Bayesian Adaptive Selection Under Prior Ignorance *

Tathagata Basu[1], Matthias C. M. Troffaes[1], and Jochen Einbeck[1]

Department of Mathematical Sciences, Durham University, UK

Abstract. Bayesian variable selection is one of the popular topics in modern day statistics. It is an important tool for high dimensional statistics, where the number of model parameters is greater than the number of observations. Several Bayesian models have been proposed for variable selection. However, a convincing robust Bayesian approach is yet to be investigated. Here in this work, we investigate sensitivity analysis over a simplex of probability measures. We sample from this simplex to get an inclusion probability of each variable. The sensitivity analysis gives us a set of posteriors instead of a single posterior. This set of posteriors gives us a behaviour of the model parameters with respect to different prior elicitations resulting in robust inferential conclusions.

Keywords: High Dimensional · Variable Selection · Bayesian Analysis · Imprecise Probability.

1 Introduction

High dimensional statistical modelling is a popular topic in modern day statistics. These type of problems are often very hard to be dealt using classical methods and we often rely on regularisation methods. There are several well-known frequentist methods which are efficient in tackling high dimensional problems. Tibshirani introduced least absolute shrinkage and selection operator or simply LASSO [11]. Fan and Li investigated asymptotic properties for variable selection and introduced SCAD [4]. Zou introduced adaptive LASSO [12], a weighted version of LASSO that gives asymptotically unbiased estimates.

High dimensional modelling is equally well investigated in a Bayesian context. George and McCulloch introduced stochastic search variable selection [5] which uses latent variables for the selection of predictors. Ishwaran and Rao used a continuous bimodal prior for hyper-variances in spike and slab model to attain sparsity [6]. Park and Casella introduced a hierarchical model using the double exponential distribution [9]. Lykou and Ntzoufras [7] proposed a double exponential distribution for the regression parameters along with bernoulli distributed latent variables. Several other works have been done.

* This work is funded by the European Commission's H2020 programme, through the UTOPIAE Marie Curie Innovative Training Network, H2020-MSCA-ITN-2016, Grant Agreement number 722734.

In this article, we follow the approach of Narisetty and He [8] to attain sparsity. Moreover, we introduce an additional imprecise beta-Bernoulli to specify the selection probability of the latent variables similar to [7]. We perform a sensitivity analysis over these sets of selection probabilities to obtain a robust Bayesian variable selection routine.

The rest of the paper is organised as follows: We first define our hierarchical model in Section 2, followed by the posterior computation in Section 3. We use an orthogonal design case to show the closed form posteriors and discuss their properties. In Section 4, we illustrate our results using both synthetic and real datasets and finally, we draw conclusions in Section 5.

2 Model

Let, $Y \coloneqq (Y_1, \cdots, Y_n)^T$ denote the responses and $\mathbf{X} \coloneqq (\mathbf{X}_1, \cdots, \mathbf{X}_n)^T$ be corresponding p-dimensional predictors. Then we define a linear model in the following way:

$$Y = \mathbf{X}\beta + \epsilon \tag{1}$$

where, $\beta \coloneqq (\beta_1, \cdots, \beta_p)^T$ is the vector of regression coefficients and $\epsilon \coloneqq (\epsilon_1, \cdots, \epsilon_n)^T$ are Gaussian noises so that for $1 \leq i \leq n$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

We define the following hierarchical model for linear models, so that for $1 \leq j \leq p$,

$$Y \mid \mathbf{X}, \beta \sim \mathcal{N}\left(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n\right) \tag{2}$$

$$\beta_j \mid \gamma_j = 1 \sim \mathcal{N}(0, \sigma^2 \tau_1^2) \tag{3}$$

$$\beta_j \mid \gamma_j = 0 \sim \mathcal{N}(0, \sigma^2 \tau_0^2) \tag{4}$$

$$\gamma_j \mid q_j \sim \text{Ber}(q_j) \tag{5}$$

$$q_j \sim \text{Beta}(s\alpha_j, s(1-\alpha_j)) \tag{6}$$

where, $s > 0$ are fixed constants.

The latent variables $\gamma \coloneqq (\gamma_1, \cdots, \gamma_p)$ in the model corresponds to spike and slab prior specification routine where γ_j represents the selection of the co-variate $\mathbf{x}_j$. We fix sufficiently small τ_0 ($1 \gg \tau_0^2 > 0$) so that $\beta_j | \gamma_j = 0$ has its probability mass concentrated around zero. Therefore, probability distribution of $\beta_j | \gamma_j = 0$ represents the spike component of our prior specification. To construct the slab component, we consider τ_1^2 to be large so that $\tau_1 \gg \tau_0$. This allows the prior for $\beta_j \mid \gamma_j = 1$ to be flat. We assume σ^2 to be known for the ease of computation. In a more generalised setting, we may choose an inverse-gamma distribution.

We use imprecise beta priors to specify the selection probabilities $q \coloneqq (q_1, \ldots, q_p)$. We use $\alpha \coloneqq (\alpha_1, \ldots, \alpha_p)$ to represent our prior expectation of the selection probabilities (q) and s to represent concentration parameter. We consider $\alpha \in \mathcal{P}$, where

$$\mathcal{P} \coloneqq (0, 1)^p . \tag{7}$$

Note that, in our model we consider a near vacuous set for the elicitation of each α_j. That is, for $1 \gg \epsilon > 0$, $\alpha_j \in [\epsilon, 1 - \epsilon]$. Therefore, the prior elicitation on the

total number of active co-variates lies between $p\epsilon$ to $p(1-\epsilon)$. More generally, we can consider the following:

$$\alpha \in \mathcal{P} \subseteq (0,1)^p. \tag{8}$$

Each α_j assign a prior selection probability for each co-variate.

3 Posterior Computation

Let $\gamma \coloneqq (\gamma_1, \ldots, \gamma_p)$ and $q \coloneqq (q_1, \ldots, q_p)$. The joint posterior of the proposed hierarchical model can be computed in the following way:

$$P(\beta, \gamma, q \mid Y, \mathbf{X}) \propto P(Y \mid \mathbf{X}, \beta) P(\beta \mid \gamma) P(\gamma \mid q) P(q). \tag{9}$$

For the ease of computation we will use orthogonal design case ie, $\mathbf{X}^T\mathbf{X} = n\mathbf{I}_p$.

3.1 Selection indicators

Using Eq. (9), we write posterior of γ as

$$P(\gamma \mid Y, \mathbf{X}) = \iint P(\beta, \gamma, q \mid Y, \mathbf{X}) dq d\beta \tag{10}$$

$$\propto \int P(Y \mid \mathbf{X}, \beta) \left(P(\beta \mid \gamma) \int P(\gamma \mid q) P(q) dq \right) d\beta. \tag{11}$$

Let $f_{\gamma_j}(\beta_j)$ be the density of $\beta_j \mid \gamma_j$ as mentioned in Eq. (3) and Eq. (4). So,

$$f_{\gamma_j}(\beta_j) \coloneqq \frac{1}{\sqrt{2\pi}\sigma\tau_{\gamma_j}} \exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_{\gamma_j}^2} \right). \tag{12}$$

Since $P(\gamma_j \mid q_j) = q_j^{\gamma_j}(1-q_j)^{1-\gamma_j}$ and q_j follows Beta distribution,

$$P(\beta \mid \gamma) \int P(\gamma \mid q) P(q) dq$$

$$= \prod_j \left([f_1(\beta_j)]^{\gamma_j} [f_0(\beta_j)]^{1-\gamma_j} \int q_j^{\gamma_j} (1-q_j)^{1-\gamma_j} P(q_j) dq_j \right) \tag{13}$$

$$= \prod_j \left([\alpha_j f_1(\beta_j)]^{\gamma_j} [(1-\alpha_j) f_0(\beta_j)]^{1-\gamma_j} \right). \tag{14}$$

Now for the orthogonal design case that is when $\mathbf{X}^T\mathbf{X} = n\mathbf{I}_p$, we have $\hat{\beta} = \mathbf{X}^T Y/n$, where $\hat{\beta} \coloneqq (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ are the ordinary least square estimates. Then,

$$P(Y \mid \mathbf{X}, \beta) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2} \|Y - \mathbf{X}\beta\|_2^2 \right) \tag{15}$$

$$= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2} \left(n\beta^T\beta - 2n\beta^T\hat{\beta} + Y^TY \right) \right). \tag{16}$$

Then combining Eq. (11), Eq. (14) and Eq. (16) we have the decomposed posterior of γ_j such that

$$P(\gamma_j \mid Y, \mathbf{X}) = M_j \int \exp\left(-\frac{n(\beta_j - \hat{\beta}_j)^2}{2\sigma^2}\right) \times [\alpha_j f_1(\beta_j)]^{\gamma_j} [(1-\alpha_j) f_0(\beta_j)]^{1-\gamma_j} d\beta_j \tag{17}$$

where, M_j is a normalisation constant independent of γ_j. Then we have,

$$P(\gamma_j = 1 \mid \mathbf{X}, Y) = M_j \alpha_j \int \exp\left(-\frac{n(\beta_j - \hat{\beta}_j)^2}{2\sigma^2}\right) f_1(\beta_j) d\beta_j. \tag{18}$$

Now, by completing the square, it can be shown that for $k \in \{0, 1\}$ and $j \in \{1, \cdots, p\}$ we have

$$\exp\left(-\frac{n(\beta_j - \hat{\beta}_j)^2}{2\sigma^2}\right) f_k(\beta_j) = w_{k,j} \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\left(\beta_j - \hat{\beta}_{k,j}\right)^2}{2\sigma_k^2}\right) \tag{19}$$

where, $\hat{\beta}_{k,j} \coloneqq \frac{n\tau_k^2 \hat{\beta}_j}{n\tau_k^2+1}$, $\sigma_k^2 \coloneqq \frac{\sigma^2\tau_k^2}{n\tau_k^2+1}$ and $w_{k,j} \coloneqq \frac{1}{\sqrt{n\tau_k^2+1}} \exp\left(-\frac{n\hat{\beta}_j^2}{2(n\sigma^2\tau_k^2+\sigma^2)}\right)$. Then using Eq. (19) we have

$$P(\gamma_j = 1 \mid \mathbf{X}, Y) = M_j \alpha_j w_{1,j} \tag{20}$$

and

$$P(\gamma_j = 0 \mid \mathbf{X}, Y) = M_j (1-\alpha_j) w_{0,j}. \tag{21}$$

Therefore,

$$\gamma_j \mid \mathbf{X}, Y \sim Ber\left(\frac{\alpha_j w_{1,j}}{\alpha_j w_{1,j} + (1-\alpha_j) w_{0,j}}\right). \tag{22}$$

Co-variate selection For the co-variate selection we investigate the posterior odds of each γ_j. We assign a co-variate to be non-active when

$$\sup_{\alpha_j \in \mathcal{P}} \left\{ \frac{P(\gamma_j = 1 \mid \mathbf{X}, Y)}{P(\gamma_j = 0 \mid \mathbf{X}, Y)} \right\} < 1, \tag{23}$$

for $j = 1, \cdots, p$. Or equivalently when,

$$\sup_{\alpha_j \in \mathcal{P}} \left\{ \frac{w_{1,j}\alpha_j}{w_{0,j}(1-\alpha_j)} \right\} < 1. \tag{24}$$

Similarly, we assign a co-variate to be active if,

$$\inf_{\alpha_j \in \mathcal{P}} \left\{ \frac{w_{1,j}\alpha_j}{w_{0,j}(1-\alpha_j)} \right\} > 1. \tag{25}$$

We define the rest to be indeterminate, as it depends on prior elicitation on the selection probability.

Properties of the posterior: The posterior odds are given by:

$$\frac{w_{1,j}\alpha_j}{w_{0,j}(1-\alpha_j)} = \frac{w_{1,j}}{w_{0,j}}\left(\frac{1}{1-\alpha_j} - 1\right). \tag{26}$$

Now, the first derivative of the posterior odds are given by:

$$\frac{w_{1,j}}{w_{0,j}}\frac{1}{(1-\alpha_j)^2} > 0. \tag{27}$$

Therefore, the posterior odds are monotone with respect to the prior selection probability α_j.

Near vacuous set: Let $1 \gg \epsilon > 0$. We define a near vacuous set for prior selection probability α_j. So that, $\alpha_j \in [\epsilon, 1-\epsilon]$. Then, because of the monotonicity of posterior odds, we can compute the posterior odds on the lower and upper bounds of the set instead of the whole interval. Alternatively,

$$\sup_{\alpha_j \in [\epsilon, 1-\epsilon]} \left\{ \frac{w_{1,j}\alpha_j}{w_{0,j}(1-\alpha_j)} \right\} = \frac{(1-\epsilon)}{\epsilon} \cdot \frac{w_{1,j}}{w_{0,j}} \tag{28}$$

and,

$$\inf_{\alpha_j \in [\epsilon, 1-\epsilon]} \left\{ \frac{w_{1,j}\alpha_j}{w_{0,j}(1-\alpha_j)} \right\} = \frac{\epsilon}{(1-\epsilon)} \cdot \frac{w_{1,j}}{w_{0,j}} \tag{29}$$

3.2 Regression coefficients

The joint posterior of regression coefficients i.e. β is given by:

$$P(\beta \mid Y, \mathbf{X}) = \sum_{\gamma} \int P(\beta, \gamma, q \mid Y, \mathbf{X}) dq \tag{30}$$

$$\propto \sum_{\gamma} \int P(Y \mid \mathbf{X}, \beta) P(\beta \mid \gamma) P(\gamma \mid q) P(q) dq \tag{31}$$

$$\propto P(Y \mid \mathbf{X}, \beta) \sum_{\gamma} \left(P(\beta \mid \gamma) \int P(\gamma \mid q) P(q) dq \right). \tag{32}$$

From Eq. (14) we have

$$P(\beta \mid \gamma) \int P(\gamma \mid q) P(q) dq = \prod_j \left([\alpha_j f_1(\beta_j)]^{\gamma_j} [(1-\alpha_j) f_0(\beta_j)]^{1-\gamma_j} \right). \tag{33}$$

Then we can write Eq. (32) as

$$P(\beta \mid Y, \mathbf{X}) \propto P(Y \mid \mathbf{X}, \beta) \sum_{\gamma} \left(\prod_j \left([\alpha_j f_1(\beta_j)]^{\gamma_j} [(1-\alpha_j) f_0(\beta_j)]^{1-\gamma_j} \right) \right).$$

Therefore, swapping sum and product operations we get,

$$P(\beta \mid Y, \mathbf{X}) \propto P(Y \mid \mathbf{X}, \beta) \prod_j \sum_{\gamma_j} \left([\alpha_j f_1(\beta_j)]^{\gamma_j} [(1-\alpha_j) f_0(\beta_j)]^{1-\gamma_j} \right) \tag{34}$$

$$\propto P(Y \mid \mathbf{X}, \beta) \prod_j [\alpha_j f_1(\beta_j) + (1-\alpha_j) f_0(\beta_j)]. \tag{35}$$

Now, combining Eq. (16) and Eq. (35) we have

$$P(\beta \mid Y, \mathbf{X}) \propto \exp\left(-\frac{1}{2\sigma^2} \left(n\beta^T\beta - 2n\beta^T\hat{\beta} \right) \right) \prod_j [\alpha_j f_1(\beta_j) + (1-\alpha_j) f_0(\beta_j)]$$

$$\propto \exp\left(-\frac{n}{2\sigma^2} \|\beta - \hat{\beta}\|_2^2 \right) \prod_j [\alpha_j f_1(\beta_j) + (1-\alpha_j) f_0(\beta_j)] \tag{36}$$

$$\propto \prod_j \exp\left(-\frac{n(\beta_j - \hat{\beta}_j)^2}{2\sigma^2} \right) [\alpha_j f_1(\beta_j) + (1-\alpha_j) f_0(\beta_j)]. \tag{37}$$

Therefore, the β_j's are a posteriori independent and for each $1 \leq j \leq p$, we have,

$$P(\beta_j \mid Y, \mathbf{X}) \propto \exp\left(-\frac{n(\beta_j - \hat{\beta}_j)^2}{2\sigma^2} \right) [\alpha_j f_1(\beta_j) + (1-\alpha_j) f_0(\beta_j)]. \tag{38}$$

Let $W_j \coloneqq \alpha_j w_{1,j} + (1-\alpha_j) w_{0,j}$. Then combining Eq. (38) and Eq. (19) we have,

$$\beta_j \mid Y, \mathbf{X} \sim \frac{\alpha_j w_{1,j}}{W_j} \mathcal{N}\left(\hat{\beta}_{1,j}, \sigma_1^2 \right) + \frac{(1-\alpha_j) w_{0,j}}{W_j} \mathcal{N}\left(\hat{\beta}_{0,j}, \sigma_0^2 \right). \tag{39}$$

Properties of the posterior: To analyse the properties of the posterior, we first consider the ratio of the weights in Eq. (39). For $1 \leq j \leq p$, ratios of the weights are given by:

$$\frac{\alpha_j w_{1,j}}{(1-\alpha_j) w_{0,j}}. \tag{40}$$

This corresponds to posterior selection probability of selection indicators. Therefore, for active co-variates this ratio becomes greater than 1 for all $\alpha_j \in (0,1)$ and $\mathcal{N}\left(\hat{\beta}_{1,j}, \sigma_1^2 \right)$ dominates the posterior. Similarly, for non-active co-variates this ratio becomes less than 1 for all values of α_j and $\mathcal{N}\left(\hat{\beta}_{0,j}, \sigma_0^2 \right)$ dominates the posterior.

An interesting case occurs when, $\tau_0 \ll 1/n$ and $\alpha_j \in [\epsilon, 1-\epsilon]$. Then, $\mathcal{N}\left(\hat{\beta}_{1,j}, \sigma_1^2 \right)$ dominates the posterior if,

$$\hat{\beta}_j^2 > \frac{\sigma^2}{n} \frac{n\tau_1^2 + 1}{n\tau_1^2} \left[2\log\left(\frac{1-\epsilon}{\epsilon} \right) + \log\left(n\tau_1^2 + 1 \right) \right], \tag{41}$$

and similarly, $\mathcal{N}\left(\hat{\beta}_{0,j}, \sigma_0^2 \right)$ dominates the posterior if,

$$\hat{\beta}_j^2 < \frac{\sigma^2}{n} \frac{n\tau_1^2 + 1}{n\tau_1^2} \left[2\log\left(\frac{\epsilon}{1-\epsilon} \right) + \log\left(n\tau_1^2 + 1 \right) \right]. \tag{42}$$

Posterior mean and variance: The posterior expectation of β_j is given by:

$$E(\beta_j \mid Y, \mathbf{X}) = \frac{\alpha_j w_{1,j}}{W_j}\hat{\beta}_{1,j} + \frac{(1-\alpha_j)w_{0,j}}{W_j}\hat{\beta}_{0,j}. \tag{43}$$

Therefore, for orthogonal design, the posterior mean of β_j is equal to the least square estimate of β_j. The posterior variance of β_j is given by:

$$\begin{aligned}
&\mathrm{Var}(\beta_j \mid Y, \mathbf{X}) \\
&= \frac{\alpha_j w_{1,j}}{W_j}\left(\sigma_1^2 + \hat{\beta}_{1,j}^2\right) + \frac{(1-\alpha_j)w_{0,j}}{W_j}\left(\sigma_0^2 + \hat{\beta}_{0,j}^2\right) \\
&\qquad\qquad - \left[\frac{\alpha_j w_{1,j}\hat{\beta}_{1,j} + (1-\alpha_j)w_{0,j}\hat{\beta}_{0,j}}{W_j}\right]^2 \qquad (44) \\
&= \frac{\alpha_j w_{1,j}\sigma_1^2 + (1-\alpha_j)w_{0,j}\sigma_0^2}{W_j} + \frac{\alpha_j w_{1,j}\hat{\beta}_{1,j}^2 + (1-\alpha_j)w_{0,j}\hat{\beta}_{1,j}^2}{W_j} \\
&\qquad\qquad - \left[\frac{\alpha_j w_{1,j}\hat{\beta}_{1,j} + (1-\alpha_j)w_{0,j}\hat{\beta}_{0,j}}{W_j}\right]^2 \qquad (45) \\
&= \frac{\alpha_j w_{1,j}\sigma_1^2 + (1-\alpha_j)w_{0,j}\sigma_0^2}{W_j} + \frac{\alpha(1-\alpha)w_{1,j}w_{0,j}(\hat{\beta}_{1,j} - \hat{\beta}_{0,j})^2}{W_j^2}. \qquad (46)
\end{aligned}$$

Therefore, we get a set of posterior variances $\mathcal{S}_j$ such that:

$$\begin{aligned}
&\mathcal{S}_j \\
&= \left\{ \frac{\alpha_j w_{1,j}\sigma_1^2 + (1-\alpha_j)w_{0,j}\sigma_0^2}{W_j} + \frac{\alpha(1-\alpha)w_{1,j}w_{0,j}(\hat{\beta}_{1,j} - \hat{\beta}_{0,j})^2}{W_j^2} : \alpha_j \in (0,1) \right\}
\end{aligned} \tag{47}$$

where, $w_{k,j}$ and σ_k are as defined before.

4 Illustration

We analyse both synthetic datasets and a real dataset to illustrate our method.

4.1 Synthetic Datasets

We use three different synthetic datasets to showcase the performance of our method in terms of variable selection.

Synthetic Dataset 1: In this dataset, we construct an orthogonal design matrix $X_{i,j}$ with 100 predictors and 100 observations. We assign the regression coefficients to be, $(\beta_1, \cdots, \beta_6) \coloneqq (100, 125, -80, 100, 200, -150)$ and $\beta_j = 0$ for $j > 6$. We consider standard normal noise to construct the response vector $y_i = \sum_{j=1}^{6} X_{i,j}\beta_j + \epsilon_i$ where, $\epsilon_i \sim N(0,1)$ for $i = 1, \cdots, 100$. This setting allows us to evaluate the performance of our method with only strong non-zero effects. We analyse this dataset with two different sets of α_j's and three different choices of τ_1. We show the summary in Table 1.

Synthetic Dataset 2: In this case, we construct a similar design matrix as of synthetic dataset 1. We assign the regression coefficients such that the first 12 β_j's represent a strong effect and the next 20 β_j's represent a mild effect. We set $\beta_j = 0$ for $j > 32$. We construct the response vector in the following way: $y_i = \sum_{j=1}^{32} X_{i,j}\beta_j + \epsilon_i$ where, $\epsilon_i \sim N(0,1)$ for $i = 1, \cdots, 100$. This type of coefficient assignment allows us to investigate both small and large effects within the model. We analyse this dataset with two different sets of α_j's and three different choices of τ_1. We show the summary in Table 1. We observe that in this case, the choice of τ_1 plays an important role.

Synthetic Dataset 3: We use the third synthetic data set to illustrate the high dimensional case. We construct the design matrix with 100 observations and 200 predictors. We assign the first 12 regression coefficients to demonstrate large effects and the next 28 as small effects. We set the rest of the regression coefficients to be zero, ie, $\beta_j = 0$ for $j > 40$. We construct the response vector in a similar fashion as for synthetic datasets 1 and 2. We use two different sets of weights. We use three different τ_1's for each set of weights. We provide the summary in Table 1.

In all of the three cases, we also provide a comparison of different variable selection methods. We use `basad` [8], `blasso` [9] and `SSLASSO` [10] along with our method. We observe that in all the cases our method gives similar results to `blasso`. However, for `blasso`, we use the median values of the posteriors to identify the variables unlike our method of computing the posterior expectation of the latent variables. We also notice that for these three synthetic datasets, fixing $\tau_1 = 10$, gives us more accurate sets of active co-variates and also the number of indeterminate variables is less. We also observe that for high dimensional case, our method is more accurate in detecting the inactive variables unlike the other two datasets. We also see that our method does not identify an inactive co-variate as an active co-variate. However, for high dimensional case, our method identifies some of the small effects as inactive for smaller values of τ_1 and for larger values of τ_1 it tends to identify variables as indeterminate, which can be understood by Eq. (41) and Eq. (42).

4.2 Real Data Analysis

We investigate the gaia dataset to illustrate our method using real data. This dataset was used for computer experiments [3, 2] prior to the launch of European

	True Active			True Inactive		
Parameter Setting/ Method	Act	Inact	Indet	Act	Inact	Indet
Dataset 1, active 6 and inactive 94						
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 1$	6	0	0	0	62	32
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 10$	6	0	0	0	86	8
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 100$	6	0	0	0	78	16
$\alpha \in [0.05, 0.95]$, $\tau_0 = 10^{-6}$, $\tau_1 = 1$	6	0	0	0	0	94
$\alpha \in [0.05, 0.95]$, $\tau_0 = 10^{-6}$, $\tau_1 = 10$	6	0	0	0	69	25
$\alpha \in [0.05, 0.95]$, $\tau_0 = 10^{-6}$, $\tau_1 = 100$	6	0	0	0	63	31
BASAD	6	0	–	1	93	–
BLASSO (Median)	6	0	–	1	93	–
SSL (Double Exponential)	6	0	–	0	94	–
Dataset 2, active 32 and inactive 68						
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 1$	12	0	20	0	56	12
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 5$	32	0	0	0	68	0
$\alpha \in [0.1, 0.9]$, $\tau_0 = 10^{-6}$, $\tau_1 = 10$	32	0	0	0	68	0
$\alpha \in [0.3, 0.95]$, $\tau_0 = 10^{-6}$, $\tau_1 = 5$	32	0	0	0	51	17
$\alpha \in [0.3, 0.95]$, $\tau_0 = 10^{-6}$, $\tau_1 = 10$	32	0	0	0	63	5
$\alpha \in [0.3, 0.95]$, $\tau_0 = 10^{-6}$, $\tau_1 = 100$	32	0	0	0	57	11
BASAD	12	20	–	0	68	–
BLASSO (Median)	32	0	–	1	67	–
SSL (Double Exponential)	12	20	–	0	68	–
Dataset 3, active 40 and inactive 160						
$\alpha \in [0.1, 0.2]$, $\tau_0 = 10^{-6}$, $\tau_1 = 1$	14	26	0	0	160	0
$\alpha \in [0.1, 0.2]$, $\tau_0 = 10^{-6}$, $\tau_1 = 5$	16	14	10	0	160	0
$\alpha \in [0.1, 0.2]$, $\tau_0 = 10^{-6}$, $\tau_1 = 10$	19	0	21	0	160	0
$\alpha \in [0.2, 0.5]$, $\tau_0 = 10^{-6}$, $\tau_1 = 5$	22	2	16	0	159	1
$\alpha \in [0.2, 0.5]$, $\tau_0 = 10^{-6}$, $\tau_1 = 10$	40	0	0	0	160	0
$\alpha \in [0.2, 0.5]$, $\tau_0 = 10^{-6}$, $\tau_1 = 100$	17	0	23	0	102	58
BASAD	12	28	–	0	160	–
BLASSO (Median)	40	0	–	0	160	–
SSL (Double Exponential)	12	28	–	0	160	–

Table 1. Summary of variable selection for three different synthetic datasets.

Space Agency's Gaia mission [1]. The data contains spectral information of 16 (p) wavelength bands, and four different stellar parameters. In this example, we take stellar-temperature (in Kelvin scale) as the response variable. This dataset contains 8286 observations which are highly correlated. We show the correlation between the co-variates in Fig. 1. We randomly sample 100 (n) of them to fit our model and 100 more to measure the prediction accuracy. We standardise the dataset for the sake of clearer interpretation.

A robust Bayesian routine needs different measure(s) of accuracy as we don't have a single posterior for prediction. We introduce a new measure which can be considered to evaluate prediction accuracy and call it minimum squared error.

Let

$$\mathcal{A}(\alpha) := \left\{ j : \left\{ \frac{P\left(\gamma_j = 1 \mid \mathbf{X}, Y\right)}{P\left(\gamma_j = 0 \mid \mathbf{X}, Y\right)} \right\} > 1 \right\}. \tag{48}$$

Therefore, $\mathcal{A}(\alpha)$ or simply, $\mathcal{A}$ denotes the set of active variables for each value of α. We define minimum squared error by:

$$\text{Minimum Squared Error} = \min_{\alpha \in \mathcal{P}} \|Y - \mathbf{X}_{\mathcal{A}} \hat{\beta}_{\mathcal{A}}^{\text{post}}\|_2^2 \tag{49}$$

where $\hat{\beta}_{\mathcal{A}}^{\text{post}} := E(\beta_{\mathcal{A}} \mid Y)$ is the posterior mean of $\beta_{\mathcal{A}}$. The sensitivity analysis also creates an indeterminacy in prediction. Therefore, we define a similar measure called maximum squared error over the set of $\alpha \in \mathcal{P}$. We use both minimum and maximum squared error to introduce a new measure to capture the indeterminacy such that:

$$\text{Indeterminacy} = \frac{\text{Maximum Squared Error} - \text{Minimum Squared Error}}{\text{Maximum Squared Error}}. \tag{50}$$

Note that, for classical methods, indeterminacy is equal to zero.

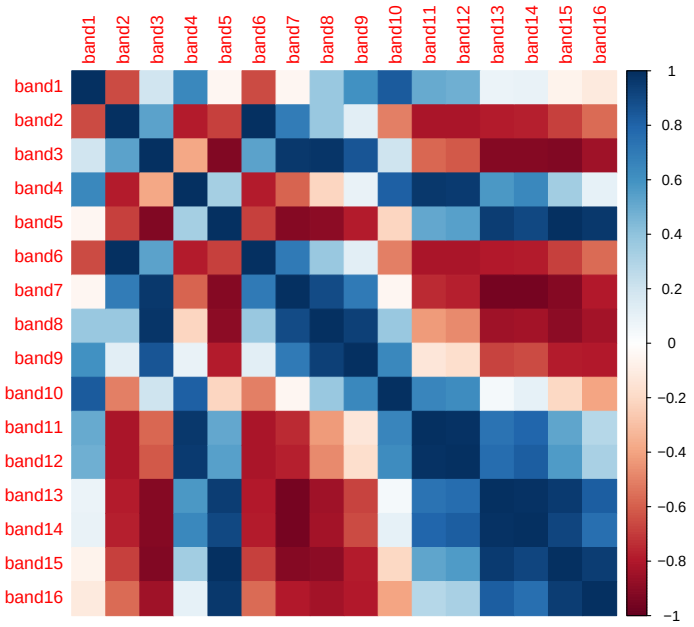

Fig. 1. Correlation plot matrix of the gaia dataset

Literature [3] suggests that this dataset contains 1-3 main contributory variables. Based on this information, we take two sets for α so that $\alpha \in [0.05, 0.2]$ and $\alpha \in [0.2, 0.4]$. We use `JAGS` to perform our analysis which we show in Table 2. We observe that for $\alpha \in [0.05, 0.2]$, our model performs better in terms of minimum squared error as well as indeterminacy. We observe that our method identifies only one active variable (band 6) irrespective of the choice of α. We also observe that unlike the synthetic datasets, we don't have a better choice of τ_1. The higher values of τ_1 result in smaller minimum squared errors. However, indeterminacy is much higher than the case where $\tau_1 = 1$. We notice that our method is in agreement with Spike and Slab lasso [10] and Bayesian Lasso [9]. For Bayesian lasso, we use the posterior median of the selected variables to fit the model instead of posterior mean. We see that for `basad` [8], it selects two active variables (band 2 and 6). This can be related to our setting where we have indeterminate co-variates contributing to higher minimum squared error.

Parameter Setting/ Method	Act	Inact	Indet	Min. SE	Indeterminacy
$\alpha \in [0.2, 0.4]$, $\tau_0 = 10^{-1}$, $\tau_1 = 1$	1	11	4	8.44	0.57
$\alpha \in [0.2, 0.4]$, $\tau_0 = 10^{-1}$, $\tau_1 = 10$	1	15	0	8.31	0.49
$\alpha \in [0.2, 0.4]$, $\tau_0 = 10^{-1}$, $\tau_1 = 50$	1	12	3	8.26	0.56
$\alpha \in [0.2, 0.4]$, $\tau_0 = 10^{-2}$, $\tau_1 = 1$	1	13	2	8.07	0.38
$\alpha \in [0.2, 0.4]$, $\tau_0 = 10^{-2}$, $\tau_1 = 10$	1	13	2	8.38	0.25
$\alpha \in [0.2, 0.4]$, $\tau_0 = 10^{-2}$, $\tau_1 = 50$	1	15	0	8.85	0.34
$\alpha \in [0.05, 0.2]$, $\tau_0 = 10^{-2}$, $\tau_1 = 1$	1	15	0	8.14	0.19
$\alpha \in [0.05, 0.2]$, $\tau_0 = 10^{-2}$, $\tau_1 = 10$	1	15	0	8.20	0.58
BASAD	2	14	–	10.83	0
BLASSO (Median)	1	15	–	8.16	0
SSL (Double Exponential)	1	15	–	8.14	0

Table 2. Comparison of different methods for the Gaia dataset.

5 Conclusion

Bayesian variable selection is a very important topic in modern statistics. In this paper, we discuss a novel and robust Bayesian variable selection routine based on the notion of spike and slab priors. The robustness within the hierarchical model is introduced using imprecise beta model which allows us to incorporate prior elicitation in a more flexible way. We inspect posterior properties of regression coefficients and selection indicators for the orthogonal design case. For the illustration of our method, we use three synthetic datasets covering different aspects of design matrices and a real life dataset to evaluate performance of our method for general cases. Under suitable scaling parameter, our method outperforms other methods in variable selection using the synthetic datasets. For the considered real dataset, it is in good agreement with other methods.

References

1. ESA science & technology: Gaia. http://sci.esa.int/gaia, accessed: 2018-02-06
2. Bailer-Jones, C.A.L.: The ILIUM forward modelling algorithm for multivariate parameter estimation and its application to derive stellar parameters from Gaia spectrophotometry. Monthly Notices of the Royal Astronomical Society **403**(1), 96–116 (2010). https://doi.org/10.1111/j.1365-2966.2009.16125.x, + http://dx.doi.org/10.1111/j.1365-2966.2009.16125.x
3. Einbeck, J., Evers, L., Bailer-Jones, C.: Representing complex data using localized principal components with application to astronomical data. In: Gorban, A.N., Kégl, B., Wunsch, D.C., Zinovyev, A.Y. (eds.) Principal Manifolds for Data Visualization and Dimension Reduction. pp. 178–201. Springer, Berlin, Heidelberg (2008)
4. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association **96**(456), 1348–1360 (2001). https://doi.org/10.1198/016214501753382273, https://doi.org/10.1198/016214501753382273
5. George, E.I., McCulloch, R.E.: Variable selection via Gibbs sampling. Journal of the American Statistical Association **88**(423), 881–889 (1993), http://www.jstor.org/stable/2290777
6. Ishwaran, H., Rao, J.S.: Spike and slab variable selection: Frequentist and bayesian strategies. Ann. Statist. **33**(2), 730–773 (04 2005). https://doi.org/10.1214/009053604000001147, https://doi.org/10.1214/009053604000001147
7. Lykou, A., Ntzoufras, I.: On Bayesian lasso variable selection and the specification of the shrinkage parameter. Statistics and Computing **23**(3), 361–390 (May 2013). https://doi.org/10.1007/s11222-012-9316-x
8. Narisetty, N.N., He, X.: Bayesian variable selection with shrinking and diffusing priors. Ann. Statist. **42**(2), 789–817 (04 2014). https://doi.org/10.1214/14-AOS1207, https://doi.org/10.1214/14-AOS1207
9. Park, T., Casella, G.: The Bayesian lasso. Journal of the American Statistical Association **103**(482), 681–686 (2008). https://doi.org/10.1198/016214508000000337
10. Ročková, V., George, E.I.: The spike-and-slab lasso. Journal of the American Statistical Association **113**(521), 431–444 (2018). https://doi.org/10.1080/01621459.2016.1260469
11. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **58**(1), 267–288 (1996), http://www.jstor.org/stable/2346178
12. Zou, H.: The adaptive lasso and its oracle properties. Journal of the American Statistical Association **101**(476), 1418–1429 (2006). https://doi.org/10.1198/016214506000000735