On Bayesian Lasso Variable Selection and the Specification of the Shrinkage Parameter

Anastasia Lykou and Ioannis Ntzoufras

Abstract

In this work, 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. By this way, a meaningful interpretation of λ is achieved that leads to a simple specification of this parameter which is of prominent importance in Lasso literature.

Keywords: Bayes factors; MCMC; Partial Correlation; Pearson Correlation; Shrinkage; Benchmark and Threshold Correlations.

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

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

for the usual regression model

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

where $\text{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution of dimension n with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, Y is a $n \times 1$ vector of the random responses, $\boldsymbol{y}$ is the corresponding vector of observed response values, $\boldsymbol{X}$ is the $n \times p$ design or data matrix with elements X_{ij} corresponding to i individual and j variable, $\boldsymbol{\beta}$ is a $n \times 1$ vector with elements the coefficients β_j of each X_j covariate, σ^2 is the error variance of the regression model, λ is the shrinkage parameter of Lasso, $\boldsymbol{s}_{\boldsymbol{\beta}}$ is a vector with elements the sign of each $\widehat{\beta}_j^{\text{lasso}}$ and $\boldsymbol{I}_n$ being 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 $\beta \sim DE(\mu, b)$ 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 $\beta_j \sim DE(0, \sigma^2/\lambda)$ will produce a posterior distribution that will be maximized under the Lasso estimates (1).

For this reason, a wide variety of Bayesian Lasso methods have been 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 (2009) 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 the double-exponential prior and sample the posterior inclusion probabilities by using a Gibbs sampler.

Griffin and Brown (2010) also investigate the Bayesian Lasso by imposing normal-gamma prior and a data depended hyperprior on 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.

Finally, the last years, pure Bayesian shrinkage methods have been also received attention in the statistical community resulting in the introduction of other prior distributions such as the horse-shoe prior (Carvalho et al. 2010) and the double generalized Parero (Armagan et al. 2010). All these approaches try to overshrink small coefficients and leave unaffected large ones (similar to lasso) but also have additional 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 faster while less shrinkage is applied for large coefficients. This is due to the diamond shaped restriction area that Lasso implements on $(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$ 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 Figure 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.

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 a 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. Usually cross-validation techniques are used to provide a sensible value for this parameter. Nevertheless, the specification of the shrinkage coefficient highly depends on the choices we make without having a solid universal methodology for defining this parameter.

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. Theses estimates will approach zero slowly but they will be never 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 prior, makes the evaluation of the posterior distribution less straightforward requiring the use of MCMC methods.

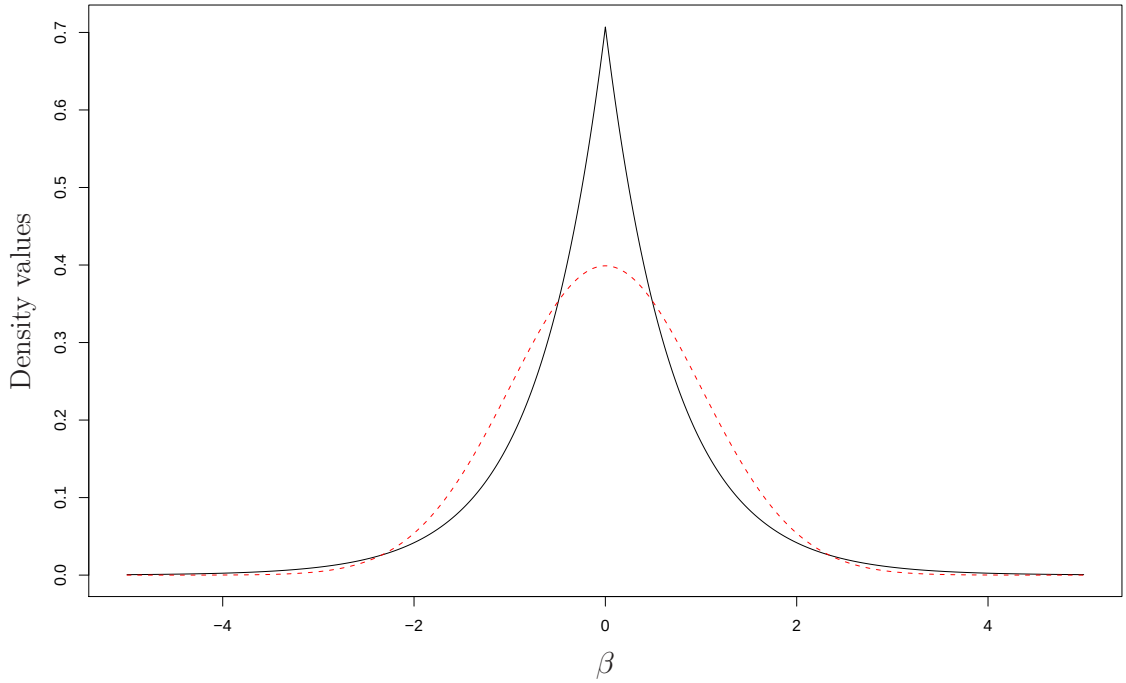

Figure 1: Plot of the density functions for the double exponential (solid line) and normal (dashed line) distributions with zero mean and variance equal to one.

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 covariance 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 that the number of predictors is smaller than the number of observations but we strongly believe that the method will efficiently find practical implementations in large p small n problems. 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 their effect on posterior model probabilities and Bayes factors. By investigating the behaviour and the sensitivity

of these measures on the choice of λ we obtain a simple and meaningful interpretation of its effect. By this way, we can a-priori specify the shrinkage level and control the variable selection procedure. With this analysis, we have traced a range of correlation values for which covariates will be never be included in the model structure whatever is the value of the shrinkage parameter is. Moving further, we specify λ by defining the levels of correlation measures that produce Bayes factors that cannot discriminate between nested models that differ by a specific covariate X_j.

The article is organized as follows. In Section 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 partials correlation coefficients. In particular, Section 3.1 focuses on the unicoviate Bayes factor comparing each simple regression model with the null and their association with the Pearson correlation coefficient. We examine and interpret this association using extensive graphical representations. In Section 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. In Section 4 we illustrate our method using two simulation studies (with low and high correlated covariates, respectively) and a real dataset. 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

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

where $\tau = 1/\sigma^2$ is the precision of the Normal regression model, $\mathrm{Bernoulli}(\pi)$ is the Bernoulli distribution with success probability π and $\mathrm{Gamma}(a, d)$ is the gamma distribution with

mean a/d and variance a/d^2.

The prior specification in the formulation above 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

$$\beta_j|\boldsymbol{y}, \tau, \boldsymbol{\beta}_{\setminus j}, \boldsymbol{\gamma}_{\setminus j}, \gamma_j = 0 \quad \sim \text{DE}\Big(0, \frac{1}{\tau\lambda}\Big) \tag{3}$$

$$\beta_j, \omega_j|\boldsymbol{y}, \tau, \boldsymbol{\beta}_{\setminus j}, \boldsymbol{\gamma}_{\setminus j}, \gamma_j = 1, \sim \omega_j\, TN(\mu_j^-, s_j^2, \beta_j < 0) + (1-\omega_j)\, TN(\mu_j^+, s_j^2, \beta_j \geq 0), \tag{4}$$

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 $A \subset \Re$ with density function

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

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

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

and

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

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

$$\mu_j^- = \frac{c_j + \lambda}{||\boldsymbol{X}_j||^2}, \quad \mu_j^+ = \frac{c_j - \lambda}{||\boldsymbol{X}_j||^2}, \quad c_j = \boldsymbol{X}_j^T(\boldsymbol{e} + \beta_j \boldsymbol{X}_j) \;\text{ and }\; s_j^2 = \frac{1}{\tau||\boldsymbol{X}_j||^2}.$$

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 $||\boldsymbol{z}||^2 = \sum_{i=1}^n z_i^2$ and $||\boldsymbol{z}|| = \sum_{i=1}^n |z_i|$ for any vector $\boldsymbol{z}$ 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

$$\begin{aligned} w_j &= \mathsf{P}(\omega_j = 1 | \, \boldsymbol{y}, \tau, \boldsymbol{\beta}_{\setminus j}, \boldsymbol{\gamma}_{\setminus j}, \gamma_j = 1) = \mathsf{P}(\beta_j < 0 \, | \, \boldsymbol{y}, \tau, \boldsymbol{\beta}_{\setminus j}, \boldsymbol{\gamma}_{\setminus j}, \gamma_j = 1) \\ &= \frac{\Phi(-\mu_j^-/s_j)/f_N(0\,;\mu_j^-, s_j^2)}{\Phi(-\mu_j^-/s_j)/f_N(0\,;\mu_j^-, s_j^2) + \Phi(\mu_j^+/s_j)/f_N(0\,;\mu_j^+, s_j^2)}. \end{aligned} \tag{5}$$

Hence, when $\gamma_j = 1$ we generate

- Generate ω_j from a Bernoulli with success probability w_j given by (5).
- If $\omega_j = 1$ generate β_j from $TN(\mu_j^-, s_j^2, \beta_j < 0)$ otherwise from $TN(\mu_j^+, s_j^2, \beta_j \geq 0)$.

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

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

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

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

2.3 Regularization Plots for Bayesian Lasso Variable Selection

Using the Gibbs sampler described in Section 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 Bayesian model averaged (BMA) versions $\beta_j^* = \gamma_j\beta_j$ of the effect of each covariate X_j. For the later two quantities, we will examine their behaviour 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 in 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

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

The proposed method is performed on this dataset for different values of λ, $\pi_j = 0.5$ for all j, $a = d = 10^{-4}$ and we consider 10000 updates after discarding additional one thousand iterations as burn-in period.

In Figure 2(a), the posterior means of $\beta_j^* = \gamma_j\beta_j$ are plotted against the values of λ while in Figure 2(b) the usual regularization plot of the Lasso estimates is depicted. As it is obvious from both of these plots, λ controls the shrinkage applied on each β_j: for $\lambda \to 0$ no shrinkage is implemented while as λ increases the coefficients are shrunk to zero.

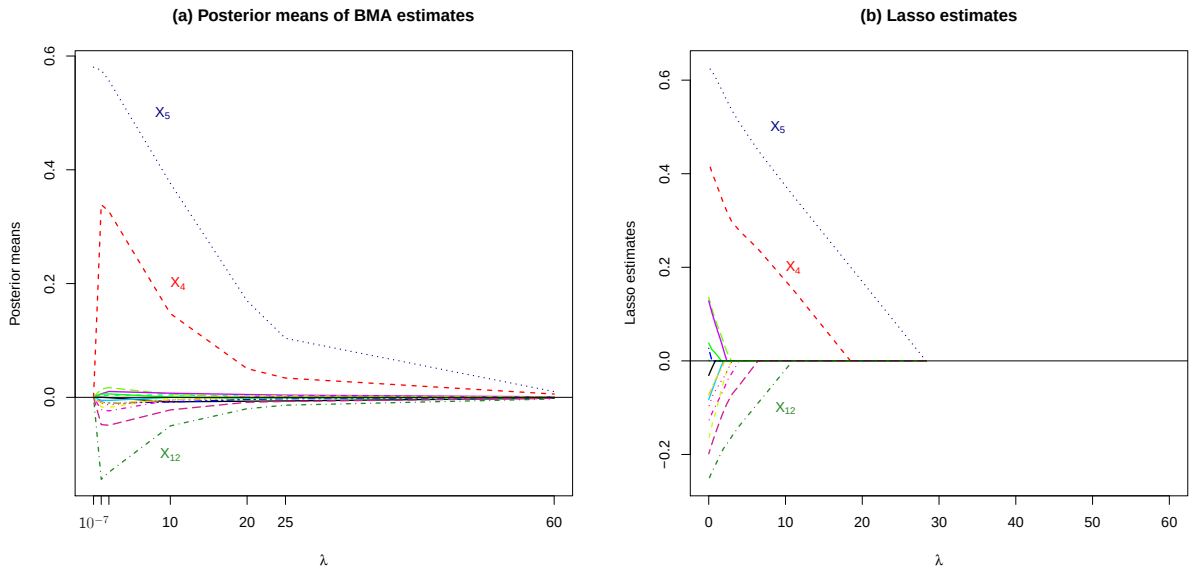

Figure 2: Regularization plots for the posterior means of $\beta_j^* = \gamma_j \beta_j$ and usual Lasso estimates against λ.

The Lasso estimates (derived by the Lars algorithm) approach the ordinary least squares estimates as λ approaches zero. Similarly, for very low values of λ (expressing high prior ignorance), we would expect the BMA posterior means of β_j^* to approach the MLE estimates. Nevertheless, due to the Lindley-Bartlett paradox (Lindley, 1957, Bartlett, 1957), this is not true since small values of λ (corresponding to large prior variance of β_j) activate the effect of the 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 Figure 3(a). Posterior medians become exactly equal to zero when $P(\gamma_j = 1|\boldsymbol{y}) < 0.5$ in contrast to the posterior means which will be small but non-zero unless $P(\gamma_j = 1|\boldsymbol{y}) = 0$. Hence, in the plot of the posterior medians, non-important variables are eliminated from the plot for all values of λ.

The second plot of Figure 3 (on the right) shows the posterior probabilities of $P(\gamma_j = 1|y)$ as a function of λ. As a result of the Lindley-Bartlett paradox, the posterior probabilities of including a variable in the model tends to zero for $\lambda \to 0$. The posterior probabilities of the unimportant variables approach the value of 0.5 as λ moves away from zero. The behaviour 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

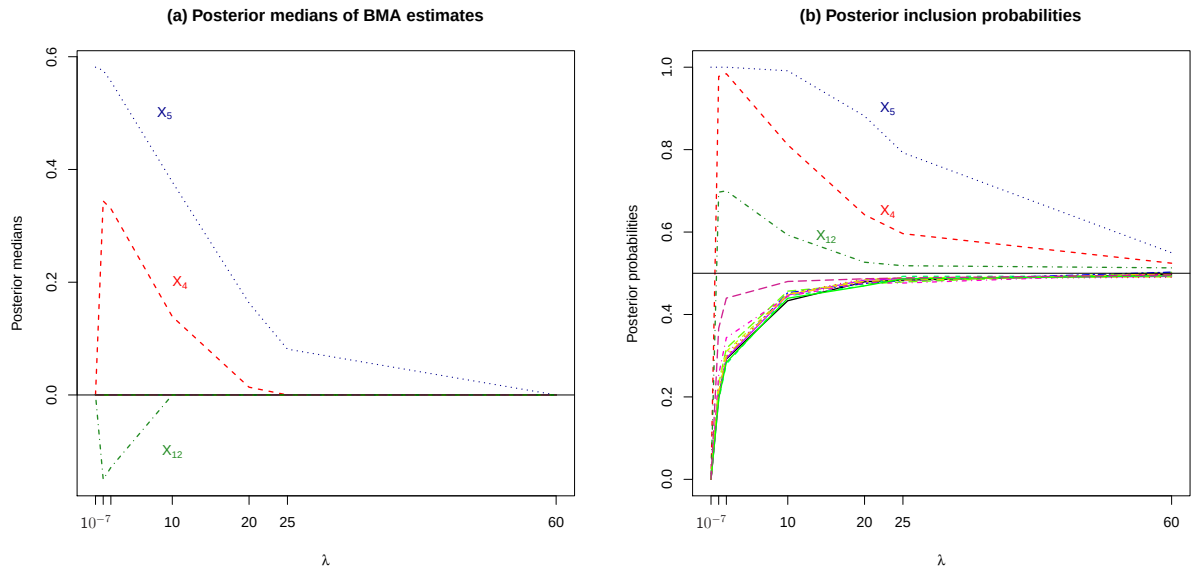

Figure 3: Regularization plots for the posterior medians of $\beta_j^* = \gamma_j \beta_j$ and the posterior variable inclusion probabilities $P(\gamma_j = 1|\boldsymbol{y})$ against λ.

model formulation under consideration.

Even in these initial illustrations, the proposed method seems to offer a very challenging approach to perform 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 are eliminated for all values of the shrinkage parameter λ. Moreover, the behaviour of the posterior inclusion probabilities for large and small values of λ can motivate the restriction of the sensible values of λ to avoid over-shrinkage (when λ is large) or the Lindley-Bartlett paradox (when λ is small). Using these observations as a starting point, in Section 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 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 will 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 covariate X_j. We will call these Bayes factors as unicovariate and facilitate results based on these simple comparisons to identify reasonable values for the choice of λ.

Definition 1 (Unicovariate Bayes Factor $BF_j^{\mathbf{un}}$) *The unicovariate Bayes factor BF_j^{un} for covariate X_j is defined as the Bayes factor that evaluates the evi-*

dence of model m_j *versus* m_0 *with*

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

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

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

where

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

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:

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

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

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

In order to interpret the behaviour of the unicovariate Bayes factors, we present their logarithms in Figure 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

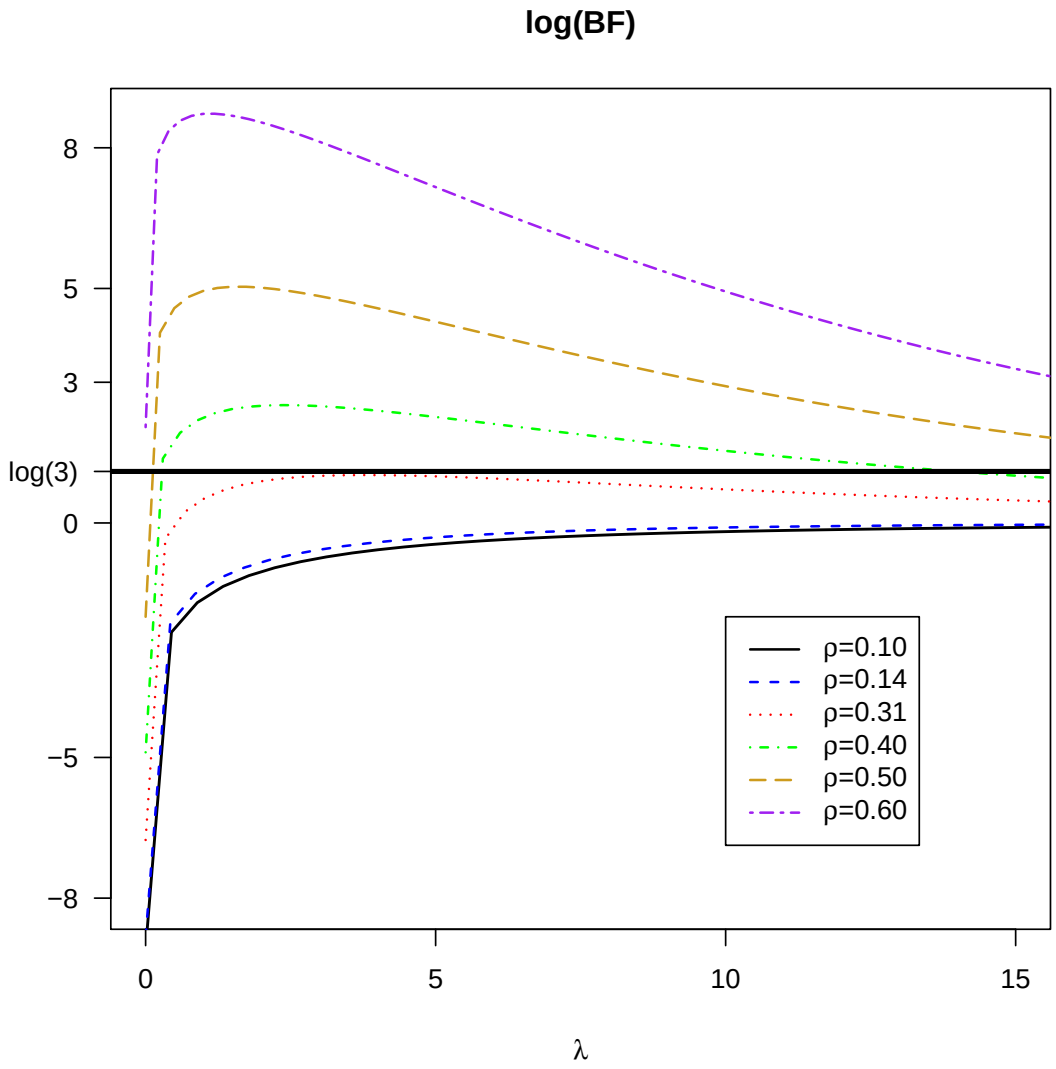

Figure 4: Logarithm $\mathrm{BF}_j^{\mathrm{un}}$ against λ for several values of the Pearson correlation coefficient ρ ([1]Sample size is fixed to $n = 50$; [2]$\mathrm{BF}_j^{\mathrm{un}}$ is the Bayes factor of model m_j (with variable $\boldsymbol{X}_j$) versus model m_0 (constant)).

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 be derived for sample of different size.

3.2 Specification of the shrinkage parameter λ

Several approaches for tuning the shrinkage levels have been proposed in the literature 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 (Eq. 9), its relation to the Pearson's sample correlation and its behaviour as illustrated in Section 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 Figure 4, we observe that, for specific values of ρ, the BF_j^{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 $\rho = 0.31$ or lower. Thus we can identify a range of sample correlations corresponding to covariates that will be never considered as "important" determinants of the response for all values of λ and fixed n.

A graphical representation of the BF_j^{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. Thus we define the non-important set of correlations using Definition 2.

> **Definition 2 (Non-important set of correlations $\mathcal{I}$)** *The "non-important" set of correlations is the set of correlations that correspond to covariates with univariate Bayes factors less than* 3 *for all possible shrinkage values λ, i.e. $\mathcal{I} = \{\rho : BF_j^{un} \leq 3$ for all $\lambda > 0\}$*

Moreover, we specify the benchmark correlation using the Definition which follows.

> **Definition 3 (Benchmark correlation ρ_b)** *The benchmark correlation ρ_b is defined as the maximum value in the "non-important" set of correlations $\mathcal{I}$. All the covariates with correlation less than this ρ_b will not be supported strongly enough by the corresponding unicovariate BFs for any shrinkage value λ.*

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

In Section 3.2.1 we identified which covariates will be never 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 Equation 10, for any given value of λ, we can identify a specific ρ for which BF_j^{un} takes a particular value. Specifically, we seek the combination of λ and ρ that produces a unicovariate Bayes factor equal to one. Covariates 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. $\rho_t = \{\rho : BF_j^{un} = 1\}$ for a given λ.*

We can now work backwards and specify a threshold level of practical significance $\rho_t \geq \rho_b$ 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 be never 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 value is $\rho_b = 0.31$. Hence, we may choose λ such that the threshold correlation is equal to $\rho_t = 0.40$ as a reasonable value. For this choice, any model including a covariate correlated with Y with $\rho = 0.4$ will be a-posteriori supported with 50% probability while this value will be increased as ρ increases. Other reasonable choices in this example might be $\rho_t = 0.35$ or $\rho_t = 0.5$. The first choice will be less strict supporting models of slightly higher dimension while the later will be more strict supporting more parsimonious models. Table 1 presents λ for $n = 50, 100$ and 500 and various values of the correlation as the threshold values.

Specification of λ	n=50 ρ_b=0.31	n=100 ρ_b=0.22	n=500 ρ_b=0.10
$\{\lambda : \rho = \rho_t, BF_j^{\text{un}} = 1\}$	$\rho_t = 0.35$, $\lambda = 0.218$	$\rho_t = 0.25$, $\lambda = 0.335$	$\rho_t = 0.15$, $\lambda = 0.060$
$\{\lambda : \rho = \rho_t, BF_j^{\text{un}} = 1\}$	$\rho_t = 0.40$, $\lambda = 0.067$	$\rho_t = 0.30$, $\lambda = 0.069$	$\rho_t = 0.20$, $\lambda = 7 \times 10^{-4}$
$\{\lambda : \rho = \rho_t, BF_j^{\text{un}} = 1\}$	$\rho_t = 0.50$, $\lambda = 0.004$	$\rho_t = 0.40$, $\lambda = 0.001$	$\rho_t = 0.30$, $\lambda = 5 \times 10^{-6}$
$\{\lambda : \rho = 0.01, BF_j^{\text{un}} = \frac{1}{150}\}$	$\rho_t = 0.42$, $\lambda = 0.038$	$\rho_t = 0.31$, $\lambda = 0.053$	$\rho_t = 0.14$, $\lambda = 0.116$

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

Figure 5 presents the benchmark correlation against the sample size. Similarly to the results in Table 1, ρ_b decreases as the sample size increases, and therefore BF_j^{un} allows less important (in terms of correlation values) covariates to enter the model. The dotted line in Figure 5 shows threshold correlation values ρ_t when the shrinkage parameter is set equal to $\lambda = 0.067$ for all the sample sizes. The value of $\lambda = 0.067$ is indicative and was selected to correspond to the threshold correlation of 0.40 for $n = 50$.

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.

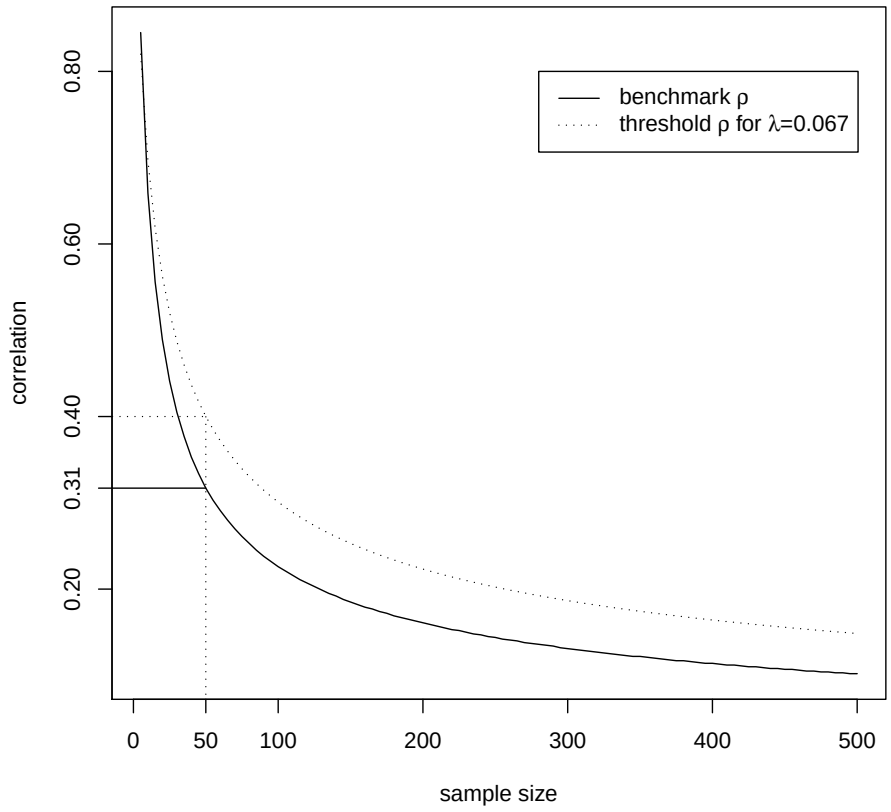

Figure 5: Benchmark and threshold correlations versus sample size; the value of $\lambda = 0.067$ corresponds to threshold value of $\rho_t = 0.4$ for sample size $n = 50$.

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 for nested model comparisons equal to one (i.e. which are the levels of partial correlation that correspond to the limits between significance and insignificance) for any given level of shrinkage λ. By 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, Chapter 5) in order to introduce some preliminary Lasso measures. Therefore we consider Y to be a $n \times 1$ 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 $\boldsymbol{\beta}^{\text{ols}}$ arises when we minimize the residual variance $\mathsf{var}(\varepsilon) = \mathsf{var}(\mathsf{Y} - \mathsf{X}\boldsymbol{\beta})$ giving $\boldsymbol{\beta}^{\text{ols}} = [\mathsf{var}(\mathsf{X})]^{-1}\,\mathsf{cov}(\mathsf{X},\mathsf{Y})$ assuming that $\mathsf{E}(\mathsf{X})$ and $\mathsf{E}(\mathsf{Y})$ are zero for simplicity. In the same analogy, the Lasso prediction coefficient $\boldsymbol{\beta}^{\text{lasso}}$ arises when we minimize a penalized version of the residual variance $\mathsf{var}(\mathsf{Y}-\mathsf{X}\boldsymbol{\beta}) + k||\boldsymbol{\beta}||$ resulting in $\boldsymbol{\beta}^{\text{lasso}} = \big[\mathsf{var}(\mathsf{X})\big]^{-1}\big(\mathsf{cov}(\mathsf{X},\mathsf{Y}) - k\boldsymbol{s}_{\boldsymbol{\beta}}\big)$; where $\boldsymbol{s}_{\boldsymbol{\beta}}$ is the sign vector of $\boldsymbol{\beta}^{\text{lasso}}$ and k is the shrinkage level when working with the variances and expectations of the random variables Y and X.

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

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

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*

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

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

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

$$\begin{aligned} R^{(\text{lasso})2}_{\mathsf{Y}|\mathsf{X}} &= 1 - \frac{\mathsf{var}_{\text{lasso}}(\mathsf{Y}|\mathsf{X}) + 2k\big|\big|\boldsymbol{\beta}^{\text{lasso}}\big|\big|}{\mathsf{var}(\mathsf{Y})} &(12)\\ &= R^{(\text{ols})2}_{\mathsf{Y}|\mathsf{X}} - 2k\frac{\big|\big|\boldsymbol{\beta}^{\text{lasso}}\big|\big|}{\mathsf{var}(\mathsf{Y})} - k^2\frac{\boldsymbol{s}_{\boldsymbol{\beta}}^T\,\mathsf{var}(\mathsf{X})^{-1}\boldsymbol{s}_{\boldsymbol{\beta}}}{\mathsf{var}(\mathsf{Y})}. &(13)\end{aligned}$$

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

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*

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

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 practice and in the illustrated examples here, the differences between the two measures are minor. As we will see in Section 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 imposed 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 $BF^{\text{mu}}_{m,j}$ which is defined as follows.

Definition 7 (Nested Multiple Lasso Bayes Factor $BF^{\mathbf{mu}}_{m,j}$**)** *For any model* m *with included covariates* $\mathsf{X}_\ell \in \mathcal{V}_m$*, the nested multiple Lasso Bayes factor* $BF^{mu}_{m,j}$ *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 the corresponding $\mathrm{BF}^{\mathrm{mu}}_{m,j}$ is approximately given by

$$\begin{aligned}
\mathrm{BF}^{\mathrm{mu}}_{m,j} &\approx \lambda c \left(\frac{|\boldsymbol{X}^T_{m_j^+} \boldsymbol{X}_{m_j^+}|}{|\boldsymbol{X}^T_{m_j^-} \boldsymbol{X}_{m_j^-}|} \right)^{-1/2} \frac{\left(||\boldsymbol{y} - \boldsymbol{X}_{m_j^+} \widehat{\boldsymbol{\beta}}^{\mathrm{lasso}}_{m_j^+}||^2 + 2\lambda ||\widehat{\boldsymbol{\beta}}^{\mathrm{lasso}}_{m_j^+}|| \right)^{-df/2}}{\left(||\boldsymbol{y} - \boldsymbol{X}_{m_j^-} \widehat{\boldsymbol{\beta}}^{\mathrm{lasso}}_{m_j^-}||^2 + 2\lambda ||\widehat{\boldsymbol{\beta}}^{\mathrm{lasso}}_{m_j^-}|| \right)^{-(df-1)/2}} \\
&= ck \left(\frac{\mathsf{var}_{\mathrm{lasso}}(\boldsymbol{y}|\boldsymbol{X}_{m_j^+}) + 2k||\widehat{\boldsymbol{\beta}}^{\mathrm{lasso}}_{m_j^+}||_1}{\mathsf{var}_{\mathrm{lasso}}(\boldsymbol{y}|\boldsymbol{X}_{m_j^-}) + 2k||\widehat{\boldsymbol{\beta}}^{\mathrm{lasso}}_{m_j^-}||_1} \right)^{-df/2} \\
&\quad \times \Big[\mathsf{var}(\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}) \big(\mathsf{var}_{\mathrm{lasso}}(\boldsymbol{y}|\boldsymbol{X}_{m_j^-}) + 2k||\widehat{\boldsymbol{\beta}}^{\mathrm{lasso}}_{m_j^-}|| \big) \Big]^{-1/2}.
\end{aligned}$$

where $c = \sqrt{\pi} \frac{\Gamma(\frac{df}{2})}{\Gamma(\frac{df-1}{2})}$, $\widehat{\boldsymbol{\beta}}^{\mathrm{lasso}}_{m_j^+}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{lasso}}_{m_j^-}$ are the Lasso estimates when regressing $\boldsymbol{y}$ on $\boldsymbol{X}_{m_j^+}$, and $\boldsymbol{X}_{m_j^-}$ respectively while $\mathsf{var}(\boldsymbol{y}|\boldsymbol{X})$ and $\mathsf{var}_{\mathrm{lasso}}(\boldsymbol{y}|\boldsymbol{X})$ 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, λ is the shrinkage level when working directly with the penalized version of the square differences between the fitted and the observed response value as described by (1). For this reason, we have that the shrinkage level used in Section 3.3.1 is given by $k = \lambda/(n-1)$.

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

$$\mathrm{BF}^{\mathrm{mu}}_{m,j} \approx ck \Big[1 - \mathsf{corr}^{(\mathrm{lasso})2}(\boldsymbol{y}, \boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}) \Big]^{-df/2} \frac{1}{\sqrt{\left(1 - R^{(\mathrm{ols})2}_{\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}}\right)\left(1 - R^{(\mathrm{lasso})2}_{\boldsymbol{y}|\boldsymbol{X}_{m_j^-}}\right)}}. \qquad (15)$$

According to our proposed method, we define the shrinkage level by setting the unicovariate $\mathrm{BF}^{\mathrm{un}}_j$ 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. By this way, we can examine the behaviour of the proposed variable selection procedure and why covariates with low Pearson correlations are finally included in most probable a-posteriori models. The behaviour of $\mathrm{BF}^{\mathrm{mu}}_{m,j}$ is depicted in Figure 6. This Figure presents threshold partial correlations against the sample size for $\lambda = 0.067$ and for various values of $R^{(\mathrm{ols})2}_{\boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}}, R^{(\mathrm{lasso})2}_{\boldsymbol{y}|\boldsymbol{X}_{m_j^-}}$. The corresponding threshold values for Pearson correlation are also presented in the solid line. The two threshold values are closer when there is a weak correlation between $\boldsymbol{X}_j$ and $\boldsymbol{X}_{m_j^-}$. For large values $R^2_{\boldsymbol{y}|\boldsymbol{X}_{m_j^-}}$, which means that the $\boldsymbol{X}_{m_j^-}$ interpret a large percentage of the response variability, the threshold values for the partial correlation decrease.

Theorem 3.1 *For any selected λ, $LBF^{un}_j = LBF^{mu}_{m,j} \Rightarrow \mathsf{corr}^{(lasso)2}(\boldsymbol{y}, \boldsymbol{X}_j|\boldsymbol{X}_{m_j^-}) \leq (\rho_j - ks)$; where LBF is the Laplace approximation of the corresponding Bayes factor.*

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

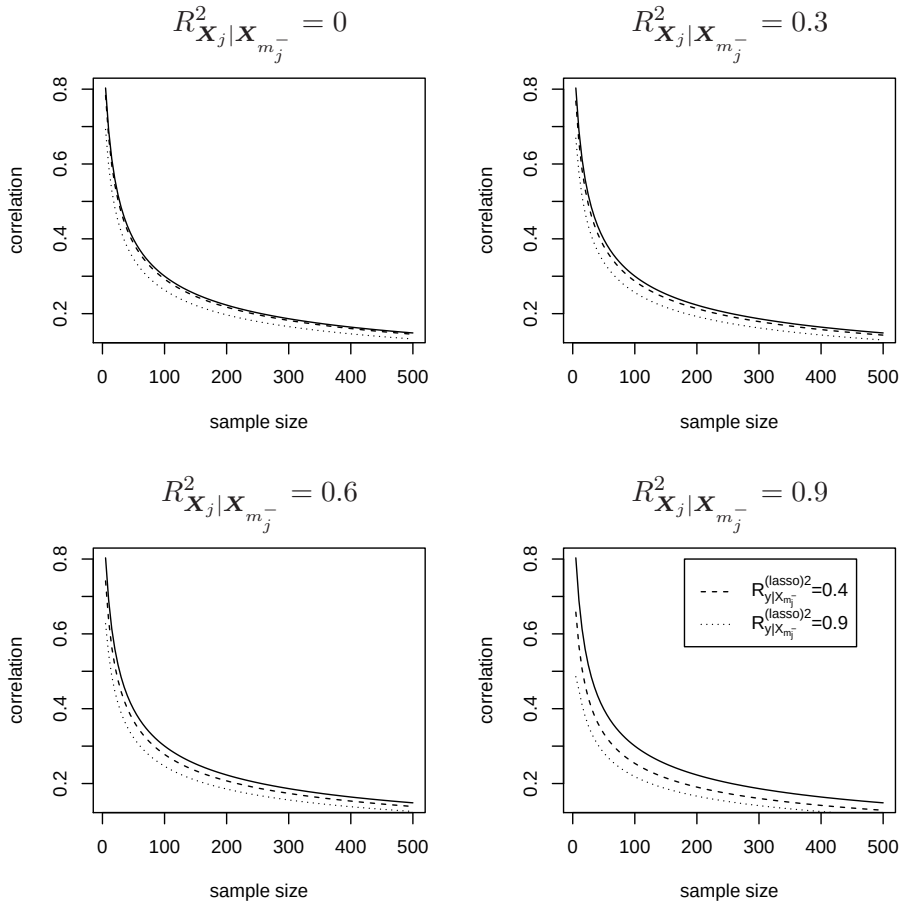

Figure 6: Plot of the threshold correlation for $\lambda = 0.067$ against the sample size and for $R^2_{\boldsymbol{X}_j|\boldsymbol{X}_{m_j^+}} = 0, 0.3, 0.6$ and 0.9. The solid line shows the threshold Pearson correlations, the dashed and dotted lines show the corresponding threshold values for the partial correlation for $R^2_{\boldsymbol{y}|\boldsymbol{X}_{m_j^+}} = 0.4$ and 0.9 respectively.

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, 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 in 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 for 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 behaviour 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 and

more important covariates enter the model. The convergence between the threshold values of the partial and Pearson correlations is also depicted in Figure 6 where their differences (as appear on the right bottom of each plot) diminish for large sample sizes.

To sum up, in this Section we have illustrated the effect of any chosen level of λ on $\mathrm{BF}^{\mathrm{mu}}_{m,j}$. 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. By 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 Illustration

4.1 Simulated exampled

We investigate the performance of the proposed method of tuning the shrinkage parameter. We use the same data set as in Section 2.3 and we choose different threshold values for the Pearson correlation as indicated in Table 1.

The results are summarized in Table 2. For threshold correlation equal to 0.35 or 0.40 the model with the maximum a posterior probability (MAP) is the true model, $\mathsf{X}_4 + \mathsf{X}_5$. Choosing a more strict shrinkage level ($\lambda = 0.004$) leads to the selection only of X_5 (which is the covariate with the highest correlation with Y), though, the true model is still visited frequently (43.3%). The idea to specify the shrinkage level such that a covariate with very low correlation should be excluded from the model with high probability also seems promising, since, in this particular example, it succeeded identifying the true model with high posterior probability (61.4%); see third row of Table 2.

		Var. incl.	Post. Incl. Prob	Prob. of model	
ρ_t	λ		$\mathsf{X}_4,\ \mathsf{X}_5,\ \mathsf{X}_{12}$	MAP	true
0.35	0.217	$\mathsf{X}_4, \mathsf{X}_5$	$0.96, 1.00, 0.38$	26.22%	
0.40	0.067	$\mathsf{X}_4, \mathsf{X}_5$	$0.85, 1.00, 0.15$	55.49%	
0.42*	0.038*	$\mathsf{X}_4, \mathsf{X}_5$	$0.78, 1.00, 0.09$	61.39%	
0.50	0.004	X_5	$0.45, 0.96, 0.01$	50.45%	43.33%

* *These values have been produced by setting* $BF^{un}_j = 1/150$ *for covariates with* $\rho = 0.01$.

Table 2: Posterior summaries for various choices of λ for Example 4.1.

Even though the unicovariate BF and the Pearson correlation between the covariates and the response have been used to specify the shrinkage level, it turns out that this choice is in accordance with the partial correlations. Table 3 shows the absolute values of the sample Pearson and partial correlations, where we observe that the selected variates (the ones in bold font) are the ones with partial correlation higher than the threshold correlations.

	X_2	X_4	X_5	X_6	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{15}
$\mathsf{corr}(\boldsymbol{y}, \boldsymbol{X}_j)$	0.03	**0.38**	**0.58**	0.01	0.08	0.06	0.02	0.03	0.22	0.10
$\mathsf{corr}(\boldsymbol{y}, \boldsymbol{X}_j \vert \boldsymbol{X}_{\setminus j})^*$	0.11	**0.51**	**0.68**	0.18	0.16	0.22	0.13	0.16	0.34	0.28

**Lasso partial correlations are equal to the ordinary partial correlations (in 3 d.p.) due to small values of the shrinkage parameters used in Table 1.*
Covariates with partial correlations ≤ 0.05 *are omitted from the table.*

Table 3: Observed Pearson and partial correlation coefficients (in absolute values) for Example 4.1.

4.2 Simulation study

We examine the performance of the proposed Bayesian Lasso and the corresponding method to specify the shrinkage parameter in a set of simulated data where some of the covariates are correlated with each other. We therefore use the simulation design study from Nott and Kohn (2005), which consists of 15 variables of 50 observations each. 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 $\varepsilon \sim N(0, 2.5^2)$.

We use the same threshold correlation levels as in the Example 4.1, since the sample size is the same. Each MCMC was updated using 20000 iterations after discarding additional 10000 observations. All results are evaluated over 100 datasets generated using the sampling scheme described above. Figure 7 shows 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 very frequently selected for all the shrinkage levels, whereas, the remaining covariates are less frequently selected, while their inclusions probabilities become smaller as the shrinkage parameter decreases.

Table 4 presents the true Pearson correlation and partial correlations for this structure of simulated data set. In fact only the covariates $\mathsf{X}_1, \mathsf{X}_7$ and X_{11}, which have the higher posterior probabilities, are the ones that have higher 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 $\mathsf{X}_5, \mathsf{X}_{11}$ and X_{13}. The Bayesian Lasso 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.

The highest a-posteriori models over 100 simulated data sets are presented in Table 5. The model with covariates $\mathsf{X}_1, \mathsf{X}_7, \mathsf{X}_{11}$ is the one most frequently indicated as the MAP model

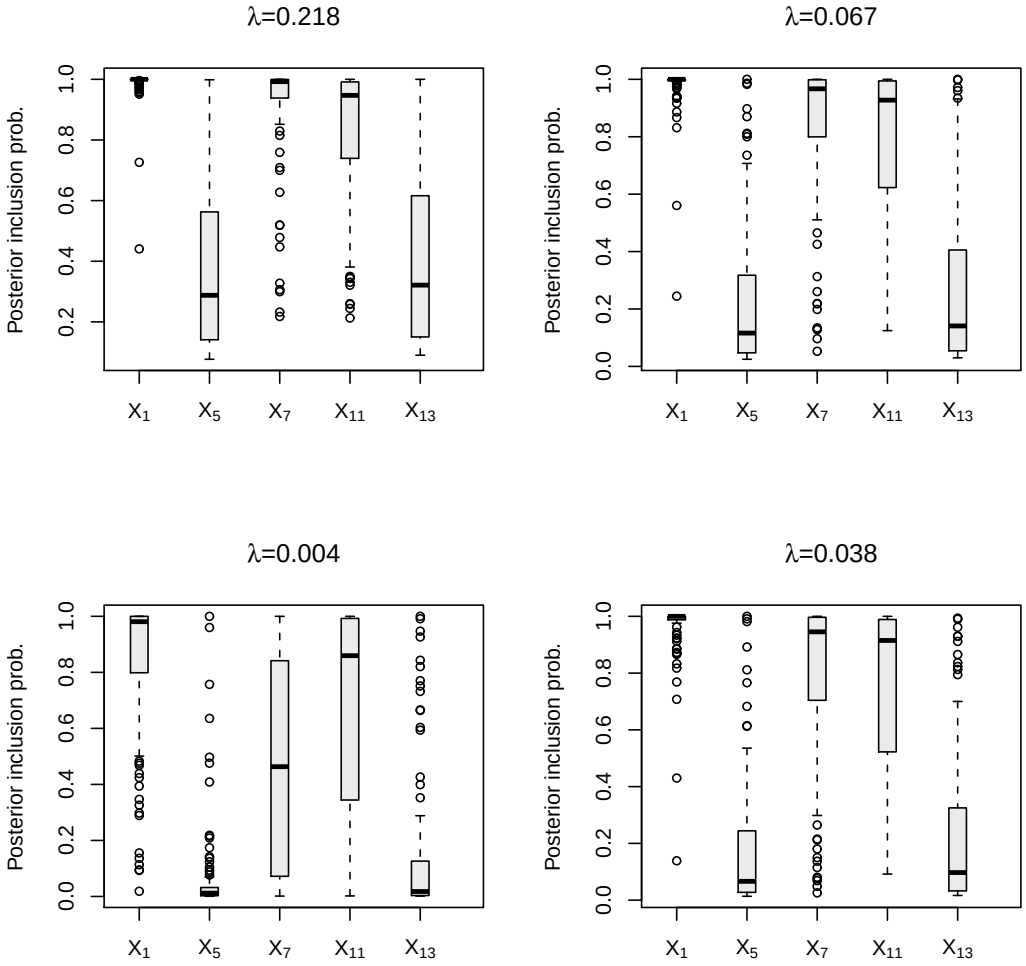

Figure 7: Boxplots of the posterior inclusion probabilities for covariates with true non-zero effects over 100 generated datasets for Example 4.2.

	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}
$\text{corr}(\boldsymbol{y}, \boldsymbol{X}_j)$	0.56	0.17	0.24	0.31	0.15	0.00	0.34	0.56	0.44	0.50
$\text{corr}(\boldsymbol{y}, \boldsymbol{X}_j \vert \boldsymbol{X}_{\setminus j})$	**0.55**	0.00	0.00	0.00	0.15	0.00	**0.52**	**0.37**	0.00	0.20

Table 4: True values of the Pearson and the partial correlation coefficients (in absolute values) for Example 4.2.

for all the shrinkage levels implemented here. As λ decreases (which implies that the selection procedure becomes more strict since the threshold correlation increases) the posterior probability of the true model decreases as expected.

The proposed method to specify the shrinkage parameter works well even for this data set with highly correlated variables. The important variables are identified and included in the MAP model apart from the variables X_5 and X_{13}. Covariate X_5 has been used to generate the response but it has low Pearson and partial with the response and it is not included in the

ρ_t	λ	$\mathsf{X}_1+\mathsf{X}_7+\mathsf{X}_{11}$	$\mathsf{X}_1+\mathsf{X}_5+\mathsf{X}_7+\mathsf{X}_{11}$	$\mathsf{X}_1+\mathsf{X}_5+\mathsf{X}_7+\mathsf{X}_{11}+\mathsf{X}_{13}$
0.35	0.218	29%	10%	11%
0.40	0.067	44%	7%	5%
0.42	0.038	46%	7%	2%
0.50	0.004	28%	1%	0%

Table 5: Percentage of samples that each model is identified as the MAP over 100 generated datasets for Example 4.2.

model. The variable X_{13} has high Pearson correlation with the response but it is not included in the model due to its high correlation with the variable X_{11} ($\mathsf{corr}(X_{11}, X_{13}) = 0.74$).

4.3 Real example – Diabetes data set

The diabetes data set was widely used in the literature to evaluate algorithms for Lasso (Efron et al., 2004). The data set contains ten baseline variables, age, sex, body mass index (bmi), average blood pressure (bp), six blood serum measurements (tc, ldl, hdl, tch, ltg, glu) and the response which is a one year measure of disease progression for 442 diabetes patients. According to Efron et al. (2004) fitting linear models are desirable in this diagnostic application, not only for future prediction but also for revealing the important factors.

The benchmark correlation for data of this size ($n = 442$) is found equal to 0.11 for this reason we choose threshold correlations equal to $0.15, 0.20$ and 0.30. We also present the threshold correlation that corresponds to the choice of $\lambda = 0.067$ ($\rho_t = 0.4$ for $n = 50$), which can be considered as a reference shrinkage value. The corresponding shrinkage levels are used to update 50000 observations through the proposed Gibbs sampler after discarding additional ten thousands iterations as burn-in period. The posterior inclusion probabilities are summarized in Table 6, where the variables with their probabilities in bold are the ones that included in the MAP model.

"Age" as well the second, fourth and sixth of the blood serum measurements have very small posterior inclusion probabilities. These are also found to be the weakest predictors in Hans (2009), whereas, Park and Casella (2008), Balakrishnan and Madigan (2009) and Li and Lin (2010) do not identify the sixth blood serum measurements among the weakest variables. The first blood serum measurement has moderate inclusion probabilities but is excluded from the model for all the selected shrinkage levels. The variables of sex, blood pressure and the third blood serum measurement are important predictors, included in the model when the shrinkage levels imposed are moderate. Nevertheless, there is strong evidence for the importance of body mass index and ltg measurement, which are included in the MAP model even when the choice of the threshold correlation is high and therefore the implied variable selection rule very strict. These are the important variables that have been also identified by Hans (2009) and Li and Lin (2010). The same conclusion is drawn if we choose $\lambda = 0.067$ (implying $\rho_t = 0.16$), which was used in the previous (smaller in size) examples.

Table 7 shows the Pearson correlations between the variables and the response, the partial correlation between the response and the candidate variable given that all the remaining variables are included in the model and, in the third row, the partial correlation between

ρ_t	λ	age	sex	bmi	bp	tc	ldl	hdl	tch	ltg	glu
0.150	0.110	0.010	**0.889**	**1.000**	**1.000**	0.364	0.240	**0.673**	0.088	**1.000**	0.017
0.157*	0.067*	0.007	**0.810**	**1.000**	**0.998**	0.305	0.191	**0.661**	0.077	**1.000**	0.011
0.200	0.002	0.000	0.002	**1.000**	**0.585**	0.031	0.001	0.030	0.000	**1.000**	0.001
0.300	4.41×10^{-6}	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	0.000	**1.000**	0.000

** The shrinkage value $\lambda = 0.067$ has been chosen by setting $BF_j^{un} = 1$ for covariates with $\rho = 0.40$ and $n = 50$.*

Table 6: Posterior variable inclusion probabilities for several shrinkage values λ for the Diabates data set (Example 4.3).

response and each candidate variable given the variables that are included in the MAP model for $\lambda = 0.067$. We observe that all variables included in the model are the ones with partial correlation higher than the selected threshold correlation. Also note that hdl should not be included in the full model (Lasso partial correlation 0.02 while the corresponding threshold value is 0.13), while it should be included in the MAP (Lasso partial correlation 0.21 with threshold value 0.15).

	age	sex	bmi	bp	tc	ldl	hdl	tch	ltg	glu
$\mathsf{corr}(\boldsymbol{y}, \boldsymbol{X}_j)$	0.19	0.04	0.59	0.44	0.21	0.17	0.40	0.43	0.57	0.38
	(0.15)	(0.15)	(0.15)	(0.15)	(0.15)	(0.15)	(0.15)	(0.15)	(0.15)	(0.15)
$\mathsf{corr}^{(\text{lasso})}(\boldsymbol{y}, \boldsymbol{X}_j \vert \boldsymbol{X}_{\setminus j})$	0.01	**0.19**	**0.35**	**0.23**	0.09	0.07	**0.02**	0.05	**0.21**	0.05
	(0.15)	(0.15)	(0.15)	(0.15)	(0.12)	(0.12)	(0.13)	(0.13)	(0.13)	(0.15)
$\mathsf{corr}^{(\text{lasso})}(\boldsymbol{y}, \boldsymbol{X}_j \vert \boldsymbol{X}_{m_j^-})^*$		**0.18**	**0.36**	**0.24**			**0.21**		**0.33**	
		(0.15)	(0.15)	(0.15)			(0.15)		(0.15)	

** 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.*

Table 7: Absolute values of the observed Pearson and partial correlation coefficients for Diabetes data set (Example 4.3); Threshold values for $\lambda = 0.067$ are given in parentheses below each correlation measure. .

5 Discussion

In this article we present 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 investigate the value of new regularization plots which depict the behaviour 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 behaviour 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 be never 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 proceed 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 exploit this relation to define λ by specifying the desired level of threshold correlation. By 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 can use these findings to interpret and understand the effect of our choice in nested multiple regression model comparisons or even specify λ. Results from our illustrations indicate that the method behaves efficiently identifying important and sensible covariate effects.

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.

Another issue that the authors of this paper are currently examining is the use of hyper-priors by exploiting active sets of λ values. These sets can be defined by eliminating prior values of no practical use such as the ones that activate Lindley's paradox or over-shrink important effects towards zero. This may lead to robust variable selection methods in the direction of the priors proposed by Liang et al. (2008). Following them, we may extend the usual Lasso method to incorporate a covariance structure for regression coefficients and propose a sensible hyper-prior for the shrinkage parameter λ.

Finally, 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.

Appendix

Details for the derivation of equation and (11)

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

Details for the derivation of equations (12) and (13)

From Definition 5 we can write

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

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

$$\begin{aligned}
R^{(\text{lasso})2}_{\mathsf{Y}|\mathsf{X}} &= 1 - \frac{\mathsf{var}(\mathsf{Y}|\mathsf{X}) + k\boldsymbol{s}_{\boldsymbol{\beta}}^{T}\,\mathsf{var}(\mathsf{X})^{-1}\left(2\,\mathsf{cov}(X, Y) - k\boldsymbol{s}_{\boldsymbol{\beta}}\right)}{\mathsf{var}(\mathsf{Y})} \\
&= 1 - \frac{\mathsf{var}(\mathsf{Y}|\mathsf{X}) + k^2\boldsymbol{s}_{\boldsymbol{\beta}}^{T}\,\mathsf{var}(\mathsf{X})^{-1}\boldsymbol{s}_{\boldsymbol{\beta}} + 2k\boldsymbol{s}_{\boldsymbol{\beta}}^{T}\,\mathsf{var}(\mathsf{X})^{-1}\left(\mathsf{cov}(\mathsf{X}, \mathsf{Y}) - k\boldsymbol{s}_{\boldsymbol{\beta}}\right)}{\mathsf{var}(\mathsf{Y})} \\
&= 1 - \frac{\mathsf{var}(\mathsf{Y}|\mathsf{X}) + k^2\boldsymbol{s}_{\boldsymbol{\beta}}^{T}\,\mathsf{var}(\mathsf{X})^{-1}\boldsymbol{s}_{\boldsymbol{\beta}} + 2k\boldsymbol{s}_{\boldsymbol{\beta}}^{T}\boldsymbol{\beta}^{\text{lasso}}}{\mathsf{var}(\mathsf{Y})} \\
&= 1 - \frac{\mathsf{var}(\mathsf{Y}|\mathsf{X}) + k^2\boldsymbol{s}_{\boldsymbol{\beta}}^{T}\,\mathsf{var}(\mathsf{X})^{-1}\boldsymbol{s}_{\boldsymbol{\beta}} + 2k\left|\left|\boldsymbol{\beta}^{\text{lasso}}\right|\right|}{\mathsf{var}(\mathsf{Y})} \qquad (16)
\end{aligned}$$

From (11) we have that

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

which is the expression (12).

Finally, substituting $R^2_{\mathsf{Y}|\mathsf{X}} = 1 - \frac{\mathsf{var}(\mathsf{Y}|\mathsf{X})}{\mathsf{var}(\mathsf{Y})}$ back in (16) gives us the result of equation (13).

Corollary 1

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

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

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

Theorem 3.1

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

$$\text{LBF}^{\text{un}}_j = ck\big[1 - \Big(\rho_j - k\boldsymbol{s}_{\widehat{\boldsymbol{\beta}}}\Big)^2\big]^{-df/2}, \tag{17}$$

where ρ_j is sample Pearson correlation between Y and X_j. Equating (15) and (17), the threshold values of the Pearson and partial correlations satisfy the following

$$(1 - \rho^2_{\boldsymbol{y},\boldsymbol{X}_j|\boldsymbol{X}_{m^-_j}}) \left[\left(1 - R^{(\text{ols})2}_{\boldsymbol{X}_j|\boldsymbol{X}_{m^-_j}}\right)\left(1 - R^{(\text{lasso})2}_{\boldsymbol{y}|\boldsymbol{X}_{m^-_j}}\right)\right]^{1/df} = 1 - \Big(\rho_j - k\boldsymbol{s}_{\widehat{\boldsymbol{\beta}}}\Big)^2, \tag{18}$$

where $\rho_{\boldsymbol{y},\boldsymbol{X}_j|\boldsymbol{X}_{m^-_j}} = \mathsf{corr}^{(\text{lasso})}\big(\boldsymbol{y}, \boldsymbol{X}_j|\boldsymbol{X}_{m^-_j}\big)$.

Since $R^{(\text{ols})2}_{\boldsymbol{X}_j|\boldsymbol{X}_{m^-_j}}$ and $R^{(\text{lasso})2}_{\boldsymbol{y}|\boldsymbol{X}_{m^-_j}}$ lie in the $[0,1]$ interval, we have that

$$\rho^2_{\boldsymbol{y},\boldsymbol{X}_j|\boldsymbol{X}_{m^-_j}} \leq \Big(\rho_j - k\boldsymbol{s}_{\widehat{\boldsymbol{\beta}}}\Big)^2 .$$

□

Corollary 2

Proof: The proof of Corollary 2, immediately follows Theorem 3.1 if we set $L\text{BF}^{\text{un}}_j = L\text{BF}^{\text{mu}}_{m,j} = 1$. □

Corollary 3

Proof: For $n \to \infty$, (18) becomes equal to

$$(1 - \rho^2_{\boldsymbol{y},\boldsymbol{X}_j|\boldsymbol{X}_{m^-_j}}) = 1 - \rho^2_j$$

since $df = n + 2a + 1 \to \infty$ and $k = \lambda/(n-1) \to 0$. When considering $L\text{BF}^{\text{un}}_j = L\text{BF}^{\text{mu}}_{m,j} = 1$ we obtain that the threshold values of the Pearson and the partial correlations are equal for $n \to \infty$. □

References

Balakrishnan, S. and Madigan, D. (2009). Priors on the variance in sparse bayesian learning: the demi-bayesian lasso. *Submitted*.

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

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

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32:407–499.

George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889.

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

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188.

Hans, C. (2009). Bayesian Lasso regression. *Biometrika*, 96(4):835–845.

Hans, C. (2010). Model uncertainty and variable selection in bayesian lasso regression. *Statistics and Computing*, 20:221–229.

Johnson, B. (2009). On lasso for censored data. *Electronic Journal of Statistics*, 3:485–506.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *The Indian Journal of Statistics*, 60:65–81.

Li, Q. and Lin, N. (2010). The bayesian elastic net. *Bayesian Analysis*, 5(1):847–866.

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103:410–423.

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

Lykou, A. and Whittaker, J. (2010). Sparse canonical correlation analysis by using the lasso. *Computational Statistics and Data Analysis*, 54:31443157.

Meier, L., Van de Geer, S., and Bhlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B*, 70:53–71.

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

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBugs*. John Wiley & Sons.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319337.

Park, M. Y. and Hastie, T. (2006). l_1 regularization path algorithm for generalized linear models. 69:659–677.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–687.

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

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

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

Yuan, M. and Lin, Y. (2005). Efficient empirical bayes variable selection and estimation in linear models. *Journal of American Statistical Association*, 100(472).

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

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594.

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

Zou, J. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67:301–320.