Robust Bayesian variable selection with sub-harmonic priors

Yuzo Maruyama and William, E. Strawderman

The University of Tokyo and Rutgers University
e-mail: maruyama@csis.u-tokyo.ac.jp; straw@stat.rutgers.edu

Abstract: This paper studies Bayesian variable selection in linear models with spherically symmetric error distributions. We give a series of proper prior distributions which converge in a certain sense to an improper prior distribution and for which the Bayes factor for each possible sub-model converges to the Bayes factor for the improper prior. This convergence justifies the use of the improper prior in variable selection. We also show that the resulting improper Bayes factors are independent of the particular sampling model when all sub-models are assumed to have the same error distribution. This gives a surprising robustness to the procedure which is analogous to that observed in certain Bayes estimation problems involving spherically symmetric error distributions. We also show that our procedure has model selection consistency as the sample size increases for fixed maximum number of predictors uniformly over the entire class of spherical error distributions. A simulation study indicates that the procedure performs well and stably relative to a BIC based alternative.

AMS 2000 subject classifications: Primary 62F15, 62F07; secondary 62A10.
Keywords and phrases: Bayes factor, Bayesian variable selection, fully Bayes method, model selection consistency, sub-harmonic prior.

1. Introduction

Suppose the linear regression model is used to relate Y to the p potential predictors $x_1, \ldots, x_p$,

$$\boldsymbol{y} = \alpha \mathbf{1}_n + \boldsymbol{X}_F \boldsymbol{\beta}_F + \boldsymbol{\epsilon}_F, \tag{1.1}$$

where the subscript F refers to the full model $\mathcal{M}_F$. In the model (1.1), α is an unknown intercept parameter, $\mathbf{1}_n$ is an $n \times 1$ vector of ones, $\boldsymbol{X}_F = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is an $n \times p$ design matrix, and $\boldsymbol{\beta}_F$ is a $p \times 1$ vector of unknown regression coefficients. The error term $\boldsymbol{\epsilon}_F$ has a spherically symmetric distribution with density $\sigma_F^{-n} f_F(\|\boldsymbol{\epsilon}\|^2/\sigma_F^2)$, which satisfies

$$\int_{\mathcal{R}^n} \frac{f_F(\|\boldsymbol{\epsilon}\|^2/\sigma_F^2)}{\sigma_F^n} d\boldsymbol{\epsilon} = \frac{\pi^{n/2}}{\Gamma(n/2)} \int_0^\infty s^{n/2-1} f_F(s) ds = 1, \tag{1.2}$$

where $\|\boldsymbol{\epsilon}\|$ denotes the Euclidean norm given by $(\epsilon_1^2 + \cdots + \epsilon_n^2)^{1/2}$. In (1.2), σ_F^2 is the variance of ϵ_i and hence $f_F(\cdot)$ satisfies

$$\frac{\pi^{n/2}}{\Gamma(n/2)} \int_0^\infty s^{n/2} f_F(s) ds = n \tag{1.3}$$

as well as (1.2). We assume that the columns of $\boldsymbol{X}_F$ have been standardized so that for $1 \leq i \leq p$, $\boldsymbol{x}_i'\mathbf{1}_n = 0$ and $\boldsymbol{x}_i'\boldsymbol{x}_i/n = 1$.

We shall be particularly interested in the variable selection problem where we would like to select an unknown subset of the effective predictors. It will be convenient throughout to index each of these 2^p possible subset choices by the vector

$$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$$

where $\gamma_i = 0$ or 1. We use $q_\gamma = \boldsymbol{\gamma}'\mathbf{1}_p$ to denote the size of the γth subset. The problem then becomes that of selecting a submodel of (1.1)

$$\boldsymbol{Y} = \alpha\mathbf{1}_n + \boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma + \boldsymbol{\epsilon}_\gamma. \tag{1.4}$$

In (1.4), $\boldsymbol{X}_\gamma$ is the $n \times q_\gamma$ matrix whose columns correspond to the γth subset of $x_1, \ldots, x_p$, $\boldsymbol{\beta}_\gamma$ is a $q_\gamma \times 1$ vector of unknown regression coefficients. Let $\mathcal{M}_\gamma$ denote the submodel given by (1.4). We allow $\boldsymbol{\epsilon}_\gamma$ to be distributed differently from $\boldsymbol{\epsilon}_F$. In particular, the error term $\boldsymbol{\epsilon}_\gamma$ has a spherically symmetric distribution with the probability density $\sigma_\gamma^{-n} f_\gamma(\|\boldsymbol{\epsilon}\|^2/\sigma_\gamma^2)$ satisfying

$$\frac{\pi^{n/2}}{\Gamma(n/2)}\int_0^\infty s^{n/2-1} f_\gamma(s)ds = 1, \quad \frac{\pi^{n/2}}{\Gamma(n/2)}\int_0^\infty s^{n/2} f_\gamma(s)ds = n, \tag{1.5}$$

but f_γ is not necessarily equal to f_F. By (1.5), σ_γ^2 is the variance of the components of $\boldsymbol{\epsilon}_\gamma$. We note that, in almost all earlier studies, error terms $\boldsymbol{\epsilon}_F$ and $\boldsymbol{\epsilon}_\gamma$ in linear models are assumed to have the same Gaussian distribution, that is, $f_F = f_\gamma = f_G$ where

$$f_G(t) = \frac{1}{(2\pi)^{n/2}}\exp(-t/2), \tag{1.6}$$

as in George and Foster (2000) and Liang *et al.* (2008).

In this paper, we assume that $n > p + 1$ (the so called classical setup) and $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p\}$ are linearly independent, which implies that

$$\text{rank } \boldsymbol{X}_F = p, \quad \text{rank } \boldsymbol{X}_\gamma = q_\gamma. \tag{1.7}$$

We also assume that the null model ($q_\gamma = 0$ or $\gamma = (0, \ldots, 0)'$) is not a possible model, that is, the number of possible models is $2^p - 1$, rather than, 2^p. In the following, we will sometimes omit γ in $\boldsymbol{X}_\gamma$, q_γ, and $\boldsymbol{\epsilon}_\gamma$ when its absence should not cause confusion.

A Bayesian approach to this problem entails the specification of prior distributions on the models $\pi_\gamma = \Pr(\mathcal{M}_\gamma)$, and on the parameters $p(\alpha, \beta, \sigma^2)$ of each model. For each such specification, of key interest is the posterior probability of $\mathcal{M}_\gamma$ given y,

$$\Pr(\mathcal{M}_\gamma|y) = \frac{\pi_\gamma m_\gamma(\boldsymbol{y})}{\sum_\gamma \pi_\gamma m_\gamma(\boldsymbol{y})} = \frac{\pi_\gamma \text{BF}[\mathcal{M}_\gamma; \mathcal{M}_F]}{\sum_\gamma \pi_\gamma \text{BF}[\mathcal{M}_\gamma; \mathcal{M}_F]}, \tag{1.8}$$

where $m_\gamma(\boldsymbol{y})$ is the marginal density under $\mathcal{M}_\gamma$. In (1.8), $\text{BF}[\mathcal{M}_\gamma; \mathcal{M}_F]$ is the Bayes factor for comparing each of $\mathcal{M}_\gamma$ to the full model $\mathcal{M}_F$ which is defined

as

$$\mathrm{BF}[\mathcal{M}_\gamma;\mathcal{M}_F]=\frac{m_\gamma(\boldsymbol{y})}{m_F(\boldsymbol{y})},$$

where $m_F(\boldsymbol{y})$ is the marginal density under the full model. In Bayesian model selection, the Bayes factor is often used as a criterion instead of employing the marginal density directly. A popular strategy is to select the model for which $\Pr(\mathcal{M}_\gamma|y)$ or $\pi_\gamma \mathrm{BF}[\mathcal{M}_\gamma;\mathcal{M}_F]$ is largest.

Our main focus in this paper is to propose and study specifications for the prior distribution of the parameters for the submodel $\mathcal{M}_\gamma$. In particular, the joint density we consider has the form

$$p(\alpha,\boldsymbol{\theta},\sigma^2)\propto\{\sigma^2\}^{-a/2-1}\|\boldsymbol{\theta}\|^{-q+a} \tag{1.9}$$

for $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_q)'=(\boldsymbol{X}'\boldsymbol{X})^{1/2}\boldsymbol{\beta}$ and $0<a<1$. Since the term including $\boldsymbol{\theta}$ in the prior above, $\|\boldsymbol{\theta}\|^{-q+a}$ for $0<a<\min(2,q)$, is known as a sub-harmonic function, that is,

$$\sum_{i=1}^{q}\frac{\partial^2}{\partial\theta_i^2}\|\boldsymbol{\theta}\|^{-q+a}>0,$$

we call the prior given by (1.9) a sub-harmonic prior. Fundamentally, a proper prior should be used for the calculation of the marginal density for all models in Bayesian model selection. The validity of the improper prior given by (1.9) will be discussed in Section 3.

The organization of this paper is as follows. In Section 2, we give details of the prior distribution. In Section 3, we show that the Bayes factor with respect to the above prior is given by

$$\mathrm{BF}[\mathcal{M}_\gamma;\mathcal{M}_F|a]=\frac{E[\|\boldsymbol{\epsilon}_\gamma\|^a]}{E[\|\boldsymbol{\epsilon}_F\|^a]}\mathrm{BF}^G[\mathcal{M}_\gamma;\mathcal{M}_F|a] \tag{1.10}$$

where

$$\mathrm{BF}^G[\mathcal{M}_\gamma;\mathcal{M}_F|a]=\frac{\int_0^\infty g^{a/2-1}(1+g)^{(n-q_\gamma-1)/2}\{g(1-R_\gamma^2)+1\}^{-(n-1)/2}\,dg}{\int_0^\infty g^{a/2-1}(1+g)^{(n-p-1)/2}\{g(1-R_F^2)+1\}^{-(n-1)/2}\,dg}, \tag{1.11}$$

for $0<a<1$ (we recommend $a=1/2$ eventually). In (1.11), $\mathrm{BF}^G[\mathcal{M}_\gamma;\mathcal{M}_F|a]$ is the Bayes factor for a Gaussian distribution (1.6) and R_γ^2 and R_F^2 are the coefficient of determination under the submodel $\mathcal{M}_\gamma$ and the full model $\mathcal{M}_F$, respectively. From (1.10), we see that if $f_\gamma=f_F$ the Bayes factor does not depend on the sampling density. Hence, even when there is no specific information about the error distribution of each model (other than spherical symmetry), but we assume they are all the same, it is not necessary to specify the exact form of the sampling density. It suffices to assume they are all Gaussian, that is, $f_\gamma=f_F=f_G$. In the case where f_γ and f_F are assumed to be different, the most typical choice of errors in linear model are probably multivariate-t mainly because the tail behavior can be controlled by a single parameter from

thin (Gaussian) to fat (Cauchy). We will show that even if the error distributions, f_γ and f_F, are multivariate-t with different (but > 3) degrees of freedom $\mathrm{BF}^G[\mathcal{M}_\gamma; \mathcal{M}_F|a]$ with $a = 1/2$ remains a good approximation to $\mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_F|a]$ with $a = 1/2$. As far as we know, in the area of Bayesian variable selection with shrinkage priors, the sampling density has been assumed to be Gaussian and this kind of robustness result has not yet been studied. Originally similar robustness results were derived by Maruyama (2003) and Maruyama and Strawderman (2005) in the problem of estimating regression coefficients with the Stein effect. In Section 4, we approximate the Bayes factor given by (1.10). In Section 5, we show that our Bayes factor $\mathrm{BF}^G[\mathcal{M}_\gamma; \mathcal{M}_F|a]$ has model selection consistency as $n \to \infty$ and p is fixed. In Section 6, we give some numerical results.

2. Prior distributions

In this section, we give a prior joint density of a form

$$p(\alpha, \boldsymbol{\beta}, \sigma^2) = p(\alpha)p(\sigma^2)p(\boldsymbol{\beta}|\sigma^2).$$

Recall that we are suppressing the subscript on $\boldsymbol{\beta}$, σ^2 and q. The natural choice of priors for location (α) and scale (σ^2) are

$$p^I_\alpha(\alpha) = I_{(-\infty,\infty)}(\alpha), \tag{2.1}$$

and

$$p^I_{\sigma^2}(\sigma^2) = (\sigma^2)^{-1} I_{(0,\infty)}(\sigma^2). \tag{2.2}$$

In (2.1) and (2.2), the superscript ^{I} means that the prior density is improper. Since (2.1) and (2.2) have invariance to location and scale transformation, respectively, they are considered by many as non-informative objective priors. The problem is that they are improper and hence determined only up to an arbitrary multiplicative constant. In this paper, the use of improper priors is justified through sequences of proper priors approaching the target improper priors (2.1) and (2.2):

$$p_\alpha(\alpha; h_\alpha) = \frac{1}{2h_\alpha} I_{(-h_\alpha, h_\alpha)}(\alpha) \tag{2.3}$$

where $h_\alpha \to \infty$ and

$$p_{\sigma^2}(\sigma^2; h_\sigma) = \frac{(\sigma^2)^{-1}}{\int_{h_\sigma^{-1}}^{h_\sigma} (\sigma^2)^{-1} d\sigma^2} I_{(h_\sigma^{-1},\ h_\sigma)}(\sigma^2) = \frac{(\sigma^2)^{-1}}{2\log h_\sigma} I_{(h_\sigma^{-1},\ h_\sigma)}(\sigma^2) \tag{2.4}$$

where $h_\sigma \to \infty$. See the beginning of Section 3 for details of the justification.

Next we give a sequence of proper conditional priors of $\boldsymbol{\beta}$ given σ^2, which approach an improper conditional prior of $\boldsymbol{\beta}$ given σ^2:

$$p_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta}|\sigma^2; h_g) = \left\{ \frac{a/2}{h_g^{a/2}} \right\} \int_0^{h_g} g^{a/2-1} \frac{|\boldsymbol{X}'\boldsymbol{X}|^{1/2}}{(2\pi\sigma^2 g)^{q/2}} \exp\left(-\frac{\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2 g} \right) dg$$

for $0 < a < q$ and $h_g > 0$. Eventually we will recommend setting $a = 1/2$, the midpoint of $(0, 1)$, because we will have reason to choose the same a for all submodels. Since this prior has the hierarchical structure

$$\boldsymbol{\beta}|\{\sigma^2 g\} \sim N_q(0, g\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}), \quad p_g(g; h_g) = \frac{a/2}{h_g^{a/2}} g^{a/2-1} I_{[0,h_g]}(g),$$

it can be interpreted a scale mixture of Zellner's g-priors. Similar priors have been considered by Liang *et al.* (2008) and Maruyama and George (2008) under the normal linear regression setup. For any fixed $h_g > 0$, the prior $p_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta}|\sigma^2; h_g)$ is clearly a proper probability density, that is,

$$\int_{\mathcal{R}^q} p_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta}|\sigma^2; h_g) d\boldsymbol{\beta} = 1.$$

As $h_g \to \infty$, the limit of a variant of $p_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta}|\sigma^2; h_g)$ is given by

$$\begin{aligned} p^I_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta}|\sigma^2) &= \lim_{h_g \to \infty} \left\{ \frac{h_g^{a/2}}{a/2} \right\} p_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta}|\sigma^2; h_g) \\ &= \int_0^\infty g^{a/2-1} \frac{|\boldsymbol{X}'\boldsymbol{X}|^{1/2}}{(2\pi)^{q/2} g^{q/2} \sigma^q} \exp\left(-\frac{\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2 g} \right) dg \\ &= \frac{\Gamma(\{q-a\}/2)}{2^{a/2}\pi^{q/2}} |\boldsymbol{X}'\boldsymbol{X}|^{1/2} (\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta})^{-(q-a)/2} \{\sigma^2\}^{-a/2}. \end{aligned}$$

In summary, the proper prior joint density, which we will use in this paper, is given by

$$p(\alpha, \boldsymbol{\beta}, \sigma^2; h_\alpha, h_g, h_\sigma) = p_\alpha(\alpha; h_\alpha) p_{\sigma^2}(\sigma^2; h_\sigma) p_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta}|\sigma^2; h_g) \tag{2.5}$$

which clearly satisfies

$$\int_{-\infty}^{\infty} \int_{\mathcal{R}^q} \int_0^\infty p(\alpha, \boldsymbol{\beta}, \sigma^2; h_\alpha, h_g, h_\sigma) d\alpha d\boldsymbol{\beta} d\sigma^2 = 1,$$

for any fixed h_α, h_g and h_σ. In Section 3, we will also use the improper joint density of α, $\boldsymbol{\beta}$ and σ^2 given by

$$\begin{aligned} &p^I(\alpha, \boldsymbol{\beta}, \sigma^2) = p^I_\alpha(\alpha) p^I_{\sigma^2}(\sigma^2) p^I_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta}|\sigma^2) \\ &= \{8/a\} \lim_{h_\alpha \to \infty} \lim_{h_\sigma \to \infty} \lim_{h_g \to \infty} h_\alpha \log h_\sigma h_g^{a/2} p(\alpha, \boldsymbol{\beta}, \sigma^2; h_\alpha, h_g, h_\sigma) \\ &= \frac{\Gamma(\{q-a\}/2)}{2^{a/2}\pi^{q/2}} |\boldsymbol{X}'\boldsymbol{X}|^{1/2} (\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta})^{-(q-a)/2} \{\sigma^2\}^{-a/2-1}. \end{aligned} \tag{2.6}$$

In this presentation of the improper joint density $p^I(\alpha, \boldsymbol{\beta}, \sigma^2)$, two facts,

1. $(\alpha, \boldsymbol{\beta})$ and σ^2 are separable,
2. the part depending on σ^2 is given by the power function $\{\sigma^2\}^{-a/2-1}$,

will be the key for calculating the marginal density in the next section.

If, in the above joint prior for $(\alpha, \boldsymbol{\beta}, \sigma^2)$, we make the change of variables, $\boldsymbol{\theta} = (\boldsymbol{X}'\boldsymbol{X})^{1/2}\boldsymbol{\beta}$, the joint prior of $(\alpha, \boldsymbol{\theta}, \sigma^2)$ becomes

$$p^I(\alpha, \boldsymbol{\theta}, \sigma^2) = \frac{\Gamma(\{q-a\}/2)}{2^{a/2}\pi^{q/2}} \|\boldsymbol{\theta}\|^{-(q-a)} \{\sigma^2\}^{-a/2-1}. \tag{2.7}$$

As noted in the introduction, the part depending on $\boldsymbol{\theta}$, $\|\boldsymbol{\theta}\|^{-(q-a)}$ for $0 < a < \min(2, q)$, is known as a subharmonic function, that is,

$$\sum_{i=1}^{q} \frac{\partial^2}{\partial \theta_i^2} \|\boldsymbol{\theta}\|^{-(q-a)} = (q-a)(2-a)\|\boldsymbol{\theta}\|^{-(q-a)-2} > 0.$$

3. Marginal density and Bayes factor

In this section we derive the marginal density under each submodel and the Bayes factor for comparing each $\mathcal{M}_\gamma$ to the full model $\mathcal{M}_F$. The marginal density of $\boldsymbol{y}$ under $\mathcal{M}_\gamma$, is given by

$$\begin{aligned} m_\gamma(\boldsymbol{y}; h_\alpha, h_g, h_\sigma) = \int_{-\infty}^{\infty} \int_{R^q} \int_0^{\infty} \frac{1}{\sigma^n} f_\gamma(\|\boldsymbol{y} - \alpha \mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta}\|^2/\sigma^2) \\ \times p(\alpha, \boldsymbol{\beta}, \sigma^2; h_\alpha, h_g, h_\sigma) d\alpha \, d\boldsymbol{\beta} \, d\sigma^2, \end{aligned} \tag{3.1}$$

where the proper joint prior $p(\alpha, \boldsymbol{\beta}, \sigma^2; h_\alpha, h_g, h_\sigma)$ is given by (2.5). In (3.1), h_α, h_g, and h_σ do not depend on the submodel, but are the same in all models. In the following, instead of $m_\gamma(\boldsymbol{y})$ directly, we consider the limit of a variant of $m_\gamma(\boldsymbol{y})$,

$$\begin{aligned} & M_\gamma(\boldsymbol{y}) \\ & = \{8/a\} \lim_{h_\alpha \to \infty} \lim_{h_\sigma \to \infty} \lim_{h_g \to \infty} h_\alpha \log h_\sigma h_g^{a/2} m_\gamma(\boldsymbol{y}; h_\alpha, h_g, h_\sigma) \\ & = \int_{-\infty}^{\infty} \int_{R^q} \int_0^{\infty} \frac{1}{\sigma^n} f_\gamma \left(\frac{\|\boldsymbol{y} - \alpha \mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta}\|^2}{\sigma^2} \right) p^I(\alpha, \boldsymbol{\beta}, \sigma^2) d\alpha \, d\boldsymbol{\beta} \, d\sigma^2, \end{aligned} \tag{3.2}$$

which is the marginal density with respect to the improper joint prior $p^I(\alpha, \boldsymbol{\beta}, \sigma^2)$ given by (2.6). The second equality in (3.2) follows from the monotone convergence theorem. We choose the same a in all submodels, and thus the Bayes factor $m_\gamma(\boldsymbol{y}; h_\alpha, h_g, h_\sigma)/m_F(\boldsymbol{y}; h_\alpha, h_g, h_\sigma)$ approaches $M_\gamma(\boldsymbol{y})/M_F(\boldsymbol{y})$ as $h_\alpha \to \infty$, $h_g \to \infty$ and $h_\sigma \to \infty$. Hence the use of the improper joint prior is justified as long as $M_\gamma(\boldsymbol{y})/M_F(\boldsymbol{y})$ is well-defined. As remarked in Section 1, the null-model is not a possible model. Since there is no $\boldsymbol{\beta}$ and hence no h_g, $M_N(\boldsymbol{y})/M_F(\boldsymbol{y})$ is not well-defined.

Let $M_\gamma^G(\boldsymbol{y})$ be the marginal density under $\mathcal{M}_\gamma$ with Gaussian errors $\boldsymbol{\epsilon}_G$, i.e., $f_\gamma = f_G$ where f_G is given by (1.6). Before proceeding with the entire calculation of the marginal density, $M_\gamma(\boldsymbol{y})$, we will provide a relationship between $M_\gamma(\boldsymbol{y})$ and $M_\gamma^G(\boldsymbol{y})$ as follows.

Lemma 3.1. *Let a be between 0 and q_γ. Assume the existence of $E[\|\boldsymbol{\epsilon}_\gamma\|^a]$. Then*

$$M_\gamma(\boldsymbol{y}) = \frac{E[\|\boldsymbol{\epsilon}_\gamma\|^a]}{E[\|\boldsymbol{\epsilon}_G\|^a]} M_\gamma^G(\boldsymbol{y}). \tag{3.3}$$

Proof. See Appendix. □

Hence $M_\gamma(\boldsymbol{y})$ depends on the error distribution $\boldsymbol{\epsilon}_\gamma$ only through the a-th moment of $\boldsymbol{\epsilon}_\gamma$, $E[\|\boldsymbol{\epsilon}_\gamma\|^a]$. Using some techniques from Strawderman (1971) and Liang *et al.* (2008), we have the following result concerning $M_\gamma^G(\boldsymbol{y})$.

Lemma 3.2. *Let a be between 0 and q_γ. Then*

$$M_\gamma^G(\boldsymbol{y}) = \frac{n^{1/2}\Gamma(\{n-1\}/2)}{\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^{n-1}\pi^{(n-1)/2}} \int_0^\infty \frac{g^{a/2-1}(1+g)^{(n-q_\gamma-1)/2}}{\{g(1-R_\gamma^2)+1\}^{(n-1)/2}}\, dg, \tag{3.4}$$

where R_γ^2 is the coefficient of determination under the submodel $\mathcal{M}_\gamma$.

Proof. See Appendix. □

Combining Lemmas 3.1 and 3.2, we have the main result of this paper.

Theorem 3.1. *Let a be between 0 and q_γ. Assume that the proper joint prior density of $(\alpha, \boldsymbol{\beta}_\gamma, \sigma_\gamma^2)$ is given by (2.5) with parameters $h_\alpha > 0$, $h_\sigma > 0$, $h_g > 0$. Assume also $E[\|\boldsymbol{\epsilon}_\gamma\|^a] < \infty$. Then the limit of the Bayes factor for comparing each of $\mathcal{M}_\gamma$ to the full model $\mathcal{M}_F$ is given by*

$$\begin{aligned}\mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_F|a] &= \lim_{h_\alpha\to\infty}\lim_{h_g\to\infty}\lim_{h_\sigma\to\infty} \frac{m_\gamma(\boldsymbol{y}; h_\alpha, h_g, h_\sigma)}{m_F(\boldsymbol{y}; h_\alpha, h_g, h_\sigma)} \\ &= \frac{E[\|\boldsymbol{\epsilon}_\gamma\|^a]}{E[\|\boldsymbol{\epsilon}_F\|^a]} \mathrm{BF}^G[\mathcal{M}_\gamma; \mathcal{M}_F|a]\end{aligned}$$

where

$$\mathrm{BF}^G[\mathcal{M}_\gamma; \mathcal{M}_F|a] = \frac{\int_0^\infty g^{a/2-1}(1+g)^{(n-q_\gamma-1)/2}\{g(1-R_\gamma^2)+1\}^{-(n-1)/2}\, dg}{\int_0^\infty g^{a/2-1}(1+g)^{(n-p-1)/2}\{g(1-R_F^2)+1\}^{-(n-1)/2}\, dg}. \tag{3.5}$$

At this point, a has not been fixed, but has to be in the open interval $(0,1)$ in order that all Bayes factors are well defined. As the default choice of a, we recommend the midpoint

$$a_* = 1/2. \tag{3.6}$$

In Sub-Section 3.2 below, we will consider the "correction term", $E[\|\boldsymbol{\epsilon}_\gamma\|^a]/E[\|\boldsymbol{\epsilon}_F\|^a]$ for $a = 1/2$ and will demonstrate $E[\|\boldsymbol{\epsilon}_\gamma\|^a]/E[\|\boldsymbol{\epsilon}_F\|^a]$ is negligible in many cases. In Section 4, we will approximate $\mathrm{BF}^G[\mathcal{M}_\gamma; \mathcal{M}_F|a]$ by the Laplace method.

Remark 3.1. Expression (3.5) again shows why the null model is not allowed as a possibility. For the null model $\mathcal{M}_N$, $R_N^2 = 0$ so the numerator of (3.5) is infinite, and hence so would be $\mathrm{BF}^G[\mathcal{M}_N; \mathcal{M}_F|a]$. This situation may be

avoided at a slight cost in complexity and in interpretability of the expressions. The required alteration in the prior distributions (proper and improper) is to treat the intercept parameter α as another $\boldsymbol{\beta}$, (and not give it a "uniform" prior). This results in replacing the improper prior in (2.6) by

$$p^I(\alpha,\boldsymbol{\beta},\sigma^2)=\frac{\Gamma(\{q+1-a\}/2)}{2^{a/2}\pi^{q/2+1/2}}|\check{\boldsymbol{X}}'\check{\boldsymbol{X}}|^{1/2}(\check{\boldsymbol{\beta}}'\check{\boldsymbol{X}}'\check{\boldsymbol{X}}\check{\boldsymbol{\beta}})^{-(q+1-a)/2}\{\sigma^2\}^{-a/2-1},$$

where $\check{\boldsymbol{\beta}}=(\alpha,\boldsymbol{\beta}')'$ and $\check{\boldsymbol{X}}=(\mathbf{1}_n|\boldsymbol{X})$. Similarly the marginal distribution in (3.4) and the Bayes factor given by (3.5) are replaced by

$$\check{M}_\gamma^G(\boldsymbol{y})=\frac{\Gamma(n/2)}{\|\boldsymbol{y}\|^n\pi^{n/2}}\int_0^\infty\frac{g^{a/2-1}(1+g)^{(n-q_\gamma-1)/2}}{\{g(1-\check{R}^2_\gamma)+1\}^{n/2}}\,dg,$$

and

$$\check{\mathrm{BF}}^G[\mathcal{M}_\gamma;\mathcal{M}_F|a]=\frac{\int_0^\infty g^{a/2-1}(1+g)^{(n-q_\gamma-1)/2}\{g(1-\check{R}^2_\gamma)+1\}^{-n/2}\,dg}{\int_0^\infty g^{a/2-1}(1+g)^{(n-p-1)/2}\{g(1-\check{R}^2_F)+1\}^{-n/2}\,dg},$$

where $\check{R}^2_\gamma=1-\mathrm{RSS}_\gamma/\|\boldsymbol{y}\|^2$, (the "coefficient of determination" of the model $\mathcal{M}_\gamma$ relative to the 0-intercept model). Hence with the substitution $R^2_\gamma\to\check{R}^2_\gamma$, $n-1\to n$, $q_\gamma\to q_\gamma+1$, $\boldsymbol{y}-\bar{y}\mathbf{1}_n\to\boldsymbol{y}$, all expressions and results in the paper remains valid and the result (Theorem 5.1) on model selection consistency in Section 5 holds also for the null model. Clearly $\check{R}^2_\gamma$ is unusual, but if model selection consistency under the null-model is desirable, we can use $\check{\mathrm{BF}}^G[\mathcal{M}_\gamma;\mathcal{M}_F|a]$.

3.1. BIC under spherically symmetric error distributions

BIC (Schwarz (1978)) is a popular criterion for model selection. We will show in this subsection that BIC has a similar distributional robustness property to the above Bayes model selection procedure. In Section 4, we will develop a Laplace approximation to our robust Bayes factors which relates them to BIC. In Section 5 we will show that both the BIC and robust Bayes model selection procedures are consistent.

BIC for the model $\mathcal{M}_\gamma$ is defined as

$$\mathrm{BIC}=-2\ln\left\{\max_{\alpha,\boldsymbol{\beta}_\gamma,\sigma^2}\frac{1}{\sigma^n}f_\gamma\left(\frac{\|\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma\|^2}{\sigma^2}\right)n^{-q/2}\right\},\tag{3.7}$$

and is derived by eliminating $O(1)$ terms from the approximate marginal densities. Here we denote the function (3.7) by $M_\gamma(\boldsymbol{y}|\mathrm{BIC})$. In general, the maximization with respect to unknown parameters in (3.7) is not always tractable. However when f_γ is decreasing (i.e. $\boldsymbol{\epsilon}_\gamma$ has a unimodal spherically symmetric distribution), the maximization is achieved by $\hat{\alpha}=\bar{y}$, $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, and

$$\hat{\sigma}^2=c_\gamma\|\boldsymbol{y}-\hat{\alpha}\mathbf{1}_n-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2=c_\gamma\|\boldsymbol{y}-\bar{y}\mathbf{1}_n\|^2(1-R^2_\gamma)\tag{3.8}$$

where c_γ is the sole solution of

$$n/2 + cf'_\gamma(c)/f_\gamma(c) = 0. \tag{3.9}$$

Hence $M_\gamma(\boldsymbol{y}|\mathrm{BIC})$ may be expressed as

$$M_\gamma(\boldsymbol{y}|\mathrm{BIC}) = \frac{c_\gamma^{-n/2} f_\gamma(c_\gamma)}{c_G^{-n/2} f_G(c_G)} M_\gamma^G(\boldsymbol{y}|\mathrm{BIC}) \tag{3.10}$$

where $M_\gamma^G(\boldsymbol{y}|\mathrm{BIC})$ is $M_\gamma(\boldsymbol{y}|\mathrm{BIC})$ with the Gaussian error, specifically

$$\begin{aligned} M_\gamma^G(\boldsymbol{y}|\mathrm{BIC}) &= c_G^{-n/2} f_G(c_G)\{\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2(1-R_\gamma^2)\}^{-n/2} n^{-q/2} \\ &= n^{-n/2} f_G(n)\{\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2(1-R_\gamma^2)\}^{-n/2} n^{-q/2} \end{aligned} \tag{3.11}$$

(since c_G is given by n). Clearly (3.10) and (3.11) correspond to (3.3) and (3.4), respectively. The Bayes factor based on the BIC is given by

$$\mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_F|\mathrm{BIC}] = \frac{M_\gamma(\boldsymbol{y}|\mathrm{BIC})}{M_F(\boldsymbol{y}|\mathrm{BIC})} = \frac{c_\gamma^{-n/2} f_\gamma(c_\gamma)}{c_F^{-n/2} f_F(c_F)} \mathrm{BF}^G[\mathcal{M}_\gamma; \mathcal{M}_F|\mathrm{BIC}] \tag{3.12}$$

where

$$\mathrm{BF}^G[\mathcal{M}_\gamma; \mathcal{M}_F|\mathrm{BIC}] = \frac{(1-R_\gamma^2)^{-n/2} n^{-q/2}}{(1-R_F^2)^{-n/2} n^{-p/2}}, \tag{3.13}$$

which corresponds to (3.5).

Hence the Bayes factor based on BIC is also independent of the sampling distribution $f_\gamma(\cdot)$ provided the unimodal error density $f_\gamma(\cdot)$ is the same for all models (c.f. Theorem 3.1).

3.2. Correction terms for Bayes factor

In this subsection, we will consider the correction terms, $E[\|\boldsymbol{\epsilon}_\gamma\|^a]/E[\|\boldsymbol{\epsilon}_F\|^a]$ in (3.5) and $\{c_\gamma^{-n/2} f_\gamma(c_\gamma)\}/\{c_F^{-n/2} f_F(c_F)\}$ in (3.12).

Case I $\boldsymbol{\epsilon}_\gamma$ and $\boldsymbol{\epsilon}_F$ have the same arbitrary distribution

Clearly both correction terms become 1 and hence we have

$$\mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_F|a] = \mathrm{BF}^G[\mathcal{M}_\gamma; \mathcal{M}_F|a]$$

and

$$\mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_F|\mathrm{BIC}] = \mathrm{BF}^G[\mathcal{M}_\gamma; \mathcal{M}_F|\mathrm{BIC}],$$

with unimodal $\boldsymbol{\epsilon}_\gamma$ and $\boldsymbol{\epsilon}_F$. Hence, even when there is no specific information about the error distribution of each model (other than spherical symmetry), but we assume they are all the same, it is not necessary to specify the exact form of the sampling density. It suffices to assume they are all Gaussian, that is, $f_\gamma = f_F = f_G$. As far as we know, in the area

of Bayesian variable selection with shrinkage priors, the sampling density has been assumed to be Gaussian and this kind of robustness results has not yet been studied. Similar robustness results have been derived by Maruyama (2003) and Maruyama and Strawderman (2005) in the problem of estimating regression coefficients with the Stein effect.

Case II $\boldsymbol{\epsilon}_\gamma$ and $\boldsymbol{\epsilon}_F$ are distributed differently

In the case where f_γ and f_F are assumed to be different, the most typical choice of error distribution in linear models is probably a multivariate-t, mainly because the tail behavior can be controlled by a single parameter from thin (Gaussian) to fat (Cauchy).

Lemma 3.3. *Let* $\boldsymbol{\epsilon}_\gamma$ *have a multivariate-t with* m *degrees of freedom, with the density given by*

$$\frac{\Gamma(\{m+n\}/2)}{\pi^{n/2}m^{n/2}\Gamma(m/2)}(1+\|\boldsymbol{\epsilon}^2\|/m)^{-\{m+n\}/2}. \tag{3.14}$$

Then

$$\frac{E[\|\boldsymbol{\epsilon}_\gamma\|^a]}{E[\|\boldsymbol{\epsilon}_G\|^a]} = \Big(\frac{m}{2}\Big)^{a/2}\frac{\Gamma(\{m-a\}/2)}{\Gamma(m/2)} \equiv g(m,a)$$

and

$$\frac{c_\gamma^{-n/2}f_\gamma(c_\gamma)}{c_G^{-n/2}f_G(c_G)} = \frac{\Gamma(\{m+n\}/2)}{\{(m+n)/2\}^{(m+n)/2}}\frac{\{m/2\}^{m/2}}{\Gamma(m/2)}e^{n/2} \equiv h(m,n).$$

Proof. See Appendix. □

From the properties of the Gamma function, $g(m,a)$ is decreasing in m, for example in the case $a=1/2$, $g(m,a)$ varies from $g(3,0.5)\approx 1.132$ to $g(\infty,0.5)=1$. On the other hand, $h(m,n)$ for any fixed n is increasing in m, for example, in the case $n=30$, $h(m,n)$ varies from $h(3,30)\approx 0.287$ to $h(\infty,30)=1$.

The next result uses the above lemma to give bounds on the ratios of Bayes factors for the case of t-distributions with at least three degrees of freedom (so that the second moments exist) and for default choice of $a=1/2$.

Theorem 3.2. *Let* $\boldsymbol{\epsilon}_\gamma$ *and* $\boldsymbol{\epsilon}_F$ *have multivariate-t distributions with different degrees of freedom, each of which is greater than* 3, *and* $a=1/2$. *Then*

$$C^{-1} < \frac{\mathrm{BF}[\mathcal{M}_\gamma;\mathcal{M}_F|1/2]}{\mathrm{BF}^G[\mathcal{M}_\gamma;\mathcal{M}_F|1/2]} < C \tag{3.15}$$

where $C\approx 1.132$ *and*

$$C_{BIC}^{-1}(n) < \frac{\mathrm{BF}[\mathcal{M}_\gamma;\mathcal{M}_F|\mathrm{BIC}]}{\mathrm{BF}^G[\mathcal{M}_\gamma;\mathcal{M}_F|\mathrm{BIC}]} < C_{BIC}(n) \tag{3.16}$$

where $C_{BIC}(30)\approx 3.486$, $C_{BIC}(50)\approx 4.426$ *and something.*

Hence $\mathrm{BF}^G[\mathcal{M}_\gamma;\mathcal{M}_F|a]$ with $a=1/2$ remains a good approximation to $\mathrm{BF}[\mathcal{M}_\gamma;\mathcal{M}_F|a]$ with $a=1/2$ even if the error distributions, f_γ and f_F, are multivariate-t with different (but >3) degrees of freedom. On the other hand, the correction term for BIC is not negligible in this case.

4. The Laplace approximation

In this section, we will approximate the function $\mathrm{BF}^G[\mathcal{M}_\gamma;\mathcal{M}_F]$ by the so-called Laplace approximation. First we provide a summary of Laplace approximations to the integral based on Tierney and Kadane (1986). For integrals of the form

$$\int_{-\infty}^{\infty}\exp(h(\tau,n))d\tau,$$

we make the use of the fully exponential Laplace approximation, based on expanding a smooth unimodal function $h(\tau,n)$ in a Taylor series expansion about $\hat{\tau}$, the mode of $h(\tau,n)$. The Laplace approximation is given by

$$\lim_{n\to\infty}\frac{\int_{-\infty}^{\infty}\exp(h(\tau,n))d\tau}{(2\pi)^{1/2}\hat{\sigma}_h\exp(h(\hat{\tau},n))}=1 \tag{4.1}$$

where

$$\hat{\sigma}_h=\left\{-\frac{\partial^2 h(\tau,n)}{\partial\tau^2}\Big|_{\tau=\hat{\tau}}\right\}^{-1/2}.$$

In the following, we will use the symbol $f(n)\approx g(n)$ $(n\to\infty)$ if

$$\lim_{n\to\infty}\frac{f(n)}{g(n)}=1. \tag{4.2}$$

Hence the approximation given by (4.1) is written as

$$\int_{-\infty}^{\infty}\exp(h(\tau,n))d\tau\approx(2\pi)^{1/2}\hat{\sigma}_h\exp(h(\hat{\tau},n)),\ (n\to\infty). \tag{4.3}$$

The next result gives approximations of the Bayes factor (3.5) in terms of the Bayes factor based on BIC given in (3.13).

Proposition 4.1.

$$\begin{aligned}&\mathrm{BF}^G[\mathcal{M}_\gamma;\mathcal{M}_F|a]\\ &\approx\left\{\frac{(q_\gamma-a)^{q_\gamma-a-1}(1-R_\gamma^2)^{-n+1+q_\gamma-a}\{nR_\gamma^2e\}^{a-q_\gamma}}{(p-a)^{p-a-1}(1-R_F^2)^{-n+1+p-a}\{nR_F^2e\}^{a-p}}\right\}^{1/2}\\ &=\left\{\frac{c(q_\gamma-a,R_\gamma^2)}{c(p-a,R_F^2)}\right\}^{1/2}\mathrm{BF}^G[\mathcal{M}_\gamma;\mathcal{M}_F|\mathrm{BIC}]\end{aligned} \tag{4.4}$$

where $c(s,R^2)=s^{s-1}(1-R^2)^{s+1}\{eR^2\}^{-s}$ and $\mathrm{BF}^G[\mathcal{M}_\gamma;\mathcal{M}_F|\mathrm{BIC}]$ is given by (3.13).

Proof. See Appendix. □

5. Model selection consistency

In this section, we consider model selection consistency in the case where p is fixed and as n approaches infinity. Let $\mathcal{M}_T$ be the true model,

$$\boldsymbol{y} = \alpha_T \mathbf{1}_n + \boldsymbol{X}_T \boldsymbol{\beta}_T + \boldsymbol{\epsilon}.$$

The consistency for model choice is defined as

$$\operatorname*{plim}_{n\to\infty} \Pr(\mathcal{M}_T|\boldsymbol{y}) = 1,$$

where plim denotes convergence in probability and the probability distribution is the sampling distribution under the true model $\mathcal{M}_T$. We will show that our criterion of general form, $\mathrm{BF}^G[\mathcal{M}_\gamma; \mathcal{M}_F|a]$ given by (3.5), has a model selection consistency. The consistency property is clearly equivalent to

$$\operatorname*{plim}_{n\to\infty} \frac{\mathrm{BF}^G[\mathcal{M}_\gamma; \mathcal{M}_F|a]}{\mathrm{BF}^G[\mathcal{M}_T; \mathcal{M}_F|a]} = 0 \quad \forall \mathcal{M}_\gamma \neq \mathcal{M}_T. \tag{5.1}$$

For the model selection consistency, we assume as follows.

A1. $U_n = \|\boldsymbol{\epsilon}\|^2/\{n\sigma^2\}$ is bounded in probability from below and from above, that is, for any $c > 0$ and any positive integer n, there exists an M such that
$$\Pr\left(M^{-1} < U_n < M\right) > 1 - c.$$

A2. The correlation coefficient of x_i and x_j, $\boldsymbol{x}_i'\boldsymbol{x}_j/n$, for any $i \neq j$ has a limit as $n \to \infty$.

A3. The limit of the correlation matrix of $x_1, \ldots, x_p$, $\lim_{n\to\infty} \boldsymbol{X}_F'\boldsymbol{X}_F/n$, is positive definite.

A1 seems more general than necessary. It appears that, by the law of large numbers, U_n ought to converge to 1 in probability, but this is not necessarily true if the error distribution is not Gaussian. In the case of a scale mixture of Gaussian, U_n approaches, in law, a random variable g which has the distribution of the mixing variable of the variance. Even when the error distribution is not a scale mixture of normals, A1 appears to be a reasonable and minimal assumption. A2 is the standard assumption which also appears in Knight and Fu (2000) and Zou (2006). A3 is natural because the columns of X_F are assumed to be linear independent. Under these mild assumptions, we have following preliminary results for proving the consistency.

Lemma 5.1. *Assume A1, A2 and A3.*

1. *For any $k > 0$ and any positive integer n, there exists a $c_1(\gamma, k) > 1$ such that*
$$\Pr\left(\frac{1}{c_1(\gamma,k)} < R_\gamma^2 < 1 - \frac{1}{c_1(\gamma,k)}\right) > 1 - k. \tag{5.2}$$

2. *Let $\gamma \supsetneq T$. Then $(1-R_T^2)/(1-R_\gamma^2) \geq 1$. Further for any $k>0$ and any positive integer n, there exists a $c_2(\gamma,T,k)>0$ such that*

$$\Pr\left(1 \leq \left(\frac{1-R_T^2}{1-R_\gamma^2}\right)^n < 1+c_2(\gamma,T,k)\right) > 1-k. \tag{5.3}$$

3. *Let $\gamma \not\supseteq T$. Then for any $k>0$ and any positive integer n, there exists a $c_3(\gamma,T,k)>1$ such that*

$$\Pr\left(\frac{1-R_T^2}{1-R_\gamma^2} < 1-\frac{1}{c_3(\gamma,T,k)}\right) > 1-k. \tag{5.4}$$

Proof. See Appendix. □

The main theorem on the consistency is as follows.

Theorem 5.1. *Assume A1, A2 and A3. Then the Bayes factor* $\mathrm{BF}^G[\mathcal{M}_\gamma;\mathcal{M}_F|a]$ *for* $0<a<1$ *is consistent for model selection.*

Proof. By Lemma 5.1, $c(q_\gamma - a, R_\gamma^2)$ for $0<a<1$ goes to a constant in probability for all models and hence (5.1) is equivalent to

$$\operatorname*{plim}_{n\to\infty} \frac{\mathrm{BF}^G[\mathcal{M}_\gamma;\mathcal{M}_F|\mathrm{BIC}]}{\mathrm{BF}^G[\mathcal{M}_T;\mathcal{M}_F|\mathrm{BIC}]} = \operatorname*{plim}_{n\to\infty} \left\{ n^{q_T-q_\gamma}\left(\frac{1-R_T^2}{1-R_\gamma^2}\right)^n\right\}^{1/2} = 0. \tag{5.5}$$

Consider the following two situations:

1. $\gamma \supsetneq T$: By the lemma, $\{(1-R_T^2)/(1-R_\gamma^2)\}^n$ is bounded in probability. Since $q_\gamma > q_T$, (5.5) is satisfied.
2. $\gamma \not\supseteq T$: By the lemma, $(1-R_T^2)/(1-R_\gamma^2)$ is strictly less than 1 in probability. Hence $\{(1-R_T^2)/(1-R_\gamma^2)\}^n$ converges to zero in probability exponentially fast with respect to n. Therefore, no matter what value $q_T - q_\gamma$ takes, (5.5) is satisfied.

These complete the proof. □

Remark 5.1. The issue of model selection consistency in our setup, is somewhat complicated by the wide choice of possible error distributions. If all errors are normally distributed, then under our assumptions A2 and A3 on the design matrix $\boldsymbol{X}_F$, imply that each R_γ^2 approaches a constant, and that $\|\boldsymbol{\epsilon}\|^2/n \to \sigma^2 < \infty$. If on the other hand, all models are variance mixtures of normals with mixture variance distributed as a positive random variable g, then $\|\boldsymbol{\epsilon}\|^2/n \to g$ a random variable, and R_γ^2 also approaches a random variable which is bounded above and below in probability provided that g is similarly bounded.

In general philosophical terms, it might be better to assume that the sequence of error terms $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)'$ are exchangeable for all n. By De finetti's Theorem, this would imply that the error terms all have a variance mixture of normal distributions. We have chosen a slightly weaken requirement on the sequence of error distributions, namely, that $\|\boldsymbol{\epsilon}\|^2/n$ remains bounded above

and below in probability, which extracts the necessary limiting behavior of the error terms to ensure consistency of model selection. Interestingly, although we attain model selection consistency with these assumptions, it is not necessarily true that $\sigma^2 = \mathrm{var}\epsilon_i = \mathrm{var}g$ is consistently estimated by $\|\boldsymbol{\epsilon}\|^2/n$.

6. Simulation Study

In this section, we compare numerical performance of our $\mathrm{BF}^G[\mathcal{M}_\gamma; \mathcal{M}_F|a]$ with BIC in a small simulation study. We generated 16 possible correlated predictors ($p = 16$) as follows:

$$\overbrace{x_1, x_2}^{\mathrm{cor}=0.5}, \underbrace{x_3, x_4}_{\mathrm{cor}=-0.4}, \overbrace{x_5, x_6}^{\mathrm{cor}=0.3}, \underbrace{x_7, x_8}_{\mathrm{cor}=-0.2}, \overbrace{x_9, x_{10}}^{\mathrm{cor}=0.1} \sim N(0,1)$$
$$x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16} \sim N(0,1).$$

Here "cor" denotes the correlation of two normal random variables. Also (x_1, x_2), (x_3, x_4), (x_5, x_6), (x_7, x_8), (x_9, x_{10}), $x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}$ are assumed to be independent. After generating pseudo random $x_1, \ldots, x_{16}$, we centered and scaled them as noted in Section 1. We assume $n = 30$ and consider 4 cases where the true predictors are

$q_T = 16$	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}$
$q_T = 12$	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}$
$q_T = 8$	$x_1, x_2, \quad x_5, x_6, \quad x_9, x_{10}, x_{11}, x_{12}$
$q_T = 4$	$x_1, x_2, \quad x_5, x_6$

(where q_T denotes the number of true predictors) and the true model is given by

$$\boldsymbol{y} = \mathbf{1}_{30} + 2 \sum_{i \in \{\mathrm{true}\}} \boldsymbol{x}_i + \sigma \times \begin{cases} N_{30}(\mathbf{0}, \boldsymbol{I}_{30}), \\ \text{Multi-}t(\mathbf{0}, \boldsymbol{I}_{30}; 3, 30), \end{cases} \tag{6.1}$$

with $\sigma = 0.5,\ 1,\ 2$. In (6.1), the density of Multi-$t(\mathbf{0}, \boldsymbol{I}_n; m, n)$ is given by (3.14). The Table 1 and 2 show that how often the true model is in the top 3 among $2^{16} - 1$ candidates when the number of replication is $N = 200$. The error distributions are normal (Table 1) and multivariate-t with 3 degrees of freedom (Table 2). For the case of normally distributed errors (See Table 1), the Bayes factor method performed well and stably for $\sigma = 0.5$ and $\sigma = 1$ and did reasonably well for $\sigma = 2$ for the smaller true models ($q_T = 4, 8$). BIC seemed, generally, to have a preference for larger models, and performed much less well than the Bayes factor method for $\sigma = 0.5$ and $\sigma = 1$ for models of smaller size ($q_T = 4, 8, 12$) For $\sigma = 2$, BIC did substantially better than BF for the largest model ($q_T = 16$) and somewhat better for $q_T = 12$.

Interestingly, for the case of a multivariate-t error distribution with 3 degrees of freedom (the minimum so that a variance exists), the numerical results were quite similar to those in the normal case for both BF and BIC, both quantitatively and qualitatively.

TABLE 1
Frequency of the true model (normal error)

q_T	16		12		8		4	
rank	1st	1st-3rd	1st	1st-3rd	1st	1st-3rd	1st	1st-3rd
$\sigma = 0.5$								
BF	1.00	1.00	0.96	1.00	0.90	1.00	0.89	0.98
BIC	1.00	1.00	0.43	0.66	0.31	0.53	0.23	0.44
$\sigma = 1$								
BF	0.82	0.90	0.89	0.99	0.85	0.95	0.80	0.93
BIC	1.00	1.00	0.43	0.66	0.31	0.53	0.23	0.44
$\sigma = 2$								
BF	0.05	0.08	0.22	0.39	0.54	0.74	0.56	0.74
BIC	0.58	0.78	0.31	0.50	0.27	0.47	0.23	0.43

TABLE 2
Frequency of the true model (multi-t with d.f. 3 error)

q_T	16		12		8		4	
rank	1st	1st-3rd	1st	1st-3rd	1st	1st-3rd	1st	1st-3rd
$\sigma = 0.5$								
BF	0.95	0.95	0.96	0.98	0.92	0.99	0.88	0.99
BIC	0.99	0.99	0.46	0.66	0.30	0.48	0.26	0.43
$\sigma = 1$								
BF	0.89	0.93	0.94	0.98	0.90	0.98	0.84	0.98
BIC	0.98	0.99	0.44	0.66	0.29	0.48	0.26	0.43
$\sigma = 2$								
BF	0.13	0.16	0.27	0.39	0.43	0.57	0.42	0.61
BIC	0.44	0.55	0.28	0.42	0.18	0.35	0.17	0.33

Appendix A: Proof of Lemma 3.1

Under the submodel $\mathcal{M}_\gamma$, the conditional marginal density of $\boldsymbol{y}$ with respect to improper prior $(\sigma^2)^{-a/2-1}$ given α and $\boldsymbol{\beta}$ is

$$\begin{aligned}
&\int_0^\infty \sigma^{-n} f_\gamma\left(\frac{\|\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\|^2}{\sigma^2}\right)(\sigma^2)^{-a/2-1}d\sigma^2 \\
&= \|\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\|^{-n-a}\int_0^\infty t^{\{n+a\}/2-1}f_\gamma(t)dt \\
&= \frac{\int_0^\infty t^{(n+a)/2-1}f_\gamma(t)dt}{\int_0^\infty t^{(n+a)/2-1}f_G(t)dt}\int_0^\infty f_G\left(\frac{\|\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\|^2}{\sigma^2}\right)\frac{(\sigma^2)^{-a/2-1}}{\sigma^n}d\sigma^2 \\
&= \frac{E[\|\boldsymbol{\epsilon}_\gamma\|^a]}{E[\|\boldsymbol{\epsilon}_G\|^a]}\int_0^\infty f_G\left(\frac{\|\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\|^2}{\sigma^2}\right)\frac{(\sigma^2)^{-a/2-1}}{\sigma^n}d\sigma^2
\end{aligned} \tag{A.1}$$

where

$$f_G(t)=\frac{1}{(2\pi)^{n/2}}\exp(-t/2)$$

provided

$$\int_0^\infty t^{(n+a)/2-1}f_\gamma(t)dt<\infty \;\Leftrightarrow\; E[\|\boldsymbol{\epsilon}_\gamma\|^a]<\infty. \tag{A.2}$$

Therefore, we have

$$M_\gamma(\boldsymbol{y})=\frac{E[\|\boldsymbol{\epsilon}_\gamma\|^a]}{E[\|\boldsymbol{\epsilon}_G\|^a]}M_\gamma^G(\boldsymbol{y})$$

where

$$\begin{aligned}
M_\gamma^G(\boldsymbol{y})=\int_{-\infty}^\infty\int_{R^q}\int_0^\infty &\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{\|\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\right) \\
&\times p^I(\alpha,\boldsymbol{\beta},\sigma^2)d\alpha\,d\boldsymbol{\beta}\,d\sigma^2.
\end{aligned} \tag{A.3}$$

Appendix B: Proof of Lemma 3.2

As in (A.3), $M_\gamma^G(\boldsymbol{y})$ is given by

$$\begin{aligned}
M_\gamma^G(\boldsymbol{y}) &= \int_{-\infty}^\infty\int_{R^q}\int_0^\infty \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{\|\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\right) \\
&\qquad\times p^I(\alpha,\boldsymbol{\beta},\sigma^2)d\alpha\,d\boldsymbol{\beta}\,d\sigma^2 \\
&= \int_{-\infty}^\infty\int_{R^q}\int_0^\infty\int_0^\infty \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{\|\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\right) \\
&\qquad\times g^{a/2-1}\{\sigma^2\}^{-1}\frac{|\boldsymbol{X}'\boldsymbol{X}|^{1/2}}{(2\pi\sigma^2)^{q/2}g^{q/2}}\exp\left(-\frac{\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2 g}\right)d\alpha\,d\boldsymbol{\beta}\,d\sigma^2\,dg.
\end{aligned} \tag{B.1}$$

In the following, we calculate the integration of $M_\gamma^G(\boldsymbol{y})$ with respect to α, $\boldsymbol{\beta}$, σ^2, and g, in this order.

By the simple relation

$$\boldsymbol{y} - \alpha \mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta} = (-\alpha + \bar{y})\mathbf{1}_n + \boldsymbol{v} - \boldsymbol{X}\boldsymbol{\beta}$$

where $\bar{y}$ is the mean of $\boldsymbol{y}$ and $\boldsymbol{v} = \boldsymbol{y} - \bar{y}\mathbf{1}_n$, we have the Pythagorean relation,

$$\|\boldsymbol{y} - \alpha \mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta}\|^2 = n(-\alpha + \bar{y})^2 + \|\boldsymbol{v} - \boldsymbol{X}\boldsymbol{\beta}\|^2,$$

since $\boldsymbol{X}$ has been already centered. Then we have

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|\boldsymbol{y} - \alpha \mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\right) d\alpha \\
&= \int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{n(\alpha - \bar{y})^2}{2\sigma^2} - \frac{\|\boldsymbol{v} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\right) d\alpha \\
&= \frac{n^{1/2}}{(2\pi\sigma^2)^{(n-1)/2}} \exp\left(-\frac{\|\boldsymbol{v} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\right).
\end{aligned}$$

Next we consider the integration with respect to $\boldsymbol{\beta}$. Note the relation of completing squares with respect to $\boldsymbol{\beta}$

$$\begin{aligned}
&\|\boldsymbol{v} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + g^{-1}\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\
&= \frac{1+g}{g}\left(\boldsymbol{\beta} - \frac{g}{1+g}\hat{\boldsymbol{\beta}}\right)' \boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta} - \frac{g}{1+g}\hat{\boldsymbol{\beta}}\right) - \frac{g}{1+g}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{v}\|^2 \\
&= \frac{1+g}{g}\left(\boldsymbol{\beta} - \frac{g}{1+g}\hat{\boldsymbol{\beta}}\right)' \boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta} - \frac{g}{1+g}\hat{\boldsymbol{\beta}}\right) + \frac{\|\boldsymbol{v}\|^2}{1+g}\left\{g(1-R^2)+1\right\}
\end{aligned}$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{v}$ and $R^2 = \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/\|\boldsymbol{v}\|^2$ is the coefficient of determination under the submodel $\mathcal{M}_\gamma$. Hence we have

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{R^q} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|\boldsymbol{y} - \alpha \mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\right) \\
&\quad \times \frac{|\boldsymbol{X}'\boldsymbol{X}|^{1/2}}{(2\pi\sigma^2)^{q/2} g^{q/2}} \exp\left(-\frac{\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2 g}\right) d\alpha\, d\boldsymbol{\beta} \qquad \text{(B.2)} \\
&\quad = \frac{n^{1/2}(1+g)^{-q/2}}{(2\pi\sigma^2)^{(n-1)/2}} \exp\left(-\frac{\|\boldsymbol{v}\|^2\{g(1-R^2)+1\}}{2\sigma^2(g+1)}\right).
\end{aligned}$$

Then we consider the integration with respect to σ^2. By (B.2), we have

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{R^q}\int_0^{\infty} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|\boldsymbol{y} - \alpha \mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\right) \\
&\quad \times \frac{|\boldsymbol{X}'\boldsymbol{X}|^{1/2}}{(2\pi\sigma^2)^{q/2} g^{q/2}} \exp\left(-\frac{\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2 g}\right) \frac{1}{\sigma^2} d\alpha\, d\boldsymbol{\beta}\, d\sigma^2 \qquad \text{(B.3)} \\
&\quad = \frac{n^{1/2}\Gamma(\{n-1\}/2)}{\pi^{(n-1)/2}\|\boldsymbol{v}\|^{n-1}} (1+g)^{(n-q-1)/2} \left\{g(1-R^2)+1\right\}^{-(n-1)/2}.
\end{aligned}$$

Finally we consider the integration with respect to g. By (B.3) we have

$$M_\gamma^G(\boldsymbol{y}) = \frac{n^{1/2}\Gamma(\{n-1\}/2)}{\pi^{(n-1)/2}\|\boldsymbol{v}\|^{n-1}} \int_0^{\infty} \frac{g^{a/2-1}(1+g)^{(n-q-1)/2}}{\{g(1-R^2)+1\}^{(n-1)/2}}\, dg. \qquad \text{(B.4)}$$

Appendix C: Proof of Lemma 3.3

Since the Jacobian for the change of variables $\|\boldsymbol{\epsilon}\|^2 = t$ is proportional to $t^{n/2-1}$, $E[\|\boldsymbol{\epsilon}_\gamma\|^a]/E[\|\boldsymbol{\epsilon}_G\|^a]$ is written as

$$\begin{aligned}
&\frac{\int t^{n/2+a-1}(1+t/m)^{-\{m+n\}/2}dt}{\int t^{n/2+a-1}(1+t/m)^{-\{m+n\}/2}dt}\frac{\int t^{n/2-1}\exp(-t/2)dt}{\int t^{n/2+a-1}\exp(-t/2)dt}\\
&= m^a\frac{B(n/2+a,m/2-a)}{B(n/2,m/2)}\times 2^{-a}\frac{\Gamma(n/2)}{\Gamma(n/2+a)}\\
&= \{m/2\}^a\frac{\Gamma(m/2-a)}{\Gamma(m/2)}.
\end{aligned}$$

Note c_γ is defined as the solution of (3.9). The solution of

$$\frac{n}{2}+c\frac{\{d/dc\}(1+c/m)^{-(m+n)/2}}{(1+c/m)^{-(m+n)/2}}=0$$

is $c=n$. Since c_G is also n, the last half of the lemma follows.

Appendix D: Proof of Proposition 4.1

Denote the numerator of $\mathrm{BF}^G[\mathcal{M}_\gamma;\mathcal{M}_F|a]$ in (3.5) by $H(n)$. When approximating $H(n)$, make the change of variables $\tau=\log g$. See Liang *et al.* (2008) for details. With this transformation, the integral becomes

$$H(n)=\int_{-\infty}^{\infty}\frac{e^{(a/2-1)\tau}(1+e^\tau)^{-q/2+(n-1)/2}}{\{e^\tau(1-R^2)+1\}^{(n-1)/2}}e^\tau\,d\tau, \tag{D.1}$$

where the extra e^τ comes from the Jacobian of the transformation of variables. Denote the logarithm of the integrand function in (D.1) by $h(\tau,n)$. We have

$$\begin{aligned}
\frac{\partial}{\partial\tau}h(\tau,n)&=\frac{1}{2}\left\{-(q-a)-\frac{n-q-1}{1+z}+\frac{n-1}{1+Az}\right\}\\
\frac{\partial^2}{\partial\tau^2}h(\tau,n)&=\frac{1}{2}\left\{(n-q-1)\frac{z}{(1+z)^2}-(n-1)\frac{Az}{(1+zA)^2}\right\}
\end{aligned}$$

where $z=e^\tau$ and $A=1-R^2$. Setting $\{\partial/\partial\tau\}h(\tau,n)=0$ gives a quadratic equation in $z=e^\tau$:

$$(q-a)Az^2+(A\{n-a-1\}-n+q-a+1)z-a=0.$$

Since $0<a<1$, only one of the roots is positive, $\hat{z}=e^{\hat{\tau}}$, which is given by

$$\begin{aligned}
\hat{z}=&\frac{1}{2(q-a)A}\Big\{-A\{n-a-1\}+n-q+a-1\\
&+\{(A\{n-a-1\}-n+q-a+1)^2+4A(q-a)a\}^{1/2}\Big\}.
\end{aligned}$$

The mode $\hat{\tau} = \log \hat{z}$ satisfies

$$\lim_{n\to\infty} \frac{\hat{z}}{n} = \frac{1-A}{(q-a)A} = \frac{R^2}{(q-a)(1-R^2)}. \tag{D.2}$$

Hence we have

$$\begin{aligned}
&e^{h(\hat{\tau},n)} \\
&= \left\{\hat{z}^a(1+\hat{z})^{n-q-1}(1+A\hat{z})^{-n+1}\right\}^{1/2} \\
&= \left\{\frac{\hat{z}^{-q+a}}{A^{n-1}}\left(1+\frac{\{n/\hat{z}\}}{n}\right)^{n-q-1}\left(1+\frac{\{n/A\hat{z}\}}{n}\right)^{-n+1}\right\}^{1/2} \\
&\approx \left\{\left(\frac{(q-a)A}{n(1-A)}\right)^{q-a} A^{-n+1} e^{\frac{(q-a)A}{1-A}-\frac{(q-a)}{1-A}}\right\}^{1/2} \\
&= \left\{\left(\frac{(q-a)(1-R^2)}{nR^2e}\right)^{q-a}(1-R^2)^{-n+1}\right\}^{1/2},
\end{aligned} \tag{D.3}$$

and

$$\begin{aligned}
\{\partial^2/\partial\tau^2\}h(\tau,n)|_{\tau=\hat{\tau}} &= \frac{1}{2}\left\{(n-q-1)\frac{\hat{z}}{(1+\hat{z})^2} - (n-1)\frac{A\hat{z}}{(1+\hat{z}A)^2}\right\} \\
&= \frac{1}{2}\left\{\frac{n-q-1}{1+\hat{z}}\frac{\hat{z}}{1+\hat{z}} - \frac{n-1}{1+\hat{z}A}\frac{A\hat{z}}{1+\hat{z}A}\right\} \\
&\approx \frac{1}{2}\left\{\frac{A(q-a)}{1-A} - \frac{q-a}{1-A}\right\} \\
&= -(q-a)/2.
\end{aligned} \tag{D.4}$$

Therefore we have

$$\begin{aligned}
H(n) &\approx (2\pi)^{1/2} e^{h(\hat{\tau},n)} \left(\{-\partial^2/\partial\tau^2\}h(\tau,n)|_{\tau=\hat{\tau}}\right)^{-1/2} \\
&\approx \left\{\frac{4\pi}{q-a}\left(\frac{(q-a)(1-R^2)}{nR^2e}\right)^{q-a}(1-R^2)^{-n+1}\right\}^{1/2}
\end{aligned} \tag{D.5}$$

as $n \to \infty$. Hence the proposition follows.

Appendix E: Proof of Lemma 5.1

Let $\mathcal{M}_T$ be the true submodel $\boldsymbol{y} = \alpha_T \mathbf{1}_n + \boldsymbol{X}_T\boldsymbol{\beta}_T + \boldsymbol{\epsilon}$ where $\boldsymbol{X}_T$ is the $n \times q_T$ true design matrix and $\boldsymbol{\beta}_T$ is the true $(q_T \times 1)$ coefficient vector.

For the submodel $\mathcal{M}_\gamma$, $1 - R^2_\gamma$ is given by $\|\boldsymbol{Q}_\gamma(\boldsymbol{y} - \bar{y}\mathbf{1}_n)\|^2/\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2$ with $\boldsymbol{Q}_\gamma = \boldsymbol{I} - \boldsymbol{X}_\gamma(\boldsymbol{X}'_\gamma\boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}'_\gamma$. The numerator and denominator are rewritten as

$$\begin{aligned}
\|\boldsymbol{Q}_\gamma(\boldsymbol{y} - \bar{y}\mathbf{1}_n)\|^2 &= \|\boldsymbol{Q}_\gamma\boldsymbol{X}_T\boldsymbol{\beta}_T + \boldsymbol{Q}_\gamma\check{\boldsymbol{\epsilon}}\|^2 \\
&= \boldsymbol{\beta}'_T\boldsymbol{X}'_T\boldsymbol{Q}_\gamma\boldsymbol{X}_T\boldsymbol{\beta}_T + 2\boldsymbol{\beta}'_T\boldsymbol{X}'_T\boldsymbol{Q}_\gamma\boldsymbol{\epsilon} + \check{\boldsymbol{\epsilon}}'\boldsymbol{Q}_\gamma\check{\boldsymbol{\epsilon}}
\end{aligned} \tag{E.1}$$

where $\check{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon} - \bar{\epsilon}\mathbf{1}_n$ and similarly

$$\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2 = \boldsymbol{\beta}_T' \boldsymbol{X}_T' \boldsymbol{X}_T \boldsymbol{\beta}_T + 2\boldsymbol{\beta}_T' \boldsymbol{X}_T' \boldsymbol{\epsilon} + \|\check{\boldsymbol{\epsilon}}\|^2.$$

Since $\check{\boldsymbol{\epsilon}}' \boldsymbol{Q}_\gamma \check{\boldsymbol{\epsilon}} \le \|\check{\boldsymbol{\epsilon}}\|^2$, $1 - R_\gamma^2$ is bounded as

$$\begin{aligned} &\frac{\boldsymbol{\beta}_T' \{\boldsymbol{X}_T' \boldsymbol{Q}_\gamma \boldsymbol{X}_T/n\} \boldsymbol{\beta}_T + 2\boldsymbol{\beta}_T' \{\boldsymbol{X}_T' \boldsymbol{Q}_\gamma \boldsymbol{\epsilon}/n\} + \sigma^2 W_\gamma V_n}{\boldsymbol{\beta}_T' \{\boldsymbol{X}_T' \boldsymbol{X}_T/n\} \boldsymbol{\beta}_T + 2\boldsymbol{\beta}_T' \{\boldsymbol{X}_T' \boldsymbol{\epsilon}/n\} + \sigma^2 W_\gamma V_n} \\ &\le 1 - R_\gamma^2 \le \frac{\boldsymbol{\beta}_T' \{\boldsymbol{X}_T' \boldsymbol{Q}_\gamma \boldsymbol{X}_T/n\} \boldsymbol{\beta}_T + 2\boldsymbol{\beta}_T' \{\boldsymbol{X}_T' \boldsymbol{Q}_\gamma \boldsymbol{\epsilon}/n\} + \sigma^2 V_n}{\boldsymbol{\beta}_T' \{\boldsymbol{X}_T' \boldsymbol{X}_T/n\} \boldsymbol{\beta}_T + 2\boldsymbol{\beta}_T' \{\boldsymbol{X}_T' \boldsymbol{\epsilon}/n\} + \sigma^2 V_n} \end{aligned} \tag{E.2}$$

where $V_n = \check{\boldsymbol{\epsilon}}' \check{\boldsymbol{\epsilon}} / \{n\sigma^2\}$ and $W_\gamma = \check{\boldsymbol{\epsilon}}' \boldsymbol{Q}_\gamma \check{\boldsymbol{\epsilon}} / \|\check{\boldsymbol{\epsilon}}\|^2 \sim Be(\{n - q_\gamma - 1\}/2, q_\gamma/2)$. In (E.2), we have the following.

- Since $E[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{var}[\boldsymbol{\epsilon}] = \sigma^2 \boldsymbol{I}_n$, $E[\boldsymbol{X}_T' \boldsymbol{\epsilon}/n] = \mathbf{0}$ and

$$\mathrm{var}\,(\boldsymbol{X}_T' \boldsymbol{\epsilon}/n) = n^{-1}\sigma^2 \{\boldsymbol{X}_T' \boldsymbol{X}_T/n\} \to \mathbf{0}. \tag{E.3}$$

  Therefore $\boldsymbol{\beta}_T' \boldsymbol{X}_T' \boldsymbol{\epsilon}/n$ approaches 0 in probability.
- When $\gamma \supseteq T$, $\boldsymbol{Q}_\gamma \boldsymbol{X}_T$ is a zero matrix. When $\gamma \not\supseteq T$, $\boldsymbol{\beta}_T' \{\boldsymbol{X}_T' \boldsymbol{Q}_\gamma \boldsymbol{\epsilon}/n\} \to 0$ in probability can be proved as (E.3).
- By the assumption A3, $\boldsymbol{X}_T' \boldsymbol{X}_T/n - \boldsymbol{X}_T' \boldsymbol{Q}_\gamma \boldsymbol{X}_T/n$ is positive-definite for any n and hence

$$\boldsymbol{\beta}_T' \{\boldsymbol{X}_T' \boldsymbol{X}_T/n\} \boldsymbol{\beta}_T > \boldsymbol{\beta}_T' \{\boldsymbol{X}_T' \boldsymbol{Q}_\gamma \boldsymbol{X}_T/n\} \boldsymbol{\beta}_T, \text{ for } \boldsymbol{\beta}_T \neq \mathbf{0}.$$

- W_γ converges to 1 in probability.
- By the assumption A1 on $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}/\{n\sigma^2\}$, V_n is also bounded in probability from below and from above.

Combining these facts, we see $0 < R_\gamma^2 < 1$ with strict inequalities in probability.

Since $\boldsymbol{Q}_\gamma \boldsymbol{X}_T = \mathbf{0}$ for $\gamma \supseteq T$ and using (E.1), $(1 - R_T^2)/(1 - R_\gamma^2)$ is given by $\|\boldsymbol{Q}_T \check{\boldsymbol{\epsilon}}\|^2 / \|\boldsymbol{Q}_\gamma \check{\boldsymbol{\epsilon}}\|^2$. Further we easily have

$$1 \le \frac{1 - R_T^2}{1 - R_\gamma^2} = \frac{\|\boldsymbol{Q}_T \check{\boldsymbol{\epsilon}}\|^2}{\|\boldsymbol{Q}_\gamma \check{\boldsymbol{\epsilon}}\|^2} \le \frac{\|\check{\boldsymbol{\epsilon}}\|^2}{\|\boldsymbol{Q}_\gamma \check{\boldsymbol{\epsilon}}\|^2} = \frac{1}{W_\gamma}.$$

Note $W_\gamma \sim Be(\{n - q_\gamma - 1\}/2, q_\gamma/2)$ is distributed as $(1 + \chi_{q_\gamma}^2/\chi_{n-q_\gamma-1}^2)^{-1}$ where $\chi_{n-q_\gamma-1}^2$ and $\chi_{q_\gamma}^2$ are independent. Hence

$$\begin{aligned} \left\{1 + \chi_{q_\gamma}^2/\chi_{n-q_\gamma-1}^2\right\}^{-n} &= \left\{1 + \left\{n/\chi_{n-q_\gamma-1}^2\right\}\left\{\chi_{q_\gamma}^2/n\right\}\right\}^{-n} \\ &\sim \exp(-\chi_{q_\gamma}^2) \text{ as } n \to \infty \end{aligned}$$

since $\chi_{n-q_\gamma-1}^2/n \to 1$ in probability. Therefore W_γ^{-n} is bounded in probability from above and hence the theorem follows.

$(1-R_T^2)/(1-R_\gamma^2)$ is written as

$$\begin{aligned}\frac{1-R_T^2}{1-R_\gamma^2} &= \frac{\|\boldsymbol{Q}_T\check{\boldsymbol{\epsilon}}\|^2}{\boldsymbol{\beta}_T'\boldsymbol{X}_T'\boldsymbol{Q}_\gamma\boldsymbol{X}_T\boldsymbol{\beta}_T+2\boldsymbol{\beta}_T'\boldsymbol{X}_T'\boldsymbol{Q}_\gamma\boldsymbol{\epsilon}+\check{\boldsymbol{\epsilon}}'\boldsymbol{Q}_\gamma\check{\boldsymbol{\epsilon}}} \\ &\le \frac{\|\check{\boldsymbol{\epsilon}}\|^2}{\boldsymbol{\beta}_T'\boldsymbol{X}_T'\boldsymbol{Q}_\gamma\boldsymbol{X}_T\boldsymbol{\beta}_T+2\boldsymbol{\beta}_T'\boldsymbol{X}_T'\boldsymbol{Q}_\gamma\boldsymbol{\epsilon}+\check{\boldsymbol{\epsilon}}'\boldsymbol{Q}_\gamma\check{\boldsymbol{\epsilon}}} \\ &= \left(\frac{\boldsymbol{\beta}_T'\{\boldsymbol{X}_T'\boldsymbol{Q}_\gamma\boldsymbol{X}_T/n\}\boldsymbol{\beta}_T+2\boldsymbol{\beta}_T'\{\boldsymbol{X}_T'\boldsymbol{Q}_\gamma\boldsymbol{\epsilon}/n\}}{\sigma^2V_n}+W_\gamma\right)^{-1}.\end{aligned} \tag{E.4}$$

Clearly $W_\gamma \to 1$ in probability. Also since $\gamma \not\supseteq T$, $\boldsymbol{\beta}_T'\{\boldsymbol{X}_T'\boldsymbol{Q}_\gamma\boldsymbol{X}_T/n\}\boldsymbol{\beta}_T > 0$ for any n. Further as $\{\boldsymbol{X}_T'\boldsymbol{Q}_\gamma\boldsymbol{\epsilon}/n\} \to \boldsymbol{0}$ in probability, $(1-R_T^2)/(1-R_\gamma^2)$ is strictly smaller than 1 in probability.

References

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747. MR1813972

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

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

Maruyama, Y. (2003). A robust generalized Bayes estimator improving on the James-Stein estimator for spherically symmetric distributions. *Statist. Decisions* **21** 69–77. MR1985652

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

Maruyama, Y. and Strawderman, W. E. (2005). A new class of generalized Bayes minimax ridge regression estimators. *Ann. Statist.* **33** 1753–1770. MR2166561

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.* **42** 385–388. MR0397939

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86. MR830567

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469