TEST (2017) 26:377–404
DOI 10.1007/s11749-016-0516-0

ORIGINAL PAPER

A mixture of g-priors for variable selection when the number of regressors grows with the sample size

Minerva Mukhopadhyay[1] · Tapas Samanta[2]

Received: 29 April 2016 / Accepted: 25 November 2016 / Published online: 20 December 2016

Abstract We consider the problem of variable selection in linear regression using mixtures of g-priors. A number of mixtures have been proposed in the literature which work well, especially when the number of regressors p is fixed. In this paper, we propose a mixture of g-priors suitable for the case when p grows with the sample size n, more specifically when $p = O(n^b)$, $0 < b < 1$. The marginal density based on the proposed mixture has a nice approximation with a closed form expression, which makes application of the method as tractable as an information criterion-based method. The proposed method satisfies fundamental properties like model selection consistency when the true model lies in the model space, and also consistency in an appropriate sense, under misspecified models setup. The method is quite robust in the sense that the above properties are not confined to normal linear models; they continue to hold under reasonable conditions for a general class of error distributions. Finally, we compare the performance of the proposed prior theoretically with that of some other mixtures of g-priors. We also compare it with several other Bayesian methods of model selection using simulated data sets. Theoretically, as well as in simulations, it emerges that unlike most of the other methods of model selection, the proposed prior is competent enough while selecting the true model irrespective of its dimension.

Keywords Model selection consistency · Misspecified models · General class of distributions of errors · Kullback–Leibler divergence

Electronic supplementary material The online version of this article (doi:10.1007/s11749-016-0516-0) contains supplementary material, which is available to authorized users.

✉ Minerva Mukhopadhyay
minervamukherjee@gmail.com

[1] Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata, India

[2] Applied Statistics Unit, Indian Statistical Institute, Kolkata, India

Mathematics Subject Classification 62F15 Bayesian inference · 62F12 Asymptotic properties of estimators

1 Introduction

We consider the regression setup with response variable y and a set of p potential regressors $x_1, x_2, \ldots, x_p$. Let $\mathbf{y}_n = (y_1, y_2, \ldots, y_n)$ be a set of n observations on y, and $\mathbf{X}_n = \left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p\right)$ be the $n \times p$ design matrix, where $\mathbf{x}_i$ is the vector of n observations on the i^{th} regressor x_i for $i = 1, 2, \ldots, p$. We write

$$\mathbf{y}_n = \boldsymbol{\mu}_n + \mathbf{e}_n, \tag{1}$$

where $\boldsymbol{\mu}_n = E(\mathbf{y}_n|\mathbf{X}_n)$ is the regression of $\mathbf{y}_n$ on $\mathbf{X}_n$ and $\mathbf{e}_n$ is the vector of random errors. If we assume the normal linear regression model, then $\boldsymbol{\mu}_n = \beta_0\mathbf{1} + \mathbf{X}_n\boldsymbol{\beta}$ and $\mathbf{e}_n \sim N_n(\mathbf{0}, \sigma^2 I)$. Here β_0 is the intercept, $\mathbf{1}$ and $\mathbf{0}$ are the $n \times 1$ vectors of ones and zeros, respectively, and $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \ldots, \beta_p\right)'$ is the vector of regression coefficients.

In this article, we study the problem of variable selection. Given a set of p available regressor variables, there are 2^p possible linear regression models. The space of all these models is denoted by $\mathcal{M}$ and indexed by γ, where each γ consists of a subset of size $p(\gamma)$ $(0 \le p(\gamma) \le p)$ of the set $\{1, 2, \ldots p\}$, indicating the regressors selected in the model. If the model M_γ corresponding to some $\gamma \in \mathcal{M}$ is assumed to be true then $\boldsymbol{\mu}_n = \beta_0\mathbf{1} + \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma$ and M_γ can be expressed as

$$M_\gamma : \mathbf{y}_n = \beta_0\mathbf{1} + \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma + \mathbf{e}_n, \tag{2}$$

where $\mathbf{X}_\gamma$ is a sub-matrix of $\mathbf{X}_n$ consisting of the $p(\gamma)$ columns specified by γ, and $\boldsymbol{\beta}_\gamma$ is the corresponding vector of regression coefficients. We assume that all the components of $\boldsymbol{\beta}_\gamma$ are non-zero. This ensures that there is at most one true model in $\mathcal{M}$.

In a Bayesian approach, each model M_γ is associated with a prior probability $P(M_\gamma)$ and the corresponding set of parameters $\boldsymbol{\theta}_\gamma = \left(\beta_0, \boldsymbol{\beta}_\gamma, \sigma^2\right)'$ involved in the model is also associated with a prior distribution $\pi(\boldsymbol{\theta}_\gamma|M_\gamma)$. Given the priors, one computes the posterior probability of M_γ as

$$P(M_\gamma|\mathbf{y}_n) = \frac{P(M_\gamma)m_\gamma(\mathbf{y}_n)}{\sum_{\gamma\in\mathcal{M}} P(M_\gamma)m_\gamma(\mathbf{y}_n)}, \tag{3}$$

where

$$m_\gamma(\mathbf{y}_n) = \int f(\mathbf{y}_n|\boldsymbol{\theta}_\gamma, M_\gamma)\pi(\boldsymbol{\theta}_\gamma|M_\gamma)d\boldsymbol{\theta}_\gamma \tag{4}$$

is the marginal density of $\mathbf{y}_n$ and $f(\mathbf{y}_n|\boldsymbol{\theta}_\gamma, M_\gamma)$ is the density of $\mathbf{y}_n$ given $\boldsymbol{\theta}_\gamma$ under M_γ. In our search for a model, $f(\mathbf{y}_n|\boldsymbol{\theta}_\gamma, M_\gamma)$ will be taken as normal. We consider the model selection procedure that selects the model in $\mathcal{M}$ with the highest posterior probability.

A prevalent conventional prior on $\boldsymbol{\beta}_\gamma$ is the g-prior due to Zellner (1986). Properties of the method based on g-prior are studied extensively in the literature (see, e.g., Liang et al. 2008; Fernández et al. 2001). This method crucially depends on the choice of the hyperparameter g (see, e.g., Berger and Pericchi 2001; Liang et al. 2008). Moreover, this method is subject to inconsistencies like *Bartlett paradox* (see Bartlett 1957; Jeffreys 1961) and *information paradox* (see Zellner 1986; Berger and Pericchi 2001). Liang et al. (2008) considered a prior on g instead of considering a fixed g to remove these inconsistencies. Subsequently, a number of mixtures of g-priors are proposed in the literature.

We work in the mixture of g-priors setup as considered in Liang et al. (2008). The complete prior specification is given by

$$\pi(\beta_0, \sigma^2 | M_\gamma) = \frac{1}{\sigma^2}, \ \boldsymbol{\beta}_\gamma | \beta_0, \sigma^2, g, M_\gamma \sim N_{p(\gamma)}(\mathbf{0}, g\sigma^2(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^{-1}), \quad g \sim \pi(g). \tag{5}$$

Here, without loss of generality, we assume that the columns of $\mathbf{X}_\gamma$ are centered so that $\mathbf{1}'\mathbf{x}_i = 0$ for all i. We do not consider any specific prior probability on the model space, rather impose some conditions on model prior probabilities, $P(M_\gamma)$, under which our results hold.

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

Bayarri et al. (2012) described some desirable properties a prior should satisfy in the context of model selection which are satisfied by the robust prior. Ley and Steel (2012) made an extensive simulation study to compare several priors. However, none of them considered the case where p increases with n. Maruyama and George (2011) proposed a prior which is applicable when $p > n$, but proved consistency of their method for the case when p is fixed. Shang and Clayton (2011) proved consistency for mixtures of g-priors for growing p, but their setup differs from the usual g-prior setup with respect to the covariance structure of the prior distribution of $\boldsymbol{\beta}_\gamma$. Wang and Sun (2014) and Xiang et al. (2016) investigated properties of different mixtures for the case with growing number of regressors. However, they only established results for Bayes factor consistency for pairwise comparison of models. Recently, Moreno et al. (2015) have studied properties of the g-prior with $g = n$ and the Zellner–Siow prior when p grows with n.

The motivation of this work is to develop a consistent and robust model selection method suitable for *'large p large n'* regime. We propose a prior $\pi(g)$ on g, suitable for the case where p increases with n at a rate $p = O(n^b)$, $0 < b < 1$ and $p < n$ (see, e.g., Sparks et al. 2015; Johnson and Rossell 2012; Wang and Sun 2014; Moreno et al. 2015 for related work in this setup). The proposed mixture belongs to the family of the Zellner–Siow prior. In a sense, we consider a modified form of the Zellner–Siow prior by choosing an appropriate scale parameter. An advantage of this prior is that it provides an approximation to the marginal density $m_\gamma(\mathbf{y}_n)$ in (4) with a closed-form expression which facilitates easy implementation and theoretical studies. Further, it satisfies many attractive consistency properties when p increases with n.

In our study, the following two situations are considered separately. First, we consider the setup where the true model belongs to the model space $\mathcal{M}$, i.e., the true regression $\boldsymbol{\mu}_n$ is as in (2) for some γ. A well-known notion of consistency in this regard is *model selection consistency* which requires that the posterior probability of the true model goes to one as $n \to \infty$. The proposed mixture is shown to be model selection consistent irrespective of the dimension of the true model. We also show that most of the other available mixtures fail to be model selection consistent for similar rate of increase in p under sparse situations. We next consider the case where the models are misspecified. Here, $\boldsymbol{\mu}_n$ can be any unknown vector, not necessarily in the span of $\{\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_p\}$. Consistency in this case refers to the property of a model selection procedure to choose that model in $\mathcal{M}$ which is closest to the unknown true model in some asymptotic sense. We investigate consistency of the proposed prior under the misspecified models setup, using an appropriate notion of consistency (see, e.g., Chakrabarti and Ghosh 2006; Chakrabarti and Samanta 2008; Lv and Liu 2014; Mukhopadhyay et al. 2015 for related work).

Presence of the information paradox in Zellner's g-prior is one of the motivations for considering mixture of g-priors. We, therefore, verify whether the proposed mixture can resolve the information paradox, i.e., is information consistent in the sense of Bayarri et al. (2012).

Finally, we compare the performance of our proposed mixture with several other well-known Bayesian methods under the setup where p grows with n and $p < n$. Our simulation results show that unlike most of the other methods which perform well for specific ranges of the dimension of the true model, the proposed method is able to select the true model irrespective of its dimension.

In Sect. 2, we define the proposed prior and discuss the motivation for considering the same. The approximation of the marginal density to a closed form expression and its rate of accuracy are also discussed here. Sections 3 and 4 of the paper deal with model selection consistency. In Sect. 3, we consider the case with normal errors. The performance of the proposed mixture is studied in Sect. 3.1, while that of several other mixtures is studied in Sect. 3.2. In Sect. 4 we relax the assumption of normality of $\mathbf{e}_n$ and prove model selection consistency for a general class of error distributions. Section 5 deals with the case of misspecified models. We find sufficient conditions under which the proposed prior is consistent in an appropriate sense for the general class of error distributions. In Sect. 6, information consistency is considered. In Sect. 7, we present our simulation studies. Section 8 contains concluding remarks. Proofs of some important results are presented in the 'Appendix', and those of the other results and a part of our simulation studies are presented in the supplementary file.

2 The proposed mixture of g-priors

We first motivate our proposal for a mixture on g. Most of the priors on g in the literature are extremely right-skewed having a unique modal point close to zero and a very flat decay. For example, the hyper-g and g/n priors are positive-tailed J-shaped with modal point at zero. The robust prior of Bayarri et al. (2012) is a truncated one which moves the support away from zero and still is extremely right-skewed having the modal point as the point of truncation.

On the other hand, if we consider popular recommendations of g in Zellner's g-prior, choices include the unit information prior ($g = n$, Kass and Raftery 1995), the choice of g related to the risk inflation criterion ($g = p^2$, see Foster and George 1994; George and Foster 2000), and the benchmark prior ($g = \max\{n, p^2\}$, Fernández et al. 2001). Recently, Mukhopadhyay et al. (2015) presented some theoretical results to explain why a relatively larger value of g yields better results, especially when p grows with n and recommended using $g = n^2$ for practical purposes. From such recommendations, it seems reasonable to put relatively higher probability masses to higher values of g for a mixture. Thus, there persists a gap in the domain of g getting relatively higher mass when a fixed g is considered compared to that of a mixture. Here, we propose a class of mixtures which gives more probability mass to a range of relatively higher values of g compared to the existing mixtures. We consider the *scaled inverse chi-square* prior $\pi(g)$ on g with scale parameter τ^2 and degrees of freedom ν, which is same as *Inv-Gamma* $(\nu/2, \tau^2\nu/2)$ prior on g, given by

$$\pi(g) = \frac{\left(\tau^2\nu/2\right)^{\nu/2}}{\Gamma(\nu/2)} \frac{\exp\left[-\tau^2\nu/(2g)\right]}{g^{1+\nu/2}}, \quad g > 0, \quad \nu > 0, \quad \tau^2 > 0. \tag{6}$$

We choose the hyperparameter $\tau = n^r$, $1 \leq r < 2$. Note that such a choice of τ ensures that the prior has a unique mode at $n^{2r}\nu/(\nu + 2)$ and a very flat decay. Although the hyperparameter ν can take any positive value, we recommend using values between 1 and p. In this paper, we will consider two extreme choices of ν, namely $\nu = 1$ and $\nu = p$. We further discuss on the choices of the hyperparameters in Sect. 8.

Inverse gamma mixtures of g-prior has been previously used by Zellner and Siow (1980). In the context of linear regression models with shrinkage priors, Park and Casella (2008) and Hans (2009) used inverse gamma priors for similar normal scale mixtures for $\boldsymbol{\beta}_\gamma$, while using the Bayesian version of lasso. The mixture we propose results in a prior for $\boldsymbol{\beta}_\gamma$ with thick tails, which is also recommended by Jeffreys (1961).

An advantage of considering the proposed mixture of g-priors is the availability of an approximation of the marginal density to a closed form expression, which makes the method theoretically tractable and application much simpler. In the following subsection, we give an explicit form of the approximation and also discuss its accuracy.

2.1 An approximation to the marginal density

For the linear model setup, the vector of parameters in the model M_γ is given by $\boldsymbol{\theta}_\gamma = \left(\beta_0, \boldsymbol{\beta}_\gamma, \sigma^2, g\right)$. The marginal density $m_\gamma(\mathbf{y}_n)$ in (4) is

$$m_\gamma(\mathbf{y}_n) = \int f(\mathbf{y}_n|\beta_0, \boldsymbol{\beta}_\gamma, \sigma^2, M_\gamma)\pi(\boldsymbol{\beta}_\gamma|\beta_0, \sigma^2, g, M_\gamma)\pi(\beta_0, \sigma^2) \\ \times \pi(g)d(\beta_0, \boldsymbol{\beta}_\gamma, \sigma^2, g),$$

where $f(\mathbf{y}_n|\beta_0, \boldsymbol{\beta}_\gamma, \sigma^2, M_\gamma)$ is the p.d.f. of the n-variate normal distribution with mean $\beta_0\mathbf{1}+\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma$, dispersion matrix $\sigma^2 I$ and $\pi(\boldsymbol{\beta}_\gamma|\beta_0, \sigma^2, g, M_\gamma), \pi(\beta_0, \sigma^2), \pi(g)$ are as in Eqs. (5) and (6).

Integrating the integrand above with respect to β_0, $\boldsymbol{\beta}_\gamma$ and σ^2, we obtain a closed form expression which leads to

$$m_\gamma(\mathbf{y}_n) = \frac{\Gamma(n-1)/2}{\pi^{(n-1)/2}\sqrt{n}} \left(\mathcal{S}_y^2\right)^{-(n-1)/2} \int_0^\infty \frac{(1+g)^{(n-1-p(\gamma))/2}}{\left[1+g(1-R_\gamma^2)\right]^{(n-1)/2}} \pi(g)\mathrm{d}g, \quad (7)$$

where $\mathcal{S}_y^2 = \|\mathbf{y}_n - \bar{y}_n\mathbf{1}\|^2/n$, $(1-R_\gamma^2) = \mathbf{y}_n'(I-P_n(\gamma))\mathbf{y}_n/(n\mathcal{S}_y^2)$, $P_n(\gamma) = Z_{n\gamma}(Z_{n\gamma}'Z_{n\gamma})^{-1}Z_{n\gamma}'$, and $Z_{n\gamma} = (\mathbf{1}\ \ \mathbf{X}_\gamma)$.

Note that the marginal density of the intercept only model $M_\gamma = M_N$: $\mathbf{y}_n = \beta_0\mathbf{1} + \mathbf{e}_n$, which will be referred to as the *null* model, does not involve the hyperparameter g. It can be obtained as a special case of the marginal in expression (7) by putting $R_\gamma^2 = 0$ and $p(\gamma) = 0$.

For models $\gamma \in \mathcal{M}\backslash\{N\}$, the marginal density given by the proposed prior (6) does not have a closed form. However, we obtain an approximation of the marginal density with a closed form expression when the proposed mixture (6) is used in (7). When p is fixed, this approximation can achieve an accuracy of order $n^{-(2r-1)}$, $1 \le r < 2$ with probability tending to 1, which is at least as good as the accuracy of the Laplace approximation (see Kass and Raftery 1995). When p increases with n, the Laplace approximation may not be valid for the integral in (7) for commonly used priors on g, since the integrand may not be *Laplace regular* (see Kass et al. 1990). When $p = O(n^b)$, $0 < b < 1$ and $\nu = 1$ (or, when ν is free of n), the approximation is accurate with an error of the order $n^{-(2r-b-1)}$. When $\nu = p$ (or, ν is of the same order of n as p), the approximation still attains an accuracy of the order $n^{-(2r-1)}$.

We first state the assumptions under which the approximation holds.

Throughout this paper, $\mathbf{y}_n$ is modeled as (1) and we assume the following:

(A1) $\boldsymbol{\mu}_n{}'\boldsymbol{\mu}_n = O(n)$ as $n \to \infty$.
Assumption (A1) holds if the μ_i's are of comparable magnitude and they do not grow too fast compared to n. We next consider a general class of distributions of errors satisfying the following assumption:

(A2) The errors $e_1, e_2, \ldots, e_n$ are *i.i.d.* with a common density having mean 0 and finite fourth-order moment.

Assumption (A2) is satisfied by a wide variety of error-distributions including the normal, Laplace, logistic, generalized normal distributions, $t_{(k)}$ with $k > 4$ and all distributions with bounded supports.

We now state the result:

Theorem 1 *Consider the set of priors* (5) *and* (6) *with* ν *varying from 1 to* p. *Under assumptions* (*A1*) *and* (*A2*), *the marginal density in* (7) *satisfies the following:*

$$m_\gamma(\mathbf{y}_n) \le \widetilde{m}_\gamma(\mathbf{y}_n)\left(1 + \frac{p}{\nu n^{2r-1}} O(1)\right)$$
$$\text{and} \quad m_\gamma(\mathbf{y}_n) \ge \widetilde{m}_\gamma(\mathbf{y}_n)\left(1 + \frac{p}{\nu n^{2r-1}} O_p(1)\right),$$

uniformly in γ, *for any* $\gamma \in \mathcal{M}\backslash\{N\}$ *as* $n \to \infty$, *where*

$$\widetilde{m}_\gamma(\mathbf{y}_n) = \frac{\Gamma((n-1)/2)\Gamma((\nu + p(\gamma))/2)}{\sqrt{n}\Gamma(\nu/2)} \times \left(\pi \mathcal{S}_y^2\left(1 - R_\gamma^2\right)\right)^{-(n-1)/2} \left(\frac{n^{2r}\nu}{2}\right)^{-p(\gamma)/2}.$$

From Theorem 1, we obtain an approximation to the marginal density

$$m_\gamma(\mathbf{y}_n) \approx \widetilde{m}_\gamma(\mathbf{y}_n), \quad \text{as} \quad n \to \infty,$$

in the sense that the ratio of $m_\gamma(\mathbf{y}_n)$ and $\widetilde{m}_\gamma(\mathbf{y}_n)$ converges to 1 in probability. This approximation holds uniformly in γ since the $O(1)$ and $O_p(1)$ terms can be made free of γ. Further, this approximation holds for a large class of error distributions satisfying (A2). From Theorem 1 it follows that if $\nu = n^s$ for some $0 \le s \le b$, then the approximation is accurate upto an order $n^{-(2r+s-b-1)}$.

3 Model selection consistency under normality

In this section, we assume that the true mean $\boldsymbol{\mu}_n$ can be expressed as a linear combination of a subset of the p regressors and $\mathbf{e}_n \sim N(\mathbf{0}, \sigma^2 I)$. Let M_{γ_c}, with $\gamma_c \in \mathcal{M}$ be the true model. An ideal model selection procedure is likely to identify the true model with high probability in this framework. This property is termed as *model selection consistency* and is achieved if posterior probability of M_{γ_c}, given by (3), converges to one in probability, i.e.,

$$P(M_{\gamma_c}|\mathbf{y}_n) \xrightarrow{p} 1 \quad \text{as} \quad n \to \infty. \tag{8}$$

In Sect. 3.1, we provide sufficient conditions under which (8) holds when the proposed prior is used. In Sect. 3.2, we theoretically investigate model selection consistency of some other mixtures of g-priors in a sparse situation.

3.1 Consistency of the proposed mixture

We split the whole space of 2^p models of $\mathcal{M}$ into three mutually exclusive and exhaustive parts as follows:

$\mathcal{M}_1 = \{\gamma \in \mathcal{M} : M_\gamma \supset M_{\gamma_c}, \gamma \ne \gamma_c\}$, $\mathcal{M}_2 = \{\gamma \in \mathcal{M} : \gamma \notin \mathcal{M}_1, \gamma \ne \gamma_c\}$ and $\{\gamma_c\}$. We assume that

(A3) $\liminf_{n\to\infty} n^s \min_{\gamma\in\mathcal{M}_2} \boldsymbol{\mu}_n{}'(I - P_n(\gamma))\boldsymbol{\mu}_n/n > \delta$ for some constants $s \in [0, 1)$ and $\delta > 0$.

We impose a general restriction on model prior probability as follows:

(A4) For all $\gamma,\ \gamma' \in \mathcal{M}$ and some constant $c > 1$,

$$P(M_\gamma)/P(M_{\gamma'}) \le c^{|p(\gamma)-p(\gamma')|}.$$

A remark on each of these assumptions is made after the following theorem:

Theorem 2 *Let* $\mathbf{y}_n$ *be as in* (1) *with* $\boldsymbol{\mu}_n$ *satisfying* (*A1*) *and* $\mathbf{e}_n \sim N_n\left(\mathbf{0}, \sigma^2 I\right)$. *If* $p = O(n^b)$, *then the prior specification given by* (5) *and* (6) *is model selection consistent for* $0 < b < 2r/5$ *when* $\nu = 1$, *and for* $0 < b < r/2$ *when* $\nu = p$, *provided* (*A3*) *holds with* $s < (1 - b)/2$ *and* (*A4*) *holds.*

Remark 1 This result is different from those obtained by Wang and Sun (2014) and Xiang et al. (2016), who have shown Bayes factor consistency for growing number of regressors. Our result deals with model selection consistency considering all the 2^p models in the model space, and it is a much stronger result than pairwise consistency.

Remark 2 Assumption (A3) with $s = 0$ has been considered previously by many authors (see, e.g., Fernández et al. 2001; Liang et al. 2008; Bayarri et al. 2012). It is the key assumption for model selection consistency which ensures that the models can be differentiated. Here, we relax this assumption by allowing $s > 0$, which is a natural extension for the situation when p grows with n.

Remark 3 Assumption (A4) indicates that the true model may not have a prior probability arbitrarily close to zero, which is necessary to achieve consistency.

This assumption is quite general and includes many popular classes of model prior probabilities such as the *uniform* prior probability (i.e., $P(M_\gamma) = 1/2^p$) and the *Bernoulli* prior probability (i.e., $P(M_\gamma) = q^{p(\gamma)}(1-q)^{(p-p(\gamma))}$, with $0 < q < 1$). In the hierarchical beta Bernoulli setup (see Ley and Steel 2009), the parameter q of the Bernoulli prior is assigned a beta hyperprior. Assumption (A4) also holds for the hierarchical beta Bernoulli prior if both the parameters of the beta hyperprior are proportional to p.

Theorem 2 can also be proved if we replace (A4) by the following weaker assumption:

(A5) For all $\gamma,\ \gamma' \in \mathcal{M}$, and some constants $c_1,\ c_2 > 1, b_0 > 0$

$$P(M_\gamma)/P(M_{\gamma'}) \le (c_1 n^{b_0}) \vee c_2^{|p(\gamma)-p(\gamma')|},$$

where $a \vee b = \max\{a, b\}$.

To prove Theorem 2, we require $b_0 < r - 5b/2$ when $\nu = 1$ and $b_0 < r - 2b$ when $\nu = p$. However, for simplicity in presentation we work with (A4).

3.2 Some other mixtures of g-priors

The use of beta prime (beta of second kind) distribution in a mixture of g-priors is a common practice (see Liang et al. 2008; Maruyama and George 2011; Bayarri et al. 2012). One of the reasons for recommending this mixture is the possibility of obtaining marginal densities that are of closed form (see, e.g., Bayarri et al. 2012, Sect. 3.3). Therefore, we study the performance of this prior and identify the ranges of hyperparameters which lead to inconsistency when p grows with n. Let g follow a beta prime distribution with parameters λ_0 and λ_1; then

$$\pi(g) = \frac{\Gamma(\lambda_0+\lambda_1)}{\Gamma(\lambda_0)\Gamma(\lambda_1)} g^{\lambda_0-1}(1+g)^{-(\lambda_0+\lambda_1)},\ g>0,\ \lambda_0>0,\ \lambda_1>0. \tag{9}$$

The following theorem states that when $p = n^b$, $0 < b < 1$, then for some inappropriate specifications of the hyperparameters λ_0 and λ_1, the model selection rule given by the set of priors (5) and (9) becomes inconsistent under the null model, M_N.

Theorem 3 *Let* $\mathbf{y}_n = \beta_0\mathbf{1} + \mathbf{e}_n$ *where* $\mathbf{e}_n \sim N_n\left(\mathbf{0}, \sigma^2 I\right)$*, and assumption* (A4) *hold. If the number of regressors* $p = n^b$, $0 < b < 1$, *then the set of priors given by* (5) *and* (9) *with* $\lambda_1 > \epsilon$*, for some* $\epsilon > 0$ *free of* n*, is inconsistent provided* $(\lambda_0/\lambda_1) = O(n^{2b})$.

Remark 4 For the hyper-g prior, $\pi(g)$ is as in (9) with $\lambda_0 = 1$ and $\lambda_1 = (a/2-1)$ for some $a > 2$ free of n. Hence, it follows from the above theorem that the hyper-g prior is inconsistent for any $b > 0$. It is shown in Liang et al. (2008) that the hyper-g prior is not consistent under the null model even for fixed p. To remove this inconsistency the authors have considered the hyper-g/n prior by changing the scale of the hyper-g prior to n. The hyper-g/n prior is consistent when p is fixed, but fails to be consistent if $p = n^b$, for any $b > 0$. The proof is in the supplementary file.

Remark 5 For the generalized g-prior of Maruyama and George (2011) $\pi(g)$ is as in (9) with $\lambda_0 = A + 1$ and $\lambda_1 = B + 1$ where the authors recommend using $A = (n - p(\gamma) - 1)/2 - B$ and some fixed $B \in (-1, -1/2)$ for the case when $p+1 < n$. Hence, it follows from Theorem 3 that the generalized-g prior is inconsistent for this recommended setting if $b \geq 1/2$.

Remark 6 The robust prior of Bayarri et al. (2012) can also be expressed as a truncated scaled beta prime distribution as $(g+B)/(\rho_\gamma(n+B))-1 \sim \mathit{beta\ prime}(1, A)$ where $A > 0,\ B > 0,\ \rho_\gamma > B/(B+n)$. The recommended choices of hyperparameters are $A = 1/2,\ B = 1$ and $\rho_\gamma = 1/(1+p(\gamma))$. It has also been suggested that ρ_γ should be free of n. This makes choice of the parameter ρ_γ difficult when $p = n^b$, since in that case the recommended choice of ρ_γ involves n. We check with two choices of ρ_γ, a constant ρ_γ free of n (say, $\rho_\gamma = \rho$, for all $\gamma \in \mathcal{M}$) and $\rho_\gamma = 1/(1+p(\gamma))$. It has been shown in the supplementary file that when $p = n^b$, a necessary condition for consistency of the robust prior under the null model is $b < 1/2$, for both choices of ρ_γ.

Remark 7 It has been shown in Moreno et al. (2015) that Zellner–Siow prior is inconsistent for $b \geq 1/2$ when Bernoulli prior is used on the model space. The increment of scale from the order of n to n^{2r} makes the prior suitable for *'large p large n'* regime. For $r > 1$, clearly the prior is consistent for some $b > 1/2$ which is not the case with the other priors.

However, similar improvement is not expected from all the priors we mentioned above. For example, if we change the scale of the hyper-g/n prior from n to n^r for any $r > 1$, it still remains inconsistent when the null model is true, and $p = n^b$, for any $0 < b < 1$. The proof is similar in idea to the proof of the result stated in Remark 4.

4 Model selection consistency for general error distributions

In this section, we extend our results to situations where the distribution of regression errors belongs to a larger class satisfying assumption (A2). We investigate the strength of our model selection procedure when the distribution of the errors is not necessarily normal and the same model selection rule (based on the normal likelihood) is used. In a sense, we study robustness of our model selection rule for non-normal errors.

Unlike the case for normal errors, here we do not consider all the 2^p models. We make an additional assumption that the number of models of each dimension in the model space is bounded by a fixed number. Thus we assume the following:

(A6) If $p^*(d)$ denotes the number of models in the model space which are of dimension d $(d = 1, 2, \ldots, p + 1)$, then

$$\max_{1\leq d\leq p+1} p^*(d) \leq m$$

for some fixed positive integer m (free of n).

Condition (A6) holds, for example, when a class of *nested models* is considered. The class of all nested models can be expressed as

$$\mathcal{M}^* = \{\{\phi\}, \{1\}, \{1, 2\}, \ldots, \{1, 2, \ldots, p\}\}, \quad \text{with} \quad \mathcal{M}^* \subset \mathcal{M}.$$

Note that $\mathcal{M}^*$ has $p+1$ different models and (A6) holds with $m = 1$. When p increases with n, the number of models in $\mathcal{M}^*$ also increases. While the cardinality of $\mathcal{M}$ is exponential in p (i.e., 2^p), for $\mathcal{M}^*$ it is linear in p.

The situation with a model space like $\mathcal{M}^*$ may occur, for example, when we have information on relative importance of the regressors, and the regressors can be ordered accordingly. Model selection in nested models has been widely studied in the Bayesian paradigm when the error distribution is normal (see, e.g., Moreno 1997; Cui and George 2008; Wang and Sun 2014). Unlike these authors, who study Bayes factor consistency, we consider model selection consistency restricted to the space of models satisfying (A6) when p is increasing.

Condition (A6) also holds for other classes of nested and non-nested models. For example, it holds if the p regressors are divided into small groups (of bounded size) which can be arranged in deceasing order of importance and there are no preferences for regressors within each group.

We summarize our findings in the following theorem:

Theorem 4 *Let $\mathbf{y}_n$ be as in* (1) *with $\boldsymbol{\mu}_n$ satisfying* (*A1*), *and $\mathbf{e}_n$ satisfying* (*A2*). *Assume conditions* (*A4*) *and* (*A6*). *If $p = O(n^b)$, then the prior specification given by* (5) *and* (6) *with $\tau = n^r$ is model selection consistent for any $0 < b < 1$ provided* (*A3*) *holds with $s < (1-b)/2$.*

Remark 8 Theorem 4 can be proved if we replace assumption (A4) by assumption (A5) stated in Remark 3 with $b_0 < r - b/2$ for $\nu = 1$, and with $b_0 < r$ for $\nu = p$.

5 Consistency in the case of misspecified models for general error distributions

In Sects. 3 and 4, we have considered situations when the true mean $\boldsymbol{\mu}_n$ in (1) belongs to the span of $\{\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_n\}$. We will now consider a more general scenario where $\boldsymbol{\mu}_n$ is any n-dimensional vector, i.e., the true model does not necessarily belong to the model space $\mathcal{M}$. Several authors have studied related problems of linear model selection under this framework (see, e.g., Li 1987; Shao 1997; Chakrabarti and Ghosh 2006; Chakrabarti and Samanta 2008; Mukhopadhyay et al. 2015). The usual notion of model selection consistency cannot be used in this scenario. To validate a model selection rule we, therefore, adopt an alternative notion of consistency suited for this case as done in Mukhopadhyay et al. (2015).

Here, consistency of a model selection procedure refers to the property of choosing the model which is closest to the unknown true model among all candidate models in $\mathcal{M}$ (in an asymptotic sense). Let the true density of $\mathbf{y}_n$ be f. We consider the Kullback–Leibler divergence as the measure of distance between two probability distributions. We define the distance $\Delta_n(\gamma)$ between the true distribution f of $\mathbf{y}_n$, and the model M_γ as the minimum Kullback–Leibler distance between f and the density under M_γ, minimized with respect to the parameters $(\beta_0, \boldsymbol{\beta}_\gamma)$. An ideal model selection criterion should choose a model M_{γ^*} which is as close as possible to the true distribution, i.e., for which $\Delta_n(\gamma^*) = \min_{\gamma \in \mathcal{M}} \Delta_n(\gamma)$.

It can be easily verified that the Kullback–Leibler distance between the true distribution, given by the density function f, and the distribution of $N(\mathbf{1}\beta_0 + \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 I)$ under M_γ equals

$$\begin{aligned}&\int f(\mathbf{y}_n) \log f(\mathbf{y}_n) d\mathbf{y}_n + \frac{n}{2}(1 + \log \sigma^2)\\ &+\frac{1}{\sigma^2}(\boldsymbol{\mu}_n - \mathbf{1}\beta_0 - \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma)'(\boldsymbol{\mu}_n - \mathbf{1}\beta_0 - \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma).\end{aligned}$$

The distance $\Delta_n(\gamma)$ between the true model f and the model M_γ, obtained by minimizing the above with respect to $(\beta_0, \boldsymbol{\beta}_\gamma)$, is given by

$$\Delta_n(\gamma) = \int f(\mathbf{y}_n) \log f(\mathbf{y}_n)\, d\mathbf{y}_n + \frac{n}{2}\left(1 + \log \sigma^2\right) + D_n(\gamma), \tag{10}$$

where

$$D_n(\gamma) = \frac{1}{2\sigma^2} \boldsymbol{\mu}_n{}' (I - P_n(\gamma))\, \boldsymbol{\mu}_n. \tag{11}$$

Note that the first two terms of $\Delta_n(\gamma)$ in (10) do not involve γ and, therefore, $\operatorname{argmin}_\gamma \Delta_n(\gamma) = \operatorname{argmin}_\gamma D_n(\gamma)$. Thus, we refer a model selection procedure to be consistent if it satisfies

$$\frac{D_n(\hat{\gamma})}{\min_{\gamma \in \mathcal{M}} D_n(\gamma)} \xrightarrow{p} 1 \text{ as } n \to \infty, \tag{12}$$

where $M_{\hat{\gamma}}$ is the model chosen by the model selection rule.

While proving consistency of the model selection procedure based on the proposed prior in the above sense, we consider the general class of error distributions satisfying assumption (A2) and the whole set of 2^p models $\mathcal{M}$ for comparison. We make the following assumption, which is analogous to assumption (A3), by replacing $\mathcal{M}_2$ in (A3) by $\mathcal{M}$ as

(A3*) $\liminf_{n\to\infty} n^s \min_{\gamma \in \mathcal{M}} \boldsymbol{\mu}_n{}'(I - P_n(\gamma))\boldsymbol{\mu}_n/n > \delta$ for some constants $s \in [0, 1)$ and $\delta > 0$.

Note that in this case $\mathcal{M}_2 = \mathcal{M}$.

The result on consistency under misspecified models setup is stated as follows:

Theorem 5 *Let* $\mathbf{y}_n$ *be as in* (1) *with* $\boldsymbol{\mu}_n$ *being any real vector in* $\mathbb{R}^n$ *satisfying* (*A1*) *and* $\mathbf{e}_n$ *satisfying* (*A2*). *Suppose the assumption* (*A4*) *holds and* (*A3**) *holds with* $s < (1 - b)/2$. *If the number of regressors* $p = O(n^b)$, $0 < b < 1$, *then the set of priors* (5) *and* (6) *is consistent in the sense that* (12) *holds for any* $0 < b < 1$.

In Theorem 5, it is only assumed that $\mathbf{y}_n$ is the sum of two components, namely the regression mean $\boldsymbol{\mu}_n$ and a random error $\mathbf{e}_n$. Here, $\boldsymbol{\mu}_n$ is allowed to be arbitrary and $\mathbf{e}_n$ can follow any distribution satisfying assumption (A2). Thus, given the additive structure of $\mathbf{y}_n$ as stated in (1), consistency is obtained in a much general setting.

Remark 9 Theorem 5 can be proved if assumption (A4) on model prior probabilities is replaced by the following much weaker assumption:

$$P(M_\gamma)/P(M_{\gamma'}) \le e^{n^t} \vee c^{|p(\gamma)-p(\gamma')|}$$

for all $\gamma, \gamma' \in \mathcal{M}$, some $t < (1 - s)$ and $c > 1$.

6 Information consistency

The criterion of information consistency has been addressed by several authors including Jeffreys (1961), Berger and Pericchi (2001), Bayarri and García-Donato (2008), Liang et al. (2008) and Bayarri et al. (2012). While comparing the null model with any model M_γ, suppose that $\|\hat{\boldsymbol{\beta}}_\gamma\|^2 \to \infty$ holds (or equivalently, the usual F statistic goes to ∞) with both n and $p(\gamma)$ fixed, $\hat{\boldsymbol{\beta}}_\gamma$ being the least squares estimator of $\boldsymbol{\beta}_\gamma$. This is considered as a very strong evidence supporting the model M_γ, and it is expected that the Bayes factor of the model M_γ relative to the null model would go to ∞. The property that the Bayes factor goes to ∞ whenever $\|\hat{\boldsymbol{\beta}}_\gamma\|^2 \to \infty$ with fixed n and $p(\gamma)$, is termed as information consistency in Bayarri et al. (2012). However, this does not hold for Zellner's g-prior. For mixtures of g-priors, Liang et al. (2008, Theorem 2) give a sufficient condition which ensures information consistency. The following result gives conditions under which the mixture proposed in (6) is information consistent:

Result 1 *Consider the set of priors* (5). *The mixture of g-priors given by* (6) *is information consistent if $n \geq p + 1$ when $\nu = 1$ and if $n \geq 2p$ when $\nu = p$.*

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

7 Performance of the proposed prior on simulated datasets

In this section, we validate the performance of the proposed set of priors (5) and (6) using some simulated datasets. We present simulation results for model selection consistency under different simulation schemes.

Choices of hyperparameters There are two hyperparameters involved in the proposed mixture (6), namely $\tau = n^r$ and ν. For practical purposes we take $r = 1$. Arguments supporting this choice of r are provided in the next section. In each case, we consider our proposed prior with two choices of ν, *viz.*, $\nu = 1$ (*proposed I*) and $\nu = p$ (*proposed II*). From our results with these choices of ν, one should get some idea about the performance of an intermediate choice of ν.

Other methods We compare the proposed method of model selection with a wide variety of Bayesian methods, which includes four other mixtures of g-priors, namely the Zellner–Siow prior, the hyper-g/n prior, the generalized-g prior and the robust prior with recommended choices of hyperparameters. Among other methods, we consider the Bayesian shrinking and defusing prior (*BASAD*) of Narisetty and He (2014). Since our simulation setup is not restricted to sparse cases, we consider a less sparse specification of the parameter K, $K = 25$, which is used as an initial choice for the dimension of the true model in Narisetty and He (2014). We also consider the non-local prior (*piMOM*) of Johnson and Rossell (2012) and the methods based on Bayesian credible region (*BCR.joint* and *BCR.marg*) due to Bondell and Reich (2012) for comparison.

Simulation setup Our studies address the case when p increases with n and, therefore, we consider moderately large p compared to n. Three choices of n ($n = 50, 100, 150$) and two choices of p ($p + 1 = 30, 50$) for each value of n have been considered. Note that for any such choices of n and p, p can be regarded as of order n^b for some $0 < b < 1$. For each combination of (n, p), we generate n values of each of the p regressors $x_1, x_2, \ldots, x_p$ and this gives the full design matrix $\mathbf{X}_n$. We choose p numbers ξ_i, for $i = 1, \ldots, p$ and for each i generate the n values of the ith regressor x_i from an $N(\xi_i, 1)$ distribution. We assume that the n values of the ith regressor are coming from a homogeneous population. In order to fix a true model, we choose its dimension $p(\gamma_c)$. We then choose the $p(\gamma_c)$ non-zero regression coefficients β_i's, the intercept β_0 in the true model and also a value for the error variance σ^2. The $p(\gamma_c)$

columns of the design matrix $\mathbf{X}_{\gamma_c}$ for the true model are chosen at random from the p columns of $\mathbf{X}_n$.

Here, $(\xi_1, \ldots, \xi_p)$ is chosen as a random permutation of $(0.2, 0.4, \ldots, 0.2 \times p)$. We choose three different values of $p(\gamma_c)$; first a sparse model where $p(\gamma_c)$ is small ($p(\gamma_c) = 5$, *Scheme 1*), second a true model with half of the regressors active ($p(\gamma_c) = [p/2]$, *Scheme 2*) and finally, a $p(\gamma_c)$ which is close to the dimension of the full model (*Scheme 3*). The $p(\gamma_c)$ non-zero regression coefficients β_j's and the intercept β_0 in the true model are chosen randomly from the set $\{-0.2, 0.4, \ldots, (-1)^{p(\gamma_c)} \times 0.2 \times p(\gamma_c)\}$. The error variance σ^2 is chosen to be 1. After choosing the dimension $p(\gamma_c)$, the coefficients $(\beta_0, \boldsymbol{\beta}_{\gamma_c})$, σ^2 and the design matrix $\mathbf{X}_n$, we generate $\mathbf{e}_n$ from $N(\mathbf{0}, \sigma^2 I)$. The vector of observations $\mathbf{y}_n$ is obtained by adding $\boldsymbol{\mu}_n = \mathbf{1}\beta_0 + \mathbf{X}_{\gamma_c}\boldsymbol{\beta}_{\gamma_c}$ and $\mathbf{e}_n$. We repeat the data generation procedure 100 times and count the proportion of times each method selects the true model.

There are two issues to be mentioned here. First, for calculation of the marginal densities $\left(m_\gamma(\mathbf{y}_n)\right)$ for the mixtures of g priors, one needs to calculate the integral in (7), which is not of closed form. We use numerical integration (available in R software) to calculate this integral for all the mixtures. Second, since p is large, calculation of the marginal densities for all the 2^p candidate models in $\mathcal{M}$ is quite infeasible. Therefore, we use Gibbs sampling technique to run a Markov chain on $\mathcal{M}$ and select the highest visited model (discarding burn-in) as the one chosen by the model selection procedure.

Measures of comparison We make our comparison on the basis of the proportion of times the true model is visited (*Prop*) by a model selection procedure. Along with this proportion, we also compute the average difference (*Diff*) between the selected model and the true model. By difference we mean the cardinality of symmetric difference between the index set of covariates of the true model and that of the model selected by the corresponding model selection procedure.

Results obtained The simulation results for schemes 1, 2 and 3 are presented in Tables 1, 2 and 3, respectively. The results obtained in *Scheme 1* show how the methods of model selection perform when the true model is sparse. As theoretically shown in Sect. 3.2, most of the mixtures of g-priors fail to perform well when the true model is sparse. On the other hand, *BASAD* and methods based on Bayesian credible region, being designed for sparse situations, work reasonably well in this setup. The proposed methods and *piMOM* also yield competitive performance in this case. For the cases with $(p + 1) = 30$ and $n = 100, 150$, *proposed II* yields the best results for both the measures *Prop* and *Diff*, which is closely followed by *proposed I*. For the case with $n = 50$ and $(p + 1) = 30$, *BCR.marg* yields the best performance in terms of *Prop*. The performance of *BCR.marg* is closely followed by *proposed II* and *BCR.joint* in terms of *Prop*, and *BCR* is outperformed by *proposed II* in terms of *Diff*. *BASAD* and *piMOM* also perform moderately in this case. For $(p + 1) = 50$, the best performance is by *BASAD*, which is closely followed by *BCR* and *piMOM* for $n = 50$ and 100. While *Prop* is maximized by *BASAD* in each of these two cases, *piMOM* tends to select a model which is closest to the true model on an average. For the case with $(n, p + 1) = (150, 50)$, both *BASAD* and *proposed II* perform well compared to other methods, and *BCR* and *piMOM* also yield comparable performance. While *BASAD* maximizes *Prop*, *proposed II* maximizes *Prop* and minimizes *Diff* in this case.

Table 1 Proportion of times the true model is selected (*Prop*) and average difference (*Diff*) of the selected models from the true model for ten different competing methods of model selection in Scheme 1

$(n, p+1)$ →	(50, 30)		(100, 30)		(150, 30)		(50, 50)		(100, 50)		(150, 50)	
Methods ↓	*Prop*	*Diff*	*Prop*	*Diff*	*Prop*	*Diff*	*Prop*	*Diff*	*Prop*	*Diff*	*Prop*	*Diff*
Zellner–Siow	0.00	9.03	0.00	9.95	0.00	5.54	0.00	18.31	0.00	15.00	0.00	13.37
Hyper g/n	0.00	9.28	0.00	7.14	0.00	5.75	0.00	18.26	0.00	15.49	0.00	14.11
Robust	0.00	9.03	0.00	7.19	0.00	6.00	0.00	18.87	0.00	15.16	0.00	14.15
Generalized-g	0.00	8.88	0.00	7.26	0.00	5.39	0.00	17.18	0.00	15.73	0.00	13.86
piMOM	0.01	2.25	0.07	1.34	0.33	0.81	0.00	**1.92**	0.06	**1.16**	0.24	0.83
BASAD	0.00	10.52	0.02	4.66	0.14	2.66	**0.07**	4.13	**0.27**	1.52	**0.45**	1.00
BCR.joint+BIC	0.20	2.67	0.46	1.00	0.53	0.63	0.00	38.52	0.25	2.05	0.30	1.43
BCR.marg+BIC	**0.22**	2.64	0.45	1.02	0.58	0.59	0.00	40.57	0.26	2.00	0.37	1.26
Proposed I	0.00	4.02	0.26	1.14	0.80	0.23	0.00	14.87	0.00	7.84	0.00	6.31
Proposed II	0.20	**1.70**	**0.72**	**0.32**	**0.90**	**0.10**	0.00	7.92	0.01	2.36	**0.45**	**0**.78

The bold values signify the best performance among all the methods in each case
Here the true model has five active regressors

Table 2 *Prop* and *Diff* for ten different competing methods of model selection in *Scheme 2*

$(n, p+1)$ →	(50, 30)		(100, 30)		(150, 30)		(50, 50)		(100, 50)		(150, 50)	
Methods ↓	*Prop*	*Diff*	*Prop*	*Diff*	*Prop*	*Diff*	*Prop*	*Diff*	*Prop*	*Diff*	*Prop*	*Diff*
Zellner–Siow	0.03	2.76	0.31	0.92	0.50	0.64	0.00	4.85	0.00	3.82	0.00	2.25
Hyper g/n	0.03	2.74	0.39	0.76	0.45	0.69	0.00	5.06	0.00	3.75	0.02	2.15
Robust	0.01	2.85	0.36	0.82	0.55	0.57	0.00	4.64	0.00	3.68	0.00	2.18
Generalized-g	0.01	2.71	0.18	1.17	0.46	0.71	0.00	3.76	0.04	2.71	0.43	1.79
piMOM	0.01	2.37	0.05	1.31	0.30	0.81	0.00	4.77	0.00	2.57	0.19	0.93
BASAD	0.00	9.20	0.23	2.55	0.62	1.04	0.00	10.60	0.00	5.26	0.00	3.92
BCR.joint+BIC	0.12	2.77	0.55	0.73	0.56	0.56	0.00	24.65	0.17	2.47	0.32	1.30
BCR.marg+BIC	0.13	3.07	0.57	0.71	0.58	0.56	0.00	25.67	0.20	2.55	0.34	1.31
Proposed I	0.07	2.17	0.70	0.37	0.82	0.20	0.00	3.51	0.05	1.69	0.79	0.23
Proposed II	**0.34**	**1.29**	**0.76**	**0.29**	**0.89**	**0.11**	**0.01**	**2.65**	**0.53**	**1.39**	**0.83**	**0.20**

The bold values signify the best performance among all the methods in each case
Here, the true model has $[(p-1)/2]$ active regressors

In *Scheme 2*, where half of the covariates are active, all the methods yield competitive performance. For all choices of p and n, the best performance is by *proposed II* with respect to both the measures. For the choice $(n, p+1) = (50, 30)$, performance of *proposed II* is followed by the methods based on credible region, and for the other two cases with $(p+1) = 30$, it is followed by *proposed I*. For $(p+1) = 50$, apart from *proposed II*, performance of *BCR*, *proposed I* and *generalized-g* are better than other methods. Being designed for sparse cases only, *BASAD* fails to perform well in this scenario.

Table 3 *Prop* and *Diff* for ten different competing methods of model selection in *Scheme 3*

$(n, p+1)$ →	(50, 30)		(100, 30)		(150, 30)		(50, 50)		(100, 50)		(150, 50)	
Methods ↓	*Prop*	*Diff*	*Prop*	*Diff*	*Prop*	*Diff*	*Prop*	*Diff*	*Prop*	*Diff*	*Prop*	*Diff*
Zellner–Siow	0.18	1.62	0.71	0.29	0.94	0.06	0.00	20.94	0.12	1.20	0.48	0.55
Hyper g/n	0.19	1.58	0.71	0.29	0.94	0.06	0.00	22.10	0.12	1.19	0.52	0.51
Robust	0.18	1.61	0.71	0.29	**0.95**	**0.05**	0.00	21.64	0.11	1.22	0.51	0.51
Generalized-g	0.17	1.69	0.77	0.23	**0.95**	**0.05**	0.00	21.74	0.48	0.63	0.63	0.50
piMOM	0.00	3.37	0.01	1.60	0.17	1.18	0.00	11.8	0.01	1.76	0.09	1.12
BASAD	0.00	4.73	0.00	4.98	0.23	2.86	0.00	34.38	0.00	16.84	0.00	11.14
BCR.joint+BIC	0.12	2.13	0.34	0.91	0.40	0.67	0.00	10.33	0.18	1.62	0.23	1.20
BCR.marg+BIC	0.11	2.22	0.34	0.91	0.41	0.65	0.00	10.14	0.18	1.60	0.22	1.18
Proposed I	0.20	1.55	**0.79**	**0.21**	**0.95**	**0.05**	0.00	11.74	0.50	0.60	**0.85**	0.17
Proposed II	**0.25**	**1.34**	0.77	0.23	0.93	0.07	0.00	**9.16**	**0.54**	**0.57**	**0.85**	**0.16**

The bold values signify the best performance among all the methods in each case
Here, the true model has 25 active regressors when $p = 30$ and 40 active regressors when $p = 50$

In *Scheme 3*, 25 covariates are chosen to be active when $(p + 1) = 30$, while 40 covariates are active when $(p + 1) = 50$. Performance of the mixtures of g priors are better in general than other methods in this case. As expected *BASAD* fails to perform well in this scenario. Among the mixtures, results for *proposed I* and *proposed II* are better than others. Finally, unlike the mixtures, *piMOM* does not show any significant improvement in this case. The methods based on credible region also yield moderate performance in this scenario.

General remarks From results of these three schemes, it is evident that the methods based on the proposed mixture are more robust than the other methods, in the sense that irrespective of the dimension of the true model, they select the same with higher probability. The other mixtures of g-priors perform well only when the dimension of the true model is not too small. On the other hand, *BASAD* requires the true model to be sparse enough to perform well. Again, *piMOM* performs well when the regression coefficients (β_is) are significant enough. As it appears from the simulation results, *piMOM* mostly misses out the covariates with regression coefficients -0.2 and 0.4. Finally, the methods based on credible region yield a moderate performance in almost all the cases, although they show a little improvement as n grows from 50 to 150 in each case. As our purpose is to propose a method which works well in cases where no prior information is available on the dimension of the true model, it is evident from our numerical results that the method based on the proposed priors fulfill our goal satisfactorily.

Median probability model The proportion of visits to the true model may not always be a good measure to look at when there are large number of competing models (see García-Donato and Martínez-Beneito 2013). In that case, one may also look at the inclusion probabilities of the regressors and obtain the median probability model rather than the highest visited model. To demonstrate the performance of the proposed method further, we find the median probability model for each of the methods except *BCR.joint* and *BCR.marg*. The last two methods do not rely on sequential visits of Markov chains in the model space, and, therefore, it is not possible to calculate the inclusion probabilities. For each of the other methods, we also calculate the average difference between the true model and the median probability model. Tables 4, 5 and 6 in the supplementary file show the performance of different methods with respect to closeness of the median probability model to the true model. From the tables it can be seen that the performance of the proposed method is uniformly better than all other methods when the true model is of small dimension (Table 4), as well as when it is of large dimension (Table 6). For the case where the true model is of moderate dimension (Table 5) it is comparable with other methods and often yields the best performance.

8 Concluding remarks

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

The prior for $\boldsymbol{\beta}_\gamma$ arising from this mixture has a very thick tail which is recommended by Jeffreys (1961). Further, the set of priors (5) has the properties like *predictive matching* and *group invariance* as described in Bayarri et al. (2012) (see Results 2–4 of Bayarri et al. 2012 in this context). The authors have explicitly justified the adoption of the form (5) in a broader context.

Choice of the hyperparameters The hyperparameter, $\tau = n^r$, $1 \leq r < 2$, acts as a scale parameter of the proposed prior. Theoretical results suggest taking higher values of r to achieve a better rate of consistency. Also, bigger values of r result in better rates of approximation of the marginal density to a closed form expression. But very large values of r would make the prior too vague and may result in singularity problem. Thus we keep r as small as possible and recommend using $r = 1$. A slightly bigger value of r may also be considered. Note that the results with $r = 1$ are reasonably good in all the simulation schemes.

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

Acknowledgements We are thankful to the reviewers for their very useful comments and suggestions that helped us improve the paper.

Appendix

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

Auxiliary results

We first state some lemmas which will help in proving our main results. The proofs of these lemmas are given in the supplementary file.

Lemma 1 *If $\mathbf{y}_n = \boldsymbol{\mu}_n + \mathbf{e}_n$ with $\boldsymbol{\mu}_n$ satisfying assumption (A1) and $\mathbf{e}_n$ satisfying assumption (A2), then the following results hold as $n \to \infty$:*

(i) $\bar{e} = \sum_{i=1}^{n} e_i/n \xrightarrow{p} 0$,

(ii) $\sum_{i=1}^{n} \mu_i e_i/n \xrightarrow{p} 0$,

(iii) $\mathcal{S}_e^2 \xrightarrow{p} \sigma^2$ *where* $n\mathcal{S}_e^2 = \sum_{i=1}^{n} (e_i - \bar{e})^2$,

(iv) $\max_{\gamma \in \mathcal{M}} \mathbf{e}_n' P_n(\gamma)\mathbf{e}_n/n = O_p(p/n)$, *where $P_n(\gamma)$ is the projection matrix onto the span of* $[\mathbf{1}, \mathbf{X}_\gamma]$.

(v) $\max_{\gamma \in \mathcal{M}} |\boldsymbol{\mu}_n{}'(I - P_n(\gamma))\mathbf{e}_n|/n = O_p(\sqrt{p/n})$ *and*

(vi) $\mathbf{e}_n'(I - P_n(\gamma))\mathbf{e}_n/n \xrightarrow{p} \sigma^2$ *uniformly in* $\gamma \in \mathcal{M}$.

(vii) $\sum (y_i - \bar{y})^2/n = O_p(1)$.

Lemma 2 *Let R_γ^2 be such that $(1 - R_\gamma^2) = y_n'(I - P_n(\gamma))y_n/(n\mathcal{S}_y^2)$. Under assumptions (A1) and (A2), there exists $\delta_0 > 0$ such that $(1 - R_\gamma^2) > \delta_0$ with probability tending to 1 uniformly in $\gamma \in \mathcal{M}$.*

Lemma 3 *For any $\gamma \in \mathcal{M}$ and $\lambda > 0$, we have*

(a)

$$\int_0^\infty e^{-w} w^{(p(\gamma)+\lambda)/2-1} \left(1 + \frac{2w}{n^{2r}\lambda}\right)^{(n-p(\gamma)-1)/2} \mathrm{d}w \\ \leq \Gamma\left(\frac{p(\gamma)+\lambda}{2}\right) \sum_{j=0}^{\infty} \left(\frac{p+\lambda}{2n^{2r}+\lambda}\right)^j,$$

(b)

$$\int_0^\infty e^{-w} w^{(p(\gamma)+\lambda)/2-1} \left(1 - \frac{2w}{\delta_0 n^{2r}\lambda}\right)^{(n-p(\gamma)-1)/2} \mathrm{d}w \\ \geq \Gamma\left(\frac{p(\gamma)+\lambda}{2}\right) \left\{1 - \sum_{j=0}^{\infty} \left(\frac{p+\lambda}{2\delta_0 n^{2r}+\lambda}\right)^j\right\} \text{ for any } \delta_0 > 0.$$

Lemma 4 *Under the setup of Theorem 4, for any fixed $R > 0$, with probability tending to one*

$$\max_{\gamma\in\mathcal{M}_1} \frac{\mathbf{e}_n'(P_n(\gamma) - P_n(\gamma_c))\mathbf{e}_n}{\sigma^2(p(\gamma) - p(\gamma_c))} \le R\log p.$$

Lemma 5 *Under the assumptions of Theorem 2 the following results hold:*

(a) *For some $\delta_1 > 0$, with probability tending to one,*

$$\max_{\gamma\in\mathcal{M}_2} \left(\frac{1-R^2_{\gamma_c}}{1-R^2_{\gamma}}\right) \le \left(1+\frac{\delta_1}{n^s}\right)^{-1}.$$

(b) *For any $R > 2$ and any $0 < \epsilon < 1$, with probability tending to one uniformly in $\gamma \in \mathcal{M}_1$, we have*

$$\frac{1-R^2_{\gamma_c}}{1-R^2_{\gamma}} \le p^{R(p(\gamma)-p(\gamma_c))/(n(1-\epsilon)^2)}.$$

(c) *For any $R > 2$ and any $0 < \epsilon < 1$, with probability tending to one uniformly in $\gamma \in \mathcal{M}$, we get*

$$\frac{nR^2_{\gamma}}{1-R^2_{\gamma}} \le \frac{Rp(\gamma)\log p}{1-\epsilon}.$$

Proof of Theorem 1

Using (6) and (7), we write

$$m_\gamma(\mathbf{y}_n) = \mathcal{C}_{1,y,n}\mathcal{I}, \tag{13}$$

where

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

and $\mathcal{I} = \int_0^\infty e^{-\tau^2\nu/(2g)} g^{-(1+\nu/2)}(1+g)^{(n-p(\gamma)-1)/2}\{1+(1-R^2_\gamma)g\}^{-(n-1)/2}\mathrm{d}g$.

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

$$\mathcal{I} = \mathcal{C}_{2,y,n}\int_0^\infty e^{-w}w^{\nu/2-1}\left(1+\frac{\tau^2\nu}{2w}\right)^{(n-p(\gamma)-1)/2}\left\{1+(1-R^2_\gamma)\frac{\tau^2\nu}{2w}\right\}^{-(n-1)/2}\mathrm{d}w, \tag{14}$$

where $\mathcal{C}_{2,y,n} = (\tau^2\nu/2)^{-\nu/2}$. Next we use the fact that for any $w > 0$

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

Use of this inequality along with multiplication and division by $(\tau^2\nu/(2w))^{(n-p(\gamma)-1)/2}$ in right-hand side of (14) gives

$$\mathcal{I} \le \mathcal{C}_{3,y,n} \int_0^\infty e^{-w} w^{(p(\gamma)+\nu)/2-1} \left(1 + \frac{2w}{\tau^2\nu}\right)^{(n-p(\gamma)-1)/2} \mathrm{d}w,$$

where $\mathcal{C}_{3,y,n} = \mathcal{C}_{2,y,n}(\tau^2\nu/2)^{-p(\gamma)/2}(1 - R_\gamma^2)^{-(n-1)/2}$. It follows from part (a) of Lemma 3 and by putting $\tau = n^r$ that

$$\mathcal{I} \le \mathcal{C}_{3,y,n}\Gamma\left(\frac{p(\gamma)+\nu}{2}\right)\sum_{j=0}^{\infty}\left(\frac{p+\nu}{2n^{2r}+\nu}\right)^j.$$

From the above inequality it follows that

$$\mathcal{I} \le \mathcal{C}_{3,y,n}\Gamma\left(\frac{p(\gamma)+\nu}{2}\right)\left(1 + \frac{p}{\nu n^{2r-1}}O(1)\right). \tag{15}$$

Next, we assign a bound on $\mathcal{I}$ from other direction and show that the difference between the two bounds is small. For this, we move back to (14) and use the inequality $(1+(\tau^2\nu)/(2w)) > (\tau^2\nu)/(2w)$ along with a multiplication and division by the factor $((1 - R_\gamma^2)(\tau^2\nu)/(2w))^{(n-1)/2}$ in the integrand of (14). It then follows that

$$\begin{aligned}\mathcal{I} &\ge \mathcal{C}_{3,y,n}\int_0^\infty w^{(\nu+p(\gamma))/2-1}e^{-w}\left(1 - \frac{1}{1 + (1 - R_\gamma^2)\tau^2\nu/(2w)}\right)^{(n-1)/2} \mathrm{d}w \\ &\ge \mathcal{C}_{3,y,n}\int_0^\infty w^{(\nu+p(\gamma))/2-1}e^{-w}\left(1 - \frac{2w}{(1 - R_\gamma^2)\tau^2\nu}\right)^{(n-1)/2} \mathrm{d}w,\end{aligned}$$

where $\mathcal{C}_{3,y,n}$ is the same as in (15). From Lemma 2, we know that for all $\gamma \in \mathcal{M}$, $(1 - R_\gamma^2)$ has a fixed positive lower bound δ_0 with probability tending to 1. This, together with part (b) of Lemma 3 and the fact that $\tau = n^r$ gives

$$\mathcal{I} \ge \mathcal{C}_{3,y,n}\Gamma\left(\frac{p(\gamma)+\nu}{2}\right)\left\{1 - \sum_{j=0}^{\infty}\left(\frac{p+\nu}{2\delta_0 n^{2r}+\nu}\right)^j\right\},$$

for some $\delta_0 > 0$ with probability tending to 1. We then have

$$\mathcal{I} \ge \mathcal{C}_{3,y,n}\Gamma\left(\frac{p(\gamma)+\nu}{2}\right)\left(1 + \frac{p}{\nu n^{2r-1}}Op(1)\right). \tag{16}$$

The theorem now follows from (13), (15) and (16). □

Proof of Theorem 2

We need to show that

$$\sum_{\gamma\in\mathcal{M}_i}\frac{P(M_\gamma)}{P(M_{\gamma_c})}\frac{m_\gamma(\mathbf{y}_n)}{m_{\gamma_c}(\mathbf{y}_n)}\xrightarrow{p}0, \tag{17}$$

as $n\to\infty$ for $i=1,2$. We prove for the cases $M_{\gamma_c}\neq M_N$ and $M_{\gamma_c}=M_N$ and also prove (17) for $i=1$ and $i=2$, separately. Considering the fact that when $M_{\gamma_c}=M_N$, then $\mathcal{M}=\mathcal{M}_1\cup\{\gamma_c\}$, we split the proof into three parts.

Case I $M_{\gamma_c}\neq M_N$ and $\gamma\in\mathcal{M}_2$. From Theorem 1, we have

$$\frac{m_\gamma(\mathbf{y}_n)}{m_{\gamma_c}(\mathbf{y}_n)}\leq\left(\frac{n^{2r}\nu}{2}\right)^{-(p(\gamma)-p(\gamma_c))/2}\left(\frac{1-R^2_{\gamma_c}}{1-R^2_\gamma}\right)^{(n-1)/2}$$
$$\frac{\Gamma\{(\nu+p(\gamma))/2\}}{\Gamma\{(\nu+p(\gamma_c))/2\}}\frac{\left\{1+pO(1)/(\nu n^{2r-1})\right\}}{\left\{1+pO_p(1)/(\nu n^{2r-1})\right\}}, \tag{18}$$

where the terms $O(1)$ and $O_p(1)$ are free of γ. To evaluate the third term of right-hand side of the above expression, we make use of the result

$$\{x/(x+s)\}^s\leq\Gamma(x+s)/\left(x^s\Gamma x\right)\leq 1$$

for $0<s<1$ and $x>0$ from Wendel (1948). It can be shown that

$$\frac{\Gamma\{(\nu+p(\gamma))/2\}}{\Gamma\{(\nu+p(\gamma_c))/2\}}\leq\left(\frac{\nu+p}{2}\right)^{|p(\gamma)-p(\gamma_c)|/2}. \tag{19}$$

Hence from inequalities (18) and (19), and part (a) of Lemma 5 we have

$$\max_{\gamma\in\mathcal{M}_2}\frac{m_\gamma(\mathbf{y}_n)}{m_{\gamma_c}(\mathbf{y}_n)}\leq\left(\frac{n^{2r}\nu}{2}\right)^{p/2}\left(1+\frac{\delta_1}{n^s}\right)^{-(n-1)/2}$$
$$\times\left(\frac{\nu+p}{2}\right)^{p/2}\frac{\left\{1+pO(1)/(\nu n^{2r-1})\right\}}{\left\{1+pO_p(1)/(\nu n^{2r-1})\right\}}.$$

Using assumption (A4), we also get an upper bound of the ratio of prior probabilities of the models. Therefore, we get

$$\sum_{\gamma\in\mathcal{M}_2}\frac{P(M_\gamma)}{P(M_{\gamma_c})}\frac{m_\gamma(\mathbf{y}_n)}{m_{\gamma_c}(\mathbf{y}_n)}\leq c^p\,2^{p+1}\left(\frac{n^{2r}\nu}{2}\right)^{p/2}$$
$$\times\left(1+\frac{\delta_1}{n^s}\right)^{-(n-1)/2}\left(\frac{\nu+p}{2}\right)^{p/2}, \tag{20}$$

with probability tending to one. It is now easy to check that the above quantity goes to 0 as $n \to \infty$ when $s < (1-b)/2$.

Case II $M_{\gamma_c} \neq M_N$ *and* $\gamma \in \mathcal{M}_1$. Consider each term of the right-hand side of the inequality in (18). From part (b) of Lemma 5, for any $R > 2$ and any $0 < \epsilon < 1$, with probability tending to one uniformly in $\gamma \in \mathcal{M}_1$,

$$\frac{1-R^2_{\gamma_c}}{1-R^2_{\gamma}} \leq p^{R(p(\gamma)-p(\gamma_c))/(n(1-\epsilon)^2)}. \tag{21}$$

Combining (18), (19), (21) and using assumption (A4), we have

$$\begin{aligned}\sum_{\gamma\in\mathcal{M}_1} \frac{P(M_\gamma)}{P(M_{\gamma_c})}\frac{m_\gamma(\mathbf{y}_n)}{m_{\gamma_c}(\mathbf{y}_n)} &\leq \sum_{\gamma\in\mathcal{M}_1}\left(\frac{c^2\,(\nu+p)p^{R/(1-\epsilon)^2}}{n^{2r}\nu}\right)^{(p(\gamma)-p(\gamma_c))/2}\\ &\leq \sum_{q=1}^{p-p(\gamma_c)}\binom{p-p(\gamma_c)}{q}\left(\frac{c\sqrt{\nu+p}\;\;p^{R/\{2(1-\epsilon)^2\}}}{n^r\sqrt{\nu}}\right)^q\\ &\leq \left\{\left(1+\frac{c\sqrt{\nu+p}\;\;p^{R/\{2(1-\epsilon)^2\}}}{n^r\sqrt{\nu}}\right)^{(p-p(\gamma_c))}-1\right\}. \end{aligned}\tag{22}$$

If the first term within the curly brackets converges to 1, then the above expression converges to 0 as $n \to \infty$. If $\nu = 1$ and $p = O(n^b)$, then the first term, for some positive constants c' and k, any $R > 2$ and any $\epsilon > 0$, is less than $(1 + c'\,n^{-[r-(R+1)b/\{2(1-\epsilon)^2\}]})^{kn^b}$. Also, when $\nu = p$ we have this term to be less than $(1 + c''\,n^{-[r-Rb/\{2(1-\epsilon)^2\}]})^{kn^b}$ for some positive constants c'' and k, any $R > 2$ and any $\epsilon > 0$. Letting $R \downarrow 2$ and $\epsilon \downarrow 0$, the last expression in (22) converges to 0 if $b < 2r/5$ when $\nu = 1$ and if $b < r/2$ when $\nu = p$.

Case III $M_{\gamma_c} = M_N$. When the null model is true, the Bayes factor of any model with respect to the null model is given by

$$\frac{m_\gamma(\mathbf{y}_n)}{m_N(\mathbf{y}_n)} = \int_0^\infty (1+g)^{(n-p(\gamma)-1)/2}\left\{1+g(1-R^2_\gamma)\right\}^{-(n-1)/2}\pi(g)\mathrm{d}g, \tag{23}$$

where $\pi(g)$ is as in (6). Now, we have

$$\begin{aligned}\left[\frac{1+g}{1+(1-R^2_\gamma)g}\right]^{(n-1)/2} &= \exp\left[\left(\frac{n-1}{2}\right)\left\{\ln(1+g)-\ln(1+(1-R^2_\gamma)g)\right\}\right]\\ &= \exp\left[\left(\frac{n-1}{2}\right)\left\{\ln(1+g)-\ln(1+g)+\frac{R^2_\gamma g}{1+g^*}\right\}\right],\end{aligned}$$

where $g^* \in [(1-R_\gamma^2)g, g]$. The above quantity is less than

$$\exp\left[\left(\frac{n-1}{2}\right)\left\{\frac{R_\gamma^2}{(1-R_\gamma^2)+1/g}\right\}\right] \le \exp\left[\left(\frac{n-1}{2}\right)\left(\frac{R_\gamma^2}{1-R_\gamma^2}\right)\right].$$

Then by part (c) of Lemma 5, for any $R > 2$ and any $0 < \epsilon < 1$,

$$\left[\frac{1+g}{1+(1-R_\gamma^2)g}\right]^{(n-1)/2} \le p^{Rp(\gamma)/(2(1-\epsilon))}$$

with probability tending to one uniformly in $\gamma \in \mathcal{M}$. From (23), we also have

$$\frac{m_\gamma(\mathbf{y}_n)}{m_N(\mathbf{y}_n)} \le p^{Rp(\gamma)/(2(1-\epsilon))} \int_0^\infty (1+g)^{-p(\gamma)/2}\pi(g)\mathrm{d}g. \tag{24}$$

Now, with $\pi(g)$ as given in (6)

$$I = \int_0^\infty (1+g)^{-p(\gamma)/2}\pi(g)\mathrm{d}g \le \frac{\left(\tau^2\nu/2\right)^{\nu/2}}{\Gamma(\nu/2)} \int_0^\infty e^{-\tau^2\nu/(2g)} g^{-(p(\gamma)+\nu)/2-1}\mathrm{d}g$$

by the fact that $(1+g)^{-1} < g^{-1}$. We then have

$$I = \begin{cases} \left(\tau^2/2\right)^{-p(\gamma)/2} \Gamma\{(p(\gamma)+1)/2\}/\Gamma(1/2) & \text{for } \nu = 1, \\ \left(\tau^2 p/2\right)^{-p(\gamma)/2} \Gamma\{(p(\gamma)+p)/2\}/\Gamma(p/2) & \text{for } \nu = p. \end{cases}$$

To evaluate these terms, we again use the results of Wendel (1948) stated before in Case I. Little algebra shows

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

By assumption (A4) and putting $\tau = n^r$, it follows from (24) that

$$\begin{aligned} \sum_{\gamma\in\mathcal{M}\setminus\{\gamma_c\}} \frac{P(M_\gamma)}{P(M_{\gamma_c})}\frac{m_\gamma(\mathbf{y}_n)}{m_{\gamma_c}(\mathbf{y}_n)} &< \begin{cases} \sum\limits_{\gamma\in\mathcal{M}\setminus\{\gamma_c\}} \left(c^2 p^{1+R/(1-\epsilon)}/n^{2r}\right)^{p(\gamma)/2}/\sqrt{\pi} & \text{for } \nu = 1, \\ \sum\limits_{\gamma\in\mathcal{M}\setminus\{\gamma_c\}} \left(2c^2 p^{R/(1-\epsilon)}/n^{2r}\right)^{p(\gamma)/2} & \text{for } \nu = p. \end{cases} \\ &= \begin{cases} \left(1 + c\, p^{\{1+R/(1-\epsilon)\}/2}/n^r\right)^p - 1 & \text{for } \nu = 1, \\ \left(1 + \sqrt{2}\, c\, p^{R/\{2(1-\epsilon)\}}/n^r\right)^p - 1 & \text{for } \nu = p, \end{cases} \end{aligned} \tag{25}$$

for any $R > 2$ and any $\epsilon > 0$.

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

Proof of Theorem 3

Let $M_{\gamma_c} = M_N$. By assumption (A4), $P(M_\gamma)/P(M_{\gamma_c}) \geq c^{-p(\gamma)}$ for all γ and, therefore, from (23) we have

$$\sum_{\gamma \in \mathcal{M}\setminus\{\gamma_c\}} \frac{P(M_\gamma)}{P(M_{\gamma_c})} \frac{m_\gamma(\mathbf{y}_n)}{m_{\gamma_c}(\mathbf{y}_n)} \geq \sum_{\gamma} \int_0^\infty c^{-p(\gamma)}(1+g)^{-p(\gamma)/2}\pi(g)\mathrm{d}g, \qquad (26)$$

where $\pi(g)$ is given by (9). Putting the prior we get the R.H.S. of (26) as

$$\sum_{\gamma \in \mathcal{M}\setminus\{\gamma_c\}} c^{-p(\gamma)} \frac{\Gamma(\lambda_0+\lambda_1)\Gamma(\lambda_1+p(\gamma)/2)}{\Gamma\lambda_1\Gamma(\lambda_0+\lambda_1+p(\gamma)/2)}.$$

Using the inequality of Wendel (1948) stated before in the proof of Theorem 2 and the fact that $\lambda_1 > \epsilon$ for some $\epsilon > 0$ free of n, it can be shown that for some constant $C' > 0$, the above expression is bigger than

$$C' \sum_{\gamma \in \mathcal{M}\setminus\{\gamma_c\}} \left(\frac{c^{-2}\lambda_1}{\lambda_0+\lambda_1}\right)^{p(\gamma)/2} = C'\left\{\left(1+c^{-1}\sqrt{\frac{\lambda_1}{\lambda_0+\lambda_1}}\right)^p - 1\right\}.$$

Thus if $p = n^b$, the R.H.S. of (26) does not go to 0 if $\lambda_0/\lambda_1 = O(n^{2b})$. □

Proof of Theorem 4

Under the setup of Theorem 4, we do not consider all the 2^p models. We denote the reduced classes of models corresponding to $\mathcal{M}$, $\mathcal{M}_1$, $\mathcal{M}_2$ by $\mathcal{M}^*$, $\mathcal{M}_1^*$, $\mathcal{M}_2^*$, respectively. We proceed as in the proof of Theorem 2, and prove (17) with $\mathcal{M}_i$ replaced by $\mathcal{M}_i^*$, for $i = 1, 2$. We consider separately the cases when the true model is null and when it is non-null.

Case I $M_{\gamma_c} \neq M_N$. First, we consider the case when $i = 2$. It can easily be seen that the part concerning the model space $\mathcal{M}_2$ in Theorem 2 (i.e., Case I of the theorem) is proved without using the assumption of normality. Since $\mathcal{M}_2^*$ is a proper subset of $\mathcal{M}_2$, the same proof works here.

Next consider (17) with $\mathcal{M}_i$ replaced by $\mathcal{M}_i^*$ and $i = 1$. Here we use Lemma 4 which will imply that part (b) of Lemma 5 holds for the model space $\mathcal{M}^*$ satisfying

(A6) a well. From (18), (19) and (21) we have, for any $R > 0$ and any $\epsilon > 0$,

$$\begin{aligned}\sum_{\gamma\in\mathcal{M}_1^*}\frac{P(M_\gamma)}{P(M_{\gamma_c})}\frac{m_\gamma(\mathbf{y}_n)}{m_{\gamma_c}(\mathbf{y}_n)} &\le \sum_{\gamma\in\mathcal{M}_1^*}\left(\frac{c^2\ (\nu+p)p^{R/(1-\epsilon)^2}}{n^{2r}\nu}\right)^{(p(\gamma)-p(\gamma_c))/2}\\ &\le m\sum_{q=1}^{p-p(\gamma_c)}\left(\frac{c\ \sqrt{\nu+p}\ \ p^{R/\{2(1-\epsilon)^2\}}}{n^r\sqrt{\nu}}\right)^q\\ &\le \frac{m\ c\ \sqrt{\nu+p}\ \ p^{R/\{2(1-\epsilon)^2\}}}{n^r\sqrt{\nu}\left(1-c\ \sqrt{\nu+p}\ \ p^{R/\{2(1-\epsilon)^2\}}/(n^r\sqrt{\nu})\right)},\end{aligned}$$

where m is as in condition (A6). For suitably chosen R and ϵ, it can be easily seen that the last expression converges to 0 for any $0 < b < 1$.

Case II $M_{\gamma_c} = M_N$. We proceed as in Theorem 2. Observe that using Lemma 4, we obtain an inequality similar to (25) with $\mathcal{M}$ replaced by $\mathcal{M}^*$. Thus, we get

$$\begin{aligned}\sum_{\gamma\in(\mathcal{M}^*\setminus\{\gamma_c\})}\frac{P(M_\gamma)}{P(M_{\gamma_c})}\frac{m_\gamma(\mathbf{y}_n)}{m_{\gamma_c}(\mathbf{y}_n)} &< \begin{cases}\displaystyle\sum_{\gamma\in(\mathcal{M}^*\setminus\{\gamma_c\})}\pi^{-1/2}\left(c^2\ p^{1+R/(1-\epsilon)}/n^{2r}\right)^{p(\gamma)/2} & \text{for } \nu=1,\\ \displaystyle\sum_{\gamma\in(\mathcal{M}^*\setminus\{\gamma_c\})}\left(2\,c^2\ p^{R/(1-\epsilon)}/n^{2r}\right)^{p(\gamma)/2} & \text{for } \nu=p.\end{cases}\\ &< \begin{cases} m\left\{c\ p^{1/2+R/\{2(1-\epsilon)\}}/\left(n^r-c\ p^{1/2+R/\{2(1-\epsilon)\}}\right)\right\} & \text{for } \nu=1,\\ m\ \left\{\sqrt{2}\,c\ p^{R/\{2(1-\epsilon)\}}/\left(n^r-\sqrt{2}\,c\ p^{R/\{2(1-\epsilon)\}}\right)\right\} & \text{for } \nu=p.\end{cases}\end{aligned}$$

Here m is as in Condition (A6). One can choose $R > 0$ and $\epsilon > 0$ suitably and show that the above quantities go to 0 for any $0 < b < 1$. □

Proof of Theorem 5

Our model selection rule is to choose a model $\hat{\gamma}$ in the model space $\mathcal{M}$, which maximizes $P(M_\gamma)m_\gamma(\mathbf{y}_n)$ with respect to γ. Now, from Theorem 1, this is equivalent to maximizing

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

where $|\varepsilon_n(\gamma)| = pO_p(1)/(n^{2r-1}\nu)$ uniformly in γ. We omit the other terms involved in the approximation given in Theorem 1, as those are free of γ. Maximizing the above is equivalent to minimizing

$$\left[P(M_\gamma)\Gamma\left(\frac{\nu+p(\gamma)}{2}\right)(1+\varepsilon_n(\gamma))\right]^{-2/(n-1)}\left(\frac{n^{2r}\nu}{2}\right)^{p(\gamma)/(n-1)} n\mathcal{S}_y^2(1-R_\gamma^2) \tag{27}$$

with respect to γ. From (11), we have $n\mathcal{S}_y^2(1-R_\gamma^2) = C_n + 2\sigma^2 D_n(\gamma)(1+\xi_n(\gamma))$, where $C_n = \mathbf{e}_n'\mathbf{e}_n$ and $\xi_n(\gamma) = \{2\boldsymbol{\mu}_n{}'(I-P_n(\gamma))\mathbf{e}_n - \mathbf{e}_n' P_n(\gamma)\mathbf{e}_n\}/(2\sigma^2 D_n(\gamma))$. If $M_{\hat{\gamma}}$ is the model for which (27) is minimized, then we get

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

where

$$b_n(\gamma) = \left\{\left(\frac{P(M_{\hat{\gamma}})}{P(M_\gamma)}\right)^2 \left(\frac{\Gamma\{(n+p(\hat{\gamma}))/2\}}{\Gamma\{(n+p(\gamma))/2\}}\right)^2 \left(\frac{n^2\nu}{2}\right)^{(p(\hat{\gamma}-p(\gamma))} \left(\frac{1+\varepsilon_n(\hat{\gamma})}{1+\varepsilon_n(\gamma)}\right)^2\right\}^{1/(n-1)}. \tag{28}$$

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

$$1 \leq \frac{D_n(\hat{\gamma})}{\min_\gamma D_n(\gamma)} \leq \frac{C_n}{2n\sigma^2(1-\xi_n)} \times \max_\gamma \frac{n(b_n(\gamma)-1)}{D_n(\gamma)} + \frac{(1+\xi_n)}{(1-\xi_n)} \times \max_\gamma b_n(\gamma). \tag{29}$$

The rest of the proof will follow from the facts stated below:

$$C_n/n \xrightarrow{p} \sigma^2, \tag{30}$$

$$\xi_n \xrightarrow{p} 0, \tag{31}$$

$$\max_\gamma \frac{n(b_n(\gamma)-1)}{D_n(\gamma)} \xrightarrow{p} 0, \tag{32}$$

$$\text{and} \qquad \max_\gamma b_n(\gamma) \xrightarrow{p} 1. \tag{33}$$

The proof of (30) is straightforward. To prove (31), we first note that

$$\xi_n \leq \frac{2\max_\gamma \boldsymbol{\mu}_n{}'(I-P_n(\gamma))\mathbf{e}_n/n - \min_\gamma \mathbf{e}_n' P_n(\gamma)\mathbf{e}_n/n}{2\min_\gamma \sigma^2 D_n(\gamma)/n} \leq \frac{O_p(\sqrt{p/n})}{\delta/n^s}.$$

This follows from parts (iv), (v) of Lemma 1, and assumption (A3*). Clearly if $s < (1-b)/2$, (31) holds.

Next, we prove (33). We show that $\log(\max_\gamma b_n(\gamma)) = \max_\gamma \log(b_n(\gamma)) \xrightarrow{p} 0$. From (28) and (19) and by assumption (A4), we have

$$\max_\gamma \log b_n(\gamma) \leq \frac{2}{n-1}\left\{\log\left(\max_\gamma \frac{P(M_{\hat{\gamma}})}{P(M_\gamma)}\right) + \log\left(\max_\gamma \frac{\Gamma\{(n+p(\hat{\gamma}))/2\}}{\Gamma\{(n+p(\gamma))/2\}}\right)\right.$$
$$\left. + \log\left(\frac{1+\varepsilon_n(\hat{\gamma})}{1-\max_\gamma|\varepsilon_n(\gamma)|}\right)\right\} + \max_\gamma \frac{p(\hat{\gamma})-p(\gamma)}{n-1}\log\left(\frac{n^{2r}\nu}{2}\right)$$

$$\leq \frac{2}{n-1}\left\{p\log c + \frac{p}{2}\log\left(\frac{p+\nu}{2}\right)\right.$$
$$\left. + \log\left(\frac{1+pO_p(1)/(n^{2r-1}\nu)}{1-pO_p(1)/(n^{2r-1}\nu)}\right) + \frac{p}{2}\log\left(\frac{n^{2r}\nu}{2}\right)\right\}.$$

We note that $\log(\max_\gamma b_{n\gamma}) \leq \max_\gamma \log b_{n\gamma}$. It is now easy to show that when $p = O(n^b)$ with $0 < b < 1$, the above expression is $O_p\left(n^{-(1-b)}\log n\right)$ and converges to 0 with probability tending to 1. Hence, (33) holds.

Finally, we prove (32). By mean value theorem, for some $z > 0$, $(e^z - 1) = ze^{z^*} < ze^z$, where $z^* \in (0, z)$. Replacing z by $\log b_n(\gamma)$, we get

$$\max_\gamma(b_n(\gamma) - 1) \leq \max_\gamma \log b_n(\gamma)\exp\{\max_\gamma \log b_n(\gamma)\}.$$

By assumption (A3*), we have

$$\max_\gamma \frac{n(b_n(\gamma)-1)}{D_n(\gamma)} \leq \frac{\max_\gamma(b_n(\gamma)-1)}{\min_\gamma D_n(\gamma)/n}$$
$$\leq n^s O_p\left(n^{-(1-b)}\log n\right)\exp\left\{O_p\left(n^{-(1-b)}\log n\right)\right\},$$

which goes to 0 with probability tending to 1, as $n \to \infty$. □

References

Bartlett M (1957) A comment on D.V. Lindley's statistical paradox. Biometrika 44(3–4):533–534

Bayarri MJ, García-Donato G (2008) Generalization of Jeffreys divergence-based priors for Bayesian hypothesis testing. J R Stat Soc Ser B Stat Methodol 70(5):981–1003. doi:10.1111/j.1467-9868.2008.00667.x

Bayarri MJ, Berger JO, Forte A, García-Donato G (2012) Criteria for Bayesian model choice with application to variable selection. Ann Stat 40(3):1550–1577. doi:10.1214/12-AOS1013

Berger JO, Pericchi LR (2001) Objective Bayesian methods for model selection: introduction and comparison. In: Model selection, IMS Lecture Notes Monogr. Ser., vol 38, Inst. Math. Statist., Beachwood, OH, pp 135–207. doi:10.1214/lnms/1215540968, with discussion by J. K. Ghosh, Tapas Samanta, and Fulvio De Santis, and a rejoinder by the authors

Bondell HD, Reich BJ (2012) Consistent high-dimensional Bayesian variable selection via penalized credible regions. J Am Stat Assoc 107(500):1610–1624. doi:10.1080/01621459.2012.716344

Chakrabarti A, Ghosh JK (2006) A generalization of BIC for the general exponential family. J Stat Plan 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, Inst. Math. Stat. Collect., vol 3, Inst. Math. Statist., Beachwood, OH, pp 138–154. doi:10.1214/074921708000000110

Cui W, George EI (2008) Empirical Bayes vs. fully Bayes variable selection. J Stat Plan Inference 138(4):888–900. doi:10.1016/j.jspi.2007.02.011

Fernández C, Ley E, Steel MFJ (2001) Benchmark priors for Bayesian model averaging. J Econom 100(2):381–427. doi:10.1016/S0304-4076(00)00076-2

Foster DP, George EI (1994) The risk inflation criterion for multiple regression. Ann Stat 22(4):1947–1975. doi:10.1214/aos/1176325766

García-Donato G, Martínez-Beneito MA (2013) On sampling strategies in Bayesian variable selection problems with large model spaces. J Am Stat Assoc 108(501):340–352. doi:10.1080/01621459.2012.742443

George EI, Foster DP (2000) Calibration and empirical Bayes variable selection. Biometrika 87(4):731–747. doi:10.1093/biomet/87.4.731

Hans C (2009) Bayesian lasso regression. Biometrika 96(4):835–845. doi:10.1093/biomet/asp047

Jeffreys H (1961) Theory of probability, 3rd edn. Clarendon Press, Oxford

Johnson VE, Rossell D (2012) Bayesian model selection in high-dimensional settings. J Am Stat Assoc 107(498):649–660. doi:10.1080/01621459.2012.682536

Kass R, Raftery A (1995) Bayes factors. J Am Stat Assoc 90(430):773–795

Kass RE, Tierney L, Kadane JB (1990) The validity of posterior asymptotic expansions based on Laplace's method. In: Geisser S, Hodges JS, Press SJ, Zellner A (eds) Bayesian and likelihood methods in statistics and econometrics. Elsevier, North-Holland, New York, pp 473–488

Ley E, Steel MFJ (2009) On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. J Appl Econom 24(4):651–674. doi:10.1002/jae.1057

Ley E, Steel MFJ (2012) Mixtures of g-priors for Bayesian model averaging with economic applications. J Econom 171(2):251–266. doi:10.1016/j.jeconom.2012.06.009

Li KC (1987) Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. Ann Stat 15(3):958–975. doi:10.1214/aos/1176350486

Liang F, Paulo R, Molina G, Clyde MA, Berger JO (2008) Mixtures of g priors for Bayesian variable selection. J Am Stat Assoc 103(481):410–423. doi:10.1198/016214507000001337

Lv J, Liu JS (2014) Model selection principles in misspecified models. J R Stat Soc Ser B Stat Methodol 76(1):141–167. doi:10.1111/rssb.12023

Maruyama Y, George EI (2011) Fully Bayes factors with a generalized g-prior. Ann Stat 39(5):2740–2765. doi:10.1214/11-AOS917

Moreno E (1997) Bayes factors for intrinsic and fractional priors in nested models. Bayesian robustness. In: L_1-statistical procedures and related topics (Neuchâtel, 1997), IMS Lecture Notes Monogr. Ser., vol 31, Inst. Math. Statist., Hayward, CA, pp 257–270. doi:10.1214/lnms/1215454142

Moreno E, Girón J, Casella G (2015) Posterior model consistency in variable selection as the model dimension grows. Stat Sci 30(2):228–241. doi:10.1214/14-STS508

Mukhopadhyay M, Samanta T, Chakrabarti A (2015) On consistency and optimality of Bayesian variable selection based on g-prior in normal linear regression models. Ann Inst Stat Math 67(5):963–997. doi:10.1007/s10463-014-0483-8

Narisetty NN, He X (2014) Bayesian variable selection with shrinking and diffusing priors. Ann Stat 42(2):789–817. doi:10.1214/14-AOS1207

Park T, Casella G (2008) The Bayesian lasso. J Am Stat Assoc 103(482):681–686. doi:10.1198/016214508000000337

Shang Z, Clayton MK (2011) Consistency of Bayesian linear model selection with a growing number of parameters. J Stat Plan Inference 141(11):3463–3474. doi:10.1016/j.jspi.2011.05.002

Shao J (1997) An asymptotic theory for linear model selection. Stat Sin 7(2):221–264 (with comments and a rejoinder by the author)

Sparks DK, Khare K, Ghosh M (2015) Necessary and sufficient conditions for high-dimensional posterior consistency under g-priors. Bayesian Anal 10(3):627–664. doi:10.1214/14-BA893

Wang M, Sun X (2014) Bayes factor consistency for nested linear models with a growing number of parameters. J Stat Plan Inference 147:95–105. doi:10.1016/j.jspi.2013.11.001

Wendel JG (1948) Note on the gamma function. Am Math Mon 55:563–564

Xiang R, Ghosh M, Khare K (2016) Consistency of Bayes factors under hyper g-priors with growing model size. J Stat Plan Inference 173:64–86. doi:10.1016/j.jspi.2016.01.001

Zellner A (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Goel PK, Zellner A (eds) Bayesian inference and decision techniques: essays in honor of Bruno de Finetti, pp 233–243

Zellner A, Siow A (1980) Posterior odds for selected regression hypotheses. In: Bernardo JM, DeGroot MH, Lindley DV, Smith AFM (eds) Bayesian statistics. Valencia University Press, Valencia, pp 585–603