Bayes Factor Consistency in Linear Models

Ruixin Guo

University of Missouri, Columbia, USA.

Paul L. Speckman

University of Missouri, Columbia, USA.

Summary. This article examines consistency of Bayes factors in the model comparison problem for linear models. Common approaches to Bayesian analysis of linear models use Zellner's g-prior, a partially conjugate normal prior on the model parameters indexed by a single parameter g. More generally, a hyper-prior can be placed on g, providing a mixture of g-priors. When comparing nested models, flat priors are often placed on the common parameters with the g-prior used for the other parameters, forcing the prior on g to be proper for a determinate Bayes factor. Even for the non-nested case, an "encompassing" approach comparing all models to a base model is often used, where the base model has a flat prior and the prior on g must be proper. In this paper, we consider the Jeffreys prior on g, an improper prior that is also the reference prior. By using this improper prior on all models under comparison, the Bayes factor is always well defined. Under nominal assumptions, the reference prior is shown to yield consistent Bayes factors. Thus we show that the reference prior, successful in many estimation problems, is also appealing for the model comparison problem. In addition, consistency is demonstrated for a large class of proper priors on g as well.

Keywords: Bayes factor, Model selection, consistency, g-prior.

1. Introduction

Model comparison, which refers to using the data to decide on the plausibility of two or model competing models, is a common problem in statistical inference. In the Bayesian framework, the approach for model selection and hypothesis testing is conceptually the same, whereas there is a significant difference in classical frequentist procedures for model selection and hypothesis testing. Non-Bayesian model selection is generally very difficult. Berger and Pericchi (2001) discussed the motivation and advantages for using a Bayesian approach to the model comparison problem over the classical approach.

One important reason for using a Bayesian approach is consistency. Consistency means that the true model (or hypothesis) will be selected if enough data are observed, assuming that one of the competing models is true. In classical hypothesis testing, the p-value is used to determine whether to reject the null hypothesis. In frequentist model selection, the classical tools such as p-values, C_p and AIC are used. However, these classical tools do not guarantee consistency (Berger and Pericchi, 2001). In contrast, Bayesian model comparison can usually provide consistency and a better interpretation. There are various procedures for Bayesian model comparison (Berger and Pericchi, 2001). A natural approach is to use the Bayes factor (Kass and Raftery, 1995), which is based on the posterior model probabilities. In this work, we consider Bayes factor consistency in linear models.

E-mail: rgd27@mizzou.edu, SpeckmanP@Missouri.edu

Consider linear models of the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)^T$ is an $n \times 1$ response vector, $\boldsymbol{X}$ is an $n \times p$ matrix of known quantities, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \delta \boldsymbol{I})$ and δ is the unknown variance. We assume that $\boldsymbol{X}$ has full rank p.

Suppose we have two such linear models, $\boldsymbol{\mathcal{M}}_1$ and $\boldsymbol{\mathcal{M}}_2$, with dimensions p_1 and p_2,

$$\boldsymbol{\mathcal{M}}_1 : \boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}, \tag{2a}$$

$$\boldsymbol{\mathcal{M}}_2 : \boldsymbol{Y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}. \tag{2b}$$

The two competing models are not necessarily nested. A natural way in the Bayesian framework is to compare models in terms of their posterior model probabilities. Let $p(\boldsymbol{\mathcal{M}}_i)$ be the prior probability for model $\boldsymbol{\mathcal{M}}_i,\ i = 1, 2$ and let $m_i(\boldsymbol{Y})$ be the marginal likelihood of $\boldsymbol{Y}$ given $\boldsymbol{\mathcal{M}}_i$. From Bayes Theorem, the posterior model probability for each model $\boldsymbol{\mathcal{M}}_i,\ i = 1, 2$, is

$$p(\boldsymbol{\mathcal{M}}_i \mid \boldsymbol{Y}) = \frac{p(\boldsymbol{\mathcal{M}}_i) m_i(\boldsymbol{Y})}{\sum_j p(\boldsymbol{\mathcal{M}}_j) m_j(\boldsymbol{Y})}. \tag{3}$$

Since $\boldsymbol{\beta}_i$ and δ are unknown, the marginal likelihood of $\boldsymbol{Y}$ given $\boldsymbol{\mathcal{M}}_i$ is

$$m_i(\boldsymbol{Y}) = p(\boldsymbol{Y} \mid \boldsymbol{\mathcal{M}}_i) = \int p(\boldsymbol{Y} \mid \boldsymbol{\beta}_i, \delta) \pi(\boldsymbol{\beta}_i, \delta)\ d\boldsymbol{\beta}_i\ d\delta, \tag{4}$$

where $\pi(\boldsymbol{\beta}_i, \delta)$ in (4) is the prior for the unknown $\boldsymbol{\beta}_i$ and δ under $\boldsymbol{\mathcal{M}}_i$. The Bayes factor in favor of $\boldsymbol{\mathcal{M}}_2$ against $\boldsymbol{\mathcal{M}}_1$, BF_{21}, is defined as the ratio of the posterior odds of $\boldsymbol{\mathcal{M}}_2$ to its prior odds (Kass and Raftery, 1995). From (3), the Bayes factor can be written in terms of the marginal likelihoods,

$$BF_{21} = \frac{m_2(\boldsymbol{Y})}{m_1(\boldsymbol{Y})}, \tag{5}$$

which evaluates the evidence in favor of $\boldsymbol{\mathcal{M}}_2$ against $\boldsymbol{\mathcal{M}}_1$ based on the data.

In order to use the Bayesian method, priors are needed for the unknown parameters $\boldsymbol{\beta}$ and δ. In general, proper priors are required for model specific parameters, since improper priors produce indeterminate Bayes factors (Berger and Pericchi, 2001). A frequently used default Bayesian prior is the Conventional Prior (CP) approach (Berger and Pericchi, 2001), with roots in Jeffreys (1961) and developed by Zellner and Siow (1980) and Zellner (1986). A widely used conventional prior for $\boldsymbol{\beta}$ in the normal linear model is Zellner's g-prior (Zellner, 1986),

$$\pi(\boldsymbol{\beta} | g, \delta) \sim \mathcal{N}(\tilde{\boldsymbol{\beta}}, g\delta(\boldsymbol{X}^T\boldsymbol{X})^{-1}), \tag{6}$$

which is appealing because it is computationally efficient and depends on only one hyper-parameter, g. A brief review is given in Section 2.1.

The choices of g are reviewed in Liang et al. (2008), where they compared mixtures of g-priors (assign a prior on g) with fixed g-priors (fix g as a constant). Simply fixing g may cause some undesirable properties (Liang et al., 2008; Berger and Pericchi, 2001), for instance, the "Information Paradox." Fortunately, many problems associated with the original formulation can be resolved by using mixtures of g-priors (Liang et al., 2008). In the CP approach, noninformative priors are used for common parameters and proper priors are used for the parameters that are in only one model. In the typical case where one model is contained in the other, the parameter g, as a hyper-parameter in the Zellner's g-prior, occurs only in the larger model, which forces the prior on g to be be proper. However, objective priors are usually improper. Marin and Robert (2007b) suggested using g-priors for model selection in a different formulation, allowing the use of an improper prior on g. In this paper, we adopt the idea of Marin and Robert (2007b) and study Bayes factor consistency properties using the reference prior (Bernardo, 1979; Berger and Bernardo, 1989, 1992) for g (see Section 3 for details).

As mentioned in the beginning, consistency refers to the ability of selecting the true model (assuming it exists) if enough data are provided. In terms of posterior model probability, consistency (Fernández et al., 2001) means that

$$P(\boldsymbol{\mathcal{M}}_T \mid \boldsymbol{Y}) \xrightarrow{P} 1 \tag{7}$$

when $\boldsymbol{\mathcal{M}}_T$ is the true model. This is equivalent to considering Bayes factor consistency in the sense that

$$BF_{jT} \xrightarrow{P} 0, \tag{8}$$

for any model $\boldsymbol{\mathcal{M}}_j$ other than the true model.

There are several other papers considering Bayes Factor consistency. Westfall and Gönen (1996) developed a new Bayes factor for testing the null hypothesis of no group differences in one-way ANOVA and evaluated its consistency properties for both fixed and growing model dimensions. Fernández et al. (2001) investigated the consistency problem for the Bayes factor with various choices of "fixed" g, which may depend on n or p or both. Berger et al. (2003) proved asymptotic consistency of Bayes factors for large model dimensions $p = n$ for the one-way ANOVA model assuming known variance δ. In addition, García-Donato and Sun (2007) showed that Bayes factors for testing in a one-way random effects model with intrinsic prior and divergence based priors are consistent for fixed and growing model dimensions p. Maruyama and George (2008) proposed a new form of Bayesian variable selection criterion for normal linear regression model with extended Zellner's g-prior for the case $p > n$ and proved consistency for fixed p as $n \to \infty$.

This paper is organized as follows. In Section 2, Zellner's g-prior is reviewed along with the two different formulations in Liang et al. (2008) and Marin and Robert (2007b). In Section 3, we present the reference prior for g and the associated Bayes factor. In Section 4, we study the consistency problem for the Bayes factor associated with both the reference prior and the usual proper priors for g. Finally, we conclude with our findings and some future work.

2. Mixtures of g-Priors

2.1. Zellner's g-Prior

One prior for linear model (1) that has gained considerable attention is Zellner's g-prior (Zellner, 1986), a special form of a normal-gamma conjugate prior,

$$\pi(\boldsymbol{\beta} \mid g, \delta) \sim \mathcal{N}(\tilde{\boldsymbol{\beta}}, g\delta(\boldsymbol{X}^T\boldsymbol{X})^{-1}),\ \pi(\delta) \propto 1/\delta. \tag{9}$$

The prior mean $\tilde{\boldsymbol{\beta}}$ may either be elicited based on an "imaginary" sample, or it may be a hypothesized value under the null hypothesis. This prior on $\boldsymbol{\beta}$ has a convenient correlation structure and only depends on one parameter g, which controls the amount of information in the prior relative to the sample. As we can see, the amount of subjectivity in Zellner's g-prior is limited to the choice of g. The problem of choosing g becomes crucial. Liang et al. (2008) gave a nice review of Zellner's g-priors along with the choice of g.

Common approaches for model comparison based on Zellner's g-prior adopt the spirit of Zellner and Siow (1980) to put flat priors on the common parameters in the two models and the g-prior on those parameters only in the more complex model for nested models. That is, for $\boldsymbol{\mathcal{M}}_1$ and $\boldsymbol{\mathcal{M}}_2$,

$$\begin{aligned} \boldsymbol{\mathcal{M}}_1 : &\quad \boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon} \\ \boldsymbol{\mathcal{M}}_2 : &\quad \boldsymbol{Y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_{-1}\boldsymbol{\beta}_{-1} + \boldsymbol{\epsilon}, \end{aligned} \tag{10}$$

where $\boldsymbol{X}_{-1}\boldsymbol{\beta}_{-1}$ refers to the part excluded in model $\boldsymbol{\mathcal{M}}_1$. Flat priors are used for $(\boldsymbol{\beta}_1, \delta)$ and the g-prior is assigned to $\boldsymbol{\beta}_{-1}$, which is only contained in $\boldsymbol{\mathcal{M}}_2$:

$$\pi(\boldsymbol{\beta}_1, \delta) \propto 1/\delta, \quad \boldsymbol{\beta}_{-1} \mid g, \delta \sim \mathcal{N}(\mathbf{0}, g\delta(\boldsymbol{X}_{-1}^T\boldsymbol{X}_{-1})^{-1}). \tag{11}$$

The prior mean $\tilde{\boldsymbol{\beta}}_{-1}$ of $\boldsymbol{\beta}_{-1}$ is set to zero since (10) is set up to test $\boldsymbol{\beta}_{-1} = 0$. This setting is appealing in testing problems, but it is less appealing for model selection because the prior specification for the more complex model depends on the simpler model under comparison.

For any two non-nested models $\boldsymbol{\mathcal{M}}_1$ and $\boldsymbol{\mathcal{M}}_2$, the "encompassing" approach of Zellner (1984) is usually used (Liang et al., 2008; Marin and Robert, 2007b; Casella and Girón, 2008)) to compute

$$BF_{21} = \frac{BF_{2b}}{BF_{1b}}, \tag{12}$$

where the subscript b indicates the base model, either the full model $\boldsymbol{\mathcal{M}}_F$ or the null model $\boldsymbol{\mathcal{M}}_N$, to make each pair nested so that (11) formulation can still be used.

In this formulation, the prior for g must be proper in order to get a determinate Bayes factor since it only appears in the more complex model.

2.2. *Mixtures of g-Priors*

Liang et al. (2008) considered the model selection problem in the linear regression model context and proposed fully Bayes approaches by using mixtures of g-priors. They used the formulation (11) and (12), which restricts that the prior for g has to be proper. Liang et al. (2008) discussed three families of proper priors for g, which they termed the Zellner-Siow, hyper-g and hyper-g/n priors:

- Zellner-Siow priors: $\pi^{ZS}(g) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} e^{-n/2g}$,
- hyper-g priors: $\pi^{g}(g) = \frac{a-2}{2}(1+g)^{-a/2}, \quad 2 < a \leq 4$,
- hyper-g/n priors: $\pi^{g/n}(g) = \frac{a-2}{2n}(1+g/n)^{-a/2}, \quad a > 2$.

The prior distribution of $\boldsymbol{\beta} \mid \delta$ based on $\pi^{ZS}(g)$ is actually the multivariate Cauchy distribution suggested in Zellner and Siow (1980). The hyper-g/n prior is a modification of the hyper-g prior in order to get a consistent Bayes factor under the null model (when there is only the intercept term). Under any non-null model, Liang et al. (2008) show consistency of the Bayes factor associated with these three types of priors, whereas under the null model consistency does not hold for hyper-g prior.

Marin and Robert (2007b) proposed a different but novel way to use mixtures of g-priors for model selection in the regression model. They compare each model $\boldsymbol{\mathcal{M}}_i$ with the full model $\boldsymbol{\mathcal{M}}_F$. For each model, they placed Zellner's g-prior on the regression coefficients $\boldsymbol{\beta}_i$ (or $\boldsymbol{\beta}$ in $\boldsymbol{\mathcal{M}}_F$) and a noninformative prior on the variance δ,

$$\boldsymbol{\mathcal{M}}_i : \boldsymbol{\beta}_i \mid g_i, \delta \sim \mathcal{N}_{p_i}(\tilde{\boldsymbol{\beta}}_i, g_i\delta(\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}), \quad \pi(\delta) \propto 1/\delta, \tag{13a}$$

$$\boldsymbol{\mathcal{M}}_F : \boldsymbol{\beta} \mid g, \delta \sim \mathcal{N}_{p}(\tilde{\boldsymbol{\beta}}, g\delta(\boldsymbol{X}^T\boldsymbol{X})^{-1}), \quad \pi(\delta) \propto 1/\delta, \tag{13b}$$

where $\tilde{\boldsymbol{\beta}}_i = (\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^T\boldsymbol{X}\tilde{\boldsymbol{\beta}}$. If one uses the same prior for g_i and g, then the prior on g can be improper but still have a determinate Bayes factor. This is encouraging, since objective priors are usually improper. Marin and Robert (2007b) introduced a discrete improper diffuse prior distribution on g (see Section 3.3.2 of their book for details). However, the infinite summation becomes a problem in practice and the values for g are restricted to be integers. Later, Marin and Robert (2007a) suggested that a continuous improper prior for g, the Jeffreys' prior,

$$\pi(g, \delta) \propto \frac{1}{\delta(1+g)}, \tag{14}$$

may be used. However, they didn't consider any large sample consistency property. In Section 3, we show that (14) is also the reference prior.

3. Bayes Factor and Priors

In this paper, for the model comparison problem (2), we adopt the idea of Marin and Robert (2007b) and place a g-prior on each model as in (13):

$$\boldsymbol{\mathcal{M}}_1 : \boldsymbol{\beta}_1 \mid g, \delta \sim \mathcal{N}_{p_1}(\mathbf{0}, g\delta(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}), \quad \pi(\delta) \propto 1/\delta, \tag{15a}$$

$$\boldsymbol{\mathcal{M}}_2 : \boldsymbol{\beta}_2 \mid g, \delta \sim \mathcal{N}_{p_2}(\mathbf{0}, g\delta(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}), \quad \pi(\delta) \propto 1/\delta, \tag{15b}$$

where $\boldsymbol{X}_i$ is the $n \times p_i$ known design matrix $\boldsymbol{\mathcal{M}}_i$, $i = 1,\ 2$. Without loss of generality, we choose $\tilde{\boldsymbol{\beta}}_i = 0$, $i = 1, 2$ in (15), where $\tilde{\boldsymbol{\beta}}_i$ is defined as in (13). One special case is important. In the model selection context when comparing non-null regression models of the form

$$\boldsymbol{Y}_i = \beta_0\mathbf{1}_n + \boldsymbol{X}_i\boldsymbol{\beta}_i + \epsilon, \quad i = 1, 2, \tag{16}$$

where β_0 denotes the intercept term and $\mathbf{1}_n$ is a $n \times 1$ vector of 1, a flat prior can be put on β_0. Then it is reasonable to set $\tilde{\boldsymbol{\beta}} = 0$ as in Liang et al. (2008). In fact, after integrating out β_0, the results in the following sections would be the same as under (15).

Under formulation (15), g is in each model so that both proper priors and improper priors on g can be used.

3.1. Bayes Factor

Let $\pi(g)$ be the prior on g. It's straightforward to show that the marginal likelihood of $\boldsymbol{Y}$ given model $\boldsymbol{\mathcal{M}}_i$ satisfies

$$p(\boldsymbol{Y} \mid \boldsymbol{\mathcal{M}}_i) \propto (\boldsymbol{Y}^T\boldsymbol{Y})^{-n/2} \int_0^\infty (1+g)^{-p_i/2}(1-\frac{g}{1+g}\frac{\boldsymbol{Y}^T H_i \boldsymbol{Y}}{\boldsymbol{Y}^T\boldsymbol{Y}})^{-n/2}\pi(g)\,dg, \tag{17}$$

where $H_i = \boldsymbol{X}_i(\boldsymbol{X}_i^T\boldsymbol{X}_i)\boldsymbol{X}_i^T$ is the $n \times n$ hat matrix for model $\boldsymbol{\mathcal{M}}_i$, $i = 1,\ 2$.

By definition, the Bayes factor for comparing model $\boldsymbol{\mathcal{M}}_1$ with model $\boldsymbol{\mathcal{M}}_2$ is

$$\begin{aligned} BF_{21} &= \frac{\int_0^\infty (1+g)^{-p_2/2}(1-\frac{g}{1+g}\widetilde{R}_2^2)^{-n/2}\pi(g)\,dg}{\int_0^\infty (1+g)^{-p_1/2}(1-\frac{g}{1+g}\widetilde{R}_1^2)^{-n/2}\pi(g)\,dg} \\ &= \frac{\int_0^\infty (1+g)^{(n-p_2)/2}[1+g(1-\widetilde{R}_2^2)]^{-n/2}\pi(g)\,dg}{\int_0^\infty (1+g)^{(n-p_1)/2}[1+g(1-\widetilde{R}_1^2)]^{-n/2}\pi(g)\,dg}, \end{aligned} \tag{18}$$

where

$$\widetilde{R}_i^2 = \boldsymbol{Y}^T H_i \boldsymbol{Y}/\boldsymbol{Y}^T\boldsymbol{Y}, \quad i = 1, 2. \tag{19}$$

Note that (18) is very similar to the formulas in Liang et al. (2008) when comparing non-null models. One difference is that $\widetilde{R}^2$ here is not the true coefficient of determination unless all models have an intercept term as in (16).

3.2. Reference Prior for g

In this paper, we focus our discussion based on an improper $\pi(g)$, the reference prior (Bernardo, 1979). It is derived based on the idea to maximize the expected Kullback-Leibler divergence of the posterior relative to the prior. The reference prior has proved to be remarkably successful for estimation problems. The purpose here is to study the properties of the Bayes factor associated with the joint reference prior on δ and g.

In order to find the reference prior, first we need to find the Fisher information matrix. For model (1) with Zellner's g prior for $\boldsymbol{\beta}$, i.e. $\boldsymbol{\beta} \mid g, \delta \sim N_p(0, g\delta(\boldsymbol{X}^T\boldsymbol{X})^{-1})$, by a straightforward calculation, the likelihood function for (g, δ) is

$$L(g,\delta) = p(\boldsymbol{Y} \mid g, \delta) \propto \delta^{-n/2}(1+g)^{-p/2}\exp\left\{-\frac{1}{2\delta}(\boldsymbol{Y}^T\boldsymbol{Y} - \frac{g}{1+g}\boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y})\right\}, \tag{20}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. It's easy to show that the Fisher information matrix for (g, δ) is

$$\boldsymbol{I}(g,\delta) = \begin{bmatrix} \frac{p}{2(1+g)^2} & \frac{p}{2\delta(1+g)} \\ \frac{p}{2\delta(1+g)} & \frac{n}{2\delta^2} \end{bmatrix}, \tag{21}$$

and

$$|\boldsymbol{I}(g,\delta)| = \frac{p(n-p)}{4\delta^2(1+g)^2}. \tag{22}$$

Hence, the Jeffreys' prior of (g, δ) is (14). In the following theorem, we show that the reference prior has the same form of the Jeffreys' prior, i.e.,

$$\pi^R(g, \delta) \propto \frac{1}{\delta(1+g)}. \tag{23}$$

THEOREM 1.

(a) The improper prior (23) is the Jeffreys' prior for (g, δ).

(b) It is also the one-at-a-time reference prior when either g or δ is of interest and the other is the nuisance parameter.

(c) It is the independent reference prior for g and δ.

(d) The Jeffreys and reference priors do not depend on the choice of $\boldsymbol{X}_i$.

The proof is given in Appendix A.1. In fact, this prior on g, $\pi(g) \propto 1/(1+g)$, is a special case of the hyper-g prior with $a = 2$, $\pi(g) = \frac{a-2}{2}(1+g)^{-a/2}$, which is discussed in Liang et al. (2008). However, they only focus on proper priors, i.e., $a > 2$. In this paper, we will discuss the asymptotic properties of Bayes factors based on the (improper) reference prior for g.

4. Bayes Factor Consistency

We consider consistency of the Bayes factors defined as follows.

DEFINITION 1. *The Bayes factor BF_{21} for comparing $\boldsymbol{\mathcal{M}}_1$ with $\boldsymbol{\mathcal{M}}_2$ is consistent if, as $n \to \infty$, $BF_{21} \xrightarrow{P} 0$ under model $\boldsymbol{\mathcal{M}}_1$ and $BF_{21} \xrightarrow{P} \infty$ under $\boldsymbol{\mathcal{M}}_2$. If both relations hold with probability 1, the Bayes factor is called almost surely consistent or strongly consistent.*

4.1. *Consistency under the Reference Prior*

Consider using the g-prior in both models as in (15) with the reference prior (23) on g, i.e., the improper prior on g for the Bayes factor (18). The following theorem shows consistency of the Bayes factor under the reference prior on g for model (2).

THEOREM 2. *Without loss of generality, suppose that $\boldsymbol{\mathcal{M}}_1$ is the true model with model dimension p_1, and $\boldsymbol{\mathcal{M}}_2$ is any other model with dimension p_2. Assume that*

$$\frac{\boldsymbol{\beta}_1^T \boldsymbol{X}_1^T \boldsymbol{X}_1 \boldsymbol{\beta}_1}{n} \to b_1 > 0, \quad n \to \infty. \tag{24}$$

Consider the Bayes factor BF_{21} associated with the reference prior (23):

(i) If $\boldsymbol{\mathcal{M}}_1 \subset \boldsymbol{\mathcal{M}}_2$, then $BF_{21} \xrightarrow{P} 0$ as $n \to \infty$. Moreover, if $p_2 - p_1 \geq 3$, almost sure convergence holds, i.e., $BF_{21} \xrightarrow{a.s.} 0$.

(ii) If $\boldsymbol{\mathcal{M}}_1 \not\subseteq \boldsymbol{\mathcal{M}}_2$ and

$$\frac{\boldsymbol{\beta}_1^T \boldsymbol{X}_1^T \boldsymbol{H}_2 \boldsymbol{X}_1 \boldsymbol{\beta}_1}{n} \to b_2, \quad 0 \leq b_2 < b_1, \tag{25}$$

then $BF_{21} \xrightarrow{a.s.} 0$.

The proof of this theorem is given in Appendix A.4. From (18), we can see that the asymptotic behavior of the Bayes factor entirely depends on the asymptotic behavior of $\widetilde{R}_i^2$, $i = 1, 2$. In general, $\widetilde{R}^2$ can be expressed in terms of quadratic forms with noncentral chi-square distributions. Fernández et al. (2001) gave some results about convergence in probability of residual sums of squares in the context of linear regression. We develop some more general and stronger results in the next two lemmas.

Lemma 1. *Suppose $Q_d \sim \chi_d^2(u_d)$, where u_d is the noncentral parameter and d is the degrees of freedom for noncentral chi-square distribution. If $u_d/d \to c_u,\ 0 \le c_u < \infty$ as $d \to \infty$, then*

$$\frac{Q_d - u_d}{d} \xrightarrow{a.s.} 1 \text{ as } d \to \infty. \tag{26}$$

The proof of this lemma is given in Appendix A.2. Based on this lemma, we can obtain the limits of $\widetilde{R}_i^2$.

Lemma 2. *Assume that $\boldsymbol{\mathcal{M}}_1$ is the true model and (24) holds.*

(i) If $\boldsymbol{\mathcal{M}}_1 \subset \boldsymbol{\mathcal{M}}_2$, then

$$\widetilde{R}_1^2 \stackrel{a.s.}{\to} \frac{b_1}{\delta + b_1}, \quad \widetilde{R}_2^2 \stackrel{a.s.}{\to} \frac{b_1}{\delta + b_1}. \tag{27}$$

(ii) When $\boldsymbol{\mathcal{M}}_1 \not\subseteq \boldsymbol{\mathcal{M}}_2$, under assumption (25),

$$\widetilde{R}_1^2 \stackrel{a.s.}{\to} \frac{b_1}{\delta + b_1}, \quad \widetilde{R}_2^2 \stackrel{a.s.}{\to} \frac{b_2}{\delta + b_1}. \tag{28}$$

See Appendix A.3 for the proof. As we can see from Lemma 2, $\widetilde{R}_1^2$ and $\widetilde{R}_2^2$ converge to the same constant almost surely if $\boldsymbol{\mathcal{M}}_1 \subset \boldsymbol{\mathcal{M}}_2$. When $\boldsymbol{\mathcal{M}}_1 \not\subseteq \boldsymbol{\mathcal{M}}_2$, the limit of $\widetilde{R}_1^2$ is greater than that of $\widetilde{R}_2^2$, which makes intuitive sense. Theorem 2 provides some theoretical justification for using the reference prior on (δ, g) for the model comparison problem (2).

4.2. Consistency for Proper Priors

In this section, we consider proper priors on g. Liang et al. (2008) showed Bayes factor consistency associated with Zellner-Siow priors, hyper-g priors and hyper-g/n priors when comparing any non-null models under the formulation (10), (11) and (12). In the following theorem, we provide a more general result for proper priors under prior structure (15). This theorem gives sufficient (but weak) conditions under which a sequence of proper priors yields a consistent Bayes factor.

We adopt the following notation for the rest part of this paper. The notation $a_n \approx b_n$ will mean $\lim_{n\to\infty} a_n/b_n$ exists and is positive almost surely.

Theorem 3. *Without loss of generality, suppose that $\boldsymbol{\mathcal{M}}_1$ is the true model with model dimension p_1, and $\boldsymbol{\mathcal{M}}_2$ is any other model with dimension p_2. Consider a sequence $\pi(g)$ of proper priors on g (which may depend on n) that satisfy two conditions:*

(a1) There is a constant $k \ge 0$ such that $\int_{a_n}^{c_0 a_n} \pi(g)\, dg \approx n^{-k}$ for any constant $c_0 > 1$ and any sequence $a_n \approx n$.

(a2) There exists a constant k_u such that $k-(p_2-p_1)/2 < k_u \leq k$ and $\int_0^\infty (1+g)^{k_u}\pi(g)\,dg \approx 1$.

Assume also that the model structure $\boldsymbol{X}_1\boldsymbol{\beta}_1$ satisfies (24).

(i) If $\boldsymbol{\mathcal{M}}_1 \subset \boldsymbol{\mathcal{M}}_2$ and (a1) and (a2) hold, then $BF_{21} \xrightarrow{P} 0$. Furthermore, strong consistency holds if $p_2 - p_1 > 2 + 2(k - k_u)$.

(ii) If $\boldsymbol{\mathcal{M}}_1 \not\subseteq \boldsymbol{\mathcal{M}}_2$ and (a1) and (25) hold, then $BF_{21} \xrightarrow{a.s.} 0$.

The proof of this theorem is in Appendix A.5. In this theorem, we assumed that $\boldsymbol{\mathcal{M}}_1$ is the true model. In fact, the same results hold if $\boldsymbol{\mathcal{M}}_2$ is true by reversing the subscripts. Therefore, any proper prior that satisfies conditions (a1) and (a2) of Theorem 3 yields consistency.

The conditions in Theorem 3 are actually quite weak and many proper priors satisfy them. The following corollaries are proved in Appendix A.6. To begin, the proper priors considered in Liang et al. (2008) yield consistent Bayes factors.

Corollary 1. *The Bayes factors BF_{21} associated with the Zellner-Siow, hyper-g, and hyper-g/n priors, are consistent as $n \to \infty$.*

The hyper-g priors and hyper-g/n priors are proper when $a > 2$. Liang et al. (2008) suggested taking $2 < a \leq 4$. These priors yield consistent Bayes factors.

Maruyama and George (2008) proposed another proper prior on g, the beta-prime prior with density

$$\pi^B(g) = \frac{g^b(1+g)^{-a-b-2}}{B(a+1,b+1)}, a > -1, b > -1. \tag{29}$$

They suggested choosing $-1 < a < -1/2$ and $b \approx n$. With their choice of parameters or with fixed b, we have the following consistency result.

Corollary 2. *The Bayes factor BF_{21} associated with the beta-prime priors* (29) *with fixed (a,b) or with $-1 < a < -1/2$ and $b \approx n$ is consistent.*

Finally, the Zellner-Siow and hyper-g are both examples of scale families of priors. In fact, any such scale family of priors yields consistent Bayes factors.

Corollary 3. *If π is any proper density on $[0,\infty)$ and the prior on g is the scale family $\pi^S(g) = \pi(g/n)/n$, $n = 1, 2, \ldots$, then the associated Bayes factor is consistent.*

5. Discussion

This paper treats consistency problems for the Bayes factor model comparison using mixtures of g-priors in the context of linear model. Usually the g-prior is only used on the more complex model, which forces the prior on g to be proper. In this paper, we adopt the idea of Marin and Robert (2007b) to assign a g-prior on each model so that an improper prior for g can be used. Reference priors are widely used for estimation prior and have proved to be remarkably successful. It is interesting to study the property of the Bayes factor associated with the reference prior on g, which is improper. We show that the reference prior is also appealing in our model comparison formulation since the associated Bayes factor is consistent. We also provide a general result for proper priors on g as stated in Theorem 3.

In this paper, we fix the model dimension, i.e., p_1 and p_2 do not depend on n. It would be interesting to consider the case when the model dimension actually depends on n. In a on-going project, we investigate the consistency problems of the Bayes factor when the model dimension p grows with n and obtain consistency or inconsistency depending on the limit behavior of p/n.

A. Appendix: Proofs

A.1. Proof of Theorem 1

Parts (a), (c) and (d) are obvious. For part (b), if g is the parameter of interest, we follow Berger and Bernardo (1989)'s reference prior algorithm. Choose a nested sequence of compact sets

$$(0, b_l) \times (c_l, d_l)$$

for (g, δ), where $b_l \to \infty$, $c_l \to 0$ and $d_l \to \infty$.

The Fisher information matrix is given in (21). It is easy to see that the conditional reference prior for δ given g is

$$\pi^l(\delta|g) = \frac{|I_{22}|^{1/2}\mathbf{1}_{(c_l,d_l)}(\delta)}{\displaystyle\int_{c_l}^{d_l} |I_{22}|^{1/2} d\delta} = \frac{\mathbf{1}_{(c_l,d_l)}(\delta)}{\delta \log(d_l/c_l)},$$

and the marginal prior of g on $(0, b_l)$ is

$$\begin{aligned} \pi^l(g) &\propto \exp\left\{\frac{1}{2}\int_{c_l}^{d_l} \pi^l(\delta|g) \log\left|\frac{\boldsymbol{I}}{I_{22}}\right| d\delta\right\} \\ &= \exp\left\{\frac{1}{2}\int_{c_l}^{d_l} \frac{1}{\delta\log(d_l/c_l)} \log\left[\frac{p(n-p)}{2n(1+g)^2}\right] d\delta\right\} \\ &\propto \frac{1}{1+g}, \quad g \in (0, b_l). \end{aligned}$$

Thus

$$\pi^l(g) = \frac{\mathbf{1}_{(0,b_l)}(g)}{(1+g)\log(b_l+1)},$$

and the reference prior is

$$\pi^R(g,\delta) = \lim_{l\to\infty} \frac{\pi^l(\delta|g)\pi^l(g)}{\pi^l(1|1)\pi^l(1)} \propto \frac{1}{\delta(1+g)}. \tag{30}$$

If δ is the parameter of interest and g is the nuisance parameter, we have

$$\pi^l(g|\delta) = \frac{|I_{11}|^{1/2}\mathbf{1}_{(0,b_l)}(g)}{\displaystyle\int_0^{b_l} |I_{11}|^{1/2} dg} = \frac{\mathbf{1}_{(0,b_l)}(g)}{(1+g)\log(b_l+1)},$$

and

$$\begin{aligned}\pi^l(\delta) &\propto \exp\left\{\frac{1}{2}\int_0^{b_l}\pi^l(g|\delta)\log\left|\frac{\boldsymbol{I}}{I_{11}}\right|dg\right\}\\ &= \exp\left\{\frac{1}{2}\int_0^{b_l}\frac{1}{(1+g)\log(b_l+1)}\log\left[\frac{n-p}{2\delta^2}\right]dg\right\}\\ &\propto \frac{1}{\delta},\quad \delta\in(c_l,d_l).\end{aligned}$$

Thus $\pi^l(\delta) = \dfrac{\mathbf{1}_{(c_l,\delta_l)}(\delta)}{\delta\log(d_l/c_l)}$,

$$\pi^R(g,\delta) = \lim_{l\to\infty}\frac{\pi^l(g|\delta)\pi^l(\delta)}{\pi^l(1|1)\pi^l(1)} \propto \frac{1}{\delta(1+g)}, \tag{31}$$

and part (b) holds.

□

A.2. Proof of Lemma 1

Since Q_d has a noncentral chi-square distribution, it can be written as

$$Q_d = (Z_{1d}+\sqrt{u_d})^2 + Z_{2d}^2 + \ldots + Z_{dd}^2,$$

where $Z_{id} \stackrel{iid}{\sim} N(0,1)$, $i = 1,\ldots,d$. Then

$$\frac{Q_d-u_d}{d} = \frac{1}{d}\sum_{i=1}^{d} Z_{id}^2 + 2\frac{Z_{1d}}{\sqrt{d}}\sqrt{\frac{u_d}{d}},$$

where $\{Z_{id}^2\}$ is actually a triangular array and the Strong Law of Large Numbers (SLLN) for i.i.d. samples does not apply. But we can still obtain $\sum_{i=1}^d Z_{id}^2/d \xrightarrow{a.s.} 1$ by the strong law of large numbers for triangular arrays (refer to Hu, and Taylor (1997), Theorem 2.1). Moreover, by the Borel Cantelli lemma, $Z_{1d}/\sqrt{d} \stackrel{a.s.}{\to} 0$. Since $\sqrt{u_d/d}$ converges to a constant, $(Q_d-u_d)/d \xrightarrow{a.s.} 1$ as $d\to\infty$.

□

A.3. Proof of Lemma 2

$$\widetilde{R}_1^2 = \frac{\boldsymbol{Y}^T\boldsymbol{H}_1\boldsymbol{Y}}{\boldsymbol{Y}^T\boldsymbol{Y}} = 1-\frac{\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H}_1)\boldsymbol{Y}/\delta}{\boldsymbol{Y}^T\boldsymbol{Y}/\delta} = 1-Q_1/Q, \tag{32}$$

$$\widetilde{R}_2^2 = \frac{\boldsymbol{Y}^T\boldsymbol{H}_2\boldsymbol{Y}}{\boldsymbol{Y}^T\boldsymbol{Y}} = 1-\frac{\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H}_2)\boldsymbol{Y}/\delta}{\boldsymbol{Y}^T\boldsymbol{Y}/\delta} = 1-Q_2/Q, \tag{33}$$

where $Q = \boldsymbol{Y}^T\boldsymbol{Y}/\delta$, $Q_1 = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H}_1)\boldsymbol{Y}/\delta$ and $Q_2 = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H}_2)\boldsymbol{Y}/\delta$.

(i) If $\boldsymbol{\mathcal{M}}_1$ is nested within model $\boldsymbol{\mathcal{M}}_2$,

$$
\begin{aligned}
Q &\sim \chi^2_n(\lambda_1), \quad \lambda_1 = \boldsymbol{\beta}_1^T \boldsymbol{X}_1^T \boldsymbol{X}_1 \boldsymbol{\beta}_1/\delta, && (34)\\
Q_1 &\sim \chi^2_{n-p_1}(0), && (35)\\
Q_2 &\sim \chi^2_{n-p_2}(0). && (36)
\end{aligned}
$$

Under assumption (24), $\lambda_1/n \to b_1/\delta$. By Lemma 1, we have $Q/n \stackrel{a.s.}{\to} 1 + b_1/\delta$, $Q_1/(n - p_1) \stackrel{a.s.}{\to} 1$, and $Q_2/(n - p_2) \stackrel{a.s.}{\to} 1$. Thus, $\widetilde{R}^2_1 \stackrel{a.s.}{\to} 1 - \delta/(\delta + b_1) = b_1/(\delta + b_1)$ and $\widetilde{R}^2_2 \stackrel{a.s.}{\to} 1 - \delta/(\delta + b_2) = b_1/(\delta + b_1)$.

(ii) For any model $\boldsymbol{\mathcal{M}}_2$ that does not contain $\boldsymbol{\mathcal{M}}_1$, under assumption (24), Q and Q_1 still have distributions (34) and (35) respectively, and $\widetilde{R}^2_1 \stackrel{a.s.}{\to} 1 - \delta/(\delta + b_1) = b_1/(\delta + b_1)$. Let $\lambda_2 = \boldsymbol{\beta}_1^T \boldsymbol{X}_1^T (\boldsymbol{I} - \boldsymbol{H}_2) \boldsymbol{X}_1 \boldsymbol{\beta}_1/\delta$, so $Q_2 \sim \chi^2_{n-p_2}(\lambda_2)$. By (25), $\lambda_2/n \to (b_1 - b_2)/\delta > 0$. Using Lemma 1 again, we have $Q_2/(n - p_2) \stackrel{a.s.}{\to} 1 + (b_1 - b_2)/\delta$. Thus, $\widetilde{R}^2_2 \stackrel{a.s.}{\to} 1 - (\delta + b_1 - b_2)/(\delta + b_1) = b_2/(\delta + b_1)$.

□

A.4. Proof of Theorem 2

We adopt the following notation for the proofs of the theorems. Recall that $a_n \approx b_n$ means that a_n/b_n converges to a positive limit almost surely. In addition, the notation $a_n \lesssim b_n$ means $\limsup_{n\to\infty} a_n/b_n$ exists and is positive almost surely.

The Bayes factor (18) is a ratio of two integrals. Let I_2 and I_1 denote the integrals in the numerator and the denominator respectively. We derive lower and upper bounds for each integral. Consider I_1 first. For simplicity, we reparameterize with $t = g/(1 + g)$. Let $\hat{t}_1 = (n\widetilde{R}^2_1 - p_1 + 2/3)/[\widetilde{R}^2_1(n - p_1 + 2/3)]$ be the maximizer of the function $(1-t)^{p_1/2-1/3}(1 - \widetilde{R}^2_1 t)^{-n/2}$. By Lemma 2, $\hat{t}_1 > 0$ for sufficiently large n almost surely. For such n and for the reference prior on g,

$$
\begin{aligned}
I_1 &= \int_0^\infty (1+g)^{-p_1/2}(1 - \frac{g}{1+g}\widetilde{R}^2_1)^{-n/2} \pi^R(g)\, dg \\
&= \int_0^\infty (1+g)^{-p_1/2-1}(1 - \frac{g}{1+g}\widetilde{R}^2_1)^{-n/2}\, dg \\
&= \int_0^1 (1-t)^{p_1/2-1}(1 - \widetilde{R}^2_1 t)^{-n/2}\, dt = \int_0^1 (1-t)^{p_1/2-1/3}(1 - \widetilde{R}^2_1 t)^{-n/2}(1-t)^{-2/3}\, dt \\
&\le (1-\hat{t}_1)^{p_1/2-1/3}(1 - \widetilde{R}^2_1 \hat{t}_1)^{-n/2} \int_0^1 (1-t)^{-2/3}\, dt \\
&= 3\left(\frac{p_1 - 2/3}{n - p_1 + 2/3}\right)^{p_1/2-1/3} \left(\frac{1-\widetilde{R}^2_1}{\widetilde{R}^2_1}\right)^{p_1/2-1/3} \left(\frac{n}{n - p_1 + 2/3}\right)^{-n/2} (1 - \widetilde{R}^2_1)^{-n/2} \\
&\approx n^{-p_1/2+1/3}(1 - \widetilde{R}^2_1)^{-n/2} \triangleq u_1(n).
\end{aligned}
$$

Next, consider a lower bound for I_1, assuming $p_1 \geq 2$:

$$\begin{aligned} I_1 &= \int_0^1 (1-t)^{p_1/2-1}(1-\widetilde{R}_1^2 t)^{-n/2}\,dt \geq \int_0^{1-\frac{1}{n}} (1-t)^{p_1/2-1}(1-\widetilde{R}_1^2 t)^{-n/2}\,dt \\ &\geq n^{1-p_1/2}\int_0^{1-\frac{1}{n}} (1-\widetilde{R}_1^2 t)^{-n/2}\,dt = \frac{n^{1-p_1/2}}{(n-2)\widetilde{R}_1^2}\left\{\left[1-(1-\frac{1}{n})\widetilde{R}_1^2\right]^{1-n/2}-1\right\} \\ &\approx n^{-p_1/2}(1-\widetilde{R}_1^2)^{-n/2} \triangleq l_1(n). \end{aligned} \tag{37}$$

When $p_1 = 1$,

$$\begin{aligned} I_1 &= \int_0^1 (1-t)^{-1/2}(1-\widetilde{R}_1^2 t)^{-n/2}\,dt \geq \int_{1-\frac{1}{n}}^1 (1-t)^{-1/2}(1-\widetilde{R}_1^2 t)^{-n/2}\,dt \\ &\geq n^{1/2}\int_{1-\frac{1}{n}}^1 (1-\widetilde{R}_1^2 t)^{-n/2}\,dt \geq n^{1/2}\left[1-\left(1-\frac{1}{n}\right)\right]\left[1-\left(1-\frac{1}{n}\right)\widetilde{R}_1^2\right]^{-n/2} \\ &\approx n^{-1/2}(1-\widetilde{R}_1^2)^{-n/2} = l_1(n). \end{aligned} \tag{38}$$

So (37) holds for $p \geq 1$.

Similarly, let $l_2(n) = n^{-p_2/2}(1-\widetilde{R}_2^2)^{-n/2}$ and $u_2(n) = n^{-p_2/2+1/3}(1-\widetilde{R}_2^2)^{-n/2}$. We have lower and upper bounds for I_2,

$$l_2(n) \lesssim I_2 \lesssim u_2(n).$$

If $\boldsymbol{\mathcal{M}}_1$ is the true model, consistency for the Bayes factor requires $BF_{21} \to 0$ as $n \to \infty$. It's sufficient to show

$$BF_{21} = \frac{I_2}{I_1} \lesssim \frac{u_2(n)}{l_1(n)} = n^{(p_1-p_2)/2+1/3}\left(\frac{1-\widetilde{R}_1^2}{1-\widetilde{R}_2^2}\right)^{n/2} \to 0, \quad n \to \infty. \tag{39}$$

Now consider case (ii) first, $\boldsymbol{\mathcal{M}}_1 \nsubseteq \boldsymbol{\mathcal{M}}_2$. From Lemma 2, as $n \to \infty$,

$$\frac{1-\widetilde{R}_1^2}{1-\widetilde{R}_2^2} \xrightarrow{a.s.} \frac{\delta}{\delta + b_1 - b_2} < 1.$$

Thus $[(1-\widetilde{R}_1^2)/(1-\widetilde{R}_2^2)]^{n/2} \to 0$ exponentially fast with respect to n. We conclude that (39) holds regardless of the sign of $p_1 - p_2$, and consequently BF_{21} is consistent.

Finally consider case (i), $\boldsymbol{\mathcal{M}}_1 \subset \boldsymbol{\mathcal{M}}_2$. We have

$$\frac{1-\widetilde{R}_1^2}{1-\widetilde{R}_2^2} = \frac{Q_1}{Q_2} = 1 + \frac{Q_1 - Q_2}{Q_2/(n-p_2)}\frac{1}{n-p_2}. \tag{40}$$

Applying Lemma 2 again, as $n \to \infty$, $Q_2/(n-p_2) \xrightarrow{a.s.} 1$ and $Q_1 - Q_2 = \boldsymbol{Y}^T(\boldsymbol{H}_2 - \boldsymbol{H}_1)\boldsymbol{Y}/\delta \sim \chi^2_{p_2-p_1}(0)$. Thus

$$\left(\frac{1-\widetilde{R}_1^2}{1-\widetilde{R}_2^2}\right)^{n/2} = \left(1 + O_P(n^{-1})\right)^{n/2} = O_P(1), \tag{41}$$

which is bounded in probability. Moreover, $\boldsymbol{\mathcal{M}}_1$ is nested within $\boldsymbol{\mathcal{M}}_2$, i.e. $p_2 - p_1 \geq 1$, so that $(p_1 - p_2)/2 + 1/3 \leq -1/2 + 1/3 < 0$ and $n^{(p_1-p_2)/2+1/3} \to 0$ as $n \to \infty$. Therefore, (39) holds in probability, and BF_{21} is consistent in this case.

If $p_2 - p_1 = 1$ or 2, it appears that no stronger general result is possible. However, if $p_2 - p_1 \geq 3$, almost sure convergence of BF_{21} for this case can be proved. We give an outline of the proof using the Borel-Cantelli Lemma. Let

$$\begin{aligned} B_n &= u_2(n)/l_1(n) = n^{(p_1-p_2)/2+1/3}\left(\frac{1-\widetilde{R}_1^2}{1-\widetilde{R}_2^2}\right)^{n/2} \\ &= n^{(p_1-p_2)/2+1/3}\left(1+\frac{Q_1-Q_2}{Q_2/(n-p_2)}\frac{1}{n-p_2}\right)^{n/2}. \end{aligned}$$

In order to show almost sure consistency, it's equivalent to show $P\{|B_n| > \epsilon \text{ i.o.}\} = 0$ for any $\epsilon > 0$. By the Borel-Cantelli Lemma, we need to show $\sum_n P\{|B_n| > \epsilon\} < \infty$. We have

$$\sum_n P(|B_n| > \epsilon) = \sum_n P\left\{\frac{Q_1-Q_2}{Q_2/(n-p_2)}\frac{1}{n-p_2} > \left(\epsilon n^{(p_2-p_1)/2-1/3}\right)^{2/n} - 1\right\}.$$

For simplicity, let $\alpha = (p_2 - p_1)/2 - 1/3$. Since $Q_2/(n-p_2) \overset{a.s.}{\to} 1$ and $Q_1 - Q_2 = \boldsymbol{Y}^T(\boldsymbol{H}_2 - \boldsymbol{H}_1)\boldsymbol{Y}/\delta \sim \chi^2_{p_2-p_1}(0)$, it suffices to prove

$$\sum_n P\left\{\chi^2_{p_2-p_1} > a_n\right\} < \infty, \tag{42}$$

where

$$\begin{aligned} a_n &= n(\epsilon^{2/n}n^{2\alpha/n} - 1) = n\left(\exp\left\{2n^{-1}\log\epsilon + 2\alpha n^{-1}\log n\right\} - 1\right) \\ &= (2\log\epsilon + 2\alpha\log n)[1 + o(1)]. \end{aligned}$$

There is an $M < \infty$ such that $x^{(p_2-p_1)/2-1}e^{-x/2} < e^{-x/(2+1/4)}$ for $x > M$. If $a_n > M$,

$$P(\chi^2_{p_2-p_1} > a_n) < \int_{a_n}^{\infty} e^{-x/(2+1/4)}dx \approx n^{-2\alpha/(2+1/4)}.$$

Thus, if $p_2 - p_1 \geq 3$, $2\alpha/(2+1/4) > 1$, and (42) holds.

□

A.5. Proof of Theorem 3

As in the proof of Theorem 2, it's sufficient to show

$$BF_{21} = \frac{I_2}{I_1} \lesssim \frac{u_2(n)}{l_1(n)} \to 0, \quad n \to \infty, \tag{43}$$

for suitable lower and upper bounds $l_1(n)$ and $u_2(n)$. Consider I_1 first; bounds for I_2 can be derived similarly.

Suppose first that case (i), $\boldsymbol{\mathcal{M}}_1 \subset \boldsymbol{\mathcal{M}}_2$, holds. Let $\pi(g)$ be any proper prior satisfying condition (a1), and let $\hat{g}_1$ be the maximizer of $(1+g)^{(n-p_1)/2}[1 + g(1-\widetilde{R}_1^2)]^{-n/2}$. It is easy to show that

$$\hat{g}_1 = \max\left\{\frac{(n-p_1)\widetilde{R}_1^2}{p_1(1-\widetilde{R}_1^2)} - 1, 0\right\}. \tag{44}$$

Since R^2 converges to a constant strictly between 0 and 1, $\hat{g}_1 \approx n$.

Then

$$\begin{aligned} I_1 &= \int_0^\infty (1+g)^{n/2-p_1/2}[1+g(1-\widetilde{R}_1^2)]^{-n/2}\pi(g)\,dg \\ &\geq \int_{\hat{g}_1}^\infty \left[\frac{1+g}{1+g(1-\widetilde{R}_1^2)}\right]^{n/2}(1+g)^{-p_1/2}\pi(g)\,dg \\ &\geq \left[\frac{1+\hat{g}_1}{1+\hat{g}_1(1-\widetilde{R}_1^2)}\right]^{n/2}\int_{\hat{g}_1}^\infty (1+g)^{-p_1/2}\pi(g)\,dg \\ &= \left(\frac{n-p_1}{n}\right)^{n/2}(1-\widetilde{R}_1^2)^{-n/2}\int_{\hat{g}_1}^\infty (1+g)^{-p_1/2}\pi(g)\,dg \\ &\geq \left(\frac{n-p_1}{n}\right)^{n/2}(1-\widetilde{R}_1^2)^{-n/2}\int_{\hat{g}_1}^{c_0\hat{g}_1} (1+g)^{-p_1/2}\pi(g)\,dg \\ &\geq \left(\frac{n-p_1}{n}\right)^{n/2}(1-\widetilde{R}_1^2)^{-n/2}[1+c_0\hat{g}_1]^{-p_1/2}\int_{\hat{g}_1}^{c_0\hat{g}_1} \pi(g)\,dg \\ &\approx n^{-p_1/2-k}(1-\widetilde{R}_1^2)^{-n/2} \triangleq l_1(n). \end{aligned} \tag{45}$$

The last "$\approx$" holds since $\hat{g}_1 \approx n$ and condition (a1).

To obtain an upper bound of I_1 under condition (a2), let $\tilde{g}_1$ be the maximizer of $(1+g)^{(n-p_1)/2-k_u}[1+g(1-\widetilde{R}_1^2)]^{-n/2}$. Then

$$\tilde{g}_1 = \max\left\{\frac{(n-p_1-2k_u)\widetilde{R}_1^2}{(p_1+2k_u)(1-\widetilde{R}_1^2)} - 1, 0\right\}, \tag{46}$$

and

$$\begin{aligned} I_1 &= \int_0^\infty (1+g)^{n/2-p_1/2}[1+g(1-\widetilde{R}_1^2)]^{-n/2}\pi(g)\,dg \\ &= \int_0^\infty (1+g)^{n/2-p_1/2-k_u}[1+g(1-\widetilde{R}_1^2)]^{-n/2}(1+g)^{k_u}\pi(g)\,dg \\ &\leq (1+\tilde{g}_1)^{n/2-p_1/2-k_u}[1+\tilde{g}_1(1-\widetilde{R}_1^2)]^{-n/2}\int_0^\infty (1+g)^{k_u}\pi(g)\,dg \\ &= \left(\frac{p_1+2k_u}{n-p_1-2k_u}\right)^{p_1/2+k_u}\left(\frac{1-\widetilde{R}_1^2}{\widetilde{R}_1^2}\right)^{p_1/2+k_u}\left(\frac{n}{n-p_1-2k_u}\right)^{-n/2} \\ &\quad (1-\widetilde{R}_1^2)^{-n/2}\int_0^\infty (1+g)^{k_u}\pi(g)\,dg \\ &\approx n^{-p_1/2-k_u}(1-\widetilde{R}_1^2)^{-n/2} \triangleq u_1(n). \end{aligned} \tag{47}$$

Similarly, with $l_2(n) = n^{-p_2/2-k}(1-\widetilde{R}_2^2)^{-n/2}$ and $u_2(n) = n^{-p_2/2-k_u}(1-\widetilde{R}_2^2)^{-n/2}$, we can get lower and upper bounds for the integral I_2,

$$l_2(n) \lesssim I_2 \lesssim u_2(n). \tag{48}$$

Hence

$$BF_{21} \lesssim \frac{u_2(n)}{l_1(n)} = n^{-(p_2-p_1)/2+k-k_u} \left(\frac{1-\widetilde{R}_1^2}{1-\widetilde{R}_2^2} \right)^{n/2}. \tag{49}$$

The proof of case (i) in Theorem 2 can now be used to show consistency. Note that in case (i), when $k_u \geq k - (p_2 - p_1)/2$, $n^{-(p_2-p_1)/2+k-k_u} \to 0$. Then consistency holds, since the last term in (49) is bounded in probability in this case. Similarly, we can obtain almost sure convergence for the non-nested case and for the nested case if $p_2 - p_1 > 2 + 2(k - k_u)$. If there exists k_u such that $k - k_u < 1/2$, then almost sure convergence holds for the nested case if $p_2 - p_1 \geq 3$. The proof is similar to the proof of Theorem 2.

Finally, consider case (ii), $\boldsymbol{\mathcal{M}}_1 \nsubseteq \boldsymbol{\mathcal{M}}_2$. The proof is similar to case (i) but simpler. Since the last term in (49) goes to 0 exponentially fast, then consistency holds for any power of n in (49). For the upper bound, just take $k_u = 0$ and use the fact that the prior is proper. For the lower bound, k only needs to be fixed and nonnegative. Thus, condition (a1) is all that is needed.

□

A.6. Proof of Corollaries

Consider Corollary 3 first. If we abuse notation slightly and write

$$\pi^S(a_n, c_0 a_n) = \int_{a_n}^{c_0 a_n} \pi^S(g)\, dg,$$

we have $\pi^S(a_n, c_0 a_n) = \pi(a_n/n, c_0 a_n/n) \approx 1$. Thus (a1) holds with $k = 0$. But (a2) is automatic with $k_u = 0$, and Corollary 3 holds. This also demonstrates Corollary 1 for the Zellner-Siow and hyper-g/n priors.

For the hyper-g prior in Corollary 1, $\int_{a_n}^{c_0 a_n} \pi^g(g)\, dg \approx n^{1-a/2}$, so let $k = a/2 - 1$ and take $k_u < a/2 - 1$.

Finally, for the beta-prime priors in Corollary 2, if a and b do not depend on n and $-1 < a < -1/2$, we have $\int_{a_n}^{c_0 a_n} \pi^B(g)\, dg \approx n^{-(a+1)}$, i.e., $k = a + 1 \in (0, 1/2)$. It suffices to let $k_u = 0$. If $-1 < a < -1/2$ and $b \approx n$, then $\int_{a_n}^{c_0 a_n} \pi^B(g)\, dg \approx 1$, i.e., $k = 0$, and one can take $k_u = 0$. Therefore, the beta-prime priors $\pi^B(g)$ satisfy the conditions of Corollary 2, and the corresponding Bayes factors are consistent as $n \to \infty$.

□

References

Alston, C., P. Kuhnert, S. Low Choy, R. McVinish, and K. Mengersen (2005). Bayesian model comparison: Review and discussion. In *International Statistical Institue, 55th Session.*

Bayarri, M. J. and G. García-Donato (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika 94*, 135–152.

Berger, J. O. and J. M. Bernardo (1989). Estimating a product of means: Bayesian analysis with reference priors. *Journal of the American Statistical Association 84*, 200–207.

Berger, J. O. and J. M. Bernardo (1992). On the development of reference priors (Disc: P49-60). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting*, pp. 35–49. Clarendon Press [Oxford University Press].

Berger, J. O. and J. M. Bernardo (1992a). Ordered group reference priors with application to the multinomial problem. *Biometrika 79*, 25–37.

Berger, J. O., J. K. Ghosh, and N. Mukhopadhyay (2003). Approximations and consistency of Bayes factors as model dimension grows. *Journal of Statistical Planning and Inference 112*(1-2), 241–258.

Berger, J. O. and L. R. Pericchi (2001). Objective Bayesian methods for model selection: Introduction and comparison (Pkg: P135-207). In *Model selection [Institute of Mathematical Statistics lecture notes-monograph series 38]*, pp. 135–193. IMS Press.

Bernardo, J. M. (1979). Reply to comments on "Reference posterior distributions for Bayesian inference". *Journal of the Royal Statistical Society, Series B: Methodological 41*, 113–147.

Billingsley, P. (1995). *Probability and Measure*. John Wiley & Sons.

Casella, G. and F. J. Girón (2008). Consistency of bayesian procedures for variable selection. *to appear in the Annals of Statistics*.

Fernández, C., C. Fernandez, E. Ley, and M. F. J. Steel (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics 100*(2), 381–427.

García-Donato, G. and D. Sun (2007). Objective priors for hypothesis testing in one-way random effects models. *The Canadian Journal of Statistics 35*(2), 303–320.

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

Hu, T.-C. and R. L. Taylor (1997). On the strong law for arrays and for the bootstrap mean and variance. *International Journal of Mathematics and Mathematical Sciences 20*(2), 375–382.

Jeffreys, H. (1961). *Theory of Probability*. London: Oxford University Press.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Kass, R. E. and L. Wasserman (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association 90*, 928–934.

Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008). Mixtures of g-priors for bayesian variable selection. *To appear in Journal of the American Statistical Association*.

Marden, J. I. (2000). Hypothesis testing: From p values to Bayes factors. *Journal of the American Statistical Association 95*(452), 1316–1320.

Marin, J.-M. and C. Robert (2007a). Bayesian variable selection in linear regression. http://3w.eco.uniroma1.it/OB07/papers/Marin.pdf. The Sixth Workshop on Objective Bayes Technology, Rome, June 8-12, 2007.

Marin, J.-M. and C. P. Robert (2007b). *Bayesian Core: A Practical Approach to Cmoputational Bayesian Statistics*. Springer-Verlag Inc.

Maruyama, Y. and E. George (2008). A g-prior extension for $p > n$. *submitted, arXiv:0801.4410v1 [stat.ME]*.

Moreno, E. and F. J. Giron (2005). Consistency of bayes factors for intrinsic priors in normal linear models. *Comptes Rendus Mathematique 340*(12), 911–914.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 81*, 82–86.

Westfall, P. H. and M. Gönen (1996). Asymptotic properties of ANOVA Bayes factors. *Communications in Statistics: Theory and Methods 25*, 3101–3123.

Zellner, A. (1984). Posterior odds ratios for regression hypotheses: General considerations and some specific results. In *Basic Issues in Econometrics*, pp. 275–305. Chicago: University of Chicago Press.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pp. 233–243. Elsevier/North-Holland [Elsevier Science Publishing Co., New York; North-Holland Publishing Co., Amsterdam].

Zellner, A. and A. Siow (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*, pp. 585–603. University of Valencia.