gBF: A Fully Bayes Factor with a Generalized g-prior

Yuzo Maruyama and Edward I. George

University of Tokyo and University of Pennsylvania
e-mail: maruyama@csis.u-tokyo.ac.jp; *e-mail:* edgeorge@wharton.upenn.edu

Abstract: For the normal linear model variable selection problem, we propose selection criteria based on a fully Bayes formulation with a generalization of Zellner's g-prior which allows for $p > n$. A special case of the prior formulation is seen to yield tractable closed forms for marginal densities and Bayes factors which reveal new model characteristics of potential interest.

AMS 2000 subject classifications: Primary 62F07, 62F15; secondary 62C10.
Keywords and phrases: Bayes factor, model selection consistency, ridge regression, singular value decomposition, variable selection.

1. Introduction

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

$$\boldsymbol{y} \sim N_n(\alpha \mathbf{1}_n + \boldsymbol{X}_F \boldsymbol{\beta}_F, \sigma^2 \boldsymbol{I}_n) \tag{1.1}$$

where α is an unknown intercept parameter, $\mathbf{1}_n$ is an $n \times 1$ vector each component of which is one, $\boldsymbol{X}_F = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is an $n \times p$ design matrix, $\boldsymbol{\beta}_F$ is a $p \times 1$ vector of unknown regression coefficients, $\boldsymbol{I}_n$ is an $n \times n$ identity matrix and σ^2 is an unknown positive scalar. (The subscript F denotes the full model). 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 important 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 $\boldsymbol{\gamma}$th subset. The problem then becomes that of selecting a submodel of (1.1) which has a density of the form

$$p(\boldsymbol{y}|\alpha, \boldsymbol{\beta}_\gamma, \sigma^2, \boldsymbol{\gamma}) = \phi_n(\boldsymbol{y}; \alpha \mathbf{1}_n + \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 \boldsymbol{I}_n) \tag{1.2}$$

where $\phi_n(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the n-variate normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In (1.2), $\boldsymbol{X}_\gamma$ is the $n \times q_\gamma$ matrix whose columns correspond to the $\boldsymbol{\gamma}$th subset of $x_1, \ldots, x_p$, $\boldsymbol{\beta}_\gamma$ is a $q_\gamma \times 1$ vector of unknown regression coefficients. We assume throughout that $\boldsymbol{X}_\gamma$ is of full rank denoted

$$r_\gamma = \min\{q_\gamma, n-1\}.$$

Lastly, let $\mathcal{M}_\gamma$ denote the submodel given by (1.2).

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 $p(\alpha, \boldsymbol{\beta}_\gamma, \sigma^2)$ of each model. For each such specification, of key interest is the posterior probability of $\mathcal{M}_\gamma$ given $\boldsymbol{y}$

$$\Pr(\mathcal{M}_\gamma|\boldsymbol{y}) = \frac{\pi_\gamma m_\gamma(\boldsymbol{y})}{\sum_\gamma \pi_\gamma m_\gamma(\boldsymbol{y})} = \frac{\pi_\gamma \mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_N]}{\sum_\gamma \pi_\gamma \mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_N]}, \tag{1.3}$$

where $m_\gamma(\boldsymbol{y})$ is the marginal density of $\boldsymbol{y}$ under $\mathcal{M}_\gamma$. In (1.3), $\mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_N]$ is so called "null-based Bayes factor" for comparing each of $\mathcal{M}_\gamma$ to the null model $\mathcal{M}_N$ which is defined as

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

where the null model $\mathcal{M}_N$ is given by $\boldsymbol{y} \sim N_n(\alpha \mathbf{1}_n, \sigma^2 \boldsymbol{I}_n)$ and $m_N(\boldsymbol{y})$ is the marginal density of $\boldsymbol{y}$ under the null model. For model selection, a popular strategy is to select the model for which $\Pr(\mathcal{M}_\gamma|\boldsymbol{y})$ or $\pi_\gamma \mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_N]$ is largest.

Our main focus in this paper is to propose and study specifications for the parameter prior for each submodel $\mathcal{M}_\gamma$, which we will consider to be of the form

$$p(\alpha, \boldsymbol{\beta}_\gamma, \sigma^2) = p(\alpha)p(\sigma^2)p(\boldsymbol{\beta}_\gamma|\sigma^2) = p(\alpha)p(\sigma^2)\int p(\boldsymbol{\beta}_\gamma|\sigma^2, g)p(g)dg, \tag{1.4}$$

where g is a hyperparameter. In Section 2, we explicitly describe our choices of prior forms for (1.4). Our key innovation there will be to use a generalization of

$$p(\boldsymbol{\beta}_\gamma|\sigma^2, g) = \phi_{q_\gamma}(\boldsymbol{\beta}; \mathbf{0}, g\sigma^2(\boldsymbol{X}_\gamma'\boldsymbol{X}_\gamma)^{-1})), \tag{1.5}$$

Zellner's (1986) g-prior, a normal conjugate form which leads to tractable marginalization, for example see George and Foster (2000); Fernández, Ley and Steel (2001); Liang *et al.* (2008). Under (1.5) and a flat prior on α, the marginal density of $\boldsymbol{y}$ given g and σ^2 under $\mathcal{M}_\gamma$ is given by

$$m_\gamma(\boldsymbol{y}|g,\sigma^2) \propto \exp\left(\frac{g}{g+1}\left\{\max_{\alpha,\boldsymbol{\beta}_\gamma}\log p(\boldsymbol{y}|\alpha,\boldsymbol{\beta}_\gamma,\sigma^2) - q_\gamma H(g)\right\}\right) \tag{1.6}$$

where $H(g) = (2g)^{-1}(g+1)\log(g+1)$, a special case of the key relation in George and Foster (2000). As they point out, for particular values of g, when σ^2 is known, the Bayesian strategy of choosing $\mathcal{M}_\gamma$ to maximize (1.6) corresponds to common fixed penalty selection criteria. For example, setting $H(g) = 2$, $\log n$ or $2\log p$ (independently of $\boldsymbol{y}$) would correspond to AIC (Akaike, 1974), BIC (Schwarz, 1978), or RIC (Foster and George, 1994), respectively. For a discussion of recommendations in the literature for choosing a fixed g depending on p and/or n, see Section 2.4 of Liang *et al.* (2008).

Although the correspondences to fixed penalty criteria are interesting, as a practical matter, it is necessary to deal with the uncertainty about g and σ^2 to obtain useful criteria. For this purpose, George and Foster (2000) proposed selecting the model maximizing $m_\gamma(\boldsymbol{y}|g,\sigma^2)$ based on an empirical Bayes estimate of g and the standard unbiased estimate of σ^2. More recently, Cui and George (2008) proposed margining out g with respect to a prior, and Liang *et al.* (2008) proposed margining out g and σ^2 with respect to priors. All of these strategies lead to criteria that can be seen as adapting to the fixed penalty criterion which would be most suitable for the data at hand. In this paper, we shall similarly follow a fully Bayes approach, but with a generalization of the g-prior (1.5) and an extension of the considered class of priors on g.

After describing our prior forms in Section 2 and then calculating the marginals and Bayes factors in Section 3, we ultimately obtain our proposed g-prior Bayes Factor (gBF), which is of the form (omitting the γ subscripts for clarity)

$$g\mathrm{BF}[\mathcal{M}_\gamma] = \begin{cases} C_{n,q}\left\{\dfrac{\bar{d}}{d_q}\right\}^{-q}\dfrac{\left\{1-R^2+d_q^2\,\|\hat{\boldsymbol{\beta}}_{LS}\|^2\right\}^{-\frac{1}{4}-\frac{q}{2}}}{(1-R^2)^{(n-q)/2-3/4}} & \text{if } q < n-1, \\ \left\{\bar{d}\times\|\hat{\boldsymbol{\beta}}_{LS}^{MP}\|\right\}^{-n+1} & \text{if } q \geq n-1, \end{cases} \tag{1.7}$$

where $C_{n,q} \equiv \frac{B(q/2+1/4,(n-q)/2-3/4)}{B(1/4,(n-q)/2-3/4)}$ using the Beta function $B(\cdot,\cdot)$, R^2 is the familiar R-squared statistic under $\mathcal{M}_\gamma$, $\bar{d}$ and d_r are respectively the geometric mean and minimum of the singular values of $\boldsymbol{X}_\gamma$, $\|\cdot\|$ is the L_2 norm, and finally, for the standardized response $(\boldsymbol{y} - \bar{y}\mathbf{1}_n)/\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|$, $\hat{\boldsymbol{\beta}}_{LS}$ is the usual least squares estimator, and $\hat{\boldsymbol{\beta}}_{LS}^{MP}$ is the least squares estimator using the Moore-Penrose inverse matrix.

Two immediately apparent features of (1.7) should be noted. First, in contrast to other fully Bayes factors for our selection problem, gBF is a closed form expression which allows for interpretation and straightforward calculation under any model. As will be seen in later sections, this transparency reveals that gBF not only rewards explained variation overall, but also rewards variation explained by the larger principal components of the design matrix. Second, gBF can be applied to all models even when the number of predictors p exceeds the number of observations n. This includes $p > n$ which is of increasing interest. This is not the case for (1.5) which requires $p \leq n-1$ so that $\boldsymbol{X}_\gamma' \boldsymbol{X}_\gamma$ will be invertible for all q_γ, (recall that $\boldsymbol{X}_\gamma$ has dimension at most $n-1$ because its columns have been centered). Note also that when $p > n-1$, penalized sum-of-squares criteria such as AIC, BIC and RIC will be unavailable for all submodels.

The organization of this paper is as follows. In Section 2, we will give priors including a special variant of g-prior. In Section 3, we derive the Bayes factor above. In Section 4, we discuss the choice of hyper-parameters which appears in the variant of g-priors. In Section 5, the estimation after selection is discussed. In Section 6, we show that gBF has consistency for model selection as $n \to \infty$. In Section 7, we give some numerical results.

2. A Fully Bayes Prior Formulation

We now proceed to describe the prior components that form $p(\alpha, \boldsymbol{\beta}_\gamma, \sigma^2)$ in (1.4). Throughout this section and the next, we will omit the subscript γ for notational simplicity when there is no ambiguity.

2.1. A generalized g-prior for $\boldsymbol{\beta}$

To motivate our proposed generalization of Zellner's g-prior, we begin with a reconsideration the original g-prior (1.5) for the case $p \leq n-1$. The covariance matrix of the g-prior, $g\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, is proportional to the covariance matrix of the least squares estimator $\hat{\boldsymbol{\beta}}_{LS}$. As a consequence of this choice, the marginal density with respect to the g-prior appealingly becomes a function only of the residual sum-of-squares, RSS.

However, from the "conditioning" viewpoint of Casella (1980, 1985) which advocates more shrinkage on higher variance estimates, the original g-prior may not be reasonable. To see why, let us rotate the problem by the $q \times q$ orthogonal matrix $\boldsymbol{W} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_q)$ which diagonalizes $\boldsymbol{X}'\boldsymbol{X}$ as

$$\boldsymbol{W}'(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{W} = \boldsymbol{D}^2 \tag{2.1}$$

where $\boldsymbol{D} = \mathrm{diag}(d_1, \dots, d_q)$ with

$$d_1 \geq \cdots \geq d_q > 0. \tag{2.2}$$

Thus

$$\boldsymbol{W}'\hat{\boldsymbol{\beta}}_{LS} \sim N_q(\boldsymbol{W}'\boldsymbol{\beta}, \sigma^2 \boldsymbol{D}^{-2}).$$

For this rotation, we consider priors on $\boldsymbol{\beta}$ for which

$$\boldsymbol{W}'\boldsymbol{\beta} \sim N_q(\boldsymbol{0}, \sigma^2 \boldsymbol{\Psi}_q)$$

where $\boldsymbol{\Psi}_q = \mathrm{diag}(\psi_1, \dots, \psi_q)$.

From a Bayesian perspective, it is more sensible to put stronger prior information on the components $\boldsymbol{w}_i'\boldsymbol{\beta}$ of $\boldsymbol{W}'\boldsymbol{\beta}$ which are estimated poorly, (that is, the components with larger sample variance). Hence, we would like to consider $\boldsymbol{\Psi}_q$ for which

$$\psi_1 \geq \cdots \geq \psi_q > 0 \tag{2.3}$$

are in descending order. In fact, a slightly weaker ordering of the form

$$d_1^2\psi_1 \geq \cdots \geq d_q^2\psi_q > 0 \tag{2.4}$$

would still be reasonable because the resulting Bayes estimator of $\boldsymbol{w}_i'\boldsymbol{\beta}$ would be of the form

$$(1 + \{d_i^2\psi_i\}^{-1})^{-1}\boldsymbol{w}_i'\hat{\boldsymbol{\beta}}_{LS},$$

so that under (2.4), the components of $\boldsymbol{W}'\hat{\boldsymbol{\beta}}_{LS}$ with larger variance would be shrunk more. We note that the original g-prior (1.5), for which $\psi_i = g d_i^{-2}$, satisfies only the extreme boundary of (2.4), namely

$$d_1^2\psi_1 = \cdots = d_q^2\psi_q = g.$$

This violates (2.3) whenever $d_i > d_{i+1}$ in which case $\psi_i < \psi_{i+1}$.

An appealing general form for $\boldsymbol{\Psi}_q$ is $\boldsymbol{\Psi}_q(g, \boldsymbol{\nu}) = \mathrm{diag}(\psi_1(g, \boldsymbol{\nu}), \dots, \psi_q(g, \boldsymbol{\nu}))$ where

$$\psi_i(g, \boldsymbol{\nu}) = (1/d_i^2)\left\{\nu_i(1+g) - 1\right\}, \tag{2.5}$$

$\boldsymbol{\nu} = (\nu_1, \dots, \nu_q)'$ and $\nu_i \geq 1$ for any i, guaranteeing $\psi_i(g, \boldsymbol{\nu}) > 0$. Note that $\boldsymbol{\Psi}_q(g, \boldsymbol{\nu})$, like the original g-prior, is controlled by a single hyperparameter

$g > 0$. When $\nu_1 = \cdots = \nu_q = 1$, $\sigma^2 \boldsymbol{\Psi}_q(g, \boldsymbol{\nu})$ becomes $g\sigma^2 \boldsymbol{D}^{-2}$, yielding the covariance structure of the original g-prior. Although (2.4) will be satisfied whenever $\nu_1 \geq \cdots \geq \nu_q \geq 1$, in subsequent sections we shall ultimately be interested in the particular design dependent choice

$$\nu_1 = d_1^2/d_q^2,\ \nu_2 = d_2^2/d_q^2, \ldots, \nu_q = 1 \tag{2.6}$$

which satisfies (2.3) as well as (2.4). In summary, when $q \leq n-1$, we propose a generalized g-prior for $\boldsymbol{\beta}$ of the form

$$p(\boldsymbol{\beta} \mid \sigma^2, g) = \phi_q(\boldsymbol{W}'\boldsymbol{\beta}; \boldsymbol{0}, \sigma^2 \boldsymbol{\Psi}_q(g, \boldsymbol{\nu})) \tag{2.7}$$

where $\nu_1 \geq \cdots \geq \nu_q \geq 1$.

When $q > n-1$ and the rank of $\boldsymbol{X}$ is $n-1$, there exists a $q \times (n-1)$ matrix $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{n-1})$ which diagonalizes $\boldsymbol{X}'\boldsymbol{X}$ as

$$\boldsymbol{W}'(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{W} = \boldsymbol{D}^2 \tag{2.8}$$

where $\boldsymbol{W}'\boldsymbol{W} = \boldsymbol{I}_{n-1}$ and $\boldsymbol{D} = \mathrm{diag}(d_1, d_2, \ldots, d_{n-1})$ with $d_1 \geq d_2 \geq \cdots \geq d_{n-1} > 0$. For this case, we propose a generalized g-prior of the form

$$p(\boldsymbol{\beta}|\sigma^2, g) = \phi_{n-1}(\boldsymbol{W}'\boldsymbol{\beta}; \boldsymbol{0}, \sigma^2 \boldsymbol{\Psi}_{n-1}(g, \boldsymbol{\nu}))\, p_{\#}(\boldsymbol{W}_{\#}'\boldsymbol{\beta}) \tag{2.9}$$

where $\boldsymbol{\Psi}_{n-1}(g, \boldsymbol{\nu}) = \mathrm{diag}(\psi_1, \ldots, \psi_{n-1})$ is again given by (2.5) and $\nu_1 \geq \cdots \geq \nu_{n-1} \geq 1$. Here, $\boldsymbol{W}_{\#}$ is an arbitrary matrix which makes the $q \times q$ matrix $(\boldsymbol{W}, \boldsymbol{W}_{\#})$ orthogonal, and $p_{\#}(\cdot)$ is an arbitrary probability density on $\boldsymbol{W}_{\#}'\boldsymbol{\beta}$, respectively. As will be seen, the choices of $\boldsymbol{W}_{\#}$ and $p_{\#}$ have no effect on the selection criteria we obtain, thus we leave them as arbitrary. As in (2.6), we shall be ultimately interested in the particular design dependent choice

$$\nu_1 = d_1^2/d_{n-1}^2,\ \nu_2 = d_2^2/d_{n-1}^2, \ldots, \nu_{n-1} = 1. \tag{2.10}$$

Combining the above two cases by letting

$$r = \min\{q, n-1\}, \tag{2.11}$$

our suggested generalized g-prior is of the form

$$\begin{aligned} &p(\boldsymbol{\beta}|g, \sigma^2) \\ &= \phi_r(\boldsymbol{W}'\boldsymbol{\beta}; \boldsymbol{0}, \sigma^2 \boldsymbol{\Psi}_r(g, \boldsymbol{\nu})) \times \begin{cases} 1 & \text{if } q \leq n-1 \\ p_{\#}(\boldsymbol{W}_{\#}'\boldsymbol{\beta}), & \text{if } q > n-1, \end{cases} \end{aligned} \tag{2.12}$$

where the $q \times r$ matrix $\boldsymbol{W}$ satisfies both $\boldsymbol{W}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{W} = \mathrm{diag}(d_1^2, \ldots, d_r^2)$ and $\boldsymbol{W}'\boldsymbol{W} = \boldsymbol{I}_r$, and $\boldsymbol{\Psi}_r(g, \boldsymbol{\nu}) = \mathrm{diag}(\psi_1(g, \boldsymbol{\nu}), \ldots, \psi_r(g, \boldsymbol{\nu}))$ with (2.5).

Remark 2.1. In (2.1) and (2.8), let

$$\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r) = (\boldsymbol{X}\boldsymbol{w}_1/d_1, \ldots, \boldsymbol{X}\boldsymbol{w}_r/d_r) = \boldsymbol{X}\boldsymbol{W}\boldsymbol{D}^{-1}. \tag{2.13}$$

Then $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}_r$ and

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{W}' = \sum_{i=1}^{r} d_i \boldsymbol{u}_i \boldsymbol{w}_i'. \tag{2.14}$$

This is the non-null part of the well-known singular value decomposition (SVD). The diagonal elements of $\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_r)$ are the singular values of $\boldsymbol{X}$, and the columns of $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r)$ are the normalized principal components of the column space of $\boldsymbol{X}$. Note that the components of the rotated vector $\boldsymbol{W}'\boldsymbol{\beta}$ are the coefficients for the principal component regression of $\boldsymbol{y}$ on $\boldsymbol{U}\boldsymbol{D}$. From the definition of $\boldsymbol{W}$ and $\boldsymbol{U}$ by (2.1), (2.8) and (2.13), the signs of $\boldsymbol{u}_i\boldsymbol{w}_i'$ are determinate although the signs of $\boldsymbol{w}_i$ and $\boldsymbol{u}_i$ for $1 \le i \le r$ are indeterminate. These indeterminacies can safely be ignored in our development.

2.2. A prior for g

Turning to the prior for the hyperparameter g, we propose

$$p(g) = \frac{g^b (1+g)^{-a-b-2}}{B(a+1, b+1)} I_{(0,\infty)}(g) \tag{2.15}$$

with $a > -1$, $b > -1$, a Pearson Type VI or *beta-prime* distribution under which $1/(1+g)$ has a Beta distribution $Be(a+1, b+1)$. Choices for the hyperparameters a and b are discussed later.

Although Zellner and Siow (1980) did not explicitly use a g-prior formulation with a prior on g, their recommendation of a multivariate Cauchy form for $p(\boldsymbol{\beta}|\sigma^2)$ implicitly corresponds to using a g-prior with an inverse Gamma prior

$$(n/2)^{1/2}\{\Gamma(1/2)\}^{-1} g^{-3/2} e^{-n/(2g)}$$

on g. Both Cui and George (2008) and Liang *et al.* (2008) proposed using g-priors with priors of the form

$$p(g) = (a+1)^{-1}(1+g)^{-a-2}, \tag{2.16}$$

the subclass of (2.15) with $b = 0$. Cases for which $b = O(n)$ will be of interest to us in what follows.

2.3. Priors for α and σ^2

For the parameter α, we ultimately use the location invariant flat prior

$$p(\alpha) = I_{(-\infty,\infty)}(\alpha), \tag{2.17}$$

but do so indirectly by first applying the proper uniform prior

$$p(\alpha; h_\alpha) = \frac{1}{2h_\alpha} I_{(-h_\alpha, h_\alpha)}(\alpha) \tag{2.18}$$

and then taking the limit as $h_\alpha \to \infty$. Similarly for σ^2, we ultimately use the scale invariant prior

$$p(\sigma^2) = (\sigma^2)^{-1} I_{(0,\infty)}(\sigma^2), \tag{2.19}$$

but do so by first applying the proper truncated version

$$p(\sigma^2; h_\sigma) = \frac{(\sigma^2)^{-1}}{\int_{h_\sigma^{-1}}^{h_\sigma} (\sigma^2)^{-1} d\sigma^2} I_{(h_\sigma^{-1},\ h_\sigma)}(\sigma^2) = \frac{(\sigma^2)^{-1}}{2\log h_\sigma} I_{(h_\sigma^{-1},\ h_\sigma)}(\sigma^2) \tag{2.20}$$

and then taking the limit as $h_\sigma \to \infty$. Because α and σ^2 appear in every model, this approach avoids the arbitrary norming constant difficulties associated with the use of improper priors in Bayesian model selection.

We note in passing that for the estimation of a multivariate normal mean, priors equivalent to (2.7), (2.15), (2.17) and (2.19) have been considered by Strawderman (1971) and extended by Maruyama and Strawderman (2005).

3. Marginal Densities and Bayes Factors

For the model $\mathcal{M}_\gamma$, the marginal density of $\boldsymbol{y}$ under the proper priors $p(\boldsymbol{\beta}|\sigma^2, g)$, $p(g)$, $p(\alpha; h_\alpha)$ and $p(\sigma^2; h_\sigma)$ given by (2.12), (2.15), (2.18) and (2.20), is obtained as

$$\begin{aligned} m_\gamma(\boldsymbol{y}) = \int_{-h_\alpha}^{h_\alpha} \int_{R^q} \int_{h_\sigma^{-1}}^{h_\sigma} \int_0^\infty & p(\boldsymbol{y}|\alpha, \boldsymbol{\beta}, \sigma^2) p(\alpha; h_\alpha) p(\boldsymbol{\beta}|\sigma^2, g) \\ & \times p(\sigma^2; h_\sigma) p(g) d\alpha\, d\boldsymbol{\beta}\, d(\sigma^2)\, dg. \end{aligned} \tag{3.1}$$

Going further, it will be useful to consider the limit of a renormalized version of $m_\gamma(\boldsymbol{y})$,

$$\begin{aligned} & M_\gamma(\boldsymbol{y}) \\ & = \lim_{h_\alpha \to \infty} \lim_{h_\sigma \to \infty} \{4 h_\alpha \log h_\sigma\}\, m_\gamma(\boldsymbol{y}) \\ & = \int_{-\infty}^{\infty} \int_{R^q} \int_0^\infty \int_0^\infty p(\boldsymbol{y}|\alpha, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}|g, \sigma^2) \frac{1}{\sigma^2} p(g) d\alpha\, d\boldsymbol{\beta}\, d(\sigma^2)\, dg, \end{aligned} \tag{3.2}$$

which is the marginal density of $\boldsymbol{y}$ with respect to the improper priors (2.17) on α and (2.19) on σ^2. Note that the second equality in (3.2) follows from the monotone convergence theorem. Thus $m_\gamma(\boldsymbol{y})/m_N(\boldsymbol{y})$, the Bayes factor for $\mathcal{M}_\gamma$ with respect to the null model $\mathcal{M}_N$, approaches $M_\gamma(\boldsymbol{y})/M_N(\boldsymbol{y})$ as $h_\alpha \to \infty$ and $h_\sigma \to \infty$. Hence the use of improper priors $p(\alpha) = 1$ and $p(\sigma^2) = 1/\sigma^2$ is formally justified because $M_\gamma(\boldsymbol{y})/M_N(\boldsymbol{y})$ is well-defined.

Defining $\boldsymbol{v} = \boldsymbol{y} - \bar{y}\mathbf{1}_n$, where $\bar{y}$ is the mean of $\boldsymbol{y}$, so that

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

we obtain

$$\int_{-\infty}^{\infty} p(\boldsymbol{y}|\alpha, \boldsymbol{\beta}, \sigma^2)d\alpha = \frac{n^{1/2}}{(2\pi\sigma^2)^{(n-1)/2}} \exp\left(-\frac{\|\boldsymbol{v} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\right). \tag{3.3}$$

We make the following orthogonal transformation when integration with respect to $\boldsymbol{\beta}$ is considered:

$$\boldsymbol{\beta} \to \begin{cases} \boldsymbol{W}'\boldsymbol{\beta} \equiv \boldsymbol{\beta}_* & \text{if } q \le n-1 \\ \begin{pmatrix} \boldsymbol{W}'\boldsymbol{\beta} \\ \boldsymbol{W}'_{\#}\boldsymbol{\beta} \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{\beta}_* \\ \boldsymbol{\beta}_{\#} \end{pmatrix} & \text{if } q > n-1, \end{cases} \tag{3.4}$$

so that

$$\begin{aligned} &\int_{-\infty}^{\infty}\int_{R^q} p(\boldsymbol{y}|\alpha, \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}|\sigma^2, g)d\alpha\, d\boldsymbol{\beta} \\ &= \frac{n^{1/2}}{(2\pi\sigma^2)^{(n-1)/2}} \frac{|\boldsymbol{\Psi}|^{-1/2}}{(2\pi\sigma^2)^{r/2}} \int_{R^r} \exp\left(-\frac{\|\boldsymbol{v} - \boldsymbol{U}\boldsymbol{D}\boldsymbol{\beta}_*\|^2}{2\sigma^2} - \frac{\boldsymbol{\beta}_*'\boldsymbol{\Psi}^{-1}\boldsymbol{\beta}_*}{2\sigma^2}\right) d\boldsymbol{\beta}_* \\ &\qquad \times \begin{cases} 1 & \text{if } q \le n-1 \\ \int_{R^{q-n+1}} p_{\#}(\boldsymbol{\beta}_{\#})d\boldsymbol{\beta}_{\#} \ (=1) & \text{if } q > n-1. \end{cases} \end{aligned}$$

Completing the square $\|\boldsymbol{v} - \boldsymbol{U}\boldsymbol{D}\boldsymbol{\beta}_*\|^2 + \boldsymbol{\beta}_*'\boldsymbol{\Psi}^{-1}\boldsymbol{\beta}_*$ with respect to $\boldsymbol{\beta}_*$, we have

$$\begin{aligned} &\|\boldsymbol{v} - \boldsymbol{U}\boldsymbol{D}\boldsymbol{\beta}_*\|^2 + \boldsymbol{\beta}_*'\boldsymbol{\Psi}^{-1}\boldsymbol{\beta}_* \\ &= \{\boldsymbol{\beta}_* - (\boldsymbol{D}^2 + \boldsymbol{\Psi}^{-1})^{-1}\boldsymbol{D}'\boldsymbol{U}'\boldsymbol{v}\}'(\boldsymbol{D}^2 + \boldsymbol{\Psi}^{-1})\{\boldsymbol{\beta}_* - (\boldsymbol{D}^2 + \boldsymbol{\Psi}^{-1})^{-1}\boldsymbol{D}'\boldsymbol{U}'\boldsymbol{v}\} \\ &\qquad - \boldsymbol{v}'\boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + \boldsymbol{\Psi}^{-1})^{-1}\boldsymbol{D}'\boldsymbol{U}'\boldsymbol{v} + \boldsymbol{v}'\boldsymbol{v}, \end{aligned} \tag{3.5}$$

where the residual term is rewritten as

$$
\begin{aligned}
&- \boldsymbol{v}'\boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2+\boldsymbol{\Psi}^{-1})^{-1}\boldsymbol{D}'\boldsymbol{U}'\boldsymbol{v}+\boldsymbol{v}'\boldsymbol{v} \\
&\quad = -\boldsymbol{v}'\left(\sum_{i=1}^{r}\boldsymbol{u}_i\boldsymbol{u}_i'\frac{d_i^2}{d_i^2+\psi_i^{-1}}\right)\boldsymbol{v}+\boldsymbol{v}'\boldsymbol{v} \\
&\quad = \frac{g}{g+1}\left\{\boldsymbol{v}'\boldsymbol{v}-\sum_{i=1}^{r}(\boldsymbol{u}_i'\boldsymbol{v})^2\right\}+\frac{1}{1+g}\left\{\boldsymbol{v}'\boldsymbol{v}-\sum_{i=1}^{r}\left(1-\frac{1}{\nu_i}\right)(\boldsymbol{u}_i'\boldsymbol{v})^2\right\}.
\end{aligned}
$$

Hence by

$$
|\boldsymbol{\Psi}|=\prod_{i=1}^{r}\frac{\nu_i+\nu_i g-1}{d_i^2}, \quad |\boldsymbol{D}^2+\boldsymbol{\Psi}^{-1}|=\prod_{i=1}^{r}\frac{d_i^2\nu_i(1+g)}{\nu_i+\nu_i g-1},
$$

we have

$$
\begin{aligned}
&\int_{-\infty}^{\infty}\int_{R^q} p(\boldsymbol{y}|\alpha,\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta}|g,\sigma^2)d\alpha\, d\boldsymbol{\beta} \\
&\quad = \frac{n^{1/2}}{(2\pi\sigma^2)^{(n-1)/2}}\frac{(1+g)^{-r/2}}{\prod_{i=1}^{r}\nu_i^{1/2}}\exp\left(-\frac{\|\boldsymbol{v}\|^2\{g(1-R^2)+1-Q^2\}}{2\sigma^2(g+1)}\right)
\end{aligned}
$$

where

$$
\begin{aligned}
R^2 &= \sum_{i=1}^{r}\frac{(\boldsymbol{u}_i'\boldsymbol{v})^2}{\boldsymbol{v}'\boldsymbol{v}}=\sum_{i=1}^{r}\{\mathrm{cor}(\boldsymbol{u}_i,\boldsymbol{y})\}^2, \\
Q^2 &= \sum_{i=1}^{r}\left(1-\nu_i^{-1}\right)\frac{(\boldsymbol{u}_i'\boldsymbol{v})^2}{\boldsymbol{v}'\boldsymbol{v}}=\sum_{i=1}^{r}\left(1-\nu_i^{-1}\right)\{\mathrm{cor}(\boldsymbol{u}_i,\boldsymbol{y})\}^2.
\end{aligned}
\tag{3.6}
$$

Note that R^2 and Q^2 are the usual and a modified version of the R-squared statistics and $\mathrm{cor}(\boldsymbol{u}_i,\boldsymbol{y})$ is the correlation of the response $\boldsymbol{y}$ and the i-th principal component of $\boldsymbol{X}$.

Next we consider the integration with respect to σ^2. By (3.6), we have

$$
\begin{aligned}
&\int_{-\infty}^{\infty}\int_{R^q}\int_{0}^{\infty} p(\boldsymbol{y}|\alpha,\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta}|g,\sigma^2)\frac{1}{\sigma^2}d\alpha\, d\boldsymbol{\beta}\, d\sigma^2 \\
&= \int_{0}^{\infty}\frac{n^{1/2}}{(2\pi\sigma^2)^{(n-1)/2}}\frac{(1+g)^{-r/2}}{\prod_{i=1}^{r}\nu_i^{1/2}}\exp\left(-\frac{\|\boldsymbol{v}\|^2\{g(1-R^2)+1-Q^2\}}{2\sigma^2(g+1)}\right)\frac{1}{\sigma^2}d\sigma^2 \\
&= \frac{K(n,\boldsymbol{y})}{\prod_{i=1}^{r}\nu_i^{1/2}}(1+g)^{-r/2+(n-1)/2}\left\{g(1-R^2)+1-Q^2\right\}^{-(n-1)/2}
\end{aligned}
\tag{3.7}
$$

where

$$K(n, \boldsymbol{y}) = \frac{n^{1/2}\Gamma(\{n-1\}/2)}{\pi^{(n-1)/2}\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^{n-1}}.$$

When $q \geq n-1$, $R^2 = 1$ and $r = n-1$ so that

$$\begin{aligned}&\int_{-\infty}^{\infty}\int_{R^q}\int_0^{\infty} p(\boldsymbol{y}|\alpha, \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}|g, \sigma^2)\frac{1}{\sigma^2}d\alpha\, d\boldsymbol{\beta}\, d\sigma^2 \\ &\quad = \frac{K(n, \boldsymbol{y})}{\prod_{i=1}^{n-1}\nu_i^{1/2}}\left\{1 - Q^2\right\}^{-(n-1)/2}\end{aligned} \tag{3.8}$$

which does not depend on g. Hence, in this case, $M_\gamma(\boldsymbol{y})$ does not depend on the prior density of g. Notice also that if $\nu_1 = \cdots = \nu_{n-1} = 1$ when $q \geq n-1$, $Q^2 = 0$ so that $M_\gamma(\boldsymbol{y}) = K(n, \boldsymbol{y})$ will not distinguish between models. Clearly, the choice of $\boldsymbol{\nu}$ matters.

When $q < n-1$, we consider the prior (2.15) of g with $-1 < a < -1/2$ and $b = (n-5)/2 - q/2 - a$, where b is guaranteed to be strictly greater than -1 for $q < n-1$. Then we have

$$\begin{aligned}&M_\gamma(\boldsymbol{y}) \\ &= \frac{K(n, \boldsymbol{y})}{\prod_{i=1}^{q}\nu_i^{1/2}B(a+1, b+1)} \\ &\quad \times \int_0^{\infty}\frac{g^b}{(1+g)^{a+b+2}}\frac{\{g(1-R^2)+1-Q^2\}^{-(n-1)/2}}{(1+g)^{q/2-(n-1)/2}}dg \\ &= \frac{K(n, \boldsymbol{y})(1-Q^2)^{-(n-1)/2}}{\prod_{i=1}^{q}\nu_i^{1/2}B(a+1, b+1)}\int_0^{\infty} g^b\left(\frac{1-R^2}{1-Q^2}g+1\right)^{-(n-1)/2} dg \\ &= \frac{K(n, \boldsymbol{y})(1-Q^2)^{-(n-1)/2+b+1}}{\prod_{i=1}^{q}\nu_i^{1/2}\{1-R^2\}^{b+1}}\frac{B(q/2+a+1, b+1)}{B(a+1, b+1)} \\ &= \frac{K(n, \boldsymbol{y})(1-Q^2)^{-q/2-a-1}}{\prod_{i=1}^{q}\nu_i^{1/2}\{1-R^2\}^{(n-q-3)/2-a}}\frac{B(q/2+a+1, (n-q-3)/2-a)}{B(a+1, (n-q-3)/2-a)}.\end{aligned} \tag{3.9}$$

In the same way, $M_N(\boldsymbol{y})$ for the null model, is obtained as

$$M_N(\boldsymbol{y}) = K(n, \boldsymbol{y}). \tag{3.10}$$

Using (3.8), (3.9) and (3.10), we have a following theorem about the Bayes factor ratio of the marginal densities under each of $\mathcal{M}_\gamma$ and $\mathcal{M}_N$.

Theorem 3.1. *Under the proper prior distributions of $\boldsymbol{\beta}|(\sigma^2, g)$, α and σ^2 given by (2.12), (2.18) and (2.20), and under the proper prior distribution* *of g given by (2.15) with $-1 < a < -1/2$ and $b = (n-5)/2 - q/2 - a$ when $q < n-1$, (when $q \geq n-1$, the prior distribution of g is arbitrary), the limit of the Bayes factor for comparing each of $\mathcal{M}_\gamma$ to the null model is*

$$\lim_{h_\alpha \to \infty} \lim_{h_\sigma \to \infty} \frac{m_\gamma(\boldsymbol{y})}{m_N(\boldsymbol{y})} = \mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_N | a, \boldsymbol{\nu}]$$

where

$$\begin{aligned} &\mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_N | a, \boldsymbol{\nu}] \\ &= \begin{cases} \displaystyle\prod_{i=1}^{q} \nu_i^{-1/2} \frac{B(\frac{q}{2} + a + 1, \frac{n-q-3}{2} - a)}{B(a+1, \frac{n-q-3}{2} - a)} \frac{(1-Q^2)^{-\frac{q}{2} - a - 1}}{(1-R^2)^{\frac{n-q-3}{2} - a}} & \text{if } q < n-1, \\ \displaystyle\prod_{i=1}^{n-1} \nu_i^{-1/2} \left(1 - Q^2\right)^{-(n-1)/2} & \text{if } q \geq n-1 \end{cases} \end{aligned} \tag{3.11}$$

where $\nu_1 \geq \cdots \geq \nu_r \geq 1$, R^2 and Q^2 are given by (3.6).

Remark 3.1. R^2 and Q^2 given by (3.6) are the usual and a modified form of the R-squared measure for multiple regression. They are here expressed in terms of $\{\mathrm{cor}(\boldsymbol{u}_1, \boldsymbol{y})\}^2, \ldots, \{\mathrm{cor}(\boldsymbol{u}_r, \boldsymbol{y})\}^2$, the squared correlations of the response $\boldsymbol{y}$ and the principal components $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$ of $\boldsymbol{X}$. For fixed q and $\boldsymbol{\nu}$, the BF criterion is increasing in both R^2 and Q^2. The former is definitely reasonable. Larger Q^2 would also be reasonable when $\nu_1 \geq \cdots \geq \nu_r$ so that Q^2 would put more weight on those components of $\boldsymbol{W}'\boldsymbol{\beta}$ for which d_i is larger and are consequently better estimated. In this sense, Q^2 would reward those models which are more stably estimated. Note that if $\nu_i = 1$ for all i (that is, the original g-prior), Q^2 becomes zero, and BF becomes a function of just R^2 and q.

Remark 3.2. The analytical simplification in (3.9) is a consequence of the choice $b = (n-5)/2 - q/2 - a$, and results in a convenient closed form for our Bayes factor. Such a reduction is unavailable for other choices of b. For example, Liang *et al.* (2008) use Laplace approximations to avoid the evaluation of the special functions that arise in the resulting Bayes factor when $b = 0$. Another attractive feature of the choice $b = (n-5)/2 - q/2 - a$ will be discussed in Section 4.2.

At this point, we are ready to consider default choices for a and $\boldsymbol{\nu}$. For a, we recommend

$$a_* = -3/4, \tag{3.12}$$

the median of the range of values $(-1, -1/2)$ for which the marginal density is well defined for any choices of $q < n-1$. In section 4, we will explicitly see the appealing consequence of this choice on the asymptotic tail behavior of $p(\boldsymbol{\beta}|\sigma^2)$. For $\boldsymbol{\nu}$, we recommend

$$\boldsymbol{\nu}_* = (d_1^2/d_r^2, d_2^2/d_r^2, \dots, 1)'$$

which yields

$$\begin{aligned} Q^2 &= R^2 - d_r^2 \sum_{i=1}^r \frac{(\boldsymbol{u}_i'\boldsymbol{v})^2}{d_i^2 \boldsymbol{v}'\boldsymbol{v}} \\ &= R^2 - d_r^2 \left\| \boldsymbol{D}^{-1}\boldsymbol{U}'\{\boldsymbol{v}/\|\boldsymbol{v}\|\} \right\|^2 \\ &= \begin{cases} R^2 - d_q^2 \|\hat{\boldsymbol{\beta}}_{LS}\|^2 & \text{if } q < n-1 \\ 1 - d_{n-1}^2 \|\hat{\boldsymbol{\beta}}_{LS}^{MP}\|^2 & \text{if } q \geq n-1, \end{cases} \end{aligned} \tag{3.13}$$

where, for the standardized response $\boldsymbol{v}/\|\boldsymbol{v}\|$ for $\boldsymbol{v} = \boldsymbol{y} - \bar{y}\mathbf{1}_n$, $\hat{\boldsymbol{\beta}}_{LS}$ is the usual LS estimator for $q < n-1$, and $\hat{\boldsymbol{\beta}}_{LS}^{MP}$ is the LS estimator based on the Moore-Penrose inverse matrix. The third equality in (3.13) follows from the fact that both $\hat{\boldsymbol{\beta}}_{LS}$ and $\hat{\boldsymbol{\beta}}_{LS}^{MP}$ for the response $\boldsymbol{v}/\|\boldsymbol{v}\|$ can be expressed as

$$\hat{\boldsymbol{\beta}} = \boldsymbol{W}\boldsymbol{D}^{-1}\boldsymbol{U}'\{\boldsymbol{v}/\|\boldsymbol{v}\|\},$$

and from the orthogonality of $\boldsymbol{W}$,

$$\|\hat{\boldsymbol{\beta}}\|^2 = \left\| \boldsymbol{D}^{-1}\boldsymbol{U}'\{\boldsymbol{v}/\|\boldsymbol{v}\|\} \right\|^2.$$

It will also be useful to define

$$\bar{d} = \left(\prod_{i=1}^r d_i \right)^{1/r}, \tag{3.14}$$

the geometric mean of the singular values $d_1, \dots, d_r$. Inserting our default choices for a and $\boldsymbol{\nu}$ into $\mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_N | a, \boldsymbol{\nu}]$, and noting that

$$\prod_{i=1}^r \nu_i^{-1/2} = (\bar{d}/d_r)^{-r}, \tag{3.15}$$

we obtain our recommended Bayes factor which we denote by gBF (g-prior

Bayes Factor):

$$
\begin{aligned}
&g\mathrm{BF}[\mathcal{M}_\gamma;\mathcal{M}_N] = \mathrm{BF}[\mathcal{M}_\gamma;\mathcal{M}_N|a_*,\boldsymbol{\nu}_*] \\
&= \begin{cases} \left\{\dfrac{\bar{d}}{d_q}\right\}^{-q} \dfrac{B(\frac{q}{2}+\frac{1}{4},\frac{n-q}{2}-\frac{3}{4})}{B(\frac{1}{4},\frac{n-q}{2}-\frac{3}{4})} \dfrac{(1-R^2+d_q^2\|\hat{\boldsymbol{\beta}}_{LS}\|^2)^{-\frac{1}{4}-\frac{q}{2}}}{(1-R^2)^{\frac{n-q}{2}-\frac{3}{4}}} & \text{if } q < n-1 \\ \left\{\bar{d}\times\|\hat{\boldsymbol{\beta}}_{LS}^{MP}\|\right\}^{-(n-1)} & \text{if } q \geq n-1, \end{cases}
\end{aligned}
\tag{3.16}
$$

which is a function of the key quantities q, R^2, the LS estimators and the singular values of the design matrix.

Remark 3.3. Like traditional selection criteria such as AIC, BIC and RIC, the gBF criterion (3.16) rewards models for explained variation through R^2. However, gBF also rewards models for stability of estimation through smaller values of $\bar{d}/d_q$ and $d_q\|\hat{\boldsymbol{\beta}}_{LS}\|$ for $q < n-1$, and through smaller values of the product $\bar{d}/d_{n-1}$ and $d_{n-1}\|\hat{\boldsymbol{\beta}}_{LS}^{MP}\|$ for $q \geq n-1$, the case where R^2 is unavailable.

To see how these various quantities bear on stable estimation, note first that

$$\bar{d}/d_r = \prod_{i=1}^{r}(d_i/d_r), \tag{3.17}$$

which gets smaller as the d_i/d_r ratios get smaller. Like the well-known condition number d_1/d_r, smaller values of (3.17) indicate a more stable design matrix $\boldsymbol{X}_\gamma$.

For $d_q\|\hat{\boldsymbol{\beta}}_{LS}\|$ and $d_{n-1}\|\hat{\boldsymbol{\beta}}_{LS}^{MP}\|$, note that each of these can be expressed as

$$d_r^2\|\hat{\boldsymbol{\beta}}\|^2 = \sum_{i=1}^{r}\left(\frac{d_r}{d_i}\right)^2\left\{\frac{(\boldsymbol{u}_i'\boldsymbol{v})}{\|\boldsymbol{u}_i\|\|\boldsymbol{v}\|}\right\}^2 = \sum_{i=1}^{r}\left(\frac{d_r}{d_i}\right)^2\{\mathrm{cor}(\boldsymbol{u}_i,\boldsymbol{y})\}^2. \tag{3.18}$$

Thus, for a given set of d_i/d_r ratios, (3.18) gets smaller if the larger correlations $\mathrm{cor}(\boldsymbol{u}_i,\boldsymbol{y})$ correspond to the larger d_i. Again, this is a measure of stability, as the largest principal components $d_i\boldsymbol{u}_i$ are the ones which are most stably estimated.

4. The Choice of Hyperparameters

In Section 3, we proposed the prior form $p(g)$ given by (2.15) with hyperparameters a and b, recommending the choices $a = -3/4$ and $b =$

$(n-q-5)/2-a$ for the case $q<n-1$ where the prior on g matters. In the following subsections, we show some appealing consequences of these choices.

4.1. The effect of a on the tail behavior of $p(\boldsymbol{\beta}|\sigma^2)$

Combining $p(\boldsymbol{\beta}|g,\sigma^2)$ in (2.12) with $p(g)$ in (2.15), the the probability density of $\boldsymbol{\beta}$ given σ^2 is given by

$$p(\boldsymbol{\beta}|\sigma^2)=\int_0^\infty \frac{\phi_q(\boldsymbol{W}'\boldsymbol{\beta};\mathbf{0},\sigma^2\boldsymbol{\Psi}_q(g,\boldsymbol{\nu}))}{B(a+1,b+1)}\frac{g^b}{(1+g)^{a+b+2}}dg. \tag{4.1}$$

To examine the asymptotic behavior of the density $p(\boldsymbol{\beta}|\sigma^2)$ as $\|\boldsymbol{\beta}\|\to\infty$, we appeal to the Tauberian theorem for the Laplace transform (see Geluk and de Haan (1987)), which tells us that the contribution of the integral (4.2) around zero becomes negligible as $\|\boldsymbol{\beta}\|\to\infty$. Thus we have only to consider the integration between ν_1 and ∞ (the major term).

Since $d_1\geq\cdots\geq d_q$, and assuming $\nu_1\geq\cdots\geq\nu_q$, we have

$$\frac{d_q^2}{(\nu_1+1)g}\leq\frac{d_i^2}{\nu_i+\nu_i g-1}\leq\frac{d_1^2}{\nu_q g}, \tag{4.2}$$

for $g\geq\nu_1$ and any i, which implies

$$C\frac{d_q^q}{(\nu_1+1)^{q/2}}\int_{\nu_1}^\infty\left(\frac{g}{g+1}\right)^{a+b+2}\left(\frac{1}{g}\right)^{q/2+a+2}\exp\left(-\frac{1}{g}\frac{d_1^2\|\boldsymbol{W}'\boldsymbol{\beta}\|^2}{2\nu_q\sigma^2}\right)dg$$

$$\leq \text{the major term of } p(\boldsymbol{\beta}|\sigma^2)$$

$$\leq C\frac{d_1^q}{\nu_q^{q/2}}\int_{\nu_1}^\infty\left(\frac{g}{g+1}\right)^{a+b+2}\left(\frac{1}{g}\right)^{q/2+a+2}\exp\left(-\frac{1}{g}\frac{d_q^2\|\boldsymbol{W}'\boldsymbol{\beta}\|^2}{2(\nu_1+1)\sigma^2}\right)dg$$

where $C=\{B(a+1,b+1)\}^{-1}(2\pi\sigma^2)^{-q/2}$. Thus, by the Tauberian theorem, there exist $C_1<C_2$ such that

$$C_1<\frac{\|\boldsymbol{\beta}\|^{q+2a+2}}{(\sigma^2)^{a+1}}p(\boldsymbol{\beta}|\sigma^2)<C_2 \tag{4.3}$$

for sufficiently large $\|\boldsymbol{\beta}\|$.

From (4.3), we see that the asymptotic tail behavior of $p(\boldsymbol{\beta}|\sigma^2)$ is determined by a and unaffected by b. Smaller a yields flatter tail behavior thereby diminishing the prior influence of $p(\boldsymbol{\beta}|\sigma^2)$. For $a = -1/2$ the asymptotic tail behavior of $p(\boldsymbol{\beta}|\sigma^2)$, $\|\boldsymbol{\beta}\|^{-q-1}$, corresponds to that of multivariate Cauchy distribution recommended by Zellner and Siow (1980). In contrast, the asymptotic tail behavior of our choice $a = -3/4$, $\|\boldsymbol{\beta}\|^{-q-1/2}$, is even flatter than that of the multivariate Cauchy distribution.

4.2. The effect of b on the implicit $O(n)$ choice of g

For implementations of the original g-prior (1.5), Zellner (1986) and others have recommended choices for which $g = O(n)$. This prevents the g-prior from asymptotically dominating the likelihood which would occur if g was unchanged as n increased. The recommendation of choosing $g = O(n)$ also applies to the choice of a fixed g for the generalized g-prior (2.12) where

$$\mathrm{tr}\{\mathrm{Var}(\boldsymbol{\beta}|g,\sigma^2)\} = \sigma^2 \sum_{i=1}^{q} \frac{\nu_i + \nu_i g - 1}{d_i^2}.$$

Since $d_i^2 = O(n)$ for $1 \le i \le q$ by Lemma 6.1, $\mathrm{tr}\{\mathrm{Var}(\boldsymbol{\beta}|g,\sigma^2)\} = gO(n^{-1})$ if ν_i is bounded. Therefore the choice $g = O(n)$ will also prevent the generalized g-prior from asymptotically dominating the likelihood, and stabilize it in the sense that $\mathrm{tr}\{\mathrm{Var}(\boldsymbol{\beta}|g,\sigma^2)\} = O(1)$.

For our fully Bayes case, where g is treated as a random variable, our choice of b, in addition to yielding a closed form for the marginal density in (3.9), also yields an implicit $O(n)$ choice of g, in the sense that

$$\begin{aligned}[\text{mode of } g] &= \frac{b}{a+2} = \frac{2(n-q)-7}{5} \\ \frac{1}{E[g^{-1}]} &= \frac{b}{a+1} = 2(n-q)-7\end{aligned}$$

for our recommended choices $a = -3/4$ and $b = (n-q-5)/2 - a$. (Note that $E[g]$ does not exist under the choice $a = -3/4$.

5. Shrinkage Estimation Conditionally on a Model

In this section, we consider estimation conditionally on a model $\mathcal{M}_\gamma$. Because $\boldsymbol{\beta}$ is not identifiable when $q > n-1$, and hence not estimable, we instead focus on estimation of $\boldsymbol{X\beta}$, the prediction of Y, which is always estimable. For this purpose, we consider estimation of $\boldsymbol{X\beta}$ under scaled quadratic loss

$(\boldsymbol{\delta}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{Q}(\boldsymbol{\delta}-\boldsymbol{X}\boldsymbol{\beta})/\sigma^2$ for positive-definite $\boldsymbol{Q}$. The Bayes estimator under this loss for any $\boldsymbol{Q}$ is of the form

$$\boldsymbol{X}\hat{\boldsymbol{\beta}}_B = \boldsymbol{X}E[\sigma^{-2}\boldsymbol{\beta}|\boldsymbol{y}]/E[\sigma^{-2}|\boldsymbol{y}]. \tag{5.1}$$

As will be seen from calculations similar to those in Section 3, under our priors, a simple closed form can be obtained for this estimator. In contrast, such a simple closed form is not available for the usual Bayes estimator, $\boldsymbol{X}E[\boldsymbol{\beta}_\gamma|\boldsymbol{y}]$, the posterior mean under $(\boldsymbol{\delta}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{Q}(\boldsymbol{\delta}-\boldsymbol{X}\boldsymbol{\beta})$ which does not scale for the variance σ^2.

We proceed by finding a simple closed form for $\hat{\boldsymbol{\beta}}_B$ in (5.1). Making use of the transformation (3.4), and by the calculation in (3.5), $E[\boldsymbol{\beta}_\#|\boldsymbol{y}] = E[\boldsymbol{\beta}_\#]$ (say $\boldsymbol{\mu}_\#$) and

$$\begin{aligned}\boldsymbol{W}\frac{E[\sigma^{-2}\boldsymbol{\beta}_*|\boldsymbol{y}]}{E[\sigma^{-2}|\boldsymbol{y}]} &= \frac{1}{E[\sigma^{-2}|\boldsymbol{y}]}E\left[\sigma^{-2}\sum_{i=1}^r \frac{\boldsymbol{u}_i'\boldsymbol{v}}{d_i}\left\{1-\frac{1}{\nu_i(1+g)}\right\}\boldsymbol{w}_i|\boldsymbol{y}\right] \\ &= \sum_{i=1}^r \frac{\boldsymbol{u}_i'\boldsymbol{v}}{d_i}\left\{1-\frac{H(\boldsymbol{y})}{\nu_i}\right\}\boldsymbol{w}_i\end{aligned}$$

where

$$H(\boldsymbol{y}) = \frac{E[\sigma^{-2}(1+g)^{-1}|\boldsymbol{y}]}{E[\sigma^{-2}|\boldsymbol{y}]}. \tag{5.2}$$

Thus

$$\hat{\boldsymbol{\beta}}_B = \sum_{i=1}^r \frac{\boldsymbol{u}_i'\boldsymbol{v}}{d_i}\left(1-\frac{H(\boldsymbol{y})}{\nu_i}\right)\boldsymbol{w}_i + \begin{cases}\boldsymbol{0} & \text{if } q \le n-1 \\ \boldsymbol{W}_\#\boldsymbol{\mu}_\# & \text{if } q > n-1.\end{cases} \tag{5.3}$$

Since $\boldsymbol{\beta}$ is not identifiable when $q \ge n-1$, it is not surprising that $\hat{\boldsymbol{\beta}}_B$ is incompletely defined due to the arbitrariness of $\boldsymbol{W}_\#\boldsymbol{\mu}_\#$. However, because $\boldsymbol{X}\boldsymbol{W}_\# = \boldsymbol{0}$, this arbitrariness is not an issue for the estimation of $\boldsymbol{X}\boldsymbol{\beta}$, for which we obtain

$$\boldsymbol{X}\hat{\boldsymbol{\beta}}_B = \sum_{i=1}^r (\boldsymbol{u}_i'\boldsymbol{v})\boldsymbol{u}_i\left(1-\frac{H(\boldsymbol{y})}{\nu_i}\right). \tag{5.4}$$

It now only remains to obtain a closed form for $H(\boldsymbol{y})$. As in (3.3), (3.6) and (3.7) in Section 3,

$$
\begin{aligned}
&\int_{-\infty}^{\infty}\int_{R^q}\int_0^{\infty}\frac{1}{\sigma^2}p(\boldsymbol{y}|\alpha,\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta}|g,\sigma^2)\frac{1}{\sigma^2}d\alpha\, d\boldsymbol{\beta}\, d\sigma^2 \\
&= \int_0^{\infty}\{\sigma^2\}^{-(n+1)/2}\frac{n^{1/2}}{(2\pi)^{(n-1)/2}}\frac{(1+g)^{-r/2}}{\prod_{i=1}^r \nu_i^{1/2}} \\
&\qquad \times \exp\left(-\frac{\|\boldsymbol{v}\|^2\{g(1-R^2)+1-Q^2\}}{2\sigma^2(g+1)}\right)\frac{1}{\sigma^2}d\sigma^2 \qquad (5.5)\\
&= \frac{2n^{1/2}\Gamma(\{n+1\}/2)}{\pi^{(n-1)/2}}\frac{\|\boldsymbol{v}\|^{-n-1}}{\prod_{i=1}^r \nu_i^{1/2}}(1+g)^{-r/2+(n+1)/2} \\
&\qquad \times \left\{g(1-R^2)+1-Q^2\right\}^{-(n+1)/2},
\end{aligned}
$$

which differs slightly from (3.7) because of the extra $1/\sigma^2$ term in the first expression. Letting

$$
L(\boldsymbol{y}|g) = (1+g)^{-r/2+(n+1)/2}\left\{g(1-R^2)+1-Q^2\right\}^{-(n+1)/2}, \qquad (5.6)
$$

we have

$$
\begin{aligned}
H(\boldsymbol{y}) &= \frac{\int_0^{\infty}(1+g)^{-1}L(\boldsymbol{y}|g)p(g)dg}{\int_0^{\infty}L(\boldsymbol{y}|g)p(g)dg} \\
&= \frac{\int_0^{\infty}(1+g)^{-r/2+(n-1)/2}\{g(1-R^2)+1-Q^2\}^{-(n+1)/2}p(g)dg}{\int_0^{\infty}(1+g)^{-r/2+(n+1)/2}\{g(1-R^2)+1-Q^2\}^{-(n+1)/2}p(g)dg}.
\end{aligned}
$$

When $q<n-1$, under the prior (2.15) used in Section 3, namely

$$
p(g) = \frac{g^b(1+g)^{-a-b-2}}{B(a+1,b+1)} = \frac{g^b(1+g)^{-(n-r-1)/2}}{B(a+1,b+1)}
$$

where $b=(n-5)/2-r/2-a$, we have

$$
\begin{aligned}
H(\boldsymbol{y}) &= \frac{\int_0^{\infty}g^b\{g(1-R^2)+1-Q^2\}^{-(n+1)/2}dg}{\int_0^{\infty}g^b(1+g)\{g(1-R^2)+1-Q^2\}^{-(n+1)/2}dg} \\
&= \left(1+\frac{\int_0^{\infty}g^{b+1}\{g(1-R^2)+1-Q^2\}^{-(n+1)/2}dg}{\int_0^{\infty}g^b\{g(1-R^2)+1-Q^2\}^{-(n+1)/2}dg}\right)^{-1} \\
&= \left(1+\frac{1-Q^2}{1-R^2}\frac{B(q/2+a+1,b+2)}{B(q/2+a+2,b+1)}\right)^{-1} \\
&= \left(1+\frac{1-Q^2}{1-R^2}\frac{(n-q-3)/2-a}{q/2+a+1}\right)^{-1}.
\end{aligned}
$$

On the other hand, when $q \geq n-1$, it follows that $R^2 = 1$, $r = n-1$, $L(\boldsymbol{y}|g) = (1+g)(1-Q^2)^{-(n+1)/2}$ and hence

$$H(\boldsymbol{y}) = \frac{\int_0^\infty p(g)dg}{\int_0^\infty (1+g)p(g)dg} = \{1+E[g]\}^{-1}. \tag{5.7}$$

Thus, when $q \geq n-1$, we must specify the mean of prior density of g, although no such specification was needed for model selection. A reasonable specification may be $E[g] = d_{n-1}^2/d_1^2$, a function of the condition number d_1/d_{n-1} of the linear equation. For extremely large values of d_1/d_{n-1}, the coefficients of the first and the last terms in (5.4) become nearly 1 and 0, respectively. See Casella (1985) and Maruyama and Strawderman (2005) for further discussion of the condition number.

Thus, for our recommended choices of hyperparameters $a = -3/4$ and $\nu_i = d_i^2/d_r^2$ for $1 \leq i \leq r$, our recommended estimator of $\boldsymbol{X\beta}$ for a given model $\mathcal{M}_\gamma$ is

$$X\hat{\boldsymbol{\beta}}_B = \sum_{i=1}^r (\boldsymbol{u}_i'\boldsymbol{v})\left(1 - \{d_r^2/d_i^2\}H(\boldsymbol{y})\right)\boldsymbol{u}_i, \tag{5.8}$$

where

$$H(\boldsymbol{y}) = \begin{cases} \left(1 + \frac{1-R^2+d_q^2\|\hat{\boldsymbol{\beta}}_{LS}\|^2}{1-R^2}\frac{n/2-q/2-3/4}{q/2+1/4}\right)^{-1} & \text{if } q < n-1 \\ \left(1+d_{n-1}^2/d_1^2\right)^{-1} & \text{if } q \geq n-1. \end{cases} \tag{5.9}$$

6. Model selection consistency

In this section, we consider the model selection consistency in the case where p is fixed and n approaches infinity. Posterior consistency for model choice means

$$\operatorname*{plim}_{n\to\infty} \Pr(\mathcal{M}_T|y) = 1 \quad \text{when } \mathcal{M}_T \text{ is the true model},$$

where plim denotes convergence in probability under the true model $\mathcal{M}_T$, namely $\boldsymbol{y} = \alpha_T\boldsymbol{1}_n + \boldsymbol{X}_T\boldsymbol{\beta}_T + \boldsymbol{\epsilon}$ 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 and $\boldsymbol{\epsilon}_n \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$.

Let us show that our general criterion, $\mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_N|a, \boldsymbol{\nu}]$ given by (3.11) with bounded ν_1, is model selection consistent. This is clearly equivalent to

$$\operatorname*{plim}_{n\to\infty} \frac{\mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_N|a, \boldsymbol{\nu}]}{\mathrm{BF}[\mathcal{M}_T; \mathcal{M}_N|a, \boldsymbol{\nu}]} = 0 \quad \forall \mathcal{M}_\gamma \neq \mathcal{M}_T. \tag{6.1}$$

Recall that we have already assumed that $\boldsymbol{x}_i'\boldsymbol{1}_n = 0$ and $\boldsymbol{x}_i'\boldsymbol{x}_i/n = 1$ for any $1 \leq i \leq p$. To obtain model selection consistency, we also assume:

A1. The correlation between x_i and x_j, $\boldsymbol{x}_i'\boldsymbol{x}_j/n$, has a limit as $n \to \infty$.
A2. 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 is the standard assumption which also appears in Knight and Fu (2000) and Zou (2006). A2 is natural because the columns of $\boldsymbol{X}_F$ are assumed to be linearly independent. Note that A2 implies that, for any model $\mathcal{M}_\gamma$, there exists a positive definite matrix $\boldsymbol{H}_\gamma$ such that

$$\lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}_\gamma'\boldsymbol{X}_\gamma = \boldsymbol{H}_\gamma. \tag{6.2}$$

Under these assumptions, we will give following lemmas (Lemma 6.1 on $\boldsymbol{X}_T$ and $\boldsymbol{X}_\gamma$ and Lemmas 6.2, 6.3 on R_T^2 and R_γ^2) for our main proof. See also Fernández, Ley and Steel (2001) and Liang *et al.* (2008).

Lemma 6.1. *1. Let $d_1[\gamma]$ and $d_q[\gamma]$ be the maximum and minimum of singular values of $\boldsymbol{X}_\gamma$. Then $\{d_1[\gamma]\}^2/n$ and $\{d_q[\gamma]\}^2/n$ approach the maximum and minimum eigenvalues of $\boldsymbol{H}_\gamma$, respectively.*

2. The $q_T \times q_T$ limit

$$\lim_{n\to\infty} n^{-1}\boldsymbol{X}_T'\boldsymbol{X}_\gamma(\boldsymbol{X}_\gamma'\boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma'\boldsymbol{X}_T = \boldsymbol{H}(T,\gamma) \tag{6.3}$$

exists.

3. When $\gamma \not\supseteq T$, the rank of $\boldsymbol{H}_T - \boldsymbol{H}(T,\gamma)$ is given by the number of non-overlapping predictors and $\boldsymbol{\beta}_T'\boldsymbol{H}_T\boldsymbol{\beta}_T > \boldsymbol{\beta}_T'\boldsymbol{H}(T,\gamma)\boldsymbol{\beta}_T$.

4. $\boldsymbol{H}_T - \boldsymbol{H}(T,\gamma) = \boldsymbol{0}$ for $\gamma \supsetneq T$.

Lemma 6.2. *Let $\gamma \not\supseteq T$. Then*

$$\operatorname*{plim}_{n\to\infty} R_\gamma^2 = \frac{\boldsymbol{\beta}_T'\boldsymbol{H}(\gamma,T)\boldsymbol{\beta}_T}{\sigma^2 + \boldsymbol{\beta}_T'\boldsymbol{H}_T\boldsymbol{\beta}_T}\left(< \frac{\boldsymbol{\beta}_T'\boldsymbol{H}_T\boldsymbol{\beta}_T}{\sigma^2 + \boldsymbol{\beta}_T'\boldsymbol{H}_T\boldsymbol{\beta}_T}\right). \tag{6.4}$$

Proof. For the submodel $\mathcal{M}_\gamma$, $1 - R_\gamma^2$ is given by

$$\|\boldsymbol{Q}_\gamma(\boldsymbol{y} - \bar{y}\boldsymbol{1}_n)\|^2/\|\boldsymbol{y} - \bar{y}\boldsymbol{1}_n\|^2$$

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

$$\begin{aligned}\|\boldsymbol{Q}_\gamma(\boldsymbol{y} - \bar{y}\boldsymbol{1}_n)\|^2 &= \|\boldsymbol{Q}_\gamma\boldsymbol{X}_T\boldsymbol{\beta}_T + \boldsymbol{Q}_\gamma\check{\boldsymbol{\epsilon}}\|^2 \\ &= \boldsymbol{\beta}_T'\boldsymbol{X}_T'\boldsymbol{Q}_\gamma\boldsymbol{X}_T\boldsymbol{\beta}_T + 2\boldsymbol{\beta}_T'\boldsymbol{X}_T'\boldsymbol{Q}_\gamma\boldsymbol{\epsilon} + \check{\boldsymbol{\epsilon}}'\boldsymbol{Q}_\gamma\check{\boldsymbol{\epsilon}}\end{aligned} \tag{6.5}$$

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

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

Hence $1 - R_\gamma^2$ can be rewritten as

$$\frac{\boldsymbol{\beta}_T'\{\boldsymbol{X}_T'\boldsymbol{Q}_\gamma\boldsymbol{X}_T/n\}\boldsymbol{\beta}_T + 2\boldsymbol{\beta}_T'\{\boldsymbol{X}_T'\boldsymbol{Q}_\gamma\boldsymbol{\epsilon}/n\} + \|\boldsymbol{Q}_\gamma\check{\boldsymbol{\epsilon}}\|^2/n}{\boldsymbol{\beta}_T'\{\boldsymbol{X}_T'\boldsymbol{X}_T/n\}\boldsymbol{\beta}_T + 2\boldsymbol{\beta}_T'\{\boldsymbol{X}_T'\boldsymbol{\epsilon}/n\} + \|\check{\boldsymbol{\epsilon}}\|^2/n}. \tag{6.6}$$

In (6.6), $\boldsymbol{\beta}_T'\boldsymbol{X}_T'\boldsymbol{\epsilon}/n$ approaches 0 in probability because $E[\boldsymbol{\epsilon}] = \mathbf{0}$, $\mathrm{var}[\boldsymbol{\epsilon}] = \sigma^2\boldsymbol{I}_n$, $E[\boldsymbol{X}_T'\boldsymbol{\epsilon}/n] = \mathbf{0}$, and

$$\mathrm{var}\left(\boldsymbol{X}_T'\boldsymbol{\epsilon}/n\right) = n^{-1}\sigma^2\{\boldsymbol{X}_T'\boldsymbol{X}_T/n\} \to \mathbf{0}. \tag{6.7}$$

Similarly $\boldsymbol{\beta}_T'\{\boldsymbol{X}_T'\boldsymbol{Q}_\gamma\boldsymbol{\epsilon}/n\} \to 0$ in probability. Further both $\|\check{\boldsymbol{\epsilon}}\|^2/n$ and $\|\boldsymbol{Q}_\gamma\check{\boldsymbol{\epsilon}}\|^2/n$ for any γ converge to σ^2 in probability.

Therefore, by parts 2 and 3 of Lemma 6.1, R_γ^2 for $\gamma \not\supseteq T$ approaches

$$\frac{\boldsymbol{\beta}_T'\boldsymbol{H}(\gamma, T)\boldsymbol{\beta}_T}{\sigma^2 + \boldsymbol{\beta}_T'\boldsymbol{H}_T\boldsymbol{\beta}_T}\left(< \frac{\boldsymbol{\beta}_T'\boldsymbol{H}_T\boldsymbol{\beta}_T}{\sigma^2 + \boldsymbol{\beta}_T'\boldsymbol{H}_T\boldsymbol{\beta}_T}\right)$$

in probability. □

Lemma 6.3. *Let $\gamma \supsetneq T$. Then*

1. *$R_\gamma^2 \geq R_T^2$ for any n and*

$$\operatorname*{plim}_{n\to\infty} R_T^2 = \operatorname*{plim}_{n\to\infty} R_\gamma^2 = \frac{\boldsymbol{\beta}_T'\boldsymbol{H}_T\boldsymbol{\beta}_T}{\sigma^2 + \boldsymbol{\beta}_T'\boldsymbol{H}_T\boldsymbol{\beta}_T}. \tag{6.8}$$

2. *$\{(1 - R_T^2)/(1 - R_\gamma^2)\}^n$ is bounded from above in probability.*

Proof. 1. When $\gamma \supsetneq T$, $\boldsymbol{Q}_\gamma\boldsymbol{X}_T = \mathbf{0}$. Hence, as in (6.6), we have

$$\begin{aligned} 1 - R_\gamma^2 &= \frac{\|\boldsymbol{Q}_\gamma\check{\boldsymbol{\epsilon}}\|^2/n}{\boldsymbol{\beta}_T'\{\boldsymbol{X}_T'\boldsymbol{X}_T/n\}\boldsymbol{\beta}_T + 2\boldsymbol{\beta}_T'\{\boldsymbol{X}_T'\boldsymbol{\epsilon}/n\} + \|\check{\boldsymbol{\epsilon}}\|^2/n} \\ 1 - R_T^2 &= \frac{\|\boldsymbol{Q}_T\check{\boldsymbol{\epsilon}}\|^2/n}{\boldsymbol{\beta}_T'\{\boldsymbol{X}_T'\boldsymbol{X}_T/n\}\boldsymbol{\beta}_T + 2\boldsymbol{\beta}_T'\{\boldsymbol{X}_T'\boldsymbol{\epsilon}/n\} + \|\check{\boldsymbol{\epsilon}}\|^2/n}. \end{aligned} \tag{6.9}$$

Since $\|\boldsymbol{Q}_T\check{\boldsymbol{\epsilon}}\|^2/n > \|\boldsymbol{Q}_\gamma\check{\boldsymbol{\epsilon}}\|^2/n$ for any n and both approach σ^2 in probability, part 1 follows.

2. By (6.9), $(1-R_T^2)/(1-R_\gamma^2)$ is given by $\|\boldsymbol{Q}_T\check{\boldsymbol{\epsilon}}\|^2/\|\boldsymbol{Q}_\gamma\check{\boldsymbol{\epsilon}}\|^2$. Further we have
$$1 \le \frac{1-R_T^2}{1-R_\gamma^2} = \frac{\|\boldsymbol{Q}_T\check{\boldsymbol{\epsilon}}\|^2}{\|\boldsymbol{Q}_\gamma\check{\boldsymbol{\epsilon}}\|^2} \le \frac{\|\check{\boldsymbol{\epsilon}}\|^2}{\|\boldsymbol{Q}_\gamma\check{\boldsymbol{\epsilon}}\|^2} = \frac{1}{W_\gamma}$$
where $W_\gamma \sim (1+\chi^2_{q_\gamma}/\chi^2_{n-q_\gamma-1})^{-1}$, for independent $\chi^2_{n-q_\gamma-1}$ and $\chi^2_{q_\gamma}$. Hence
$$\begin{aligned}\left\{1+\chi^2_{q_\gamma}/\chi^2_{n-q_\gamma-1}\right\}^{-n} &= \left\{1+\left\{n/\chi^2_{n-q_\gamma-1}\right\}\left\{\chi^2_{q_\gamma}/n\right\}\right\}^{-n} \\ &\sim \exp(-\chi^2_{q_\gamma}) \text{ as } n\to\infty\end{aligned}$$
since $\chi^2_{n-q_\gamma-1}/n \to 1$ in probability. Therefore W_γ^{-n} is bounded in probability from above and part 2 follows.

□

Our main consistency theorem is as follows.

Theorem 6.1. *Under assumptions A1 and A2, if ν_1 is bounded, then* $\mathrm{BF}[\mathcal{M}_\gamma;\mathcal{M}_N|a,\boldsymbol{\nu}]$ *is consistent for model selection.*

Note that our recommended choice $\nu_1 = d_1^2/d_q^2$ is bounded by Lemma 6.1.

Proof. Note that
$$\nu_1^{-1} \le 1-Q_\gamma^2 \le 1$$
by (3.6),
$$\nu_1^{-q/2} \le \prod_{i=1}^{q}\nu_i^{-1/2} \le 1$$
because the ν_i's are descending,
$$\frac{B(q/2+a+1,(n-q-3)/2-a)}{B(a+1,(n-q-3)/2-a)} = \frac{\Gamma(q/2+a+1)}{\Gamma(a+1)}\frac{\Gamma(\{n-q-1\}/2)}{\Gamma(\{n-1\}/2)},$$
and
$$\lim_{n\to\infty}(n/2)^{q/2}\frac{\Gamma(\{n-q-1\}/2)}{\Gamma(\{n-1\}/2)} = 1$$
by Stirling's formula. Then, by (3.11), there exist $c_1(\gamma) < c_2(\gamma)$ (which do not depend on n) such that
$$c_1(\gamma) < \left\{n^{q_\gamma}(1-R_\gamma^2)^n\right\}^{1/2}\frac{\mathrm{BF}[\mathcal{M}_\gamma;\mathcal{M}_N|a,\nu]}{(1-R_\gamma^2)^{(q_\gamma+3)/2+a}} < c_2(\gamma)$$

for sufficiently large n. By Lemmas 6.2 and 6.3, R^2_γ goes to some constant in probability. Hence, to show consistency, it suffices to show that

$$\operatorname*{plim}_{n\to\infty} n^{q_T - q_\gamma} \left(\frac{1 - R^2_T}{1 - R^2_\gamma} \right)^n = 0. \tag{6.10}$$

Consider the following two situations:

1. $\gamma \not\supseteq T$: By Lemma 6.2 and 6.3, $(1 - R^2_T)/(1 - R^2_\gamma)$ is strictly less than 1 in probability. Hence $\{(1-R^2_T)/(1-R^2_\gamma)\}^n$ converges to zero in probability exponentially fast with respect to n. Therefore, no matter what value $q_T - q_\gamma$ takes, (6.10) is satisfied.
2. $\gamma \supsetneq T$: By Lemma 6.3, $\{(1 - R^2_T)/(1 - R^2_\gamma)\}^n$ is bounded in probability. Since $q_\gamma > q_T$, (6.10) is satisfied.

□

7. Simulated Performance Evaluations

In this section, we report on a number of simulated performance comparisons between our recommended Bayes factor $g\mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_N]$ and the following selection criteria:

$$\begin{aligned}
\mathrm{ZE} &= (1 - R^2)^{-(n-q)/2+3/4} \frac{B(q/2 + 1/4, (n-q)/2 - 3/4)}{B(1/4, (n-q)/2 - 3/4)} \\
\mathrm{EB} &= \max_g m_\gamma(\boldsymbol{y}|g, \hat{\sigma}^2) \\
\mathrm{AIC} &= -2 \times \text{maximum log likelihood} + 2(q+2) \\
\mathrm{AICc} &= -2 \times \text{maximum log likelihood} + 2(q+2)\frac{n}{n-q-3} \\
\mathrm{BIC} &= -2 \times \text{maximum log likelihood} + q \log n.
\end{aligned}$$

Here, ZE is the special case of $\mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_N]$ with $a = -3/4$ and $\nu_1 = \cdots = \nu_q = 1$ (corresponding to Zellner's g-prior). Note that comparisons of gBF with ZE should reveal the effect of our choice of descending ν. EB is the empirical Bayes criterion of George and Foster (2000) in (1.6), also based on the original g-prior, with $\hat{\sigma}^2 = \mathrm{RSS}_\gamma/(n - q_\gamma - 1)$ plugged-in. Finally, AICc is the well-known correction of AIC proposed by Hurvich and Tsai (1989).

For these comparisons, we consider data generated by submodels (1.2) of (1.1) with $p = 16$ potential predictors for two different choices of the underlying design matrix $\boldsymbol{X}_F$. For the first choice, which we refer to as the correlated case, each row of the 16 predictors are generated as $x_1, \ldots, x_{13} \sim$

$N(0,1)$, and $x_{14}, x_{15}, x_{16} \sim U(-1,1)$ (the uniform distribution) with the following pairwise correlations

$$\overbrace{x_1, x_2}^{\text{cor}=0.9},\ \underbrace{x_3, x_4}_{\text{cor}=-0.7},\overbrace{x_5, x_6}^{\text{cor}=0.5},\ \underbrace{x_7, x_8}_{\text{cor}=-0.3},\overbrace{x_9, x_{10}}^{\text{cor}=0.1} \tag{7.1}$$

and independently otherwise. For the second choice, which we refer to as the simple case, each row of the 16 predictors are generated as $x_1, \dots, x_{16}$ iid $\sim N(0,1)$.

For our first set of comparisons, we set $n = 30$ (larger than $p = 16$) and considered 4 submodels where the true predictors are

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

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

$$Y = 1 + 2 \sum_{i \in \{\text{true}\}} x_i + \{\text{normal error term } N(0,1)\}. \tag{7.2}$$

Table 1 compares the criteria by how often the true model was selected as best, or in the top 3, among the 2^{16} candidate models across the $N = 500$ replications. We note the following

- In the correlated cases, EB, ZE and gBF were very similar for $q_T = 4, 8$, but gBF was much better for $q = 12, 16$.
- In the simple cases, gBF, ZE and EB were very similar suggesting no effect of our extension of Zellner's g-prior with descending ν.
- In both the correlated and simple cases, AIC and BIC were poor for all cases except $q_T = 16$.
- In both the correlated and simple cases, AICc was poor for $q_T =$ 16 and 4 but good for $q_T = 8, 12$.

Overall, Table 1 suggests that gBF is stable and good for most cases, and that our generalization of Zellner's g-prior is effective in the correlated case.

On data from the same setup with $n = 30$ and $N = 500$, Table 2 compares the models selected by each criterion based on their (in-sample) predictive error

$$\frac{(\hat{y}_* - \alpha_T 1_n - X_T \boldsymbol{\beta}_T)'(\hat{y}_* - \alpha_T 1_n - X_T \boldsymbol{\beta}_T)}{n\sigma^2}$$

TABLE 1
Rank of the true model

q_T	16		12		8		4	
rank	1st	1st-3rd	1st	1st-3rd	1st	1st-3rd	1st	1st-3rd
correlated case								
gBF	0.71	0.91	0.73	0.94	0.69	0.87	0.66	0.86
ZE	0.40	0.70	0.63	0.89	0.68	0.89	0.67	0.87
EB	0.41	0.71	0.63	0.90	0.67	0.88	0.66	0.85
AIC	0.95	0.99	0.23	0.38	0.09	0.17	0.05	0.08
AICc	0.25	0.45	0.67	0.90	0.52	0.75	0.25	0.44
BIC	0.88	0.98	0.41	0.65	0.31	0.43	0.23	0.42
simple case								
gBF	0.98	0.99	0.83	0.97	0.75	0.93	0.67	0.85
ZE	0.94	0.98	0.87	0.97	0.78	0.95	0.69	0.88
EB	0.95	0.98	0.87	0.98	0.76	0.95	0.65	0.87
AIC	1.00	1.00	0.22	0.37	0.08	0.13	0.05	0.08
AICc	0.82	0.87	0.85	0.97	0.55	0.80	0.24	0.46
BIC	0.99	1.00	0.41	0.65	0.27	0.46	0.22	0.39

TABLE 2
Prediction error comparisons

	16		12		8		4	
	mean	(LQ, UQ)	mean	(LQ,UQ)	mean	(LQ,UQ)	mean	(LQ,UQ)
correlated case								
oracle	0.57	(0.43, 0.68)	0.43	(0.31, 0.53)	0.30	(0.20, 0.38)	0.17	(0.09, 0.22)
gBF	0.70	(0.44, 0.78)	0.52	(0.32, 0.61)	0.37	(0.22, 0.47)	0.26	(0.11, 0.35)
ZE	1.02	(0.53, 1.20)	0.59	(0.35, 0.71)	0.41	(0.23, 0.53)	0.27	(0.11, 0.37)
EB	1.00	(0.52, 1.16)	0.58	(0.35, 0.70)	0.41	(0.23, 0.53)	0.27	(0.11, 0.37)
AIC	0.56	(0.42, 0.67)	0.54	(0.40, 0.65)	0.51	(0.37, 0.62)	0.48	(0.33, 0.59)
AICc	1.29	(0.65, 1.65)	0.56	(0.34, 0.68)	0.42	(0.25, 0.52)	0.36	(0.22, 0.47)
BIC	0.58	(0.42, 0.69)	0.53	(0.38, 0.64)	0.46	(0.31, 0.58)	0.39	(0.23, 0.51)
simple case								
oracle	0.57	(0.43, 0.68)	0.43	(0.31, 0.53)	0.30	(0.20, 0.38)	0.17	(0.09, 0.22)
gBF	0.57	(0.41, 0.67)	0.45	(0.33, 0.56)	0.35	(0.21, 0.45)	0.25	(0.12, 0.33)
ZE	0.66	(0.42, 0.70)	0.45	(0.32, 0.56)	0.34	(0.21, 0.44)	0.24	(0.12, 0.32)
EB	0.65	(0.42, 0.69)	0.45	(0.32, 0.56)	0.35	(0.21, 0.45)	0.25	(0.12, 0.34)
AIC	0.56	(0.42, 0.67)	0.54	(0.39, 0.65)	0.51	(0.37, 0.63)	0.48	(0.32, 0.60)
AICc	0.98	(0.45, 0.83)	0.46	(0.33, 0.55)	0.39	(0.25, 0.50)	0.35	(0.20, 0.47)
BIC	0.56	(0.42, 0.67)	0.52	(0.37, 0.64)	0.45	(0.30, 0.57)	0.38	(0.21, 0.50)

where X_T, α_T and $\boldsymbol{\beta}_T$ are the true $n \times q_T$ design matrix, the true intercept and the true coefficients. The prediction $\hat{y}_*$ for each selected model is given by $\bar{y}1_n + X_{\gamma*}\hat{\boldsymbol{\beta}}_{\gamma*}$ where $X_{\gamma*}$ is the selected design matrix, $\hat{\boldsymbol{\beta}}_{\gamma*}$ is the Bayes estimator for gBF, ZE and EB, and is the least squares estimator for AIC, BIC and AICc. To aid in gauging these comparisons, we also included the "oracle" prediction error, namely that based on the least squares estimate under the true model.

The summary statistics reported in Table 2 are the mean predictive error, and the lower quantile (LQ) and upper quantile (UQ) of the predictive errors. In terms of predictive performance, the comparisons are similar to those in Table 1. Overall, we see that gBF works well in this setting.

For our final evaluations, we use data again simulated from the simple form (7.2), but now with $x_1, x_2, \ldots, x_{12}, x_{14}, x_{15}$ as the true predictors ($q_T = 14$), and a small sample size $n = 12$ (smaller than $p = 16$). Since $p > q_T > n$, the true model is not identifiable here. Furthermore, AIC, BIC, AICc, ZE and EB cannot even be computed (because $p > n$) and so we confine our evaluations to gBF.

For this very difficult situation, gBF did not pick the true model as best even once across the $N = 500$ iterations. Nonetheless, we did find evidence that gBF is often partially correct. Tables 3 and 4 report the observed gBF selection frequencies of model sizes and individual predictors in this experiment. These show that gBF always chose as best, a model with less than 12 predictors (i.e., an identifiable model), and that the true individual predictors (designated by (T) in the table) were usually selected more often than not. Note that the selection of a parsimonious identifiable model is reasonable and often desirable.

Remark 7.1. The only variables that were under-selected by gBF in Table 4 were (x_3, x_4) and (x_{14}, x_{15}) in the correlated case. Although x_3 and x_4 are true predictors, their under-selection may be explained by the high negative correlation between them. Interestingly, the under-selection of x_{14} and x_{15} is not explained by correlation (as they are independent in both the correlated and simple cases). Rather, it suggests that selection of uniformly distributed predictors may be more difficult than normally distributed predictors (they are uniform in the correlated case and normal in the simple case).

Although gBF did not select the true model as best in this experiment, it often ranked the true model relatively high. This can be seen in Table 5, which summarizes the relative rank of the true model (rank/2^{16}) over the $N = 500$ iterations. (Note that smaller is better). The mean relative rank of the true model was 0.035 in the correlated case and 0.039 in the simple

TABLE 3
Model size frequencies in the many predictors case

	0–6	7	8	9	10	11	12–16
correlated	0.10	0.11	0.22	0.34	0.16	0.07	0.00
simple	0.11	0.15	0.21	0.33	0.14	0.06	0.00

TABLE 4
Predictor frequencies in the many predictors case

	x_1 (T)	x_2 (T)	x_3 (T)	x_4 (T)	x_5 (T)	x_6 (T)
correlated	0.65	0.63	0.44	0.46	0.62	0.60
simple	0.54	0.54	0.54	0.54	0.54	0.57
	x_7 (T)	x_8 (T)	x_9 (T)	x_{10} (T)	x_{11} (T)	x_{12} (T)
correlated	0.56	0.56	0.59	0.58	0.58	0.60
simple	0.55	0.55	0.54	0.56	0.52	0.50
	x_{13} (F)	x_{14} (T)	x_{15} (T)	x_{16} (F)		
correlated	0.40	0.43	0.45	0.40		
simple	0.34	0.55	0.57	0.39		

structure case. In both of these cases, these mean ranks were the largest mean ranks of all 2^{16} candidate models!

Although gBF ranked the true model as highest in an average sense, a smaller identifiable model was selected as best for each particular sample. To shed light on how gBF performed among the larger models, Table 6 reports the frequency with which the true model was ranked highly among the $(16 \times 15)/2 = 120$ candidate models with exactly 14 predictors. Here, gBF is clearly performing well. To our knowledge, we know of no other analytical selection criterion for choosing between models with $R^2 = 1$, which is the case here.

TABLE 5
The relative rank of the true model

	Min	LQ	Median	Mean	UQ	Max
correlated	0.001	0.012	0.023	0.035	0.042	0.518
simple	0.001	0.013	0.023	0.039	0.043	0.555

Table 6
Frequency that the true model was ranked highly among models with 14 *predictors*

	1st	1st-2nd	1st-3rd
correlated	0.14	0.22	0.26
simple	0.13	0.20	0.26

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control* **AC-19** 716–723. System identification and time-series analysis. MR0423716

Casella, G. (1980). Minimax ridge regression estimation. *Ann. Statist.* **8** 1036–1056. MR585702

Casella, G. (1985). Condition numbers and minimax ridge regression estimators. *J. Amer. Statist. Assoc.* **80** 753–758. MR803264

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

Fernández, C., Ley, E. and Steel, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics* **100** 381–427. MR1820410

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975. MR1329177

Geluk, J. L. and de Haan, L. (1987). *Regular variation, extensions and Tauberian theorems*. *CWI Tract* **40**. Stichting Mathematisch Centrum Centrum voor Wiskunde en Informatica, Amsterdam.

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

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76** 297–307. MR1016020

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. 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

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian inference and decision techniques. Stud. Bayesian Econometrics Statist.* **6** 233–243. North-Holland, Amsterdam. MR881437

Zellner, A. and Siow, A. (1980). Posterior Odds Ratios for Selected Regression Hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 585–603. University of Valencia.

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