# Consistent high-dimensional Bayesian variable selection via penalized credible regions

By: Howard D. Bondell and Brian J. Reich

*Department of Statistics, North Carolina State University,*
*Box 8203, Raleigh, NC 27695, U.S.A.*

Correspondence Author: Howard D. Bondell
E-mail: bondell@stat.ncsu.edu
Telephone: (919) 515-1914
Fax: (919) 515-1169

July 11, 2012

# Consistent high-dimensional Bayesian variable selection via penalized credible regions


## Abstract

For high-dimensional data, particularly when the number of predictors greatly exceeds the sample size, selection of relevant predictors for regression is a challenging problem. Methods such as sure screening, forward selection, or penalized regressions are commonly used. Bayesian variable selection methods place prior distributions on the parameters along with a prior over model space, or equivalently, a mixture prior on the parameters having mass at zero. Since exhaustive enumeration is not feasible, posterior model probabilities are often obtained via long MCMC runs. The chosen model can depend heavily on various choices for priors and also posterior thresholds. Alternatively, we propose a conjugate prior only on the full model parameters and use sparse solutions within posterior credible regions to perform selection. These posterior credible regions often have closed-form representations, and it is shown that these sparse solutions can be computed via existing algorithms. The approach is shown to outperform common methods in the high-dimensional setting, particularly under correlation. By searching for a sparse solution within a joint credible region, consistent model selection is established. Furthermore, it is shown that, under certain conditions, the use of marginal credible intervals can give consistent selection up to the case where the dimension grows exponentially in the sample size. The proposed approach successfully accomplishes variable selection in the high-dimensional setting, while avoiding pitfalls that plague typical Bayesian variable selection methods.



*Key Words*: Bayesian variable selection; Consistency; Credible region; LASSO; Stochastic search.

# 1 Introduction

Consider the usual linear regression model, $Y = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon$ with a data set of $n$ observations and $p$ predictor variables. Let the response, $Y$, be a scalar random variable, and the predictor, $\mathbf{x}$, be a vector of length $p$. The errors are assumed independent with constant variance $\sigma^2$. Variable selection in this context is accomplished via setting $\beta_j = 0$ for some $j \in \{1, ..., p\}$. If $\beta_j = 0$ then variable $j$ is effectively removed from the model. In modern data analysis it is often the case that the number of candidate predictors can be extremely large. Of particular interest for modern data analysis is the high-dimensional situation with $p >> n$. Clearly, simply performing ordinary least squares is not feasible in this high-dimensional setting. Variable selection to reduce the large dimension of the candidate predictors is required.

Many variable selection methods have been developed, from discrete subset selection methods to more recent approaches. Penalization methods have gained popularity, particularly with the ever-increasing dimensionality of current data. These methods include nonnegative garrote (Breiman, 1995), least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996), smoothly clipped absolute deviation (SCAD, Fan and Li, 2001), elastic-net (Zou and Hastie, 2005), fused LASSO (Tibshirani et al., 2005), adaptive LASSO (Zou, 2006), group LASSO (Yuan and Lin, 2006), octagonal shrinkage and clustering algorithm for regression (OSCAR, Bondell and Reich, 2008).

In parallel, Bayesian approaches to variable selection have also been gaining popularity (George and McCulloch, 1993; George and Foster, 2000; Brown et al., 2002; Tadesse et al., 2005; Kinney and Dunson, 2007; Dunson et al., 2008; Liang et al., 2008). For some reviews of Bayesian approaches see George and McCulloch (1997) and O'Hara and Sillanpää (2009).

A Bayesian approach places a prior distribution on $\boldsymbol{\beta}$, denoted by $\Pi(\boldsymbol{\beta})$. It is typically the case that only an unknown subset of the coefficients $\beta_j$ are truly relevant,

so in the context of variable selection we begin by indexing each candidate model with one binary vector $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_p)^T$ where each element $\delta_j$ takes the value 1 or 0 depending on whether variable $j$ is included or excluded from the model, that is,

$$\delta_j = \begin{cases} 1 & \text{if } x_j \text{ is included in the model,} \\ 0 & \text{if } x_j \text{ is excluded from the model.} \end{cases} \tag{1}$$

Writing $\boldsymbol{\beta}^{(\delta)}$ to denote the vector of coefficients only for the predictors where $\delta_j = 1$, prior distributions are commonly specified via $p(\boldsymbol{\delta})$ and as $\Pi(\boldsymbol{\beta}^{(\delta)} \mid \boldsymbol{\delta})$. The prior $\Pi(\boldsymbol{\beta}^{(\delta)} \mid \boldsymbol{\delta})$ is often chosen for convenience, such as conjugacy, or to represent ignorance, i.e. a noninformative prior. Note that this is equivalent to a 'spike and slab' type prior where each coefficient in the full model is a mixture of a continuous distribution and mass at zero, where the mass at zero is exactly given by $p(\boldsymbol{\delta})$.

For linear regression, a simple choice is that $\Pi(\boldsymbol{\beta}^{(\delta)} \mid \boldsymbol{\delta}, \sigma^2, \tau)$ be given by $N(0, \sigma^2/\tau)$, where $\sigma^2$ is the error variance and $\tau$ is a precision parameter. The error variance, $\sigma^2$ is typically given a diffuse inverse Gamma prior, while the hyperparameter $\tau$ is either chosen to be a fixed small value, or given a prior distribution such as a gamma. Another popular choice is Zellner's g-prior (Zellner, 1986; Liang et al., 2008), which is computationally convenient by using the crossproduct matrix $X^T X$ in the prior covariance. However, when $p > n$, this matrix is singular, and hence the prior would be a singular normal. The independence prior $N(0, \sigma^2/\tau)$ avoids this difficulty.

A common prior for the inclusion indicators is, $p(\boldsymbol{\delta}) = \pi^{p_\delta}(1-\pi)^{p-p_\delta}$ (George and McCulloch, 1993, 1997; George and Foster, 2000) where $p_{\boldsymbol{\delta}} = \sum_{j=1}^{p} \delta_j$ is the number of predictors in the model defined by $\boldsymbol{\delta}$, and $\pi$ is the prior inclusion probability for each covariate. We can see this being equivalent to placing independent Bernoulli($\pi$) priors on $\delta_j$ and thereby giving equal weight to any pair of equally-sized models. Setting $\pi = 1/2$ yields the popular uniform prior over model space and, under this prior

specification, the posterior model probability for a given model is proportional to the marginal likelihood for that model. Brown et al. (2002) suggest to set $\pi$ at an apriori guess at the proportion of important predictors out of the total. Instead of fixed hyperparameter, $\pi$, an alternative is to use a beta prior to allow the degree of sparsity to be determined by the data. Scott and Berger (2010) examines these issues in the high-dimensional setting and suggest a default uniform prior on $\pi$. Other choices of priors over model space are possible and may lead to different results in terms of the model choices.

Bayesian variable selection proceeds by computing the posterior probability of each model using an approach such as the highest posterior model, or examining each predictor's marginal posterior inclusion probability (Barbieri and Berger, 2004). In practice, it is not feasible to compute the posterior model probabilities for all $2^p$ models, and hence Stochastic Search Variable Selection (SSVS) methods have been proposed (George and McCulloch, 1993) to sample through model space and avoid enumerating all models. These MCMC methods can be computationally prohibitive in high dimensions. Also, the resulting estimates of the high-dimensional discrete posterior of the vector $\boldsymbol{\delta}$ are unstable.

Traditional Bayesian variable selection methods that rely on computing model probabilities can be highly sensitive to prior choices. These standard methods require: 1) a proper prior distribution, 2) a choice of prior inclusion probabilities for each predictor, 3) a choice of posterior threshold to decide which predictors to include, and 4) MCMC sampling to estimate the posterior quantities. The selection procedure can be highly sensitive to each choice in the prior specification. For example, it is well known that letting $\tau \to 0$ favors the null model. Often, particularly in high-dimensional settings, no reliable prior information is available. Another problem is the choice of prior over model space. For large $p$, a seemingly noninformative uniform prior over model space

places almost all of its prior mass on models around size $p/2$, which is likely far away from a reasonable model size.

An intuitive approach to Bayesian variable selection is proposed in this paper based on the notion of the posterior credible region. The approach separates variable selection and model fitting. First, the full model is fit using all predictors with a continuous prior. A decision-theoretic type approach is then adopted to select a sparse estimate of $\boldsymbol{\beta}$ based on its posterior distribution. This sparse vector then determines the selected model. This approach does not require long runs of MCMC chains to search over model space. By using joint credible regions and incorporating correlation in the posterior when constructing the credible regions, high dimensional variable selection can be accomplished efficiently with improved performance over SSVS approaches. For the special case of $p < n$ and a flat prior on $\boldsymbol{\beta}$, the proposed Bayesian estimator is shown to be equivalent to the popular adaptive LASSO (Zou, 2006). However, in general, it is shown that the proposed approach outperforms its frequentist counterpart, particularly when $p \approx n$. Furthermore, the proposed approach is valid even in the case where $p >> n$.

Joint credible regions refer to posterior contours using the full joint posterior distribution, such as high density regions, while marginal credible regions consider only the marginal posterior distributions (based on the full joint posterior, however), resulting in a union of univariate credible intervals. It is shown that by using the joint credible regions, consistent model selection can be accomplished. By consistent model selection, we mean that the correct model is selected with probability tending to one, as $n \to \infty$. In addition, using marginal credible regions can accomplish consistent model selection in the ultra-high dimensional setting up to the case where the dimension can grow exponentially fast relative to the sample size. This ultra-high dimensional consistency in selection matches the best available rate for model selection, and hence shows that

this simple proposed approach becomes a useful tool for ultra-high dimensional data.

The remainder of the paper is organized as follows. Section 2 introduces the credible region approach to selection and discusses computational aspects. Section 3 presents the asymptotic selection consistency results for both the joint regions with fixed $p$ and the marginal regions, with quickly diverging $p$. Simulation results are presented in Section 4, while Section 5 gives the analysis of a real-time PCR dataset. All proofs are given in the appendix.

# 2 Penalized credible regions

## 2.1 The approach

The idea behind the proposed approach is to build a sequence of credible regions for the model parameters based on a chosen prior, which may be improper (if $p < n$), as long as the resulting posterior distribution is proper. Given a credible region at a particular level, one can imagine that any point in the region can be considered a feasible value for the parameter vector. We propose to choose the model represented by the point within the region having the sparsest representation. For example, one may say that the highest posterior density region containing 50% of the probability is the region that would be considered for our feasible models. As the desired probability content increases, the model will become more sparse, as the region will expand to cover more and more of the space. This sequence of credible sets can then be viewed as a sequence of selected models.

In many cases, with a simple choice of prior, such as a default non-informative prior, the posterior has a known analytic form. In such cases, no MCMC sampling is needed to compute the sequence of models. The results of the SSVS methods have been shown to be highly sensitive to the choices of prior distributions on the parameters and

the inclusion probabilities, along with the posterior threshold. Additionally, variable selection is based on thresholding the marginal posterior inclusion probabilities, which does not fully take into account the joint posterior distribution. For instance, in the case of three highly correlated predictors, in each MCMC draw only one of the three may be nonzero, hence the marginal posterior inclusion probability for each may only be around 33%, yet at least one of them is included the model at nearly 100% of the draws. Using the common posterior threshold of 50% would yield a final model that did not include any of the three.

For the proposed approach, a prior distribution is specified for the full model only, $\Pi(\boldsymbol{\beta})$ without a prior distribution over model space, and without the need for a 'spike and slab' type prior. In particular, we suggest the use of a simple choice of congugate prior.

Let $||\boldsymbol{\beta}||_0$ denote the $L_0$ norm of the vector $\boldsymbol{\beta}$, i.e. the number of nonzero elements, and let $\mathcal{C}_\alpha$ denote a credible region containing $(1-\alpha)\times 100\%$ of the posterior probability. The proposed estimate is

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{\beta}||_0 \text{ subject to } \boldsymbol{\beta} \in \mathcal{C}_\alpha, \tag{2}$$

with the chosen model then given by the set of indices with non-zero coefficients, $A_n = \{j \mid \tilde{\beta}_j \neq 0\}$.

Typically $\mathcal{C}_\alpha$ is chosen as the highest posterior density region, but it can be any posterior region having coverage $(1-\alpha)$. For the linear model, with fixed $\tau$ and $\{\boldsymbol{\beta} \mid \sigma^2, \tau\} \sim N(0, \sigma^2/\tau\ I_p)$, the posterior of $\boldsymbol{\beta}$ is elliptical, with density of the form $\Pi(\boldsymbol{\beta}|\text{ Data }) = H\left[(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\Sigma^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right]$, for some monotone decreasing function $H$, where $\hat{\boldsymbol{\beta}}$ and $\Sigma$ are the posterior mean and covariance matrix, respectively. Note that the form of $H$ depends on the prior for $\sigma^2$. For example, if $\sigma^2$ is given an inverse gamma prior, then $\Pi(\boldsymbol{\beta}|\text{ Data })$ is a multivariate t-distribution. Hence the highest density

region in this case, and others as well, is of the form $\mathcal{C}_\alpha = \{\boldsymbol{\beta} : H((\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\Sigma^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})) \geq K_\alpha\}$ for some $K_\alpha$, which is equivalent to $\{\boldsymbol{\beta} : (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\Sigma^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \leq C_\alpha\}$, for some $C_\alpha$. In general, placing a prior distribution on $\tau$ no longer maintains the elliptical contours in the posterior distribution. However, credible sets can still be constructed using elliptical contours, although they will no longer match the highest density region.

As defined, there are difficulties with (2). First, the solution requires a search over a possibly high dimensional region, and even in low dimensions, it represents a combinatorial search. Second, in most cases the solution is non-unique. To circumvent these difficulties, the $L_0$ criterion is replaced by a criterion proposed by Lv and Fan (2009), a smooth homotopy between $L_0$ and $L_1$. The proposed criterion is $\sum_{j=1}^p \rho_a(|\beta_j|)$, where

$$\rho_a(t) = \frac{(a+1)t}{a+t} = \left(\frac{t}{a+t}\right) I(t \neq 0) + \left(\frac{a}{a+t}\right) t, \qquad t \geq 0, a > 0 \tag{3}$$

and

$$\rho_0(t) = \lim_{a \to 0^+} \rho_a(t) = I(t \neq 0) \text{ and } \rho_\infty(t) = \lim_{a \to \infty} \rho_a(t) = t. \tag{4}$$

Interest here focuses on $\rho_a(t)$ for $a \approx 0$, in which case $\sum_{j=1}^p \rho_a(|\beta_j|) \approx ||\boldsymbol{\beta}||_0$.

The optimization problem becomes

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{j=1}^p \rho_a(|\beta_j|) \text{ subject to } \boldsymbol{\beta} \in \mathcal{C}_\alpha. \tag{5}$$

This is a nonconvex optimization problem. However, this will become a more tractable optimization problem.

A recent, and increasingly popular, approach to optimization with non-convex penalty functions is a local linear approximation (Zou and Li, 2008). Letting $\hat{\boldsymbol{\beta}}$ be

the posterior mean or mode, a local linear approximation of $\rho_a(|\beta_j|)$ is given by

$$\rho_a(|\beta_j|) \approx \rho_a(|\hat{\beta}_j|) + \rho'_a(|\hat{\beta}_j|)\left(|\beta_j| - |\hat{\beta}_j|\right), \tag{6}$$

with

$$\rho'_a(|\hat{\beta}_j|) = \frac{a(a+1)}{\left(a + |\hat{\beta}_j|\right)^2}. \tag{7}$$

In practice, we choose the posterior mean, as it is often easier to compute than the mode, and are, of course, equal if the posterior is elliptically symmetric.

Since $\mathcal{C}_\alpha = \{\boldsymbol{\beta} : (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \Sigma^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq C_\alpha\}$ is a convex set, the proposed linearized optimization problem is then equivalent to its Lagrangian optimization problem given by

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \Sigma^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \lambda \sum_{j=1}^{p} \left(a + |\hat{\beta}_j|\right)^{-2} |\beta_j|, \tag{8}$$

where $\lambda_\alpha$ is a Lagrange multiplier that corresponds to the optimization problem for a given choice of $\alpha$. Note that, for any given dataset there is a one-to-one correspondence between the chosen coverage, $1 - \alpha$, and the value of $\lambda_\alpha$. However, this relationship is highly nonlinear and data-dependent.

Letting $a \to 0$ in (8) yields

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \Sigma^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \lambda \sum_{j=1}^{p} |\hat{\beta}_j|^{-2} |\beta_j|. \tag{9}$$

The proposed sequence of solutions is then given by the solution to (9) as a function of $\lambda$. Note that this results in a single parameter indexing the path. This index (or tuning) parameter can be any one of $\lambda$, $\alpha$, or $C_\alpha$ as they are each in a one-to-one correspondence to one another.

As an example, suppose $p < n$ and that one were to choose an (improper) flat prior for $\boldsymbol{\beta}$, so that $\Pi(\boldsymbol{\beta}) \propto 1$, then posterior credible sets will align with frequentist

confidence intervals based on the likelihood function. For this choice of improper prior, $\hat{\boldsymbol{\beta}}$ is exactly the maximum likelihood estimator and $\Sigma$ is its covariance matrix. Then some algebra shows that $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \Sigma^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$ is exactly the likelihood function plus a constant that is independent of $\beta$. The problem then reduces to a penalized likelihood, with penalty function given by a particular choice of weighting, $|\hat{\beta}_j|^{-2}$ for an Adaptive LASSO estimator (Zou, 2006). Hence a noninformative choice of prior distribution for the proposed problem results in a special case of a previously proposed model selection approach. However, an improper prior is not feasible in the case where $p > n$ and even when $p < n$, some regularization by the prior should improve the procedure.

We note here a connection between the proposed approach and the Dantzig Selector (Candes and Tao, 2007). Similarly to the proposed approach, the Dantzig selector also seeks to find a sparse solution within a constrained region. The Dantzig selector is defined via

$$\tilde{\boldsymbol{\beta}}_{DS} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{\beta}||_1 \text{ subject to } \boldsymbol{\beta} \in \mathcal{D},$$

where $\mathcal{D} = \{\boldsymbol{\beta} : \sup_{1 \leq j \leq p} \mid \mathbf{X}_j^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \mid\}$. Although both approaches seek a sparse solution within a constrained region, that is where the similarities end. The proposed measure of sparsity is via an approximation to $L_0$ while the Dantzig selector uses $L_1$. However, the more important distinction is the form of the constraint region. The proposed constraint region arises from a Bayesian formulation where the constraint stems from a posterior credible set. In stark contrast, the Dantzig selector region is proposed based on bounding the maximum correlation between a predictor and the residual vector, or equivalently the $L_\infty$ norm of the score function. This resulting region does not coincide with any typical posterior credible set.

## 2.2 Computation

In general, computation of the proposed sequence of solutions to (9) can be directly accomplished with existing software. Letting $X^* = \Sigma^{-1/2} D$ and $Y^* = \Sigma^{-1/2}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ and $\Sigma$ are the posterior mean and covariance, respectively, and the matrix $D$ is a diagonal matrix with $j^{th}$ element given by $\hat{\beta}_j^2$. Solving a standard L1 penalized regression using $X^*$ and $Y^*$ for a given $\lambda$ yields $\boldsymbol{\beta}^*$. The solution to (9) is then given by $\tilde{\boldsymbol{\beta}} = D\boldsymbol{\beta}^*$. The entire sequence of solutions can thus be computed efficiently using the LARS algorithm (Efron et al., 2004).

Before computing the sequence of solutions, it is necessary to obtain $\hat{\boldsymbol{\beta}}$ and $\Sigma$, the posterior mean and covariance matrix. For fixed hyperparameter, $\tau$, both can be obtained in closed-form (Carlin and Louis, 2009). However, when placing a prior on $\tau$, they are no longer available in closed form. A simple short MCMC run is used to sample from the posterior and estimate the mean and covariance matrix from this relatively small number of MCMC samples. The conjugate form of the priors make the MCMC updates in closed form so Gibbs sampling can be used. Unless $p$ is extremely high-dimensional the entire vector, $\beta$ can be updated simultaneously, so that convergence of the MCMC chain is virtually immediate. In the ultra-high dimensional setting, blocks of the parameter vector can instead be updated iteratively to save on very large matrix computations. This still gives fast convergence. Once convergence of the chain occurs, a large number of draws is not needed, even though they are quick to generate, as it is only to estimate the mean and covariance matrix that is needed.

# 3 Asymptotic Theory

## 3.1 Selection consistency of joint credible sets

Let $\tilde{\boldsymbol{\beta}}^{\alpha}$ denote the sequence of solutions to the credible set optimization problem as a function of the level $\alpha$. Suppose that, for a given sample, we choose the solution on the sequence corresponding to level $\alpha_n$. Denote the estimated and true sets of non-zero coefficients as $A_n = \{j : \tilde{\beta}_j^{\alpha_n} \neq 0\}$ and $A = \{j : \beta_j^0 \neq 0\}$, respectively, where $\boldsymbol{\beta}^0$ is the true parameter vector. It will now be shown that the sequence of penalized credible sets is a consistent model selection procedure. In particular, there exists a sequence $\alpha_n$ such that $P(A_n = A) \to 1$.

Consider a sequence of credible sets such that the coverage, $1 - \alpha_n$, increases with $n$. Let the sequence of sets be defined by the sequence of posterior thresholds, such that $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \Sigma^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq C_n$. Note that there is a one-to-one relationship between $\alpha_n$ and $C_n$ for each fixed sample and $C_n \to \infty$ implies that the coverage goes to one (see the Appendix).

First consider the case for fixed dimension, $p$. For this case, the following regularity conditions are assumed.

(A1) The error terms are iid with mean zero and finite variance.

(A2) The matrix $\mathrm{X}^T\mathrm{X}/n \to D$, for some positive definite matrix, $D$.

(A3) The prior precision satisfies $\tau = o(n)$.

(A4) $\min\{|\beta_j^0| : j \in A\} > c_1/\sqrt{n}$, for some $c_1 > 0$.

THEOREM 1 *Under conditions (A1) - (A4), if $C_n \to \infty$ and $n^{-1}C_n \to 0$, then the credible set method is consistent in variable selection, i.e. $P(A_n = A) \to 1$.*

Although Theorem 1 guarantees that, with probability tending to one, the correct model is found by a choice of credible set with coverage $1 - \alpha_n$. In practice, varying $\alpha \in (0, 1)$ defines a set of models from the full model down to the null model. Note that it is not necessarily a nested set of models that continue to reduce in size. As noted in Efron et al. (2004), the LASSO solution path may have a variable appear in an earlier model, but then be removed at another stage. The path of models determined by the nested credible set approach has the same property.

This selection consistency property allows for us to expect that the sequence of nested credible sets will contain the true model with high probability, at least as the sample size increases. With this in mind, we shall consider the sequence of credible sets as our set of candidate models to choose from.

Note that the condition of Theorem 1 deals with fixed dimension $p$. Since the situation considered here is the high-dimensional case with $p >> n$, the fixed $p$ asymptotics may not be appropriate. Modifications can be made to address the situation in which $p_n = o(n)$. However, in fact, one would hope to consider $p_n/n \to \infty$ in that the sample size in many high dimensional problems will never practically dominate the number of predictors. This is more in line with today's high dimensional data, in that no matter how much data is actually collected, the sample size will never overwhelm the dimension.

For a given threshold $C_n$, define $E_{C_n}$ to denote the event that the credible region given by $\{\boldsymbol{\beta} : (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \Sigma^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq C_n\}$ contains at least one point such that $\beta_j = 0$ for all $j \in A^c$ and also contains no points such that $\beta_j = 0$ for some $j \in A$. More specifically, the credible set covers the zero vector for all $j \in A^c$ and zero is outside the region for all $j \in A$. Now, if there exists an $\epsilon > 0$ such that $P(E_{C_n}) < 1 - \epsilon$ for all $n$ regardless of the choice of $C_n$, then it is not possible to obtain selection consistency in this case.

THEOREM 2 *If $p_n/n \to c > 0$ then the posterior mean, $\hat{\boldsymbol{\beta}}$ is not mean square consistent for $\boldsymbol{\beta}^0$. In particular $E\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)^T \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right) \to K$, for $0 < K \leq \infty$.*

COROLLARY 1 *If $p_n/n \to c > 0$ then there does not necessarily exist a sequence $C_n$, such that $P(E_{C_n}) \to 1$.*

The corollary follows directly from the theorem due to the fact that in order to guarantee that the credible set will contain the true zeros for $j \in A^c$, the sequence $C_n$ cannot go to zero (due to the non-mean square consistency). However, unless the region shrinks around the true nonzeros for $j \in A$, there is no guarantee of the region not containing zero for $j \in A$.

## 3.2 Ultra-high dimensional selection consistency of marginal credible sets

Although the joint credible set approach is selection consistent, it is not consistent if $p_n$ diverges at order $n$ or even faster. Now, instead of constructing elliptical credible sets from the posterior, consider constructing rectangular credible sets based on the same posterior distribution. Specifically, the selection rule takes the form $A_n = \{j : |\hat{\beta}_j| > t_{n,j}\}$ for some set of thresholds $t_{n,j}$, where $\hat{\boldsymbol{\beta}}$ is the posterior mean. Note that these thresholds will again be dependent on the choice of $\alpha$. Typically, one would choose $t_{n,j} = s_j t_n$, where $s_j$ is proportional to the posterior standard deviation of $\beta_j$. A reasonable choice of $s_j$ is the posterior standard deviation relative to the minimum standard deviation over all $j$, so that $1 \leq s_j < \infty$ for all $j$. Note that this is still using the full joint posterior, but looking at the marginal credible sets.

It will now be shown that this simple rectangular credible set approach is selection consistent up to the case where $\log p_n = O(n^c)$ for some $0 < c < 1$. Let $q_n = |A|$ denote the number of predictors with nonzero true coefficients, and $r_n \leq \min\{n, p_n\}$ denote

the rank of $X^TX/n$. Further, let $\Gamma_n$ denote the $p_n \times p_n$ matrix whose columns are the eigenvectors of $X^TX/n$ ordered by decreasing eigenvalue and let $d_1^n \geq ... \geq d_{r_n}^n > 0 = d_{r_n+1}^n = ... = d_{p_n}^n$ denote the eigenvalues. Let $\Gamma_{n,1}$ denote the first $r_n$ columns of $\Gamma_n$, i.e. those corresponding to the non-zero eigenvalues. Since $\Gamma_n$ is an orthonormal basis of rank $p_n$, write $\boldsymbol{\beta}^0 = \Gamma_n\eta = \Gamma_{n,1}\eta_1 + \Gamma_{n,2}\eta_2$, where $\Gamma_{n,2}$ are the remaining $p_n - r_n$ columns of $\Gamma$, and $\eta_1, \eta_2$ is the appropriate partitioning of $\eta$.

In the high-dimensional setup, the previous assumptions (A1) - (A4) no longer apply. Instead, we replace the low dimensional assumptions by the following (B1) - (B5).

(B1) There exists $1 \leq n_0 < \infty$ such that, for all $n \geq n_0$, we have $d_1 > d_1^n \geq ... \geq d_{r_n}^n > d_2 > 0 = d_{r_n+1}^n = ... = d_{p_n}^n$, for some $0 < d_2 \leq d_1 < \infty$.

(B2) Let the threshold sequence $t_{n,j} = s_j t_n$ with $1 \leq s_j < \infty$ for all $j, n$, and $t_n \to 0$ with $t_n^{-1}\frac{\tau_n\sqrt{q_n}}{n} \to 0$ and $t_n^{-1}\sqrt{\frac{\log p_n}{\tau_n}} \to 0$.

(B3) $\max_j |\beta_j^0| < \infty$ and $t_n / \min_{j\in A} |\beta_j^0| \to 0$.

(B4) $||\Gamma_{n,2}\eta_2||_\infty = O(n^{-1}\tau_n\sqrt{q_n})$.

(B5) The error terms are iid $N(0, \sigma^2)$.

THEOREM 3 *Let $A_n = \{j : \hat{\beta}_j > t_{n,j}\}$. Then under conditions (B1) - (B5) listed above, $P(A_n \neq A) \leq 4\sqrt{\sigma^2}\frac{p_n}{t_n\sqrt{\tau_n}}\exp\left(-\frac{t_n^2\tau_n}{8\sigma^2}\right) \to 0$. Hence it follows that the marginal posterior thresholding approach is consistent in selection, i.e. $P(A_n = A) \to 1$.*

Condition (B1) is a typical condition on the nonzero eigenvalues of $\mathrm{X}^T\mathrm{X}/n$ that bounds the condition number of the non-singular part of the matrix. Condition (B2) gives the rate at which the threshold may decrease to zero, while still allowing for the exclusion of all unimportant predictors. Note that the scalars, $s_j$, can essentially be

taken to be the ratio of standard deviations of each estimate relative to the minimum standard deviation. Condition (B3) gives the smallest rate on the magnitude of the true nonzero coefficients so that they remain identifiable from zero, and also assumes that all true coefficients are finite.

Since condition (B2) implies that $n^{-1}\tau_n\sqrt{q_n} \to 0$, it follows that conditions (B2) and (B4) together state that for large enough $n$, the true parameter is approximately in the linear space spanned by $X^TX$. This is a basic identifiability condition that is needed so that the true parameter can be found. Note that the true parameter actually lies in a subspace of dimension $q_n$. In the high-dimensional setting, it is typically assumed that $q_n = o(n^\gamma)$ for some $0 \leq \gamma < 1$. However, here it follows that $q_n/n$ may even diverge, as long as the resulting vector is approximately estimable within the linear space. So it allows for the case that $\beta$ may actually be sparse in some linear transformed space, but not very sparse in the original space.

Note that assumption (B4) always holds if $p_n = o(n)$ and $X^TX/n$ converges to a positive definite matrix. For comparison, this is actually not true of some conditions required by other methods, for example the Irrepresentability Condition for the LASSO (Meinshausen and Buhlmann, 2006; Zhao and Yu, 2006; Zou, 2006), or Partial Orthogonality (Huang, Horowitz, and Ma, 2008; Huang, Ma, and Zhang, 2008). However, when $p_n > n$, the span of the observed $X^TX$ must be further analyzed. As a class of examples of a structure that would satisfy condition (B4), suppose that $q_n < n$ and, WLOG, assume that the first $q_n$ predictors were those with the non-zero coefficients, and the remainder were null. Let $j_n > q_n > 0$ and $k_n > 0$, with $j_n + k_n = n$. For example, in the simulation design shown later, we use $j_n = k_n = n/2$. Let the $p_n \times 1$ covariate vector for subject $i$ be $X_i = Z_i + W_i$ for each $i = 1, ..., n$, where the vector $W_i$ has $\text{Cov}(W)$ such that its maximum eigenvalue is $O(1/n)$. The vector $Z_i$ comes from a covariance structure with 2 blocks. The first block of size $j_n$ is any arbitrary full-rank

covariance with smallest eigenvalue bounded away from zero. The remaining block of size $p_n - j_n$ is a rank $k_n$ covariance matrix. Note that this resulting $\text{Cov}(X)$ is of full rank, $p_n$, but has $n$ dominant eigenvalues. In addition, the eigenspace spanned by these dominant eigenvalues contains the subspace of order $q_n$ with arbitrary entries for the first $q_n$ components and zeros elsewhere (hence $\beta$ is in this subspace). This type of covariance structure allows for the important predictors to be among a growing group of predictors of size $j_n = O(n)$ with arbitrary correlation among the variables within this group. The remaining $p_n - j_n$ predictors come from a factor-type model so that they are a transformation of an underlying random vector of dimension $k_n = O(n)$. The cross-covariance between these 2 groups is small and controlled by $\text{Cov}(W)$. In practice, most common covariance structures can be well approximated by one of this form by allowing the loadings in the factor model to functionally vary. Jung and Marron (2009) discussed the consistency of the estimated eigenspace in this framework with diverging $p_n$, but with $n$ fixed. Yata and Aoshima (2009) extended this work to allow $n$ to also diverge and non-Gaussian variables. Based on these results, under this setup, if the covariates are Gaussian and $p_n = o(n^2)$ then the space spanned by the first $n$ sample eigenvectors is consistent for the true space (Yata and Aoshima, 2009). Hence the span of the sample eigenvectors will approach that of the true eigenvectors and assumption (B4) will be satisfied. Note that this is only one possible formulation for which the assumption holds, but this is not the only way to satisfy the assumption.

Condition (B2) also implies a rate at which the prior precision, $\tau_n$, must diverge with $n$. Note that if the prior precision diverges too quickly, then the bias of the posterior mean will not vanish quickly enough. However, since the case considered here is $p > n$, if the prior precision does not diverge fast enough, the smallest eigenvalue of $X^T X + \tau_n I$ will go to zero too quickly.

The assumption of normality in condition (B5) is needed to get the exact rate on

both the divergent dimension and the rate for the selection consistency in the theorem. It is possible to remove that assumption, which would result in less tight bounds for the convergence rates.

In the following corollary, we establish the rate for $\tau_n$ (and hence for $t_n$) that satisfies (B1) - (B5) and allows $p_n$ to diverge at its fastest possible rate while still yielding selection consistency.

COROLLARY 2 *Let* $\tau_n \to \infty$ *and* $\tau_n = O\left(\left(\frac{n^2 \log p_n}{q_n}\right)^{1/3}\right)$ *and assume conditions (B1) - (B5). Then it follows that the posterior thresholding approach is consistent in selection when the dimension* $p_n$ *satisfies* $\log p_n = O\left(\left(n/\sqrt{q_n}\right)^c\right)$ *for some* $0 < c < 1$.

Note that in the case that $q_n$ does not grow with $n$, i.e. the true number of important predictors remains fixed, this allows for exponential growth of the dimension. In particular, one can allow $\log p_n = O\left(n^c\right)$ for $0 < c < 1$.

If $\tau_n = O\left(\left(\frac{n^2 \log p_n}{q_n}\right)^{1/3}\right)$ then conditions (B2) and (B4) reduce to

(B2*) $t_n \to 0$ with $t_n^{-1}\left(\frac{\sqrt{q_n}\log p_n}{n}\right)^{1/3} \to 0$.

(B4*) $||\Gamma_{n,2}\eta_2||_\infty = O((\left(\frac{\sqrt{q_n}\log p_n}{n}\right)^{1/3})$.

Although it has been shown in Theorem 2 that the posterior mean is not consistent in $L_2$ norm unless $p_n = o(n)$, the following theorem shows that the posterior mean is consistent in $L_\infty$ norm.

THEOREM 4 *Let* $\hat{\boldsymbol{\beta}}$ *be the posterior mean and* $\boldsymbol{\beta}^0$ *denote the true parameter. Assume that* $\frac{\tau_n\sqrt{q_n}}{n} \to 0$ *and* $\frac{\log p_n}{\tau_n} \to 0$ *along with (B1), (B4), and (B5). Then it follows that* $\max_j |\hat{\beta}_j - \beta_j^0| \to 0$, *as* $n \to \infty$.

Note that the condition in Theorem 4 is weaker than (B2) in that it does not depend on the rate of the threshold, $t_n$. Hence the terms $\frac{\log p_n}{\tau_n}$ and $\frac{\tau_n\sqrt{q_n}}{n}$ are allowed to go to zero more slowly, while achieving $L_\infty$ consistency.

# 4 Simulations

## 4.1 Comparison of variable ranking

A simulation study is conducted to examine the performance of the sequence of credible set approach and compare with other common variable selection methods. In each case, 200 datasets are simulated from the linear model with $Y = \mathrm{X}^T\boldsymbol{\beta}+\varepsilon$, where $\varepsilon \sim N(0, I_n)$. The sample size was taken to be $n = 60$ to match the real data example in Section 5, and the number of candidate predictors is varied in $p \in \{50, 500, 2000\}$. The predictors, $\mathrm{X}_{ij}$, are standard normal with the correlation between $x_{ij_1}$ and $x_{ij_2}$ given by $\rho^{|j_1-j_2|}$, for $\rho = 0.5, 0.9$, to represent a relatively low correlation setting and a relatively high setting.

The true coefficient vector $\boldsymbol{\beta}$ is given by $\left(\mathbf{0}_{10}^T, \mathbf{B1}^T, \mathbf{0}_{20}^T, \mathbf{B2}^T, \mathbf{0}_{p-40}^T\right)^T$ for $p = \{50, 500, 2000\}$ with $\mathbf{0}_k$ denoting the zero vector of length $k$. The randomly generated coefficient vectors **B1** and **B2** are each of length 5 and generated uniformly on the interval $(0, 5)$, componentwise.

As the credible set level, $\alpha$, is varied, an ordered set of predictors is created. Analogously, as the posterior inclusion threshold in SSVS, and the tuning parameter in the penalized methods are varied, ordered sets are created. To assess the performances of the approaches, consider the induced ordering of predictors. Ideally, the relevant predictors will appear towards the beginning of the ordering. To measure the reliability of the ordering, consider the resulting model at each step of the ordering. For each point on the ordering we denote the true positives (TP) as those variables that are included in the model correctly. False positives (FP) are those variables that are incorrectly included in the model. True negatives (TN) correspond to variables correctly omitted from the model, and false negatives (FN) refer to variables incorrectly omitted from the model. The Receiver-Operating Characteristic (ROC) curve plots the false

positive rate (FPR or 1-Specificity) on the x-axis and the true positive rate (TPR or Sensitivity) on the y-axis. Note that the denominators in the FPR and the TPR are the total number of irrelevant variables, and the total number of important variables, respectively. This is related to the tradeoff between type I error and power.

In addition, since in high-dimensional data, one may be concerned also with the False Discovery Rate (FDR), and not simply Type I error, the Precision-Recall (PRC) curve gives a complimentary picture which instead relates to the FDR (Davis and Goadrich, 2006). The Precision-Recall curve instead plots Recall on the x-axis and Precision on the y-axis. Recall is exactly the same as TPR, or Sensitivity. Precision instead looks at the ratio of true positives to the total number declared as positive, and hence the Precision-Recall curve examines the tradeoff between power and False Discovery Rate.

The joint credible set approach is compared with the approach of varying the threshold in marginal credible sets, i.e. using rectangular credible regions. Also included for comparison is 4 versions of SSVS. In each case (including both the joint and marginal credible sets), the $N(0, \sigma^2/\tau)$ is used as a prior for the nonzero coefficients, and for SSVS, the Bernoulli prior with probability $\pi$ is used for the inclusion indicators. For the joint and marginal credible sets, we used a Gamma(.01, .01) prior for $\tau$, while sensitivity to this choice is discussed later.

Each of the hyperparameters in the SSVS approach can be either treated as fixed, or further treated as random. For $\tau$, we fixed its value at $\tau_0$ and also gave it a Gamma(.01, .01) prior. For $\pi$, we used a Uniform(0,1) prior (Scott and Berger, 2010) and we also treated it as fixed. Various choices of the fixed value were used, but in the end, the known proportion of true nonzero coefficients was used as the fixed value. Note that in practice that is not possible, but for the simulation, this gives the SSVS approach its best chance to compete. In addition the fixed value of $\tau_0$ was varied. To again give it its

best value, we finally set it to $\tau_0 = 25/3$. This value was chosen due to the fact that the prior for the nonzero coefficients is centered at zero with second moment given by $\tau_0$. Since in the simulation, the true non-zero coefficients are generated from Uniform(0, 5), it follows that the true $E\beta^2 = 25/3$ for the non-zero coefficients. The 4 versions of SSVS represent the 4 combinations of the fixed and random hyperparameters. For each dataset using SSVS, we ran the MCMC chain for 100,000 iterations, dropping the first 5,000 for burn-in. In addition, the Adaptive LASSO is also used (for the one case with $p < n$), which as mentioned earlier, corresponds to the joint credible set approach with an improper prior on the coefficients. Also for the case of $p < n$ we include the Ordinary Least Squares (OLS) approach where the ordering is defined by the magnitude of the individual standardized coefficients. The frequentist approaches, LASSO and Dantzig selector, are also shown for comparison, along with the sequence of predictors generated from the Forward Selection path (which has the sure screening property as discussed in Wang, 2009). For the Dantzig Selector, the DASSO algorithm (James, Radchenko, and Lv, 2009) was used for computation of the full solution path.

Overall, it is clear that the ordering of the predictors from the proposed approach yields an ordering that tends to give competitive performance in terms of both the ROC curve and the PRC curve. This performance benefit increases significantly as the correlation is increased.

Table 1 gives the mean and standard error for the area under the ROC curve along with the area under the PRC curve for the $p = 50$ predictor case. In addition, the time (in seconds) for computation of each approach for a single dataset is given. All methods are somewhat similar in this case except for the poor performance of OLS and Adaptive LASSO. This is to be expected as $p \approx n$ gives unstable OLS estimates, and hence poor initial weights. In addition, the performance of the LASSO degrades under higher correlation, while the performance of the Forward Selection path also performs

poorly under higher correlation, while it also does not perform as well as the others even under the lower correlation. Figure 1 plots the mean ROC and PRC curves for the proposed approach along with two versions of SSVS. The upper panel shows that for $\rho = 0.5$, the proposed method is competitive with the SSVS approach. However, the lower panel demonstrates the increased benefit for $\rho = 0.9$.

Table 2 and Figure 2 gives the results for the $p = 500$ predictor case. For this case, the Adaptive LASSO and OLS are not included, since $p > n$. Since the LASSO and Dantzig selector can only choose at most $\min\{n, p\}$ predictors (Zou and Hastie, 2005), the ROC and PRC curves cannot be fully constructed, but can only go out to the case where 60 predictors are included out of the 500. This is also the case for Forward Selection. Due to this, the area under the curves cannot be compared directly, and hence are also omitted. So for this setting the comparisons focus only on SSVS. In this moderately high-dimensional case, the proposed method gives a substantial improvement over SSVS. For example, with $\rho = 0.9$ (Figure 2, bottom right) and recall $= 0.8$, the credible region approach has precision of 0.8 compared to 0.2 for SSVS, at a fraction of the computation time. Note that the computational time for SSVS on a single dataset is now becoming very prohibitive.

For the $p = 2000$ case, each run of SSVS for a given dataset takes over 5 hours so running a full simulation is not feasible. Meanwhile, the credible set approach takes under 10 minutes for each dataset. To show competing methods, we use the LASSO and Dantzig selector, which can still apply to the case where $p$ is large. As mentioned previously, since the LASSO and Dantzig selector can only choose at most $\min\{n, p\}$ predictors, the ROC and PRC curves can only go out to the case where 60 predictors are included out of the 2000. Hence, Figure 3 plots the mean of these partial ROC and PRC curves. Again in this situation, the performance of the proposed credible set approach competes well with the frequentist approaches, especially in the $\rho = 0.9$

case. This is to be expected, as both the LASSO and the Dantzig selector suffer from correlation among the predictors (see Candes and Tao, 2007 and the discussion therein).

To examine how the ordering of predictors translates into prediction, we considered the $p = 500$ setting and an independent test set is generated along with each of the 200 datasets. So for each of the 200 runs, two datasets were generated identically. The first dataset was used to construct the ordering of the predictors. The second dataset was then used to evaluate the prediction error at each point along the path. Figure 4 plots the prediction error for the methods above. Note that the minimum prediction error for the proposed credible set approach in all cases shows up right around 10 predictors being included. This is due to the fact that when 10 predictors are included, most of them are the truly important ones. Meanwhile, the minimum prediction error for SSVS occurs much earlier due to the fact that more irrelevant predictors are now being included. In addition, the overall minimum prediction error is smaller for the credible set approach than for SSVS.

Finally, we note that in all the cases we considered, the joint region approach outperformed the marginal interval approach. This is likely because the joint approach uses associations between variables more efficiently. However, the marginal approach does seem to decrease the gap as the sample size is increased. The theoretical results in Section 3.2 suggest that the marginal approach is consistent in the ultra-high dimensional case.

## 4.2 Selection Properties

Additional simulations were conducted to examine the variable selection properties of the proposed method along with some methods that are also known to be consistent in selection. The first situation examines the case with $p = 50$ predictors and various sam-

ple sizes. For each of the 200 datasets, the predictors are generated from a multivariate normal whose covariance is first randomly generated from a Wishart distribution centered on the identity with degrees of freedom $p = 50$. Since the covariance matrix was generated randomly for each dataset, without loss of generality, the first $q = 3$ coefficients were considered non-zero and generated from $U(0, 1)$ while the remaining $p - q = 47$ were null. The responses were then generated from a normal distribution with error variance of 1. This results in a theoretical $R^2 = 0.5$. The proposed credible set method using both joint credible sets and the marginal credible sets were compared along with the LASSO and SCAD. In all methods, the tuning parameters were chosen via BIC, as this has been shown in many cases to give the best performance in terms of selection. The results for $n = \{60, 100, 200, 500, 1000, 2000\}$ are shown in Table 3 reporting correct selection proportion, which is the proportion of times that the correct model is selected exactly and coverage proportion shows the proportion of times that the selected model covers the true model. Also reported is the average model size, and average number of important predictors (out of the 3) included in the model.

From the results in Table 3 it shows that both the joint and marginal credible set methods perform very well compared to the existing approaches. While both of the proposed methods have better selection properties in terms of perfect selection and coverage, it is also seen that they do so while still keeping the model size small. Overall, as expected this correct selection increases with the sample size, and shows the excellent performance of the proposed approaches.

To examine the ultra-high dimensional setup, a simulation was performed with $p = 10,000$. In this setting the marginal credible set approach was compared with using Sure Independence Screening coupled with SCAD (SIS + SCAD). This was implemented using the R package SIS. As discussed in the assumptions needed for selection consistency in this case, one way to generate data is to use a factor model

as described below assumption (B4). Specifically, the first $n/2$ predictors are generated from a multivariate normal whose covariance is again randomly generated from a Wishart distribution, centered on the identity, with degrees of freedom $n/2$. The remaining predictors are generated by an underlying $n/2$ latent standard normal factors. The loadings were then randomly generated also from standard normal distributions. As noted earlier, this factor model is one way to ensure that condition (B4) is satisfied.

For computation of the marginal credible set approach, the MCMC updates were done by updating the parameter vector in blocks of size $1,000$ to avoid large matrix operations. After obtaining draws from the posterior, it is only necessary to estimate the componentwise means and marginal variances, so it was found that sampling $5,000$ draws from the posterior was sufficient to obtain reliable estimates of the posterior mean and variances.

As in the previous scenario, the first $q = 3$ coefficients were considered non-zero and generated from $U(0, 1)$ yielding a theoretical $R^2 = 0.5$. The results for $p = 10,000$ and $n = \{100, 200, 500, 1000, 2000\}$ are shown in Table 4, again reporting the same quantities. In this ultra-high dimensional setting, it is clear that the marginal credible set method not only performs well, but clearly outperforms the SIS + SCAD method in terms of correct selection, and coverage while again maintaining a small model size. Note that as $n$ increases, the SIS + SCAD method appears to select a larger model on average. This phenomenon is likely attributed to the fact that the size of the initial screened model increases with $n$ as recommended in Fan and Lv (2008).

# 5 Analysis of PCR data

An experiment to examine the genetics of two inbred mouse populations (B6 and BTBR) was conducted by Lan et al. (2006). A total of 60 arrays were used to monitor the expression levels of 22,575 genes of 31 female and 29 male mice. Some physiolog-

ical phenotypes, including numbers of phosphoenopyruvate carboxykinase (PEPCK), glycerol-3-phosphate acyltransferase (GPAT), and stearoyl-CoA desaturase 1 (SCD1) were also measured by quantitative real-time PCR. Zhang, Lin, and Zhang (2009) used orthogonal components regression to predict each of these 3 phenotypes based on the gene expression data. The gene expression data and the phenotypic data are available at GEO (http://www.ncbi.nlm.nih.gov/geo; accession number GSE3330).

To reduce the number of candidate predictors for each of the three regressions, the first step was to screen down from the 22,575 genes to the 1,999 genes whose magnitude of marginal correlation with the response was largest. After this screening, the dataset for each of the 3 regressions consisted of $p = 2,000$ predictors (gender along with the 1,999 genes) and $n = 60$ observations. Note that the top 1,999 genes differed for each of the 3 responses. After screening, the LASSO estimator and both the joint and marginal credible set approaches were then used.

In addition, the full dataset using all 22,575 genes was fit using the marginal credible set approach to examine its performance in this ultra-high dimensional setting. For comparison, SIS + SCAD was also fit to the full set of predictors.

To allow for evaluation of the performance of the proposed selection approach, the sample of size 60 was first randomly split into a training set of size 55 and a holdout test set of size 5. The stopping rule for all methods was again chosen via BIC. This model was then used to predict the remaining 5 observations and the prediction error was used as comparison. The entire process was then repeated 100 times in order to reliably compare the resulting prediction errors. Table 5 gives the mean squared prediction error (along with its standard error) based on 100 random splits of the data into training sets of size 55 and testing set of size 5. Also included is the mean number of included predictors (with its standard error).

Overall, the results show that the proposed credible set approach performs well.

When using the full data, the marginal credible set approach has significantly smaller MSPE than SIS + SCAD for both the first response, PEPCK, and the second response, GPAT. For the third response, SCD1, the SIS + SCAD approach is slightly better in MSPE than the marginal credible sets, but the difference is small compared to standard errors. After screening to the top $p = 2,000$ predictors, it is seen that the joint credible set approach outperforms the LASSO on all 3 responses, while the marginal approach does slightly better than the joint approach for PEPCK. Overall, for this dataset, the proposed approaches generally have the better performance, whether it be using the full data or screening down to $p = 2,000$.

While the credible set approach uses default priors, the hyper-prior on the precision, $\tau$, was chosen as Gamma(0.01, 0.01). We now examine the sensitivity to this choice. For the first response, PEPCK, we examined the ordering of predictors induced by the credible set method using Gamma$(\alpha, \alpha)$ for $\alpha \in \{0.001, 0.01, 0.1, 1\}$. Each choice of $\alpha$ gives an ordered sequence of predictors. For each stopping point $K = 1, ..., p$, each choice of $\alpha$ gives a chosen model of size $K$. To assess agreement, for each $K$ we compared the model chosen by $\alpha = 0.01$ to the model chosen by an alternative choice of $\alpha$, and recorded the number of variables that appeared in both models. Figure 5 plots the number of variables that agree against the size of the model, $K$, for each of the 3 comparisons to $\alpha = 0.01$. All cases have very close agreement, so we conclude that the results are not sensitive to hyperparameter choice for these data.

# 6 Conclusion

This paper has introduced a new approach for variable ordering and selection based on Bayesian credible sets. It was shown that the joint credible set approach is consistent in selection for the fixed dimensional case, whereas the marginal credible set approach is consistent in selection even when the dimension increases exponentially

in the sample size. The finite sample performance of the method appears to be more stable than existing Bayesian methods, while requiring less prior specification and computing time. The proposed approach successfully accomplishes variable selection in the high-dimensional setting, while avoiding pitfalls that plague typical Bayesian variable selection methods. The approach is also competitive with the existing frequentist penalization approaches, and can often exhibit better performance than these methods in the high-dimensional case.

The proposed approach to consistent variable selection avoids the computation of posterior model probabilities, or the use of Bayes factors to exhaustively compare the set of models. In the context of computing posterior model probabilities, Moreno, Girón, and Casella (2010) have shown that the use of Bayes factors with intrinsic priors can achieve selection consistency only up to $p_n = O(n)$. They also show that using the Schwartz approximation is consistent only when $p_n = o(n)$.

In this paper, the linear regression model with the normal likelihood was used with a simple choice of conjugate prior. However, the general approach of the credible set method can be used for any model and prior specification as long as the posterior distribution is well-approximated by elliptical contours. Alternative choices of sparsity measure such as $L_1$ instead of an approximation to $L_0$ may be used. However, investigation of the selection properties for these choices would need further work.

# Acknowledgements

The authors are grateful to the editor, an associate editor, and three anonymous referees for their valuable comments. Bondell's research was partially supported by NSF grant DMS 1005612 and NIH grants P01-CA-142538-01 and R01-MH-084022-01. Reich's research was partially supported by NIH grant R01-ES-014843-02. The authors would like to thank Gareth James for providing the DASSO code for the Dantzig selector.

# References


Barbieri, M. and Berger, J. O. (2004). Optimal predictive model selection. *Annals of Statistics* **32**, 870-897.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. *Biometrics* **64**, 115-123.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373-384.

Brown, P.J., Vannucci, M. and Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society, Series B* **64**, 519-536.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with Discussion). *Annals of Statistics* **35**, 2313-2351.

Carlin, B. P. and Louis, T. A. (2009). Bayesian Methods for Data Analysis, CRC Press, New York.

Davis, J., and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd international conference on machine learning*, pp. 233240.

Dunson, D.B., Herring, A.H. and Engel, S.M. (2008). Bayesian selection and clustering of polymorphisms in functionally-related genes. *Journal of the American Statistical Association* **103**, 534-546.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). *Least angle regression. The Annals of Statistics* **32**, 407499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348-1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B* **70**, 849-911.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731-747.

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

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339-373.

Huang, J., Horowitz, J. L., and Ma, S. G. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics* **36**, 587-613.

Huang, J., Ma, S. G., and Zhang, C.-H. (2008). Adaptive Lasso for Sparse High-Dimensional Regression Models. *Statistica Sinica* **18**, 1603-1618.

James, G. M., Radchenko, P., and Lv, J. (2009). DASSO: connections between the Dantzig selector and lasso. *Journal of the Royal Statistical Society, Series B* **71**, 127-142.

Jung, S. and Marron, J. S. (2009). PCA consistency in High dimension, low sample size context. *Annals of Statistics* **37**, 4104-4130.

Kinney, S. and Dunson, D.B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics* **63**, 690-698.

Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E. T., Flowers, M. T., Schueler, K. L., Manly, K. F., Williams, R. W., Kendziorski, K., and Attie, A. D. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics* **2**, e6.

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

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics* **37**, 3498-3528.

Meinshausen, N. and Buhlmann, P. (2006). High dimensional graphs and variable selection with the lasso. *Annals of Statistics* **34**, 1436-1462.

Moreno, E., Girón, F. J., and Casella, G. (2010). Consistency of objective Bayes factors as the model dimension grows. *Annals of Statistics* **38**, 1937-1952.

O'Hara, R. B. and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis* **4**, 85-118.

Scott, J. G. and Berger, J. O. (2010). Bayes and Empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics* **38**, 2587-2619.

Tadesse, M.G., Sha, N. and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association* **100**, 602-617.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B* **67**, 91-108.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* **104**, 1512-1524.

Yata, K. and Aoshima, M. (2009). PCA Consistency for non-Gaussian data in high dimension, low sample size context. *Communications in Statistics - Theory and Methods* **38**.

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.

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

Zhang, D., Lin, Y. and Zhang, M. (2009). Penalized orthogonal-components regression for large p small n data. *Electronic Journal of Statistics* **3**, 781-796.

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.

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.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Annals of Statistics* **36**, 15091566.

# Appendix

Proof of Theorem.

To prove the theorem, we will need the following 4 lemmas.

LEMMA 1 *If $C_n \to \infty$ then $1 - \alpha_n \to 1$, i.e. the coverage increases to 1.*

**Proof of Lemma 1**

The true parameter vector is contained in the region if $n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T \frac{\Sigma^{-1}}{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \leq C_n$. Now, since $\tau_n = o(n)$, posterior consistency implies that $n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) = O_p(1)$. This, together with condition (A2), yields that $n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T \frac{\Sigma^{-1}}{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) = O_p(1)$. Hence if $C_n \to \infty$, the true parameter is contained in the region with probability tending to 1.

LEMMA 2 *If $C_n \to \infty$ and $n^{-1}C_n \to c$, for $0 < c \leq \infty$, then $(\tilde{\beta}_j - \beta_j^0)^{-1} = O_p(1)$, for some $j$, where $\tilde{\boldsymbol{\beta}}$ is the solution to the optimization problem for that choice of $C_n$. In particular, $\tilde{\boldsymbol{\beta}}$ is not a consistent estimator.*

**Proof of Lemma 2**

We will assume that $(\tilde{\beta}_j - \beta_j^0) \to 0$ for all $j$ and arrive at a contradiction.

We know that the solution occurs on the boundary of the credible set. So we have $(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^T \Sigma^{-1} (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) = C_n$. Now, $\hat{\boldsymbol{\beta}} = \left(X^T X + \tau_n I\right)^{-1} X^T Y$ and $\Sigma^{-1} = \tilde{s}^{-1}\left(X^T X + \tau_n I\right)$, where $\tilde{s} \to \sigma^2$. Dividing through on both sides by $n$, on the right hand side we have $\tilde{s} n^{-1} C_n \to \sigma^2 c$. For the left hand side, we have $(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^T \frac{\left(X^T X + \tau_n I\right)}{n}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})$.

Since $\tau_n = o(n)$, it follows that $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \to 0$. Furthermore, since $(\tilde{\beta}_j - \beta_j^0) \to 0$ for all $j$, we have $(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \to 0$. Hence $(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^T(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \to 0$. This, together with condition (A2) yields that $(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^T \frac{\left(X^T X + \tau_n I\right)}{n}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \to 0$.

Hence the left hand side goes to zero, while the right hand side remains bounded away from zero, and thus $P\left((\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^T\frac{\Sigma^{-1}}{n}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})=n^{-1}C_n\right)\to 0$. Hence, we obtain a contradiction.

LEMMA 3 *Let* $0<\varepsilon<1$. *If* $n^{-\varepsilon}C_n\to c$, *where* $0<c<\infty$, *then* $n^{1-\varepsilon}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^T\frac{\Sigma^{-1}}{n}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})=O_p(1)$, *and* $[n^{1-\varepsilon}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^T\frac{\Sigma^{-1}}{n}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})]^{-1}=O_p(1)$.

**Proof of Lemma 3**

We shall again proceed with proof by contradiction. Suppose $n^{-\varepsilon}C_n\to c$, and either $n^{1-\varepsilon}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^T\frac{\Sigma^{-1}}{n}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})\to 0$ or $\sqrt{n^{1-\varepsilon}}(\tilde{\beta}_j-\hat{\beta}_j)\to\infty$ for some $j$. Since the solution occurs on the boundary of the credible set, we have $n(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^T\frac{\Sigma^{-1}}{n}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})=C_n$. Multiplying both sides by $n^{-\varepsilon}$, on the right hand side we have $n^{-\varepsilon}C_n\to c$.

For the left hand side, we have $n^{1-\varepsilon}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^T\frac{\left(X^TX+\tau_nI\right)}{n}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})$. By assumption, $\lim n^{1-\varepsilon}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^T\frac{\left(X^TX+\tau_nI\right)}{n}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})$ is either 0 or $\infty$. Then, $P\left(n^{1-\varepsilon}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^T\frac{\Sigma^{-1}}{n}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})=n^{-\varepsilon}C_n\right)\to 0$, which yields a contradiction.

LEMMA 4 *Let* $0<\varepsilon<1$. *Then if* $n^{-\varepsilon}C_n\to c$, *it follows that* $\sqrt{n}(\tilde{\beta}_j-\hat{\beta}_j)\to\infty$ *can be true only for* $j\in A$.

**Proof of Lemma 4**

From Lemma 3, it follows that $\sqrt{n}(\tilde{\beta}_j-\hat{\beta}_j)\to\infty$ for some $j$. It will now be shown that it cannot be true for $j\in A^c$.

Suppose the true parameters are $\beta_1=\ldots=\beta_k=0$ and $\beta_{k+1}=\ldots=\beta_p\neq 0$. Let $S(\boldsymbol{\beta})=\sum_{j=1}^p|\hat{\beta}_j|^{-2}|\beta_j|$. The solution is the minimizer of $S(\boldsymbol{\beta})$ among the points within the given credible set.

Suppose $\tilde{\boldsymbol{\beta}}$ is the minimizer of $S(\boldsymbol{\beta})$, and suppose that $\sqrt{n}(\tilde{\beta}_1-\hat{\beta}_1)\to\infty$. Since $\beta_1\in A^c$, it follows that $\sqrt{n}\hat{\beta}_1=O_p(1)$. Hence, it must be that $\sqrt{n}\tilde{\beta}_1\to\infty$. It also follows that $\sqrt{n}\mid\hat{\beta}_1\mid^2\to 0$. So, the first term of $S(\tilde{\boldsymbol{\beta}})$, $\frac{\sqrt{n}|\tilde{\beta}_1|}{\sqrt{n}|\hat{\beta}_1|^2}\to\infty$. Therefore, $S(\tilde{\boldsymbol{\beta}})\to\infty$.

Now let $\tilde{\boldsymbol{\beta}}^0 = \{0, \ldots, 0, \tilde{\beta}^*_{k+1}, \ldots, \tilde{\beta}^*_p\}$ be the minimizer of $S(\boldsymbol{\beta})$, within the credible set while setting the first $k$ components to 0. First we need to show that for large enough $n$, there exists $\tilde{\boldsymbol{\beta}}^0$ of this form in the credible set. Since from Lemma 1, the coverage probability is going to 1, for large enough $n$ it must be that 0 will be covered for all $j \in A^c$. Hence, there exits a $\tilde{\boldsymbol{\beta}}^0$ of the above form within the credible set.

Now it will be shown that $S(\tilde{\boldsymbol{\beta}}^0) < S(\tilde{\boldsymbol{\beta}})$, and hence achieve a contradiction.

Since the credible set is shrinking around $\hat{\beta}$, it follows that $\frac{|\tilde{\beta}_j|}{|\hat{\beta}_j|^2} \rightarrow \mid \beta_j \mid^{-1}$, for all $j \in A$. Let $T = \sum_{j=k+1}^{p} \mid \hat{\beta}_j \mid^{-1}$, and let $\delta > 0$.

Now, since $S(\tilde{\boldsymbol{\beta}}) \rightarrow \infty$, there exists an $n_1$ such that for all $n > n_1$, it follows that $S(\tilde{\boldsymbol{\beta}}) > T + \delta$. Furthermore, there exists an $n_2$ such that for all $n > n_2$,

$$S(\tilde{\boldsymbol{\beta}}^0) = \frac{\mid \tilde{\beta}^*_{k+1} \mid}{\mid \hat{\beta}_{k+1} \mid^2} + ... + \frac{\mid \tilde{\beta}^*_p \mid}{\mid \hat{\beta}_p \mid^2} \leq T + \delta \tag{10}$$

Hence, there exists large enough $n$ so that $S(\tilde{\boldsymbol{\beta}}^0) < S(\tilde{\boldsymbol{\beta}})$ and thus $\tilde{\boldsymbol{\beta}}$ with $\sqrt{n}(\tilde{\beta}_1 - \hat{\beta}_1) \rightarrow \infty$ cannot be the minimizer. Therefore, $\sqrt{n}(\tilde{\beta}_j - \hat{\beta}_j) \rightarrow \infty$ can be true only for $j \in A$.

**Proof of Theorem 1**

In order to prove Theorem 1, we need to show that $P(A \bigcap A^c_n) \rightarrow 0$ and $P(A^c \bigcap A_n) \rightarrow 0$. By Lemma 3, it follows that the credible set is shrinking around $\hat{\boldsymbol{\beta}}$ when $C_n = o(n)$. Hence, we know that $\tilde{\beta}_j \rightarrow_p \beta_j$, for all $j$ and this gives $P(A^c \bigcap A_n) \rightarrow 0$.

We now want to show that $P(A \bigcap A^c_n) \rightarrow 0$. WLOG assume $\beta_1 = 0$, $\beta_p \neq 0$ and by Lemma 4, let $\sqrt{n}(\tilde{\beta}_p - \beta_p) \rightarrow \infty$. Let $\sigma_{ij}$ denote the $ij^{th}$ element of $\Sigma^{-1}$.

Assume $\tilde{\beta}_1 \neq 0$, and we will obtain a contradiction. Again, we have $(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^T \Sigma^{-1} (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) = C_n$. Then $\tilde{\beta}_p$ can be implicitly defined as a function of the remaining $p-1$ regression parameters. Since $\tilde{\beta}_1 \neq 0$, the minimizer of $\sum_{j=1}^{p} \frac{1}{\tilde{\beta}_j^2} \mid \beta_j \mid$ with respect to $\beta_1$

should then satisfy

$$\frac{1}{\mid \hat{\beta}_1 \mid^2} sign(\tilde{\beta}_1) + \frac{1}{\mid \hat{\beta}_p \mid^2} sign(\tilde{\beta}_p) \frac{\partial \beta_p}{\partial \beta_1}\Big|_{\tilde{\beta}} = 0. \tag{11}$$

Consider the first term on the left hand side. Since $\beta_1 = 0$, we have $\hat{\beta}_1 \to 0$, then $\frac{1}{|\hat{\beta}_1|^2} \to \infty$. Hence, $\frac{1}{|\hat{\beta}_1|^2} sign(\tilde{\beta}_1) \to \infty$ if $\tilde{\beta}_1 \neq 0$. For the second term, since $\beta_p \neq 0$, $\frac{1}{|\hat{\beta}_p|^2} = O_p(1)$.

Now, differentiating $(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^T \Sigma^{-1} (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) = C_n$ with respect to $\beta_1$ yields

$$\frac{\partial \beta_p}{\partial \beta_1}\Big|_{\tilde{\boldsymbol{\beta}}} = -\frac{\sum_{j=1}^p \sigma_{1j}(\tilde{\beta}_j - \hat{\beta}_j)}{\sum_{j=1}^p \sigma_{pj}(\tilde{\beta}_j - \hat{\beta}_j)} = -\frac{\sum_{j=1}^p \sigma_{1j}\sqrt{n^{1-\varepsilon}}(\tilde{\beta}_j - \hat{\beta}_j)}{\sum_{j=1}^p \sigma_{pj}\sqrt{n^{1-\varepsilon}}(\tilde{\beta}_j - \hat{\beta}_j)}, \tag{12}$$

where $0 < \varepsilon < 1$ is such that $n^{-\varepsilon} C_n \to c$. Note that both numerator and denominator in (12) are $O_p(1)$ by Lemma 3. Note also that using Lemma 4, the denominator cannot be 0, due to the presence of $(\tilde{\beta}_p - \hat{\beta}_p)$. Hence, $\frac{\partial \beta_p}{\partial \beta_1}\Big|_{\tilde{\beta}} = O_p(1)$. Then, $\frac{1}{|\hat{\beta}_p|^2} sign(\tilde{\beta}_p) \frac{\partial \beta_p}{\partial \beta_1}\Big|_{\tilde{\beta}} = O_p(1)$. Hence, the left hand side of (11) diverges, yielding a contradiction.

**Proof of Theorem 2**

It suffices to show this result in the orthogonal predictor case with $p_n = cn$, for $0 < c < 1$. In this case, we have $\hat{\beta} = \frac{n}{n+\tau_n} \frac{X^T}{n} (X\beta^0 + \boldsymbol{\varepsilon})$. So $\hat{\beta} - \beta^0 = -\frac{\tau_n}{n+\tau_n}\beta^0 + \frac{\tau_n}{n+\tau_n}\frac{X^T}{n}\boldsymbol{\varepsilon}$. Hence $E \sum_{j=1}^{p_n} \left(\hat{\beta} - \beta^0\right)^2 = \frac{\tau_n^2}{(n+\tau_n)^2} \sum_{j=1}^{p_n} \left(\beta_j^0\right)^2 + p_n \frac{\sigma^2}{n+\tau_n}$.

For mean square consistency, it must be that both bias and variance go to zero. In order for the first term to go to zero, we must have $\tau_n/n \to 0$. However, if $\tau_n/n \to 0$, then the second term is $O(p_n/n)$. Hence if $p_n = cn$ it follows that the right hand side cannot go to zero.

**Proof of Theorem 3**

The event $\{A_n \neq A\}$ is equivalent to the event $\{|\hat{\beta}_j| \leq s_j t_n \text{ for some } j \in A\} \bigcup \{|\hat{\beta}_j| > s_j t_n \text{ for some } j \in A^c\}$.

Now,

$P(|\hat{\beta}_j| \leq s_j t_n \text{ for some } j \in A) \leq \sum_{j\in A} P(|\hat{\beta}_j| \leq s_j t_n) = \sum_{j\in A} P(|\beta_j^0| - |\hat{\beta}_j| \leq |\beta_j^0| - s_j t_n) \leq \sum_{j\in A} P(\left||\beta_j^0| - |\hat{\beta}_j|\right| \leq |\beta_j^0| - s_j t_n)$. From the reverse Triangle Inequality, we have that $\sum_{j\in A} P(\left||\beta_j^0| - |\hat{\beta}_j|\right| \leq |\beta_j^0| - s_j t_n) \leq \sum_{j\in A} P(|\hat{\beta}_j - \beta_j^0| \geq |\beta_j^0| - s_j t_n)$.

In addition, letting $s^* = \max_{j\in A}\{|s_j|\} < \infty$, we have $\sum_{j\in A} P(|\hat{\beta}_j - \beta_j^0| \geq |\beta_j^0| - s_j t_n) \leq \sum_{j\in A} P(|\hat{\beta}_j - \beta_j^0| \geq \min_{j\in A}\{|\beta_j^0|\} - s^* t_n) \leq \sum_{j\in A} P(|\hat{\beta}_j - \beta_j^0| \geq s^* t_n) \leq \sum_{j=1}^{p_n} P(|\hat{\beta}_j - \beta_j^0| \geq t_n)$.

where the last inequality holds since $s_j \geq 1$ and the penultimate inequality holds for large enough $n$ since from (B3), for large $n$, we have that $\min_{j\in A}\{|\beta_j^0|\} > 2s^* t_n$. Furthermore,

$P(|\hat{\beta}_j| > s_j t_n \text{ for some } j \in A^c) \leq \sum_{j\in A^c} P(|\hat{\beta}_j - \beta_j^0| \geq s_j t_n) \leq \sum_{j=1}^{p_n} P(|\hat{\beta}_j - \beta_j^0| \geq t_n)$, which again holds since $s_j \geq 1$.

So

$P(A_n \neq A) \leq \sum_{j=1}^{p_n} P(|\hat{\beta}_j - \beta_j^0| \geq t_n)$.

Now,

$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 = \left[\left(\frac{X^T X}{n} + \frac{\tau_n}{n} I\right)^{-1} \frac{X^T X}{n} - I\right] \boldsymbol{\beta}^0 + \left(\frac{X^T X}{n} + \frac{\tau_n}{n} I\right)^{-1} \frac{X^T}{n} \boldsymbol{\varepsilon}$.

Note that $\left[\left(\frac{X^T X}{n} + \frac{\tau_n}{n} I\right)^{-1} \frac{X^T X}{n} - I\right] = -\frac{\tau_n}{n}\left(\frac{X^T X}{n} + \frac{\tau_n}{n} I\right)^{-1}$,

and thus $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \sim N(\mathbf{m}, V)$, where

$\mathbf{m} = -\frac{\tau_n}{n}\left(\frac{X^T X}{n} + \frac{\tau_n}{n} I\right)^{-1} \boldsymbol{\beta}^0$

and

$V = \frac{\sigma^2}{n}\left(\frac{X^T X}{n} + \frac{\tau_n}{n} I\right)^{-1} \frac{X^T X}{n} \left(\frac{X^T X}{n} + \frac{\tau_n}{n} I\right)^{-1}$.

Letting $m^* = \max_j\{|m_j|\}$ and $V^* = \max_j\{V_{jj}\}$, we have

$\sum_{j=1}^{p_n} P(|\hat{\beta}_j - \beta_j^0| \geq t_n) = \sum_{j=1}^{p_n} P(\frac{|\hat{\beta}_j - \beta_j^0| - |m_j|}{\sqrt{V^*}} \geq \frac{t_n - |m_j|}{\sqrt{V^*}}) \leq \sum_{j=1}^{p_n} P(\frac{|\hat{\beta}_j - \beta_j^0 - m_j|}{\sqrt{V^*}} \geq \frac{t_n - m^*}{\sqrt{V^*}}) \leq \sum_{j=1}^{p_n} P(\frac{|\hat{\beta}_j - \beta_j^0 - m_j|}{\sqrt{V_{jj}}} \geq \frac{t_n - m^*}{\sqrt{V^*}}) = p_n P(|Z| \geq \frac{t_n - m^*}{\sqrt{V^*}}) \leq 2\frac{p_n\sqrt{V^*}}{t_n - m^*} \exp\left(-\frac{(t_n - m^*)^2}{2V^*}\right)$,

where the last inequality follows from the tail probability bound using normality. Note that we must have $t_n > m^*$ otherwise $P(|Z| \geq \frac{t_n - m^*}{\sqrt{V^*}}) = 1$ and the bound becomes $p_n$, and clearly would not have selection consistency. It will be shown later, that this will

be true under the given assumptions.

The next step is to bound both the maximum bias, $m^*$, and the maximum variance, $V^*$.

To bound the variance,

$V = \frac{\sigma^2}{n}\left(\frac{X^TX}{n} + \frac{\tau_n}{n}I\right)^{-1}\frac{X^TX}{n}\left(\frac{X^TX}{n} + \frac{\tau_n}{n}I\right)^{-1} = \frac{\sigma^2}{n}\left\{\left(\frac{X^TX}{n} + \frac{\tau_n}{n}I\right)^{-1} - \frac{\tau_n}{n}\left(\frac{X^TX}{n} + \frac{\tau_n}{n}I\right)^{-2}\right\}$.

So

$V^* = \frac{\sigma^2}{n}|| \text{ Diag } \left\{\left(\frac{X^TX}{n} + \frac{\tau_n}{n}I\right)^{-1}\right\} - \frac{\tau_n}{n} \text{ Diag } \left\{\left(\frac{X^TX}{n} + \frac{\tau_n}{n}I\right)^{-2}\right\}||_\infty$
$\leq \frac{\sigma^2}{n}|| \text{ Diag } \left\{\left(\frac{X^TX}{n} + \frac{\tau_n}{n}I\right)^{-1}\right\}||_\infty \leq \frac{\sigma^2}{nh_{p_n,n}}$, where $h_{p_n,n}$ denotes the smallest eigenvalue of $\frac{X^TX}{n} + \frac{\tau_n}{n}I$.

Since $h_{p_n,n} \geq \frac{\tau_n}{n}$, it follows that

$V^* \leq \frac{\sigma^2}{\tau_n}$.

To bound the bias,

$m^* = ||\frac{\tau_n}{n}\left(\frac{X^TX}{n} + \frac{\tau_n}{n}I\right)^{-1}\boldsymbol{\beta}^0||_\infty = ||\Gamma_n D_n \Gamma_n^T \boldsymbol{\beta}^0||_\infty$, where $D_n = \text{ Diag } \left(\frac{\tau_n}{nd_1+\tau_n}, ..., \frac{\tau_n}{nd_{r_n}+\tau_n}, 1, ..., 1\right)$.

So

$m^* = ||\Gamma_n D_n \Gamma_n^T \boldsymbol{\beta}^0||_\infty = ||\Gamma_n D_n \Gamma_n^T \Gamma_{n,1}\eta_1 + \Gamma_n D_n \Gamma_n^T \Gamma_{n,2}\eta_2||_\infty \leq ||\Gamma_n D_n \Gamma_n^T \Gamma_{n,1}\eta_1||_\infty +$
$||\Gamma_n D_n \Gamma_n^T \Gamma_{n,2}\eta_2||_\infty = ||\Gamma_n D_n \left(\eta_1^T, \mathbf{0}^T\right)^T||_\infty + ||\Gamma_n D_n \left(\mathbf{0}^T, \eta_2^T\right)^T||_\infty = ||\Gamma_n D_n \left(\eta_1^T, \mathbf{0}^T\right)^T||_\infty +$
$||\Gamma_{n,2}\eta_2^T||_\infty$.

From condition (B4), the second term is $O(n^{-1}\tau_n\sqrt{q_n})$.

For the first term, we have

$||\Gamma_n D_n \left(\eta_1^T, \mathbf{0}^T\right)^T||_\infty \leq ||\Gamma_n D_n \left(\eta_1^T, \mathbf{0}^T\right)^T||_2 = ||D_n \left(\eta_1^T, \mathbf{0}^T\right)^T||_2 \leq \frac{\tau_n}{nd_{r_n}+\tau_n}||\eta_1||_2 =$
$\frac{\tau_n}{nd_{r_n}+\tau_n}||\Gamma_{n,1}\eta_1||_2 \leq \frac{\tau_n}{nd_{r_n}+\tau_n}||\boldsymbol{\beta}^0||_2$.

Since all components of $\boldsymbol{\beta}^0$ are finite and there are $q_n$ non-zero, it follows that $\frac{\tau_n}{nd_{r_n}+\tau_n}||\boldsymbol{\beta}^0||_2 =$ $O(n^{-1}\tau_n\sqrt{q_n})$ as well.

Hence, from condition (B2), it follows that $t_n > 2m^*$ for sufficiently large $n$.

Thus, we have that $P(A_n \neq A) \leq 2\frac{p_n\sqrt{V^*}}{t_n/2}\exp\left(-\frac{(t_n/2)^2}{2V^*}\right) \leq 2\frac{p_n\sqrt{\sigma^2/\tau_n}}{t_n/2}\exp\left(-\frac{(t_n/2)^2}{2\sigma^2/\tau_n}\right) =$ $4\sqrt{\sigma^2}\frac{p_n}{t_n\sqrt{\tau_n}}\exp\left(-\frac{t_n^2\tau_n}{8\sigma^2}\right)$.

Now, under condition (B2), the right hand side converges to zero, and hence $P(A_n \neq A) \leq 4\sqrt{\sigma^2}\frac{p_n}{t_n\sqrt{\tau_n}} \exp\left(-\frac{t_n^2\tau_n}{8\sigma^2}\right) \to 0$.

**Proof of Theorem 4**

First note that $P(\max_j |\hat{\beta}_j - \beta_j^0| \geq \epsilon) \leq \sum_{j=1}^{p_n} P(|\hat{\beta}_j - \beta_j^0| \geq \epsilon)$. During the course of proving Theorem 3, it was shown that $\sum_{j=1}^{p_n} P(|\hat{\beta}_j - \beta_j^0| \geq \epsilon) \leq 2\frac{p_n\sqrt{V^*}}{\epsilon - m^*} \exp\left(-\frac{(\epsilon - m^*)^2}{2V^*}\right)$, where $m^* = \max_j\{|m_j|\}$ and $V^* = \max_j\{V_{jj}\}$. It was also shown that $V^* \leq \frac{\sigma^2}{\tau_n}$ and $m^* = O(n^{-1}\tau_n\sqrt{q_n})$. Hence the result follows directly.

Table 1: Mean area under the ROC curve and the PRC curve for $p = 50$ predictors, $n = 60$ observations, based on 200 datasets with standard errors in parentheses. The 4 choices of SSVS (using 10,000 MCMC runs) represent the possible combinations of prior variance for the regression parameters along with the prior inclusion probabilities. For example, SSVS (fixed, random) denotes SSVS with a fixed prior variance for the coefficients along with a random prior inclusion probability. The Adaptive Lasso, Lasso, Dantzig Selector, and the OLS estimator are also shown for comparison. Also reported is CPU time (in sec) for a single dataset.

| | ROC Area | | PRC Area | | CPU Time |
|---|---|---|---|---|---|
| | $\rho = 0.5$ | $\rho = 0.9$ | $\rho = 0.5$ | $\rho = 0.9$ | |
| Joint Credible Sets | 0.920 (0.005) | 0.911 (0.005) | 0.879 (0.006) | 0.852 (0.006) | 1.52 |
| Marginal Credible Sets | 0.897 (0.005) | 0.886 (0.005) | 0.834 (0.007) | 0.803 (0.007) | 1.52 |
| SSVS (fixed, fixed) | 0.913 (0.004) | 0.904 (0.004) | 0.845 (0.006) | 0.804 (0.007) | 40.07 |
| SSVS (random, fixed) | 0.919 (0.004) | 0.918 (0.003) | 0.855 (0.006) | 0.826 (0.006) | 40.07 |
| SSVS (fixed, random) | 0.915 (0.004) | 0.908 (0.004) | 0.851 (0.006) | 0.811 (0.007) | 40.07 |
| SSVS (random, random) | 0.916 (0.004) | 0.914 (0.004) | 0.854 (0.006) | 0.823 (0.006) | 40.07 |
| Adaptive LASSO | 0.798 (0.006) | 0.705 (0.006) | 0.725 (0.007) | 0.536 (0.008) | 0.03 |
| LASSO | 0.912 (0.005) | 0.874 (0.005) | 0.866 (0.006) | 0.790 (0.008) | 0.03 |
| Dantzig | 0.922 (0.004) | 0.912 (0.005) | 0.881 (0.006) | 0.858 (0.007) | 0.44 |
| Forward | 0.843 (0.006) | 0.755 (0.007) | 0.782 (0.007) | 0.651 (0.007) | 0.44 |
| OLS | 0.689 (0.007) | 0.582 (0.007) | 0.521 (0.009) | 0.338 (0.008) | $< 0.01$ |

Table 2: Mean area under the ROC curve and the PRC curve for $p = 500$ predictors, $n = 60$ observations, based on 200 datasets with standard errors in parentheses. The 4 choices of SSVS represent the possible combinations of prior variance for the regression parameters along with the prior inclusion probabilities. For example, SSVS (fixed, random) denotes SSVS with a fixed prior variance for the coefficients along with a random prior inclusion probability. In the case of fixed prior inclusion probability, this value was optimally set to the true proportion of non-zero coefficients, 0.02. Also reported is CPU run time for a single dataset.

| | ROC Area | | PRC Area | | CPU Time (sec) |
|---|---|---|---|---|---|
| | $\rho = 0.5$ | $\rho = 0.9$ | $\rho = 0.5$ | $\rho = 0.9$ | |
| Joint Credible Sets | 0.946 (0.004) | 0.989 (0.001) | 0.708 (0.011) | 0.873 (0.007) | 20.93 |
| Marginal Credible Sets | 0.932 (0.004) | 0.979 (0.002) | 0.687 (0.011) | 0.862 (0.007) | 20.93 |
| SSVS (fixed, fixed) | 0.902 (0.005) | 0.924 (0.004) | 0.620 (0.011) | 0.634 (0.010) | 1222.91 |
| SSVS (random, fixed) | 0.929 (0.004) | 0.957 (0.003) | 0.672 (0.010) | 0.693 (0.009) | 1222.91 |
| SSVS (fixed, random) | 0.897 (0.005) | 0.924 (0.004) | 0.615 (0.011) | 0.656 (0.010) | 1222.91 |
| SSVS (random, random) | 0.925 (0.005) | 0.955 (0.003) | 0.665 (0.010) | 0.692 (0.009) | 1222.91 |

Table 3: Selection performance for $p = 50$ for various choices of $n$ based on 200 datasets. The entries in the table denote Correct Selection Proportion (CS), Coverage Proportion (COV), Average Model Size (MS), and Average Number of Important Predictors out of the 3 Included (IP).

| | $n = 60$ | | | | $n = 100$ | | | | $n = 200$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CS | COV | MS | IP | CS | COV | MS | IP | CS | COV | MS | IP |
| Joint Sets | 16.5 | 43.0 | 4.15 | 2.29 | 25.5 | 49.0 | 3.45 | 2.37 | 44.5 | 66.0 | 3.16 | 2.62 |
| Marginal Sets | 17.0 | 43.0 | 4.48 | 2.29 | 24.5 | 43.0 | 3.28 | 2.30 | 44.5 | 58.0 | 2.97 | 2.52 |
| LASSO | 3.5 | 46.0 | 5.09 | 2.28 | 10.5 | 53.0 | 4.20 | 2.40 | 23.0 | 71.5 | 4.08 | 2.67 |
| SCAD | 11.0 | 35.0 | 2.49 | 1.86 | 20.5 | 39.5 | 3.06 | 2.18 | 29.5 | 50.5 | 3.89 | 2.44 |
| | $n = 500$ | | | | $n = 1000$ | | | | $n = 2000$ | | | |
| | CS | COV | MS | IP | CS | COV | MS | IP | CS | COV | MS | IP |
| Joint Sets | 55.5 | 71.0 | 3.02 | 2.68 | 62.0 | 76.5 | 3.00 | 2.75 | 74.0 | 82.0 | 2.96 | 2.81 |
| Marginal Sets | 50.5 | 61.5 | 3.03 | 2.58 | 56.0 | 70.0 | 2.99 | 2.66 | 70.0 | 78.0 | 2.96 | 2.76 |
| LASSO | 20.0 | 74.0 | 4.00 | 2.72 | 28.0 | 79.5 | 3.86 | 2.77 | 33.0 | 85.0 | 3.90 | 2.84 |
| SCAD | 45.5 | 77.5 | 4.03 | 2.76 | 52.0 | 80.5 | 3.58 | 2.79 | 58.0 | 81.0 | 3.71 | 2.72 |

Table 4: Selection performance for $p = 10,000$ for various choices of $n$ based on 100 datasets. The entries in the table denote Correct Selection Proportion (CS), Coverage Proportion (COV), Average Model Size (MS), and Average Number of Important Predictors out of the 3 Included (IP).

| | $n = 100$ | | | | $n = 200$ | | | | $n = 500$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CS | COV | MS | IP | CS | COV | MS | IP | CS | COV | MS | IP |
| Marginal Sets | 9.0 | 31.0 | 3.22 | 2.06 | 24.0 | 47.0 | 3.37 | 2.38 | 39.0 | 54.0 | 3.01 | 2.49 |
| SIS + SCAD | 1.0 | 15.0 | 4.08 | 1.82 | 5.0 | 35.0 | 6.06 | 2.28 | 6.0 | 59.0 | 11.62 | 2.56 |
| | $n = 1000$ | | | | $n = 2000$ | | | | | | | |
| | CS | COV | MS | IP | CS | COV | MS | IP | | | | |
| Marginal Sets | 45.0 | 61.0 | 2.98 | 2.58 | 62.0 | 74.0 | 2.89 | 2.71 | | | | |
| SIS + SCAD | 12.0 | 64.0 | 14.62 | 2.62 | 23.0 | 79.0 | 17.96 | 2.78 | | | | |

Table 5: Mean squared prediction error and model size based on 100 random splits of the real data, with standard errors in parenthesis. The 3 response variables are PEPCK, GPAT, and SCD1.

| | PEPCK | | GPAT | | SCD1 | |
|---|---|---|---|---|---|---|
| | MSPE | Model Size | MSPE | Model Size | MSPE | Model Size |
| Marginal Sets ($p = 22,576$) | 2.14 (0.15) | 7.1 (0.41) | 4.70 (0.45) | 9.3 (0.59) | 3.54 (0.26) | 7.6 (0.54) |
| SIS + SCAD ($p = 22,576$) | 2.82 (0.18) | 2.3 (0.09) | 5.88 (0.44) | 2.6 (0.10) | 3.44 (0.22) | 3.2 (0.14) |
| Joint Sets ($p = 2,000$) | 2.03 (0.14) | 9.6 (0.46) | 3.83 (0.34) | 4.2 (0.43) | 3.04 (0.22) | 22.0 (0.56) |
| Marginal Sets ($p = 2,000$) | 1.84 (0.14) | 23.3 (0.67) | 5.33 (0.41) | 21.8 (0.72) | 3.27 (0.21) | 19.1 (0.71) |
| LASSO ($p = 2,000$) | 3.03 (0.19) | 7.7 (0.96) | 5.03 (0.42) | 3.3 (0.79) | 3.25 (0.31) | 19.7 (0.77) |

Figure 1: Plot of the mean ROC and PRC curves over the 200 datasets for $p = 50$ predictors, $n = 60$ observations. The first column is the ROC curve, the second column is the PRC curve. The first row is for $\rho = 0.5$, the second row is for $\rho = 0.9$.

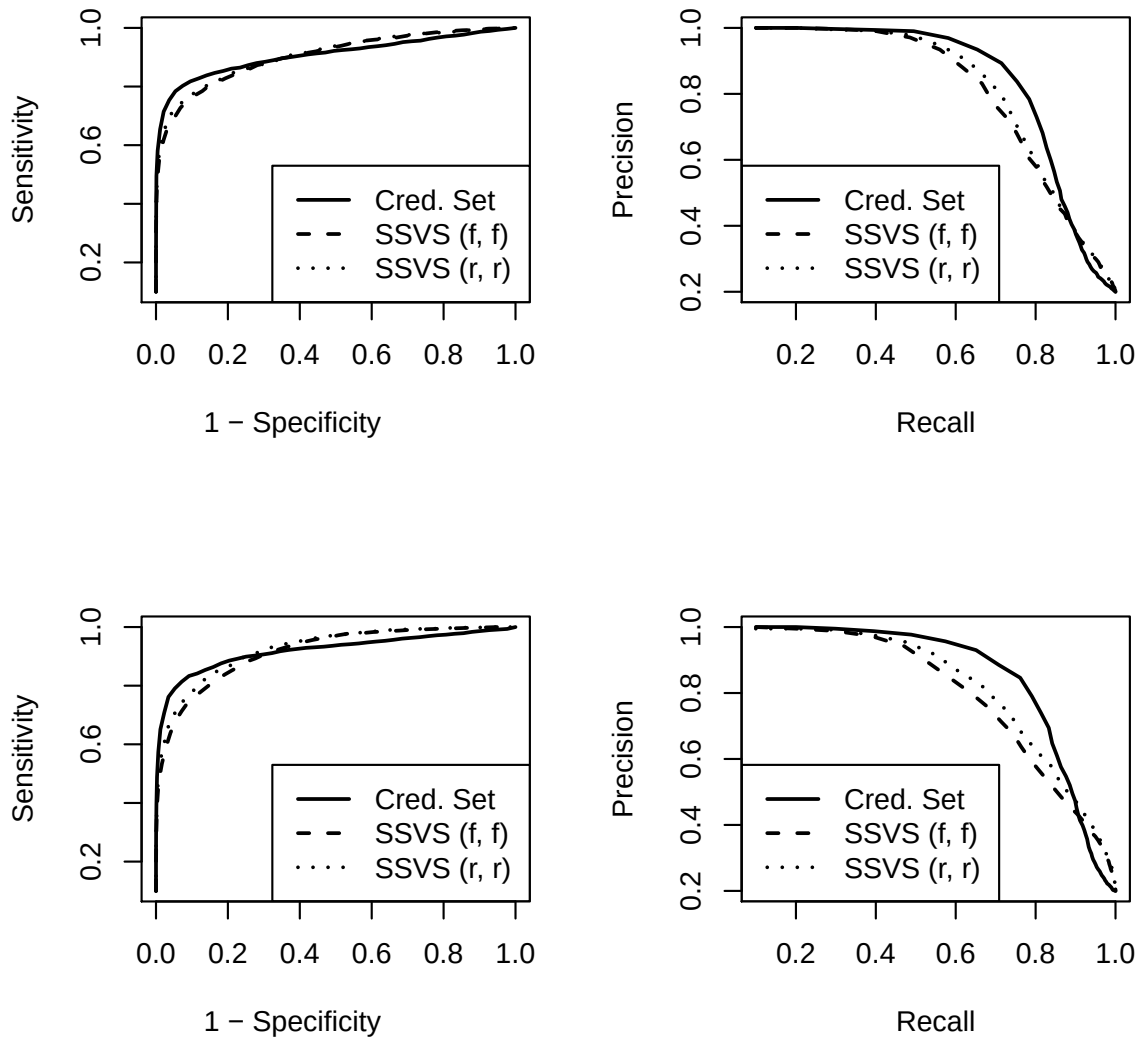

Figure 2: Plot of the mean ROC and PRC curves over the 200 datasets for $p = 500$ predictors, $n = 60$ observations. The first column is the ROC curve, the second column is the PRC curve. The first row is for $\rho = 0.5$, the second row is for $\rho = 0.9$.

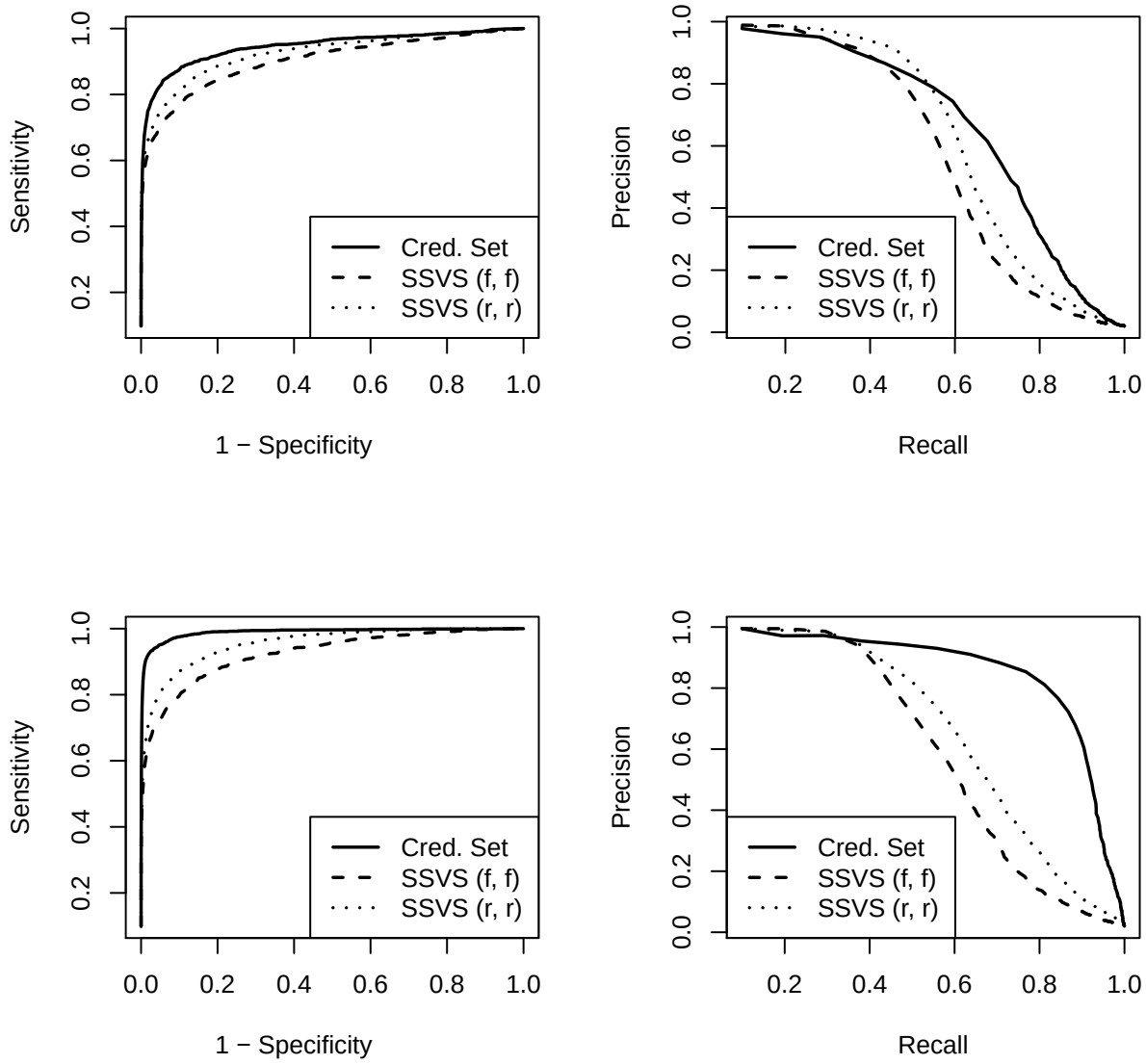

Figure 3: Plot of the mean ROC and PRC curves over the 200 datasets for $p = 2000$ predictors, $n = 60$ observations. The first column is the ROC curve, the second column is the PRC curve. The first row is for $\rho = 0.5$, the second row is for $\rho = 0.9$.

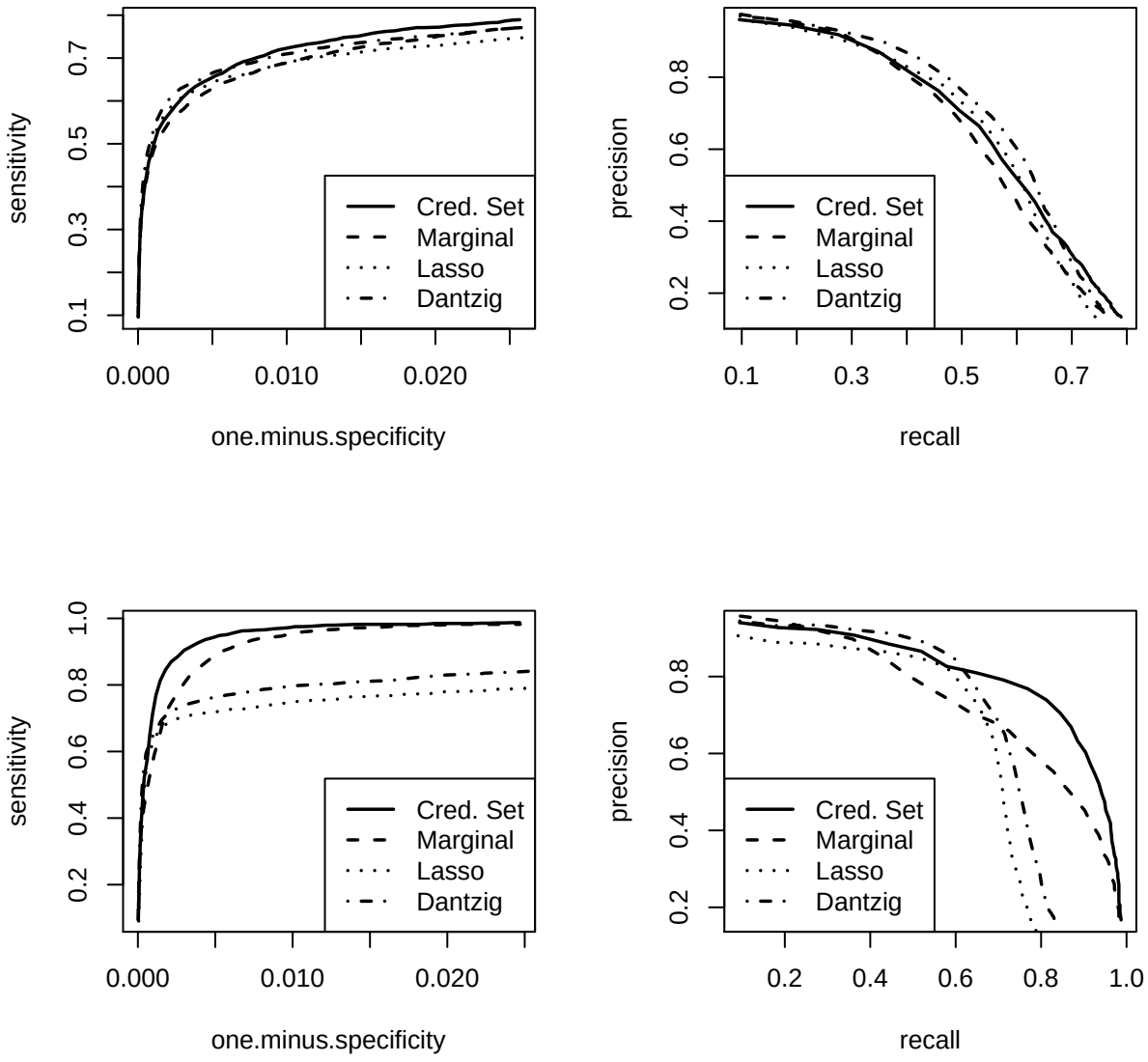

Figure 4: Plot of the mean prediction error over the 200 datasets as a function of the number of predictors included in the model for $n = 60$ observations. The first column is for $p = 50$, the second column is for $p = 500$. The first row is for $\rho = 0.5$, the second row is for $\rho = 0.9$.

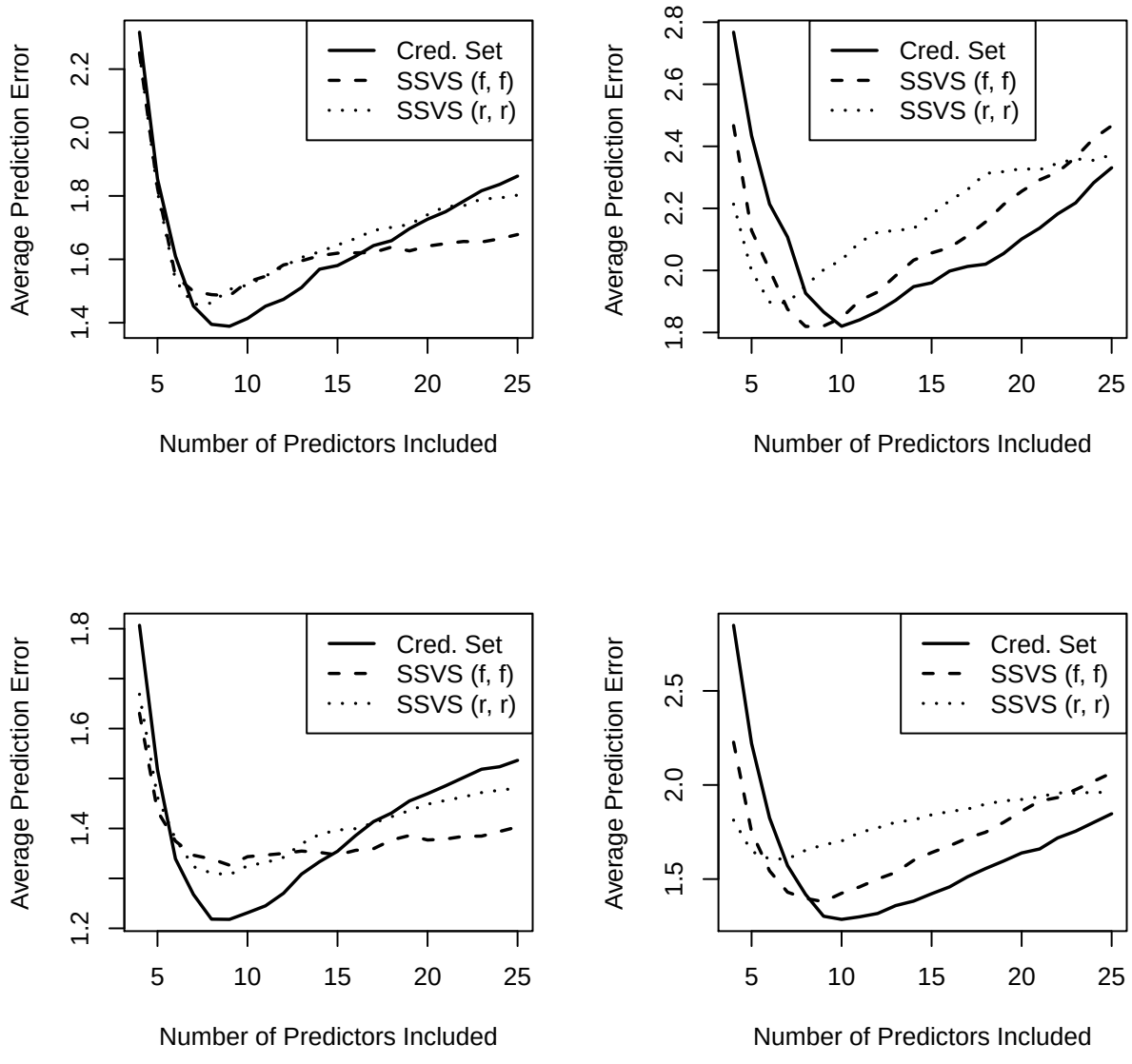

Figure 5: Sensitivity of variable ordering to prior precision. The horizontal axis denotes the number of variables selected, while the vertical axis represents the number that match from each method. The parameter of the inverse gamma distribution is varied. Each of 0.001, 0.1, and 1 are compared relative to the choice of 0.01 in terms of agreement.