alpha)\mathbf{y}_n\right)'\left(\boldsymbol{\mu}_n - a_nP_n(\alpha)\mathbf{y}_n\right) + (1-a_n)^2n\overline{y}^2 \\ &\quad -2(1-a_n)\overline{y}\mathbf{1}_n'\left(\boldsymbol{\mu}_n - a_nP_n(\alpha)\mathbf{y}_n\right) \end{aligned}$$

where

$$\begin{aligned} &\left(\boldsymbol{\mu}_n - a_nP_n(\alpha)\mathbf{y}_n\right)'\left(\boldsymbol{\mu}_n - a_nP_n(\alpha)\mathbf{y}_n\right) \\ &\quad = \boldsymbol{\mu}_n'(I_n - P_n(\alpha))\boldsymbol{\mu}_n + a_n^2\mathbf{e}_n'P_n(\alpha)\mathbf{e}_n + (1-a_n)^2\boldsymbol{\mu}_n'P_n(\alpha)\boldsymbol{\mu}_n \\ &\quad\quad -2a_n(1-a_n)\boldsymbol{\mu}_n'P_n(\alpha)\mathbf{e}_n. \quad \left(\text{Putting } \mathbf{y}_n = \boldsymbol{\mu}_n + \mathbf{e}_n\right) \end{aligned}$$

Therefore,

$$\begin{aligned} nL_n(\alpha) &= \boldsymbol{\mu}_n'\boldsymbol{\mu}_n - a_n(2-a_n)\boldsymbol{\mu}_n' P_n(\alpha)\boldsymbol{\mu}_n - 2a_n(1-a_n)\boldsymbol{\mu}_n{\prime} P_n(\alpha)\mathbf{e}_n \\ &\quad + a_n^2\mathbf{e}_n' P_n(\alpha)\mathbf{e}_n + (1-a_n)^2 n\overline{y}^2 - 2(1-a_n)\overline{y}\sum_{i=1}^n \mu_i \\ &\quad + 2a_n(1-a_n)\overline{y}\mathbf{1}_n'\mathbf{y}_n, \end{aligned} \tag{36}$$

as $\mathbf{1}_n' P_n(\alpha) = \mathbf{1}_n'$. Then from (11) we have

$$\begin{aligned} &a_n\mathbf{y}_n'(I_n - P_n(\alpha))\mathbf{y}_n - \frac{1}{(2-a_n)} nL_n(\alpha) \\ &= 2a_n\boldsymbol{\mu}_n'\mathbf{e}_n + a_n\mathbf{e}_n'\mathbf{e}_n - \frac{(1-a_n)^2}{2-a_n}\boldsymbol{\mu}_n'\boldsymbol{\mu}_n + (1-a_n)^2 n\overline{y}^2 - 2(1-a_n)\overline{y}\sum_{i=1}^n \mu_i \\ &-\frac{2a_n}{2-a_n}\boldsymbol{\mu}_n' P_n(\alpha)\mathbf{e}_n - \frac{2a_n}{2-a_n}\mathbf{e}_n' P_n(\alpha)\mathbf{e}_n - \frac{2a_n(1-a_n)}{2-a_n} n\overline{y}^2 \end{aligned}$$

and therefore,

$$\begin{aligned} &(1-a_n)\sum_{i=1}^n (y_i - \overline{y})^2 + a_n\mathbf{y}_n'(I_n - P_n(\alpha))\mathbf{y}_n \\ &= C_n' + \frac{2a_n}{(2-a_n)}\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\mathbf{e}_n - \frac{2a_n}{(2-a_n)}\mathbf{e}_n' P_n(\alpha)\mathbf{e}_n + \frac{1}{(2-a_n)} nL_n(\alpha), \end{aligned} \tag{37}$$

where

$$\begin{aligned} C_n' &= (1-a_n)\sum_{i=1}^n (y_i - \overline{y})^2 + \frac{2a_n(1-a_n)}{(2-a_n)}\boldsymbol{\mu}_n'\mathbf{e}_n + a_n\mathbf{e}_n'\mathbf{e}_n - \frac{(1-a_n)^2}{(2-a_n)}\boldsymbol{\mu}_n'\boldsymbol{\mu}_n \\ &\quad + (1-a_n)^2 n\overline{y}^2 - 2(1-a_n)\overline{y}\sum_{i=1}^n \mu_i - \frac{2a_n(1-a_n)}{(2-a_n)} n\overline{y}^2. \end{aligned} \tag{38}$$

It follows from (8) and (37) that

$$\Psi(\alpha) = [p(M_\alpha)]^{-2/(n-1)}(1+g_n)^{p_n(\alpha)/(n-1)}\left[C_n' + \frac{1}{(2-a_n)} nL_n(\alpha)(1+\xi_n(\alpha))\right],$$

where

$$\xi_n(\alpha) = \frac{2a_n\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\mathbf{e}_n - 2a_n\mathbf{e}_n' P_n(\alpha)\mathbf{e}_n}{nL_n(\alpha)}. \tag{39}$$

The proof now proceeds along the lines of proof of Theorem 1, proving similar results in terms of $nL_n(\alpha)$ (in place of $D_n(\alpha)$ as in Theorem 1). Towards that, we first note that for all $\alpha \in \boldsymbol{\mathcal{A}}$, by following similar arguments as in proving (15) for Theorem 1, we have

$$
\begin{aligned}
1 &\le \frac{L_n(\hat{\alpha})}{\min_{\alpha\in\boldsymbol{\mathcal{A}}} L_n(\alpha)} \\
&\le \frac{C_n'/n}{(1-\xi_n)/(2-a_n)} \times \max_{\alpha} \frac{n(b_{n\alpha}-1)}{nL_n(\alpha)} + \frac{1+\xi_n}{1-\xi_n} \times \max_{\alpha} b_{n\alpha}, \qquad (40)
\end{aligned}
$$

where $\xi_n = \max_\alpha |\xi_n(\alpha)|$ and $b_{n\alpha}$ is as given in (16).

To prove (28), it therefore suffices to show that the right hand side of the inequality in (40) converges to 1 in probability for both the cases when p is fixed and when $p = p_n$ varies with n. This in turn can be proved by showing that

$$
\begin{aligned}
&C_n' = O_p(n), && (41)\\
&\xi_n \xrightarrow{p} 0, && (42)\\
&\max_{\alpha} \frac{n(b_{n,\alpha}-1)}{nL_n(\alpha)} \xrightarrow{p} 0, && (43)\\
\text{and} \quad &\max_{\alpha} b_{n\alpha} \to 1 && (44)
\end{aligned}
$$

as $n \to \infty$. First note that (44) has already been proved while proving Theorem 1 under both the cases when p is fixed and when $p = p_n$ grows with n.

To prove (41), recall the definition of C_n' as in (38). From (17), the first and third terms in (38) are of order $O_p(n)$. By assumption (A.1), $\boldsymbol{\mu}_n'\boldsymbol{\mu}_n = O(n)$, $\sum_{i=1}^n \mu_i/n = O(1)$ and $\boldsymbol{\mu}_n'\mathbf{e}_n/n \xrightarrow{p} 0$. As shown in the proof of (17), $\sum_{i=1}^n y_i^2/n = O_p(1)$ which implies $\overline{y}^2 = O_p(1)$ and $\overline{y} = O_p(1)$. As a_n is either fixed or $a_n \to 1$, all these imply (41). This proof covers both the cases when p is fixed and when $p = p_n$ grows with n.

Now we prove (42) and (43). From (39) we have,

$$
\max_{\alpha} |\xi_n(\alpha)| \le \frac{2a_n}{n} \frac{\max_\alpha \left|\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\mathbf{e}_n\right|}{\min_\alpha L_n(\alpha)} + \frac{2a_n}{n} \frac{\left(\mathbf{e}_n' P_n(\alpha)\mathbf{e}_n\right)}{\min_\alpha L_n(\alpha)}. \qquad (45)
$$

The proof of (42) will be complete by showing that under the assumptions of Theorem 3, for some constant $\delta_0 > 0$,

$$
\min_{\alpha} L_n(\alpha) \ge \frac{\delta_0}{n^s} \qquad (46)
$$

with probability tending to *one* as $n \to \infty$, where $s = 0$ in case $p = p_n$ remains fixed as $n \to \infty$ and $s \in \big[0, (1-b)/2\big)$ in case $p = p_n \to \infty$ as $n \to \infty$. This is so since (46) and the fact that $0 < a_n \le 1$ together imply that with probability tending to *one*, the right hand side of (45) is bounded above by

$$
\frac{2\max_\alpha \left|\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\mathbf{e}_n\right|/n}{\delta_0/n^s} + \frac{2\max_\alpha \left(\mathbf{e}_n' P_n(\alpha)\mathbf{e}_n\right)/n}{\delta_0/n^s}
$$

and by Lemma 1 the above expression is

$$\frac{O_p(\sqrt{p_n/n})}{\delta_0/s}+\frac{O_p(p_n/n)}{\delta_0/s}.$$

This tends to zero in probability when p is constant and $s=0$ and when $p=p_n=O\left(n^b\right)$ with $0<b<1$ and $s\in\left[0,(1-b)/2\right)$.

Let us now prove (46). Note that from (36) we have

$$\begin{aligned}L_n(\alpha)&=\frac{1}{n}\boldsymbol{\mu}_n'(I_n-P_n(\alpha))\boldsymbol{\mu}_n+(1-a_n)^2\frac{1}{n}\boldsymbol{\mu}_n'P_n(\alpha)\boldsymbol{\mu}_n+a_n^2\frac{1}{n}\mathbf{e}_n'P_n(\alpha)\mathbf{e}_n\\&\quad+(1-a_n^2)\overline{y}^2+2a_n(1-a_n)\frac{1}{n}\boldsymbol{\mu}_n'(I_n-P_n(\alpha))\mathbf{e}_n-2a_n(1-a_n)\frac{1}{n}\boldsymbol{\mu}_n'\mathbf{e}_n\\&\quad-2(1-a_n)\overline{y}\,\frac{1}{n}\sum_{i=1}^{n}\mu_i+2a_n(1-a_n)n\overline{y}^2.\end{aligned}$$

Now note that by taking $\delta_0'=\min(\Delta,\delta)$ where Δ and δ are as in assumptions (A.2) and (A.2)* respectively,

$$\min_{\alpha\in\boldsymbol{\mathcal{A}}}\frac{1}{n}\boldsymbol{\mu}_n'(I_n-P_n(\alpha))\boldsymbol{\mu}_n>\frac{\delta_0'}{n^s}$$

for all sufficiently large n, where $s=0$ in case of fixed p and $s\in\left[0,(1-b)/2\right)$ when p varies with n. Observe now that the second, third, fourth and eighth terms in the above expression for $L_n(\alpha)$ are always non-negative. From Lemma 1,

$$\max_{\alpha}\left|2a_n(1-a_n)\frac{1}{n}\boldsymbol{\mu}_n'(I_n-P_n(\alpha))\mathbf{e}_n\right|=O_p\left(\sqrt{\frac{p_n}{n}}\right)=o_p\left(n^{-s}\right)$$

both in the case $p=p_n$ is fixed and s = 0 and in the case $p=p_n=O(n^b), 0<b<1$ and $s<(1-b)/2$. Also, as $(1-a_n)=(1+g_n)^{-1}=O_p(n^{-r})$, $\boldsymbol{\mu}_n'\mathbf{e}_n/n\xrightarrow{p}0$, $\overline{y}=O_p(1)$, $\overline{\mu}=O_p(1)$ and $s<r$, the sixth and seventh terms are of order $O_p(n^{-r})$ and hence of the order $o_p(n^{-s})$.

Combining all these facts, one immediately gets (46) by taking $\delta_0=\delta_0'/2$.

The assertion (43) follows from (18), (46) and assumption (A.3), by noting that

$$\left|\max_{\alpha}\frac{n(b_{n\alpha}-1)}{nL_n(\alpha)}\right|\leq\frac{\max_\alpha|n(b_{n\alpha}-1)|}{\min_\alpha nL_n(\alpha)}.$$

The result now follows from (40)–(44). This proof covers the situation under parts **(a)** and **(b)** of the theorem. □

Proof of (28) for the least squares estimator. We describe how the above proof of Theorem 3 can be suitably adapted to deal with the case when one uses $\hat{\boldsymbol{\mu}}_n(\alpha)$ as the least squares estimator of $\boldsymbol{\mu}_n$ under model α and (28) can be shown to hold under this situation as well. We note that $\hat{\boldsymbol{\mu}}_n(\alpha) = P_n(\alpha)\mathbf{y}_n$ and therefore, $nL_n(\alpha) = \|\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n(\alpha)\|^2 = \|\boldsymbol{\mu}_n - P_n(\alpha)\mathbf{y}_n\|^2$. A little algebra then shows that

$$(1-a_n)\sum_{i=1}^{n}(y_i-\overline{y})^2 + a_n\mathbf{y}_n'(I_n - P_n(\alpha))\mathbf{y}_n = C_n' + a_n nL_n(\alpha)(1+\xi_n(\alpha)),$$

$$\text{where } C_n' = (1-a_n)\sum_{i=1}^{n}(y_i-\overline{y})^2 + a_n\mathbf{e}_n'\mathbf{e}_n$$

$$\text{and } \xi_n(\alpha) = \frac{2\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\mathbf{e}_n - 2\mathbf{e}_n'P_n(\alpha)\mathbf{e}_n}{nL_n(\alpha)}.$$

Therefore,

$$\Psi(\alpha) = (p(M_\alpha))^{-2/(n-1)}(1+g_n)^{p_n(\alpha)/(n-1)}\left(C_n' + a_n nL_n(\alpha)(1+\xi_n(\alpha))\right).$$

Using similar arguments as in proving (40), one can then show that for all $\alpha \in \boldsymbol{\mathcal{A}}$,

$$1 \le \frac{L_n(\hat{\alpha})}{\min_{\alpha\in\boldsymbol{\mathcal{A}}} L_n(\alpha)} \le \frac{C_n'/n}{a_n(1-\xi_n)} \times \max_\alpha \frac{n(b_{n\alpha}-1)}{nL_n(\alpha)} + \frac{1+\xi_n}{1-\xi_n} \times \max_\alpha b_{n\alpha},$$

where $\xi_n = \max_\alpha |\xi_n(\alpha)|$ and $b_{n\alpha}$ is as in the proof of Theorem 2.3. The desired result follows by observing that (41) through (44) continue to hold when one replaces the definitions of C_n', $\xi_n(\alpha)$ and $L_n(\alpha)$ used there with the corresponding ones written above in this particular case. The proof of (44) remains completely unchanged while those of (41) through (43) are only routine modifications of the corresponding proofs obtained above. The relatively simple details are left to the reader. □

Proof of Theorem 5. We first consider the following lemma.

Lemma 2 *Consider the setup of Theorem 5. Then for any $R > 2$, with probability tending to one,*

$$\max_{\alpha\in\boldsymbol{\mathcal{A}}_1} \frac{\mathbf{e}_n'(P_n(\alpha) - P_n(\alpha_c))\mathbf{e}_n}{\sigma^2(p_n(\alpha) - p_n(\alpha_c))} \le R\log p_n. \tag{47}$$

The proof of the Lemma 2 is given after the proof of Theorem 5.

Main Proof of Theorem 5. We shall prove that

$$\sum_{\alpha\in\boldsymbol{\mathcal{A}}_i} \frac{p(M_\alpha)m_\alpha(\mathbf{y}_n)}{p(M_{\alpha_c})m_{\alpha_c}(\mathbf{y}_n)} \xrightarrow{p} 0, \quad i = 1, 2. \tag{48}$$

Then the result will follow from (29) and (48).

Let M_N denote the null model under which $\boldsymbol{\mu}_n = \mathbf{1}_n\beta_0$. We will consider the two cases $M_{\alpha_c} \neq M_N$ and $M_{\alpha_c} = M_N$ separately. Also, we will prove part **(b)** of the theorem and then make a remark on the proof of part **(a)**.

Case 1: $M_{\alpha_c} \neq M_N$

Let us first prove (48) with $i = 2$. From (7) and (11), we have for $\alpha \in \boldsymbol{\mathcal{A}}_2$,

$$\frac{m_\alpha(\mathbf{y}_n)}{m_{\alpha_c}(\mathbf{y}_n)} = (1+g_n)^{(p_n(\alpha_c)-p_n(\alpha))/2} \times \left[\frac{1+a_n\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\boldsymbol{\mu}_n/nC_n + U_{n\alpha}}{1-a_n\mathbf{e}_n'P_n(\alpha_c)\mathbf{e}_n/nC_n}\right]^{-(n-1)/2}, \quad (49)$$

where $$C_n = (1-a_n)\sum(y_i - \overline{y})^2/n + a_n\mathbf{e}_n'\mathbf{e}_n/n \quad (50)$$

and $$U_{n\alpha} = a_n\left(2\mathbf{e}_n'(I_n - P_n(\alpha))\boldsymbol{\mu}_n - \mathbf{e}_n'P_n(\alpha)\mathbf{e}_n\right)/nC_n. \quad (51)$$

Note that $a_n \to 1$ and $\sum(y_i - \overline{y})^2/n = O_p(1)$ (see the proof of (17) in the Appendix) and therefore,

$$C_n \overset{p}{\to} \sigma^2. \quad (52)$$

Then by assumption (B.2)*,

$$\min_{\alpha\in\boldsymbol{\mathcal{A}}_2} \frac{a_n}{nC_n}\boldsymbol{\mu}_n'(I_n - P_n(\alpha))\boldsymbol{\mu}_n > \frac{c_1}{n^s} \quad (53)$$

with probability tending to *one* for some constant $c_1 > 0$ and by Lemma 1

$$\max_{\alpha\in\boldsymbol{\mathcal{A}}_2}\left|\frac{2a_n}{nC_n}\mathbf{e}_n'(I_n - P_n(\alpha))\boldsymbol{\mu}_n\right| = O_p\left(\sqrt{\frac{p_n}{n}}\right) \quad (54)$$

$$\text{and} \quad \max_{\alpha\in\boldsymbol{\mathcal{A}}_2}\frac{a_n}{nC_n}\mathbf{e}_n'P_n(\alpha)\mathbf{e}_n = O_p\left(\frac{p_n}{n}\right) \quad (55)$$

with probability tending to *one*. As $(1-b)/2 < s$, from (49), (51), (53), (54) and (55) it follows that for all $\alpha \in \boldsymbol{\mathcal{A}}_2$,

$$\frac{m_\alpha(\mathbf{y}_n)}{m_{\alpha_c}(\mathbf{y}_n)} \leq (1+g_n)^{p_n/2}\left(1+\frac{\delta_0}{n^s}\right)^{-(n-1)/2}$$

for some $\delta_0 > 0$, not depending on α. Then by assumption (31), we have with probability tending to *one*

$$\sum_{\alpha\in\boldsymbol{\mathcal{A}}_2}\frac{p(M_\alpha)m_\alpha(\mathbf{y}_n)}{p(M_{\alpha_c})m_{\alpha_c}(\mathbf{y}_n)} \leq 2^{p_n}k_0n^{b_0}(1+g_n)^{p_n/2}\left(1+\frac{\delta_0}{n^s}\right)^{-(n-1)/2} \quad (56)$$

which goes to *zero* as $b < (1-s)$. To see this, note that logarithm of the right hand side of (56) is equal to

$$p_n\log 2 + \log k_0 + b_0\log n + \frac{p_n}{2}\log(1+kn^r) - \frac{(n-1)}{2}\times\frac{\delta_0}{n^s}\times\frac{1}{\xi_n},$$

where ξ_n lies between 1 and $1 + \delta_0/n^s$ and $p_n = O(n^b)$. The value of the above expression goes to $-\infty$. Thus (48) is proved with i $= 2$.

We now prove (48) with i $= 1$. We have borrowed some ideas used in the proof of Theorem 2.2 of Shang and Clayton (2011) for proving (59) below. As $P_n(\alpha)\boldsymbol{\mu}_n = \boldsymbol{\mu}_n$ for all $\alpha \in \boldsymbol{\mathcal{A}}_1$, from (7) and (11) we have for all $\alpha \in \boldsymbol{\mathcal{A}}_1$,

$$\begin{aligned}\frac{m_\alpha(\mathbf{y}_n)}{m_{\alpha_c}(\mathbf{y}_n)} &= \frac{1}{(1+g_n)^{(p_n(\alpha)-p_n(\alpha_c))/2}} \times \left[\frac{nC_n - a_n\mathbf{e}_n' P_n(\alpha)\mathbf{e}_n}{nC_n - a_n\mathbf{e}_n' P_n(\alpha_c)\mathbf{e}_n}\right]^{-(n-1)/2} \\ &= \frac{1}{(1+g_n)^{(p_n(\alpha)-p_n(\alpha_c))/2}} \times \left[1 - \frac{\mathbf{e}_n'(P_n(\alpha) - P_n(\alpha_c))\mathbf{e}_n}{nC_n/a_n - \mathbf{e}_n' P_n(\alpha_c)\mathbf{e}_n}\right]^{-(n-1)/2};\end{aligned} \tag{57}$$

where C_n is as defined above in (50). Note that $\mathbf{e}_n' P_n(\alpha_c)\mathbf{e}_n/n \xrightarrow{p} 0$ and therefore, from (52)

$$\frac{1}{n}\left(\frac{n}{a_n}C_n - \mathbf{e}_n' P_n(\alpha_c)\mathbf{e}_n\right) \xrightarrow{p} \sigma^2 \tag{58}$$

which implies that for any $0 < \delta_1 < 1$, $\left(nC_n/a_n - \mathbf{e}_n' P_n(\alpha_c)\mathbf{e}_n\right)/n > \sigma^2(1-\delta_1)$ with probability tending to *one*.

Also, consider the following inequality,

$$(1-x) \geq e^{-x/(1-\delta_1)} \quad \text{for } 0 < x < \delta_1 < 1.$$

Now, using the above inequality along with (57), (58), Lemma 2 and assumption (31), we have with probability tending to *one*, for any $R > 2$ and any $0 < \delta_1 < 1$,

$$\begin{aligned}&\sum_{\alpha \in \boldsymbol{\mathcal{A}}_1} \frac{p(M_\alpha)m_\alpha(\mathbf{y}_n)}{p(M_{\alpha_c})m_{\alpha_c}(\mathbf{y}_n)} \\ &\leq \sum_{\alpha \in \boldsymbol{\mathcal{A}}_1} \frac{k_0 n^{b_0}}{(\sqrt{1+g_n})^{p_n(\alpha)-p_n(\alpha_c)}} \left[1 - \frac{\mathbf{e}_n'(P_n(\alpha) - P_n(\alpha_c))\mathbf{e}_n}{n\sigma^2(1-\delta_1)}\right]^{-(n-1)/2} \\ &\leq \sum_{\alpha \in \boldsymbol{\mathcal{A}}_1} \frac{k_0 n^{b_0}}{(\sqrt{1+g_n})^{p_n(\alpha)-p_n(\alpha_c)}} \left[1 - \frac{R(p_n(\alpha) - p_n(\alpha_c))\log p_n}{n(1-\delta_1)}\right]^{-(n-1)/2} \\ &\leq \sum_{\alpha \in \boldsymbol{\mathcal{A}}_1} \frac{k_0 n^{b_0}}{(\sqrt{1+g_n})^{p_n(\alpha)-p_n(\alpha_c)}} \exp\left[\frac{R(p_n(\alpha) - p_n(\alpha_c))\log p_n}{2(1-\delta_1)^2}\right] \\ &= \sum_{\alpha \in \boldsymbol{\mathcal{A}}_1} \frac{k_0 n^{b_0}}{(\sqrt{1+g_n})^{p_n(\alpha)-p_n(\alpha_c)}} p_n^{\left[R(p_n(\alpha)-p_n(\alpha_c))/2(1-\delta_1)^2\right]}\end{aligned}$$

$$= k_0 n^{b_0} \sum_{q=1}^{p_n - p_n(\alpha_c)} \binom{p_n - p_n(\alpha_c)}{q} \left(\frac{p_n^{R/2(1-\delta_1)^2}}{\sqrt{1+g_n}} \right)^q$$

$$= \left[\left(1 + \frac{p_n^{R/2(1-\delta_1)^2}}{\sqrt{1+g_n}} \right)^{p_n - p_n(\alpha_c)} - 1 \right] k_0 n^{b_0}. \tag{59}$$

Using the mean value theorem, the last expression in (59) can be shown to be bounded above by

$$k_0 n^{b_0} \frac{{p_n}^{R/2(1-\delta_1)^2}}{\sqrt{1+g_n}} p_n \left(1 + \frac{{p_n}^{R/2(1-\delta_1)^2}}{\sqrt{1+g_n}} \right)^{p_n}. \tag{60}$$

As $p_n = O(n^b)$ and $g_n = kn^r$,

$$\left(1 + \ {p_n}^{R/2(1-\delta_1)^2} / \sqrt{1+g_n} \right)^{p_n} \to 1 \tag{61}$$

$$\text{if} \qquad r/2 - Rb/2(1-\delta_1)^2 > b$$

$$\text{that is, if} \qquad r/b > 2 + R/(1-\delta_1)^2. \tag{62}$$

Also,

$$k_0 n^{b_0} \ p_n \ {p_n}^{R/2(1-\delta_1)^2} / \sqrt{1+g_n} \to 0 \tag{63}$$

if

$$b_0 < r/2 - b - Rb/2(1-\delta_1)^2 = \left[r - b \left(2 + R/(1-\delta_1)^2 \right) \right] /2. \tag{64}$$

Under the assumptions of the theorem, $r > 4b$ and $b_0 < (r - 4b)/2$. As $2 + R/(1 - \delta_1)^2 \to 4$ when $R \downarrow 2$ and $\delta_1 \downarrow 0$, there exist $R > 2$ and $0 < \delta_1 < 1$ such that (62) and (64) hold and hence (61) and (63) hold. Then from (59) and (60) with these R and δ_1, (48) follows with $i = 1$. This completes the proof of part **(b)** of the theorem for Case 1, that is, when $M_{\alpha_c} \neq M_N$.

In order to prove part **(a)** of the theorem it is enough to prove that for each fixed $\alpha \in \boldsymbol{\mathcal{A}}$, $\alpha \neq \alpha_c$,

$$\frac{p(M_\alpha) m_\alpha(\mathbf{y}_n)}{p(M_{\alpha_c}) m_{\alpha_c}(\mathbf{y}_n)} \xrightarrow{p} 0.$$

The proof is much easier and is essentially contained in the above proof where we use assumption (B.2)* with $s = 0$ which is the same as assumption (B.2).

Case 2: $M_{\alpha_c} = M_N$

In this case $\boldsymbol{\mathcal{A}}_2$ is empty and therefore, we need to prove (48) only for $\boldsymbol{\mathcal{A}}_1$. Under the null model $\boldsymbol{\mu}_n = \mathbf{1}_n \beta_0$, $\mathbf{y}_n = \mathbf{1}_n \beta_0 + \mathbf{e}_n$ and $P_n(\alpha_c) = \mathbf{1}_n [\mathbf{1}_n' \mathbf{1}_n]^{-1} \mathbf{1}_n'$. Therefore, $P_n(\alpha_c) \mathbf{y}_n = \mathbf{1}_n \overline{y}$ and $\mathbf{y}_n'(I_n - P_n(\alpha_c)) \mathbf{y}_n = \sum_{i=1}^n (y_i - \overline{y})^2$. As $\mathbf{1}_n' P_n(\alpha) = \mathbf{1}_n'$, for

$\alpha \in \boldsymbol{\mathcal{A}}_1$, $\mathbf{y}_n'(I_n - P_n(\alpha))\mathbf{y}_n = \mathbf{e}_n'\mathbf{e}_n - \mathbf{e}_n' P_n(\alpha)\mathbf{e}_n$. Then from (7), for $\alpha \in \boldsymbol{\mathcal{A}}_1$ we have

$$\frac{m_\alpha(\mathbf{y}_n)}{m_{\alpha_c}(\mathbf{y}_n)} = \frac{1}{(1+g_n)^{p_n(\alpha)/2}} \times \left[\frac{(1-a_n)\sum_{i=1}^n (y_i - \overline{y})^2 + a_n \mathbf{e}_n'\mathbf{e}_n - a_n \mathbf{e}_n' P_n(\alpha)\mathbf{e}_n}{\sum_{i=}^n (y_i - \overline{y})^2}\right]^{-(n-1)/2}.$$

As

$$\mathbf{e}_n'\mathbf{e}_n = \sum_{i=1}^n (y_i - \beta_0)^2 \geq \sum_{i=1}^n (y_i - \overline{y})^2,$$

that is,

$$(1-a_n)\sum_{i=1}^n (y_i - \overline{y})^2 + a_n \mathbf{e}_n'\mathbf{e}_n \geq \sum_{i=1}^n (y_i - \overline{y})^2,$$

we have

$$\frac{m_\alpha(\mathbf{y}_n)}{m_{\alpha_c}(\mathbf{y}_n)} \leq \frac{1}{(1+g_n)^{p_n(\alpha)/2}} \left[1 - \frac{\mathbf{e}_n' P_n(\alpha)\mathbf{e}_n}{(1-a_n)\sum_{i=1}^n (y_i - \overline{y})^2/a_n + \mathbf{e}_n'\mathbf{e}_n}\right]^{-(n-1)/2}. \tag{65}$$

Noting that

$$\left((1-a_n)\sum_{i=1}^n (y_i - \overline{y})^2/a_n + \mathbf{e}_n'\mathbf{e}_n\right)/n \xrightarrow{p} \sigma^2 \tag{66}$$

and comparing (65) and (66) with (57) and (58), one can see that the rest of the proof of the result in this case is exactly similar to that in Case 1. □

Proof of Lemma 2. As $P_n(\alpha) - P_n(\alpha_c)$ is a projection matrix for $\alpha \in \boldsymbol{\mathcal{A}}_1$,

$$\frac{1}{\sigma^2}\mathbf{e}_n'(P_n(\alpha) - P_n(\alpha_c))\mathbf{e}_n \sim \chi^2_{p_n(\alpha)-p_n(\alpha_c)}.$$

Let $Y \sim \chi^2_\gamma$. Then for any $R > 0$ and any $0 < \delta < 1$,

$$\begin{aligned}
&P\left(Y > R\gamma \log(p_n)\right) \\
&= \int_{R\gamma \log(p_n)}^{\infty} \frac{1}{2^{\gamma/2}\Gamma(\gamma/2)} e^{-y/2} y^{\gamma/2-1} dy \\
&= \int_{R\gamma \log(p_n)}^{\infty} \frac{1}{2^{\gamma/2}\Gamma(\gamma/2)} e^{-y(1-\delta+\delta)/2} y^{\gamma/2-1} dy \\
&\leq \frac{1}{\delta^{\gamma/2}} e^{-R\gamma(1-\delta)\log(p_n)/2} \int_{R\gamma \log(p_n)}^{\infty} \frac{\delta^{\gamma/2}}{2^{\gamma/2}\Gamma(\gamma/2)} e^{-y\delta/2} y^{\gamma/2-1} dy \\
&\leq \frac{1}{\delta^{\gamma/2}} e^{-R\gamma(1-\delta)\log(p_n)/2} \int_{0}^{\infty} \frac{\delta^{\gamma/2}}{2^{\gamma/2}\Gamma(\gamma/2)} e^{-y\delta/2} y^{\gamma/2-1} dy \\
&= \frac{1}{\delta^{\gamma/2}} p_n^{-\frac{R\gamma(1-\delta)}{2}}.
\end{aligned}$$

Using this result we have

$$
\begin{aligned}
&P\left(\max_{\alpha\in\boldsymbol{\mathcal{A}}_1}\frac{\mathbf{e}_n'(P_n(\alpha)-P_n(\alpha_c))\mathbf{e}_n}{\sigma^2(p_n(\alpha)-p_n(\alpha_c))}>R\log(p_n)\right)\\
&\quad\le\sum_{\alpha\in\boldsymbol{\mathcal{A}}_1}P\left(\frac{\mathbf{e}_n'(P_n(\alpha)-P_n(\alpha_c))\mathbf{e}_n}{\sigma^2(p_n(\alpha)-p_n(\alpha_c))}>R\log(p_n)\right)\\
&\quad=\sum_{\alpha\in\boldsymbol{\mathcal{A}}_1}P\left(\frac{\mathbf{e}_n'(P_n(\alpha)-P_n(\alpha_c))\mathbf{e}_n}{\sigma^2}>R(p_n(\alpha)-p_n(\alpha_c))\log(p_n)\right)\\
&\quad\le\sum_{\alpha\in\boldsymbol{\mathcal{A}}_1}{p_n}^{-R(p_n(\alpha)-p_n(\alpha_c))(1-\delta)/2}/\delta^{(p_n(\alpha)-p_n(\alpha_c))/2}\\
&\quad=\sum_{q=1}^{(p_n-p_n(\alpha_c))}\binom{p_n-p_n(\alpha_c)}{q}\left(\frac{1}{\sqrt{\delta}\;p_n^{R(1-\delta)/2}}\right)^q\\
&\quad=\left(1+\frac{1}{\sqrt{\delta}\;p_n^{R(1-\delta)/2}}\right)^{(p_n-p_n(\alpha_c))}-1.
\end{aligned}
\tag{67}
$$

As $p_n=O(n^b),\ b>0$, the above expression goes to zero if $R(1-\delta)b/2>b$, that is, if $R>2/(1-\delta)$. If $R>2$, there exits $0<\delta<1$ such that $R>2/(1-\delta)$. Using (67) with these R and δ, we have the result. □

Acknowledgments We are thankful to the associate editor and the referees for their very useful comments and suggestions that helped us improve the paper.

References

Bayarri, M. J., Berger, J. O., Forte, A., García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, *40*(3), 1550–1577. doi:10.1214/12-AOS1013.

Bornn, L., Doucet, A., Gottardo, R. (2010). An efficient computational approach for prior sensitivity analysis and cross-validation. *The Canadian Journal of Statistics La Revue Canadienne de Statistique*, *38*(1), 47–64. doi:10.1002/cjs.10045.

Chakrabarti, A., Ghosh, J. K. (2006). A generalization of BIC for the general exponential family. *Journal of Statistical Planning and Inference*, *136*(9), 2847–2872. doi:10.1016/j.jspi.2005.01.005.

Chakrabarti, A., 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*, Institute of Mathematical Statistics Collections, vol 3. Institute of Mathematical Statistics, Beachwood, OH, pp. 138–154, doi:10.1214/074921708000000110.

Chaturvedi, A., Hasegawa, H., Asthana, S. (1997). Bayesian analysis of the linear regression model with non-normal disturbances. *The Australian Journal of Statistics*, *39*(3), 277–293. doi:10.1111/j.1467-842X.1997.tb00692.x.

Chipman, H., George, E.I., McCulloch, R.E. (2001). The practical implementation of Bayesian model selection. In: *Model selection, IMS Lecture Notes - Monograph Series, vol 38*, Institute of Mathematical Statistics, Beachwood, OH, pp. 65–134, doi:10.1214/lnms/1215540964, with discussion by M. Clyde, Dean P. Foster, and Robert A. Stine, and a rejoinder by the authors.

Consonni, G., Veronese, P. (2008). Compatibility of prior specifications across linear models. *Statistical Science A Review Journal of the Institute of Mathematical Statistics*, *23*(3), 332–353. doi:10.1214/08-STS258.

Fernández, C., Ley, E., Steel, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, *100*(2), 381–427. doi:10.1016/S0304-4076(00)00076-2.

Foster, D. P., George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, *22*(4), 1947–1975. doi:10.1214/aos/1176325766.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association*, *95*(452), 1304–1308. doi:10.2307/2669776.

George, E. I., Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, *87*(4), 731–747. doi:10.1093/biomet/87.4.731.

Kass, R. E., Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association, 90*(431), 928–934. http://links.jstor.org/sici?sici=0162-1459(199509)90:431<928:ARBTFN>2.0.CO;2-B&origin=MSN.

Krishna, A., Bondell, H. D., Ghosh, S. K. (2009). Bayesian variable selection using an adaptive powered correlation prior. *Journal of Statistical Planning and Inference*, *139*(8), 2665–2674. doi:10.1016/j.jspi.2008.12.004.

Li, K. C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics*, *15*(3), 958–975. doi:10.1214/aos/1176350486.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481), 410–423. doi:10.1198/016214507000001337.

Maruyama, Y., George, E. I. (2011). Fully Bayes factors with a generalized g-prior. *The Annals of Statistics*, *39*(5), 2740–2765. doi:10.1214/11-AOS917.

Miller, A. (2001). *Subset selection in regression* (2nd ed.). New York: Chapman and Hall.

Shang, Z., Clayton, M. K. (2011). Consistency of Bayesian linear model selection with a growing number of parameters. *Journal of Statistical Planning and Inference*, *141*(11), 3463–3474. doi:10.1016/j.jspi.2011.05.002.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, *7*(2), 221–264, with comments and a rejoinder by the author.

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.), pp. 233–243. Amsterdam: North-Holland.