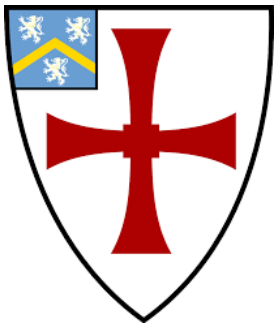


Durham University
Department of Mathematical Sciences

BAYESIAN ENRICHED POST-SELECTION MODELS FOR HIGH DIMENSIONAL DATA

Daniel Harvey Liddell

Supervisor:
Dr Reza Drikvandi
Durham University, Department of Mathematical Sciences

College of St Hild and St Bede
Durham University

Bayesian enriched post-selection models for high dimensional data

Daniel Liddell
Department of Mathematical Sciences, Durham University, Durham UK
Email: daniel.h.liddell@durham.ac.uk

Abstract

Over the past few years, regularisation approaches have become the key methods for analysing high dimensional data. The development of Bayesian statistics has only helped to further this research, with a range of Bayesian methods now available to analyse data in the high dimensional setting. However, being able to obtain accurate estimates and predictions alongside statistical inference remains a major and problematic challenge. In this paper, we introduce an enriched post-selection method within the Bayesian setting for analysing high-dimensional data. The main idea of the enriched post-selection approach, is to first use a variable selection procedure to split the model into the selected and unselected covariates. We then use these two sets to construct a post-selection model that includes an asymptotically accurate approximation of the unselected covariates. This is then used to enrich the selected model and eliminate bias in high dimensions. The variable selection step will first be examined using the lasso regression technique. We will then make use of the random properties of this variable selection to condition our model on the specific covariates which we select. This allows us to then develop a valid approximation to the unselected covariates. We then continue the theory further to make valid statistical inference by iliciting prior distributions to these sets, which we can then use to form posterior distributions. We then compare and evaluate it's performance with existing bayesian regularisation and shrinkage methods using simulation and real-data analysis. The enriched bayesian post-selection idea can be extended further to develop other areas too, including more generalised bayesian regression models.

Keywords: High dimensional data; *Horseshoe*; *Post-selection Model*; *Bayesian regularisation*; *Sparsity.*

1 Introduction

The term 'high dimensional' refers to the setting where the number of covariates p is much larger than the number of data points n (in other words, $p >> n$). Such situations are now commonplace due to a range of different technologies and methods collecting a large amount of covariates and data to attempt to better understand a given event, scenario or

situation of interest. There is a huge array of applications for high dimensional data, with uses varying from large-scale healthcare analytics to text/image analysis and astronomy (amongst many others). It is a well known statistical problem that classical regression techniques do not work in high dimensional situations.

In this paper, we consider the high dimensional linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad p >> n, \tag{1}$$

where $\mathbf{Y} = (y_1, ..., y_n)^T \in \mathbb{R}^n$ is the response vector, $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_p] \in \mathbb{R}^{n \times p}$ is the matrix of covariates, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T \in \mathbb{R}^p$ is the $p \times 1$ vector of regression coefficients and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ is the $n \times 1$ vector of random errors.
High dimensional regression analysis often requires a variable selection step in which we select a model with fewer covariates, say q ($q < n$), based on the given data. As mentioned in the abstract, we use the lasso approach in this paper. When following the lasso approach to linear regression (Tibshirani, 1996) we aim to solve the equation

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{2n} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda_{Lasso} ||\boldsymbol{\beta}||_1 \right\}, \tag{2}$$

where $||\mathbf{a}||_d := (\sum_{i=1}^t |\mathbf{a}_i|^d)^{1/d}$ is the standard L_d norm for a t-dimensional vector $\mathbf{a}$, and $\lambda_{Lasso} \geq 0$ is a regularisation parameter which shrinks the parameters towards zero in the shrinkage process. The properties of the lasso for estimation and variable selection are well understood, with details given in Buhlmann and Van De Geer (2011) and Sirimongkolkasem and Drikvandi (2019).

The main goal of this paper is to to formulate a bayesian the enriched post-selection model that aims to improve the parameter estimation and prediction capabilities of existing Bayesian regression techniques. Furthermore, we attempt to develop statistical inferences such as credible intervals in high dimensional situations. The key idea of this enriched post-selection approach is to use both the selected and unselected covariates obtained form the variable selection process (i.e the lasso) to construct a post-selection model that includes an asymptotically accurate approximation of the unselected covariates in a way which does not encounter the usual problems of the high dimensional setting. In the enriched model,

only the covariates selected by the variable selection process are used as regressors, but information from the unselected covariates is used to enrich and improve the selected model. We will show that within the bayesian setting, incorporating this appropriate approximation of the unselected covariates by iliciting appropriate prior distributions will lead to significant improvements in the estimates of selected covariates coefficients, and accurate predictions when compared to existing bayesian regression techniques.

2 Enriched post-selection models

When working in high-dimensional regression analysis, we often use a regularisation or variable selection process to select a small number of covariates (say $q,\ q < n$) whilst a large number of covariates ($p - q,\ p - q >> n$) are left out in the selected model. We thus say that the regularised variable selection method selects $\mathbf{X}_s\boldsymbol{\beta}_s$ to model the mean response, whilst it eliminates $\mathbf{X}_u\boldsymbol{\beta}_u$ from the model. We call the former the *core part* and the latter part the *ignored part* of the model. Hence our original model can be re-written as

$$\text{Response} = \text{Core Part} + \text{Ignored Part} + \text{Error (i.e, } \mathbf{Y} = \mathbf{X}_s\boldsymbol{\beta}_s + \mathbf{X}_u\boldsymbol{\beta}_u + \boldsymbol{\epsilon}), \tag{3}$$

whilst the selected model after variable selection is

$$\text{Response} = \text{Core Part} + \text{Error (i.e, } \mathbf{Y} = \mathbf{X}_s\boldsymbol{\beta}_s + \boldsymbol{\epsilon}). \tag{4}$$

The problem with this method, is that it is very likely that $\boldsymbol{\beta}_u$ contains a number of truly non-zero parameters, especially when the data is not sparse, or when p is very large when compared to n. As such, when carrying out the regression step, eliminating the ignored part $\mathbf{X}_u\boldsymbol{\beta}_u$ will affect both estimation and prediction. Examples of this can be found in Javanmard and Montanari (2019) and Sirimongkolkasem and Drikvandi (2019). The direct incorporation of the ignored part $\mathbf{X}_u\boldsymbol{\beta}_u$ into the selected model is potentially problematic because of the high-dimensional setting, even with the use of the regression techniques. We thus wish to approximately account for the ignored part so that the relevant covariates information of the unselected variables $\mathbf{X}_u$ is not fully ignored. In the context of linear regression, an enriched post selection model in the frequentist world is in the form of

$$\mathbf{Y} = \mathbf{X}_s\boldsymbol{\beta}_s + \text{Approx}\{\mathbf{X}_u\boldsymbol{\beta}_u\} + \boldsymbol{\epsilon}_s^*, \tag{5}$$

where Approx$\{\mathbf{X}_u\boldsymbol{\beta}_u\}$ is an appropriate approximation of the ignored part, and $\boldsymbol{\epsilon}_s^*$ denotes the error of the approximate model which depends on the selection outcome (Drikvandi 2022). As a result the likelihood in the bayesian setting is given by

$$\mathbf{Y}|\boldsymbol{\beta}_s, \mathbf{X}_s, \text{Approx}\{\mathbf{X}_u\boldsymbol{\beta}_u\}, {\sigma_s^*}^2 \sim N(\mathbf{X}_s\boldsymbol{\beta}_s + \text{Approx}\{\mathbf{X}_u\boldsymbol{\beta}_u\}, {\sigma_s^*}^2) \tag{6}$$

where Approx$\{\mathbf{X}_u\boldsymbol{\beta}_u\}$ is an appropriate bayesian approximation of the ignored part, and ${\sigma_s^*}^2$ is the variance of the approximate model which depends on the selection outcome.
One can consider any appropriate approximation of the ignored part of the model, however we will aim to use a consistent approximation of $\mathbf{X}_u\boldsymbol{\beta}_u$ (which is asymptotically accurate) instead of simply applying a prior to the vector of unselected covariates.
We follow three main steps to construct the enriched post-selection model, which are outlined in algorithm 1. The details of each step are shown below:

Algorithm 1: Construction of the enriched Bayesian post-selection model
1. We apply a bayesian variable selection procedure (in our case the lasso) to identify the selected covariates $\mathbf{X}_s$ and the unselected covariates $\mathbf{X}_u$ based on the observed data.
2. We ilicit a prior distribution to $\boldsymbol{\beta}_u$ (we shall use a horseshoe prior in this paper) which accounts for the sparsity of this unselected vector.
3. We build the enriched Bayesian post-selection model by incorporating the distributions of the core part $\mathbf{X}_s\boldsymbol{\beta}_s$ and ignored part $\mathbf{X}_u\boldsymbol{\beta}_u$, in order to form a valid posterior distribution from which we can make valid inferences.

We utilise the lasso selection as our variable selection procedure, but the proposed idea is general and the variable selection can be performed using other standard methods such as the elastic net, LARS and SCAD. Similarly, we can even bayesian methods such as the bayesian lasso (Park and Casella 2008) and the bayesian elastic net (Li and Lin 2010).

We let S denote the active set of the lasso, which is defined as

$$S = \{j : \hat{\beta}_{j,Lasso} \neq 0\}, \tag{7}$$

where $\hat{\boldsymbol{\beta}}_{Lasso}$ is the solution to equation (2). Clearly, we see that S is a random set, and so we can easily see that the outcome of lasso variable selection is random. In order to

make valid statistical statements and theory about our findings, we need to account for this randomness. We let $\{S = s\}$ denote the outcome of the lasso variable selection. We then follow the technique of Lee et al (2016), whereby we condition on the selected model $\{S = s\}$ to account for the randomness of this variable selection procedure. Without loss of generality. we suppose the selected covariates are $\mathbf{x}_1, ..., \mathbf{x}_q$. Then $y_i = \mathbf{X}_{s(i)}\boldsymbol{\beta}_s + \epsilon_{s(i)},\ i = 1, ..., n$ represents the selected model given $\{S = s\}$, and so the likelihood is given by

$$y_i|\mathbf{X}_{s(i)}, \boldsymbol{\beta}_s, \sigma^2 \sim N(\mathbf{X}_{s(i)}\boldsymbol{\beta}_s, \sigma^2). \tag{8}$$

We note that here, $\mathbf{X}_{s(i)} = (x_{i1}, ..., x_{iq})$ denotes the $1 \times q$ row vector of selected covariates for the ith data point and $\boldsymbol{\beta}_s$ are their associated regression parameters. In matrix notation, the lasso selected model given $\{S = s\}$ can be represented as $\mathbf{Y} = \mathbf{X}_s\boldsymbol{\beta}_s + \epsilon_s$, where $\mathbf{Y} = (y_1, ..., y_n)^T$ and $\mathbf{X}_s = [\mathbf{X}^T_{s(1)}, ..., \mathbf{X}^T_{s(n)}]^T$. We then write $\mathbf{X}_{u(i)} = (x_{i,q+1}, ..., x_{ip})$ to denote the $1 \times p - q$ row vector of unselected covariates for data point i. Thus, $\mathbf{X}_u = [\mathbf{X}^T_{u(1)}, ..., \mathbf{X}^T_{u(n)}]^T$ would be the $n \times p - q$ matrix of unselected covariates for all n data points. We can thus write $\mathbf{X} = [\mathbf{X}_s, \mathbf{X}_u]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}^T_s, \boldsymbol{\beta}^T_u)^T$.

Since $q << p$ in the high-dimensional setting, $\mathbf{X}_u$ contains most of the covariates. However, many of the covariates in $\mathbf{X}_u$ may not truly contribute to the unknown true underlying model. This tells us that the resulting matrix $\mathbf{X}_u$ will be sparse, hence we wish to assign an appropriate prior to $\boldsymbol{\beta}_u$ which accounts for this sparsity.

2.1 Parameter estimation

Due to the fact that the lasso selection set S is random, we know that the set of selected covariates $\mathbf{X}_s$ and the set of unselected covariates $\mathbf{X}_u$ are also random (and hence so are $\boldsymbol{\beta}_s$ and q). However, following the methods of Tibshirani (2013) and Lee et al. (2016) we know that all of these values will become deterministic if we condition on the selection event $\{S = s\}$. Conditioning on this event also allows us to avoid "using the data twice" by removing the information already used up during the variable selection process. We thus estimate the parameters of the enriched post-selection model by conditioning on $\{S = s\}$.

Remark 1. Note that in Lee et al. (2016) and Tibshirani (2013), the conditioning approach was used to obtain the conditional distribution of a given test statistic, whereas

we are using this condition to carry out parameter estimation within the enriched post-selection model. It is also useful to note that Lee et al. (2016) dealt with a univariate normal distribution, whereas in this paper we shall deal with a multivariate normal distribution instead.

Since we use the lasso selection to determine the selected and unselected covariates, it becomes easier to first condition on both the selected signs and the selected model of the estimates. Thus we first calculate the conditional distribution $f(\mathbf{Y}|S = s, \mathbf{\Psi} = \boldsymbol{\psi})$, in which the random vector $\mathbf{\Psi}$ denotes the signs of the lasso estimates, whose j-th element is defined as $\Psi_j = 1$ if $\hat{\boldsymbol{\beta}}_{j,Lasso} > 0$, $\Psi_j = -1$ if $\hat{\boldsymbol{\beta}}_{j,Lasso} < 0$, and $\Psi_j \in [-1, 1]$ if $\hat{\boldsymbol{\beta}}_{j,Lasso} = 0$. Now, following Tibshirani (2013) and Lee et al (2016), we consider the equicorrelation set where $|\Psi_j| = 1$.
it is known that

$$\{S = s, \mathbf{\Psi} = \boldsymbol{\psi}\} = \{\mathbf{A}(s, \boldsymbol{\psi})\mathbf{Y} \leq \mathbf{b}(s, \boldsymbol{\psi})\}, \tag{9}$$

with the matrix $\mathbf{A}(s, \boldsymbol{\psi})$ and vector $\mathbf{b}(s, \boldsymbol{\psi})$ given in Theorem 4.3 of Lee et al. (2016) amongst others.
The event $\{S = s\}$ is then the union of the above events over all possible sign patterns which is as follows

$$\{S = s\} = \bigcup_{\boldsymbol{\psi} \in \{-1,1\}^q} \{\mathbf{A}(s, \boldsymbol{\psi})\mathbf{Y} \leq \mathbf{b}(s, \boldsymbol{\psi})\}. \tag{10}$$

The above result is known as the polyhedral lemma, and tells us that the lasso selected model can be written as a union of polyhedra.

It is straightforward to find that the unconditional distribution is a multivariate normal distribution as follows

$$\mathbf{Y} \sim N(\mathbf{X}_s\boldsymbol{\beta}_s, \sigma^2\mathbf{I}_n). \tag{11}$$

Now, because equation (9) is an affine constraint on $\mathbf{Y}$, the conditional distribution in the proposed selection model is a truncated multivariate normal distribution on $\mathcal{R}_{s,\boldsymbol{\psi}}$ as follows

$$\mathbf{Y}|\{S = s, \mathbf{\Psi} = \boldsymbol{\psi}\} \sim TN(\mathbf{X}_s\boldsymbol{\beta}_s, \sigma^2\mathbf{I}_n, \mathcal{R}_{s,\boldsymbol{\psi}}), \tag{12}$$

where $\mathcal{R}_{s,\boldsymbol{\psi}} = \{\mathbf{A}(s,\boldsymbol{\psi})\mathbf{Y} \leq \mathbf{b}(s,\boldsymbol{\psi})\}$ is the polhedron for given s and $\boldsymbol{\psi}$. Then, similar to the argument in Lee et al. (2016), it is easy to see that the conditional distribution $f(\mathbf{Y}|S=s)$ is a truncated multivariate normal distribution on $\mathcal{R}_s$ given by

$$\mathbf{Y}|\{S=s\} \sim TN(\mathbf{X}_s\boldsymbol{\beta}_s, \sigma^2\mathbf{I}_n, \mathcal{R}_s), \tag{13}$$

where $\mathcal{R}_s = \cup_{\boldsymbol{\psi}\in\{-1,1\}^q}\{\mathbf{A}(s,\boldsymbol{\psi})\mathbf{Y} \leq \mathbf{b}(s,\boldsymbol{\psi})\}$is the union of polyhedra for given s.

The likelihood function in the post-selection approach is based on this above multivariate normal distribution. However, the analytic form of this distribution is difficult to characterise due to the complications of finding $\mathcal{R}_s$. We hence need an algorithm which allows us to characterise $\mathcal{R}_s$. We thus use an algorithm implemented in the programming language R to simulate this region, which we then use in our Bayesian analysis..

2.2 Bayesian Methodology

In the previous section, we detailed the frequentist variable selection procedure that we use to find $\mathbf{X}_s$ and $\mathbf{X}_u$. We thus now wish to provide our bayesian enrichment to the post-selection model. This involves assigning relevant prior distributions to $\boldsymbol{\beta}_s$, $\boldsymbol{\beta}_u$ and to the relevant variances in the model. First however, we remember that the likelihood of the response variable $\mathbf{Y}$ conditional on the lasso selection $\{S=s\}$ is a truncated multivariate normal distribution as follows

$$\mathbf{Y}|\{S=s\} \sim TN(\mathbf{X}_s\boldsymbol{\beta}_s, \sigma^2\mathbf{I}_n, \mathcal{R}_s). \tag{14}$$

Before assigning the priors which we wish to use in our model, we note that we have to be careful. Ideally, the priors we choose should accurately reflect any pre-existing knowledge we have about the model we are working in. Indeed, Bayesian analysis is found to outperform frequentist analysis if priors are accurately chosen to reflect the given model (Van Erp et al. 2018). However, iliciting priors is a time-consuming task, and even experts are prone to overstating the certainty in their choices (Garthwaite et al. 2005, Tversky 1974). Furthermore, it is often difficult to specify subjective priors due to the many parameters used, some of which will not be easily interpretable. For this reason, we aim to choose priors which are best suited to high dimensional setting.

Now, as we begin to assign our priors, we recall that $\boldsymbol{\beta}_s$ are the main parameters of interest, as these were originally selected by the lasso. We can hence use a normal prior for this data, as specified in Park and Casella (2008)

$$\boldsymbol{\beta}_s \sim N(\mathbf{0}, \sigma^2 \mathbf{D}_\kappa), \tag{15}$$

where $\mathbf{D}_\kappa = \text{diag}(\kappa_1^2, ..., \kappa_p^2)$ is a diagonal matrix which can be chosen in order to aid computations without loss of generality (Park and Casella, 2008). In our case, in order to make computations as easy as possible, we shall assume without loss of generality that all of the variances are independent and identically distributed so that $\mathbf{D}_\kappa = \mathbf{I}$.

Now, we wish to assign a prior to an asymptotic approximation of $\boldsymbol{\beta}_u$. Since $q << p$ in the high-dimensional setting, $\mathbf{X}_u$ contains most of the covariates. In the regression context, our post selection model is hence of the form

$$\mathbf{Y} = \mathbf{X}_s \boldsymbol{\beta}_s + \text{Approx}\{\mathbf{X}_u \boldsymbol{\beta}_u\} + \boldsymbol{\epsilon}_s^*, \tag{16}$$

where $\text{Approx}\{\mathbf{X}_u \boldsymbol{\beta}_u\}$ is the appropriate approximation to the ignored part mentioned previously, and $\boldsymbol{\epsilon}$ denotes the error of the approximate model which is again dependent on the selection outcome.

One can use any appropriate approximation of $\mathbf{X}_u \boldsymbol{\beta}_u$, but in this paper we will aim to use a consistent approximation which is asymptotically accurate. Furthermore, we know that as $\mathbf{X}_u$ contains the unselected covariates, many of the elements in $\mathbf{X}_u$ will have little or no effect on the response variable $\mathbf{Y}$. Despite this, $\mathbf{X}_u$ will also contain most of the covariates ($p - q,\ p - q >> n$ covariates to be precise), and hence calculations with $\mathbf{X}_u$ could be both inefficient and time consuming. It therefore makes sense to look for a lower dimensional approximation of $\mathbf{X}_u$ to use for $\text{Approx}\{\mathbf{X}_u \boldsymbol{\beta}_u\}$. As we know that everything we are working with, including the choice of underlying model, is random, we can obtain such an approximation of $\mathbf{X}_u$ using the singular value decomposition (SVD). We thus apply the SVD to the matrix $\mathbf{X}_u$ to get $\mathbf{X}_u = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\mathbf{U} = [u_1, ..., u_{q-p}]$ is an $n \times p - q$ orthogonal matrix satisfying $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{p-q}$, $\mathbf{V} = [v_i, ..., v_{p-q}]$ is a $p - q \times p - q$ orthogonal matrix satisfying $\mathbf{V}^T\mathbf{V} = \mathbf{I}_{p-q}$, and $\mathbf{D} = \text{diag}(d_1, ..., d_{p-q})$ is a $p-q \times p-q$ diagonal matrix such that $d_1 \geq d_2 \geq ... \geq d_{p-q} \geq 0$ are the ordered singular values. For any $k \leq p-q$, define

$\mathbf{U}_k = [u_1, ..., U_k]$, $\mathbf{D} = \text{diag}(d_1, ..., d_k)$, and let $\mathbf{V}_k$ be the $k \times k$ matrix containing the first k rows and columns of the loadings matrix $\mathbf{V}$. We then define $\mathbf{P}_k = \mathbf{U}_k\mathbf{D}_k$, which contains the first k principle components of $\mathbf{X}_u$. Then we have that $\mathbf{P}_k\mathbf{V}_k^T \rightarrow \mathbf{X}_u$ as $k \rightarrow p - q$. It is also clear from this that $\mathbf{P}_{p-q}\mathbf{V}_{p-q}^T = \mathbf{X}_u$. Next, we denote $\text{rank}(\mathbf{X}_u) = r_n$, which depends on the sample size n. As we are focussed on the high-dimensional setting, we always have $r_n \leq n$. This means that only the first r_n singular values of $\mathbf{X}_u$ are non-zero, i.e. $d_1 \geq d_2 \geq ... \geq d_{r_n} \geq 0, d_{r_n+1} = d_{r_n+2} = ... = d_{p-q} = 0$. Hence in practice, we can choose $k \leq r_n$ to use in our approximation. From a theoretical perspective, we can actually allow k to diverge, as $r_n \leq n \leq p$.

Now, we wish to utilise a lower dimensional approximation $\mathbf{P}_k\mathbf{V}_k^T$ of $\mathbf{X}_u$. To do this, we hence need a lower dimensional $k \times 1$ vector of parameters, which we denote $\boldsymbol{\theta}_k$, instead of $\boldsymbol{\beta}_u$. We thus use $\mathbf{P}_k\mathbf{V}_k^T\boldsymbol{\theta}_k$ as an approximation to $\mathbf{X}_u\boldsymbol{\beta}_u$. We have hence proposed the following enriched post-selection model

$$\mathbf{Y} = \mathbf{X}_s\boldsymbol{\beta}_s + \mathbf{P}_k\mathbf{V}_v^T\boldsymbol{\theta}_k + \boldsymbol{\epsilon}_{s,k}^*, \tag{17}$$

where $\boldsymbol{\epsilon}_{s,k}^*$ is indexed by k in order to be dependent on the approximation based on the given k. We assume this $\boldsymbol{\epsilon}_{s,k}^* \sim N(\mathbf{0}, \sigma_k^2\mathbf{I}_n)$ as we ideally wish to be able to say that $\mathbf{P}_k\mathbf{V}_k^T\boldsymbol{\theta}_k + \boldsymbol{\epsilon}_{s,k}^* \approx \mathbf{X}_u\boldsymbol{\beta}_u + \boldsymbol{\epsilon}$.

It is worth noting however that many of the covariates in $\mathbf{X}_u$ may not truly contribute to the unknown true underlying model. This tells us that the resulting matrix $\mathbf{X}_u$ will be sparse, hence we wish to assign an appropriate prior to $\boldsymbol{\theta}_k$ which accounts for this sparsity. In Van Erp et al. (2019), it was found that for high dimensional data, the best prior to use for sparse data is the horseshoe method (Carvalho et al. 2008), as this method allows for small coefficients to be heavily shrunken towards zero, whilst large coefficients remain large. In comparison, methods such as the bayesian lasso (Park and Casella 2008) will shrink all coefficients, which means that large coefficients which affect the data greatly are shrunk so that they appear to have a smaller effect, which is not a desirable property. For this reason, we shall use a horseshoe prior to select our variables.
The horseshoe prior has no analytic form, but can instead be represented as below:

$$(\boldsymbol{\theta}_k)_i|\lambda_i, \tau \sim N(0, \lambda_i^2\tau^2) \tag{18}$$

$$\lambda_i \sim \text{Half-Cauchy}(0, 1) \tag{19}$$

or, equivalently

$$(\boldsymbol{\theta}_k)_i|\delta^2 \sim N(0, \delta^2) \tag{20}$$

$$\delta|\omega \sim \text{Inverse-Gamma}(1/2, \omega) \tag{21}$$

$$\omega|\lambda \sim \text{Gamma}(1/2, \lambda_i^2) \tag{22}$$

$$\lambda_i^2|\gamma \sim \text{Inverse-Gamma}(1/2, \gamma) \tag{23}$$

$$\gamma|\tau^2 \sim \text{Gamma}(1/2, \tau^2). \tag{24}$$

with $i = 1, ..., p - q$, and where here, we have assigned λ_i a Half-Cauchy distribution (Gelman, 2006) in order to ensure a fully bayesian approach is being used from this point. It is however important to know that a cross validation method can also be used to specify λ_i (see Vehtari et al. 2018) but we shall use the Half-Cauchy distribution in our calculations. We also note here that in this case λ_i are the local shrinkage parameters, and τ are the global shrinkage parameters. However, one problem with this approach is that the Horseshoe prior has no analytic form, and hence the above equation for the posterior has no analytic form either. Thankfully however, we can however use the R package *Horseshoe* (Van der Pas et al. 2019) to evaluate the above information to form the posterior. The package is very useful as it allows us to specify a valid prior for τ^2 and a Half-Cauchy prior for λ_i. Given this information, we can use the horseshoe prior as a valid technique to form posterior values; and from this we are able to make bayesian inferences such as posterior means and credible intervals (Carvalho et al. 2010). As we are working in the multivariate setting, and to make calculations easier, we note that the λ_i are independent and identically distributed Half-Cauchy random variables, and thus we can find their joint probability density function as below:

$$\pi(\boldsymbol{\lambda}) = \prod_{i=1}^{p-q} \pi(\lambda_i) = \prod_{i=1}^{p-q} \pi^{-1}(1 + \lambda_i)^{-2}. \tag{25}$$

We can then rewrite the horseshoe prior for $\boldsymbol{\theta}_k|\tau, \boldsymbol{\lambda}$ in the multivariate setting as

$$\boldsymbol{\theta}_k|\tau, \boldsymbol{\lambda} \sim N(\mathbf{0}, \boldsymbol{\lambda}^T\boldsymbol{\lambda}\tau^2\mathbf{I}_{p-q}), \tag{26}$$

where $\boldsymbol{\lambda} = (\lambda_i, ..., \lambda_{p-q}^T)$.

Finally, we need to specify our prior distributions for the variance of the core part σ_s^2, and the variance of the ignored part τ^2 (remembering we are allowing each covariate to have independent and identically distributed variance). We follow the convention set out in Park and Casella (2008), and take both variances to have an exponential distribution. Hence we have that

$$\pi(\sigma_s) = e^{-\lambda\sigma_s} \tag{27}$$

$$\pi(\tau) = e^{-\lambda\tau}. \tag{28}$$

In order to carry out robust Bayesian analysis, we need to find the posterior distribution of our enriched post-selection model. We know that

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}. \tag{29}$$

we wish therefore to use the priors specified above to give the posterior distribution. In general, our posterior distribution will be given by

$$\boldsymbol{\beta}_s, \boldsymbol{\beta}_u | \mathbf{Y}, \sigma_s^2, \tau^2 \sim f(\mathbf{Y}|\{S = s\})\pi(\sigma^2)\pi(\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_u|\boldsymbol{\lambda}, \tau)\pi(\boldsymbol{\lambda})\pi(\tau). \tag{30}$$

We can then substitute in the density functions which we know,

$$\pi(\boldsymbol{\beta}_s) = \det(2\pi\sigma_s^2\mathbf{I})^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\boldsymbol{\beta}_s^T(\sigma_s^2\mathbf{I})^{-1}\boldsymbol{\beta}_s\right\} \tag{31}$$

$$\pi(\boldsymbol{\beta}_u|\boldsymbol{\lambda}, \tau) = \det(2\pi\boldsymbol{\lambda}^T\boldsymbol{\lambda}\tau^2\mathbf{I})^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\boldsymbol{\beta}_u^T(\boldsymbol{\lambda}^T\boldsymbol{\lambda}\tau^2\mathbf{I})^{-1}\boldsymbol{\beta}_u\right\} \tag{32}$$

$$\pi(\boldsymbol{\lambda}) = \frac{1}{\prod_{i=1}^{p-q}(1+\lambda_i)^2} \tag{33}$$

$$\pi(\sigma_s) = \exp\{-\lambda\sigma_s\} \tag{34}$$

$$\pi(\tau) = \exp\{-\lambda\tau\}. \tag{35}$$

Thus we can calculate the posterior as being

$$\rho(\boldsymbol{\beta}_s, \boldsymbol{\theta}_k|\mathbf{Y}, \sigma_s^2, \tau^2, \boldsymbol{\lambda}) \propto \frac{\det(2\pi\sigma^2\mathbf{I})^{-\frac{1}{2}}\det(2\pi\sigma_s^2\mathbf{I})^{-\frac{1}{2}}\det(2\pi\boldsymbol{\lambda}^T\boldsymbol{\lambda}\tau^2\mathbf{I})^{-\frac{1}{2}}}{\left(\prod_{i=1}^{k}\pi\,(1+\lambda_i)^2\right)\left(\boldsymbol{\Phi}(\mathbf{y}^{upper}) - \boldsymbol{\Phi}(\mathbf{y}^{lower})\right)} \times$$
$$\exp\left\{-\frac{1}{2}\left[(\mathbf{Y} - \mathbf{X}_s\boldsymbol{\beta}_s)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{Y} - \mathbf{X}_s\boldsymbol{\beta}_s) + \boldsymbol{\beta}_s^T(\sigma_s^2\mathbf{I})^{-1}\boldsymbol{\beta}_s + \boldsymbol{\theta}_k^T(\boldsymbol{\lambda}^T\boldsymbol{\lambda}\tau^2\mathbf{I})^{-1}\boldsymbol{\theta}_k\right] - \lambda(\sigma_s + \tau)\right\}. \tag{36}$$

, where $\mathbf{y}^{lower}$ and $\mathbf{y}^{upper}$ are the limits of $\mathcal{R}_s$, and $\Phi(.)$ is the cumulative distribution function of a standard normal distribution. From this distribution, we are able to make valid inferences and form credible intervals as desired.

3 Bayesian post-selection inference

As discussed in the introduction, it is often very difficult to conduct statistical inference for regression parameters in high-dimensional situations. However, the Bayesian enriched selection approach allows us to do just that, as it allows us to construct valid credible intervals and hypothesis tests for parameters in these high dimensional linear regression models. The reason these inferences are valid, is due wholly to the fact that our inferences are conditioned on the selection event $\{S = s\}$. In order to represent this, we consider first parameters of the form $\mathbf{a}^T\boldsymbol{\beta}_s$, where $\mathbf{a}$ is an arbitrary vector whose components are either 0 or 1. In particular, we focus on the j-th regression coefficient $\boldsymbol{\beta}_{s(j)}$, $j = 1, ..., q$. We then develop a simultaneous post-selection testing procedure for a family of hypotheses using the bayesian enriched post selection method.

3.1 Individual Bayesian Inference

First, we suppose that the parameter of interest is of the form $\mathbf{a}^T\boldsymbol{\beta}_s$, and we know that the estimate of the asymptotic variance of $\boldsymbol{\beta}_s^*$ conditional on the selection event $\{S = s\}$ is calculated from the posterior distribution as $\boldsymbol{\Omega}_n^{*\,-1}$. We let the (j, j')-th element of this variance matrix be denoted by $\boldsymbol{\Omega}^{*-1}_{n(,j,j')}$. Hence, an asymptotically valid post-selection equal tailed credible interval for $\mathbf{a}^T\boldsymbol{\beta}_s$ with significance level $1 - \alpha$ is given by

$$\mathbb{P}[\mathbf{a}^T\boldsymbol{\beta}^* - \Phi(1-\alpha/2)\sqrt{\mathbf{a}^T\boldsymbol{\Omega}_n^{*\,-1}\mathbf{a}} \leq \mathbf{a}^T\boldsymbol{\beta}_s \leq \mathbf{a}^T\boldsymbol{\beta}_s^* + \Phi(1-\alpha/2)\sqrt{\mathbf{a}^T\boldsymbol{\Omega}_n^{*\,-1}\mathbf{a}}|S = s] \to 1 - \alpha, \tag{37}$$

where here Φ denotes the CDF of the standard normal distribution.

As a special case of equation (37) as asymptotically valid post-selection equal-tailed credible interval for an individual regression coefficient $\boldsymbol{\beta}_{s(j)}$ with significance level $1 - \alpha$ would be $\boldsymbol{\beta}_s^* \pm \Phi(1-\alpha/2)\sqrt{\boldsymbol{\Omega}^{*-1}_{n(j,j)}}$.

An important property which we wish for these credible intervals to have, is that the length of the intervals, L_n, will shrink towards zero as $n \to \infty$. We will show in our simulation studies that this desirable property does actually turn out to be achieveable.

4 Simulation studies

In this section, we first conduct simulation studies to evaluate the finite-sample properties of the enriched Bayesian post-selection method. We compare its estimation and prediction performance in comparison with both lasso and ridge regression. Additionally, we also investigate how the model compares with the naive least squares approach on the lasso selected variables. We can then evaluate the finite-sample behaviour of the post selection posterior distribution and credible intervals based on this approach.

In the simulations, we generate 100 datasets from the high dimensional linear regression model $y_i = \sum_{j=1}^{(p-1)} \beta_j x_{ij} + \epsilon$, $i = 1, ..., n$ for each of the combinations of $n \in \{100, 200, 300\}$ and $p \in \{500, 1000, 5000, 10000\}$. For each dataset, the covariates are generated from a standard normal distribution, and the random errors are sampled from $N(0, \sigma^2)$ where we set $\sigma_0^2 = 1$. We generate $s_0 = |S_0|$ of the true regression coefficients from this standard normal distribution, and set the other $p - s_0$ coefficients to be zero. Now, unlike many previous works (including Van de Geer et al. 2014 and Zhang and Zhang, 2014) that use a small sparsity index, we shall focus on using relatively large values for the sparsity index in our simulations. Specifically, we shall use $s_0 = 20$, $s_0 = 50$ and $s_0 = 100$ in our simulations. It is important to note that our simulations include the cases with notably large dimensions of $p = 5000$ and $p = 10000$. Now, in the theory of high dimensional data, any data which we have is called sparse if $s_0 \leq n/\log(p)$, and data is called non-sparse otherwise. Thus in our setting, $s_0 = 20$ roughly represents sparse data, whereas $s_0 = 50$ and $s_0 = 100$ can be referred to as non-sparse.

To assess the accuracy of parameter estimation and prediction, we use the mean absolute bias (MAB) over all the regression parameter estimates. We also use the mean squared prediction error (MSPE) on unseen test data of size $T = 0.33n$ respectively. We define

these values as below:

$$\text{MAB} = \frac{1}{|S_{\text{fit}} \cap S_0|} \sum_{j \in S_{\text{fit}} \cap S_0} |\beta_j - \beta_j^0|, \tag{38}$$

$$\text{MSPE} = \frac{1}{T} \sum_{i=1}^{T} (y_i^* - y_i)^2, \tag{39}$$

where here β_j^0 are the true parameter values for β_j, and y_i^* are the posterior predictive values for y_i. Also, S_{fit} and S_0 are, respectively, the active sets of the fitted and true underlying models. We compute the average of MAB and MSPE over the 100 simulation replications.

4.1 Simulations on estimation and prediction

For simplicity of exposition, we here write BE-PoS instead of Bayesian enriched post-selection. We also note that, for the sake of space, all of the figures mentioned in this section are to be found in Appendix A. The simulation results under the above settings calculate the mean absolute bias of the proposed BE-PoS model, along with those of the lasso and ridge regression for both the lower dimensions ($p = 500$ and $p = 1000$) and higher dimensions ($p = 5000$ and $p = 10000$). The results of these simulations are presented in Figure 1 and Figure 2, respectively. We can see from these results that the proposed BEPoS approach produces a smaller MAB then the lasso and ridge regression in the scenarios considered. The bias of the method decreases as the sample size increases, which is in line with our theoretical results in Section 5.

Similarly, the simulation results for the mean squared prediction error of the aforementioned methods for lower dimensions $p = 500$ and $p = 1000$ and higher dimensions $p = 5000$ and $p = 10000$ are shown in Figure 3 and Figure 4 respectively. The simulations calculate the MSPE for the new BE-PoS method, as well as for the lasso, bootstrapped lasso (Chatterjee and Lahiri, 2011), principle components regression and ridge regression. It is seen that the BE-PoS method has a smaller prediction error than these other methods, even in the non-sparse cases of $s_0 = 50$ and $s_0 = 100$. This is represented in Figure 4e and 4f particularly clearly.

Figure 5 shows the estimates of the predicted variance, σ^2 over all 100 replications for each simulation setting using the BE-PoS approach. We can see that σ^2 is very close to the true

value σ_0^2 in the sparse case of $s_0 = 20$. For the non-sparse cases of $s_0 = 30$ and $s_0 = 100$ we see that σ^2 gets closer to the true value as n increases. It is however noticeable that the rate of convergence is lower when the sparsity index is very large ($s_0 = 50$ and $s_0 = 100$) and the dimension is high ($p = 5000$ and $p = 10000$). However, this is to be expected, seen as in these cases there are many truly non-zero and zero parameters in the model, and hence a larger sample size is required for accurate estimation of the error variance.

4.2 Simulations on statistical inference

We evaluate the empirical performance of the BE-PoS method for statistical inference on the regression coefficients corresponding to the active set S. As we are working in the bayesian setting, we thus focus on credible intervals for the individual coefficients $\boldsymbol{\beta}_{s(j)}$. In the simulations, we calculate the average length of the credible intervals for each $\boldsymbol{\beta}_{s(j)}$ at a significance level of $\alpha = 0.95$. The empirical average length of a credible interval CrI for the parameter $\boldsymbol{\beta}_{s(j)}$ over 100 simulation replications is defined as

$$AvgL_n = \frac{1}{100}\sum_{i=1}^{100}\left\{q^{-1}\sum_{j\in S} L_n(CrI_j)\right\} \tag{40}$$

The results in Table 2 (located in Appendix B) show that the BE-PoS approach provides credible intervals with relatively short lengths, which is a promising performance, especially for the non-sparse cases ($s_0 = 50$ and $s_0 = 100$ where confidence intervals are generally high. Furthermore, we see that the lengths of the credible intervals decrease as the sample size n increases, as was mentioned to be preferable in Section 6.

5 Real-data example

We apply the proposed Bayesian enriched post-selection model to the Riboflavin data made publicly available by Buhlmann, Kalisch and Meier (2014). The data was obtained from a high-throughput genomic study which focused on the production of riboflavin (otherwise known as vitamin B_2) by bacillus subtilis. The objective of the study, was to find out which genes are associated with the production rate of vitamin B_2. The data contains $n = 71$ samples and $p = 4088$ covariates, each of which corresponds to 1 of 4088 genes. The

response variable $\mathbf{Y}$ is also contained, and represents the logarithm of riboflavin production rate.

We first compare the prediction performance of the BE-PoS method on the riboflavin data in comparison with the Lasso, bootstrapped Lasso, principle components regression (PCR) and ridge regression. In order to do this, we partition the data 100 times into training (90%) and test (10%) sets. We apply all five methods to each of the training sets, and compute the mean squared prediction errors (MSPEs, given in equation 54) on the corresponding test sets. The average MSPE over all replications, of the proposed BE-PoS model (along with that of the four other methods) is given in Table 1.

The average MSPE on the 100 test data sets				
BE-PoS	Ridge	Lasso	Bootstrap Lasso	PCR
0.23	0.25	0.39	0.51	0.45

Table 1: The BE-PoS simultaneous post-selection credible intervals for the regression coefficients associated with the 17 selected genes in the riboflavin data.

The results indicate that the BE-PoS model and ridge regression perform relatively equally, and produce a smaller prediction error compared to the lasso, bootstrapped lasso and PCR. The bootstrapped lasso and PCR show the worst performance amongst all methods considered. It is however important to note here that the sample size of the training sets is $n = 63$, which may be too small compared to $p = 4088$ for a fair prediction comparison with these other methods.

Next, we model the riboflavin data using the BE-PoS approach and apply the proposed simultaneous post-selection testing procedure to find out which genes are significantly associated with the vitamin B_2 production rate. To determine the selected and unselected genes for the BE-PoS model, we fit the Bayesian Lasso using the R function *glmnet*, with λ_{Lasso} chosen using the strategy in Negahban et al. (2012). This results in $\lambda_{Lasso} = 0.156$, with which the lasso selects 17 genes, as well as an intercept term. Then, we apply our post-selection simultaneous testing procedure to simultaneously test which of the genes are significant. The BE-PoS model (with the horseshoe prior specified in Section 2) finds 3 significant genes, YEBB-at, YLXD-at and YLXE-at. For comparison purposes, the boot-

strapped lasso finds no significant gene, and the multisample splitting method (proposed by Meinshausen et al., 2009) finds one significant gene, YXLD-at, at the FWER-adjusted 5% significance level. Furthermore, the testing procedure developed by Javanmard and Montanari (2014) finds two significant genes, YXLD-at and YXLE-at, as reported in their paper. Thus we see that these methods seem to be more conservative than the BE-PoS method for simultaneous inference and do not take model selection into account. Figure 6 (located in Appendix B in order to save space) shows the (95%) simultaneous post-selection confidence intervals the the coefficients of the 17 selected genes based on the BE-PoS model. These genes have an average credible interval length of 1.564 using the BE-PoS method.

6 Conclusion

The Bayesian enriched post-selection method is a new approach for estimation, prediction and valid statistical inference within the Bayesian setting for high dimensional data. Our simulation results suggest that the newly proposed method performs well and is superior to the Bayesian lasso, Bootstrapped lasso and ridge regression in terms of parameter estimation and prediction performance. This is particularly true when the data is non-sparse ($s_0 = 100$ in our simulation settings) as well as in higher dimensions ($p = 5000$ and $p = 10000$ in our simulations). The incorporation of an appropriate prior distribution allows us to enhance the model with new information, without treating the unselected covariates as regressors. This hence allows us to retain interpretability, which proves to be important in the often confusing setting of high dimensional analysis.

One key use of the Bayesian enriched post-selection method is statistical inference, especially for simultaneous inference for regression parameters $\boldsymbol{\beta}_s$. Our empirical results demonstrate the good performance of the post-selection credible intervals based on our method. This idea was further solidified by our real data analysis which provided short credible interval lengths.

Acknowledgements

The author would like to thank the Laidlaw foundation for providing the funding nec-

essary to complete this research project. Special thanks also go to Dr Reza Drikvandi from Durham University for providing invaluable insight and new ideas as supervisor for this project. Finally, thanks go to the department of mathematical sciences at Durham University for the use of the Hamilton 8 supercomputer to run the necessary simulations for this paper.

7 References

Berger, J. O. (2006). The case for objective bayesian analysis. Bayesian Analysis, 3, 385-402.

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013), 'Valid post-selection inference', The Annals of Statistics 41, 802–837.

Bühlmann, P., Kalisch, M. and Meier, L. (2014) High-dimensional statistics with a view towards applications in biology. Annual Review of Statistics and its Applications 1, 255–278

Buhlmann, P., Meier, L. and van de Geer, S. (2014), 'Discussion: A significance test for the lasso', The Annals of Statistics 42, 469–477.

Buhlmann, P. and Van De Geer, S. (2011), Statistics for high-dimensional data, Springer Series in Statistics. Springer, New York.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2008). Handling Spar- sity via the Horseshoe. Journal of Machine Learning Research, W and CP 5 73– 80.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. Biometrika, 97(2):465–480.

Chatterjee, A., and Lahiri, S. N. (2011). Bootstrapping Lasso Estimators. Journal of the American Statistical Association, 106(494), 608–625. http://www.jstor.org/stable/41416396

Dezeure, R., Buhlmann, P., Meier, L. and Meinshausen, N. (2015), 'High-dimensional infer- ence: confidence intervals, p-values and R-software hdi', Statistical Science 30, 533–558.

Drikvandi, R. (2022). 'Enriched post-selection models for high dimensional data'.

Fan, J., Guo, S. and Hao, N. (2012), 'Variance estimation using refitted cross-validation in ultrahigh dimensional regression', Journal of the Royal Statistical Society: Series B 74, 37– 65.

Friedman, J., Hastie, T. and Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', Journal of Statistical Software 33, 1.

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. Journal of the American Statistical Association, 100(470), BSEM: SENSITIVITY TO THE PRIOR 63 680–701.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). Bayesian Analysis, 1(3):515–534.

Ghosh, P., Tang, X., Ghosh, M. and Chakrabati, A. (2013). Asymptotic properties of Bayes risk of a general class of shrinkage Priors in Multiple Hypothesis Testing under Sparsity. arXiv: Statistics Theory.

Griffin, J. E. and Brown, P. J. (2005). Alternative prior distributions for vari- able selection with very many more variables than observations. University of Warwick. Centre for Research in Statistical Methodology.

Hobert, J. P., and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. Journal of the American Statistical Association, 91(436), 1461-1473.

Hsiang, T. C. (1975). A bayesian view on ridge regression. The Statistician, 24(4):267.

Javanmard, A. and Montanari, A. (2014), 'Confidence intervals and hypothesis testing for high-dimensional regression', Journal of Machine Learning Research 15, 2869–2909.

Knight, K., Fu, W. "Asymptotics for lasso-type estimators." Ann. Statist. 28 (5) 1356 - 1378, October2000. https://doi.org/10.1214/aos/1015957397

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and bayesian lassos. Bayesian Analysis, 5(2):369–411.

Lee, J., MacEachern, S.N. (2011) Consistency of Bayes estimators without the assumption that the model is correct J. Statist. Plann. Inference, 141 (2) (2011), pp. 748-757

Lee, J.D., Sun, D.L., Sun, Y., Taylor, J.E. (2016) "Exact post-selection inference, with application to the lasso." Ann. Statist. 44 (3) 907 - 927, June 2016. https://doi.org/10.1214/15-AOS1371

Leeb, H., Potscher, B. M. and Ewald, K. (2015), 'On various confidence intervals post-model- selection', Statistical Science 30, 216–227.

Li, Q. and Lin, N. (2010). The bayesian elastic net. Bayesian Analysis, 5(1):151–170.

Meinshausen, N., Meier, L. and Buhlmann, P. (2009), 'p-values for high-dimensional regres- sion', Journal of the American Statistical Association 104, 1671–1681.

Minnier, J., Tian, L. and Cai, T. (2011), 'A perturbation method for inference on regularized regression estimates', Journal of the American Statistical Association 106, 1371–1382.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012), 'A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers', Statistical Science 27, 538–557.

Ning, Y. and Liu, H. (2017), 'A general theory of hypothesis tests and confidence regions for sparse high dimensional models', The Annals of Statistics 45, 158–195.

Panigrahi, S., J. Taylor, and A. Weinstein (2016). Bayesian post-selection inference in the linear model. arXiv preprint arXiv:1605.08824.

Park, T. and Casella, G. (2008). The bayesian lasso. Journal of the American Statistical Association, 103(482):681–686.

Rigollet, P. and Tsybakov, A. (2011), 'Exponential screening and optimal rates of sparse estimation', The Annals of Statistics 39, 731–771.

Rinaldo, A., Wasserman, L. and G'Sell, M. (2019), 'Bootstrapping and sample splitting for high-dimensional, assumption-lean inference', The Annals of Statistics 47, 3438–3469.

Sirimongkolkasem, T. and Drikvandi, R. (2019), 'On regularisation methods for analysis of high dimensional data', Annals of Data Science 6, 737–763.

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

Tibshirani, R. J. (2013), 'The lasso problem and uniqueness', Electronic Journal of Statistics 7, 1456–1490.

Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2016), 'Exact post-selection inference for sequential regression procedures', Journal of American Statistical Association 111, 600–620.

Tversky, A. (1974). Assessing uncertainty. Journal of the Royal Statistical Society. Series B (Methodological), 148–159.

Van de Geer, S. (2007), The deterministic lasso, In JSM proceedings, 2007, 140. American Statistical Association.

Van de Geer, S., Buhlmann, P., Ritov, Y. and Dezeure, R. (2014), 'On asymptotically optimal confidence regions and tests for high-dimensional models', The Annals of Statistics 42, 1166– 1202.

Van der Pas, Scott, J., Chakraborty, A., Bhattacharya, A. (2019). S., https://CRAN.R-project.org/package=horseshoe

Van Erp, S., Mulder, J., and Oberski, D. L. (2018). Prior sensitivity analysis in default bayesian structural equation modeling. Psychological Methods, 23(2):363–388.

Van Erp, S., Oberski, D. L., and Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. Journal of Mathematical Psychology, 89, 31–50. https://doi.org/10.1016/j.jmp.2018.12.004

Vehtari, A., Gabry, J., Yao, Y., and Gelman, A. (2018). loo: Efficient leave- one-out cross-validation and waic for bayesian models. R package version 2.0.0.

Yekutieli, D. (2012). Adjusted bayesian inference for selected parameters. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74(3):515-541.

Wasserman, L. and Roeder, K. (2009), 'High dimensional variable selection', The Annals of Statistics 37, 2178–2201.

Yuan, M. and Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', Journal of the Royal Statistical Society: Series B 68, 49–67.

Zhang, C. H. and Zhang, S. S. (2014), 'Confidence intervals for low dimensional parameters in high dimensional linear models', Journal of the Royal Statistical Society: Series B 76, 217– 242.

Zhao, S., Witten, D. and Shojaie, A. (2021), 'In defense of the indefensible: A very naive approach to high-dimensional inference', Statistical Science 36, 562–577.

Zhu, Y. and Bradic, J. (2018), 'Significance testing in non-sparse high-dimensional linear models', Electronic Journal of Statistics 12, 3312–3364.

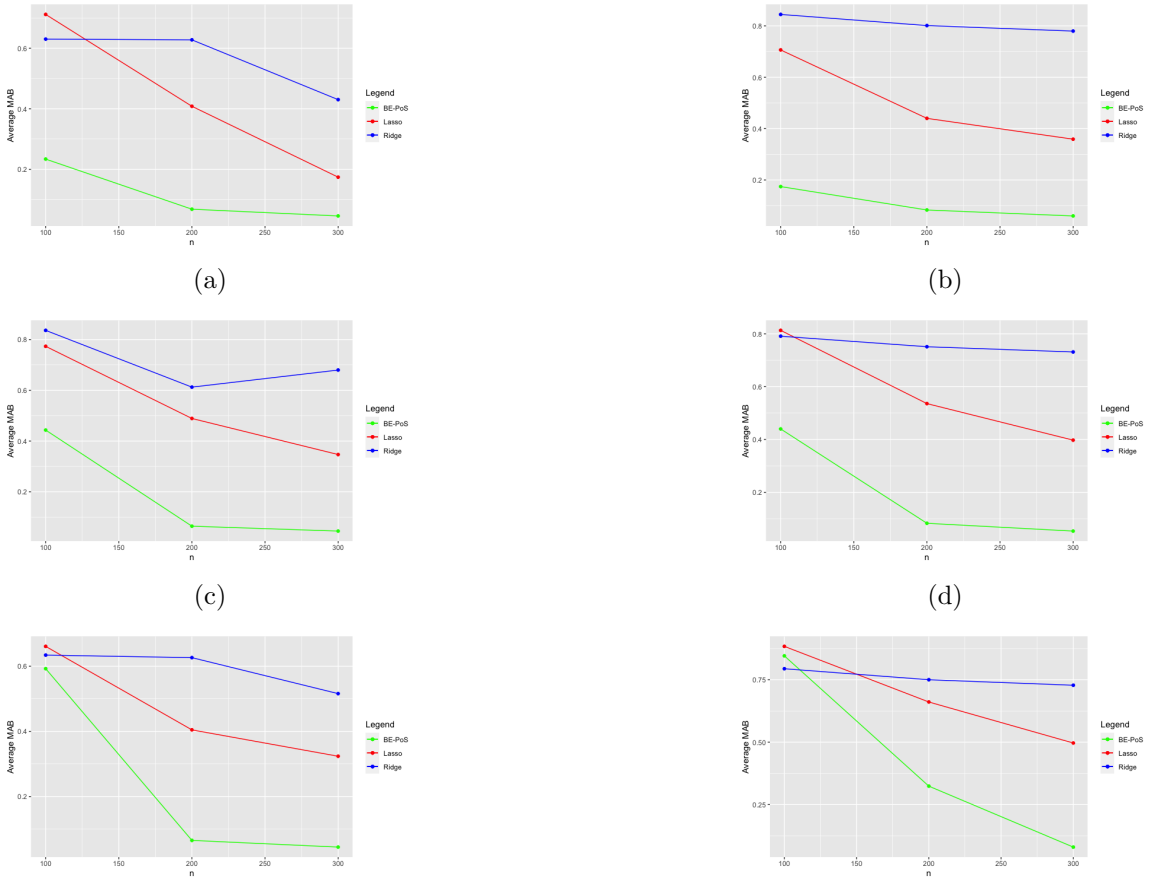
Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', Journal of the Royal Statistical Society: Series B 67, 301–320.

Zhao, P. and Yu, B. (2006), 'On model selection consistency of lasso', Journal of Machine Learning Research 7, 2541–2563.

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

Appendices

A Simulation results



(a) (b) (c) (d) (e) (f)

Figure 1: Average mean absolute bias (MAB) of the proposed BE-PoS method, lasso, bootstrapped lasso, ridge regression and PCR in the lower dimensions $p = 500$ and $p = 1000$. (a) $s_0 = 20,\ p = 500$, (b) $s_0 = 20,\ p = 1000$, (c) $s_0 = 50,\ p = 500$, (d) $S_0 = 20,\ p = 1000$, (e) $s_0 = 100,\ p = 500$, (f) $s_0 = 100,\ p = 1000$.

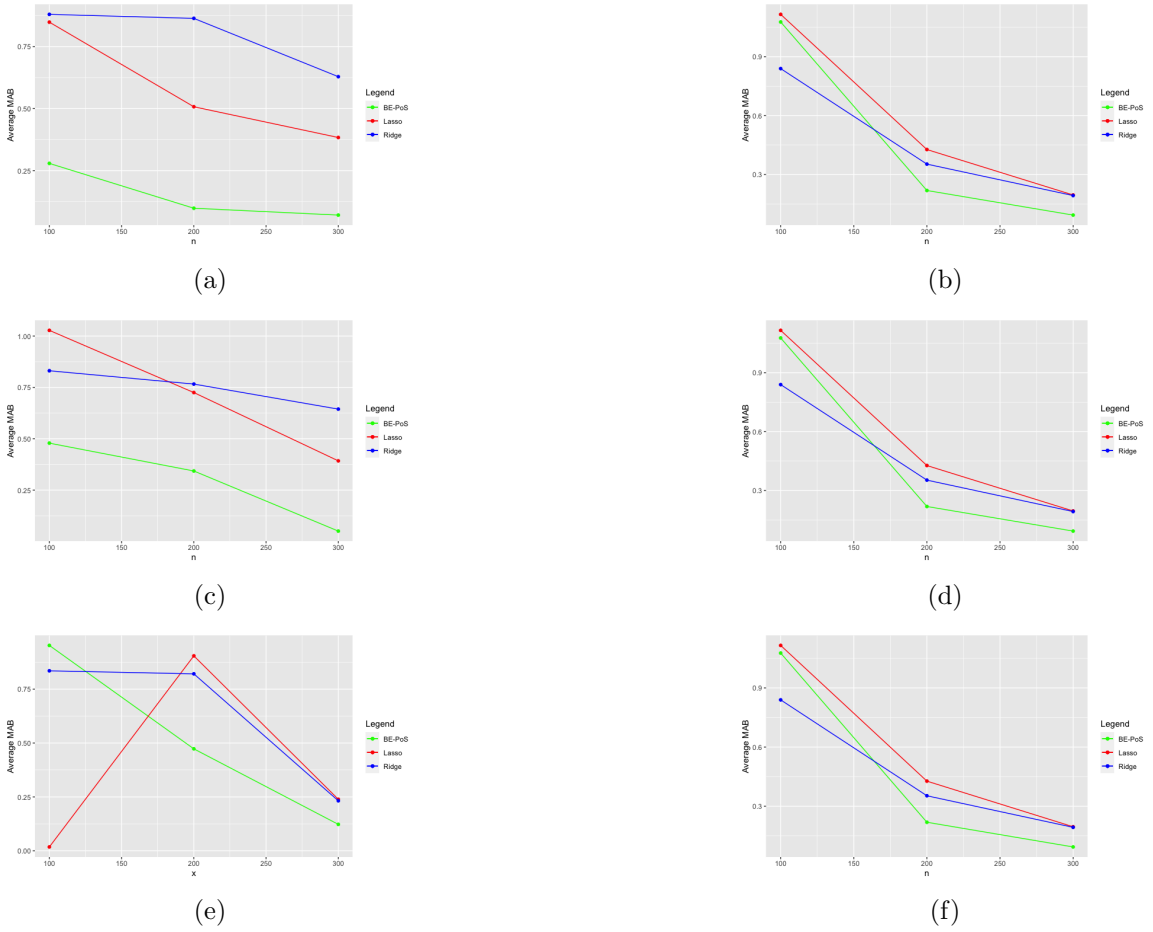


Figure 2: Average mean absolute bias (MAB) of the proposed BE-PoS method, lasso, bootstrapped lasso, ridge regression and PCR in the higher dimensions $p = 5000$ and $p = 10000$. (a) $s_0 = 20,\ p = 5000$, (b) $s_0 = 20,\ p = 10000$, (c) $s_0 = 50,\ p = 5000$, (d) $S_0 = 20,\ p = 10000$, (e) $s_0 = 100,\ p = 5000$, (f) $s_0 = 100,\ p = 10000$.

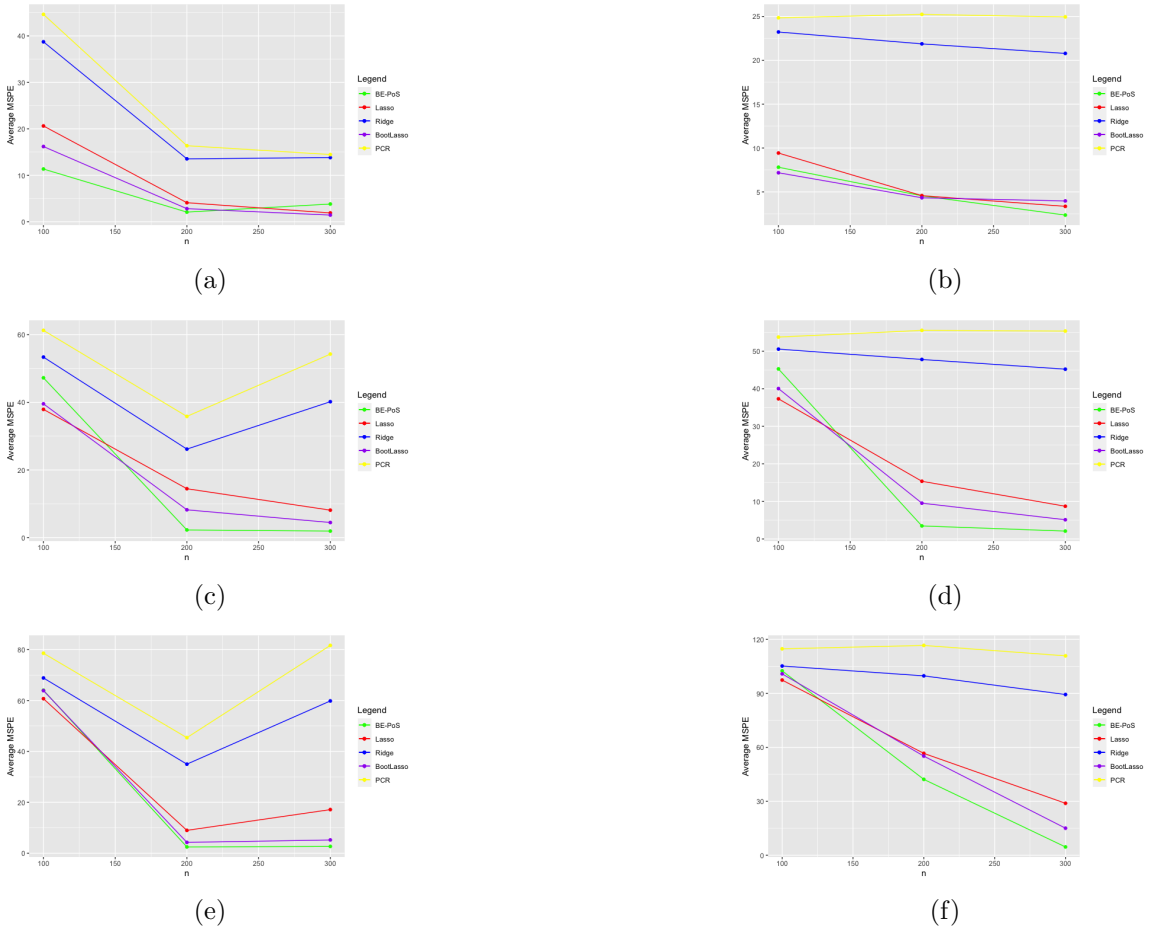


Figure 3: Average mean squared prediction error (MSPE) of the proposed BE-PoS method, lasso, bootstrapped lasso, ridge regression and PCR in the lower dimensions $p = 500$ and $p = 1000$. (a) $s_0 = 20,\ p = 500$, (b) $s_0 = 20,\ p = 1000$, (c) $s_0 = 50,\ p = 500$, (d) $S_0 = 20,\ p = 1000$, (e) $s_0 = 100,\ p = 500$, (f) $s_0 = 100,\ p = 1000$.

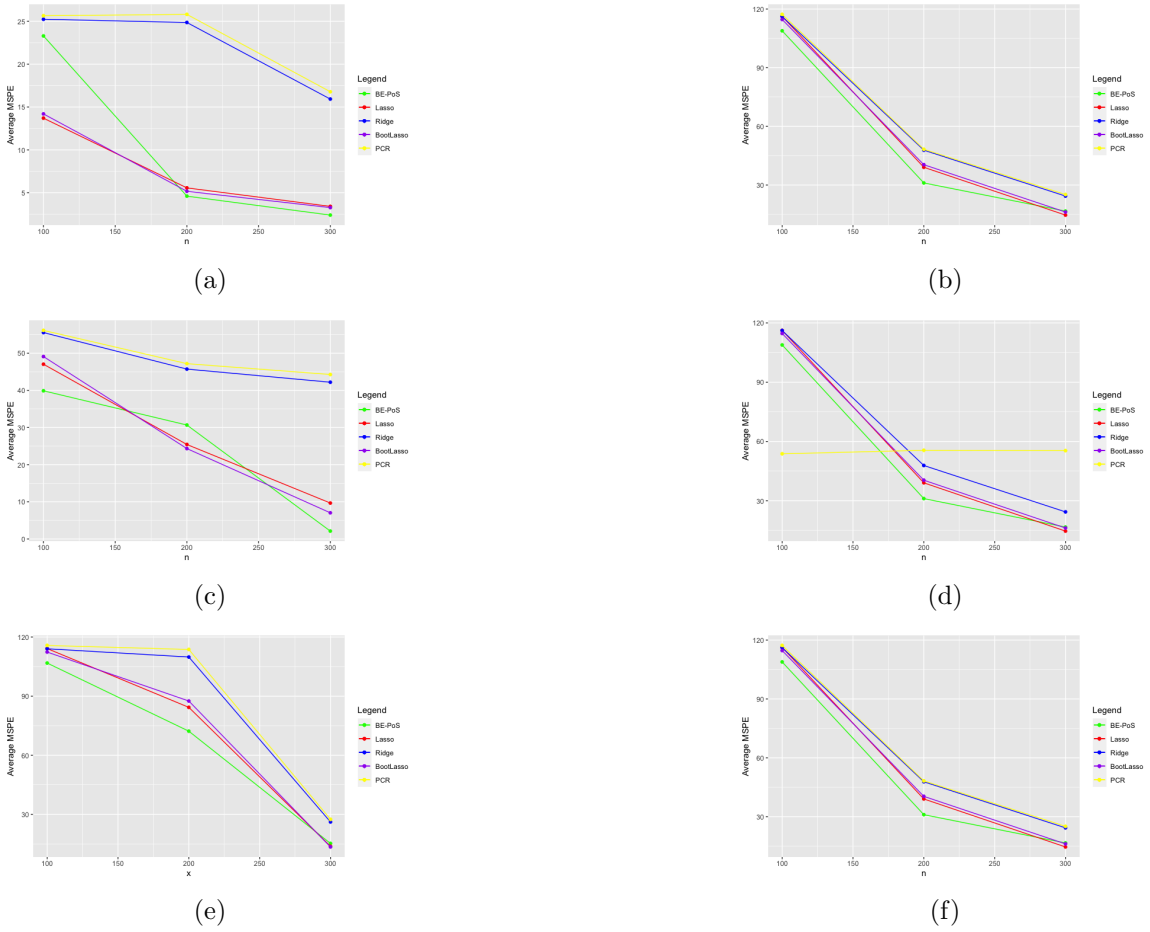
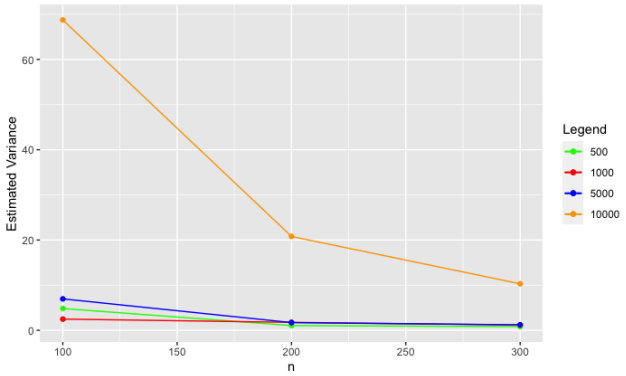
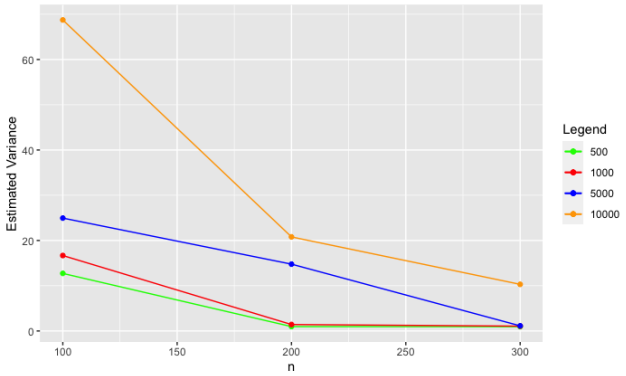


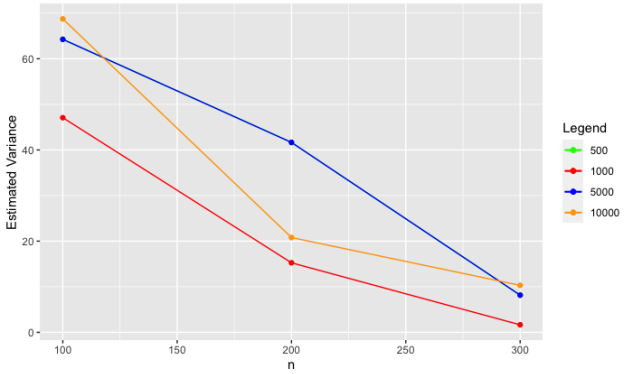
Figure 4: Average mean squared prediction error (MSPE) of the proposed BE-PoS method, lasso, bootstrapped lasso, ridge regression and PCR in the higher dimensions $p = 5000$ and $p = 10000$. (a) $s_0 = 20,\ p = 5000$, (b) $s_0 = 20,\ p = 10000$, (c) $s_0 = 50,\ p = 5000$, (d) $S_0 = 20,\ p = 10000$, (e) $s_0 = 100,\ p = 5000$, (f) $s_0 = 100,\ p = 10000$.



(a)



(b)



(c)

Figure 5: Average estimate of the variance σ^2 over all 100 possible replications using the BE-PoS method. (a) $s_0 = 20$, (b) $s_0 = 50$, (c) $s_0 = 100$. In the simulations, the true error variance is set to 1.

n	s_0	500	1000	5000	10000
100	20	3.03	3.39	5.81	5.08
100	50	5.59	5.58	7.09	7.62
100	100	5.13	8.21	10.67	11.34
200	20	1.83	2.57	3.25	3.95
200	50	1.47	2.08	4.13	3.67
200	100	1.71	3.84	6.64	6.63
300	20	1.72	2.43	2.54	2.32
300	50	1.37	1.73	1.82	1.09
300	100	0.97	1.60	1.32	1.30

Table 2: The average length of credible intervals for individual coefficients using the BE-PoS approach, at a nominal significance level of 0.05.

B Results from real data examples

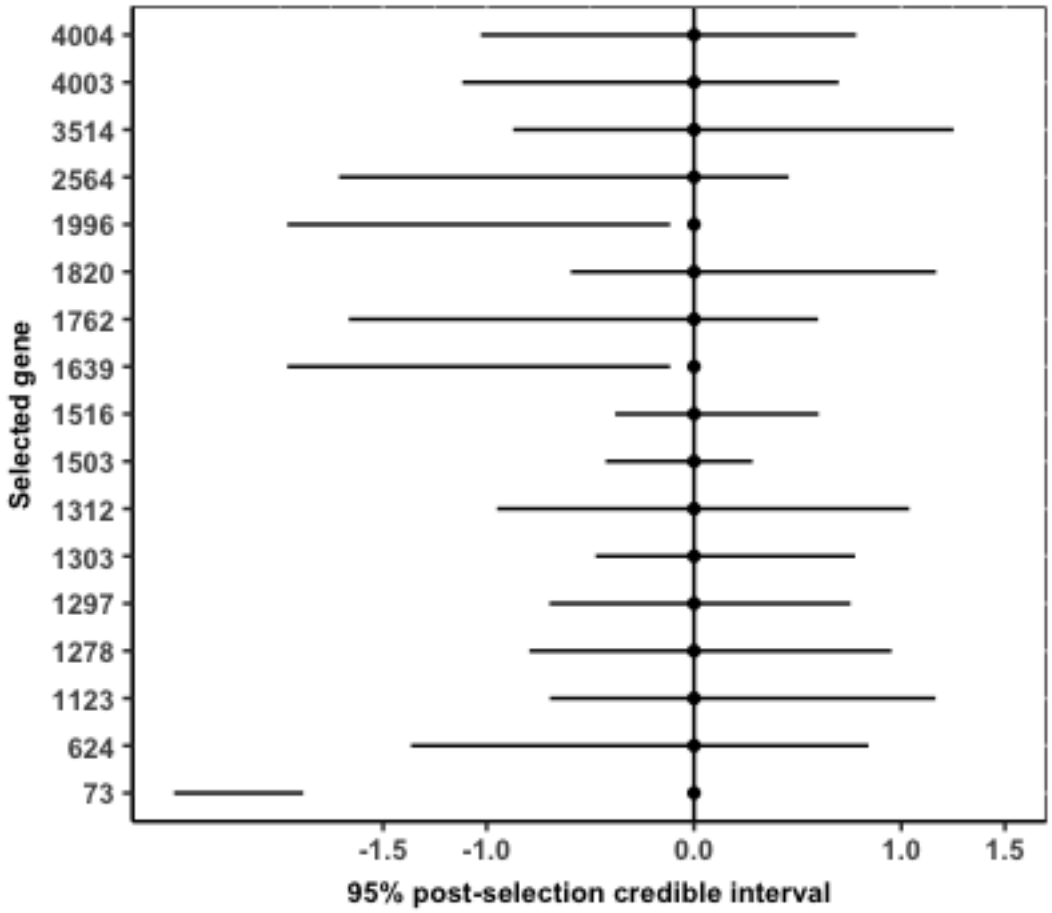


Figure 6: The prediction results for applying various methods to the riboflavin data