Taylor & Francis
Taylor & Francis Group

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Bayesian variable selection via a benchmark in normal linear models

Jun Shao, Kam-Wah Tsui & Sheng Zhang

To cite this article: Jun Shao, Kam-Wah Tsui & Sheng Zhang (2021) Bayesian variable selection via a benchmark in normal linear models, Statistical Theory and Related Fields, 5:1, 70-81, DOI: 10.1080/24754269.2020.1744074

To link to this article: https://doi.org/10.1080/24754269.2020.1744074

Published online: 27 Mar 2020.

STATISTICAL THEORY AND RELATED FIELDS
2021, VOL. 5, NO. 1, 70–81
https://doi.org/10.1080/24754269.2020.1744074

Taylor & Francis
Taylor & Francis Group

Bayesian variable selection via a benchmark in normal linear models

Jun Shao[a,b], Kam-Wah Tsui[b] and Sheng Zhang[b]

[a]KLATASDS-MOE, School of Statistics, East China Normal University Shanghai, People's Republic of China; [b]Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA

ABSTRACT
With increasing appearances of high-dimensional data over the past two decades, variable selections through frequentist likelihood penalisation approaches and their Bayesian counterparts becomes a popular yet challenging research area in statistics. Under a normal linear model with shrinkage priors, we propose a benchmark variable approach for Bayesian variable selection. The benchmark variable serves as a standard and helps us to assess and rank the importance of each covariate based on the posterior distribution of the corresponding regression coefficient. For a sparse Bayesian analysis, we use the benchmark in conjunction with a modified BIC. We also develop our benchmark approach to accommodate models with covariates exhibiting group structures. Two simulation studies are carried out to assess and compare the performances among the proposed approach and other methods. Three real datasets are also analysed by using these methods for illustration.

ARTICLE HISTORY
Received 13 August 2019
Revised 9 March 2020
Accepted 15 March 2020

KEYWORDS
Benchmark; lasso; sparse Bayes; shrinkage prior

1. Introduction

Over the past two decades, with advanced data collection techniques, a large amount of high-dimensional data continues to appear in various biological, medical, social, and economical studies. A typical example is the microarray data, where thousands or even millions of genes are involved in the data collection but only as few as hundreds or even fewer sampled subjects are available. Researchers believe that the majority of the genes are redundant and only a small subset is useful to predict the response of interest. Hence, it is desired to eliminate the unrelated genes and select important ones, for more accurate prediction as well as better interpretation. Such high-dimensional problems in practice impose great challenge to statistical analysis and motivate various variable selection techniques.

Lots of attempts have been made to solve these problems by regularisation methods, which achieve parameter estimation and variable selection simultaneously, mainly via frequentist approaches. These methods typically involve adding a penalty term on regression coefficients to the loss function, with the purpose of either parameter estimator variance stabilisation or variable selection; see, for example, the ridge regression by Hoerl and Kennard (1970), lasso by Tibshirani (1996), smoothly clipped absolute deviation (SCAD) by Fan and Li (2001), elastic net by Zou and Hastie (2005), fused lasso by Tibshirani et al. (2005), adaptive lasso by Zou (2006), COSSO by Lin and Zhang (2006), SICA by Lv and Fan (2009), MCP by Zhang (2010), truncated L1 by Shen et al. (2011), SELO by Dicker et al. (2011), and references therein.

On the other hand, variable selection via Bayesian approaches is also very active, started with the well-known Bayesian information criterion (BIC) (Schwarz, 1978). There exist three types of commonly used Bayesian approaches in variable selection. The first type works on information criterion, such as the BIC and its improvement PBIC proposed by Bayarri et al. (2019). The second type includes the indicator model selection (see, for example, Brown et al., 1998; Dellaportas et al., 1997; George & McCulloch, 1993; Kuo & Mallick, 1998; Yuan & Lin, 2005), the stochastic search method (e.g., O'Hara & Sillanpää, 2009), and the model space method by Green (1995). The third type, which is considered in the current paper, is to apply priors on the regression coefficients that promotes the shrinkage of coefficients towards 0. This last type of approaches is intrinsically connected with frequentist methods in the sense that such priors play the same role as the assumption that the coefficients are sparse for the frequentist approach. Typical examples of this type include the Bayesian lasso (Park & Casella, 2008) and Bayesian counterparts for elastic net, group lasso, and fused lasso (Kyung et al., 2010).

The shrinkage prior approach, however, does not provide sparse estimates of regression coefficients in general. A Bayesian analysis based on a subset of covariates with size considerably less than the original dimensionality, which is referred to as sparse Bayesian analysis, may produce better results than

CONTACT Jun Shao ✉ shao@stat.wisc.edu

the Bayesian analysis based on all covariates. Several attempts have been made to obtain sparse Bayesian estimates based on shrinkage priors. For instance, Hoti and Sillanpää (2006) proposed a method based on thresholding; however, the method is based on certain approximations and the choice of threshold is ad hoc. Another example is the sparse Bayesian learning by Tipping (2001), but it involves complicated nonconvex optimisation and assumes that the variance of the error term is known.

Under the framework of shrinkage priors, in this paper, we propose a Bayesian variable selection in a normal linear model via a benchmark variable that serves as a standard and helps us to assess and rank the importance of each covariate based on the posterior distribution of the corresponding regression coefficient. For a sparse Bayesian analysis, we propose a variable selection using benchmark in conjunction with a modified BIC. Furthermore, we develop our benchmark approach to accommodate normal linear models with covariates exhibiting group structures. An additional step is implemented to identify important individual variables within the selected groups. Some simulation studies are carried out to assess and compare the performances among the proposed approach and other methods. Three real datasets are also analysed by using these methods for illustration.

2. Methodology

Let $\boldsymbol{y}$ be an n-dimensional vector of responses and, without loss of generality, let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ be p centralised n-dimensional vectors of covariates. Conditional on $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, $\boldsymbol{y}$ is assumed to be distributed as multivariate normal $N(\beta_0\mathbf{1} + \boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, $\boldsymbol{a}'$ denotes the transpose of $\boldsymbol{a}$, $\beta_0, \beta_1, \ldots, \beta_p$ are $p+1$ unknown parameters, σ is an unknown positive parameter, $\mathbf{1}$ is the n-dimensional vector with all components 1, and $\boldsymbol{I}$ is the identity matrix of order n. Note that components of $\boldsymbol{X}$ can be individual covariate vectors as well as vectors having interaction effects on $\boldsymbol{y}$ such as product terms and, hence, components of $\boldsymbol{\beta}$ are main effects and interaction effects.

There are various choices of priors that shrink the regression coefficients, components of $\boldsymbol{\beta}$, towards 0. The most popular one is the Laplace prior considered by Park and Casella (2008) for their Bayesian lasso:

$$p(\boldsymbol{\beta}|\sigma^2) = \prod_{i=1}^{p} \frac{\lambda}{2\sigma} \exp\left(-\frac{\lambda|\beta_i|}{\sigma}\right) \tag{1}$$

where $\lambda > 0$ is a hyperparameter. For β_0 and σ^2 that are not involved with variable selection, we consider noninformative priors, i.e., the prior of β_0 is the Lebesgue measure and the prior of σ^2 has improper density σ^{-2}.

2.1. Benchmark

If the posterior distribution of β_i is nearly the same as that from a noise variable centred at 0, then it is natural to eliminate $\boldsymbol{x}_i$ as an unimportant covariate. However, the question is how to quantify whether a posterior distribution to be close to that of a noise.

To illustrate our idea, let us first consider an artificial case where a covariate $\boldsymbol{z}$ exists and is known to have no effect on $\boldsymbol{y}$, i.e., $\boldsymbol{y}$ conditioned on $(\boldsymbol{X}, \boldsymbol{z})$ is distributed as $N(\boldsymbol{z}\beta_z + \mathbf{1}\beta_0 + \boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ with $\beta_z = 0$. Although we know $\boldsymbol{z}$ is redundant, we still put a prior on β_z such that β_z and β_i's are independently identically distributed conditioning on σ^2. Under this setting, $\boldsymbol{x}_i$ could be treated as an unimportant variable if the posterior of β_i is similar to the posterior of β_z. In other words, the variable $\boldsymbol{z}$ serves as a benchmark in measuring the importance of $\boldsymbol{x}_i$'s.

To be more rigorous, a nonzero vector $\boldsymbol{z}$ is defined as a valid benchmark if it satisfies the following two conditions:

(C1) The posterior distribution of $\boldsymbol{\beta}$ given $(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}, \beta_z, \sigma^2)$ is the same as the posterior distribution of $\boldsymbol{\beta}$ given $(\boldsymbol{y}, \boldsymbol{X}, \sigma^2)$.

(C2) The posterior distribution of β_z given $(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}, \sigma^2)$ is centred at 0.

Condition (C1) ensures that the presence of a benchmark variable would not affect the Bayesian analysis concerning unknow $\boldsymbol{\beta}$, while (C2) guarantees that the benchmark can be used as a standard to assess the importance of covariates in terms of the posterior distributions of β_i, $i = 1, \ldots, p$.

How do we find a benchmark variable when we do not have a redundant variable at hand? We now show that a universal solution of $\boldsymbol{z}$ simultaneously satisfying (C1) and (C2) does exist. Under the Bayesian framework with column-wisely centralised $\boldsymbol{X}$, the density of $\boldsymbol{y}$ given $(\boldsymbol{X}, \boldsymbol{z}, \beta_0, \boldsymbol{\beta}, \beta_z, \sigma^2)$ is proportional to

$$\frac{1}{\sigma^n}\exp\left(-\frac{\|\boldsymbol{y} - \boldsymbol{z}\beta_z - \mathbf{1}\beta_0 - \boldsymbol{X\beta}\|^2}{2\sigma^2}\right)$$
$$= \frac{1}{\sigma^n}\exp\left(-\frac{\|\tilde{\boldsymbol{y}} - \boldsymbol{X\beta}\|^2 + \|\boldsymbol{z} - \bar{z}\mathbf{1}\|^2\beta_z^2 - 2\beta_z \boldsymbol{z}'(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta}) + n(\beta_0 - \bar{y} + \beta_z\bar{z})^2}{2\sigma^2}\right)$$

where $\bar{y}$ is the average of the components of $\boldsymbol{y}$, $\bar{z}$ is the average of the components of $\boldsymbol{z}$, $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \bar{y}\mathbf{1}$, and $\|\boldsymbol{a}\|^2 = \boldsymbol{a}'\boldsymbol{a}$. For the prior of $(\beta_z, \beta_0, \boldsymbol{\beta}, \sigma^2)$, we consider it to be $\propto \sigma^{-3}\exp(-\lambda|\beta_z|/\sigma)p(\boldsymbol{\beta}|\sigma^2)$, where $p(\boldsymbol{\beta}|\sigma^2)$ is given by (1).

Since the intercept β_0 is not of interest, we integrate it out from the posterior density $p(\beta_0, \boldsymbol{\beta}, \beta_z|\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{y}, \sigma^2)$.

Then,

$$p\left(\boldsymbol{\beta}, \beta_z | \boldsymbol{X}, \boldsymbol{z}, \boldsymbol{y}, \sigma^2\right) \propto \frac{1}{\sigma^{n+p+1}} \times \exp\left(-\frac{\left\|\tilde{\boldsymbol{y}} - \boldsymbol{X\beta}\right\|^2 + 2\beta_z \boldsymbol{z}'\boldsymbol{X\beta}}{2\sigma^2} - \frac{\lambda}{\sigma}\sum_{j=1}^{p}|\beta_j|\right) \times \exp\left(-\frac{\|\boldsymbol{z} - \bar{z}\mathbf{1}\|^2 \beta_z^2 - 2\boldsymbol{z}'\tilde{\boldsymbol{y}}\beta_z}{2\sigma^2} - \frac{\lambda}{\sigma}|\beta_z|\right) \quad (2)$$

Note that marginalisation over β_0 is equivalent to centralising the response $\boldsymbol{y}$. After integrating out β_0, the posterior inferences are drawn from the centralised response $\tilde{\boldsymbol{y}}$ instead of the original $\boldsymbol{y}$. The reason that we introduce β_0 in the model and then integrate it out, instead of eliminating it at the very beginning and directly building a linear regression model as $\tilde{\boldsymbol{y}} = \boldsymbol{z}\beta_z + \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, is mainly for the mathematical rigorousness, as $\tilde{\boldsymbol{y}}$ is not of full rank and has a degenerate distribution.

The conditional posterior density in (2) implies that conditioned on $(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}, \sigma^2)$, $\boldsymbol{\beta}$ and β_z are independent if and only if $\boldsymbol{z}'\boldsymbol{X} = 0$, and β_z has mean zero if and only if $\boldsymbol{z}'\tilde{\boldsymbol{y}} = 0$. In other words, (C1) and (C2) both hold if and only if $\boldsymbol{z}$ is orthogonal to $(\boldsymbol{X}, \tilde{\boldsymbol{y}})$. Clearly, $\boldsymbol{z} = \mathbf{1}$ is a direct solution and could be used as a benchmark to assess the importance of $\boldsymbol{x}_i$'s. Note that when $\boldsymbol{z} = \mathbf{1}$, the posterior density of β_z remains the same as its prior, and the posterior density of $(\boldsymbol{\beta}, \beta_z, \sigma^2)$ is simplified to

$$p\left(\boldsymbol{\beta}, \beta_z, \sigma^2 | \boldsymbol{X}, \boldsymbol{y}\right) \propto \frac{1}{\sigma^{n+p+3}} \times \exp\left(-\frac{\left\|\tilde{\boldsymbol{y}} - \boldsymbol{X\beta}\right\|^2}{2\sigma^2} - \frac{\lambda}{\sigma}\sum_{j=1}^{p}|\beta_j| - \frac{\lambda}{\sigma}|\beta_z|\right) \quad (3)$$

The fact that $\boldsymbol{z} = \mathbf{1}$ can be used as a benchmark does not rely on the form of prior given in (1). If the prior in (1) is replaced by a multivariate normal prior, then the result is related with ridge regression, rather than lasso or Bayesian lasso. Computation might be an issue when the prior is non-normal.

The idea of benchmark in Bayesian framework is similar to the application of pseudo variables in frequentist approach (Breiman 2001, Wu et al. 2007). The only requirement for a pseudo variable is its independence with $(\boldsymbol{X}, \boldsymbol{y})$. Such a pseudo variable is not applicable here since it is likely that the pseudo variable does not satisfy (C1) due to the fact that orthogonality is a stronger assumption than independence in general.

2.2. Example

Even without a well-defined variable selection, we now consider a real data example to illustrate how we utilise a benchmark to assess importance of covariates.

The prostate cancer data originally came from a research conducted by Stamey et al. (1989), and it was studied by Tibshirani (1996) and Zou and Hastie (2005). The goal of the research was to explore the relation between the level of prostate-specific antigen and several clinical measures in men before their hospitalisation for radical prostatectomy. The dataset contains 97 patients with the logarithm of prostate-specific antigen (lpsa) as the response and eight covariates, logarithm of cancer volume (lcavol), logarithm of prostate weight (lweight), age, logarithm of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), logarithm of capsular penetration (lcp), Gleason score (gleason), and percentage Gleason score 4 or 5 (pgg45).

Figure 1 visualises the posteriors. The leftmost boxplot is based on the posterior samples of the coefficient for the benchmark $\boldsymbol{z} = \mathbf{1}$. It is distributed symmetrically around 0 as expected. Other box plots represent the posterior distributions of the coefficients associated

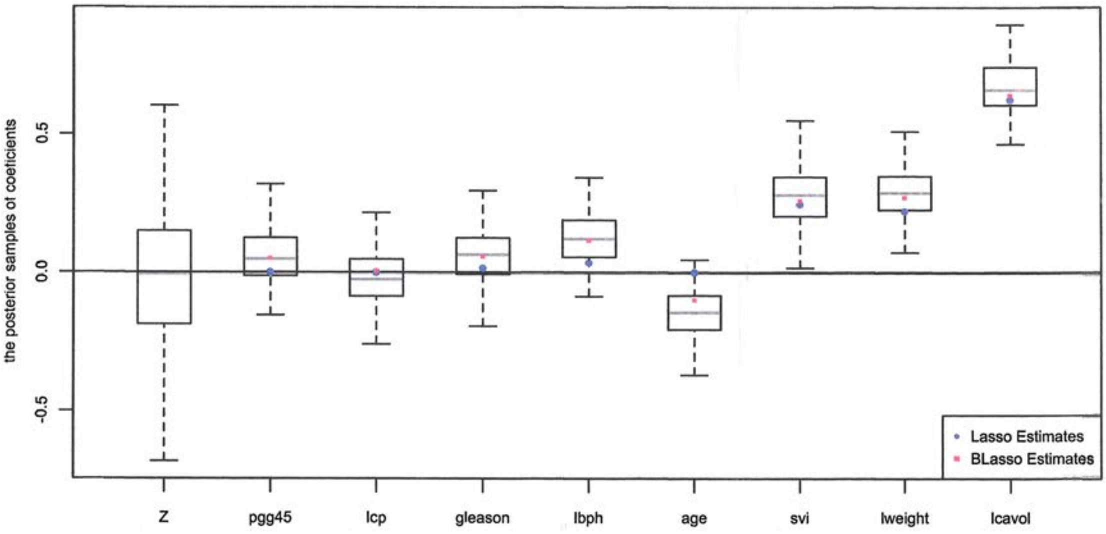

Figure 1. Posterior plots with the prostate cancer data.

with eight covariates. It can be seen that the three posteriors plotted in the far right of Figure 1 are clearly different from the posterior of the benchmark and, hence, we may conclude that the corresponding three covariates, svi, lweight, and lcavol, are useful for the response. On the other hand, the posteriors of three covariates next to the benchmark in Figure 1 are not different from the benchmark posterior and, hence, the covariates pgg45, lcp, and gleason are not useful. The posteriors of lbph and age are just marginally different from that of the benchmark, and we may still consider them to be not useful covariates.

Figure 1 also includes lasso and Bayesian lasso estimates of each coefficients, marked as circles and squares in the figure. The lasso estimates are zero for pgg45, lcp, and age, nonzero for the other five covariates. Thus, the lasso approach agrees with our approach for covariates pgg45, lcp, age, svi, lweight, and lcavol, but does not agree on gleason and lbph. Since the magnitudes of lasso estimates for gleason and lbph are small, another thresholding added to lasso will result in the same conclusion with ours. Meanwhile, the Bayesian lasso evaluates all the coefficients to be nonzero as it does not select variables to promote model sparsity.

2.3. Variable selection

The benchmark serves as a measure to assess the importance of each covariate. To compare the effect of each $\boldsymbol{x}_i$ with that of the benchmark $\boldsymbol{z}$, we define the importance score d_i for each $\boldsymbol{x}_i$ based on the following conditional posterior probability:

$$d_i = P\left(\frac{|\beta_i|}{\sqrt{V\left(\beta_i|\boldsymbol{y},\mathbf{X},\sigma^2\right)}} > \frac{|\beta_z|}{\sqrt{V\left(\beta_z|\boldsymbol{y},\mathbf{X},\sigma^2\right)}}\middle|\boldsymbol{y},\mathbf{X},\sigma^2\right) \quad (4)$$

where $V(\xi\,|\mathbf{A})$ denotes the posterior variance of ξ given $\mathbf{A}$. This probability could be evaluated either numerically or theoretically, depending on which prior is put on $\boldsymbol{\beta}$. The standardisation over the variances is necessary for the purpose of fair comparison. Intuitively, a d_i close to 0.5 indicates the effect of $\boldsymbol{x}_i$ is not much different from the effect of the benchmark and therefore $\boldsymbol{x}_i$ could be treated as an unimportant variable. With the availability of the estimated importance scores, the covariates $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p)$ could be ranked from the most important to the least important as $(\boldsymbol{x}_{(1)},\ldots,\boldsymbol{x}_{(p)})$, where $\boldsymbol{x}_{(1)}$ associates with the greatest estimated importance score, $\boldsymbol{x}_{(2)}$ associates with the second largest importance score and etc. It is desired to select covariates that are assessed to be the most important.

Naturally, the next question to be addressed is how to determine the cutoff point m^* such that only the top m^* variables $(\boldsymbol{x}_{(1)},\ldots,\boldsymbol{x}_{(m^*)})$ are selected. To avoid arbitrary thresholding on the estimated importance scores, we adopt a slighted modified BIC criterion (Chen & Chen, 2008). For each integer $m = 1,\ldots,p$, the m most important covariates $\boldsymbol{x}_{(1)},\ldots,\boldsymbol{x}_{(m)}$ are considered in a candidate model with $\boldsymbol{X}_m = (\boldsymbol{x}_{(1)},\ldots,\boldsymbol{x}_{(m)})$. The desired cutoff point m^* is the one that minimises

$$\text{BIC}\ (m) = \log(\left\|\tilde{\boldsymbol{y}} - \boldsymbol{X}_m\widehat{\boldsymbol{\beta}}_m\right\|^2/n) + \frac{m}{n}\left(\log n + \log p\right) \quad (5)$$

over m, where $\widehat{\boldsymbol{\beta}}_m$ is the posterior mean of the regression parameter under model m. The original BIC in Chen and Chen (2008) uses $2\log p$ instead of $\log p$ in (5). This slight modification does not alter the asymptotic properties established in Chen and Chen (2008) but has better simulation performance in our study.

For the prostate cancer example in Section 2.2, we compute d_i's and BIC(m) and show them in Table 1. It can be seen that BIC(m) reaches its minimum value -0.54 when $m^* = 3$, i.e., lcavol, lweight, and svi are selected as important covariates, or equivalently, we select covariates whose d_i values are over 0.9 in this example.

2.4. Computation

The Laplace prior in (1) is a shrinkage prior, but it is not conjugate and, hence, Bayesian computation is complicated. Fortunately, we can follow the approach in Park and Casella (2008) to carry out Bayesian computation using Gibbs sampler and to estimate λ using marginal likelihood. This is based on the fact that the Laplace distribution is a scale mixture of normal distributions where the mixing is through an exponential distribution as follows (Andrews & Mallows, 1974),

$$\frac{a}{2}\exp\left(-a|z|\right) = \int_0^\infty \frac{1}{\sqrt{2\pi s}}\exp\left(-\frac{z^2}{2s}\right)\frac{a^2}{2} \times \exp\left(-\frac{a^2}{2}s\right)\mathrm{d}s \quad (6)$$

Using $\mathbf{1}$ as benchmark and applying (6), we obtain that the posterior density in (3) is proportional to

$$\frac{1}{\sigma^{n+p+3}}\exp\left(-\frac{\left\|\tilde{y}-\boldsymbol{X\beta}\right\|^2}{2\sigma^2}\right)\prod_{i=z,1,\ldots,p}\int_0^\infty\frac{1}{\tau_i} \times \exp\left(-\frac{\beta_i^2}{2\sigma^2\tau_i^2}-\frac{\lambda^2\tau_i^2}{2}\right)\mathrm{d}\tau_i^2$$

Table 1. Values of d_i and BIC(m) in prostate cancer example.

	lcavol	lweight	svi	age	lbph	gleason	lcp	pgg45
d_i	1.00	0.98	0.93	0.80	0.71	0.61	0.56	0.54
BIC(m)	−0.38	−0.47	−0.54	−0.49	−0.46	−0.41	−0.33	−0.26

which gives the following conditional distributions for Gibbs sampler:

$$\begin{aligned}
\beta_z \mid \text{all others} &\sim N\left(0, \tau_z^2\sigma^2\right)\\
\boldsymbol{\beta} \mid \text{all others} &\sim N\left(\boldsymbol{A}^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}}, \sigma^2\boldsymbol{A}^{-1}\right)\\
\sigma^2 \mid \text{all others} &\sim \text{Inv-Gamma}\ ((n+p)/2,\\
&\qquad ||\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}||^2/2 + \boldsymbol{\beta}'\boldsymbol{D}_\tau^{-1}\boldsymbol{\beta}/2 + \beta_z^2/2\tau_z^2)\\
\tau_i^{-2} \mid \text{all others} &\sim \text{Inv-Gaussian}\ \left(|\lambda\sigma/\beta_j|, \lambda^2\right)\\
\tau_z^{-2} \mid \text{all others} &\sim \text{Inv-Gaussian}\ \left(|\lambda\sigma/\beta_z|, \lambda^2\right)
\end{aligned}$$

where $\boldsymbol{A} = \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_\tau^{-1}$ and $\boldsymbol{D}_\tau$ is a $p \times p$ diagonal matrix with $\tau_1^2, \dots, \tau_p^2$ as diagonal components. In the kth iteration of the Gibbs sampler, the λ value estimated from the $(k-1)$th iteration is used to get the kth sample and is then updated by the kth sample as

$$\widehat{\lambda}^{(k)} = \sqrt{\frac{2(p+1)}{E_{\widehat{\lambda}^{(k-1)}}\left[\tau_z^2 + \sum_{j=1}^{p}\tau_j^2|\tilde{\boldsymbol{y}}, \boldsymbol{X}\right]}}$$

where the conditional expectation is evaluated by the average from Gibbs samples. The derivation is omitted since it is similar to that in Park and Casella (2008).

Once the posterior samples of β_z and $\boldsymbol{\beta}$ are obtained, the importance score d_i for each $\boldsymbol{x}_i$ specified in (4) can be approximated by the corresponding relative frequency $\widehat{d}_i$. The ranked $\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)}, \dots, \boldsymbol{x}_{(p)}$ can be obtained by sorting $\widehat{d}_i$'s descendingly. Finally, we can find the cutoff point m^* by minimising BIC in (5), with $\widehat{\boldsymbol{\beta}}_m$ being the posterior mean of the regression coefficient vector when $\boldsymbol{X}_m = (\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)}, \dots \boldsymbol{x}_{(m)})$.

2.5. Covariates with group structures

In some studies, the covariates exhibit certain group structure. It is then desired to capture the intrinsic relation among variables within a group. In this section, we extend the idea of using a benchmark for variable selection under the Bayesian framework to accommodate the group structures. We perform variable selection in both group and individual variable levels.

Suppose that p covariates can be partitioned into G groups with sizes $p_1, \dots, p_G$, respectively, where $\sum_{g=1}^{G} p_g = p$. The matrix $\boldsymbol{X}$ could be written as $\boldsymbol{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_G)$, where $\boldsymbol{X}_g = (\boldsymbol{x}_{g1}, \dots, \boldsymbol{x}_{gp_g})$ is a $n \times p_g$ matrix for the gth group, $g = 1, \dots, G$. The vector of associated regression coefficients can be written as $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \dots, \boldsymbol{\beta}_G')$, where each $\boldsymbol{\beta}_g' = (\beta_{g1}, \dots, \beta_{gp_g})$ is a vector of length p_g, $g = 1, \dots, G$.

The prior in (1) does not take the group structure into consideration. Instead, as inspired by the penalty term of group lasso (Yuan & Lin, 2005), we consider the following prior density which encourages shrinkage on group level:

$$p(\boldsymbol{\beta}|\sigma^2) = \prod_{g=1}^{G} \frac{\lambda}{2\sigma} \exp\left(-\frac{\lambda\sqrt{p_g\boldsymbol{\beta}_g'\boldsymbol{\beta}_g}}{\sigma}\right) \tag{7}$$

The idea of benchmark can be extended to accommodate group level variable selection. Since a benchmark could be regarded as an individual group with a single covariate, we can assign a Laplace prior to β_z as in Section 2.1 and consider joint prior of $(\beta_z, \beta_0, \boldsymbol{\beta}, \sigma^2)$ as

$$\frac{1}{\sigma^2} p(\boldsymbol{\beta}|\sigma^2) \frac{\lambda}{2\sigma} \exp\left(-\frac{\lambda\sqrt{\beta_z^2}}{\sigma}\right)$$

where the prior of β_z matches the form of prior for $\boldsymbol{\beta}_g$ in (7), $g = 1, \dots, G$. Since the prior does not affect the fact that $\mathbf{1}$ is a benchmark as long as the prior of β_z has mean 0, we can still use $\mathbf{1}$ as a benchmark for group variable selection. It follows from (6) that

$$\begin{aligned}
&\exp\left(-\frac{\lambda\sqrt{p_g\boldsymbol{\beta}_g'\boldsymbol{\beta}_g}}{\sigma}\right)\\
&\quad = \int_0^\infty \frac{\lambda}{\sqrt{2\pi}\tau_g} \exp\left(-\frac{p_g\boldsymbol{\beta}_g'\boldsymbol{\beta}_g}{2\sigma^2\tau_g^2}\right)\\
&\qquad \exp\left(-\frac{\lambda^2}{2}\tau_g^2\right) d\tau_g^2
\end{aligned}$$

Then, after integrating out β_0, we obtain that the posterior density of $(\boldsymbol{\beta}, \beta_z, \sigma^2$ is proportional to

$$\begin{aligned}
&\frac{1}{\sigma^{n+G+3}} \exp\left(-\frac{\|\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\right) \prod_{g=z,1,\dots,G} \int_0^\infty \frac{1}{\tau_g}\\
&\quad \times \exp\left(-\frac{p_g\boldsymbol{\beta}_g'\boldsymbol{\beta}_g}{2\sigma^2\tau_g^2} - \frac{\lambda^2\tau_g^2}{2}\right) d\tau_g^2
\end{aligned}$$

which gives the following full conditional distributions:

$$\begin{aligned}
\beta_z \mid \text{all others} &\sim N(0, \tau_z^2\sigma^2)\\
\boldsymbol{\beta} \mid \text{all others} &\sim N\Big((\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_{p\tau})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}},\\
&\qquad (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_{p\tau})^{-1}\sigma^2\Big)\\
1/\tau_g^2 \mid \text{all others} &\sim \text{Inv-Gausian}\ \Big(\lambda\sigma(p_g\boldsymbol{\beta}_g'\boldsymbol{\beta}_g)^{-1/2},\\
&\qquad \lambda^2\Big)\\
1/\tau_z^2 \mid \text{all others} &\sim \text{Inv-Gaussian}\ \left(\lambda\sigma/|\beta_z|, \lambda^2\right)\\
\sigma^2 \mid \text{all others} &\sim \text{Inv-Gamma}\ \Bigg(\frac{n+G}{2}, \frac{||\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}||^2}{2}\\
&\qquad + \frac{\beta_z^2}{2\tau_z^2} + \sum_{g=1}^{G}\frac{p_g\boldsymbol{\beta}_g'\boldsymbol{\beta}_g}{2\tau_g^2}\Bigg)
\end{aligned}$$

where $\boldsymbol{D}_{p\tau}$ is a diagonal matrix with each p_g/τ_g^2 repeating p_g times in order as the diagonal components,

$g = 1, \ldots, G$. The hyperparameter λ is estimated as in Section 2.4 with p replaced by G.

Let $b_g = \sqrt{\boldsymbol{\beta}_g' \boldsymbol{\beta}_g}$, which can be regarded as a measure of the effect of group g. The gth group effect is compared with the benchmark and ranked by d_g defined as (4) with β_i replaced by b_g. These posterior probabilities can be evaluated once the posterior samples for β_z and $\boldsymbol{\beta}_g$, $g = 1, \ldots, G$ are generated form Gibbs sampling. Based on these, the importance order of groups can be obtained. Like before, a BIC criterion specified in Equation (5) can be applied to eliminate groups of covariates that are unimportant.

In the procedure described above, groups are selected in an all-in-all-out fashion. However, not all of the covariates have influence on $\boldsymbol{y}$ within an selected group. Hence, it is desired to carry out variable level selection within chosen groups. Let $\mathcal{I}$ be the index set of groups selected in the group level selection and let $\boldsymbol{X}_{\mathcal{I}} = (\boldsymbol{X}_g, g \in \mathcal{I})$. We can apply the variable selection procedure described in Section 2.3 to the covariate vector $\boldsymbol{X}_{\mathcal{I}}$. Let $\boldsymbol{X}_{m^*}$ be the vector of finally selected covariates. It could happen that some groups in $\boldsymbol{X}_{\mathcal{I}}$ are entirely eliminated in the variable level selection, i.e., some $\boldsymbol{X}_g$'s with $g \in \mathcal{I}$ are entirely not in $\boldsymbol{X}_{m^*}$. These groups are then further excluded.

Even if there is no group structure in covariates, this group level selection followed by a variable level selection can be applied for variable selection when p is very large to reduce dimensionality in a fast way because group level selection may eliminate several groups of unimportant covariates simultaneously.

3. Simulation studies

Monte Carlo simulations are carried out to compare the performance of the proposed Bayesian variable selection method via a benchmark, as well as Bayesian lasso (B-lasso) by Park and Casella (2008) and frequentist lasso by Tibshirani (1996), where the penalty parameter is tuned by 10-folds cross-validation.

In the first study, there is no group structure in covariates. Three sets of n and p with increasing ratio of p/n are considered, $n = 50, p = 10$, $n = 50, p = 100$, and $n = 100, p = 500$. The matrix $\boldsymbol{X}$ is generated from multivariate normal distribution $N(\boldsymbol{0}, \Sigma)$, where the (i, j)th element of Σ is $.5^{|i-j|}$, $i, j = 1, \ldots, p$. Given $\boldsymbol{X}$, the response vector $\boldsymbol{y}$ is generated from $N(\boldsymbol{X}\boldsymbol{\beta}_0, \sigma_0^2 \boldsymbol{I})$, where $\boldsymbol{\beta}_0 = (1.5, 3, 0, 0, 2, 0, \ldots, 0)'$ is p-dimensional with only three non-zero components (the first, second, and fifth), and σ_0 is chosen so that $\|\boldsymbol{\beta}_0\|/\sigma_0$, the signal-to-noise ratio, is $3, 5, 10$ when $n = 50$ and $3, 4, 5, 6$ when $n = 100$. Note that the intercept β_0 is set to be 0. The covariates corresponding to non-zero $\boldsymbol{\beta}_0$ components are called important covariates; otherwise they are unimportant.

We consider the following performance measures of the proposed, lasso, and B-lasso methods:

$$\text{model size} = \text{number of selected covariates} \tag{8}$$

$$\text{sensitivity} = \frac{\text{number of of selected important covariates}}{3} \tag{9}$$

$$\text{specificity} = \frac{\text{number of removed unimportant covariates}}{p - 3} \tag{10}$$

$$\text{PMSE} = \frac{\|\boldsymbol{y}_{\text{test}} - \bar{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2}{n} \tag{11}$$

where PMSE is estimated prediction mean square error based on a test response vector on the test response vector $\boldsymbol{y}_{\text{test}}$ that is independent of $\boldsymbol{y}$ generated from $N(\boldsymbol{X}\boldsymbol{\beta}_0, \sigma_0^2 \boldsymbol{I})$ with the same $\boldsymbol{X}$, and $\widehat{\boldsymbol{\beta}}$ is the posterior mean under the selected model.

The averages of quantities in (8)–(11) over 1000 simulations are presented in Table 2, with simulation standard deviations given in parenthesis. In addition, the rate in 1000 simulations of selecting exactly three important covariates are also included in Table 2.

The results in Table 2 illustrate substantial advantages of the proposed variable selection over the other two methods, in terms of measures in (8)–(11) and the rate of selecting exactly the three important covariates. The lasso selects much more covariates than the proposed method in all cases without improving the prediction error. The B-lasso does not select covariates, has sensitivity 1 and specificity and rate 0 and does not perform well in prediction especially when p/n is large.

In the second simulation study, a group structure is added to covariates and the proposed method in Section 2.5 is considered with a group selection followed by an individual variable selection. For comparison, we include three existing methods, the group lasso (glasso) proposed by Yuan and Lin (2006), which carries out the group level selection in an 'all-in-all-out' fashion, the group bridge (gbridge) proposed by Huang et al. (2009) and Zhou and Zhu (2010), which selects groups as well as individual variables, and the sparse-group lasso (sglasso) proposed by Simon et al. (2012).

Similar to the first simulation study, we generate $\boldsymbol{X}$ from $N(\boldsymbol{0}, \Sigma)$ and given $\boldsymbol{X}$, we generate $\boldsymbol{y}$ from $N(\boldsymbol{X}\boldsymbol{\beta}_0, \sigma_0^2 \boldsymbol{I})$ with $\|\boldsymbol{\beta}_0\|/\sigma_0 = 3$. The group structure is from the covariance matrix Σ: components of $\boldsymbol{X}$ within the same group have pairwise correlation 0.5, while components of $\boldsymbol{X}$ from different groups are independent. Two cases with different sample size n, dimension p, and group structures are considered.

Case I. $n = 100$ and $p = 90$. There are six groups with group sizes 10, 20, 10, 20, 10, and 20, respectively. Each of groups 1, 3 and 5 contains two important covariates whose regression coefficients are

Table 2. Results for simulation Study 1.

n	p	$\frac{\|\beta_0\|}{\sigma_0}$	Method	Model size	Sensitivity	Specificity	PMSE	Rate
50	10	3	Proposed	2.973 (0.501)	0.955 (0.117)	0.984 (0.051)	8.022 (1.831)	0.784
			Lasso	6.032 (2.114)	0.998 (0.028)	0.566 (0.301)	8.306 (1.797)	0.104
			B-lasso	10.00 (0.000)	1.000 (0.000)	0.000 (0.000)	8.303 (1.771)	0.000
		5	Proposed	3.096 (0.356)	0.995 (0.041)	0.984 (0.048)	4.665 (0.966)	0.884
			Lasso	6.015 (2.090)	1.000 (0.011)	0.569 (0.299)	4.955 (1.044)	0.102
			B-lasso	10.00 (0.000)	1.000 (0.000)	0.000 (0.000)	4.947 (1.017)	0.000
		10	Proposed	3.123 (0.366)	1.000 (0.000)	0.982 (0.052)	2.341 (0.481)	0.890
			Lasso	5.939 (2.095)	1.000 (0.000)	0.580 (0.299)	2.502 (0.531)	0.110
			B-lasso	10.00 (0.000)	1.000 (0.000)	0.000 (0.000)	2.503 (0.515)	0.000
50	100	3	Proposed	2.891 (1.072)	0.819 (0.227)	0.996 (0.007)	9.379 (2.731)	0.327
			Lasso	17.73 (10.80)	0.990 (0.058)	0.848 (0.111)	9.758 (2.405)	0.004
			B-lasso	100.0 (0.000)	1.000 (0.000)	0.000 (0.000)	15.74 (2.935)	0.000
		5	Proposed	3.272 (0.950)	0.936 (0.142)	0.995 (0.008)	5.231 (1.665)	0.538
			Lasso	17.60 (10.50)	0.999 (0.021)	0.849 (0.108)	5.771 (1.440)	0.003
			B-lasso	100.0 (0.000)	1.000 (0.000)	0.000 (0.000)	11.86 (2.083)	0.000
		10	Proposed	3.313 (0.681)	0.996 (0.038)	0.997 (0.007)	2.374 (0.542)	0.761
			Lasso	18.58 (10.66)	1.000 (0.000)	0.839 (0.110)	2.925 (0.716)	0.002
			B-lasso	100.0 (0.000)	1.000 (0.000)	0.000 (0.000)	8.850 (1.322)	0.000
100	500	3	Proposed	3.150 (0.672)	0.970 (0.096)	1.000 (0.001)	7.588 (1.171)	0.750
			Lasso	39.08 (28.90)	1.000 (0.000)	0.928 (0.058)	9.343 (2.052)	0.000
			B-lasso	500.0 (0.000)	1.000 (0.000)	0.000 (0.000)	15.93 (1.957)	0.000
		4	Proposed	3.230 (0.601)	0.997 (0.033)	1.000 (0.001)	5.535 (0.763)	0.820
			Lasso	40.73 (28.65)	1.000 (0.000)	0.924 (0.058)	7.091 (1.580)	0.000
			B-lasso	500.0 (0.000)	1.000 (0.000)	0.000 (0.000)	13.68 (1.615)	0.000
		5	Proposed	3.180 (0.500)	1.000 (0.000)	1.000 (0.001)	4.416 (0.611)	0.860
			Lasso	42.16 (28.65)	1.000 (0.000)	0.921 (0.058)	5.693 (1.261)	0.000
			B-lasso	500.0 (0.000)	1.000 (0.000)	0.000 (0.000)	12.29 (1.400)	0.000
		6	Proposed	3.150 (0.479)	1.000 (0.000)	1.000 (0.001)	3.676 (0.507)	0.890
			Lasso	41.87 (28.24)	1.000 (0.000)	0.922 (0.057)	4.737 (1.065)	0.000
			B-lasso	500.0 (0.000)	1.000 (0.000)	0.000 (0.000)	11.34 (1.257)	0.000

Notes: Numbers in parentheses are standard deviations; the true model size is 3; model size, sensitivity, specificity, and PMSE are defined in (8)–(11); rate = rate of selecting exactly three important covariates.

Table 3. Results for simulation Study 2.

		Group selection				Variable selection				
Case	Method	Size	Sens	Spec	Rate	Size	Sens	Spec	Rate	PMSE
I	Proposed	3.14	1.00	0.95	0.89	6.29	1.00	1.00	0.79	3.57
		(0.46)	(0.00)	(0.15)		(0.63)	(0.00)	(0.01)		(0.52)
	glasso	3.85	1.00	0.72	0.38	46.93	1.00	0.51	0.00	5.42
		(0.89)	(0.00)	(0.30)		(17.90)	(0.00)	(0.21)		(0.89)
	gbridge	3.59	1.00	0.80	0.64	22.46	1.00	0.80	0.00	4.15
		(0.94)	(0.00)	(0.31)		(21.53)	(0.00)	(0.26)		(0.91)
	sglasso	5.50	1.00	0.17	0.04	14.66	1.00	0.90	0.01	4.56
		(0.78)	(0.00)	(0.26)		(4.97)	(0.01)	(0.06)		(0.76)
II	Proposed	1.96	0.96	1.00	0.89	3.06	0.70	1.00	0.24	13.95
		(0.32)	(0.14)	(0.03)		(1.10)	(0.24)	(0.01)		(4.49)
	glasso	3.39	1.00	0.83	0.04	33.9	1.00	0.69	0.00	17.15
		(0.64)	(0.00)	(0.08)		(6.39)	(0.00)	(0.07)		(3.84)
	gbridge	4.80	1.00	0.65	0.12	26.1	0.98	0.77	0.00	16.25
		(1.47)	(0.00)	(0.18)		(8.57)	(0.07)	(0.09)		(3.66)
	sglasso	3.89	1.00	0.76	0.35	8.14	0.90	0.95	0.05	14.13
		(2.18)	(0.00)	(0.27)		(4.77)	(0.15)	(0.05)		(3.16)

Notes: Numbers in parentheses are standard deviations; under variable selection, size = model size, sen = sensitivity, spec = specificity, defined in (8)–(10); under group selection, size, sen, spec are defined by (8)–(10) with covariates replaced by groups; PMSE is defined by (11); rate = rate of selecting exactly the true numbers of important groups and covariates.

1.5 and 2, and 8 unimportant covariates. Groups 2, 4 and 6 contain all unimportant covariates. Thus, there are three important groups and a total of six important covariates.

Case II. $n = 50$ and $p = 100$. There are 10 groups, each with 10 covariates. Each of groups 1 and 3 has two important covariates whose regression coefficients are 1.5 and 3, and 8 unimportant covariates. All other eight groups contain unimportant covariates. Thus, there are two important groups and a total of four important covariates.

The averages of quantities in (8)–(11) over 200 simulations are presented in Table 3 for both group and individual variable levels when (8)–(10) are considered, with simulation standard deviations given in parenthesis. The rate in 200 simulations of selecting exactly number of important groups and number of important individual covariates are also included in Table 3.

The results in Table 3 demonstrate the advantage of our method in both prediction and variable selection, compared to other three methods.

4. Real data examples

For illustration, in this section, we apply the proposed method to three real datasets and compare it with other methods.

4.1. Prostate cancer

This example is introduced in Section 2.2, with variable selection illustrated in Section 2.3. To check the performance of proposed variable selection and make comparisons, we randomly split the dataset with 97 patients into 2 subsets of sizes 78 and 19, use the subset of size 78 as the training set to carry out variable selection and build regression model, and use the subset of size 19 as the test set to validate the prediction performance in terms of PMSE defined by (11). We independently repeat random splitting 100 times and obtain the empirical results of 100 replications in Table 4.

The results in Table 4 elucidates that the proposed method outperforms lasso and B-lasso. First, the proposed selection method highly concentrates on selecting three important covariates as indicated in Figure 1 and Table 1. The average model size is 2.86. Although lasso agrees with the proposed method in selecting the three most important variables, it tends to select some redundant variables without improving PMSE, the prediction accuracy. Although Bayes lasso has a small PMSE, it does not perform variable selection.

4.2. CCT8 in a genome-wide association study

Research on linking genetic variations and phenotypic variations such as susceptibility to certain disorders is important in genomics as it helps to accelerate the understanding of genetic basis and may shed light on new medical treatments. We consider a high-dimensional dataset with $p > n$ from a genome-wide association study, the expression quantitative trait locus (eQTL) mapping. The performance of high-resolution eQTL mapping on nucleotide level is based on the measurements of genome-wide single nucleotide polymorphism (SNP). Here we consider the eQTL mapping for the gene CCT8 measured by microarray as the response from 90 individuals, 45 Han Chinese from Beijing, China, and 45 Japanese from Tokyo, Japan. The analysis is to detect which SNPs are associated with the CCT8 expression level, from a total of 200 SNPs after an initial screening of many SNPs.

Table 5. Results based on 100 random splits for the CCT8 example.

	Selection rate for five SNPs with the highest rates				
Method	rs965951	rs2245431	rs2832321	rs16983706	rs468619
Proposed	0.73	0.48	0.44	0.28	0.34
Lasso	0.97	0.94	0.82	0.75	0.84
B-lasso	1.00	1.00	1.00	1.00	1.00
	average model size			PMSE	
Proposed	3.394 (2.571)			0.035 (0.021)	
Lasso	58.98 (22.93)			0.034 (0.016)	
B-lasso	200.0 (0.000)			0.064 (0.037)	

Note: Numbers in parentheses are standard deviations.

Results based on 100 random splits of the dataset similar to those in the previous example (Table 4) are given in Table 5, with 80 people in the training set and 10 in the test set. In terms of the average model size and PMSE, the results in Table 5 exhibit quite similar yet more dramatic pattern compared with the results in Table 4 for the prostate cancer data. Our variable selection method significantly promotes model sparsity by selecting only around 3.4 variables on average, whereas the lasso method selects nearly 59 variables on the average. The PMSE under our approach is not jeopardised by the simplicity of model, as it is nearly the same as the PMSE for lasso. The Bayesian lasso results in the greatest PMSE, indicating that including all 200 predictors (compared with only 8 variables in the prostate cancer example) without variable selection leads to serious prediction errors when the number of unimportant variables is overwhelming in the model.

Over 100 random data splits, the top five most frequently selected SNPs by our approach are shown in Table 5. The highest selection frequency of SNP rs965951 suggests its relevance with the response CCT8, which is in accord with the results from some previous studies (Bradic et al., 2009; Deutsch et al., 2005; Fan et al., 2012). The second most frequently selected SNP rs2245431 was also selected by Bradic et al. (2009). All findings obtained by statistical

Table 4. Results based on 100 random splits for the prostate cancer example.

	Selection rate for each covariate							
Method	lcavol	lweight	svi	age	lbph	gleason	lcp	pgg45
Proposed	1.00	0.87	0.91	0.00	0.08	0.00	0.00	0.00
Lasso	1.00	0.99	1.00	0.51	0.88	0.46	0.23	0.75
B-lasso	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		Average model size				PMSE		
Proposed		2.860 (0.377)				0.594 (0.189)		
Lasso		5.820 (1.553)				0.600 (0.216)		
B-lasso		8.000 (0.000)				0.580 (0.188)		

Note: Numbers in parentheses are standard deviations.

methodologies are yet to be further validated by relevant genomical analysis.

4.3. ACS breast cancer patient OWB data

Breast cancer is a worldwide common cancer and remains the leading cause of mortality for women. With the continuously improved survival rate and prolonged life expectancy granted by advanced modern therapies, increasing efforts have been devoted to investigating the quality of life for breast cancer patients, as the quality of life plays an important role throughout the treatment and survivorship and, hence, the relevant studies may shed light on innovative intervention designs for disease control and quality of life improvement.

We consider a dataset from a large-scale breast cancer study conducted by American Cancer Society (ACS) at the School of Nursing in Indiana University. We focus on a subset of this study with 623 seniors who were 55–70 years old at diagnosis and were surveyed 3–8 years after completion of chemotherapy and surgery, with or without radiation therapy. The response of interest is overall well being (OWB), a measure captured by Campbell's index of quality of life, which is based on seven questionnaire items (Campbell et al., 2008). The objective of this study is to identify the psychological, social, and behaviour factors having important impacts on the well being of the survivors, and to establish the association between these factors and OWB.

The total 57 covariates under consideration include 3 demographic variables and 54 social or behaviour scores quantified by questionnaires which are well studied in literature (Frank-Stromborg & Olsen, 2003). The 54 social or behaviour variables are divided to 8 non-overlapping groups, which are personality, physical health, psychological health, spiritual health, active coping, passive coping, social support, and self efficacy. Each contains 4 to 12 individual covariates describing the same aspect of the social or behaviour status from different perspectives. The three demographic variables are treated as three individual groups, which are age at diagnosis, years of education, and number of months the patients were in their initial breast cancer treatment.

As in the first two examples, we randomly split the data set to a training set of size 499 and a test set of size 124, and then show the average or frequency based on 100 splits in Table 6. Similar to the second simulation study, we compare our proposed method to the glasso, gbridge, and sglasso designed for group variable selection.

The first part of Table 6 shows the rate (over 100 random splits) of selecting groups, where the three individual demographic variables are treated as three groups with size 1. The psychological health group is always selected by every method, which strongly suggests its association with the response OWB. It makes intuitive sense as a diagnosis of breast cancer is the most devastating thing a woman can hear, and it is often accompanied with fear of death, loss of control, isolation, and depression (Knobf, 2007; Yoo et al., 2010), all of which make considerably negative impacts on OWB. The other group that is always selected by our method, glasso and sglasso is social support, which is characterised as combination of emotional, tangible, and informational support (Cohen et al., 2000), from any formal, informal, social, professional, structured or unstructured resources (House & Khan, 1985). Reviews on the relevant literature reveal that it has been long recognised that social support may affect the OWB of patients in chronic and life-threatening health conditions like breast cancer (Cohen & Syme, 1985). Besides the above two groups, our method also selects the spiritual health group at a relatively low frequency, while barely including any other remaining groups. Purnell and Andersen (2009) pointed out that spiritual well-being was significantly associated with quality of life and traumatic stress after controlling for disease and demographic variables. Furthermore, spirituality is regarded as a resource regularly used by patients with cancer coping with diagnosis and treatment Gall et al. (2005).

Our proposed method selects variables within each selected groups. Over 100 random data splits, the middle part of Table 6 shows the rates of top seven most frequently selected variables within the three selected groups. The selected psychological health group contains six variables, five of which are selected with high rates. In Table 6, tstatAnx and ttraiAnx are short for S-anxiety and T-anxiety scales, respectively, which are used to capture the anxiety level of patients based on 20 questions like 'I feel nervous and restless'; tbodimg stands for body image total score and is summarised from eight questions such as 'I am satisfied with the appearance of my body' and 'others find me attractive'; tcesd represents the total score for situations during the past week, and the questions associated with this construct are something like 'I was bothered by things that usually don't bother me' or 'my appetite was poor'. In the social support group, only one variable is selected with high frequency, tcommnow, which quantifies the communication quality between the patients and physicians, based on questions like 'I have a health care provider I trust' and 'I have a health care provider who knows me personally'. The high selection frequency of this variable is in accord with the existing research results, which suggests that although the older women obtain information regarding breast cancer from a variety of sources, they often reply heavily on their primary care physicians for support and information (Silliman et al., 1998).

While the previous analysis focuses on group and individual variable selection, the last part of Table 6 shows the advantage of our proposed method in terms of the averages of selected groups and variables, and the PMSE over 100 random splits. On the average, our method promotes model sparsity by picking only around 2.4 groups and further reduces model complexity by including less than six variables in selected groups. In contrary, both glasso and sglasso select nearly twice many groups while produce comparable PMSE. The gbridge also chooses significantly more groups than our method, while leads to a slight smaller.

Finally, as we discussed in Section 2.2, our proposed benchmark approach can also be applied by visualising the posteriors. Figure 2 illustrates how to visualise the

Table 6. Results based on 100 random splits for the OWB example.

Method	Individual demographic variable selection				
	Age	Treatment Months	Years of edu		
Proposed	0.00	0.02	0.00		
glasso	0.00	0.79	0.01		
gbridge	0.00	0.04	0.00		
sglasso	0.00	0.15	0.00		
	group selection				
	Personality size = 5	Physical health size = 4	Psych health size = 6	Spiritual health size = 12	
Proposed	0.00	0.00	1.00	0.21	
glasso	0.01	0.90	1.00	0.20	
gbridge	0.00	0.00	1.00	0.92	
sglasso	0.03	0.46	1.00	0.69	
	Active coping size = 7	Passive coping size = 5	Social support size = 4	Self-efficacy size = 11	
Proposed	0.04	0.05	1.00	0.05	
glasso	0.33	0.02	1.00	0.43	
gbridge	0.02	0.06	0.81	0.26	
sglasso	0.60	0.13	1.00	0.68	
	Selection within psychological health group				
	ttraiAnx	tstatAnx	tbodimg	tcesd	tthinkbc
Proposed	1.00	0.98	0.79	0.64	0.70
glasso	1.00	1.00	1.00	1.00	1.00
gbridge	1.00	1.00	1.00	1.00	1.00
sglasso	1.00	1.00	0.99	1.00	0.32
	Selection within social support tcommnow	Selection within spiritual health sprview6			
Proposed	1.00	0.20			
glasso	1.00	0.20			
gbridge	0.81	0.91			
sglasso	1.00	0.68			
	Average of selected groups	Average of selected individual variables	PMSE		
Proposed	2.370 (0.646)	5.500 (1.439)	2.789 (0.448)		
glasso	4.690 (0.837)	23.99 (6.889)	2.787 (0.425)		
gbridge	3.110 (0.601)	10.48 (2.172)	2.713 (0.452)		
sglasso	24.740 (2.008)	10.70 (4.150)	2.837 (0.424)		

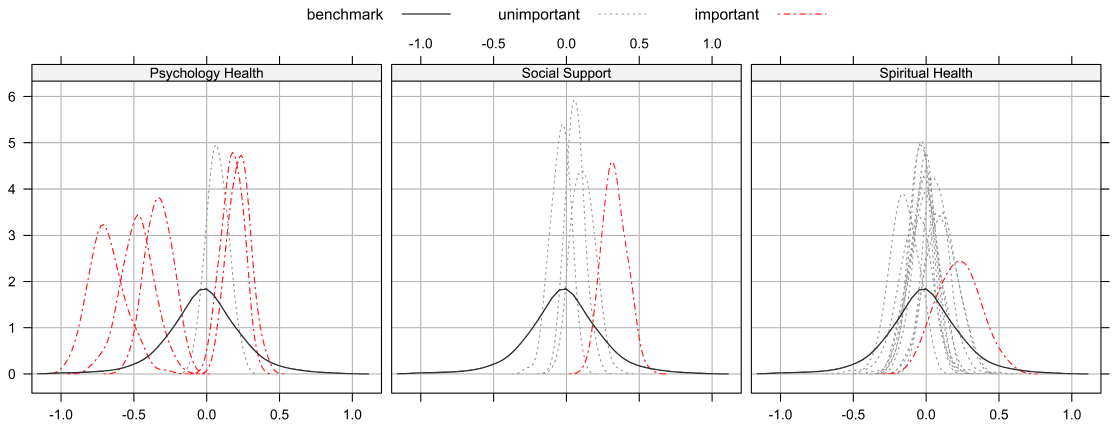

Figure 2. The posteriors of regression coefficients in three groups.

importance of variables within each three groups, based on the whole data set with 623 patients.

Acknowledgements

Our research was supported by the National Natural Science Foundation of China (11831008) and the U.S. National Science Foundation (DMS-1612873 and DMS-1914411). We would like to thank a referee for a careful review.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Our research was supported by the National Natural Science Foundation of China [grant number 11831008] and the U.S. National Science Foundation [grant numbers DMS-1612873 and DMS-1914411].

Notes on contributors

Dr. Jun Shao holds a PhD in statistics from the University of Wisconsin-Madison. He is a Professor of Statistics at the University of Wisconsin-Madison. His research interests include variable selection and inference with high dimensional data, sample surveys, and missing data problems.

Dr. Kam-Wah Tsui is an Emeritus Professor of Statistics at the University of Wisconsin–Madison. His research interests include Bayesian analysis, sample surveys, and general statistical methodology.

Dr. Sheng Zhang holds a Ph.D. in statistics from University of Wisconsin-Madison. She is now a data scientist at Google, Mountain View, California.

References

Andrews, D. F., & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B*, *36*, 99–102.

Bayarri, M. J., Berger, J. O., Jang, W., Ray, S., Pericchi, L. R., & Visser, L. (2019). Prior-based Bayesian information criterion (PBIC). *Statistical Theory and Related Fields*, *3*(1), 2–13. https://doi.org/10.1080/24754269.2019.1582126

Bradic, J., Fan, J., & Wang, W. (2009). Penalized composite quasi-Likelihood for ultrahigh-Dimensional variable selection. *Journal of the Royal Statistical Society, Series B*, *73*(3), 325–349. https://doi.org/10.1111/rssb.2011.73.issue-3

Brown, P. J., Vannucci, M., & Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B*, *60*(3), 627–641. https://doi.org/10.1111/rssb.1998.60.issue-3

Campbell, A., Converse, P., & Rodgers, W. (2008). *The quality of American life: Perceptions, evaluations, and satisfactions*. Russell Sage Foundation.

Chen, J., & Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model spaces. *Biometrika*, *95*, 759–771. https://doi.org/10.1093/biomet/asn034

Cohen, S., Gottlieb, B., & Underwood, L. (2000). Social relationships and health. In S. Cohen, L. Underwood, & B. Gottlieb (Eds.), *Social support measurement and intervention*. Oxford University Press.

Cohen, S., & Syme, L. (1985). *Social support and health* (Tech. Rep.). Academic.

Dellaportas, P., Forster, J. J., & Ntzoufras, I. (1997). *On Bayesian model and variable selection using MCMC* (Tech. Rep.). Department of Statistics, Athens University of Economics and Business.

Deutsch, S., Lyle, R., Dermitzakis, E., Attar, H., Subrahmanyan, L., Gehri, C., Parand, L., Gagnebin, M., Rougemont, J., Jongeneel, C., & Antonarakis, S. (2005). Gene expression variation and expression quantitative trait mapping of human chromosome 21 genes. *Human Molecular Genetics*, *14*(23), 3741–3749. https://doi.org/10.1093/hmg/ddi404

Dicker, L., Huang, B., & Lin, X. (2011). Variable selection and estimation with the seamless-l0 penalty. *Statistica Sinica*,

Fan, J., Han, X., & Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association*, *107*(499), 1019–1035. https://doi.org/10.1080/01621459.2012.720478

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Frank-Stromborg, M., & Olsen, S. (2003). *Instruments For Clinical Health-Care Research (Jones and Bartlett Series in Oncology, 3rd edition)* (Tech. Rep.). Jones & Bartlett Learning.

Gall, T., Charbonneau, C., Clarke, N., Grant, K., Joseph, A., & Shouldice, L. (2005). Understanding the nature and role of spirituality in relation to coping and health: A conceptual framework. *Canadian Psychology*, *46*(2), 88–104. https://doi.org/10.1037/h0087008

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *85*, 398–409.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732. https://doi.org/10.1093/biomet/82.4.711

Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hoti, F., & Sillanpää, M. J. (2006). Bayesian mapping of genotype x expression interactions in quantitative and qualitative traits. *Heredity*, *97*(1), 4–18. https://doi.org/10.1038/sj.hdy.6800817

House, J., & Khan, R. (1985). Measures and concepts of social support. In S. Cohen & S. L. Syme (Eds.), *Social support and health* (Tech. Rep.).

Huang, J., Ma, S., Xie, H., & Zhang, C. (2009). A group bridge approach for variable selection. *Biometrika*, *96*, 339–355. https://doi.org/10.1093/biomet/asp020

Knobf, M. (2007). Psychological responses in brest cancer survivors. *Seminars in Oncology Nursing*, *23*(1), 71–83. https://doi.org/10.1016/j.soncn.2006.11.009

Kuo, L., & Mallick, B. (1998). Variable selection for regression models. *Sankhya Series B*, *60*, 65–81.

Kyung, M., Gilly, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, *5*(2), 369–411. https://doi.org/10.1214/10-BA607

Lin, Y., & Zhang, H. (2006). Component selection and smoothing in smoothing spline analysis of variance models. *The Annals of Statistics*, *34*(5), 2272–2297. https://doi.org/10.1214/009053606000000722

Lv, J., & Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, *37*(6A), 3498–3528. https://doi.org/10.1214/09-AOS683

O'Hara, R. B., & Sillanpää, M. J. (2009). Review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, *4*(1), 85–117. https://doi.org/10.1214/09-BA403

Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Purnell, J., & Andersen, B. (2009). Religious practice and spirituality in the psychological adjustment of survivors of breast cancer. *Counseling and Values*, *53*(3), 165–182. https://doi.org/10.1002/(ISSN)2161-007X

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shen, X., Pan, W., & Zhu, Y. (2011). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*.

Silliman, R., Dukes, K., Sullivan, L., & Kaplan, S. (1998). Breast cancer care in older women: Sources of information, social support, and emotional health outcomes. *Cancer*, *83*, 706–711. https://doi.org/10.1002/(ISSN)1097-0142

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2012). A sparse-group lasso. *Journal of Computational and Graphical Statistics*.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., & Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate.II. radical prostatectomy treated patients. *Journal of Urology*, *141*(5), 1076–1083. https://doi.org/10.1016/S0022-5347(17)41175-X

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91–108. https://doi.org/10.1111/rssb.2005.67.issue-1

Tipping, M. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal OfMachine Learning*, *1*, 211–244.

Yoo, G., Levine, E., Aviv, C., Ewing, C., & Au, A. (2010). Older women, breast cancer, and social support. *Supportive Care in Cancer*, *18*, 121521–1530. https://doi.org/10.1007/s00520-009-0774-4

Yuan, M., & Lin, Y. (2005). Efficient empirical bayes variable selection and esimation in linear models. *Journal of the American Statistical Association*, *100*(472), 1215–1225. https://doi.org/10.1198/016214505000000367

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67. https://doi.org/10.1111/rssb.2006.68.issue-1

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942. https://doi.org/10.1214/09-AOS729

Zhou, N., & Zhu, J. (2010). Group variable selection via a hierarchical lasso and its oracle property. *Statistics and Its Inference*, *3*, 557–574.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, *67*(2), 301–320. https://doi.org/10.1111/rssb.2005.67.issue-2