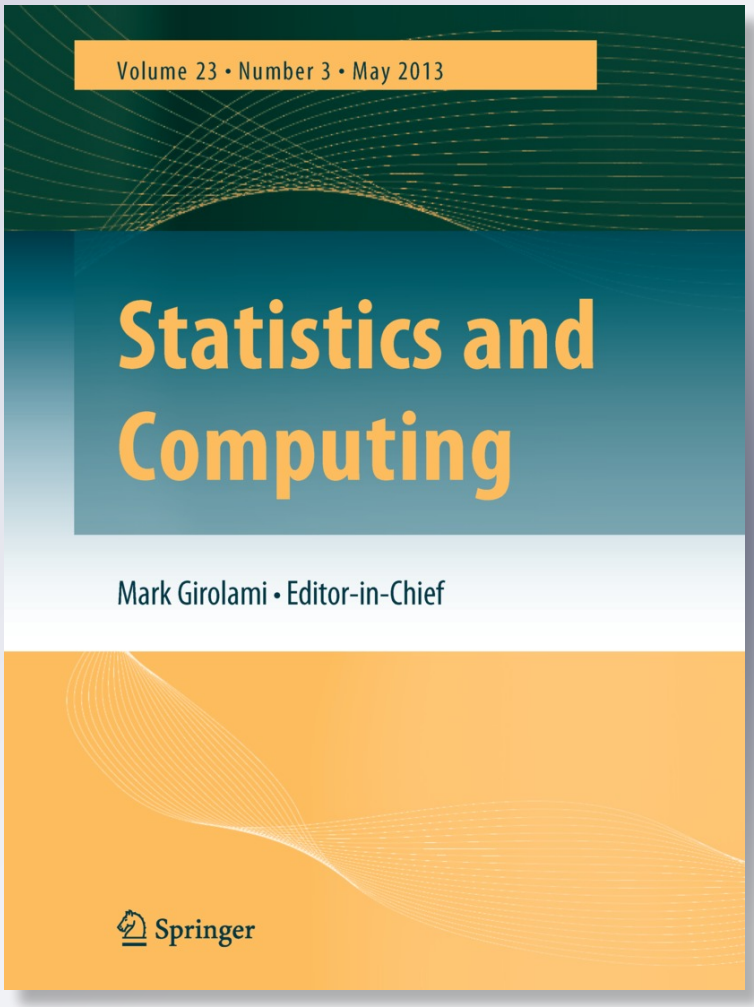
On Bayesian lasso variable selection and the specification of the shrinkage parameter

Anastasia Lykou & Ioannis Ntzoufras

Statistics and Computing

ISSN 0960-3174
Volume 23
Number 3

Stat Comput (2013) 23:361-390
DOI 10.1007/s11222-012-9316-x



Springer

Stat Comput (2013) 23:361–390
DOI 10.1007/s11222-012-9316-x

On Bayesian lasso variable selection and the specification of the shrinkage parameter

Anastasia Lykou · Ioannis Ntzoufras

Received: 16 March 2011 / Accepted: 30 January 2012 / Published online: 14 February 2012

Abstract We propose a Bayesian implementation of the lasso regression that accomplishes both shrinkage and variable selection. We focus on the appropriate specification for the shrinkage parameter λ through Bayes factors that evaluate the inclusion of each covariate in the model formulation. We associate this parameter with the values of Pearson and partial correlation at the limits between significance and insignificance as defined by Bayes factors. In this way, a meaningful interpretation of λ is achieved that leads to a simple specification of this parameter. Moreover, we use these values to specify the parameters of a gamma hyperprior for λ. The parameters of the hyperprior are elicited such that appropriate levels of practical significance of the Pearson correlation are achieved and, at the same time, the prior support of λ values that activate the Lindley-Bartlett paradox or lead to over-shrinkage of model coefficients is avoided. The proposed method is illustrated using two simulation studies and a real dataset. For the first simulation study, results for different prior values of λ are presented as well as a detailed robustness analysis concerning the parameters of the hyperprior of λ. In all examples, detailed comparisons with a variety of ordinary and Bayesian lasso methods are presented.

Keywords Bayes factors · MCMC · Gamma hyperprior for λ · Partial correlation · Pearson correlation · Shrinkage · Benchmark and threshold correlations

A. Lykou
Department of Mathematics and Statistics, Lancaster University, Lancaster, UK
e-mail: a.lykou@lancaster.ac.uk

I. Ntzoufras (✉)
Department of Statistics, Athens University of Economics and Business, Athens, Greece
e-mail: ntzoufras@aueb.gr

1 Introduction

Least absolute shrinkage and selection operator or lasso for short (Tibshirani 1996) is a shrinkage method that was originally used for the selection of variables in the linear regression problem. Its use was also extended to other problems such as multivariate models, generalized linear models (Meier et al. 2008) and survival methods (Tibshirani 1997; Johnson 2009). It imposes the L_1 norm on the least squares problem and shrinks the coefficients towards zero. The lasso estimates are given by

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{\text{lasso}} &= \underset{\beta}{\operatorname{argmin}}\left\{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})+\lambda\sum_{j=1}^{p}|\beta_j|\right\}\\ &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{Y}-\lambda\boldsymbol{s}_{\widehat{\boldsymbol{\beta}}}\right),\end{aligned} \tag{1}$$

for the usual regression model

$$\mathsf{Y}\sim \mathrm{N}_n\left(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{I}_n\sigma^2\right),$$

where $\mathrm{N}_n(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the n-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$. Moreover, we denote by $\boldsymbol{X}$ the $n\times p$ design or data matrix with elements X_{ij} corresponding to i individual and j covariate, by Y the $n\times 1$ vector of random responses and by $\boldsymbol{y}$ the corresponding observed (response) values. Finally, the coefficients β_j of each covariate X_j form the parameter vector $\boldsymbol{\beta}$; σ^2 stands for the error variance of the regression model; λ denotes the usual shrinkage parameter of lasso; $\boldsymbol{s}_{\boldsymbol{\beta}}$ is a vector with elements the sign of each $\widehat{\beta}_j^{\text{lasso}}$ and $\boldsymbol{I}_n$ is the $n\times n$ identity matrix.

The shrinkage parameter λ controls the amount of shrinkage imposed on the coefficients, where some weak effects are forced to be exactly zero if the shrinkage level is large enough. This shrinkage property makes lasso popular as a

variable selection method since there is no need to search the model space but only to fit the full model. Moreover, it is more stable than the stepwise subset selection methods and is computationally feasible for high-dimensional data under appropriate conditions (Osborne et al. 2000; Efron et al. 2004; Zhang and Huang 2008). These advantages have stimulated many researchers to propose extensions and improvements of the method; see, for example, Tibshirani (1997), Zou and Hastie (2005), Park and Hastie (2006), Zou (2006), Meier et al. (2008), Johnson (2009), and Lykou and Whittaker (2010).

1.1 Background of Bayesian lasso

Lasso has also a straightforward Bayesian interpretation since its estimates can be derived as the posterior mode when independent double exponential prior distributions are used for $\boldsymbol{\beta}$. The density of the double exponential (or Laplace) distribution for $\beta \sim \mathrm{DE}(\mu, b)$ is given by

$$f(\beta|\mu, b) = \frac{1}{2b}\exp\left(-\frac{|\beta-\mu|}{b}\right)$$

with mean μ and variance $2b^2$. Thus, a prior $\beta_j \sim \mathrm{DE}(0, \sigma^2/\lambda)$ will produce a posterior distribution that will be maximized under the lasso estimates (1). Although the posterior mode of this Bayesian model is the same as in ordinary lasso (inheriting all its properties), the posterior means and medians, which are usually in the centre of Bayesian inference, do not have the attractive property of setting non-important coefficients equal to zero. Another approach can be based on posterior credible intervals identifying non-important covariates when zero lies within these intervals; see for example in Fahrmeir et al. (2010). Nevertheless, this simple technique depends both on the selection of the posterior probability attached to such intervals and the way that they are constructed (since they are not unique). Moreover, this approach does not take into account model uncertainty and it does not quantify the importance of each covariate.

Due to the above Bayesian interpretation of ordinary lasso estimates, a wide variety of Bayesian lasso methods has been developed and published over the past years. Yuan and Lin (2005) incorporate a prior distribution of a mixture of a mass at zero and of the double exponential distribution into a linear model and they prove that the model with the highest posterior probability is the lasso solution. The choice of the shrinkage parameter is achieved through the empirical Bayes criterion CML. Park and Casella (2008) illustrate the Bayesian lasso regression by adopting the double exponential prior as a mixture of normal and exponential prior. However, this approach does not directly implement covariate selection but performs only shrinkage of the regression coefficients towards zero. They also propose a hierarchical model where a gamma distribution is imposed on the shrinkage parameter. Balakrishnan and Madigan (2010) combine the sparse Bayesian learning and the Bayesian lasso, by proposing the demi-Bayesian lasso, where a mixture of normal-exponential prior is imposed and the mixing parameter is estimated by maximizing the marginal data likelihood. Zero values in the mixing parameter denote which variables are excluded from the model, whereas, the shrinkage parameter is specified through cross-validation methods.

Hans (2009) imposes directly the double exponential prior on the lasso regression coefficients and a gamma prior on the shrinkage parameter and focuses on the problem of predicting future observations. Model uncertainty is addressed in Hans (2010) by computing exactly the marginal posterior probabilities for small model spaces. He handles the cases of large model spaces by imposing a mixture of a mass at zero and of a double exponential prior and estimates the posterior inclusion probabilities by using a Gibbs sampler.

Griffin and Brown (2010) discuss the normal-gamma prior, which is a generalization of the Bayesian lasso and has adaptive properties in terms of the shrinkage imposed on the coefficients. They also suggest a data-dependent prior for the shrinkage parameter. The Bayesian version of the Elastic net (Zou and Hastie 2005) has been introduced by Li and Lin (2010), where the prior information is a compromise between Normal and double exponential priors and the penalty parameters are chosen through an empirical method that maximizes the data marginal likelihood.

Fahrmeir et al. (2010) propose shrinkage, smoothing and selection priors for additive regression models, where the Bayesian lasso is included as a special case. This prior structure takes its name from the mixture of Normal and inverse gamma distribution (NMIG), while it imposes a spike and slab prior on the variance of the coefficients. Finally, in the last years, pure Bayesian shrinkage methods have also received attention in the statistical community resulting in the introduction of other prior distributions such as the horseshoe prior (Carvalho et al. 2010) and the double generalized Pareto (Armagan et al. 2012). All these approaches try to over-shrink small coefficients and leave as unaffected as possible large ones retaining also some consistency properties.

1.2 Merits and defects of Bayesian lasso

The main advantage of all shrinkage methods is the fact that they can be directly implemented in the full model and no model search is needed. The coefficients of covariates with weak effect on the response are immediately set equal to zero and, therefore, they are eliminated from the model structure.

Lasso is clearly better than ridge regression in terms of shrinkage since small coefficients are shrunk towards zero

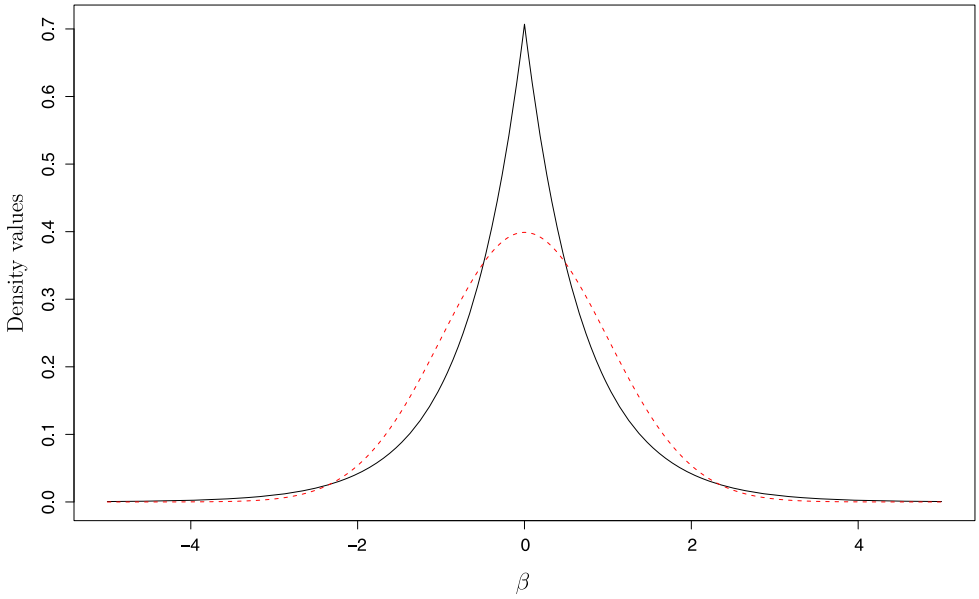


Fig. 1 Plot of the density functions for the double exponential (*solid line*) and normal (*dashed line*) distributions with zero mean and variance equal to one

faster, while less shrinkage is applied to large coefficients. This is due to the diamond-shaped restriction area that lasso implements on $(\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta})$ when it is written as a problem of constrained maximization in contrast to the corresponding n-dimensional sphere restriction area implemented by ridge regression.

From the Bayesian point of view, the double exponential prior has a considerably higher spike at zero giving higher probability to values in neighbourhoods close to zero. For example, for a double exponential centred at zero with variance equal to one, the probability of $\beta \in (-0.5, 0.5)$ is 0.507 for the double exponential and 0.383 for the normal model; see Fig. 1 for a graphical comparison of the corresponding density plots. Similarly, the distribution has slightly thicker tails, and for this reason, it has less shrinking effect on large coefficients, relatively speaking.

On the other hand, there are some disadvantages or problems that need further consideration when using lasso. First of all, lasso is a fast efficient method for selecting a single model but it does not allow to estimate model uncertainty which is important within Bayesian framework especially if prediction is the main aim.

Another problem is the selection of the shrinkage parameter λ. This actually controls the whole procedure. If we select a small value for λ, then no shrinkage (and therefore selection) will be performed, while if this value is too high, then all coefficients will be shrunk to zero. The regularization plot, which depicts the estimated lasso coefficients for different levels of the shrinkage parameter, provides valuable information about the order of decay of each coefficient and hence, the order of importance of each covariate but still it does not solve the problem. Cross-validation is the only methodology that provides a reliable and sensible way to specify this parameter. Nevertheless, here, we treat λ in a purely Bayesian manner specified a priori and without any use of data. Therefore, we choose the value of this prior parameter by looking the behavior of Bayes factors and by trying to avoid large and small shrinkage values that may overshrink the coefficients or activate the Lindley-Bartlett paradox (Lindley 1957; Bartlett 1957) respectively. The latter refers to the well-known behavior of the Bayes factors which support the most parsimonious model (from the ones under comparison) for suitably large sample size or suitably large prior variances of any additional parameters. The effect of this "paradox" is more apparent in nested model comparisons.

Within the Bayesian framework, usually interest lies in the whole posterior distribution and the posterior means and medians are often used as point estimates instead of the posterior modes. These estimates will approach zero slowly but they will never be exactly equal to zero as the posterior modes and the lasso estimates. Therefore, some of the properties of the original lasso are diminished when using this approach. Moreover, using the double exponential prior instead of the conjugate normal–inverse-gamma makes the evaluation of the posterior distribution less straightforward requiring the use of MCMC methods.

Finally, the double exponential prior in lasso formula a priori assumes independence of β_js. Therefore it does not account for the structure of the covariates as for example in Zellner's (1986) g-prior, where coefficients a priori assumed to be normally distributed with prior variance matrix equal to $g(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2$, in order to have similar structure with the OLS estimates.

1.3 Paper structure and contribution

In this paper, we combine the properties of the lasso through the use of the double exponential prior distribution with the advantages of usual variable selection techniques within the Bayesian framework. For this reason, we utilize the binary variable inclusion indicators introduced by George and McCulloch (1993) and widely used thereafter such as in Kuo and Mallick (1998) and Dellaportas et al. (2002). We focus on the case where the number of predictors is smaller than the number of observations. We use MCMC methods to estimate the posterior parameter estimates, the posterior model probabilities as well as the posterior variable inclusion probabilities. We can now additionally have meaningful regularization plots based on posterior variable inclusion probabilities and model-averaged medians of the regression coefficients.

We then focus on the specification of the shrinkage parameter using its effect on posterior model probabilities and Bayes factors. By investigating the behavior and the sensitivity of these measures on the choice of λ we obtain a simple and meaningful interpretation of its effect. In this way, we can a priori specify the shrinkage level and control the variable selection procedure at the same time. The proposed methodology can be described by the following steps:

1. Firstly, we trace a range of correlation values between the covariates and the response, for which the covariates will never be a posteriori supported whatever the value of the shrinkage parameter is. Therefore, covariates with such Pearson correlation measures will have Bayes factors measuring the evidence in favour of the addition of this covariate to the constant model lower than one for all values of λ. Following the same logic, we additionally identify the range of correlation values where the inclusion of this covariate is never a posteriori supported highly enough, i.e. the corresponding Bayes factor never becomes higher than a specified level α for all λ.
2. Moving further, we may specify λ either as constant quantity or as random variable (using hyperpriors). In the following, we propose methodology for the specification of the shrinkage parameter for each of these two choices.
   (a) *Treating λ as constant*: We specify λ by defining the levels of practical significance (i.e. when Bayes factors are equal to one) based on the correlation measures produced in step 1.
   (b) *Treating λ as random*: We utilize the produced range of correlations by specifying a sensible hyperprior for λ. With the specification of the hyperprior using values from step 1 we avoid using non-informative prior distributions for the shrinkage parameter, which unnecessarily support small or large values of λ activating the Lindley-Bartlett paradox or overshrinking model coefficients respectively.

The article is organized as follows. In Sect. 2 we introduce the model structure for the Bayesian lasso variable selection framework. Then, a simple Gibbs sampler scheme is described for the estimation of the posterior parameters, posterior variable inclusion probabilities and posterior model probabilities. The section closes with a short illustration and a discussion about new regularization plots based on the posterior medians of model averaged regression coefficients and posterior variable inclusion probabilities obtained by the Bayesian lasso variable selection. Section 3 provides an in-depth analysis about the relationship of the shrinkage parameter, the Bayes factors and their association with the Pearson and partial correlation coefficients. In particular, Sect. 3.1 focuses on the unicovariate Bayes factor, comparing each simple regression model with the null and its relation with the Pearson correlation coefficient. We examine and interpret this association using graphical representations. In Sect. 3.2 we provide arguments based on practical values of significance for the Pearson correlation for the specification of the shrinkage level λ. Section 3 concludes with an analysis about the effect of λ on the partial correlations and the Bayes factors of nested multiple lasso regression models. Section 4 proceeds, providing final recommendations about the prior value of λ and specifying a meaningful gamma hyperprior that avoids λ values which activate the Lindley-Bartlett paradox or overshrink model coefficients. In Sect. 5 we illustrate our method using two simulation studies and a real dataset. For the first simulation study we present results for λ values based on different threshold values and a detailed robustness analysis concerning the parameters of the gamma hyperprior of λ. For all illustrations, a detailed comparison between our proposed method and a variety of ordinary and Bayesian lasso methods is presented. The paper closes with a small discussion about open problems and further research on the topic.

2 Bayesian variable selection and lasso

2.1 Model structure

To set-up the Bayesian lasso variable selection we consider the usual likelihood of the normal model incorporating also the usual binary variable inclusion indicators $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ as in Kuo and Mallick (1998) and Dellaportas et al. (2002). We further assume a set of independent double exponential prior distributions for each model coefficient β_j, in order to implement a lasso type of shrinkage within each model. Hence, the model can be summarized by

the following expressions

$$\begin{aligned}
&\mathrm{Y}|\boldsymbol{\beta},\tau,\boldsymbol{\gamma} \sim \mathrm{N}_n\big(\mathbf{X}\mathbf{D}_{\boldsymbol{\gamma}}\boldsymbol{\beta},\tau^{-1}I_n\big),\\
&\quad \text{where } \mathbf{D}_{\boldsymbol{\gamma}} = \mathrm{diag}(\gamma_1,\ldots,\gamma_p),\\
&\boldsymbol{\beta}_j|\tau \sim \mathrm{DE}\left(0,\frac{1}{\tau\lambda}\right), \quad \text{for } j=1,\ldots,p, \qquad (2)\\
&\gamma_j \sim \mathrm{Bernoulli}(\pi_j),\\
&\tau \sim \mathrm{Gamma}(c,d),
\end{aligned}$$

where $\tau = 1/\sigma^2$ is the precision of the Normal regression model, Bernoulli(π) is the Bernoulli distribution with success probability π and Gamma(c,d) is the gamma distribution with mean c/d and variance c/d^2.

Using the above model formulation, inference will be based on posterior model probabilities $f(\boldsymbol{\gamma}|\boldsymbol{y})$, for any model indicator $\boldsymbol{\gamma}$, and the corresponding posterior variable inclusion probabilities $f(\gamma_j = 1|\boldsymbol{y})$, for $j = 1,\ldots,p$, quantifying the posterior importance of each covariate. In our approach, the relative posterior model probabilities comparing two rival models $\boldsymbol{\gamma}_{(1)}$ and $\boldsymbol{\gamma}_{(2)}$ play a central role. More specifically, the posterior model odds of model $\boldsymbol{\gamma}_{(1)}$ versus model $\boldsymbol{\gamma}_{(2)}$ are given by

$$PO_{12} = \frac{f(\boldsymbol{\gamma}_{(1)}|\boldsymbol{y})}{f(\boldsymbol{\gamma}_{(2)}|\boldsymbol{y})} = \frac{f(\boldsymbol{y}|\boldsymbol{\gamma}_{(1)})}{f(\boldsymbol{y}|\boldsymbol{\gamma}_{(2)})} \times \frac{f(\boldsymbol{\gamma}_{(1)})}{f(\boldsymbol{\gamma}_{(2)})}.$$

When prior model probabilities are considered equal for all competing models, then pairwise comparisons are solely based on Bayes factors $\mathrm{BF}_{12} = \frac{f(\boldsymbol{y}|\boldsymbol{\gamma}_{(1)})}{f(\boldsymbol{y}|\boldsymbol{\gamma}_{(2)})}$. Kass and Raftery (1995) thoroughly present Bayes factors and their importance in Bayesian inference. Following the arguments of Jeffreys (1961), they suggest that the data provide substantial evidence in favour of model $\boldsymbol{\gamma}_{(1)}$ for BF > 3. The evidence becomes stronger for higher values of the Bayes factor (and posterior odds).

A prior specification similar to (2) was also used by Hans (2009). However, Hans (2009) did not consider the variable inclusion indicators in his approach since he did not address the variable selection problem in that publication.

The level of the posterior shrinkage towards zero for each β_j is controlled via λ since the prior distribution becomes more and more informative as λ increases. In the remaining of the paper we assume that both the covariates and the response are standardized and therefore the constant term in the linear model is eliminated throughout this paper.

2.2 A simple Gibbs sampler for Bayesian lasso variable selection

In this work, we use the Kuo and Mallick (1998) approach to estimate the posterior densities. However, any equivalent algorithm such as the GVS (Dellaportas et al. 2002) or the RJMCMC (Green 1995) will provide similar results. Thus, the conditional posterior distribution of β_j coincides with the prior distribution for $\gamma_j = 0$, while it is a mixture of truncated normal distributions when $\gamma_j = 1$, that is

$$\beta_j|\boldsymbol{y},\tau,\boldsymbol{\beta}_{\setminus j},\boldsymbol{\gamma}_{\setminus j},\gamma_j = 0 \sim \mathrm{DE}\left(0,\frac{1}{\tau\lambda}\right) \qquad (3)$$

$$\begin{aligned}
&\beta_j|\boldsymbol{y},\tau,\boldsymbol{\beta}_{\setminus j},\boldsymbol{\gamma}_{\setminus j},\gamma_j = 1,\\
&\Omega_j \sim \Omega_j\, TN\big(\mu_j^-,s_j^2,\beta_j<0\big) \qquad (4)\\
&\qquad + (1-\Omega_j)\, TN\big(\mu_j^+,s_j^2,\beta_j\geq 0\big),
\end{aligned}$$

where $\boldsymbol{\beta}_{\setminus j},\boldsymbol{\gamma}_{\setminus j}$ are vectors $\boldsymbol{\beta},\boldsymbol{\gamma}$ without β_j and γ_j respectively, $I(A)$ is the indicator function taking the value of one when A is true and zero otherwise, and $TN(\mu,\sigma^2,A)$ is the normal distribution truncated in the subset $A \subset \mathbb{R}$ with density function

$$f_{TN}\big(x;\mu,\sigma^2,A\big) = \frac{f_N(x;\mu,\sigma^2)}{\int_A f_N(x;\mu,\sigma^2)dx} I(x\in A)$$

with $f_N(x;\mu,\sigma^2)$ denoting the density of a normal distribution with mean μ and variance σ^2 evaluated at x. Hence, the densities of the truncated normal distributions appearing in (4) are given by

$$f_{TN}\big(\beta_j;\mu_j^-,s_j^2,\beta_j<0\big) = \frac{f_N(\beta_j;\mu_j^-,s_j^2)}{\Phi(-\mu_j^-/s_j)} I(\beta_j<0)$$

and

$$f_{TN}\big(\beta_j;\mu_j^+,s_j^2,\beta_j\geq 0\big) = \frac{f_N(\beta_j;\mu_j^+,s_j^2)}{\Phi(\mu_j^+/s_j)} I(\beta_j\geq 0),$$

respectively, with $\Phi(x)$ being the cdf of the standardized normal distribution. The means and variance of the truncated normal distributions are computed by the following expressions

$$\mu_j^- = \frac{h_j+\lambda}{\|\boldsymbol{X}_j\|^2}, \qquad \mu_j^+ = \frac{h_j-\lambda}{\|\boldsymbol{X}_j\|^2}, \quad h_j = \boldsymbol{X}_j^T(\boldsymbol{e}+\beta_j\boldsymbol{X}_j)$$

$$\text{and} \quad s_j^2 = \frac{1}{\tau\|\boldsymbol{X}_j\|^2}$$

with $\boldsymbol{X}_j$ denoting the jth column of matrix $\boldsymbol{X}$, $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ denoting the vector of residual values with elements $e_i = y_i - \sum_j X_{ij}\beta_j$ (for $i = 1,\ldots,n$), while $\|z\|^2 = \sum_{i=1}^n z_i^2$ for any vector z of length n.

Additionally, Ω_j is a binary parameter specifying the sign of β_j. The full conditional posterior probability of $\Omega_j = 1$ is given by

$$w_j = \mathrm{P}(\Omega_j = 1 \mid \boldsymbol{y}, \tau, \boldsymbol{\beta}_{\setminus j}, \boldsymbol{\gamma}_{\setminus j}, \gamma_j = 1)$$
$$= \mathrm{P}(\beta_j < 0 \mid \boldsymbol{y}, \tau, \boldsymbol{\beta}_{\setminus j}, \boldsymbol{\gamma}_{\setminus j}, \gamma_j = 1)$$
$$= \frac{\Phi(-\mu_j^-/s_j)/f_N(0;\mu_j^-,s_j^2)}{\Phi(-\mu_j^-/s_j)/f_N(0;\mu_j^-,s_j^2) + \Phi(\mu_j^+/s_j)/f_N(0;\mu_j^+,s_j^2)}. \quad (5)$$

See also Hans (2009) for the multivariate case when $\boldsymbol{\gamma}$ is fixed. Hence, when $\gamma_j = 1$

- Generate Ω_j from a Bernoulli with success probability w_j given by (5).
- Generate β_j from

$$\begin{cases} TN(\mu_j^-, s_j^2, \beta_j < 0) & \text{if } \Omega_j = 1, \\ TN(\mu_j^+, s_j^2, \beta_j \geq 0) & \text{if } \Omega_j = 0. \end{cases}$$

The full conditional posterior distributions for the remaining parameters are the following

$$\tau|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{y} \sim \mathrm{Gamma}\left(\frac{n}{2} + p + c, \frac{\|\boldsymbol{y} - \mathbf{X}\mathbf{D}_{\boldsymbol{\gamma}}\boldsymbol{\beta}\|^2}{2} + \lambda\|\boldsymbol{\beta}\| + d\right) \quad (6)$$

$$\gamma_j|\boldsymbol{\beta}, \tau, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y} \sim \mathrm{Bernoulli}\left(\frac{O_j}{1+O_j}\right) \quad (7)$$

with

$$O_j = \frac{f(\gamma_j = 1|\boldsymbol{\gamma}_{\setminus j}, \boldsymbol{\beta}, \tau^2, \boldsymbol{y})}{f(\gamma_j = 0|\boldsymbol{\gamma}_{\setminus j}, \boldsymbol{\beta}, \tau^2, \boldsymbol{y})} = \frac{f(\boldsymbol{y}|\boldsymbol{\beta}, \tau, \boldsymbol{\gamma}_{\setminus j}, \gamma_j = 1)}{f(\boldsymbol{y}|\boldsymbol{\beta}, \tau, \boldsymbol{\gamma}_{\setminus j}, \gamma_j = 0)} \frac{\pi(\boldsymbol{\gamma}_{\setminus j}, \boldsymbol{\gamma}_j = 1)}{\pi(\boldsymbol{\gamma}_{\setminus j}, \boldsymbol{\gamma}_j = 0)}. \quad (8)$$

2.3 Regularization plots for Bayesian lasso variable selection

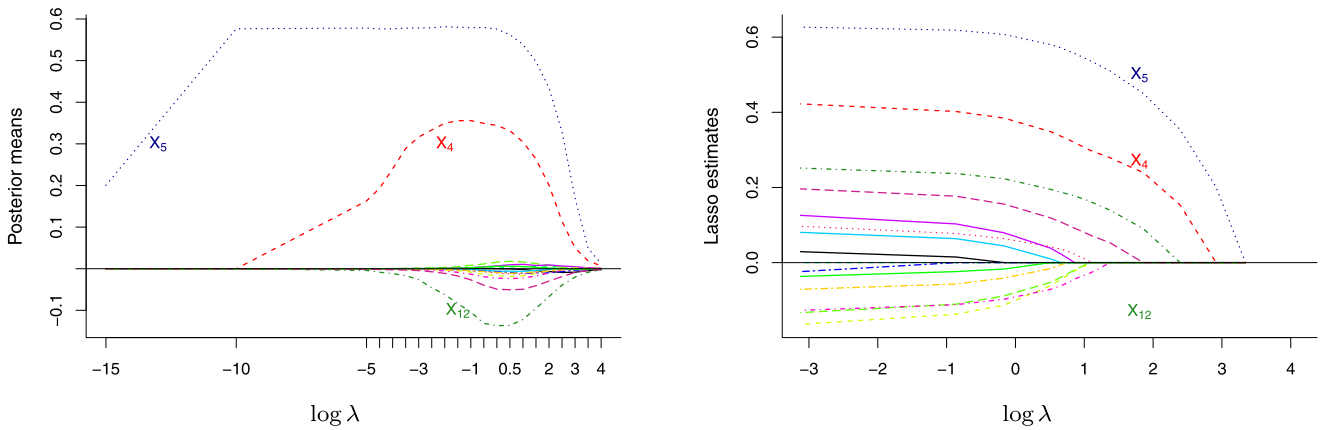
Using the Gibbs sampler described in Sect. 2.2, we obtain a posterior sample $(\boldsymbol{\beta}^{(t)}, \tau^{(t)}, \boldsymbol{\gamma}^{(t)})$ for $t = 1, 2, \ldots, T$. From this output we can estimate not only the posterior model probability $f(\boldsymbol{\gamma}|\boldsymbol{y})$ of each model $\boldsymbol{\gamma}$, but also the posterior inclusion probabilities $f(\gamma_j = 1|\boldsymbol{y})$ for each covariate X_j, as well as posterior summaries for the Bayesian model averaged (BMA) versions of the effect of each covariate X_j given by $\beta_j^* = \gamma_j\beta_j$. For the latter two quantities, we will examine their behavior using different levels of prior variances and, therefore, different levels of the shrinkage parameter λ. This sensitivity analysis is depicted using graphs equivalent to the regularization plots obtained in traditional lasso techniques.

Here, we illustrate these visual representations by considering the first simulated dataset of Dellaportas et al. (2002), which is available on the website of the book written by Ntzoufras (2009). This dataset consists of $n = 50$ observations and $p = 15$ covariates generated from a standardised normal distribution and the response from

$$\mathsf{Y}_i \sim \mathrm{N}\big(X_{i4} + X_{i5}, 2.5^2\big), \quad \text{for } i = 1, \ldots, 50.$$

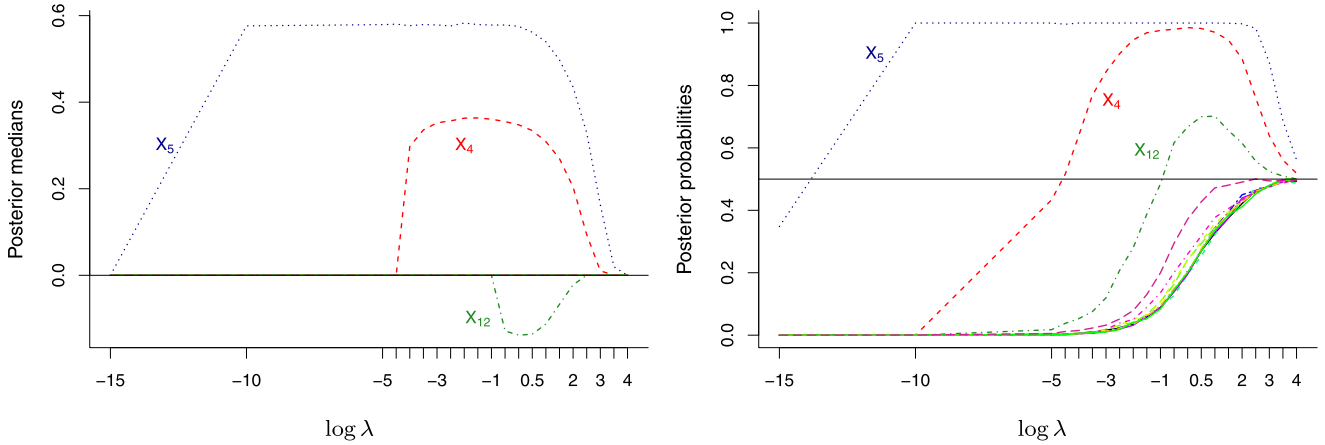
The proposed method is performed on this dataset for different values of λ, $\pi_j = 0.5$ for all j, $c = d = 10^{-4}$ and we consider 10,000 updates after discarding 1,000 additional iterations as burn-in period.

In Fig. 2(a), the posterior means of $\beta_j^* = \gamma_j\beta_j$ are plotted against the values of λ in log-scale while in Fig. 2(b) the usual regularization plot of the lasso estimates is depicted. The grid of the logarithm of lambda values has been chosen to be from -15 to -5 with an increment of 5 and from -5 to 4 with an increment of 0.5.



(a) Posterior means of BMA estimates (b) Lasso estimates

Fig. 2 (**a**) Regularization plots for the posterior means of $\beta_j^* = \gamma_j\beta_j$ and (**b**) usual lasso estimates against $\log\lambda$



(a) Posterior medians of BMA estimates (b) Posterior inclusion probabilities

Fig. 3 (**a**) Regularization plots for the posterior medians of $\beta_j^* = \gamma_j \beta_j$ and (**b**) the posterior variable inclusion probabilities $f(\gamma_j = 1|\boldsymbol{y})$ against $\log \lambda$

The lasso estimates, derived by the Lars algorithm, are controlled by λ which is the shrinkage applied on each coefficient β_j: for small values of λ no shrinkage is implemented, while as λ increases all coefficients are shrunk to zero with different rate, depending on their significance; see Fig. 2(b). Moreover, these estimates approach the ordinary least squares estimates as λ approaches zero (or $\log \lambda \to -\infty$). Although the behavior of the BMA posterior means of β_j^* is similar for large values of λ [see the right side of Fig. 2(a)] this is not the case for the remaining of the graph since for very low values of λ (expressing high prior ignorance) they shrunk back to zero instead of approaching the MLE estimates as in ordinary lasso. This is due to the Lindley-Bartlett paradox (Lindley 1957; Bartlett 1957), since small values of λ (corresponding to large prior variance of β_j) activate the effect of this paradox, leading to posterior model odds that fully support the most parsimonious model and therefore a posteriori restricting β_j^* to zero. As λ moves away from zero, the posterior means of the most important coefficients increase rapidly (in absolute value) until β_j^* is maximized. After this point, shrinkage is effective and, as expected, all coefficients gradually approach zero in a similar manner as in the original lasso. For moderate values of λ, the coefficients of the unimportant covariates have posterior means close to zero, which slowly decay to zero as λ becomes larger.

Nevertheless, the covariates that should be ultimately selected are highlighted in a more obvious way when plotting the posterior medians of β_j^*; see Fig. 3(a). Posterior medians become exactly equal to zero when $f(\gamma_j = 1|\boldsymbol{y}) < 0.5$ in contrast to the posterior means which will be small but non-zero unless $f(\gamma_j = 1|\boldsymbol{y}) = 0$. Hence, in the plot of the posterior medians, non-important covariates are eliminated from the plot for all values of λ.

The second plot of Fig. 3 (on the right) shows the posterior probabilities of $f(\gamma_j = 1|\boldsymbol{y})$ as a function of $\log \lambda$. As a result of the Lindley-Bartlett paradox, the posterior probabilities of including a variable in the model tends to zero for $\lambda \to 0$ (approximately $\log \lambda \to -15$). The posterior probabilities of the unimportant variables approach the value of 0.5 as λ moves away from zero. The behavior of the important covariates is different since they sharply increase as soon as λ moves away to zero. As λ increases, the prior variance becomes smaller and the posterior distributions of the coefficients are forced to be a posteriori close to zero. In such case, the data (in comparison to the prior) are not strong enough to provide evidence for the status of a covariate in the model formulation under consideration.

Even in these initial illustrations, the proposed method seems to offer a challenging approach for performing both shrinkage and variable selection. The regularization plot based on the medians of the BMA estimates is more efficient than the corresponding lasso plot since the effect of unimportant covariates is eliminated for all values of the shrinkage parameter λ. Moreover, the behavior of the posterior inclusion probabilities for large and small values of λ can motivate the restriction of the sensible values of λ to avoid overshrinkage (when λ is large) or the Lindley-Bartlett paradox (when λ is small). Using these observations as a starting point, in Sect. 3 we work on the choice of λ providing reasonable interpretation and insight about which values are sensible for the variable selection procedure.

3 Specification of the shrinkage parameter based on Bayes factors and practical significance values

3.1 Bayes factors for simple lasso regression and Pearson correlations

In this section, we focus on the Bayes factors comparing two simple models: the null (or constant) model m_0 versus a model m_j, which includes in the linear predictor only the covariate X_j. We will call these Bayes factors "unicovariate" and facilitate results based on these simple comparisons to identify reasonable values for the choice of λ.

Definition 1 (Unicovariate Bayes Factor $\mathrm{BF}_j^{\mathrm{un}}$) The unicovariate Bayes factor $\mathrm{BF}_j^{\mathrm{un}}$ for covariate X_j is defined as the Bayes factor that evaluates the evidence of model m_j versus m_0 with

$$\mathsf{Y}|\boldsymbol{\beta},\tau,m_j \sim \mathrm{N}_n\big(\mathbf{X}_j\beta_j,\tau^{-1}I_n\big)$$
$$\text{and}\quad \mathsf{Y}|\boldsymbol{\beta},\tau,m_0 \sim \mathrm{N}_n\big(\mathbf{0},\tau^{-1}I_n\big).$$

Under the prior setup described in the general model formulation (2), the Bayes factor of model m_j against m_0 is given by

$$\begin{aligned}\mathrm{BF}_j^{\mathrm{un}} &= \frac{f(\boldsymbol{y}|m_j)}{f(\boldsymbol{y}|m_0)}\\ &= \lambda\sqrt{\frac{\pi}{\|\boldsymbol{X}_j\|^2}}\frac{\Gamma(\frac{df}{2})}{\Gamma(\frac{df-1}{2})}\\ &\quad\times\frac{C_{j-}^{-\frac{df}{2}}P(\beta_{j-}<0)+C_{j+}^{-\frac{df}{2}}P(\beta_{j+}>0)}{(\|\boldsymbol{y}\|^2+2d)^{-(\frac{n}{2}+a)}},\end{aligned}\tag{9}$$

where

$$C_{j-}=\big(C_j-M_{j-}^2\big)\|\boldsymbol{X}_j\|^2,$$
$$C_{j+}=\big(C_j-M_{j+}^2\big)\|\boldsymbol{X}_j\|^2,$$
$$C_j=\frac{\|\boldsymbol{y}\|^2+2d}{\|\boldsymbol{X}_j\|^2},$$
$$\beta_{j-}\sim t_{df}\left(M_{j-},\frac{C_{j-}}{\|\boldsymbol{X}_j\|^2 df}\right),\qquad M_{j-}=\frac{\boldsymbol{y}^T\boldsymbol{X}_j+\lambda}{\|\boldsymbol{X}_j\|^2},$$
$$df=n+2a+1,$$
$$\beta_{j+}\sim t_{df}\left(M_{j+},\frac{C_{j+}}{\|\boldsymbol{X}_j\|^2 df}\right),\qquad M_{j+}=\frac{\boldsymbol{y}^T\boldsymbol{X}_j-\lambda}{\|\boldsymbol{X}_j\|^2},$$

where $T\sim t_\nu(\mu,\sigma^2)$ is a random variable such that $(T-\mu)/\sigma$ follows the Student's t distribution with ν degrees of freedom.

Assuming that all data are standardized and $d\to 0$, a simplified version of $\mathrm{BF}_j^{\mathrm{un}}$ can be expressed in terms of the shrinkage parameter λ and ρ_j, i.e. the sample estimate of the Pearson correlation coefficient between Y and the candidate predictor X_j:

$$\begin{aligned}\mathrm{BF}_j^{\mathrm{un}} &= \frac{\lambda}{n-1}\sqrt{\pi}\frac{\Gamma(\frac{df}{2})}{\Gamma(\frac{df-1}{2})}\left\{\left(1+\frac{t_{j-}^2}{df}\right)^{\frac{df}{2}}F_{t_{df}}(t_{j-})\right.\\ &\quad\left.+\left(1+\frac{t_{j+}^2}{df}\right)^{\frac{df}{2}}F_{t_{df}}(t_{j+})\right\}\\ &= \frac{\lambda}{n-1}\frac{df-1}{2\sqrt{df}}\left\{\left(1+\frac{t_{j-}^2}{df}\right)^{-\frac{1}{2}}\frac{F_{t_{df}}(t_{j-})}{f_{t_{df}}(t_{j-})}\right.\\ &\quad\left.+\left(1+\frac{t_{j+}^2}{df}\right)^{-\frac{1}{2}}\frac{F_{t_{df}}(t_{j+})}{f_{t_{df}}(t_{j+})}\right\}\end{aligned}\tag{10}$$

where F_{t_ν}, $f_{t_{df}}$ is the cdf and the density function of a Student's t random variable with ν degrees of freedom and

$$t_{j-}=-\frac{M_{j-}\sqrt{df}}{\sqrt{1-M_{j-}^2}},\qquad t_{j+}=\frac{M_{j+}\sqrt{df}}{\sqrt{1-M_{j+}^2}},$$
$$M_{j-}=\rho_j+\frac{\lambda}{n-1},\qquad M_{j+}=\rho_j-\frac{\lambda}{n-1}.$$

In the above computations, without loss of generality, we assume that ρ_j is positive. Moreover, the Bayes factor in (10) is available only when $0<\lambda<(n-1)(1-\rho_j)$ since quantities $(1-M_{j-}^2)$ and $(1-M_{j+}^2)$ appearing in the square roots of t_{j-} and t_{j+} respectively, should be non-negative. Empirical examination of the corresponding Bayes factors for values $\lambda>(n-1)(1-\rho_j)$ using MCMC has revealed a decreasing behavior of BF as λ increases, converging to one (as expected) for large λ since, in such case, both models assume that β is equal or very close to zero.

In order to interpret the behavior of the unicovariate Bayes factors, we present their logarithms in Fig. 4 as a function of the shrinkage parameter λ for different values of the Pearson correlation coefficient ρ_j and fixed sample size $n=50$. The sensitivity of such Bayes factors on different values of λ is clearly depicted.

As expected, the Bayes factors provide stronger evidence against the null model as the Pearson correlation between the response and the candidate variable increases. We focus on the thick dark horizontal line ($\mathrm{BF}_j^{\mathrm{un}}=3$), which, according to the interpretation tables of Kass and Raftery (1995) indicates the boundary between covariates for which the $\mathrm{BF}_j^{\mathrm{un}}$ provides or not evidence strong enough in favour of their inclusion in the model. We clearly see that the $\log\mathrm{BF}_j^{\mathrm{un}}$ never overcomes this threshold for Pearson correlation equal to 0.31 or lower. For these values, as the correlation increases, the overall values of $\mathrm{BF}_j^{\mathrm{un}}$ increase; however, it is always smaller than 3, implying that there is only weak evidence in favour of m_j for any value of λ.

For $\rho > 0.31$, the Bayes factor increases substantially, providing stronger evidence against the null model for some values of λ. For high correlations ($\rho > 0.6$), the unicovariate BF provides very strong evidence in favour of m_j for all values of λ. Furthermore, the shrinkage value that provides the strongest evidence against m_0 (i.e. maximizes $\mathrm{BF}_j^{\mathrm{un}}$) decreases when ρ increases. Similar figures can come up for samples of different size.

3.2 Specification of the shrinkage parameter λ using Bayes factors and correlations

Several approaches for tuning the shrinkage levels have been proposed, based on the generalized cross-validation techniques (Tibshirani 1996) or the C_p selection criterion (Efron et al. 2004). Here, we use the unicovariate Bayes factor (see (9)), its relation to the Pearson's sample correlation and its behavior as illustrated in Sect. 3.1 to specify a reasonable value for the shrinkage parameter λ.

3.2.1 Identifying the set of "non-important" covariates under all shrinkage values

Starting from Fig. 4, we observe that, for specific values of ρ, the $\mathrm{BF}_j^{\mathrm{un}}$ is lower than 3 for all the values of λ. In particular, for $n = 50$, the unicovariate Bayes factor will never support strongly enough models including any covariate correlated with the response with $\rho = 0.316$ or lower. Thus, we can identify a range of sample correlations corresponding to covariates that will never be considered as "important" determinants of the response for all values of λ and fixed n.

A graphical representation of the $\mathrm{BF}_j^{\mathrm{un}}$ against the values of ρ and λ will reveal the range of "non-important" correlations corresponding to covariates that will not be supported in the simple regression model for all the shrinkage levels.

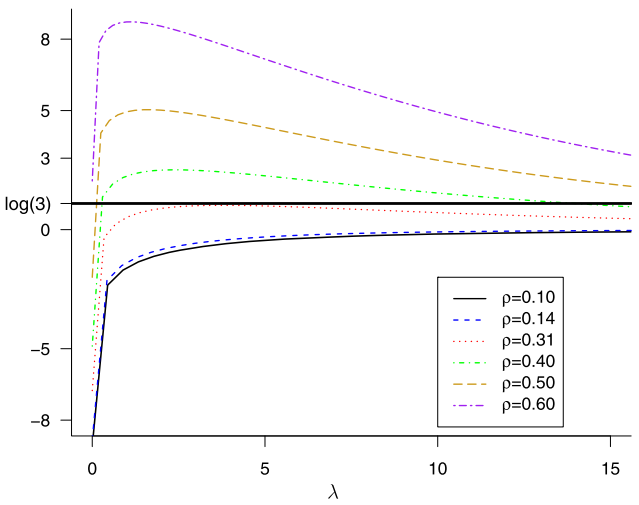


Fig. 4 Logarithm $\mathrm{BF}_j^{\mathrm{un}}$ against λ for several values of the Pearson correlation coefficient ρ

Thus, we define the non-important set of correlations using Definition 2.

Definition 2 (Non-important set of correlations $\mathcal{I}_\alpha$) The "non-important" set of correlations is the set of correlations that corresponds to covariates with unicovariate Bayes factors less than α for all possible shrinkage values λ, i.e. $\mathcal{I} = \{\rho : \mathrm{BF}_j^{\mathrm{un}} \leq \alpha \text{ for all } \lambda > 0\}$; where $\alpha \geq 1$.

Moreover, we specify the benchmark correlation using the following definition.

Definition 3 (Benchmark correlation ρ_{b_α}) The benchmark correlation ρ_{b_α} is defined as the maximum value in the "non-important" set of correlations $\mathcal{I}_\alpha$. All the covariates with correlation less than this ρ_{b_α} will not be supported strongly enough with Bayes factor bounded at the value of α for all shrinkage values λ.

An obvious selection would have been $\alpha = 1$, indicating that ρ_{b_1} is the level of correlation, where the Bayes factor does not support the inclusion of covariate with such correlation, for all possible values of λ. That is, the corresponding posterior probability of m_j will be ≤ 0.5 for all λ when comparing the constant and the simple regression model with covariate X_j (m_0 and m_j respectively).

Nevertheless, this choice is rather conservative, leading to benchmark correlations that are rather low. Moreover, since for important covariates we expect this posterior probability to be close to one (if not exactly equal to one) for a wide range of λ values, it seems intuitively appropriate to increase this value up to a reasonable level. We propose to use $\alpha = 3$ based on the rule of thumb of Kass and Raftery (1995), which also corresponds to pairwise posterior probability equal to 0.75. Moreover, simulations and empirical results show that this value seems to be quite a good trade-off between shrinkage and selection leading to parsimonious models without loosing predictive ability or overshrinking important coefficients.

3.2.2 Specifying λ via levels of practical significance for the Pearson correlation

In Sect. 3.2.1, we identified which covariates will never be supported strongly enough using Bayes factors that compare a simple regression model with the null model. Here, we specify λ via setting up the levels of practical significance for the Pearson correlation.

Returning back to (10), for any given value of λ, we can identify a specific ρ for which $\mathrm{BF}_j^{\mathrm{un}}$ takes a particular value. Specifically, we seek the combination of λ and ρ that produces a unicovariate Bayes factor equal to one. Covariates

Table 1 Shrinkage levels that correspond to BF = 1 for various values of ρ and n

Specification of λ	$n=50$ $\rho_{b_3}=0.316$		$n=100$ $\rho_{b_3}=0.225$		$n=500$ $\rho_{b_3}=0.101$	
$\{\lambda:\rho=\rho_t, \mathrm{BF}_j^{\mathrm{un}}=1\}$	$\rho_t=0.316,$	$\lambda=0.446$	$\rho_t=0.226,$	$\lambda=0.656$	$\rho_t=0.102,$	$\lambda=1.514$
$\{\lambda:\rho=\rho_t, \mathrm{BF}_j^{\mathrm{un}}=1\}$	$\rho_t=0.350,$	$\lambda=0.217$	$\rho_t=0.250,$	$\lambda=0.333$	$\rho_t=0.150,$	$\lambda=0.060$
$\{\lambda:\rho=\rho_t, \mathrm{BF}_j^{\mathrm{un}}=1\}$	$\rho_t=0.400,$	$\lambda=0.067$	$\rho_t=0.300,$	$\lambda=0.069$	$\rho_t=0.200,$	$\lambda=4.5\times10^{-4}$
$\{\lambda:\rho=0.01, \mathrm{BF}_j^{\mathrm{un}}=\frac{1}{150}\}$	$\rho_t=0.423,$	$\lambda=0.037$	$\rho_t=0.308,$	$\lambda=0.053$	$\rho_t=0.141,$	$\lambda=0.116$

with such correlations will be at the limits between significance and insignificance, since the Bayes factor cannot separate the competing models. This correlation will be called the threshold value ρ_t and its formal definition follows.

Definition 4 (Threshold correlation ρ_t) Threshold correlation ρ_t is the correlation that produces a unicovariate Bayes factor equal to one, i.e. $\rho_t=\{\rho:\mathrm{BF}_j^{\mathrm{un}}=1\}$ for a given λ.

We can now work backwards and specify a threshold level of practical significance $\rho_t\geq\rho_{b_\alpha}$ for $\alpha>1$ and obtain the corresponding shrinkage level λ. The choice of $\lambda=\lambda(\rho_t)$ implements a variable selection procedure in which covariates with Pearson correlation lower than ρ_t will never be supported in unicovariate comparisons.

Therefore, the threshold correlations can be used to specify the shrinkage parameter. This value of λ results in a Bayes factor that gives posterior weight of 50% to the model with a covariate with correlation equal to ρ_t and 50% to the constant model, i.e. it will not be able to separate between these two models. The choice of different threshold correlations, where the Bayes factor cannot decide which model is (even slightly) better, controls the shrinkage parameter λ and the sparsity of our finally selected model.

For example, for $n=50$, the benchmark values are $\rho_{b_1}=0.148$ and $\rho_{b_3}=0.316$. If we set $\rho_t=0.316$, any model including a covariate correlated with Y with $\rho=0.316$ will be a posteriori supported with 50% probability while this value will be increased as ρ increases. Selecting $\rho_t=0.25$ will be less strict, supporting models of slightly higher dimension, or $\rho_t=0.35$ will be more strict supporting more parsimonious models. Table 1 presents λ for $n=50, 100$ and 500 and various threshold correlation values.

Finally, an alternative way to exploit the relation between λ and ρ through the unicovariate Bayes factors is to specify the shrinkage parameter in such way that covariates with very low correlations are strongly not supported. Thus, we may specify λ such that a covariate with, for example, $\rho=0.01$ will result in a Bayes factor equal to $1/150$ in favour of the constant model. The shrinkage values, as well as the corresponding threshold correlation values for this setup for $n\in\{50, 100, 500\}$ are provided in the last row of Table 1.

3.3 Bayes factors for multiple lasso regression

In this section, we examine the sensitivity of the Bayes factors on the choice of the shrinkage parameters when performing multiple lasso regression. In particular, we investigate which is the level of the lasso partial correlation that corresponds to Bayes factor equal to one (i.e. which are the levels of partial correlation that correspond to the limits between significance and insignificance) for nested model comparisons and for any given level of shrinkage λ. In this way, we have a more general overview of the effect of the selected shrinkage level on our variable selection procedure. Before proceeding, we need to introduce some measures for lasso regression that are equivalent to the ones used in the ordinary regression analysis.

3.3.1 Preliminaries: lasso regression measures

Here, we follow the approach and the notation of Whittaker (1990, Chap. 5), in order to introduce some preliminary lasso measures. Therefore, we consider Y to be a $n\times1$ vector of random responses, X a $n\times p$ to be matrix of random variables that correspond to the explanatory variables and $\boldsymbol{\beta}$ to be fixed to a given value. Following this approach, the ordinary least squares prediction coefficient $\boldsymbol{\beta}^{\mathrm{ols}}$ arises when we minimize the residual variance $\mathsf{var}(\varepsilon)=\mathsf{var}(\mathsf{Y}-\mathsf{X}\boldsymbol{\beta})$ giving $\boldsymbol{\beta}^{\mathrm{ols}}=[\mathsf{var}(\mathsf{X})]^{-1}\,\mathsf{cov}(\mathsf{X},\mathsf{Y})$ assuming that $\mathsf{E}(\mathsf{X})$ and $\mathsf{E}(\mathsf{Y})$ are zero for simplicity. In the same analogy, the lasso prediction coefficient $\boldsymbol{\beta}^{\mathrm{lasso}}$ arises when we minimize a penalized version of the residual variance $\mathsf{var}(\mathsf{Y}-\mathsf{X}\boldsymbol{\beta})+k\|\boldsymbol{\beta}\|$ resulting in $\boldsymbol{\beta}^{\mathrm{lasso}}=[\mathsf{var}(\mathsf{X})]^{-1}(\mathsf{cov}(\mathsf{X},\mathsf{Y})-k\boldsymbol{s}_{\boldsymbol{\beta}})$; where $\boldsymbol{s}_{\boldsymbol{\beta}}$ is the sign vector of $\boldsymbol{\beta}^{\mathrm{lasso}}$ and k is the shrinkage level when working with the variances and expectations of the random variables Y and X.

We denote by $\mathsf{var}(\mathsf{Y}|\mathsf{X})=\mathsf{var}(\mathsf{Y}-\mathsf{X}\boldsymbol{\beta}^{\mathrm{ols}})$ the residual variance for the ordinary least squares regression model, which will be called the partial variance of Y with regressors defined by the columns of X; see Sect. 5.5 in Whittaker (1990) for a formal definition. In a similar way, we can introduce the lasso partial variance, denoted by $\mathsf{var}_{\mathrm{lasso}}(\mathsf{Y}|\mathsf{X})=\mathsf{var}(\mathsf{Y}-\mathsf{X}\boldsymbol{\beta}^{\mathrm{lasso}})$, which can be written as a function of the ordinary partial variance by the following expression

$$\mathsf{var}_{\mathrm{lasso}}(\mathsf{Y}|\mathsf{X})=\mathsf{var}(\mathsf{Y}|\mathsf{X})+k^2\boldsymbol{s}_{\boldsymbol{\beta}}^T\,\mathsf{var}(\mathsf{X})^{-1}\boldsymbol{s}_{\boldsymbol{\beta}}. \tag{11}$$

Following the definition in Whittaker (1990, p. 132), we introduce the lasso version of R^2 coefficient.

Definition 5 (Lasso R^2) The lasso R^2 is the coefficient determination of a lasso regression model, measuring the proportion of the variability of the response explained by the fitted lasso model and is given by

$$R^{(\text{lasso})2}_{\mathsf{Y}|\mathsf{X}} = \frac{\mathsf{var}(\mathsf{X}\boldsymbol{\beta}^{\text{lasso}})}{\mathsf{var}(\mathsf{Y})},$$

where $\mathsf{X}\boldsymbol{\beta}^{\text{lasso}}$ provides the vector of the fitted lasso values.

The above defined lasso multiple correlation coefficient can now be rewritten in terms of the lasso partial variance and the ordinary least squares R^2 via the expressions

$$R^{(\text{lasso})2}_{\mathsf{Y}|\mathsf{X}} = 1 - \frac{\mathsf{var}_{\text{lasso}}(\mathsf{Y}|\mathsf{X}) + 2k\|\boldsymbol{\beta}^{\text{lasso}}\|}{\mathsf{var}(\mathsf{Y})} \tag{12}$$

$$= R^{(\text{ols})2}_{\mathsf{Y}|\mathsf{X}} - 2k\frac{\|\boldsymbol{\beta}^{\text{lasso}}\|}{\mathsf{var}(\mathsf{Y})} - k^2\frac{\boldsymbol{s}_{\boldsymbol{\beta}}^T\,\mathsf{var}(\mathsf{X})^{-1}\boldsymbol{s}_{\boldsymbol{\beta}}}{\mathsf{var}(\mathsf{Y})}, \tag{13}$$

where $\|\cdot\|$ is the L_1 norm of the corresponding vector.

Corollary 1 *The lasso multiple correlation is always less than the ordinary multiple correlation, i.e.* $R^{(\text{lasso})2}_{\mathsf{Y}|\mathsf{X}} \leq R^{(\text{ols})2}_{\mathsf{Y}|\mathsf{X}} \leq 1$.

For any model m with covariates $\mathsf{X}_\ell \in \mathcal{V}_m$, we may define the model m_j^- with covariates in $\mathsf{X}_\ell \in \mathcal{V}_m \setminus \{\mathsf{X}_j\}$, which is nested to model m_j^+ with covariates $\mathsf{X}_\ell \in \mathcal{V}_m \cup \{\mathsf{X}_j\}$. Hence, covariate X_j is included in the linear predictor of model m_j^+ and excluded from the linear predictor of model m_j^-. Therefore, for any model configuration m (and the corresponding m_j^- and m_j^+) we can define the lasso version of the partial correlation coefficient using the following definition.

Definition 6 (Lasso Partial Correlation Coefficient) For any pair $(\mathsf{Y}, \mathsf{X}_j)$, we define the lasso partial correlation coefficient given a set of regressors $\mathsf{X}_{m_j^-}$ as the decrease of the percentage of unexplained response variability between model m_j^+ and m_j^- expressed as a proportion of the corresponding variability of the latter model. Therefore, the lasso partial correlation coefficient is given by

$$\mathsf{corr}^{(\text{lasso})}(\mathsf{Y}, \mathsf{X}_j|\mathsf{X}_{m_j^-}) = \sqrt{\frac{(1 - R^{(\text{lasso})2}_{\mathsf{Y}|\mathsf{X}_{m_j^-}}) - (1 - R^{(\text{lasso})2}_{\mathsf{Y}|\mathsf{X}_{m_j^+}})}{1 - R^{(\text{lasso})2}_{\mathsf{Y}|\mathsf{X}_{m_j^-}}}}$$

$$= \sqrt{1 - \frac{1 - R^{(\text{lasso})2}_{\mathsf{Y}|\mathsf{X}_{m_j^+}}}{1 - R^{(\text{lasso})2}_{\mathsf{Y}|\mathsf{X}_{m_j^-}}}}. \tag{14}$$

The above definition of the lasso partial correlation is based on a property of the ordinary partial correlation (see Whittaker 1990, p. 140). From (13) we see that for $k \to 0$ then the above defined lasso partial correlation tends to the ordinary partial correlation. Moreover, for the range of values of the shrinkage parameter we use in practise and in the illustrated examples here, the differences between the two measures are minor. As we will see in Sect. 3.3.2, the sample estimate of $\mathsf{corr}^{(\text{lasso})}(\mathsf{Y}, \mathsf{X}_j|\mathsf{X}_{m_j^-})$ appears in the Bayes factors when comparing two models that differ by a covariate X_j and we will use this property to identify the shrinkage levels separating important and non-important covariates in such pairwise model comparisons.

3.3.2 *Bayes factors as functions of lasso regression measures*

We now focus on the comparison of any two nested models that differ by a covariate X_j. For any given model structure m, this comparison is evaluated by $\text{BF}^{\text{mu}}_{m,j}$, which is defined as follows.

Definition 7 (Nested Multiple Lasso Bayes Factor $\text{BF}^{\text{mu}}_{m,j}$) For any model m with included covariates $\mathsf{X}_\ell \in \mathcal{V}_m$, the nested multiple lasso Bayes factor $\text{BF}^{\text{mu}}_{m,j}$ is defined as the Bayes factor that evaluates the evidence of model m_j^+ with covariates $\mathsf{X}_\ell \in \mathcal{V}_m \cup \{\mathsf{X}_j\}$ versus model m_j^- with covariates $\mathsf{X}_\ell \in \mathcal{V}_m \setminus \{\mathsf{X}_j\}$

In the following, $\boldsymbol{y}$ is the $n \times 1$ vector of observed responses, $\boldsymbol{X}_j$ is the $n \times 1$ vector of observed values for covariate X_j and $\boldsymbol{X}_m$ is the data matrix with columns $\boldsymbol{X}_\ell$ for $\mathsf{X}_\ell \in \mathcal{V}_m$. The variances, correlations and R^2 for $\boldsymbol{y}$, $\boldsymbol{X}_j$ and $\boldsymbol{X}_m$ refer to the corresponding sample estimates.

We use the Laplace approximation to integrate out $\boldsymbol{\beta}$ and, for $d \to 0$, the corresponding $\text{BF}^{\text{mu}}_{m,j}$ is approximately given by

$$\text{BF}^{\text{mu}}_{m,j} \approx \lambda c_{\text{mu}}\left(\frac{|\boldsymbol{X}^T_{m_j^+}\boldsymbol{X}_{m_j^+}|}{|\boldsymbol{X}^T_{m_j^-}\boldsymbol{X}_{m_j^-}|}\right)^{-1/2} \times \frac{(\|\boldsymbol{y} - \boldsymbol{X}_{m_j^+}\widehat{\boldsymbol{\beta}}^{\text{lasso}}_{m_j^+}\|^2 + 2\lambda||\widehat{\boldsymbol{\beta}}^{\text{lasso}}_{m_j^+}\|)^{-df_{\text{mu}}/2}}{(\|\boldsymbol{y} - \boldsymbol{X}_{m_j^-}\widehat{\boldsymbol{\beta}}^{\text{lasso}}_{m_j^-}\|^2 + 2\lambda\|\widehat{\boldsymbol{\beta}}^{\text{lasso}}_{m_j^-}\|)^{-(df_{\text{mu}}-1)/2}}$$

$$= kc_{\text{mu}}\left(\frac{\mathsf{var}_{\text{lasso}}(\boldsymbol{y}|\boldsymbol{X}_{m_j^+}) + 2k\|\widehat{\boldsymbol{\beta}}^{\text{lasso}}_{m_j^+}\|_1}{\mathsf{var}_{\text{lasso}}(\boldsymbol{y}|\boldsymbol{X}_{m_j^-}) + 2k\|\widehat{\boldsymbol{\beta}}^{\text{lasso}}_{m_j^-}\|_1}\right)^{-df_{\text{mu}}/2} \times \big[\mathsf{var}(\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-})\big(\mathsf{var}_{\text{lasso}}(\boldsymbol{y}|\boldsymbol{X}_{m_j^-}) + 2k\|\widehat{\boldsymbol{\beta}}^{\text{lasso}}_{m_j^-}\|\big)\big]^{-1/2},$$

where $c_{\text{mu}} = \sqrt{\pi}\frac{\Gamma(\frac{df_{\text{mu}}}{2})}{\Gamma(\frac{df_{\text{mu}}-1}{2})}$, $df_{\text{mu}} = n + 2\alpha + p$, $\widehat{\boldsymbol{\beta}}^{\text{lasso}}_{m_j^+}$ and $\widehat{\boldsymbol{\beta}}^{\text{lasso}}_{m_j^-}$ are the lasso estimates when regressing $\boldsymbol{y}$ on $\boldsymbol{X}_{m_j^+}$, and $\boldsymbol{X}_{m_j^-}$ respectively; $\text{var}(\boldsymbol{y}|\boldsymbol{X})$ and $\text{var}_{\text{lasso}}(\boldsymbol{y}|\boldsymbol{X})$ are the sample estimates of the partial variances for the ordinary and the lasso (respectively) regression model with response $\boldsymbol{y}$ and data matrix $\boldsymbol{X}$. Moreover, k, which is the shrinkage level referring to the penalized minimization of the variance as described in Sect. 3.3.1, is set equal to $\lambda/(n-1)$, since λ is the shrinkage level when working with the usual lasso representation (1).

Therefore, using (12) and (14), the $\text{BF}^{\text{mu}}_{m,j}$ can be expressed in terms of the lasso partial correlation by the expression

$$\text{BF}^{\text{mu}}_{m,j} \approx k c_{\text{mu}}\big[1 - \text{corr}^{(\text{lasso})2}(\boldsymbol{y}, \boldsymbol{X}_j|\boldsymbol{X}_{m_j^-})\big]^{-df_{\text{mu}}/2} \times \frac{1}{\sqrt{(1 - R^{(\text{ols})2}_{\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}})(1 - R^{(\text{lasso})2}_{\boldsymbol{y}|\boldsymbol{X}_{m_j^-}})}}. \quad (15)$$

According to our proposed method, we define the shrinkage level by setting the unicovariate BF^{un}_j equal to one for a given level of threshold correlation. Using (15) we can identify the corresponding threshold partial correlation level imposed by any selected level of λ. In this way, we can examine the behavior of the proposed variable selection procedure and why covariates with low Pearson correlations are finally included in the a posteriori most probable models.

Theorem 1 *For any selected* λ *and for large* n, $\text{LBF}^{\text{un}}_j = \text{LBF}^{\text{mu}}_{m,j} \Rightarrow \text{corr}^{(\text{lasso})2}(\boldsymbol{y}, \boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}) \leq (\rho_j - ks)$; *where* LBF *is the Laplace approximation of the corresponding Bayes factor.*

The proof of the theorem can be found in the Appendix.

Corollary 2 *The threshold value for the lasso partial correlation is upper-bounded by a penalized expression of the corresponding value of the Pearson correlation.*

Corollary 3 *For large sample size* n *and for any* $\lambda/(n-1) \to 0$, *the threshold partial lasso correlation is approximately equal to the corresponding values for the Pearson correlation.*

Corollary 3 helps us to approximately identify the threshold levels imposed on the comparison of multiple regression models. For large sample sizes, the threshold values of the partial correlations will be the same as the ones imposed on the Pearson correlation, while for small sample sizes it will be lower and bounded by a penalized version of the threshold value of the Pearson correlation. Moreover, this behavior justifies why covariates with low Pearson correlation are finally added in the most probable a posteriori models, since, for responses that depend on a large number of covariates, the partial correlations will increase as more important covariates are included in the model formulation. The shrinkage value λ that corresponds to $\rho_t = \rho_{b_\alpha}$ makes $\text{BF}^{\text{un}}_j = 1$ for covariates with correlation equal to ρ_{b_α}. Hence, for $\alpha > 1$, we have that $0 < \lambda < \lambda_{\max}$, where $\lambda_{\max}$ corresponds to $\text{BF}^{\text{un}}_j = \alpha$ for covariates with correlation equal to ρ_{b_α} (i.e. it is the value that maximizes the BF^{un}_j for ρ_{b_α}). The value of $\lambda_{\max}$ can be obtained by maximizing the Laplace approximation BF^{un}_j and it turns out that $0 < \lambda/(n-1) < \lambda_{\max}/(n-1) \to 0$ and thus, our proposed λ has the approximating behavior (and the corresponding attributes) described in Corollary 3.

To sum up, in this section we have illustrated the effect of any chosen level of λ on $\text{BF}^{\text{mu}}_{m,j}$. To describe and interpret this result, we have identified the value of (lasso) partial correlation which separates important and non-important covariates for any nested model comparison. In this way, we can understand how the method works in more complicated model comparisons and justify why covariates with low (or high) Pearson correlations are included in (or excluded from) models with high posterior probabilities.

4 Final recommendations for the specification of λ

4.1 Subjective selection of λ

The proposed approach offers a full understanding of the behavior of the unicovariate Bayes factors for different values of the shrinkage parameter λ. This approach can be useful in cases where practitioners have preferences or clear recommendations about the levels of practical significance. In such cases, the practitioner can select a desirable level of practical significance for the Pearson correlation (which is widely understood) and set λ such as the lasso-based Bayes factor gives equal posterior probabilities for both the constant and the corresponding simple regression model. Moreover, the benchmark values offer reasonable lower acceptable bounds for the selection of such values for λ.

If there is no preference about the chosen levels of practical significance, we suggest to set the threshold correlation equal to the benchmark correlation ρ_{b_3} as a reference value based on the rule of thumb of Kass and Raftery (1995). Simple regression models with ρ_{b_3} correlation between the response and the included covariate will have posterior probability against the null model that will not exceed the 75% *for any value of* λ.

We suggest to specify λ such that, the threshold correlation is equal to ρ_{b_3}, that is, $\rho_t = \rho_{b_3}$. Therefore, covariates with correlation equal to ρ_{b_3} will be supported with posterior

probabilities equal to 50% in pairwise comparisons of models m_0 and m_j. Covariates with higher or lower correlations will have higher or lower than 50% posterior probabilities respectively, in similar pairwise model comparisons.

4.2 Using a gamma hyperprior for λ

In Sect. 4.1, we described how we can specify λ using subjective preferences on the level of practical significance. In this section, we propose the use of a gamma hyperprior for the shrinkage parameter λ with the parameters specified by the threshold and benchmark correlations. The specification of this hyperprior avoids the prior support of extremely low and large values of λ that activate the Lindley-Bartlett paradox or overshrink model coefficient. In the first case, posterior inclusion probabilities degenerate to zero (for any available dataset), while in the latter they converge towards 0.5. Both cases give nonsensical results for the variable selection procedure and should be excluded from the hyperprior. Moreover, the hyperprior described in this section is relatively robust to different hyperparameter values following the same logic described below; see Sect. 5 for details.

The Gamma(c_λ, d_λ) is a natural choice as a hyperprior here, since, it is conditional conjugate for the likelihood of the lasso regression model. Hence, this choice results in adding a simple step in the Gibbs algorithm described in Sect. 2.2, where we update λ from a gamma full conditional posterior distribution given by

$$\lambda|\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau, \boldsymbol{y} \sim \text{Gamma}\left(p + c_\lambda, \tau \sum_{j=1}^{p} |\beta_j| + d_\lambda\right).$$

The specification of c_λ and d_λ will be based on the benchmark and threshold correlations and their corresponding λ values. In particular, for a given $\alpha > 1$ we set the hyperprior mean equal to the value of λ that corresponds to $\rho_t = \rho_{b_\alpha}$. The prior variance depends on the choice of $u > \alpha$ and is set equal to $(\mathcal{R}_{\alpha,u}/4)^2$ such that $P(\mu_\alpha - \mathcal{R}_{\alpha,u} < \lambda < \mu_\alpha + \mathcal{R}_{\alpha,u}) \geq 0.9375$ from the Chebyshev inequality. Hence, $\mathcal{R}_{\alpha,u}$ will be the half-range of an at least 93.75% probability interval of the prior distribution. To specify $\mathcal{R}_{\alpha,u}$, we select u large enough to represent the lower bound of an at least 93.75% probability interval that will be a priori supported by our hyperprior. The choice of u controls the range of values of the prior distribution, since higher values of u will correspond to greater $\mathcal{R}_{\alpha,u}$. Hence, we set

$$\mu_\alpha = \{\lambda : \rho_t = \rho_{b_\alpha}\} \quad \text{and} \quad \mathcal{R}_{\alpha,u} = \mu_\alpha - \{\lambda : \rho_t = \rho_{b_u}\},$$

and the corresponding hyperparameters c_λ and d_λ are given by

$$c_\lambda = \left(\frac{\mu_\alpha}{\mathcal{R}_{\alpha,u}/4}\right)^2 \quad \text{and} \quad d_\lambda = \frac{\mu_\alpha}{(\mathcal{R}_{\alpha,u}/4)^2}.$$

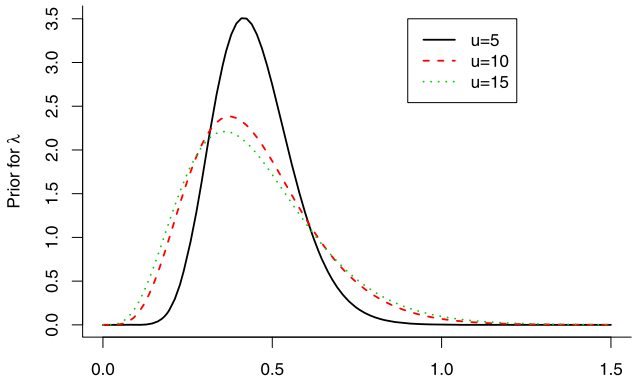


Fig. 5 Densities of the gamma hyperprior for a variety of u values and $\alpha = 3$

When λ is equal to the prior mean μ_α, the lasso based Bayes factor assigns, in pairwise comparisons of m_0 and m_j, 50% posterior probability to the model including a covariate X_j with correlation ρ_{b_α}. The value of $\mathcal{R}_{\alpha,u}$ expresses the prior uncertainty around this shrinkage value μ_α (and the corresponding level of practical significance is expressed by ρ_{b_α}). By setting $\mathcal{R}_{\alpha,u}$ as defined above, we allow λ to take values which assign levels of practical significance higher than ρ_{b_u} with very low prior probability (lower than 0.0625).

We suggest as default choices $\alpha = 3$ and $u = 10$ and perform sensitivity analysis around these values. This choice specifies that the threshold between significance and insignificance will take values in an interval $(\rho_{b_\ell}, \rho_{b_{10}})$ with ρ_{b_ℓ} being the threshold correlation when $\lambda = \mu_3 + \mathcal{R}_{3,10}$ and therefore $\rho_{b_\ell} < \rho_{b_3}$ and $1 < \ell < 3$. These values of λ correspond to covariates that are supported with posterior probabilities less than $\frac{\ell}{\ell+1}$ and 0.91 respectively for all values of λ with $\frac{\ell}{\ell+1} \in (0.5, 0.75)$. For example, for $n = 50$, $\alpha = 3$ and $u = 10$ the prior mean is set $\mu_3 = 0.446$ and assigns 50% chance to covariates with correlation 0.316. The interval $\lambda \in (0.086, 0.806)$ corresponds to threshold correlations taking values in the interval $(0.287, 0.390)$. Hence, the threshold correlation between important and non-important covariates will take values from 0.287 to 0.39 with at least 93% prior probability.

Details for the hyperparameters specified by setting $\alpha = 3$ and $u = 5$, 10, 15 when $n = 50$ are presented in Table 2. The corresponding prior densities are depicted in Fig. 5 with all of them having similar shape but different variances. Detailed sensitivity analysis for the simulation study of Sect. 5.1 is presented in Sect. 5.1.3. Getting a quick preview of the posterior results, we observe that all averages (over 50 generated datasets) of posterior inclusion probabilities are similar for the three different choices of u; see Fig. 7(c). Moreover, the corresponding average posterior inclusion probabilities for the "low-information" Gamma$(0.01, 0.01)$ hyperprior are also presented for reference. From this comparison it is evident that

Table 2 Prior values for gamma hyperprior on λ for $\alpha = 3$ and $u = 5, 10, 15$

u	ρ_{b_α}	μ_α	σ^2_α	$\mu_\alpha - \mathcal{R}_{\alpha,u}$	$\mu_\alpha + \mathcal{R}_{\alpha,u}$	$P(\lvert\lambda - \mu_\alpha\rvert < \mathcal{R}_{\alpha,u})$	c_λ	d_λ
5	0.316	0.446	0.014	0.212	0.680	0.958	14.551	32.610
10	0.316	0.446	0.032	0.086	0.806	0.959	6.144	13.768
15	0.316	0.446	0.039	0.052	0.840	0.959	5.131	11.499

μ_α and $\sigma^2_\alpha = (\mathcal{R}_{\alpha,u}/4)^2$: prior mean and variance of λ for $\alpha = 3$; $\mu_\alpha - \mathcal{R}_{\alpha,u}$: lower bound used to define the variance; c_λ and d_λ: shape and rate parameters of the gamma hyperprior

our approach provides similar posterior inclusion probabilities for important covariates and much lower posterior inclusion probabilities for the non-important covariates removing redundant model uncertainty introduced when using the Gamma$(0.01, 0.01)$ due to the support of unnecessarily large values of λ making the latter posterior probabilities to shrink towards 0.5. For a thorough robustness analysis see Sect. 5.1.3.

5 Illustrations

In this section, we present extensive results from two simulation studies and from a real dataset. The section is divided into two main sub-sections. In the first one, we present results for the first simulation study for a variety of λ values coming from different threshold correlations, as well as a robustness analysis concerning the gamma hyperprior using different prior means and variances, while in Sect. 5.2, we perform an extensive comparison of our proposed methods with a variety lasso methods that are available in the bibliography.

5.1 Simulation study 1: results for different prior set-ups of λ

We start our illustrations concerning the performance of the proposed Bayesian lasso by presenting results for a variety of values for λ specified from different threshold correlations in Sect. 5.1.2 for the simulation study described in Sect. 5.1.1. A robustness analysis over different parameter values of the gamma hyperprior follows in Sect. 5.1.3.

5.1.1 The design of the simulation study 1

In this illustration, we adopt the simulation design of Nott and Kohn (2005), which consists of 15 variables of 50 observations each. Specifically, the first 10 variables follow independent standard normal distribution and the last 5 variables are generated as follows,

$$(\mathsf{X}_{11}, \ldots, \mathsf{X}_{15}) = (\mathsf{X}_1, \ldots, \mathsf{X}_5) \times (0.3, 0.5, 0.7, 0.9, 1.1)^T \times (1, 1, 1, 1, 1) + \boldsymbol{E},$$

where $\boldsymbol{E}$ consists of 5 independent $N(0, 1)$. Under this design, the last five variables are highly correlated, whereas, they are moderately correlated with the first five variables. The response is generated as

$$\mathsf{Y} = 2\mathsf{X}_1 - \mathsf{X}_5 + 1.5\mathsf{X}_7 + \mathsf{X}_{11} + 0.5\mathsf{X}_{13} + \varepsilon,$$

where $\varepsilon \sim N(0, 2.5^2)$. This set of simulated data comprises of covariates that are correlated with each other and therefore, variable selection is not straightforward since some true effects are covered by the effects of collinear covariates. The signal-to-noise ratio (SNR) is equal to 2.15; where SNR is defined as the ratio of the variance of the linear predictor over the error variance. The resulting pairwise true correlations between the covariates are:

$$\mathsf{corr}(X_1, X_j) = 0.15, \qquad \mathsf{corr}(X_2, X_j) = 0.26,$$
$$\mathsf{corr}(X_3, X_j) = 0.36, \qquad \mathsf{corr}(X_4, X_j) = 0.46,$$
$$\mathsf{corr}(X_5, X_j) = 0.56 \quad \text{and}$$
$$\mathsf{corr}(X_l, X_j) = 0.74 \quad \text{for } l, j = 11, \ldots, 15 \text{ and } l \neq j.$$

5.1.2 Results using different threshold correlation values

Here, we present results using different threshold correlation values. Specifically, we use the shrinkage values presented in Table 1 for $n = 50$. Each MCMC was updated using 10,000 iterations after discarding 1,000 additional observations. All results are evaluated over 100 datasets generated using the sampling scheme described in Sect. 5.1.1. Figure 6 depicts the boxplots (over the 100 simulated datasets) of the posterior inclusion probabilities for covariates $\mathsf{X}_1, \mathsf{X}_5, \mathsf{X}_7, \mathsf{X}_{11}, \mathsf{X}_{13}$, which are actually used to generate Y. Covariates $\mathsf{X}_1, \mathsf{X}_7$ and X_{11} are indicated as important ones (with posterior inclusion probabilities > 0.5) with very high frequency for all the shrinkage levels, whereas, the remaining covariates are less frequently picked as important ones. The inclusion probabilities for these two variables naturally become smaller as the shrinkage parameter decreases, while another characteristic is the large variability of the inclusion probabilities over the generated datasets. The posterior inclusion probabilities for the remaining variables are considerable lower ranging from 0.15

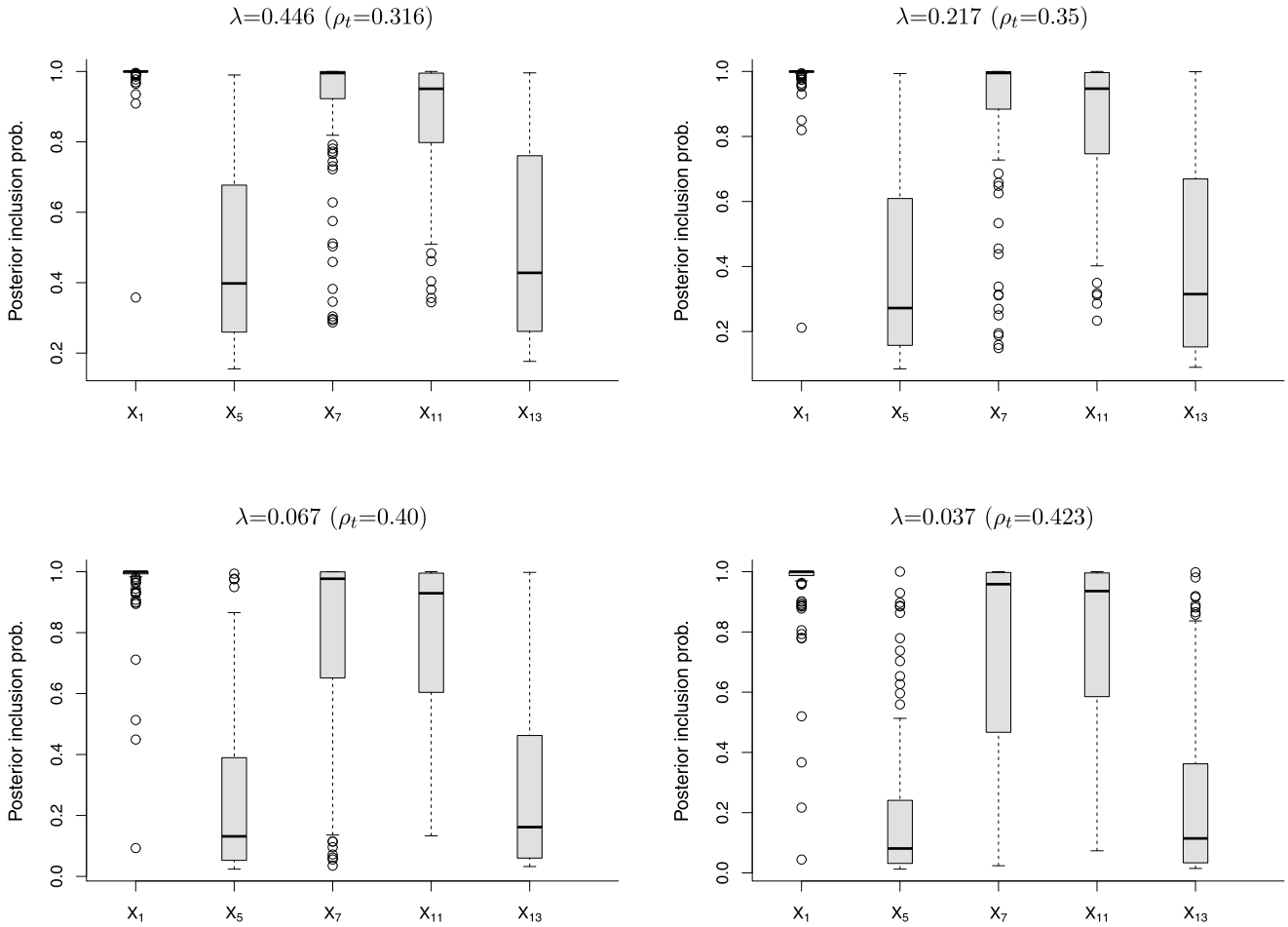


Fig. 6 Boxplots of the posterior inclusion probabilities for covariates with true non-zero effects over 100 generated datasets for simulated study 1 described in Sect. 5.1.1

Table 3 True values of the Pearson and the partial correlation coefficients (in absolute values) for Example 5.1.1

	X_1	X_2	X_3	X_4	X_5	X_6, X_8, X_9, X_{10}	X_7	X_{11}	X_{12}, X_{14}, X_{15}	X_{13}
corr$(\boldsymbol{y}, \boldsymbol{X}_j)$	0.56	0.17	0.24	0.31	0.15	0.00	0.34	0.56	0.44	0.50
corr$(\boldsymbol{y}, \boldsymbol{X}_j \vert \boldsymbol{X}_{\setminus j})$	**0.55**	0.00	0.00	0.00	0.15	0.00	**0.52**	**0.37**	0.00	0.20

to 0.22 for $\lambda = 0.446$, from 0.06 to 0.13 for $\lambda = 0.217$, from 0.03 to 0.10 for $\lambda = 0.067$ and from 0.19 to 0.27 for $\lambda = 0.037$.

Table 3 presents the true Pearson correlation and partial correlations for this structure of simulated dataset. Covariates X_1, X_7 and X_{11}, which have the highest posterior probabilities, are the ones with the highest partial correlations with the response conditional on all the remaining variables. While the covariates X_5 and X_{13} have been used to generate the response, their corresponding partial correlations are low due to the high correlations among the variables X_5, X_{11} and X_{13}. Our proposed Bayesian lasso procedure tends to select only one of the three highly correlated covariates and thus, the inclusion probabilities of X_5 and X_{13} are as expected low.

All illustrated results presented in this short section suggest that our proposed Bayesian lasso method works efficiently and according to the behavior examined previously in this manuscript. Important covariates X_1, X_7 and X_{11} are identified by both the maximum a posteriori (MAP) model and the posterior inclusion probabilities. Covariates X_5 and X_{13} with non-zero true effects cannot be traced by our proposed method due to the collinearities. In particular, X_5 has low Pearson and partial with the response while X_{13} has not supported by our procedure although it is highly correlated with the response (Pearson correlation = 0.5) due to its high correlation with the covariate X_{11} (corr$(X_{11}, X_{13}) = 0.74$)

resulting to low partial correlation (equal to 0.2) when the latter in included in the model.

5.1.3 Robustness analysis for the gamma hyperprior

Before proceeding with the comparison of the proposed methods with similar methods available in the lasso literature, we present a sensitivity analysis for a variety of hyperparameter values of the gamma hierarchical model introduced in Sect. 4.2.

As reference we consider the choices $\alpha = 3$ and $u = 10$ and we present two sensitivity analyses. In the first one we fix the mean equal to the λ obtained using threshold correlation equal to ρ_{b_3} and we change the variance using $u = 5$, 10 and 15 while in the second we fix the variance and we change the mean value using λ values obtained from different threshold correlations. We also present the results from the prior distribution Gamma$(0.01, 0.01)$, which can be considered as a low-information prior (LIP) since its variance is high (equal to 100). Figures 7 and 8 present results from these two sensitivity analyses. More specifically, each figure presents boxplots of (a) the root mean square errors (RMSEs) of the fitted values (with respect to the actual responses) and (b) the RMSEs of the estimates of β (with respect to the actual β), (c) the averages (over 50 simulated datasets) of the posterior inclusion probabilities $f(\gamma_j = 1|\boldsymbol{y})$ with $j = 1, \ldots, 15$ as well as (d) boxplots of the posterior means of the shrinkage parameter for the generated datasets.

RMSEs are calculated using the following expressions

$$RMSE(\widehat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \quad \text{and}$$
$$RMSE(\widehat{\beta}) = \sqrt{\frac{1}{n}\sum_{j=1}^{p}(\beta_j - \widehat{\beta}_j)^2} \tag{16}$$

for the fitted values and the estimated β respectively; $\widehat{\beta}_j$ are set equal to the posterior medians of β_j^* while $\widehat{y}_i = \sum_{j=1}^{p} X_{ij}\widehat{\beta}_j$.

From the results of the first analysis (Fig. 7), we observe that RMSEs are similarly distributed for all different choices of u. Posterior inclusion probabilities present only minor differences while some differences are observed in the distribution of the posterior means of λ which eventually do not have a dominant effect on posterior inclusion probabilities and RMSEs.

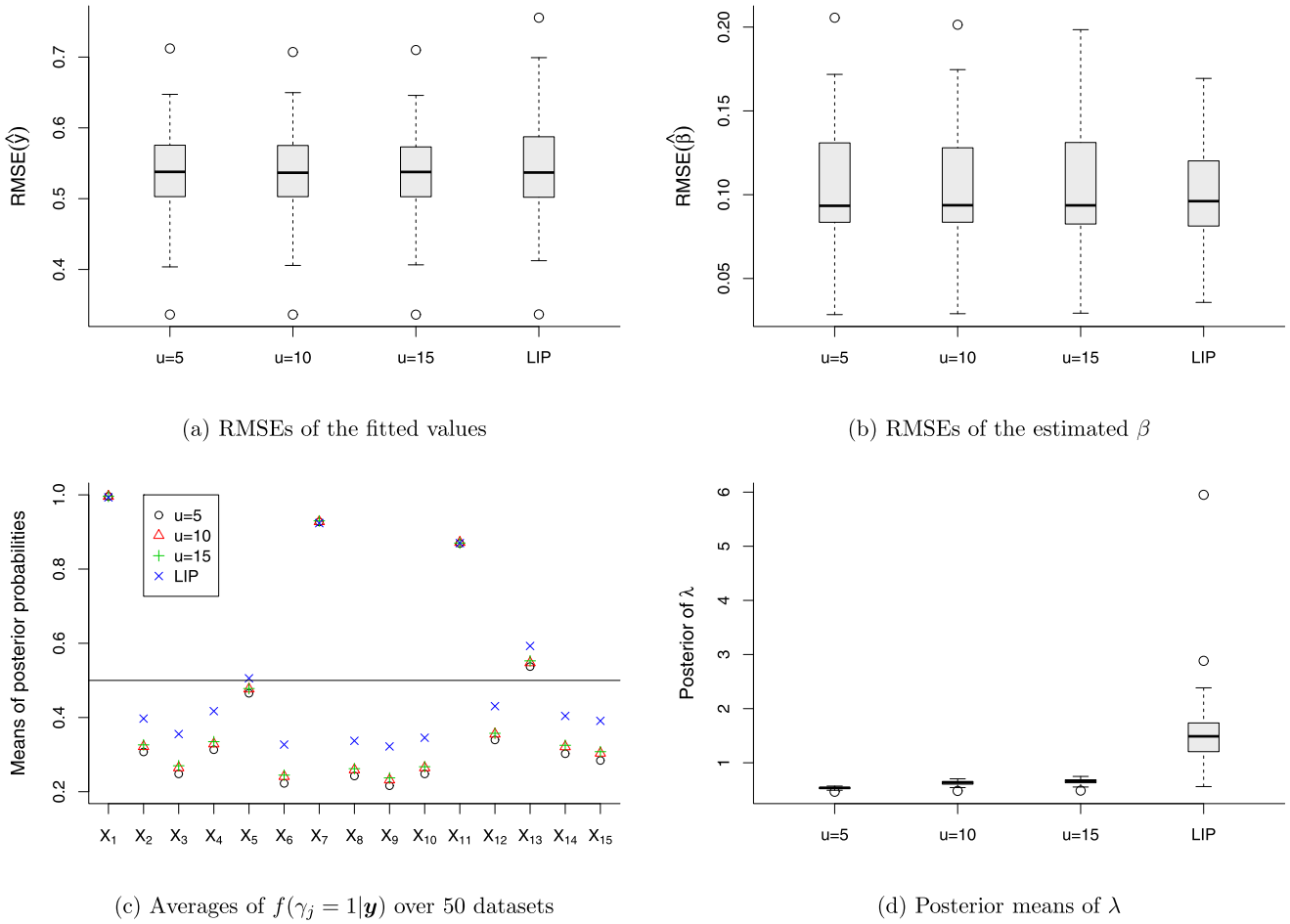
For the second sensitivity analysis (Fig. 8), the prior mean is set equal to the λ values that correspond to the threshold correlation ρ_{b_α} with $\alpha = 2, 3, 5, 7$, whereas the prior variance remains constant to the one computed for the default values suggested in Sect. 4.2. The picture of the results presented in Fig. 8 is similar to the first sensitivity analysis with RMSEs and posterior inclusion probabilities to be quite robust for all values of α. Differences are again observed in the distribution of posterior means of λ with more striking the one for $\alpha = 2$. This latter hyperprior setup also influences mildly the posterior inclusion probabilities inflating them slightly towards larger values, especially for the non-important covariates. This might be an indication of overshrinkage which is also confirmed by the higher posterior means of λs in the corresponding boxplots. Finally, the posterior medians of coefficients β_j^* are quite robust for the hyperparameters values illustrated in Fig. 8(b).

Finally, we close this section with a short comparison of the above results with the ones obtained by a Gamma$(0.01, 0.01)$ hyperprior. Using this hyperprior, RMSEs are of equivalent scale with the ones presented in Figs. 7 and 8 but posterior inclusion probabilities of the non-important covariates are all inflated towards 0.5 with an percentage increase ranging from 7%–10% (average posterior inclusion probabilities range: 0.32–0.43 versus 0.23–0.36 for the default values of our approach). This is due to the unnecessary prior support of λ values that cause overshrinkage and convergence of the corresponding Bayes factors towards one (posterior means of λ are ranging from 0.56 to 5.95 with median value 1.49 in contrast to 0.48–0.70 and 0.63 respectively, for the default values of our approach).

5.2 Comparison with different lasso methods

In this section, we proceed with extended comparisons of our proposed methods with a variety of lasso methods presented in related literature. All methods are presented in the simulation study 1 of Sect. 5.1.1 and additionally in a second simulation study and a real dataset presented in Sects. 5.2.2 and 5.2.3 respectively.

In all illustrations, which follow, we present results for our Bayesian lasso approach with λ specified by the threshold correlation value equal to ρ_{b_3} and using gamma hyperprior at the default values ($\alpha = 3$, $u = 10$) denoted as BLVS and BLVS-G respectively with the normal mixture of inverse gamma (NMIG) prior (Scheipl 2010), the Normal-gamma (NG) prior (Griffin and Brown 2010), the horseshoe (HS) prior (Carvalho et al. 2010), the Lars-lasso algorithm (Efron et al. 2004) with a 10-fold cross validation (CV) method to tune the shrinkage parameter. Lars was implemented using the homonym `R` package (R Development Core Team 2011), NMIG using the `spikeSlabGAM R` package (Scheipl 2011), and NG and HS using the `monomvn` package (Gramacy 2010) in two versions: using the full model only (i.e. as a purely shrinkage method with no direct variable selection) and in combination with



(a) RMSEs of the fitted values

(b) RMSEs of the estimated β

(c) Averages of $f(\gamma_j = 1|\boldsymbol{y})$ over 50 datasets

(d) Posterior means of λ

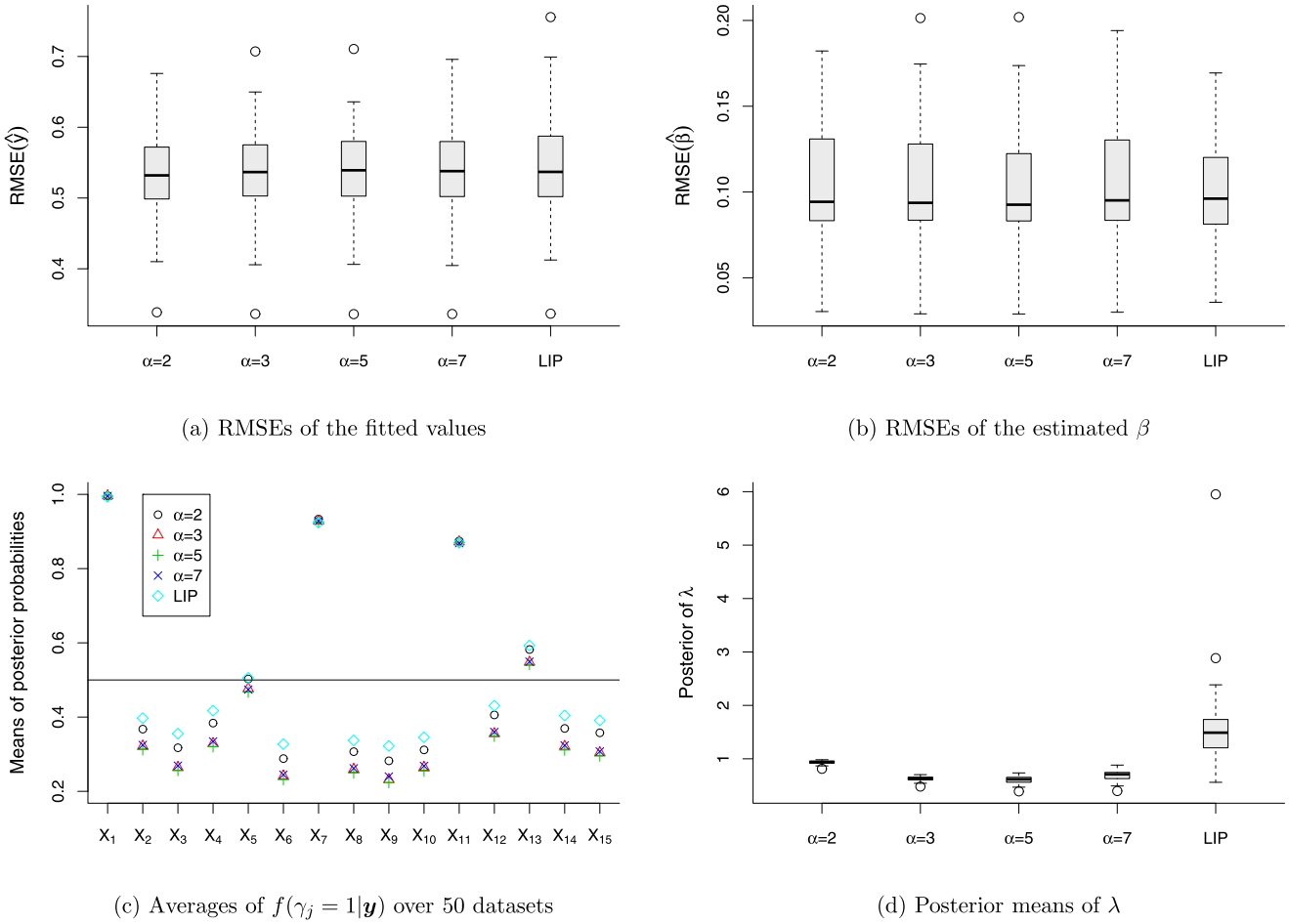
RMSE: Root mean square errors calculated by (16); $\widehat{\beta}_j$ is set equal to the posterior median of $\beta_j^* = \gamma_j \beta_j$, $\widehat{y}_i = \sum_{j=1}^{p} X_{ij} \widehat{\beta}_j$ and LIP: Gamma$(0.01, 0.01)$ *"low information" prior.*

Fig. 7 Results for the first sensitivity analysis for simulation study 1 of Sect. 5.1.1 [three different prior set-ups were considered with fixed mean and different variances ($\alpha = 3$ for all analyses, $u \in \{5, 10, 15\}$); 50 generated datasets for each prior set-up]

reversible jump MCMC (i.e. implementing also direct variable selection). NG has been implemented using the prior values suggested by Griffin and Brown (2010), while all the remaining methods have been applied using the default values of their corresponding R functions. For each comparison, we have used 100 generated datasets. All methods are compared using RMSEs of the fitted values and estimated βs as described in Sect. 5.1.3. For $\widehat{\beta}$ we have used the posterior medians for our methods and the reversible jump versions of NG and HS while for the rest of the Bayesian methods we have used the posterior means. For the ordinary lasso, we use the estimates obtained by the Lars algorithm. All fitted values were calculated as in Sect. 5.1.3. All details and acronyms are also summarized in Table 4 and are used as reference for all tables and figures of this section.

5.2.1 *Simulation study 1 (cont.)*

The distribution of RMSEs for the fitted values and the estimated β over 100 generated datasets using the simulation scheme of Sect. 5.1.1 is presented in Fig. 9 for all methods summarized in Table 4. The first five boxplots in each figure correspond to the methods that perform direct variable selection. Moreover, Lars (last boxplot) implements direct variable selection in the sense that non-important covariates are constrained to zero; nevertheless, no variable inclusion or model probabilities are available for this method and therefore model uncertainty cannot be evaluated. All RMSEs seem to be of similar scale with small differences (taking into consideration the overall variability). Nevertheless, the original HS and (especially) NG methods, which are implemented on the full model and do not implement direct variable selection but shrinkage, have the lowest RMSEs of



(a) RMSEs of the fitted values

(b) RMSEs of the estimated β

(c) Averages of $f(\gamma_j = 1|\boldsymbol{y})$ over 50 datasets

(d) Posterior means of λ

RMSE: Root mean square errors calculated by (16); $\widehat{\beta}_j$ is set equal to the posterior median of $\beta_j^* = \gamma_j \beta_j$, $\widehat{y}_i = \sum_{j=1}^{p} X_{ij}\widehat{\beta}_j$ and LIP: Gamma$(0.01, 0.01)$ *"low information" prior.*

Fig. 8 Results for the second sensitivity analysis for simulation study 1 of Sect. 5.1.1 [four different prior set-ups were considered with fixed variances and different means: mean is set for $\alpha \in \{2, 3, 5, 7\}$ and the variance is fixed to the one using the default values ($\alpha = 3, u = 10$) for all analyses; 50 generated datasets for each prior set-up]

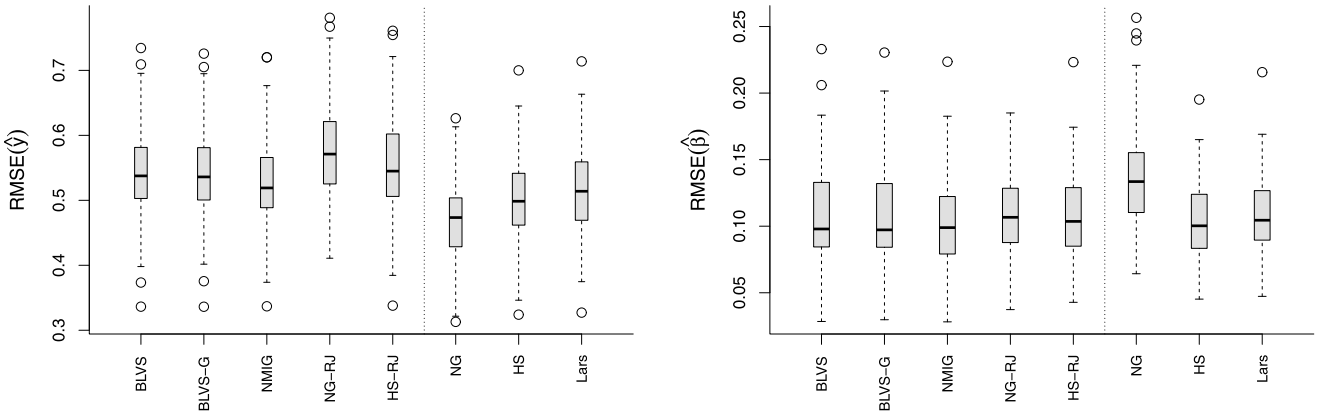
Table 4 Acronyms and details for the implemented methods

	Acronym	Method
1	BLVS	Bayesian lasso variable selection with λ corresponding to $\rho_t = \rho_{b_3}$
2	BLVS-G	Bayesian lasso variable selection with gamma hyperprior with $\alpha = 3$ and $u = 10$
3	NMIG	Normal mixture of inverse gamma (NMIG) prior (Scheipl 2010)
4	NG-RJ	Reversible jump with Normal-gamma (NG) prior (Griffin and Brown 2010)
5	HS-RJ	Reversible jump with the horseshoe (HS) prior (Carvalho et al. 2010)
6	NG	Full model using the Normal-gamma (NG) prior (Griffin and Brown 2010)
7	HS	Full model using the Horseshoe (HS) prior (Carvalho et al. 2010)
8	Lars	Lars-lasso algorithm with a 10-fold cross validation (CV) method to tune the shrinkage parameter

3 was implemented using `spikeSlabGAM`, 4–7 using `monomvn`, 8 using `Lars` packages in `R`

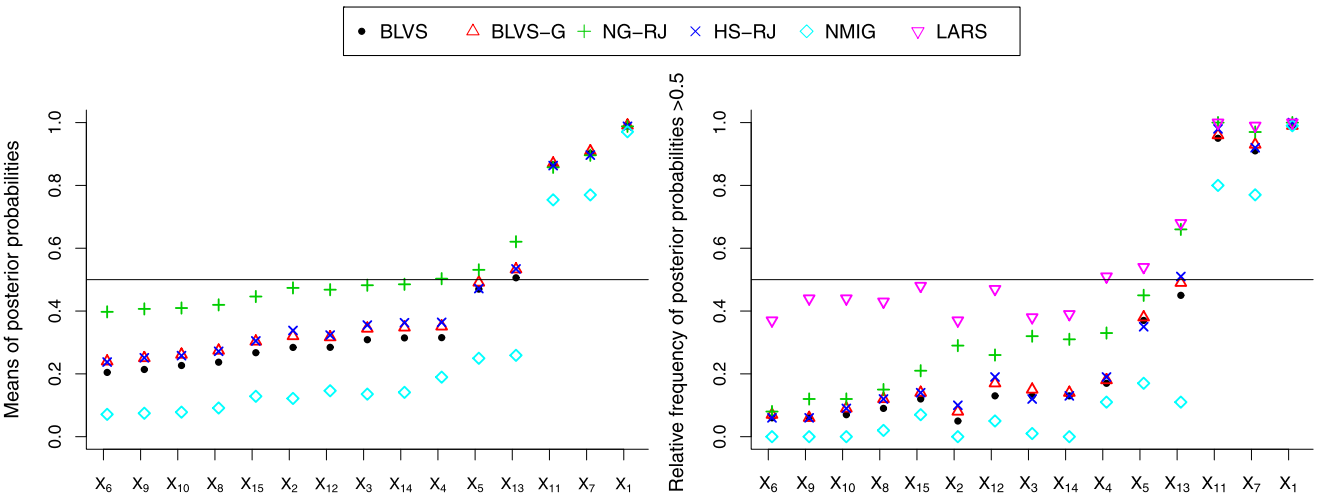
$\widehat{\beta}_j$: Posterior medians for methods 1,2, 4 and 5; Posterior means for 3,6 and 7

$\widehat{y}_i = \sum_{j=1}^{p} X_{ij}\widehat{\beta}_j$



All acronyms are available in Table 4; RMSEs are calculated using (16) with $\widehat{\beta}_j$ and $\widehat{y}_i$ described in Table 4.

Fig. 9 Boxplots of the root mean square errors (RMSEs) of the fitted values and the estimates of β over 100 generated datasets in simulation study 1 of Sect. 5.1.1



(a) Averages of $f(\gamma_j = 1|\boldsymbol{y})$ over 100 datasets

(b) Proportions with $f(\gamma_j = 1|\boldsymbol{y}) > 0.5$

All acronyms are available in Table 4; for Lars only the proportion of datasets with non-zero effects is depicted.

Fig. 10 Summaries of posterior inclusion probabilities $f(\gamma_j = 1|\boldsymbol{y})$ with $j = 1, \ldots, 15$ over 100 generated datasets in simulation study 1 of Sect. 5.1.1

the fitted values in contrast to their reversible jump versions which have the highest. In terms of RMSEs of the estimated β, NG now performs worse than the rest methods for which all RMSEs are almost indistinguishable.

Figure 10 displays (a) the averages (over the 100 generated datasets) of the posterior inclusion probabilities of each variable and (b) the proportion of datasets with posterior inclusion probabilities of each X_j greater than 0.5. From Fig. 10(a) we observe the following:

- Variables $\mathsf{X}_1, \mathsf{X}_7, \mathsf{X}_{11}$ are clearly identified as important ones by all implemented methods. The posterior inclusion probabilities for these covariates are identical for all methods (except for NMIG which are slightly lower).
- The posterior inclusion probabilities of NMIG are systematically lower than the rest of the methods for all covariates except for X_1 which is picked as the most important by all methods. For this reason this method systematically supports more parsimonious models.
- The posterior inclusion probabilities of NG-RJ are systematically higher than the rest of the methods. The difference seems to decrease with the importance of each covariate and is diminished for the three covariates that are indicated as important by all methods. Specifically, for the

non-important covariates, all posterior inclusion probabilities are close to 0.5, which considerably increase model uncertainty and force the method to support unnecessarily complicated models including redundant covariates.
- The proposed methods (BLVS and BLVS-G) and HS-RJ produce similar posterior inclusion probabilities (on average) for all covariates.

Similar is the picture and the conclusions from the proportions of datasets with posterior inclusion probabilities higher than 0.5. Differences between methods for all methods are minor for less important covariates ($\mathsf{X}_6, \mathsf{X}_8, \mathsf{X}_9$ and X_{10}) and are growing for the rest of the non-important covariates. In this figure, we also present the proportion of generated datasets with non-zero effects in Lars method. As we can observe, Lars included much more frequently non-important covariates in the model structure (in approximately 50% of the datasets) supporting models much more complicated than needed.

Concerning our proposed methods, both BLVS and BLVS-G identify correctly both the important covariates with posterior inclusion probabilities similar to the rest of the methods (except NMIG) and the non-important ones with low posterior probabilities (away from 0.5). With this behavior, both proposed methods support models of appropriate dimension including only the important covariates and identifying the best model with increased precision.

Finally, Table 5 presents the average number of correctly identified covariates using the criterion of the posterior inclusion probabilities being higher than 0.5. All methods identify on average 12–13 covariates correctly. More differences are observed on the false negative and positive selections with NMIG missing on average two covariates that should have been included in the model, while the rest of the methods select, on average, one additional covariate and miss one important one.

Table 5 Averages (over 100 generated samples) of the number of covariates that are correctly and falsely included and excluded from the linear predictor for the simulation study 1 of Sect. 5.1.1

	Correct	False Positive	False Negative
BLVS	12.72	0.91	1.37
BLVS-G	12.63	1.05	1.32
NMIG	12.52	0.29	2.19
NG-RJ	12.45	1.28	1.27
HS-RJ	12.73	0.90	1.37

Covariates are included in the model if the corresponding posterior inclusion probabilities > 0.5

False positive: a covariate is included in the model although its true effect is zero

False negative: a covariate is not included in the model although its true effect is non-zero

All acronyms are available in Table 4

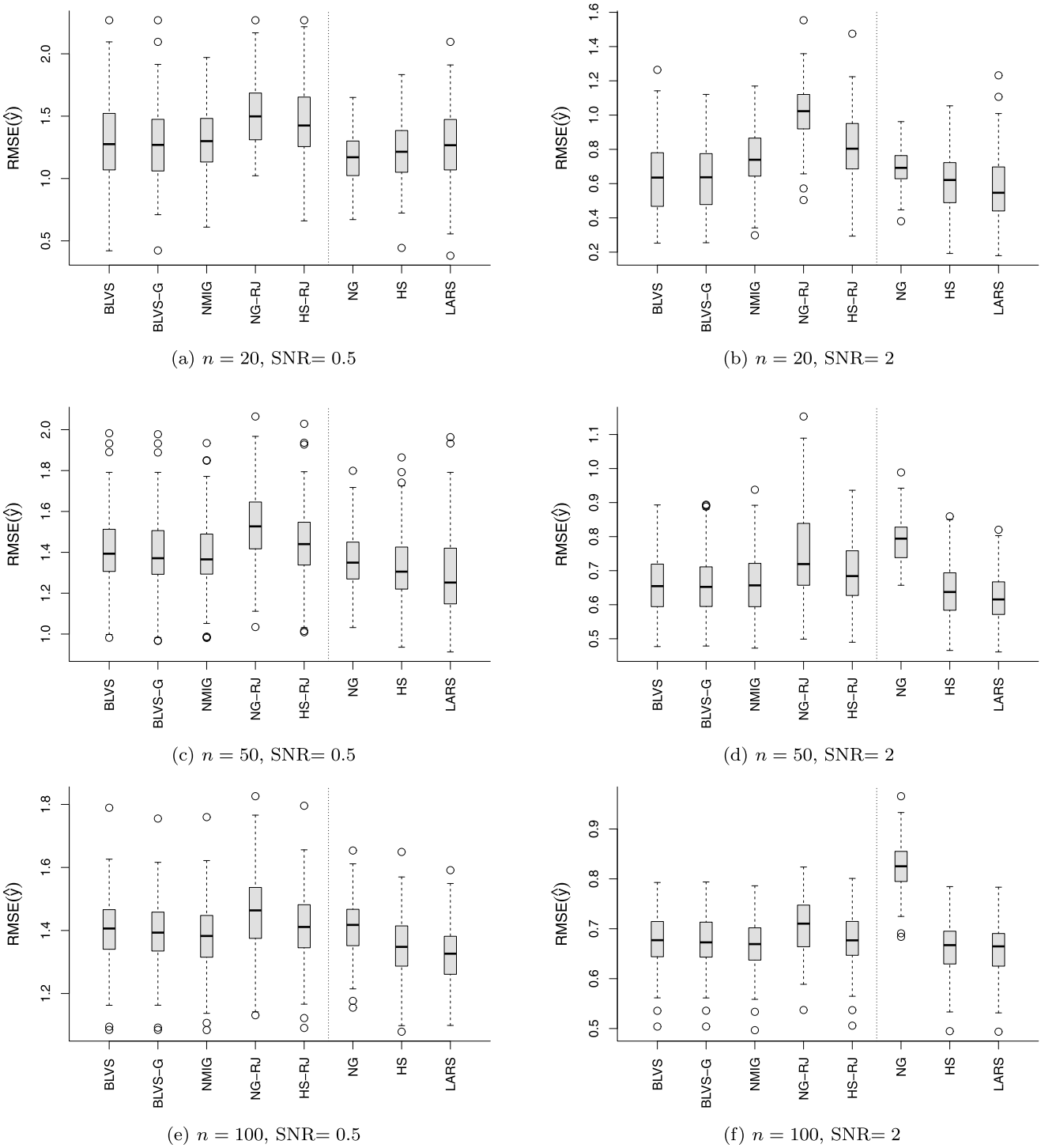
To sum up, both BLVS and BLVS-G perform satisfactorily achieving RMSEs equivalent to the rest of the methods under consideration. Both approaches select parsimonious models of appropriate dimension by setting the posterior inclusion for the non-important covariates low (around 0.30) and for the important covariates high (similarly to the rest of the methods).

5.2.2 *Simulation study 2*

Here, we illustrate and compare the implemented methods using a different (in perspective) simulation study. We use the simulation design of Scheipl (2010) where all 10 covariates have non-zero effects. Nevertheless, some of these effects are negligible not offering much in the model prediction. Therefore, all methods here are examined not for their ability to trace non-zero effects, but for their effectiveness in tracing 'well-fitted' parsimonious models that describe sufficiently the true generating mechanism. In particular, following Scheipl (2010), we generate ten independent covariates from Uniform$(-2, 2)$, which are then standardized to have zero mean and standard deviation of 0.5. The linear predictor $\eta = X\beta$ is then calculated using true effects from 0.1 to 1 changing with a step equal to 0.1; that is $\beta = (0.1, 0.2, \ldots, 1)$. Finally, the response is generated through $\mathsf{Y}_i \sim \mathrm{N}(\eta_i, \mathsf{sd}^2_\eta/\mathrm{SNR})$, where SNR is a selected signal-to-noise ratio and $\mathsf{sd}^2_\eta = \frac{1}{n-1}\sum_{i=1}^n(\eta_i - \overline{\eta})$ is the sample variance of the linear predictor with $\overline{\eta} = \frac{1}{n}\sum_{i=1}^n \eta_i$ denoting the corresponding sample mean. Following Scheipl (2010), we present results for two SNR values: 0.5 and 2 and three different sample sized $n = 20, 50$ and 100 observations.

Figures 11 and 12 present the boxplots of the RMSEs of the fitted values and the β estimates over 100 of simulated datasets. For RMSEs of the fitted values, we generally observe that our proposed methods (BLVS and BLVS-G) perform really well providing RMSEs of similar scale with most of the methods. Their RMSEs seem to improve when the signal is stronger and the size of the data grows bigger. NG and NG-RJ seem to perform systematically worst than the rest of the methods (surprisingly when the signal is strong and the size of the data bigger), while HS and Lars slightly better than the rest of them (which is expected due to the support of less parsimonious models).

In Fig. 12, the picture is less clear with more varying results between different methods especially, when information regarding the association of the covariates and the response is low due to small sample size and SNR. In particular, BLVS and BLVS-G seem to have higher RMSEs of $\widehat{\beta}$ than the rest of the methods, which is natural, since, the method is supporting more parsimonious models setting a large number of coefficients equal to zero (which here are



(a) $n = 20$, SNR$= 0.5$ (b) $n = 20$, SNR$= 2$

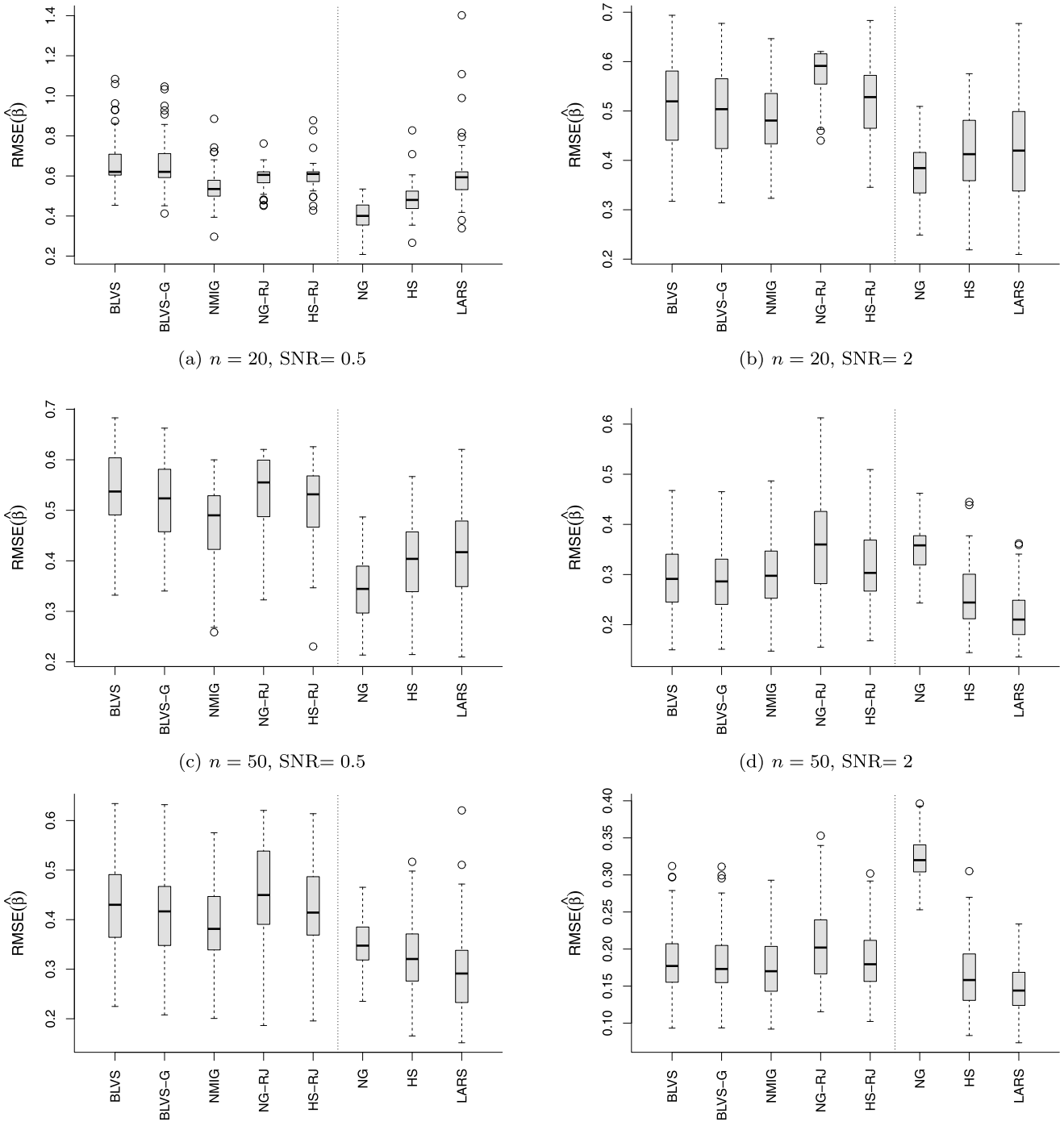
(c) $n = 50$, SNR$= 0.5$ (d) $n = 50$, SNR$= 2$

(e) $n = 100$, SNR$= 0.5$ (f) $n = 100$, SNR$= 2$

All acronyms are available in Table 4 RMSEs are calculated using (16) with $\widehat{\beta}_j$ and $\widehat{y}_i$ described in Table 4; SNR: signal-to-noise ratio; n: sample size.

Fig. 11 Root mean square error of the fitted values over 100 datasets for simulation study 2 of Sect. 5.2.2

non-zero). Nevertheless, this behavior is desirable for variable selection methods since, when no information is available (as here both SNR and sample size is small) less complicated theories (i.e. more parsimonious models) should be adopted. Note, that both these methods provide results close to the original Lars-lasso methodology. Moreover, due to the large variability of their RMSEs across different data (which is expected when low amount of information is available)



(a) $n = 20$, SNR$= 0.5$ (b) $n = 20$, SNR$= 2$

(c) $n = 50$, SNR$= 0.5$ (d) $n = 50$, SNR$= 2$

(e) $n = 100$, SNR$= 0.5$ (f) $n = 100$, SNR$= 2$

All acronyms are available in Table 4; RMSEs are calculated using (16) with $\widehat{\beta}_j$ and $\widehat{y}_i$ described in Table 4; SNR: signal-to-noise ratio; n: sample size.

Fig. 12 Root mean square error of the β estimates over 100 datasets for simulation study 2 of Sect. 5.2.2

we cannot clearly discriminate the performance of the methods (except for NG). On the contrary, NG seems to achieve clearly lower levels of RMSE than, at least, some of the remaining methods. This is due to the fact that this method considers non-zero values for all β_j even when information is low as here; see also Fig. 12(a). As the SNR and the sample size increase, differences between the methods are diminished with the exception of NG (in both versions)

that do not perform as efficiently as the remaining methods.

Figure 13 displays the medians (over 100 generated datasets) of the posterior estimates of β_j for all implemented methods along with the true β_j. The behavior of NG is strikingly different than the rest of the methods which, for low SNR and sample size, overshrink all coefficients. This behavior is expected since the overall information about the association of each covariate with the response is negligible. All methods converge towards the true effects (with different rate) as SNR and n increases. Especially, for $n = 100$ and SNR $= 2$ all methods (except NG) provide similar estimates and close to the true effects (also NG-RJ gives slightly smaller estimates than the rest of the methods). Generally, Lars-lasso seems to converge faster to the true coefficients when n or SNR increases. This can be also confirmed by previous analysis where we observed that Lars-lasso supports more complicated models than the rest of the methods. Hence, in this simulation study where all effects are non-zero, Lars-lasso moves faster towards the true model. On the other hand, the posterior estimates of β_j using NMIG converge slower towards the true effects indicating possibly that more shrinkage than necessary is implemented. Finally, our proposed methods (BLVS and BLVS-G) perform coherently setting small coefficients equal to zero and implementing small or negligible shrinkage on large effects (depending on information available and the size of the coefficient); see Figs. 13(d) and (f) for a clear picture of this behavior.

From Fig. 14 we can study the averages (over the simulated datasets) of the posterior inclusion probabilities for the methods implementing Bayesian variable selection. A first clear conclusion is that the posterior inclusion probabilities for NG-RJ and HS-RJ are very close (on average) on each other. Moreover, it is striking that for $n = 20$ and SNR $= 0.5$ (where low information about the model structure is available) these two methods give posterior probabilities around 0.5 for all covariates (with small deviations) while the remaining methods support more parsimonious models without suggesting the inclusion of any covariate. This behavior is retained also in the remaining of the graphs (where more information becomes available) for each covariate that is not supported with high posterior inclusion probabilities; see for example Fig. 14(f) with $n = 100$ and SNR $= 2$ where X_1 and X_2 with coefficients 0.1 and 0.2 respectively are not supported by any method but they have probabilities around 0.5 for NG-RJ and HS-RJ and much lower (<0.4) for BLVS, BLVS-G and NMIG. When sufficient information is available, all methods seem to converge (on average) on the same posterior inclusion probabilities for the important non-zero effects.

Another characteristic of our proposed methods (BLVS and BLVS-G) and of NMIG is that the averages of the posterior inclusion probabilities have bigger dispersion than the corresponding values of HS-RJ and NG-RJ making the discrimination of non-important and important covariates clearer. In particular, the increase in the standard deviation of the average posterior inclusion probabilities for BLVS and BLVS-G varies from ≈100% for Fig. 14(a), to ≈50% for 14(d) and ≈30% for 14(f) with all ranges being systematically higher for the proposed methods in all cases examined; for example, for Fig. 14(b) BLVS and BLVS-G ranges are 0.26–0.69 and 0.29–0.71 in contrast to 0.47–0.63 and 0.47–0.69 for HS-RJ and NG-RJ respectively.

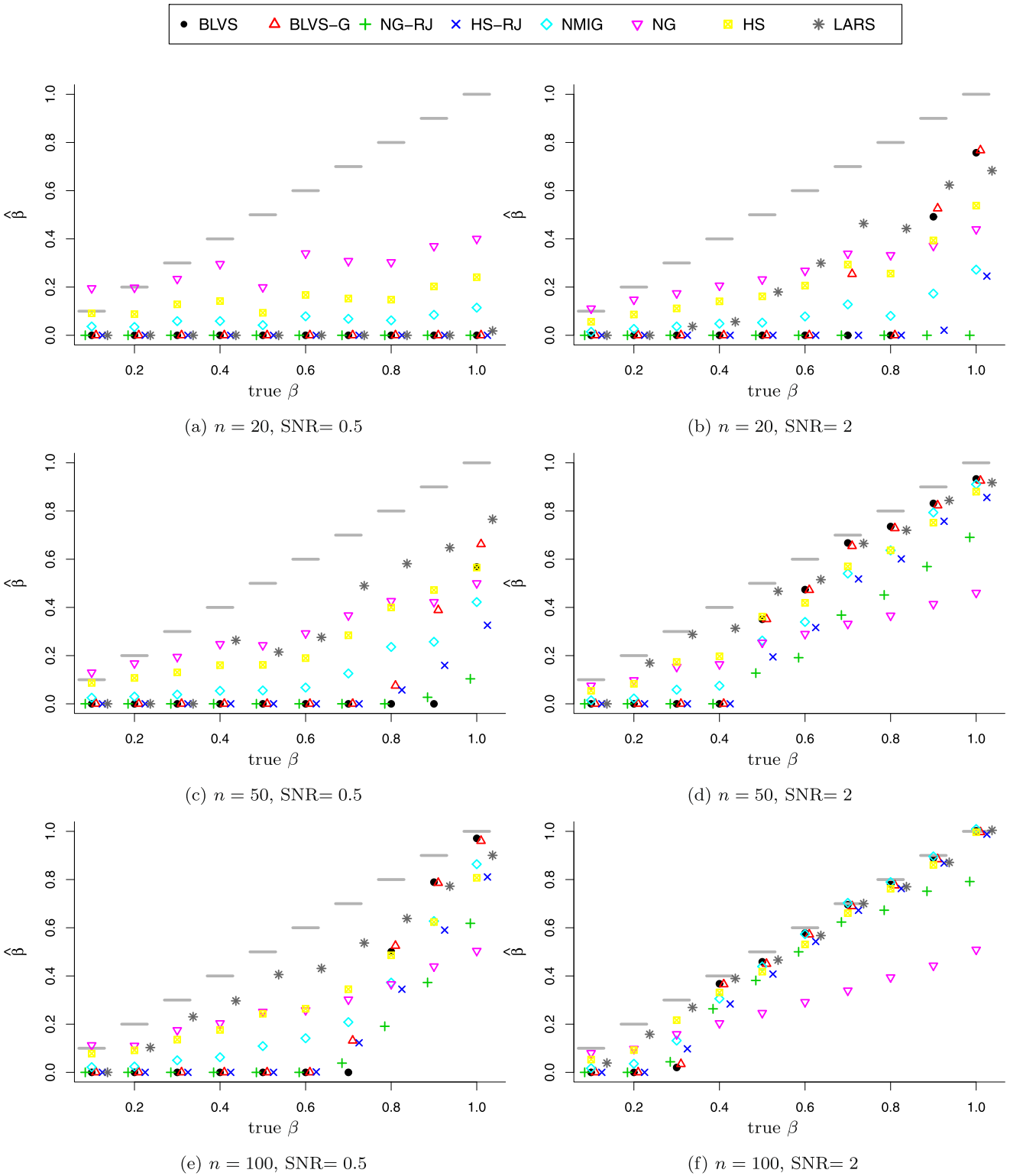
To conclude, our methods perform satisfactorily as variable selection methods identifying parsimonious "well-fitted" models. Covariates are not included in the model structure, unless enough information is provided by the data, overshrinkage and inflation (towards 0.5) of posterior inclusion probabilities for non-important or negligible covariate effects is avoided without loosing much in terms of RMSEs. Moreover, our methods seem to provide a clearer distinction between important and non-important covariates giving high probabilities to the first and low probabilities (much lower than 0.5) to the latter.

5.2.3 *Real example—diabetes dataset*

The diabetes dataset was previously used in lasso literature to illustrate the efficiency of the related methods and algorithms; see, for an example, in Efron et al. (2004). The response is a one year measurement of disease progression for 442 diabetes patients. The dataset also contains 10 baseline covariates: age, sex, body mass index (bmi), average blood pressure (bp) and six blood serum measurements (tc, ldl, hdl, tch, ltg, glu). Fitting linear regression models in data from such diagnostic studies is desirable for revealing the important determinants of the response as well as obtaining accurate predictions, as noted by Efron et al. (2004).

For BLVS we use λ corresponding to threshold correlation equal to ρ_{b_3} which here, for $n = 442$, is equal to 0.108. The hyperparameters in BLVS-G have been set at their default proposed values (i.e. $\alpha = 3$ and $u = 10$). We compare the performance of these methods with NMIG, NG, HS, and Lars using 20,000 updates after discarding 1,000 additional iterations as burn-in period for each MCMC. Posterior inclusion probabilities for all covariates are summarized in Table 6. From this table we observe that:

- "Age" as well the second, fourth and sixth blood serum measurements (ldl, thc and glu) have very low posterior inclusion probabilities in all the methods apart from some distinct cases (glu has probability equal to 0.79 in NG-RJ, ldl and tch have probabilities ≈ 0.55 in HS-RJ). These covariates were also indicated as the weakest predictors by Hans (2010). The sixth blood serum measurement (glu)



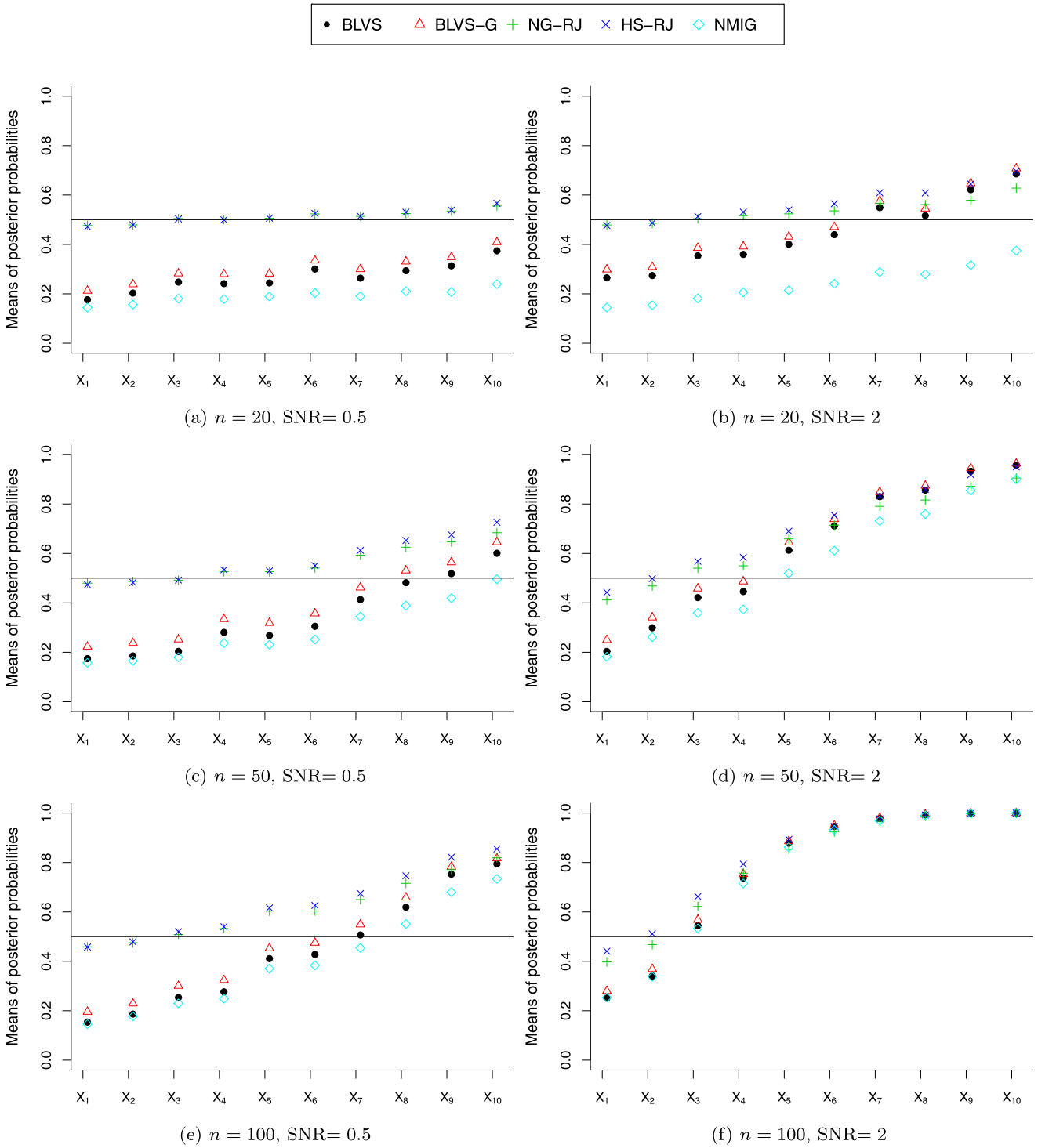
(a) $n = 20$, SNR$= 0.5$ (b) $n = 20$, SNR$= 2$

(c) $n = 50$, SNR$= 0.5$ (d) $n = 50$, SNR$= 2$

(e) $n = 100$, SNR$= 0.5$ (f) $n = 100$, SNR$= 2$

All acronyms are available in Table 4; $\widehat{\beta}_j$ for each method are defined in Table 4; Grey horizontal lines (on the diagonal of the graph) indicate true β_j; SNR: signal-to-noise ratio; n: sample size.

Fig. 13 Medians, over 100 simulated datasets, of the posterior estimates of β_j for simulation study 2 of Sect. 5.2.2



(a) $n = 20$, SNR$= 0.5$ (b) $n = 20$, SNR$= 2$

(c) $n = 50$, SNR$= 0.5$ (d) $n = 50$, SNR$= 2$

(e) $n = 100$, SNR$= 0.5$ (f) $n = 100$, SNR$= 2$

All acronyms are available in Table 4; Horizontal lines indicate value 0.5; SNR: signal-to-noise ratio; n: sample size.

Fig. 14 Averages, over 100 simulated datasets, of the posterior inclusion probabilities $f(\gamma_j = 1|\boldsymbol{y})$ for simulation study 2 of Sect. 5.2.2

picked by NG-RJ to be of relevant significance (probability 0.79) was also supported in Park and Casella (2008), Balakrishnan and Madigan (2010) and Li and Lin (2010).

- Covariates sex, body mass index, blood pressure and ltg are highly supported by all methods (posterior inclusion probabilities > 0.93).

Table 6 Posterior variable inclusion probabilities for the covariates of the diabetes dataset of Sect. 5.2.3

	age	sex	bmi	bp	Blood serum measurements					
					tc	ldl	hdl	tch	ltg	glu
BLVS	0.106	0.993	1.000	1.000	0.665	0.421	0.679	0.394	1.000	0.193
BLVS-G	0.121	0.990	1.000	1.000	0.680	0.459	0.682	0.413	1.000	0.222
NMIG	0.044	0.930	1.000	0.996	0.515	0.335	0.650	0.201	0.973	0.084
NG-RJ	0.370	0.980	1.000	1.000	0.448	0.428	0.995	0.484	1.000	0.786
HS-RJ	0.361	0.981	1.000	1.000	0.685	0.550	0.834	0.549	1.000	0.458

All method acronyms are available in Table 4; bmi: body mass index; bp: average blood pressure

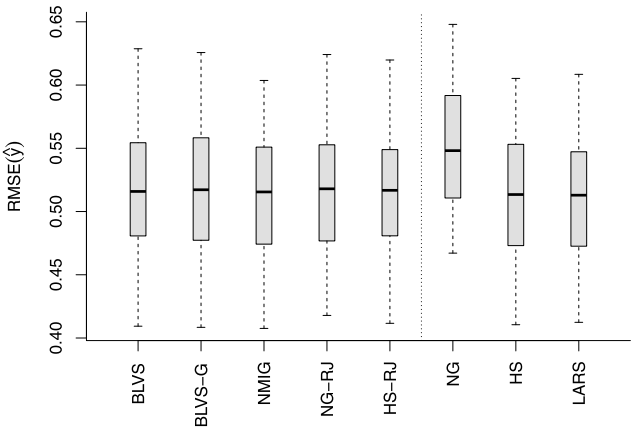


Fig. 15 Boxplots of root mean square errors of the fitted values for the diabetes dataset (Sect. 5.2.3) using 50 splits into training and test samples comprised by 70% and 30% respectively, of the total sample size

All acronyms are available in Table 4; RMSEs are calculated using (16) with $\widehat{\beta}_j$ and $\widehat{y}_i$ described in Table 4.

- The first and third blood measurements (tc and hdl) are relatively significant ranging from 0.45–0.68 and 0.65–0.99 respectively.
- BLVS and BLVS-G behave in a similar manner as NMIG but their probabilities are systematically higher. Non-important covariates are excluded with probabilities ranging from 0.10 (for age) to 0.46 (for ldl). Important covariates (sex, bmi, bp and ltg) are highly supported with posterior inclusion probabilities higher than 0.99 while the remaining two covariates (lc and hdl) are supported with probabilities around 0.68.
- HS-RJ is attributing posterior inclusion probabilities at the important covariates similar to BLVS and BLVS-G, while the non-important covariates are inflated towards 0.5 ranging from 0.36 (for age) to 0.55 (for ldl and tch).
- NG-RJ has different behavior than the rest of the methods confirming previous results. It assigns high posterior inclusion probabilities to covariates picked as important by the remaining methods and it additionally supports the inclusion of glu measurement (attributing probability 0.79). It supports more complicated models than the rest of the methods with the posterior inclusion probabilities of all non-important covariates inflated towards 0.5 (as in HS-RJ) and their values ranging from 0.37 (for age) to 0.48 (for tch).

Figure 15 displays boxplots of the RMSEs of the fitted values computed for 50 different splits of the original data into training and test sub-samples. Each training set consists of 309 patients (70% of the total sample) used to implement all methods under comparison while the remaining set of observation is used to estimate the prediction error. All RMSEs are notably close for all the methods, apart from NG that performs worse than the rest methods.

Table 7 shows the Pearson correlations between the response and the available covariates, as well as the corresponding partial correlations given that all the remaining covariates are included in the model and, in the third row, the partial correlations given the remaining covariates included in the MAP model of BLVS and BLVS-G. All covariates supported by these two methods have partial correlations higher than the selected threshold correlation ($\rho_t = 0.108$) or the corresponding range $(0.086, 0.125)$ of threshold correlations a posteriori supported by the BLVS-G approach using a gamma hyperprior for λ. Blood serum measurements tc and hdl have lasso partial correlation 0.09 and 0.02 respectively which are lower than the selected threshold correlation. However, they are finally included in the model since

Table 7 Absolute values of the observed Pearson and partial correlation coefficients for Diabetes dataset (Example 5.2.3)

	age	sex	bmi	bp	Blood serum measurements					
					tc	ldl	hdl	tch	ltg	glu
$\text{corr}(\boldsymbol{y}, \boldsymbol{X}_j)$	0.19	0.04	0.59	0.44	0.21	0.17	0.40	0.43	0.57	0.38
$\text{corr}^{(\text{lasso})}(\boldsymbol{y}, \boldsymbol{X}_j \vert \boldsymbol{X}_{\setminus j})$	0.01	**0.19**	**0.35**	**0.23**	**0.09**	0.07	**0.02**	0.05	**0.21**	0.05
$\text{corr}^{(\text{lasso})}(\boldsymbol{y}, \boldsymbol{X}_j \vert \boldsymbol{X}_{m_j^-})$[a]		**0.18**	**0.37**	**0.24**	**0.10**		**0.16**		**0.33**	

[a]Model m corresponds to the MAP model here and m_j^- to the model with covariates the ones included in the MAP except X_j

bmi: body mass index; bp: average blood pressure

their lasso partial correlations in the MAP model are 0.10 and 0.16 respectively.

6 Discussion

In this article, we presented a Bayesian lasso based model formulation which exploits the advantages of both shrinkage and variable selection methods. Shrinkage is attained via the use of a product of independent double exponential prior distributions for the regression coefficients while variable selection is achieved via the usual binary variable inclusion indicators included in the linear predictor. Estimation of the posterior distributions (including posterior model and variable inclusion probabilities) is achieved via a simple MCMC scheme. We also investigated the value of new regularization plots, which depict the behavior of the BMA based posterior summaries of regression coefficients, the Bayes factors and the variable inclusion probabilities for different values of the shrinkage parameter λ. These plots have motivated us to examine the behavior of unicovariate Bayes factors (which are available in closed form) and their relation with Pearson correlation measures. Following this lead, we have concluded to the definition of "benchmark" correlations. These measures identify the covariates that will never be supported strongly by the Bayes factor, evaluating evidence in favour of a simple regression versus the null model, whatever is the level of λ. We proceeded further, by defining the threshold correlation which identifies, for any given λ, the level or correlation at the limit between significant and insignificant covariates for this unicovariate Bayes factor. Then, we exploited this relation to define λ by specifying the desired level of threshold correlation. In this way, we achieve a simple and clear way to define the level of the shrinkage parameter λ. We have further examined which is the effect of this choice on nested multiple regression model comparisons, which evaluate the inclusion of a single covariate obtaining similar arguments based on partial correlations. We used these findings to interpret and understand the effect of our choice in nested multiple regression model comparisons or even specify λ.

We additionally constructed a gamma hyperprior with parameter values based on the threshold and benchmark correlations investigated in this article. This hyperprior eliminates any prior support to values of λ of no practical use such as the ones that activate Lindley-Bartlett paradox or overshrink important effects towards zero causing at the same time the posterior inclusion probabilities to be inflated towards 0.5 for the non-important effects. Results of our proposed methods (using fixed or varying λ values based on the methodology proposed in this article) are presented using two simulation studies and a real dataset. All results are compared with a variety of related methods previously introduced in lasso literature. Our Bayesian lasso methods work effectively in all examples tracing important effects with high posterior probabilities and eliminating non-important covariates with low posterior probabilities avoiding the Lindley-Bartlett paradox, overshrinkage of effects and inflation towards 0.5 of posterior inclusion probabilities of non-important effects without loosing in terms RMSEs.

The ideas presented in this work are more general and can be implemented in any Bayesian variable selection method. For example, it is interesting to see how ridge regression method (and its Bayesian analog) behaves and how we can specify prior parameters using similar arguments based on benchmark and threshold correlations. Another intriguing research direction, is to link the classical method of lasso with the Pearson and partial correlation limits between significance and insignificance. The existence of a relation between these values and the corresponding ones in the Bayesian approach may lead to the use of the simple lasso method for indirectly finding the MAP model or even produce reasonable approximations for posterior variable inclusion probabilities.

Extensions of this approach for generalized linear models, models for categorical data or for ANOVA models are also open issues that the authors intend to investigate in the near future.

Acknowledgements The authors would like to thank the anonymous referees, the associate editor and the editor for the constructive comments on the previous versions of this article. This research was funded by the Research Centre of the Athens University of Economics and Business.

Appendix

A.1 Details for the derivation of (11)

$$
\begin{aligned}
&\mathrm{var}_{\mathrm{lasso}}(\mathsf{Y}|\mathsf{X}) \\
&\quad = \mathrm{var}_{\mathrm{lasso}}\big(\mathsf{Y} - \mathsf{X}\boldsymbol{\beta}^{\mathrm{lasso}}\big) \\
&\quad = \mathrm{var}(\mathsf{Y}) + \mathrm{var}\big(\mathsf{X}\boldsymbol{\beta}^{\mathrm{lasso}}\big) - 2\mathrm{cov}\big(\mathsf{Y}, \mathsf{X}\boldsymbol{\beta}^{\mathrm{lasso}}\big) \\
&\quad = \mathrm{var}(\mathsf{Y}) + \boldsymbol{\beta}^{\mathrm{lasso}}\mathrm{var}(\mathsf{X})\boldsymbol{\beta}^{\mathrm{lasso}} - 2\mathrm{cov}(\mathsf{Y}, \mathsf{X})\boldsymbol{\beta}^{\mathrm{lasso}} \\
&\quad = \mathrm{var}(\mathsf{Y}) + \big(\boldsymbol{\beta}^{\mathrm{lasso}}\mathrm{var}(\mathsf{X}) - 2\,\mathrm{cov}(\mathsf{Y}, \mathsf{X})\big)\boldsymbol{\beta}^{\mathrm{lasso}} \\
&\quad = \mathrm{var}(\mathsf{Y}) - \big(\mathrm{cov}(\mathsf{Y}, \mathsf{X}) + k\boldsymbol{s}_{\boldsymbol{\beta}}^T\big)\mathrm{var}(\mathsf{X})^{-1}\big(\mathrm{cov}(\mathsf{X}, \mathsf{Y}) - k\boldsymbol{s}_{\boldsymbol{\beta}}\big) \\
&\quad = \mathrm{var}(\mathsf{Y}) - \mathrm{cov}(\mathsf{Y}, \mathsf{X})\mathrm{var}(\mathsf{X})^{-1}\mathrm{cov}(\mathsf{X}, \mathsf{Y}) \\
&\qquad + \mathrm{cov}(\mathsf{Y}, \mathsf{X})\mathrm{var}(\mathsf{X})^{-1}k\boldsymbol{s}_{\boldsymbol{\beta}} \\
&\qquad - k\boldsymbol{s}_{\boldsymbol{\beta}}^T\mathrm{var}(\mathsf{X})^{-1}\mathrm{cov}(\mathsf{X}, \mathsf{Y}) + k^2\boldsymbol{s}_{\boldsymbol{\beta}}^T\mathrm{var}(\mathsf{X})^{-1}\boldsymbol{s}_{\boldsymbol{\beta}} \\
&\quad = \mathrm{var}(\mathsf{Y}|\mathsf{X}) + k^2\boldsymbol{s}_{\boldsymbol{\beta}}^T\mathrm{var}(\mathsf{X})^{-1}\boldsymbol{s}_{\boldsymbol{\beta}}.
\end{aligned}
$$

A.2 Details for the derivation of (12) and (13)

From Definition 5 we can write

$$
\begin{aligned}
1 - R^{(\mathrm{lasso})2}_{\mathsf{Y}|\mathsf{X}} &= \frac{\mathrm{var}(\mathsf{Y}) - \mathrm{var}(X\boldsymbol{\beta}^{\mathrm{lasso}})}{\mathrm{var}(\mathsf{Y})} \\
&= \frac{\mathrm{var}(\mathsf{Y}) - (\mathrm{cov}(\mathsf{Y}, \mathsf{X}) - k\boldsymbol{s}_{\boldsymbol{\beta}}^T)\mathrm{var}(\mathsf{X})^{-1}\mathrm{var}(X)\,\mathrm{var}(\mathsf{X})^{-1}(\mathrm{cov}(\mathsf{X}, \mathsf{Y}) - k\boldsymbol{s}_{\boldsymbol{\beta}})}{\mathrm{var}(\mathsf{Y})} \\
&= \frac{\mathrm{var}(\mathsf{Y}) - \mathrm{cov}(\mathsf{Y}, \mathsf{X})\mathrm{var}(\mathsf{X})^{-1}\mathrm{cov}(\mathsf{X}, \mathsf{Y}) + 2k\boldsymbol{s}_{\boldsymbol{\beta}}^T\mathrm{var}(\mathsf{X})^{-1}\mathrm{cov}(\mathsf{X}, \mathsf{Y}) - k^2\boldsymbol{s}_{\boldsymbol{\beta}}^T\mathrm{var}(\mathsf{X})^{-1}\boldsymbol{s}_{\boldsymbol{\beta}}}{\mathrm{var}(\mathsf{Y})}
\end{aligned}
$$

From Corollary 5.5.2 of Whittaker (1990, p. 136), we have that $\mathrm{var}(\mathsf{Y}|\mathsf{X}) = \mathrm{var}(\mathsf{Y}) - \mathrm{cov}(\mathsf{Y}, \mathsf{X})\mathrm{var}(\mathsf{X})^{-1}\,\mathrm{cov}(\mathsf{X}, \mathsf{Y})$ resulting in

$$
\begin{aligned}
R^{(\mathrm{lasso})2}_{\mathsf{Y}|\mathsf{X}} &= 1 - \frac{\mathrm{var}(\mathsf{Y}|\mathsf{X}) + k\boldsymbol{s}_{\boldsymbol{\beta}}^T\,\mathrm{var}(\mathsf{X})^{-1}(2\,\mathrm{cov}(X, Y) - k\boldsymbol{s}_{\boldsymbol{\beta}})}{\mathrm{var}(\mathsf{Y})} \\
&= 1 - \frac{\mathrm{var}(\mathsf{Y}|\mathsf{X}) + k^2\boldsymbol{s}_{\boldsymbol{\beta}}^T\mathrm{var}(\mathsf{X})^{-1}\boldsymbol{s}_{\boldsymbol{\beta}} + 2k\boldsymbol{s}_{\boldsymbol{\beta}}^T\mathrm{var}(\mathsf{X})^{-1}(\mathrm{cov}(\mathsf{X}, \mathsf{Y}) - k\boldsymbol{s}_{\boldsymbol{\beta}})}{\mathrm{var}(\mathsf{Y})} \\
&= 1 - \frac{\mathrm{var}(\mathsf{Y}|\mathsf{X}) + k^2\boldsymbol{s}_{\boldsymbol{\beta}}^T\mathrm{var}(\mathsf{X})^{-1}\boldsymbol{s}_{\boldsymbol{\beta}} + 2k\boldsymbol{s}_{\boldsymbol{\beta}}^T\boldsymbol{\beta}^{\mathrm{lasso}}}{\mathrm{var}(\mathsf{Y})} \\
&= 1 - \frac{\mathrm{var}(\mathsf{Y}|\mathsf{X}) + k^2\boldsymbol{s}_{\boldsymbol{\beta}}^T\mathrm{var}(\mathsf{X})^{-1}\boldsymbol{s}_{\boldsymbol{\beta}} + 2k\|\boldsymbol{\beta}^{\mathrm{lasso}}\|}{\mathrm{var}(\mathsf{Y})}.
\end{aligned} \tag{17}
$$

From (11) we have that

$$
R^{(\mathrm{lasso})2}_{\mathsf{Y}|\mathsf{X}} = 1 - \frac{\mathrm{var}_{\mathrm{lasso}}(\mathsf{Y}|\mathsf{X}) + 2k\|\boldsymbol{\beta}^{\mathrm{lasso}}\|}{\mathrm{var}(\mathsf{Y})},
$$

which is the expression (12).

Finally, substituting $R^2_{\mathsf{Y}|\mathsf{X}} = 1 - \frac{\mathrm{var}(\mathsf{Y}|\mathsf{X})}{\mathrm{var}(\mathsf{Y})}$ back in (17) gives us the result of (13).

Proof of Corollary 1 The $R^2_{\mathsf{Y}|\mathsf{X}}$ of the lasso regression is related with the ordinary multiple correlation through

$$
\begin{aligned}
R^{(\mathrm{lasso})2}_{\mathsf{Y}|\mathsf{X}} &= 1 - \mathrm{var}_{\mathrm{lasso}}(\mathsf{Y}|\mathsf{X}) - 2k\big\|\boldsymbol{\beta}^{\mathrm{lasso}}\big\|_1 \\
&= 1 - \mathrm{var}(\mathsf{Y}|\mathsf{X}) - k^2\boldsymbol{s}_{\boldsymbol{\beta}}^T\mathrm{var}(\mathsf{X})^{-1}\boldsymbol{s}_{\boldsymbol{\beta}} - 2k\big\|\boldsymbol{\beta}^{\mathrm{lasso}}\big\|_1 \\
&= R^{(\mathrm{ols})2}_{\mathsf{Y}|\mathsf{X}} - k^2\boldsymbol{s}_{\boldsymbol{\beta}}^T\,\mathrm{var}(\mathsf{X})^{-1}\boldsymbol{s}_{\boldsymbol{\beta}} - 2k\big\|\boldsymbol{\beta}^{\mathrm{lasso}}\big\|_1.
\end{aligned}
$$

Hence, $R^{(\mathrm{lasso})2}_{\mathsf{Y}|\mathsf{X}} \leq R^{(\mathrm{ols})2}_{\mathsf{Y}|\mathsf{X}} \leq 1$, which also implies that $R^{(\mathrm{lasso})2}_{\mathsf{Y}|\mathsf{X}}$ cannot exceed 1. □

Proof of Theorem 1 We consider the Laplace approximation of the unicovariate BF, which gives

$$
\mathrm{LBF}^{\mathrm{un}}_j = ck\big[1 - (\rho_j - k\boldsymbol{s}_{\widehat{\boldsymbol{\beta}}})^2\big]^{-df/2}, \tag{18}
$$

where ρ_j is sample Pearson correlation between Y and X_j. Equating (15) and (18), the threshold values of the Pearson and partial correlations satisfy the following

$$
\begin{aligned}
&c\big[1-(\rho_j - k s_{\widehat{\boldsymbol{\beta}}})^2\big]^{-df/2} \\
&\quad = c_{\mathrm{mu}}\big(1-\rho^2_{\boldsymbol{y},\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}}\big)^{-df_{\mathrm{mu}}/2} \\
&\qquad \times\big[\big(1-R^{(\mathrm{ols})2}_{\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}}\big)\big(1-R^{(\mathrm{lasso})2}_{\boldsymbol{y}|\boldsymbol{X}_{m_j^-}}\big)\big]^{-1/2} \\
&1-(\rho_j - k s_{\widehat{\boldsymbol{\beta}}})^2 \\
&\quad = \left(\frac{c}{c_{\mathrm{mu}}}\right)^{2/df}\big(1-\rho^2_{\boldsymbol{y},\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}}\big)^{df_{\mathrm{mu}}/df} \\
&\qquad \times\big[\big(1-R^{(\mathrm{ols})2}_{\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}}\big)\big(1-R^{(\mathrm{lasso})2}_{\boldsymbol{y}|\boldsymbol{X}_{m_j^-}}\big)\big]^{1/df} \qquad (19)
\end{aligned}
$$

where $\rho_{\boldsymbol{y},\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}} = \mathrm{corr}^{(\mathrm{lasso})}(\boldsymbol{y},\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-})$.

The ratio $\frac{c}{c_{\mathrm{mu}}}$ is approximately equal to $(\frac{n}{n+p-1})^{1/2} \le 1$ for large n and thus, $(\frac{c}{c_{\mathrm{mu}}})^{2/df} \le 1$. Moreover,

$$
\begin{aligned}
&\big[\big(1-R^{(\mathrm{ols})2}_{\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}}\big)\big(1-R^{(\mathrm{lasso})2}_{\boldsymbol{y}|\boldsymbol{X}_{m_j^-}}\big)\big]^{1/df} \\
&\quad \le \big(1-R^{(\mathrm{ols})2}_{\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}}\big)\big(1-R^{(\mathrm{lasso})2}_{\boldsymbol{y}|\boldsymbol{X}_{m_j^-}}\big) \le 1,
\end{aligned}
$$

since $R^{(\mathrm{ols})2}_{\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}}$ and $R^{(\mathrm{lasso})2}_{\boldsymbol{y}|\boldsymbol{X}_{m_j^-}}$ lie in the [0, 1] interval. Using these two inequalities in (19), we obtain that $(1-\rho^2_{\boldsymbol{y},\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}})^{df_{\mathrm{mu}}/df} \le 1-\rho^2_{\boldsymbol{y},\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}}$ finally concluding to

$$
\rho^2_{\boldsymbol{y},\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}} \le (\rho_j - k s_{\widehat{\boldsymbol{\beta}}})^2. \qquad \square
$$

Proof of Corollary 2 The proof of Corollary 2, immediately follows Theorem 1 if we set $\mathrm{LBF}^{\mathrm{un}}_j = \mathrm{LBF}^{\mathrm{mu}}_{m,j} = 1$. $\square$

Proof of Corollary 3 For $n \to \infty$, in (19) we have that $k = \lambda/(n-1) \to 0$, $(\rho_j - k s_{\widehat{\boldsymbol{\beta}}}) \to \rho_j$, $df \to \infty$, $df_{\mathrm{mu}}/df \to 1$ and $(\frac{c}{c_{\mathrm{mu}}})^{2/df} \to 1$. Returning back to (19), for large n we obtain that $\rho^2_{\boldsymbol{y},\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}} \approx \rho^2_j$. $\square$

References

Armagan, A., Dunson, D., Lee, J.: Bayesian generalized double Pareto shrinkage. arXiv:1104.0861v3 [stat.ME] (2012)

Balakrishnan, S., Madigan, D.: Priors on the variance in sparse Bayesian learning: the demi-Bayesian lasso. In: Chen, M.-H., Muller, P., Sun, D., Ye, K. (eds.) Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger, pp. 346–359. Springer, Berlin (2010)

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

Carvalho, C., Polson, N., Scott, J.: The horseshoe estimator for sparse signal. Biometrika **97**, 465–480 (2010)

Dellaportas, P., Forster, J., Ntzoufras, I.: On Bayesian model and variable selection using MCMC. Stat. Comput. **12**, 27–36 (2002)

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Stat. **32**, 407–499 (2004)

Fahrmeir, L., Kneib, T., Konrath, S.: Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. Stat. Comput. **20**, 203–219 (2010)

George, E., McCulloch, R.: Variable selection via Gibbs sampling. J. Am. Stat. Assoc. **88**, 881–889 (1993)

Gramacy, R.: monomvn: Estimation for multivariate normal and Student-t data with monotone missingness. R package version 1.8-3 (2010)

Green, P.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82**, 711–732 (1995)

Griffin, J.E., Brown, P.J.: Inference with normal-gamma prior distributions in regression problems. Bayesian Anal. **5**, 171–188 (2010)

Hans, C.: Bayesian Lasso regression. Biometrika **96**, 835–845 (2009)

Hans, C.: Model uncertainty and variable selection in Bayesian lasso regression. Stat. Comput. **20**, 221–229 (2010)

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

Johnson, B.: On lasso for censored data. Electron. J. Stat. **3**, 485–506 (2009)

Kass, R., Raftery, A.: Bayes factors. J. Am. Stat. Assoc. **90**, 773–795 (1995)

Kuo, L., Mallick, B.: Variable selection for regression models. Sankhyā B **60**, 65–81 (1998)

Li, Q., Lin, N.: The Bayesian elastic net. Bayesian Anal. **5**, 847–866 (2010)

Lindley, D.: A statistical paradox. Biometrika **44**, 187–192 (1957)

Lykou, A., Whittaker, J.: Sparse canonical correlation analysis by using the lasso. Comput. Stat. Data Anal. **54**, 3144–3157 (2010)

Meier, L., Van de Geer, S., Bühlmann, P.: The group lasso for logistic regression. J. R. Stat. Soc. B **70**, 53–71 (2008)

Nott, D., Kohn, R.: Adaptive sampling for Bayesian variable selection. Biometrika **92**, 747–763 (2005)

Ntzoufras, I.: Bayesian Modeling Using WinBugs. Wiley, New York (2009)

Osborne, M.R., Presnell, B., Turlach, B.A.: On the lasso and its dual. J. Comput. Graph. Stat. **9**, 319–337 (2000)

Park, M.Y., Hastie, T.: l_1 regularization path algorithm for generalized linear models. J. R. Stat. Soc. B **69**, 659–677 (2006)

Park, T., Casella, G.: The Bayesian lasso. J. Am. Stat. Assoc. **103**, 681–687 (2008)

R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2011)

Scheipl, F.: Normal-mixture-of-inverse-gamma priors for Bayesian regularization and model selection in structured additive regression models. Technical Report 84, Department of Statistics, University of Munich (2010); available at http://epub.ub.uni-muenchen.de/11785/

Scheipl, F.: Spikeslabgam: Bayesian variable selection, model choice and regularization for generalized additive mixed models in R. J. Stat. Softw. **43**(14), 1–24 (2011)

Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B **58**, 267–288 (1996)

Tibshirani, R.: The lasso method for variable selection in the cox model. Stat. Med. **16**, 385–395 (1997)

Whittaker, J.: Graphical Models in Applied Multivariate Statistics. Wiley, New York (1990)

Yuan, M., Lin, Y.: Efficient empirical Bayes variable selection and estimation in linear models. J. Am. Stat. Assoc. **100**, 1215–1225 (2005)

Zellner, A.: On assessing prior distributions and Bayesian regression analysis using g-prior distributions. In: Goel, P., Zellner, A. (eds.) Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, pp. 233–243. North-Holland, Amsterdam (1986)

Zhang, C.-H., Huang, J.: The sparsity and bias of the lasso selection in high-dimensional linear regression. Ann. Stat. **36**, 1567–1594 (2008)

Zou, H.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**, 1418–1429 (2006)

Zou, J., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. B **67**, 301–320 (2005)