Bayesian Analysis (2016) **11**, Number 2, pp. 353–380

A Two-Component G-Prior for Variable Selection

Hongmei Zhang[*], Xianzheng Huang[†], Jianjun Gan[‡], Wilfried Karmaus[§],
and Tara Sabo-Attwood[¶]

Abstract. We present a Bayesian variable selection method based on an extension of the Zellner's g-prior in linear models. More specifically, we propose a two-component G-prior, wherein a tuning parameter, calibrated by use of pseudo-variables, is introduced to adjust the distance between the two components. We show that implementing the proposed prior in variable selection is more efficient than using the Zellner's g-prior. Simulation results also indicate that models selected using the method with the two-component G-prior are generally more favorable with smaller losses compared to other methods considered in our work. The proposed method is further demonstrated using our motivating gene expression data from a lung disease study, and ozone data analyzed in earlier studies.

Keywords: Bayes factor, measurement error, mean squared loss, pseudo variables, tuning parameter.

1 Introduction

The present work is motivated by a genetic and epidemiologic study related to the severity of ideopathic pulmonary fibrosis (IPF). IPF is a chronical lung disease characterized by scarring of the supporting framework of lungs. The incidence and mortality continue to increase with more men being afflicted than women. However, reasons for this gender difference are unclear. The goal of that study is to identify genes whose expression levels are associated with the severity of IPF, or genes such that the association is gender-specific. For this type of applications, variable selection on linear regression models is commonly used. Besides the classical methods such as those built on Akaike information criterion (AIC) and Bayesian information criterion (BIC), many other variable selection methods have been proposed to select variables. For example, the adaptive LASSO in Zou (2006) is aimed to reduce the estimation bias in the original LASSO developed in Tibshirani (1996) to achieve the oracle properties, meaning that the method will correctly select the model as if the correct submodel were known. Another variable selection method also enjoying the oracle properties is the nonconcave penalized likelihood method developed by Fan and Li (2001), where a smoothly clipped absolute deviation (SCAD) penalty is introduced.

[*] Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Memphis, TN 38152, hzhang6@memphis.edu

[†] Department of Statistics, University of South Carolina, Columbia, SC 29208, huang@stat.sc.edu

[‡] GlaxoSmithKline, Research Triangle Park, NC 27709, jianjun.x.gan@gsk.com

[§] Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Memphis, TN 38152, karmaus1@memphis.edu

[¶] Department of Environmental and Global Health, University of Florida, Gainesville, FL 32611, sabo@phhp.ufl.edu

DOI: 10.1214/15-BA953

Bayesian methods for variable selection are discussed rigorously in the literature as well. In this work, we focus on variable selection in this framework. Compared to the frequentist approaches for variable selection, one major advantage of Bayesian methods exists in their ability to incorporate prior knowledge into the selection process. In addition, rather than selecting a unique set of variables as in frequentist approaches, Bayesian methods estimate the posterior probabilities for all models under consideration. Most work in this area uses g-priors proposed in Zellner (1986) and the spike and slab models (Mitchell and Beauchamp, 1988) for the prior distribution of regression coefficients (George and McCulloch, 1993, 1997, 1999; George and Foster, 2000; George, 2000; Smith and Kohn, 1996; Fernández et al., 2001; Lee et al., 2003; Ishwaran and Rao, 2005a,b). In this article, our focus is on the g-prior, a multivariate normal distribution with mean zero and covariance being a proportion g to the covariance of regression coefficient maximum likelihood estimators.

Choosing either too big or too small g will lead to exclusion of important variables (Lindley, 1957). This Lindley's paradox of Zellner's g-prior in variable selection confesses the importance of proper choice of g (Lindley, 1957; Bartlett, 1957). To this end, Liang et al. (2008) proposed using mixtures of g-priors in variable selection. Maruyama and George (2011), in light of an appropriate distribution of prior information on different variables, advocated a generalized g-prior based on the singular value decomposition of the design matrix. However, a single g is a global shrinkage factor that applies to all predictors and will equally shrink each coordinate. This choice is reasonable if we do not know anything about the predictors or if they are equivalent in some sense as noted in a recent study (Som et al., 2014). In this article, we extend the Zellner's g-prior from a single g to a diagonal matrix G to incorporate information in the predictors. We introduce a tuning parameter to optimize the prior. This tuning parameter is adaptively selected with the help of carefully created pseudo-variables. Compared to the existing methods, a major contribution of this study exists in the extension of the classical g-prior and the inclusion of pseudo-variables to optimize the extended g-prior.

It is not unusual that gene expressions are measured with error (Rocke and Durbin, 2001). Given the high impact of measurement errors on inference and variable selection as discussed in Higgins et al. (1997), Carroll et al. (2006), Liu and Wu (2007), Liang and Li (2009), Ma and Li (2010), among others, we extend the method by incorporating measurement error modeling into the selection process.

The structure of the article proceeds as follows. In Section 2, we present a two-component G method for variable selection. The formulation of a two-component G-prior, the efficiency of the prior in terms of variable selection, the prior distributions of other parameters, and the posterior distributions are discussion in this section. An adaptive method to determine the value of a tuning parameter is presented in Section 3. The extension of the method with measurement error accounted for is discussed in Section 4. In Section 5, simulations are used to demonstrate and evaluate the proposed method, and comparisons among different variable selection methods are discussed. We apply our method to the motivating gene expression data and a data set related to ozone levels in Section 6. Finally, we summarize and discuss our findings in Section 7.

2 The Two-Component G Method

Denote by $Y_i, i = 1, \ldots, n$, the response from individual i, $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{ki}, \ldots, X_{pi})^T$ a vector of predictors. We consider variable selection via the following classical linear regression model,

$$Y_i \quad = \quad \boldsymbol{X}_i^T \boldsymbol{\beta} + \epsilon_i, \; i = 1, \ldots, n,$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\epsilon_i \sim N(0, \sigma^2)$. Since intercept is not involved in the process of variable selection, without loss of generality, we assume that the data are centered and henceforth the intercept is excluded from the model. For the purpose of variable selection, we introduce parameter $\boldsymbol{\gamma}$, a vector of length p composed of 1s and 0s indicating the inclusion or exclusion of a variable, respectively. Denote by $\Lambda_{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^T \mathbf{1}$ the size of a model determined by $\boldsymbol{\gamma}$. To infer $\boldsymbol{\gamma}$ and other parameters of interest, we propose a fully Bayesian approach.

2.1 Prior Distributions

In this section, we focus on the prior distribution of $\boldsymbol{\beta}$. The prior and hyper-prior distributions of other parameters are listed at the end of this section with details included in Appendix A.

Zellner's g-prior is defined as $\boldsymbol{\beta}|g, \sigma^2 \sim N(0, g(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2)$, where g is an unknown scalar parameter to be specified (Zellner, 1986). In the literature, various studies have discussed the selection of g in Zellner's prior, for instance, AIC, BIC, or risk inflation criterion (RIC) proposed in Foster and George (1994) are commonly used to calibrate g. George and Foster (2000) applied empirical Bayes to select g based on a marginal likelihood of g. Fernández et al. (2001) compared a variety of choices of g and recommended choosing g as $\max(\sqrt{n}, p)$. Liang et al. (2008) inferred g by introducing hyper prior distributions to g. In all these methods, the (variable selection) functionality of g is fulfilled through its overall influence on the variance components in the prior of nonzero coefficients $\boldsymbol{\beta}$.

In this article, we extend Zellner's g-prior and construct the prior distribution of $\boldsymbol{\beta}$ in two steps. We first extend Zellner's g-prior to a two-component g-prior. We assume that g in the prior is distributed following a mixture of two point masses, i.e., $p(g) = qI(\{g = g_l\}) + (1 - q)I(\{g = g_s\})$, where $I(\cdot)$ is an indicator variable and $I(\cdot) = 1$ if $(\cdot)$ is true, q is the probability of $g = g_l$, and $g_l, g_s > 0$ such that $g_l = bf_1(n), g_s = bf_2(n)$ with $f_1(n) = O(n), f_2(n) = O(n^{\psi})$, $1/2 < \psi < 1$, implying $f_2(n) = o(f_1(n))$ as $n \to \infty$. We thus have, as $n \to \infty$, $g_s \to \infty, g_l \to \infty$, while $g_s/g_l \to 0$. Both $f_1(n)$ and $f_2(n)$ are assumed to be known. For instance, we can take $f_1(n) = n$, and $f_2(n) = n^{0.55}$. The definitions of $f_1(n)$ and $f_2(n)$ follow the generic suggestion for the choice of g as stated in Fernández et al. (2001). The parameter $b > 1$ is a tuning parameter. For a given $f_1(n)$ and $f_2(n)$, b determines the distance between g_l and g_s. Next, instead of using one g to shrink all, we extend the application of $p(g)$ to each coordinate and define G as $\boldsymbol{G} = \mathrm{diag}(g_1, \ldots, g_p)$, a diagonal matrix such that each g_k is chosen according to $p(g_k) = q_k I(\{g_k = g_l\}) + (1 - q_k)I(\{g_k = g_s\}), k = 1, \ldots, p$. The prior distribution of $\boldsymbol{\beta}$

now is a two-component $\boldsymbol{G}$-prior,

$$\begin{aligned}\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{X}_{\boldsymbol{\gamma}},\boldsymbol{\gamma},\sigma^2,\boldsymbol{G} &\sim N\left\{\mathbf{0},\sigma^2\left[\left(\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}\right)^T\left(\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}\right)\right]^{-1}\right\}, \\ &= N\left\{\mathbf{0},\sigma^2\boldsymbol{G}_{\boldsymbol{\gamma}}\left(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\right)^{-1}\boldsymbol{G}_{\boldsymbol{\gamma}}\right\} \qquad (1)\end{aligned}$$

where $\boldsymbol{X}_{\boldsymbol{\gamma}}$ is an $n \times \Lambda_{\boldsymbol{\gamma}}$ model matrix with columns chosen by $\boldsymbol{\gamma}$, and $\boldsymbol{G}_{\boldsymbol{\gamma}}$ is a $\Lambda_{\boldsymbol{\gamma}} \times \Lambda_{\boldsymbol{\gamma}}$ diagonal matrix. Note that $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ instead of $\boldsymbol{\beta}$ is used in (1) to indicate the dependence of $\boldsymbol{\beta}$ on $\boldsymbol{\gamma}$. Only coefficients of selected variables are in $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, and the coefficients of unselected variables are zeros with probability 1 (George and McCulloch, 1997). Let $\boldsymbol{c} = (c_1, \ldots, c_p)$ be a vector of 1s and 0s denoting the choice between g_l and g_s, respectively. Thus we have $g_k = g_l c_k + g_s(1 - c_k), k = 1, \ldots, p$, with $p(c_k = 1) = q_k$. If $g_k = g_s$ for all $k = 1, \ldots, p$, then the prior is essentially the Zellner's g-prior, $N(0, g_s^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2)$; so is the situation if $g_k = g_l$ for all k's. The ultimate functionality of b in the extended prior (1) is to optimize g_k to eliminate unimportant variables. The choice of b is discussed in detail in Section 3. We denote this approach as the "two-component $\boldsymbol{G}$" method. The following proposition indicates that using a two-component $\boldsymbol{G}$ can be more efficient in identifying truly important variables than when a scalar g is used.

Proposition 1. *Denote by M_t the underlying true model and M_u a model such that $M_u \neq M_t$. Define $R_{u,t}(g_s) = Pr(M_u|\boldsymbol{Y}, \boldsymbol{X}, g_s)/Pr(M_t|\boldsymbol{Y}, \boldsymbol{X}, g_s)$, the ratio of posterior model probabilities (M_u over M_t) under the Zellner's g-prior with $g_k = g_s$ for all k, $k = 1, \ldots, p$, and $R_{u,t}(\boldsymbol{G}_{\boldsymbol{\gamma}}) = Pr(M_u|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{G}_{\boldsymbol{\gamma}})/\ Pr(M_t|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{G}_{\boldsymbol{\gamma}})$, the ratio of posterior model probabilities using the proposed prior as opposed to the Zellner's g-prior. Assume $\boldsymbol{X}^T\boldsymbol{X}/n \to \Sigma_X$ as $n \to \infty$ with Σ_X being the covariance of $\boldsymbol{X}$ and Σ_X positive definite. As $n \to \infty$, the ratios $R_{u,t}(g_s)$ and $R_{u,t}(\boldsymbol{G}_{\boldsymbol{\gamma}})$ (conditional on $\boldsymbol{X}$) satisfy $R_{u,t}(g_s) \to 0$, $R_{u,t}(\boldsymbol{G}_{\boldsymbol{\gamma}}) \to 0$, and $R_{u,t}(\boldsymbol{G}_{\boldsymbol{\gamma}}) = o(R_{u,t}(g_s))$.*

Proof. See Appendix B. □

The ratios in Proposition 1, $R_{u,t}(g_s)$ and $R_{u,t}(\boldsymbol{G}_{\boldsymbol{\gamma}})$, are Bayes factors when taking uniform priors for both models M_u and M_t. In summary, Proposition 1 states that, compared to the ratio of posterior model probability formulated in Zellner's g-prior, the one built upon two-component $\boldsymbol{G}$ converges in probability to zero more quickly for any model M_u other than the true model M_t.

The priors of the remaining parameters are vague priors and are listed below:

$$\begin{aligned}\gamma_k &\sim Ber(\pi_k) \text{ with } \boldsymbol{\pi} = (\pi_1, \ldots, \pi_p),\ \pi_k{\sim}uniform\ (0,1), \\ c_k &\sim Ber(q_k) \text{ with } \boldsymbol{q} = (q_1, \ldots, q_p),\ q_k \sim uniform\ (0,1),\ k = 1, \ldots, p, \\ \sigma^2 &\sim Inv\text{-}Gam(\delta_{\sigma,1}, \delta_{\sigma,2}),\end{aligned}$$

where we assume that the hyper-parameters $(\delta_{\sigma,1}, \delta_{\sigma,2})$ in the inverse gamma distributions are small and known. Detailed discussions on the choices of these priors are given in Appendix A.

2.2 The Joint Posterior Distribution and Its Computation

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{c}, \sigma^2, \boldsymbol{\pi}, \boldsymbol{q})$ denote a collection of parameters. The joint posterior distribution of $\boldsymbol{\theta}$ is, up to a normalizing constant,

$$p(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X}, b) \propto p(\boldsymbol{Y}|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2) p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{c}, b) p(\boldsymbol{\gamma}|\boldsymbol{\pi}) p(\boldsymbol{c}|\boldsymbol{q}) p(\sigma^2) p(\boldsymbol{\pi}) p(\boldsymbol{q}). \tag{2}$$

Posterior inference of $\boldsymbol{\theta}$ is obtained by successively simulating values from their full conditional posterior distributions through the Gibbs sampling scheme. These distributions are presented next. The full conditional posterior distribution of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is

$$\begin{aligned} p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{Y}, \boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{G}_{\gamma}) &= N(\boldsymbol{\Sigma}_{\beta_{\boldsymbol{\gamma}}} \boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{Y}/\sigma^2, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}), \\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}} &= \sigma^2 (\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}} + \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1})^{-1}, \end{aligned} \tag{3}$$

where we assume $\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}} + \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}$ is positive definite. Note that (3) is for selected variables; for unselected variables, the coefficients are zero and thus not included in the model.

Each component in the vector $\boldsymbol{\gamma}$ will be sampled sequentially. The full conditional posterior distribution of γ_k is, up to a normalizing constant,

$$\begin{aligned} &Pr(\gamma_k|\boldsymbol{Y}, \boldsymbol{X}_{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}_{(-k)}, \sigma^2, \boldsymbol{c}, b, \pi_k) \\ &\quad \propto p(\boldsymbol{Y}|\boldsymbol{X}_{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \gamma_k, \boldsymbol{\gamma}_{(-k)}, \sigma^2) p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{X}_{\gamma}, \gamma_k, \boldsymbol{\gamma}_{(-k)}, \sigma^2, \boldsymbol{c}, b) Pr(\gamma_k|\pi_k) \\ &\quad \propto p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{X}_{\gamma}, \gamma_k, \boldsymbol{\gamma}_{(-k)}, \sigma^2, \boldsymbol{c}, b) Pr(\gamma_k|\pi_k), \end{aligned} \tag{4}$$

where $\boldsymbol{\gamma}_{(-k)}$ is with dimension $p-1$ resulting from the exclusion of γ_k from $\boldsymbol{\gamma}$. The independence of (4) on $\boldsymbol{Y}$ is due to the hierarchical structure where $\boldsymbol{\gamma}$ influences $\boldsymbol{Y}$ only through $\boldsymbol{\beta}$ (Morris, 1987; George and McCulloch, 1993). This implies $Pr(\gamma_k|\boldsymbol{Y}, \boldsymbol{X}_{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}_{(-k)}, \sigma^2, \boldsymbol{c}, b, \pi_k) \propto Pr(\gamma_k|\boldsymbol{X}_{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}_{(-k)}, \sigma^2, \boldsymbol{c}, b, \pi_k)$. It is worth noting that the full conditional posterior distribution of γ_k depends on g_k. This implies that controlling the variation of nonzero coefficients influences the inclusion of variables. Since γ_k only takes value 0 or 1, the full conditional posterior distribution of γ_k is Bernoulli, so is the conditional posterior distribution of c_k. Their probability parameters in the Bernoulli distribution are given in Appendix A.

The full conditional posterior distribution of σ^2 is

$$\begin{aligned} p(\sigma^2|\boldsymbol{Y}, \boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) &= \textit{Inv-Gam}\left(\frac{n}{2} + \delta_{\sigma,1}, \frac{S_G}{2} + \delta_{\sigma,2}\right), \\ S_G &= (\boldsymbol{Y} - \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}})^T (\boldsymbol{Y} - \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}), \end{aligned} \tag{5}$$

which is an updated inverse gamma distribution from the prior of σ^2.

Since $\pi_k \sim \textit{uniform}(0,1)$ is conjugate to the Bernoulli prior selected for γ_k given π_k, the full conditional posterior distribution of π_k only depends on γ_k and is given by $p(\pi_k|\gamma_k) = Beta(\gamma_k + 1, 2 - \gamma_k)$. The full conditional posterior distribution of q_k is in the same format, $p(q_k|c_k) = Beta(c_k + 1, 2 - c_k)$. The independence of the full conditionals of $\boldsymbol{\gamma}$, $\boldsymbol{\pi}$, and $\boldsymbol{q}$ on $\boldsymbol{Y}$ allows for faster convergence of the sampling process.

A Markov chain Monte Carlo (MCMC) sequence will be obtained by repeatedly sampling from the full conditional posterior distributions. All the aforementioned conditional posterior distributions are standard, which facilitates easy sampling and quick convergence compared to non-standard distributions. Following the suggestions in Gelman et al. (2003), multiple chains will be simulated for the purpose of convergence evaluation and parameter inferences.

3 Calibration of b

We calibrate b by controlling Bayesian false model selection rate with respect to unimportant variables. In what follows, we define Bayesian false-model selection rate and outline the tuning steps.

3.1 Definition of Bayesian False-Model Selection Rate

A Bayesian false-model selection rate (BFSR) is defined as the relative frequency of selecting models that contain unimportant variables during the MCMC process. BFSR is defined utilizing the irreducibility property of MCMC simulations on stationary distributions such that the stochastic process walks through the state space of all possible models; models containing unimportant variables are likely to be visited less frequently than models with important ones.

The calculation of BFSR is rooted in Wu et al. (2007). Let T denote the number of MCMC iterations, out of which T_U iterations visited models that include real unimportant variables. BFSR is defined as the ratio between T_U and T, $\text{BFSR} = T_U/T$. Following Wu et al. (2007), we estimate BFSR by use of pseudo-variables satisfying the following two properties: (A1) the inclusion of pseudo-variables does not affect the inclusion of real important variables, and (A2) the probability of selecting pseudo-variables is the same as the probability of including real unimportant ones, that is, $E\{I_P(b)\} = E\{I_U(b)\}$, where b is the unknown parameter determining the distance between g_s and g_l, $I_P(b)$ is an indicator denoting the inclusion of any pseudo-variables given the value of b, and $I_U(b)$ denotes the inclusion of any real unimportant variables. Property (A2) states that if a model includes a real unimportant variable, then it is equally likely that the model also has a pseudo-variable. Since $E\{I_P(b)\} = E\{I_U(b)\}$ (A2), with T fixed, it is easy to verify $E(\text{BFSR}) = E(T_P/T)$, where T_P is the number of models with pseudo-variables included among the T iterations. An estimator of BFSR can then be defined as $\widehat{\text{BFSR}} = T_P/T$.

Wu et al. (2007) proposed an approach using $\boldsymbol{X}$ to generate pseudo-variables with the above two properties asymptotically held. Their approach is adopted here. It starts from the permutation of the rows of $\boldsymbol{X}$ design matrix. Denote by $\boldsymbol{P_X}$ the new matrix. The pseudo-variables are then formed by the residuals through regressing $\boldsymbol{P_X}$ on $\boldsymbol{X}$. Specifically, the values of the pseudo-variables are given by $\boldsymbol{(I - H_X)P_X}$ with $\boldsymbol{H_X} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Permuting the rows of $\boldsymbol{X}$ maintains correlations between different variables in the generated pseudo-variables and ensures (A2) held asymptotically. Projecting $\boldsymbol{P_X}$ through $\boldsymbol{I - H_X}$ reduces the effect of pseudo-variables on the selection of real important variables and thus makes (A1) approximately satisfied. After introducing the pseudo-variables, data $(\boldsymbol{Y}; \boldsymbol{X})$ is augmented to $(\boldsymbol{Y}; \boldsymbol{X}, \boldsymbol{R_P})$ with $\boldsymbol{R_P} = \boldsymbol{(I - H_X)P_X}$.

3.2 The Adaptive Calibration of b and the MCMC Simulation Process

The proposed adaptive selection process is motivated by the work of Browne and Draper (2006), in which an adaptive approach is introduced at the beginning of MCMC posterior sampling to select variances for jumping proposals. In our work, we use this idea to select b. Due to the indirect influence of b noted earlier on the inclusion of variables, b is calibrated to achieve a tolerable BFSR. Assume that the target BFSR is ω and the initial value of b is b_0. Usually b_0 takes a relatively small value to potentially include more variables. We list the tuning procedure as follows:

Step 1. Assume the current value is $b_m, m = 0, \ldots, M$.

- At iteration $t = 1, \ldots, T$, posterior samples of parameters are drawn using the methods described in Section 2.2. Define $I_t = 1$, if at least one pseudo-variable is selected, and $I_t = 0$ otherwise.
- Estimate BFSR as $\widehat{\text{BFSR}} = \sum_{t=1}^{T} I_t/T$.

Step 2. If $\widehat{\text{BFSR}}$ is greater than zero but lower than a pre-specified tolerance bound, $\omega+\omega_0$, then set $b = b_m$. The value of ω_0 is a pre-specified positive small number.

If $\widehat{\text{BFSR}} > \omega+\omega_0$, then increase the value to $b_{m+1} = b_m\{2-(1-\widehat{\text{BFSR}})/(1-\omega)\}$ to increase the distance between g_s and g_l in the two-component $\boldsymbol{G}$ method. This is to potentially decrease $\widehat{\text{BFSR}}$. Increase m by 1 and go to Step 1.

The tolerance bound is constructed such that being slightly away from the targeted BFSR does not severely disturb the variable selection process. The targeted BFSR can be set at 0.5, indicating random inclusion of unimportant variables. For the same tolerance bound, the tuning process is expected to be longer if a smaller ω is chosen. Alternatively, ω and ω_0 can be combined to represent one single threshold if no interest is in the target BFSR. We require BFSR be strictly positive to avoid the occurrence of too large g_k values. Once b is tuned, it is fixed. We then monitor the variation of $\widehat{\text{BFSR}}$ at the selected b value for a certain number of iterations to check the stability of BFSR. After this step, the pseudo-variables are removed from the pool of candidate variables and the MCMC simulation then starts the usual burn-in process; after the burn-in the MCMC simulation continues to draw samples for the purpose of posterior inferences. The motivation of calibrating b indicates that the above procedure can also be applied to tune $g_k = g$ in the Zellner's g-prior. Zellner's g-prior with g tuned is used as a competing method in the section of simulations.

4 Dealing with Measurement Errors

As noted earlier, the motivating example is to identify genes whose expression levels are associated with lung functions. It is known that measurement errors are common in gene expression levels (Vannucci et al., 2012). Ignoring measurement errors can cause biased inferences and consequently impact the selection of variables (Higgins et al., 1997; Carroll et al., 2006; Liu and Wu, 2007; Liang and Li, 2009; Ma and Li, 2010). Under this consideration, we further generalized the method by incorporating a measurement error model into the variable selection process. The measurement error model is defined as

$$\boldsymbol{W}_{ij} = \boldsymbol{X}_i + \boldsymbol{U}_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, J_i,$$

where $\boldsymbol{W}_i = (\boldsymbol{W}_{i1}, \ldots, \boldsymbol{W}_{iJ_i})^T$ is a matrix containing corresponding observed values of $\boldsymbol{X}_i$ measured at time $j = 1, \ldots, J_i$ for subject i, and $\boldsymbol{U}_{ij}$ is the measurement error. We assume $\boldsymbol{U}_{ij} \perp \boldsymbol{X}_i$ and $\boldsymbol{U}_{ij} \sim N(0, \boldsymbol{\Sigma_u})$ with $\boldsymbol{\Sigma_u}$ being an diagonal matrix. The reliability ratio is defined as $\Sigma_{xx}/(\Sigma_{xx} + \Sigma_{uu})$ (Carroll et al., 2006), where Σ_{xx} and Σ_{uu} denote the diagonal elements of corresponding covariance matrices of $\boldsymbol{X}$ and $\boldsymbol{U}$. The lower the reliability ratio, the severer the measurement errors will be. Assuming the measurement error is non-differential, the updated likelihood function becomes

$$p(\boldsymbol{\theta}, \boldsymbol{X_\gamma} | \boldsymbol{Y}, \boldsymbol{W}, b) \ \propto\ p(\boldsymbol{Y} | \boldsymbol{\theta}, \boldsymbol{X_\gamma}) p(\boldsymbol{W} | \boldsymbol{\theta}, \boldsymbol{X_\gamma}) p(\boldsymbol{X_\gamma} | \boldsymbol{\theta}),$$

where $X_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_X})$ with $\boldsymbol{\Sigma_X} = \mathrm{diag}(\sigma^2_{x_1}, \ldots, \sigma^2_{x_p})$ and $\boldsymbol{\theta}$ is expanded with parameters $\boldsymbol{\Sigma_u}$ and $\boldsymbol{\Sigma_X}$ included. The priors for each component of $\boldsymbol{\Sigma_u}$ and that of $\boldsymbol{\Sigma_X}$ are inverse gamma with known parameters. Then following the similar procedure, we can derive the full conditionals of all parameters. Details are given in Appendix C. Note that since $\boldsymbol{U}_{ij} \perp \boldsymbol{X}_i$, Proposition 1 given earlier and the method to generate pseudo-variables still apply. The only difference is that, in this case, the pseudo-variables are formed by projecting $\boldsymbol{P_W}$ through $\boldsymbol{I} - \boldsymbol{H_W}$, where $\boldsymbol{P_W}$ is the matrix obtained by permuting the rows of $\boldsymbol{W}$, and $\boldsymbol{H_W} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T$.

5 Simulations

Through simulations, in this section we demonstrate the proposed variable selection approach and compare its performance with four competing methods corresponding to four different settings of g_k. In all these four settings, g_k takes the same value for all k's, $g_k = g_0$. In the first setting, $g_0 = 10$ as in Smith and Kohn (1996). The second setting follows the suggestion in Fernández et al. (2001) and takes $g_0 = \max(\sqrt{n}, p)$. In the third setting, instead of fixing g_0 at a specific value, we calibrate g_0 using the procedure listed in Section 3.2. The fourth competing method is recommended in Liang et al. (2008), where a hyper-g-prior with parameter a in the prior taken as $a = 3$. In addition to these four competing methods, we also compared our method with the generalized g-prior recently proposed by Maruyama and George (2011), where a Bayes factor is used for variable selection. In total, we compare our proposed method with five competing ones.

5.1 Simulation Design

Example 1. Data in this example are simulated roughly following the study design in our motivating example. A data set of size n is generated. Five uncorrelated variables $\boldsymbol{X}$ are generated from $N(\boldsymbol{0}, \boldsymbol{I})$, among which the first one is important with coefficients $\beta_1 = 1$.

This variable also interacts with a two-level categorical variable and the coefficient for the interaction part is 2, which gives $\boldsymbol{\beta} = (1, \mathbf{0}_{1\times 5}, 2, \mathbf{0}_{1\times 4})$. For the two-level categorical variable, half of the n observations are assigned to level 1. The variance of the random error is $\sigma^2 = 1$.

Example 2. This is to demonstrate the performance of the method when the number of covariates is relatively large and those covariates are correlated. A data set of size n with 20 candidate variables is simulated, among the 20 variables, the first 4 are important with regression coefficients $(\beta_1, \ldots, \beta_4) = (1, 1.5, 2, 1.8)$. The correlation between every two candidate variables is $0.6^{|i-j|}, i, j = 1, \ldots, 20$. Other settings are the same as in Example 1.

We consider various sample sizes $n = 30, 100, 200, 400$, and 600 for each example, and for each sample size, 1000 data sets were simulated. Sample size of 30 is similar to that in our motivating example. In Example 1, after including pseudo variables, the maximum number of covariates is 22. In Example 2, when $n = 30$ the number of true candidate variables plus that of the pseudo-variables exceed the sample size. In this case, the conventional all-subsets regression strategies for variable selection are not feasible (Miller, 2002; George and McCulloch, 1993). In all the simulated data, the pseudo-variables are generated based on the criteria described in Section 3.

The following statistics are recorded and used to evaluate the performance of the methods:

1. The proportions of correct/over-fitted/under-fitted models in 1000 simulated data sets. Correct models are defined as models that only include all truly important variables; over-fitted models are the models that include all important variables plus at least one unimportant variable; and under-fitted models are those that include a subset of the important variables. The sum of the three proportions is 1.

2. Mean squared loss (MSL) assesses the estimation power of a selected model and is defined as MSL $= \frac{1}{n}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} - \boldsymbol{\beta}_{\boldsymbol{\gamma}})^T \boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}} (\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} - \beta_{\boldsymbol{\gamma}})$, where $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the posterior mean of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. In the proposed method, for large n, the distribution of MSL is approximately proportional to $\chi^2_{\Lambda_{\boldsymbol{\gamma}}}$ with $\Lambda_{\boldsymbol{\gamma}}$ degrees of freedom.

5.2 Results and Comparisons

We applied the proposed methods to the above simulated data sets. For each data set, two Markov chains are simulated for the purpose of convergence assessment (Gelman et al., 2003). Each chain starts from a certain number of iterations with an initial value of b_0. This is then followed by the process of calibrating b as described in Section 3.2 with the target BFSR set at $\omega = 0.5$ and the tolerance upper bound at $\omega + \omega_0 = 0.6$. After b is selected, we continue to run 20,000 MCMC iterations to monitor the change of $\widehat{\text{BFSR}}$. After the tuning process, the MCMC simulation process runs for 20,000 iterations with 10,000 iterations used as burn-in.

Based on one simulated data set of each example, Figure 1 is an illustration of the tuning (TP) and monitoring processes (MP). The figure displays the trend of $\widehat{\text{BFSR}}$

with the change of b, where each $\widehat{\text{BFSR}}$ is calculated based on $T = 1000$ iterations for a given b. As indicated in the figure, b can be quickly tuned when the number of variables is small (Example 1). When the number of variables is large, it may take a large number of MCMC iterations to find an appropriate b (Example 2). In addition, the results from the monitoring process indicate that once b is selected, $\widehat{\text{BFSR}}$ is relatively stable, implying that we are ready for the regular MCMC sampling. The posterior inferences discussed below are based on 10,000 posterior samples from one chain after 10,000 burn-in of two chains.

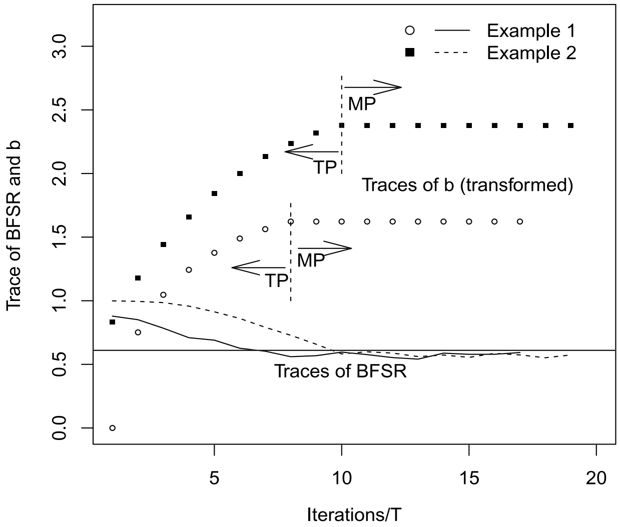


Figure 1: Illustration of the tuning process of b in the two examples ($n = 400$). The transformed b is calculated as $\sqrt{\log(b)}$, "TP" denotes "tuning process" and "MP" denotes "monitoring process", and the solid horizontal line indicates the tolerance upper bound of BFSR.

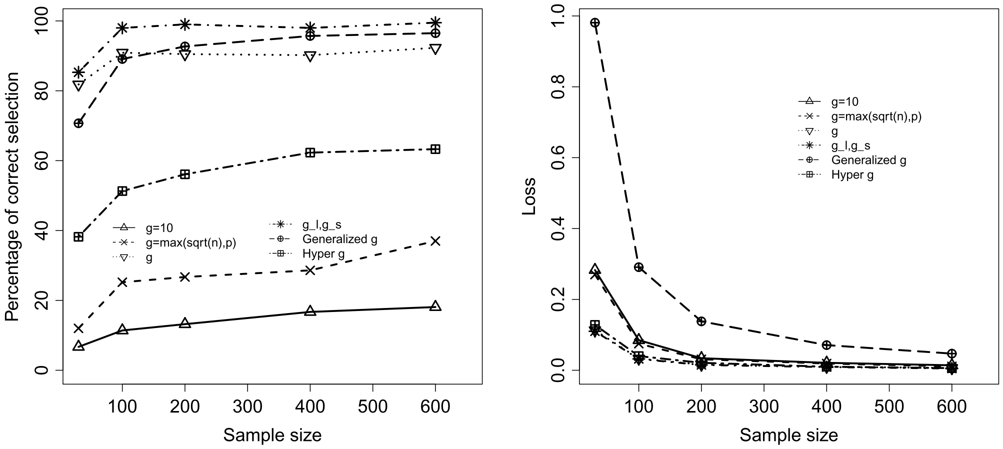
As indicated in Table 1, the findings based on finite samples are consistent with the asymptotical property given in Section 2.1, in that the proposed method selects the correct models more efficiently compared to other competing methods built upon a scalar g_0. In these simulations, we take $f_1(n) = n$, and $f_2(n) = n^{\psi}$ with $\psi = 0.55$. As noted in Proposition 1, the asymptotical property holds as long as $f_1(n) = O(n)$, $f_2(n) = O(n^{\psi})$, $1/2 < \psi < 1$. This was also supported by our additional simulations (results not shown). Overall, the method based on the two-component G-prior gives higher percentages of correct selections and lower losses. These patterns are observed as well when the number of predictors increased from 11 to 20 (see Figure 2).

5.3 Accounting for Measurement Errors

Due to the motivating example of the proposed method, we further extended the approach to incorporate measurement errors as discussed in Section 4. To examine its

$n = 30$	$g_0 = 10$	$g_0 = \max(\sqrt{n}, p)$	g_0	G_{g_l,g_s}	Hyper-g	Generalized g
% correct	32.4	36.1	68.8	61.9	54.6	65.2
% overfit	58.0	54.6	9.0	3.4	38.0	32.0
Med. b	–	–	224.12	164.15	–	–
Med. MSL	0.137	0.133	0.073	0.092	0.125	0.30
$n = 100$	$g_0 = 10$	$g_0 = \max(\sqrt{n}, p)$	g_0	G_{g_l,g_s}	Hyper-g	Generalized g
% correct	40.4	44.9	88.7	98.1	63.9	85.3
% overfit	59.6	55.1	11.2	1.8	36.1	11.2
Med. b	–	–	157.97	114.11	–	–
Med. MSL	0.031	0.028	0.016	0.014	0.027	0.085
$n = 600$	$g_0 = 10$	$g_0 = \sqrt{n}$	g_0	G_{g_l,g_s}	Hyper-g	Generalized g
% correct	43.9	64.9	89.9	98.1	78.8	95.2
% overfit	56.1	35.1	10.1	1.9	21.2	3.1
Med. b	–	–	145.02	86.11	–	–
Med. MSL	0.010	0.006	0.003	0.004	0.003	0.012

Table 1: Variable selection statistics for Example 1: Med. b, median of the tuning parameter; Med. MSL, Median mean squared loss.



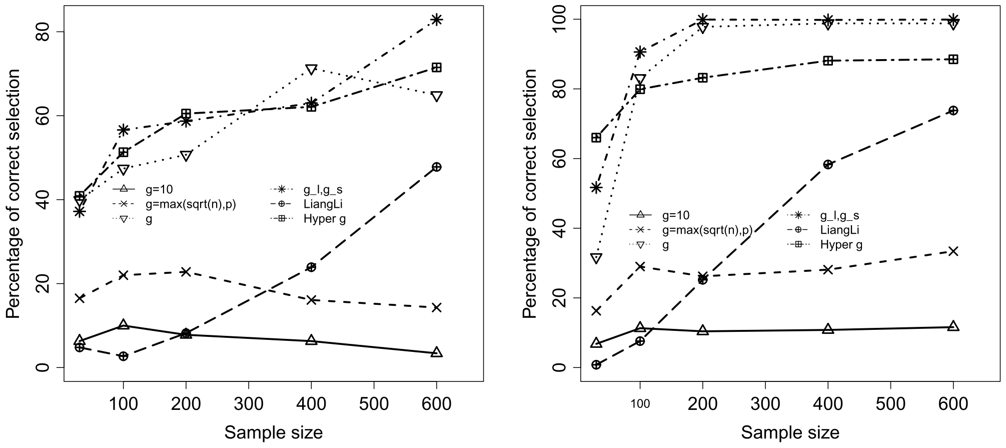
(a)

(b)

Figure 2: Patterns of percentages of correct selection on important variables (a) and mean squared losses (b) across different sample sizes. In both figures, g_l, g_s represents the two-component G-prior.

performance, we modify the simulation scenario of Example 2 by adding noise onto $\boldsymbol{X}$. Specifically, we consider $\boldsymbol{\Sigma_u} = 0.25\boldsymbol{I}$ and $0.11\boldsymbol{I}$, corresponding to reliability ratios of $a = 0.8$ and 0.9, respectively. We compare the two-component $\boldsymbol{G}$ method with four of the competing methods where the measurement errors are incorporated. The generalized g-prior is not considered in this comparison as it selects a model via Bayes factors not designed for predictors with measurement errors. A recently developed frequentist approach for variable selection in partially linear measurement error models is considered instead and included as the fifth competing method (Liang and Li, 2009). It utilizes modified least squares combined with the SCAD penalty to perform variable selection. The modified least squares is built upon mis-measured predictors and yields consistent estimators. In that method, a tuning parameter critical to the inclusion of a variable is selected using BIC pretending no measurement errors. It has been shown that with an appropriately chosen tuning parameter, their proposed method will yield a variable selection procedure that enjoys the celebrated oracle property.

We summarize our findings in Figure 3 for percentages of correct selections and Figure 4 for mean squared loss. Results related to percentages of over-fitting along with other statistics are given in Appendix D. In terms of variable selection, both the proposed method and the method of Liang and Li (2009) have finite sample performance consistent with the asymptotical properties in variable selection. However, the proposed method more often produces the highest percentages of correct selection among all the methods.



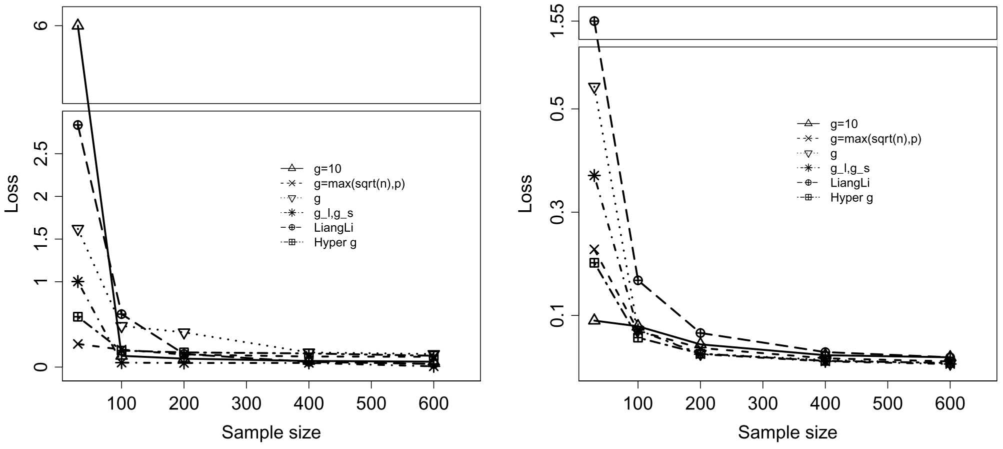
(a)

(b)

Figure 3: Patterns of percentages of correct selection on important variables across different sample sizes at different reliability ratios: (a) $a = 0.8$; (b) $a = 0.9$. In both figures, g_l, g_s represents the two-component G-prior.

6 Applications

In this section, we apply the method to two real data sets: gene expression and lung function data, and ozone data. The first data set has 12 variables with sample size of 27, representing a situation of small sample size with relatively large number of candidate predictors. Also, the gene expression levels were measured twice for the purpose of measurement error correction. The second data set has a larger number of variables



(a) (b)

Figure 4: Patterns of mean squared losses across different sample sizes at different reliability ratios: (a) $a = 0.8$; (b) $a = 0.9$. In both figures, g_l, g_s represents the two-component G-prior.

(44 variables) with sample size of 330. In both applications below, we take the same setting of $f_1(n)$ and $f_2(n)$ as in simulations.

Application 1. The gene expression and lung function data, which motivated our study, were provided by tissues from the Lung Tissue Research Consortium (LTRC). It was from an association study examining whether IPF severity (measured by forced vital capacity, FVC) is associated with age, gender, expression of five genes, and their interaction with gender. The five genes are CCR3, Collagen_IIIα1, GMCSF, IL_6, and Tenascin_C. All of these gene targets are recognized to play a role in the pathogenesis of pulmonary fibrosis and are particularly related to immune and extra cellular matrix (ECM) remodeling processes (Sabo-Attwood et al., 2011). Collagen_IIIα1 is a major component of the ECM, and Tenascin_C is a glycoprotein involved in matrix turnover. Genes IL_6, GMCSF and CCR3 are regulators that govern the functions of the immune system by modulating pulmonary recruitment of immune cells (Knight et al., 2003; Christensen et al., 2000; Huaux et al., 2005). In total $n = 27$ subjects are involved in the study. The expression of each gene is the mean of two repeated measures for each subject, denoted below as “true” measures.

The model to be evaluated includes all main effects of the five genes, gender, and age along with the gene and gender interaction effects, which results in 12 variables. Pseudo-variables are generated using the method given in Section 3, and data are centered (treating all variables as continuous variables). We fit the data using the two-component $\boldsymbol{G}$ method along with the five competing approaches discussed in Section 5. Table 2 lists the top two models selected by each approach along with the posterior probabilities of each selection (Bayes factor is provided for the method by Maruyama and George (2011)).

Models	Selected variables	Selection statistics
		Post. prob. of selection
	$g = 10$	
1	Collagen_IIIα1, GMCSF × gender	0.13
2	Tenascin_C, Collagen_IIIα1 × gender, GMCSF × gender, IL_6 × gender	0.07
	$g = 12$	
1	Collagen_IIIα1, GMCSF	0.10
2	GMCSF, Tenascin_C	0.07
	g tuned	
1	Collagen_IIIα1, GMCSF × gender	0.12
2	Collagen_IIIα1	0.11
	G	
1	Collagen_IIIα1	0.26
2	Collagen_IIIα1, GMCSF × gender	0.21
	hyper-g-prior	
1	Collagen_IIIα1, GMCSF × gender	0.0049
2	Collagen_IIIα1, GMCSF, GMCSF × gender	0.0041
		Bayes Factor ($\log(BF)$)
	generalized g-prior	
1	Collagen_IIIα1, GMCSF × gender	4.52
2	Collagen_IIIα1, GMCSF	4.31

Table 2: Selected top two models for the two-component G and 5 competing methods. The models are sorted by posterior probabilities of selection or Bayes factor.

The two-component $\boldsymbol{G}$ method and the method by tuning a single $g_k = g$ select the same top two models, and both models are with similar posterior probabilities of selection. As seen in the simulations, these two methods select the correct models more often with the two-component $\boldsymbol{G}$ method enjoying the highest correctness rates. All methods except for taking $g = \max(\sqrt{n}, p)$ select a model which includes Collagen_IIIα1 and the interaction of GMCSF with gender. With $g = 10$, it has the tendency to select more variables, while taking $g = \max(\sqrt{n}, p)$ selects two models both with main effects only. Based on our simulations, the correctness rates of these two approaches are low and they also tend to over fit the data. These observations combined with the findings from simulations indicate that it is likely that, among all the genes, Collagen_IIIα1 is the most important, and it is likely that GMCSF interacts with gender to influence the lung function. The findings in previous studies support the selection of Collagen-IIIα1; Collagen-IIIα1 has shown to be highly upregulated in rodent models of pulmonary fibrosis (Sabo-Attwood et al., 2011) and is differentially expressed in patients with idiopathic pulmonary fibrosis compared with control subjects (Brass et al., 2007). However, the interaction of GMCSF with gender was not identified before, which deserves further investigation to exclude the effect of other risk factors such as occupation and environmental factors. It is worth noting that the variables listed in Table 2 are selected by treating each interaction variable as a variable independent from others. In practice, the hierarchical rule in variable selection is suggested to be followed, e.g., in the model selected by most methods that involves the interaction GMCSF×gender, both GMCSF and gender should be included in the model as well.

Given the availability of two repeated measures of gene expression levels for each gene, we performed the selection procedure by use of two repeated measures of gene expression level for each gene and each subject, using the method discussed in Section 4, which has the ability to account for measurement errors. We compare the results from the proposed method with the results from all other methods including Liang and Li (2009) except for the generalized g-prior as noted in Section 5.3. Utilizing the repeated measures of gene expression levels, we estimated the reliability ratios of each gene which range from 0.88 to 0.95 using the method stated in Section 4.4.2 (p. 70–72) of Carroll et al. (2006). With measurement error incorporated, for the two-component $\boldsymbol{G}$ method, the top two models are the same as those when "true" measures are used. The posterior probabilities are 0.23 (Model 1) and 0.15 (Model 2), respectively. The method by tuning a single g also identified the same top models (with posterior probabilities 0.10 for Model 1 and 0.16 for Model 2, respectively), plus a model with a comparable posterior probability of 0.11 that includes main effects of two genes, Collagen_IIIα1 and GMCSF. The most popular model including Collagen_IIIα1 and the interaction of GMCSF with gender is kept by the method utilizing hyper-g-prior as the top 1 model (posterior probability 0.028). In addition, it identified a second best model which includes a main effect of GMCSF, and interaction effects of GMCSF with gender and Collagen_IIIα1 with gender (posterior probability 0.02). The selection results from other methods based on a scalar g are far from being consistent with those based on "true" measures, which is likely to reflect their low correctness rates as seen in the simulations. Regarding the frequentist approach in Liang and Li (2009) designed for variable selection in measurement error models, it selected a model that includes the main effect of Collagen_IIIα1 and interaction effect of GMCSF with gender, which is Model 2 by use of the two-component $\boldsymbol{G}$ method. Overall, in the situation of measurement errors, the methods based on the two-component G prior, the g-prior with g tuned, the hyper-g-prior, and the frequentist approach identified the same best model. Only the two-component G method showed consistent findings with the situation of measurement error free.

Application 2. This application is to select variables associated with ozone levels. The data was analyzed and discussed in different studies (Breiman and Friedman, 1985; Casella and Moreno, 2006; Miller, 2002; Liang et al., 2008). It consists of daily measurements of the maximum ozone concentration near Los Angeles and 8 meteorological variables (a description of the variables is in Appendix E). As done in other studies (Miller, 2002; Casella and Moreno, 2006; Liang et al., 2008), the full linear regression model contains the 8 meteorological variables, two-way interactions, and squares of each variable. This gives in total 44 possible predictors. We applied the proposed method to

Selected variables	Selection statistics	R^2
Full model		0.789
	Post. prob. of selection	
$g = 10$		
temp, ibh, humid*temp, humid*dpg, humid*ibt, temp*ibh,temp*vis, ibh*ibt, ibh*vis, ibt*vis, vh^2, humid^2, dpg^2	0.033	0.775
$g = 44$		
humid, dpg, vh*ibt, humid*ibh, humid*ibt, temp*ibt, dpg^2,	0.066	0.767
g tuned		
vh*ibt, humid*ibh, humid*ibt, temp*ibt	0.333	0.756
G		
vh*ibt, humid*ibh, humid*ibt, temp*ibt	0.527	0.756
hyper-g-prior		
dpg, ibt, vh*ibh, humid*dpg, humid*ibt, temp*ibt, ibh*vis, humid^2, dpg^2	0.0008	0.771
	Bayes Factor ($\log(BF)$)	
generalized g-prior[#]		
ibh, humid*temp, humid*ibh, temp*ibt, ibh*ibt	208.60	0.750

[#] Due to the large number of possible models, exhaustive search based on Bayes factor is practically impossible. Candidate predictors are chosen such that these variables are selected by at least one of the other five methods.

Table 3: Selected top two models along with the posterior probabilities of selection. The models are sorted by posterior probabilities of selection.

the data, and compared the results with those from the 5 competing approaches discussed in Section 5. Besides the posterior probabilities of selection or Bayes factors for the best model selected, we also calculated R^2 for each selected model and the full model that includes all 44 variables. As shown in Table 3, the two-component G-prior selected a model that includes four terms and has $R^2 = 0.756$, close to $R^2 = 0.775$ from the full model. Compared to the models selected by other methods, except for the method using g-prior with g tuned which identified the same model, all the other methods selected complex models with slightly higher R^2 (except for the generalized g-prior, which gives lower R^2). As in Application 1, we suggest that in practice the final model determination follow the hierarchical rule, e.g., for the model selected using the two component G-prior, it should also include the main effects of each variable selected.

7 Discussion

We proposed a Bayesian variable selection approach in linear regressions based on an extension of the Zellner's g-prior. Instead of using a universal single g as in the classic Zellner's g-prior, we recommended the use of two different g_k's (the two-component $\boldsymbol{G}$-prior) to distinguish important variables from unimportant ones. The distance between the two components was adjusted by a tuning parameter b. The tuning parameter was selected adaptively by controlling a Bayesian false-model selection rate. The impact of measurement errors on parameter inferences and the availability of repeated measures in real data motivated us to further extend the method to incorporate measurement errors in variable selection.

We theoretically justified the efficiency of the proposed method and further demonstrated and evaluated the method via simulations. The simulation results showed that the proposed method can efficiently identify the correct models regardless of the existence of measurement errors. We compared the method with five methods developed based on traditional g-priors and one additional method constructed in the frequentist framework. In general, the competing methods tend to over-fit the data and the proposed method performs the best especially with respect to the correct selection rates.

The two real data applications further strengthened the findings from simulations and demonstrated that the proposed method has the ability to choose parsimonious models with R^2 close to that from the full model. In our first real data application, by use of the proposed method, we also performed variable selection with measurement error ignored. No difference was observed in terms of variables selected. However, as indicated in Liang and Li (2009), any appropriate variable selection procedures may falsely classify important variables if one ignores measurement errors. To further strengthen this statement, we simulated a data set using the scenario in Example 2 but set the first 4 important regression coefficients as $(\beta_1, \ldots, \beta_4) = (1, 0.5, 2, 1.8)$. For the purpose of illustration, we only considered reliability ratio $a = 0.8$ and sample size $n = 200$. Our results from the two-component $\boldsymbol{G}$ method indicated that ignoring measurement error does have a potential to cause false variable selection. Specifically, out of 1000 simulated data sets, the percentage of correct selection when ignoring measurement errors was 66% and the percentage increased to 72% if measurement error was addressed. In most cases, the second variable with coefficient 0.5 was classified as an unimportant variable and consequently caused the low correctness rate. For large effect sizes $(\beta_1, \ldots, \beta_4) = (1, 1.5, 2, 1.8)$ as in our Example 2, with $a = 0.8$, such increase in correctness rate after accounting for measurement error was not observed. However, it is expected that when measurement error is severe, the need of accounting for measurement error will be more obvious even in the situation of large effect sizes. In real data applications, because the effect size in general is unknown and measurement error can be large, it is safer to correct for measurement errors whenever possible. Additionally, in this article, we only considered two components in $\boldsymbol{G}$. However, moderately important variables may exist. In this case, a multi-component $\boldsymbol{G}$-prior may perform better. A potential challenge in this direction is how to deal with the complexity of the model and consequently prevent from potential power loss due to the increase of the number of components. This is our ongoing research work. Finally, it is worth noting that the methods can be easily extended to other types of statistical models including generalized linear models and models applied to survival data analysis, although the posterior distributions may be more involved.

Appendix A

The Prior Distributions The prior distributions of γ_k and c_k are both selected to be independent Bernoulli distributions, $\gamma_k \sim Ber(\pi_k)$ with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$ and $c_k \sim Ber(q_k)$ with $\boldsymbol{q} = (q_1, \ldots, q_p)$. Parameter π_k indicates the prior probability of including a variable in the model, and q_k is the prior probability of choosing g_l.

We infer π_k based on information in the data and thus assume a vague prior distribution, $\pi_k \sim$ *uniform* $(0, 1)$. The same choice of hyper prior distribution is applied to $\boldsymbol{q}$.

The Posterior Distributions The conditional posterior distribution of γ_k is Bernoulli with

$$\begin{aligned} Pr(\gamma_k = 1 | \boldsymbol{Y}, \boldsymbol{X}_\gamma, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}_{(-k)}, \sigma^2, \boldsymbol{c}, b, \pi_k) &= \frac{a_0}{a_0 + b_0}, \\ a_0 = p(\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \boldsymbol{X}_{\boldsymbol{\gamma}}, \gamma_k = 1, \boldsymbol{\gamma}_{(-k)}, \sigma^2, \boldsymbol{c}, b, \pi_k) Pr(\gamma_k = 1 | \pi_k), \\ b_0 = p(\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \boldsymbol{X}_{\boldsymbol{\gamma}}, \gamma_k = 0, \boldsymbol{\gamma}_{(-k)}, \sigma^2, \boldsymbol{c}, b, \pi_k) Pr(\gamma_k = 0 | \pi_k). \end{aligned}$$

Note that in the calculations of a_0 and b_0, for each $\boldsymbol{\gamma}$ given (with $\gamma_k = 1$ in the calculation of a_0 and $\gamma_k = 0$ in the calculation of b_0), the coefficients for variables not included in the model are zeros with probability 1, and do not contribute to the determination of a_0 and b_0.

The construction of the full conditional for $\boldsymbol{c}$ follows the same way as that for $\boldsymbol{\gamma}$,

$$\begin{aligned} Pr(c_k = 1 | \boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{c}_{(-k)}, b, q_k) &= \frac{a_1}{a_1 + b_1}, \\ a_1 = p(\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2, c_k = 1, \boldsymbol{c}_{(-k)}, b, q_k) Pr(c_k = 1 | q_k), \\ b_1 = p(\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2, c_k = 0, \boldsymbol{c}_{(-k)}, b, q_k) Pr(c_k = 0 | q_k). \end{aligned}$$

Appendix B

To prove the Proposition presented in Section 2.1, we need the following lemma:

Lemma 1. *Assume $\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}} / n \to \Sigma_{\boldsymbol{X}_{\boldsymbol{\gamma}}}$ with $\Sigma_{\boldsymbol{X}_{\boldsymbol{\gamma}}}$ being the covariance of $\boldsymbol{X}_{\boldsymbol{\gamma}}$ and $\Sigma_{\boldsymbol{X}_{\boldsymbol{\gamma}}}$ positive definite. As $n \to \infty$, given all other parameters, $c_k = 0$ with probability 1 if $\beta_k \neq 0$.*

Proof. In all the proofs in this Appendix, the limits are taken with respect to n, i.e., $n \to \infty$.

To show that c_k taking the value of 0 with probability 1 given the values of all other parameters is equivalent to showing $Pr(c_k = 0|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{c}_{(-k)}, b, q_k) \to 1$ as $n \to \infty$. Recall when $c_k = 1$, $g_k = g_l$ and $g_l \to \infty$ as $n \to \infty$. The natural logarithm of the conditional posterior of c_k is

$$\begin{aligned}\log(Pr(c_k|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{c}_{(-k)}), b, q_k) = C_0 &- \frac{1}{2}\log|\boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}\boldsymbol{G}_{\boldsymbol{\gamma}}| \\ &- \frac{\boldsymbol{\beta}_{\boldsymbol{\gamma}}^T\boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}})\boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\beta}_{\boldsymbol{\gamma}}}{2\sigma^2} \\ &+ c_k\log q_k + (1-c_k)\log(1-q_k), \qquad (6)\end{aligned}$$

where q_k is the prior probability of $c_k = 1$ and C_0 is a constant. Simplifying the second term,

$$\begin{aligned}-\frac{1}{2}\log|\boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}\boldsymbol{G}_{\boldsymbol{\gamma}}| &= -\frac{1}{2}\log\left\{|\boldsymbol{G}_{\boldsymbol{\gamma}}|^2|\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}|^{-1}\right\} \\ &= -\frac{1}{2}\log\left\{\left[\prod_{k'\neq k} g_{k'}^2\right] g_k^2|\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}|^{-1}\right\} \\ &= -\sum_{k'\neq k}\log g_{k'} - \log g_k + \frac{1}{2}\log|\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}|. \qquad (7)\end{aligned}$$

Simplifying the third term by highlighting information related to the kth variable, we have

$$\begin{aligned}&-\frac{\boldsymbol{\beta}_{\boldsymbol{\gamma}}^T\boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}})\boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\beta}_{\boldsymbol{\gamma}}}{2\sigma^2} \\ &= C_1 - \frac{\beta_k g_k^{-1}(\sum_{k'\neq k}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}[k,k']\beta_{k'}g_{k'}^{-1}) + \beta_k^2 g_k^{-2}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}[k,k]}{2\sigma^2}, \qquad (8)\end{aligned}$$

where C_1 is constant with respect to the kth variable.

Based on (7) and (8), assuming $\beta_k \neq 0$, we have

$$\begin{aligned}&\log(Pr(c_k = 1|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{c}_{(-k)}, b, q_k)) - \log(Pr(c_k = 0|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{c}_{(-k)}, b, q_k)) \\ &= \log(g_s/g_l) - \frac{(g_l^{-1} - g_s^{-1})\beta_k(\sum_{k'\neq k}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}[k,k']\beta_{k'}g_{k'}^{-1})}{2\sigma^2} \\ &\quad - \frac{(g_l^{-2} - g_s^{-2})\beta_k^2\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}[k,k]}{2\sigma^2}. \qquad (9)\end{aligned}$$

In (9), as $n \to \infty$, $\log(g_s/g_l) \to -\infty$ in the speed of $\log(n^{a_0})$ with $a_0 < 1$. The numerator of the second term is

$$\begin{aligned}&(g_l^{-1} - g_s^{-1})n\beta_k(\sum_{k'\neq k}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}[k,k']/n\beta_{k'}g_{k'}^{-1}) \\ &= \beta_k\{\sum_{k'\neq k}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}[k,k']/n\beta_{k'}(n/(g_{k'}g_l) - n/(g_{k'}g_s))\}.\end{aligned}$$

Similarly, the third term in (9) is $(n/g_l^2 - n/g_s^2)\beta_k^2 \boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}}[k,k]/n$. Since $g_l \to \infty$ in the order $O(n)$, $g_s \to \infty$ in the order of $O(n^{\psi})$ with $1/2 < \psi < 1$, and $\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}}/n \to \Sigma_{\boldsymbol{X}_{\boldsymbol{\gamma}}}$ with $\Sigma_{\boldsymbol{X}_{\boldsymbol{\gamma}}}$ positive definite, the last two terms in (9) are finite. Thus

$$\begin{aligned}&\log(Pr(c_k = 1|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{c}_{(-k)}, b, q_k))\\ &-\log(Pr(c_k = 0|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{c}_{(-k)}, b, q_k)) \to -\infty\end{aligned}$$

as $n \to \infty$. This implies that if variable k is important, then g_k takes the value of g_l with probability approaching to 0 as $n \to \infty$. Consequently, if $\beta_k \neq 0$, then conditional on other parameters, $c_k = 0$ (that is, $g_k = g_s$) with probability approaching to 1 as $n \to \infty$. □

Following the above conclusion, it is straightforward that

$$Pr(\gamma_k = 0|\boldsymbol{Y}, \boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}_{(-k)}, \sigma^2, \quad c_k = 1, \boldsymbol{c}_{(-k)}, b, \pi_k) \to 1$$

as $n \to \infty$, which implies that if a variable k takes $g_k = g_l$, then this variable is more likely to be treated as an unimportant variable and will be excluded from the model. The choice between g_l and g_s will be determined by the data.

Now we prove Proposition 1.

Proof. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2)$ denote a collection of parameters for a given model determined by $\boldsymbol{\gamma}$. The joint distribution of $\boldsymbol{\theta}_0$ and $\boldsymbol{Y}$ conditional on $\boldsymbol{X}_{\boldsymbol{\gamma}}$ and $\boldsymbol{G}_{\boldsymbol{\gamma}}$ is

$$p(\boldsymbol{\theta}_0, \boldsymbol{Y}|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \boldsymbol{G}_{\boldsymbol{\gamma}}) = p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{Y}, \boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{G}_{\boldsymbol{\gamma}})p(\sigma^2|\boldsymbol{Y}, \boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \boldsymbol{G}_{\boldsymbol{\gamma}})p(\boldsymbol{Y}|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \boldsymbol{G}_{\boldsymbol{\gamma}}). \tag{10}$$

Note that by taking $g_k = g_s$ in $\boldsymbol{G}_{\boldsymbol{\gamma}}$, (10) is reduced to the setting under the Zellner's g-prior.

The first term is the full conditional posterior distribution of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and is given in (3). The second part is marginal posterior distribution of σ^2 under a specific model. It is obtained by integrating out β from (10)

$$\begin{aligned}p(\sigma^2|\boldsymbol{Y}, \boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \boldsymbol{G}_{\boldsymbol{\gamma}}) &\propto (\sigma^2)^{-(1+n/2)} \exp\left\{-Y^T Q_{\boldsymbol{\gamma}} Y/(2\sigma^2)\right\}\\ &= \mathit{Inv\text{-}Gam}\left(\frac{n}{2}, \frac{Y^T Q_{\boldsymbol{\gamma}} Y}{2}\right),\\ Q_{\boldsymbol{\gamma}} &= \boldsymbol{I} - \boldsymbol{X}_{\boldsymbol{\gamma}}(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}} + \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1})^{-1} \boldsymbol{X}_{\boldsymbol{\gamma}}^T.\end{aligned}$$

The last term in (10) is related to the calculation of Bayes factor:

$$p(\boldsymbol{Y}|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \boldsymbol{G}_{\boldsymbol{\gamma}}) \propto (Y^T Q_{\boldsymbol{\gamma}} Y)^{-n/2} \frac{|\boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}})\boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}|^{1/2}}{|\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}} + \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}})\boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}|^{1/2}}. \tag{11}$$

From Lemma 1, if variable k is important, then for a given σ^2 under a model defined by $\boldsymbol{\gamma}$, c_k takes the value of 0 with probability 1 as $n \to \infty$, equivalently, $g_k = g_s$ with probability 1 as $n \to \infty$ if $\beta_k \neq 0$. Thus if $\boldsymbol{\gamma}$ specifies the true model t, then $g_k \to g_s$ in probability 1 for all k in model t as $n \to \infty$. Thus we have $Pr(M_t|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{G_\gamma}) - Pr(M_t|\boldsymbol{Y}, \boldsymbol{X}, g_s) \to 0$ as $n \to \infty$. We denote this as Result A.

Define $R_u(\boldsymbol{G_\gamma}, g_s)$ as

$$R_u(\boldsymbol{G_\gamma}, g_s) = \frac{Pr(M_u|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{G_\gamma})}{Pr(M_u|\boldsymbol{Y}, \boldsymbol{X}, g_s)} \propto \frac{Pr(\boldsymbol{Y}|\boldsymbol{X_\gamma}, \gamma, \boldsymbol{G_\gamma})}{Pr(\boldsymbol{Y}|\boldsymbol{X_\gamma}, \gamma, g_s)}.$$

From (11),

$$\begin{aligned} R_u(\boldsymbol{G_\gamma}, g_s) &= \left[\frac{Y^T Q_{\boldsymbol{\gamma}, g_s} Y}{Y^T Q_{\boldsymbol{\gamma}, \boldsymbol{G_\gamma}} Y}\right]^{n/2} \frac{|X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}}|^{1/2}(1+g_s^{-2})^{k_0/2}}{(g_s^{-1})^{k_0}|X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}}|^{1/2}} \frac{|\boldsymbol{G}_{\boldsymbol{\gamma}}^{-1} X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}|^{1/2}}{|X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}} + \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1} X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}|^{1/2}} \\ &\leq \left[\frac{Y^T Q_{\boldsymbol{\gamma}, g_s} Y}{Y^T Q_{\boldsymbol{\gamma}, \boldsymbol{G_\gamma}} Y}\right]^{n/2} \frac{|X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}}|^{1/2}(1+g_s^{-2})^{k_0/2}}{|X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}}|^{1/2} + |\boldsymbol{G}_{\boldsymbol{\gamma}}^{-1} X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}|^{1/2}} \frac{(g_l^{-1})^m (g_s^{-1})^{k_0-m}}{(g_s^{-1})^{k_0}} \\ &= \left[\frac{Y^T Q_{\boldsymbol{\gamma}, g_s} Y}{Y^T Q_{\boldsymbol{\gamma}, \boldsymbol{G_\gamma}} Y}\right]^{n/2} \frac{(g_l^{-1})^m g_s^m}{1+(g_l^{-1})^m (g_s^{-1})^{k_0-m}} (1+g_s^{-2})^{k_0/2}, \end{aligned} \tag{12}$$

where $Q_{\boldsymbol{\gamma}, g_s}$ is $Q_{\boldsymbol{\gamma}}$ by taking all $g_k = g_s$, $Q_{\boldsymbol{\gamma}, \boldsymbol{G_\gamma}}$ is $Q_{\boldsymbol{\gamma}}$ by using the two-component $\boldsymbol{G}$, and we assume model M_u has k_0 variables. The inequality in (12) is due to the fact that

$$|X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}} + \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1} X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}|^{1/2} = \left[|X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}} + \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1} X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}|^{1/n}\right]^{n/2},$$

and by the Minkowski determinant theorem (Marcus, 1992)

$$|X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}} + \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1} X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}|^{1/n} \geq |X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}}|^{1/n} + |\boldsymbol{G}_{\boldsymbol{\gamma}}^{-1} X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1}|^{1/n}.$$

Let

$$A_{g_s} = \frac{1}{1+g_s^{-2}} \boldsymbol{X_\gamma}(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X_\gamma})^{-1} \boldsymbol{X}_{\boldsymbol{\gamma}}^T, \text{ and } A_{\boldsymbol{G_\gamma}} = \boldsymbol{X_\gamma}(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X_\gamma} + \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X_\gamma} \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1})^{-1} \boldsymbol{X}_{\boldsymbol{\gamma}}^T.$$

Combined with the conditional posterior mean of $\boldsymbol{\beta_\gamma}$, we can see that $\boldsymbol{Y}^T(I - A_{g_s} A_{g_s})\boldsymbol{Y}$ and $\boldsymbol{Y}^T(I - A_{\boldsymbol{G_\gamma}} A_{\boldsymbol{G_\gamma}})\boldsymbol{Y}$ are the sum of squared errors under the model with $g_k = g_s$ and the model with $\boldsymbol{G_\gamma}$, respectively. Write $A_{\boldsymbol{G_\gamma}}$ as

$$A_{\boldsymbol{G_\gamma}} = \boldsymbol{X_\gamma}(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X_\gamma} + g_s^{-2} \mathrm{diag}(1, g_s/g_l) \boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X_\gamma} \mathrm{diag}(1, g_s/g_l))^{-1} \boldsymbol{X}_{\boldsymbol{\gamma}}^T,$$

with $\mathrm{diag}(1, g_s/g_l)$ denoting the diagonal elements of $\boldsymbol{G_\gamma}/g_s$. Since g_s/g_l approaches zero as $n \to \infty$, we have as n becomes large, $\boldsymbol{Y}^T(I - A_{g_s} A_{g_s})\boldsymbol{Y} < \boldsymbol{Y}^T(I - A_{\boldsymbol{G_\gamma}} A_{\boldsymbol{G_\gamma}})\boldsymbol{Y}$. Consequently, we have $\boldsymbol{Y}^T(I - A_{g_s})\boldsymbol{Y} < \boldsymbol{Y}^T(I - A_{\boldsymbol{G_\gamma}})\boldsymbol{Y}$, because $\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X_\gamma}$ in A_{g_s} and $(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X_\gamma} + \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X_\gamma} \boldsymbol{G}_{\boldsymbol{\gamma}}^{-1})$ in $A_{\boldsymbol{G_\gamma}}$ are positive definite. Bringing this inequality to the first factor in (12), we have

$$\left[Y^T Q_{\boldsymbol{\gamma},g_s} Y/(Y^T Q_{\boldsymbol{\gamma},\boldsymbol{G}_{\boldsymbol{\gamma}}} Y)\right]^{n/2} \to 0,$$

as $n \to \infty$. For the second factor in (12),

$$(g_l^{-1})^m g_s^m / \left(1 + (g_l^{-1})^m (g_s^{-1})^{k_0 - m}\right) = (g_s/g_l)^m / \left(1 + g_s^{k_0}(g_s/g_l)^m\right) \to 0,$$

as $n \to \infty$. The last factor in (12), $(1 + g_s^{-2})^{k_0/2}$, is clearly finite given $k_0 < \infty$. The results of these three factors indicate that $R_u(\boldsymbol{G}_{\boldsymbol{\gamma}}, g_s) \to 0$ as $n \to \infty$. We denote this as Result B.

Combining Results A and B, we have as $n \to \infty$,

$$R_{u,t}(\boldsymbol{G}_{\boldsymbol{\gamma}})/R_{u,t}(g_s) \quad = \quad \left[\frac{Pr(M_u|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{G}_{\boldsymbol{\gamma}})}{Pr(M_t|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{G}_{\boldsymbol{\gamma}})}\right] / \left[\frac{Pr(M_u|\boldsymbol{Y},\boldsymbol{X},g_s)}{Pr(M_t|\boldsymbol{Y},\boldsymbol{X},g_s)}\right] \to 0,$$

that is, $R_{u,t}(\boldsymbol{G}_{\boldsymbol{\gamma}}) = o(R_{u,t}(g_s))$.

From Fernández et al. (2001), we have $R_{u,t}(g_s) \to 0$ as $n \to \infty$. Finally, since $R_{u,t}(\boldsymbol{G}_{\boldsymbol{\gamma}}) = o(R_{u,t}(g_s))$, it is straightforward that $R_{u,t}(\boldsymbol{G}) \to 0$ asymptotically. □

Appendix C

The Prior and Joint Posterior Distribution with Measurement Errors in $\boldsymbol{X}$ Incorporated When measurement error presents, prior distributions for $\boldsymbol{X}$ and $\boldsymbol{\Sigma_u}$ are needed. Following Carroll et al. (2006), the prior distribution of $\boldsymbol{X}_i$ is given by $\boldsymbol{X}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_X)$ with $\boldsymbol{\Sigma}_X = \text{diag}(\sigma^2_{x_1}, \ldots, \sigma^2_{x_p})$. The hyper-prior distribution of $\sigma^2_{x_k}, k = 1, \ldots, p$, is assumed to be an inverse gamma distribution, $\sigma^2_{x_k} \sim \textit{Inv-Gam}(\delta_{\sigma_{x_k},1}, \delta_{\sigma_{x_k},2})$ with small $\delta_{\sigma_{x_k},1}$ and $\delta_{\sigma_{x_k},2}$. In this article, we assume independent measurement errors and $\boldsymbol{\Sigma_u} = \text{diag}(\sigma^2_{u_1}, \ldots, \sigma^2_{u_p})$ with $\sigma^2_{u_k} \sim \textit{Inv-Gam}(\delta_{\sigma_u,1}, \delta_{\sigma_u,2})$. Caution should be given to the choices of small shape and scale parameters. As Gelman (2006) pointed out, if the sample size is small or values of the variance are small, posterior inferences will become sensitive to the choice of the shape and scale parameters, and these prior distributions will be nowhere noninformative.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{c}, \boldsymbol{\Sigma}_X, \boldsymbol{\Sigma_u}, \pi_k, q_k)$ denote a collection of parameters. The joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{X}$ is, up to a normalizing constant,

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{X}_{\boldsymbol{\gamma}}|\boldsymbol{Y}, \boldsymbol{W}, b) \quad &\propto \quad p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{X}_{\boldsymbol{\gamma}})p(\boldsymbol{W}|\boldsymbol{\theta}, \boldsymbol{X}_{\boldsymbol{\gamma}})p(\boldsymbol{\theta}, \boldsymbol{X}_{\boldsymbol{\gamma}}) \\ &= \quad p(\boldsymbol{Y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{X}_{\boldsymbol{\gamma}})p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \sigma^2, \boldsymbol{c}, b, \boldsymbol{X}_{\boldsymbol{\gamma}})p(\boldsymbol{W}|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_u) \\ &\quad \cdot p(\boldsymbol{X}_{\boldsymbol{\gamma}}|\boldsymbol{\Sigma}_X)p(\sigma^2)p(\boldsymbol{\Sigma}_X)p(\boldsymbol{\Sigma}_u)p(\boldsymbol{\gamma}|\boldsymbol{\pi})p(\boldsymbol{c}|\boldsymbol{q})p(\boldsymbol{\pi})p(\boldsymbol{q}). \end{aligned}$$

In the following, we only list the full conditional posterior distributions that depend on mis-measured candidate variables. The full conditional posterior distributions of $\sigma^2_{x_k}$

and $\sigma^2_{u_k}$ are

$$\begin{aligned} p(\sigma^2_{x_k}|\boldsymbol{X}_k) &= \textit{Inv-Gam}\left(\frac{n}{2}+\delta_{\sigma_{x_k},1}, \frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{2}+\delta_{\sigma_{x_k},2}\right), \\ p(\sigma^2_{u_k}|\boldsymbol{W}_k,\boldsymbol{X}_k) &= \textit{Inv-Gam}\left\{\frac{nJ}{2}+\delta_{\sigma_u,1}, \frac{(\boldsymbol{W}_k-\boldsymbol{X}_k)^T(\boldsymbol{W}_k-\boldsymbol{X}_k)}{2}+\delta_{\sigma_u,2}\right\}, \end{aligned}$$

where $\boldsymbol{W_k}$ and $\boldsymbol{X_k}$ denote the observed and true measure of variable k, respectively, and J denotes the number of repeated measures. The full conditional posterior distribution of $\boldsymbol{X}_i$ is

$$\begin{aligned} p(X_{k,i}|\boldsymbol{Y},\boldsymbol{W},\boldsymbol{\beta}_\gamma,\boldsymbol{\gamma},\sigma^2,\sigma^2_{x_k},\sigma^2_{u_k}) &= N(\mu_{k,i},\sigma^2_k), \\ \sigma^2_k &= 1/(\beta^2_k/\sigma^2+1/\sigma^2_{x_k}+J/\sigma^2_{u_k}), \\ \mu_{k,i} &= \sigma^2_k\left\{\beta_k/\sigma^2(Y_i-\sum_{l\neq k}X_{l,i}\beta_l)+J\overline{W}_{k,i}/\sigma^2_{u_k}\right\}, \\ \overline{\boldsymbol{W}} &= \sum_{j=1}^{J}\boldsymbol{W}_j/J. \end{aligned}$$

Appendix D

Measurement Error Incorporated Variable Selection Summary Statistics

$n=30$	$g=10$	$g=\sqrt{n}$	g	G	Hyper-g	Liang & Li
% correct	6.3	16.5	39.9	37.2	40.9	4.8
% overfit	76.3	71.5	52.5	50.3	49.5	21.3
Med. b	–	–	672.67	1599.83	–	–
Med. MSL	6.060	0.271	1.62	1.002	0.59	2.837
$n=200$	$g=10$	$g=\sqrt{n}$	g	G	Hyper-g	Liang & Li
% correct	7.8	12.8	50.7	58.7	60.5	8.2
% overfit	92.2	87.2	49.3	41.3	39.3	90.7
Med. b	–	–	2090.94	1286.88	–	–
Med. MSL	0.099	0.205	0.407	0.048	0.172	0.153
$n=600$	$g=10$	$g=\sqrt{n}$	g	G	Hyper-g	Liang & Li
% correct	3.4	14.3	64.9	82.9	71.5	47.8
% overfit	96.6	85.7	35.1	17.1	23.2	52.2
Med. b	–	–	2107.47	1331.63	–	–
Med. MSL	0.063	0.121	0.153	0.009	0.133	0.037

Table 4: Variable selection statistics for Example 2 ($n = 30, 200, 600$), reliability ratio 0.8. Med. b, median of the tuning parameter; Med. MSL, Median mean squared loss.

$n = 30$	$g = 10$	$g = \sqrt{n}$	g	G	Hyper-g	Liang & Li
% correct	6.8	16.3	31.7	51.7	66.0	0.8
% overfit	87.7	79.7	51.1	45.1	32.1	63.3
Med. b	–	–	769.77	469.52	–	–
Med. MSL	0.090	0.228	0.543	0.371	0.202	1.551
$n = 200$	$g = 10$	$g = \sqrt{n}$	g	G	Hyper-g	Liang & Li
% correct	10.4	19.4	97.8	99.9	83.2	25.2
% overfit	89.6	80.6	2.2	0.1	16.8	74.8
Med. b	–	–	662.94	436.32	–	–
Med. MSL	0.044	0.037	0.026	0.025	0.026	0.066
$n = 600$	$g = 10$	$g = \sqrt{n}$	g	G	Hyper-g	Liang & Li
% correct	11.6	33.4	98.8	99.9	88.5	73.8
% overfit	88.4	66.6	1.2	0.1	11.5	26.2
Med. b	–	–	627.21	355.39	–	–
Med. MSL	0.019	0.011	0.007	0.006	0.008	0.019

Table 5: Variable selection statistics for Example 2 ($n = 30, 200, 600$), reliability ratio 0.9. Med. b, median of the tuning parameter; Med. MSL, Median mean squared loss.

Appendix E

Description of the Variables in the Ozone Data

ozone Daily ozone concentration (maximum one hour average, parts per million) at Upland, CA;

vh Vandenburg 500 millibar pressure height (m);

wind Wind speed (mph) at Los Angeles International Airport (LAX);

hum Humidity (percent) at LAX;

temp Sandburg Air Force Base temperature ($F°$);

ibh Inversion base height at LAX;

ibt Inversion base temperature at LAX;

dpg Daggett Pressure gradient (mm Hg) from LAX to Daggett, CA;

vis Visibility (miles) at LAX.

References

Bartlett, M. (1957). "A comment on D. V. Lindley's statistical paradox." *Biometrika*, 44: 533–534. 354

Brass, D. M., Tomfohr, J., Yang, I. V., and Schwartz, D. A. (2007). "Using Mouse Genomics to Understand Idiopathic Interstitial Fibrosis." *Proceedings of the American Thoracic Society*, 4: 92–100. 366

Breiman, L. and Friedman, J. H. (1985). "Estimating optimal transformations for multiple regression and correlation." *Journal of the American Statistical Association*, 80: 580–598. MR0803258. 367

Browne, W. J. and Draper, D. (2006). "A Comparison of Bayesian and Likelihood-based Methods for Fitting Multilevel Models (Pkg: P473-550)." *Bayesian Analysis*, 1(3): 473–514. MR2221283. 359

Carroll, R., Ruppert, D., Stefanske, L. A., and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC Press. MR2243417. doi: `http://dx.doi.org/10.1201/9781420010138`. 354, 360, 367, 374

Casella, G. and Moreno, E. (2006). "Objective Bayesian variable selection." *Journal of the American Statistical Association*, 101: 157–167. MR2268035. doi: `http://dx.doi.org/10.1198/016214505000000646`. 367

Christensen, P. J., Bailie, M. B., Goodman, R. E., O'Brien, A. D., Toews, G. B., and Paine, R. (2000). "Role of diminished epithelial GM-CSF in the pathogenesis of bleomycin-induced pulmonary fibrosis." *American Journal of Physiology – Lung Cellular and Molecular Physiology*, 279: L487–L495. 365

Fan, J. and Li, R. (2001). "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties." *Journal of the American Statistical Association*, 96(456): 1348–1360. MR1946581. doi: `http://dx.doi.org/10.1198/016214501753382273`. 353

Fernández, C., Ley, E., and Steel, M. (2001). "Benchmark priors for Bayesian model averaging." *Journal of Econometrics*, 100(2): 381–427. MR1820410. doi: `http://dx.doi.org/10.1016/S0304-4076(00)00076-2`. 354, 355, 360, 374

Foster, D. P. and George, E. I. (1994). "The risk inflation criterion for multiple regression." *Annals of Statistics*, 22: 1947–1975. MR1329177. doi: `http://dx.doi.org/10.1214/aos/1176325766`. 355

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1: 515–533. MR2221284. 374

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman & Hall/CRC. MR1385925. 358, 361

George, E. I. (2000). "The Variable Selection Problem." *Journal of the American Statistical Association*, 95: 1304–1308. MR1825282. doi: `http://dx.doi.org/10.2307/2669776`. 354

George, E. I. and Foster, D. P. (2000). "Calibration and Empirical Bayes Variable Selection." *Biometrika*, 87: 731–747. MR1813972. doi: `http://dx.doi.org/10.1093/biomet/87.4.731`. 354, 355

George, E. I. and McCulloch, R. E. (1993). "Variable Selection via Gibbs Sampling." *Journal of the American Statistical Association*, 88: 881–889. 354, 357, 361

— (1997). "Approaches for Bayesian Variable Selection." *Statistica Sinica*, 7: 339–374. 354, 356

— (1999). "Comment on "Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-based Prior"." *Journal of the American Statistical Association*, 94: 798–799. MR1723272. doi: http://dx.doi.org/10.2307/2669990. 354

Higgins, K. M., Davidian, M., and Giltinan, D. M. (1997). "A Two-step Approach to Measurement Error in Time-dependent Covariates in Nonlinear Mixed-effects Models, with Application to IGF-I Pharmacokinetics." *Journal of the American Statistical Association*, 92: 436–448. 354, 360

Huaux, F., Gharaee-Kermani, M., Liu, T., Morel, V., McGarry, B., Ullenbruch, M., Kunkel, S. L., Wang, J., Xing, Z., and Phan, S. H. (2005). "Role of Eotaxin-1 (CCL11) and CC Chemokine Receptor 3 (CCR3) in Bleomycin-Induced Lung Injury and Fibrosis." *The American Journal of Pathology*, 167: 1485–1496. 365

Ishwaran, H. and Rao, J. S. (2005a). "Spike and Slab Gene Selection for Multigroup Microarray Data." *Journal of the American Statistical Association*, 100: 764–780. MR2201009. doi: http://dx.doi.org/10.1198/016214505000000051. 354

— (2005b). "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies." *The Annals of Statistics*, 33: 730–773. MR2163158. doi: http://dx.doi.org/10.1214/009053604000001147. 354

Knight, D., Ernst, M., Anderson, G., Moodley, Y., and Mutsaers, S. (2003). "The role of gp130/IL-6 cytokines in the development of pulmonary fibrosis: critical determinants of disease susceptibility and progression?" *Pharmacology & Therapeutics*, 99: 327–338. 365

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., and Mallick, B. K. (2003). "Gene selection: a Bayesian variable selection approach." *Bioinformatics*, 19: 90–97. 354

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g priors for Bayesian Variable Selection." *Journal of the American Statistical Association*, 103: 410–423. MR2420243. doi: http://dx.doi.org/10.1198/016214507000001337. 354, 355, 360, 367

Liang, H. and Li, R. (2009). "Variable Selection for Partially Linear Models with Measurement Errors." *Journal of the American Statistical Association*, 104: 234–248. MR2504375. doi: http://dx.doi.org/10.1198/jasa.2009.0127. 354, 360, 364, 367, 369

Lindley, D. V. (1957). "A Statistical Paradox." *Biometrika*, 44: 187–192. MR0087273. 354

Liu, W. and Wu, L. (2007). "Simultaneous Inference for Semiparametric Nonlinear Mixed-Effects Models with Covariate Measurement Errors and Missing Responses." *Biometrics*, 63: 342–350. MR2370792. doi: http://dx.doi.org/10.1111/j.1541-0420.2006.00687.x. 354, 360

Ma, Y. and Li, R. (2010). "Variable Selection in Measurement Error Models." *Bernoulli*, 16: 274–300. MR2648758. doi: http://dx.doi.org/10.3150/09-BEJ205. 354, 360

Marcus, M. (1992). *A survey of matrix theory and matrix inequalities*, Vol. 14. Mineola, NY: Courier Dover Publications, Inc. MR1215484. 373

Maruyama, Y. and George, E. I. (2011). "Fully Bayes Factor with a Generalized g-prior." *The Annals of Statistics*, 39: 2740–2765. MR2906885. doi: http://dx.doi.org/10.1214/11-AOS917. 354, 360, 365

Miller, A. (2002). *Subset Selection in Regression*. New York: Chapman and Hall. MR2001193. doi: http://dx.doi.org/10.1201/9781420035933. 361, 367

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian Variable Selection in Linear Regression (C/R: P1033-1036)." *Journal of the American Statistical Association*, 83: 1023–1032. MR0997578. 354

Morris, C. N. (1987). "Comments on "The Calculation of Posterior Distributions by Data Augmentation"." *Journal of the American Statistical Association*, 82: 542–543. MR0898357. 357

Rocke, D. M. and Durbin, B. (2001). "A Model for Measurement Error for Gene Expression Arrays." *Journal of Computational Biology*, 8: 557–569. 354

Sabo-Attwood, T., Ramos-Nino, M. E., Eugenia-Ariza, M., MacPherson, M. B., Butnor, K. J., Vacek, S. P., P. C.and McGee, Clark, J. C., Steele, C., and Mossman, B. T. (2011). "Osteopontin Modulates Inflammation, Mucin Production, and Gene Expression Signatures After Inhalation of Asbestos in a Murine Model of Fibrosis." *The American Journal of Pathology*, 178: 1975–1985. 365, 366

Smith, M. and Kohn, R. (1996). "Nonparametric Regression Using Bayesian Variable Selection." *Journal of Econometrics*, 75: 317–343. 354, 360

Som, A., Hans, C., and MacEachern, S. (2014). "Bayesian Modeling with Blockwise Hyper-g Priors." arXiv:1406.6419. 354

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society, Series B (Methodological)*, 58: 267–288. MR1379242. 353

Vannucci, M., Do, K., and Müller, P. (2012). *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press. 359

Wu, Y., Boos, D. D., and Stefanski, L. A. (2007). "Controlling variable selection by the addition of pseudo variables." *Journal of the American Statistical Association*, 102: 235–243. MR2345541. doi: http://dx.doi.org/10.1198/016214506000000843. 358

Zellner, A. (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distributions." In: Goel, P. K. and Zellner, A. (eds.), *Bayesian*

Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, 233–243. Elsevier/North-Holland [Elsevier Science Publishing Co., New York; North-Holland Publishing Co., Amsterdam]. MR0881437. 354, 355

Zou, H. (2006). "The Adaptive LASSO and Its Oracle Properties." *Journal of the American Statistical Association*, 101: 1418–1429. MR2279469. doi: http://dx.doi.org/10.1198/016214506000000735. 353

Acknowledgments

This project was supported by NIH grants 1R03HL095429-01 (Principle Investigators H. Zhang and T. Sabo-Attwood), 1R21AI099367-01 (Principle Investigator H. Zhang), and NSF grant DMS-1006222 (Principle Investigator X. Huang). The authors sincerely thank the referees, Associate Editor, and Editor for their constructive and insightful comments, which improved the manuscript substantially, Professor Edward George at the University of Pennsylvania and Professor Robert McCulloch at the University of Chicago for their C codes and Professor Yuzo Maruyama for his R codes; Professor Hua Liang at the University of Rochester and Professor Raymond Carroll at the Texas A&M University for their comments and suggestions.