Bayesian Analysis (2018) **13**, Number 3, pp. 823–844

Variable Selection via Penalized Credible Regions with Dirichlet–Laplace Global-Local Shrinkage Priors

Yan Zhang* and Howard D. Bondell†

Abstract. The method of Bayesian variable selection via penalized credible regions separates model fitting and variable selection. The idea is to search for the sparsest solution within the joint posterior credible regions. Although the approach was successful, it depended on the use of conjugate normal priors. More recently, improvements in the use of global-local shrinkage priors have been made for high-dimensional Bayesian variable selection. In this paper, we incorporate global-local priors into the credible region selection framework. The Dirichlet–Laplace (DL) prior is adapted to linear regression. Posterior consistency for the normal and DL priors are shown, along with variable selection consistency. We further introduce a new method to tune hyperparameters in prior distributions for linear regression. We propose to choose the hyperparameters to minimize a discrepancy between the induced distribution on R-square and a prespecified target distribution. Prior elicitation on R-square is more natural, particularly when there are a large number of predictor variables in which elicitation on that scale is not feasible. For a normal prior, these hyperparameters are available in closed form to minimize the Kullback–Leibler divergence between the distributions.

Keywords: variable selection, posterior credible region, global-local shrinkage prior, Dirichlet–Laplace, posterior consistency, hyperparameter tuning.

1 Introduction

High dimensional data has become increasingly common in all fields. Linear regression is a standard and intuitive way to model dependency in high dimensional data. Consider the linear regression model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{X}$ is the $n \times p$ high-dimensional set of covariates, $\boldsymbol{Y}$ is the n scalar responses, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ is the p-dimensional coefficient vector, and $\boldsymbol{\varepsilon}$ is the error term assumed to have $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$. Ordinary least squares is not feasible when the number of predictors p is larger than the sample size n. Variable selection is necessary to reduce the large number of candidate predictors. The classical variable selection methods include subset selection, criteria such as Akaike Information Criterion (AIC; Akaike, 1973) and Bayesian Information Criterion (BIC; Schwarz et al., 1978), and penalized methods such as the least absolute shrinkage and selection operator (Lasso; Tibshirani, 1996), smoothly clipped absolute deviation (SCAD; Fan and Li, 2001), the elastic net

*Department of Biostatistics, Johns Hopkins University, yzhan284@jhu.edu
†Department of Statistics, North Carolina State University, bondell@stat.ncsu.edu

© 2018 International Society for Bayesian Analysis https://doi.org/10.1214/17-BA1076

(Zou and Hastie, 2005), adaptive Lasso (Zou, 2006), the Dantzig selector (Candes and Tao, 2007), and octagonal shrinkage and clustering algorithm for regression (OSCAR; Bondell and Reich, 2008).

In the Bayesian framework, approaches for variable selection include: stochastic search variable selection (SSVS) (George and McCulloch, 1993), Bayesian regularization (Park and Casella, 2008; Li et al., 2010; Polson et al., 2013; Leng et al., 2014), empirical Bayes variable selection (George and Foster, 2000), spike and slab variable selection (Ishwaran and Rao, 2005), and global-local (GL) shrinkage priors. Those traditional Bayesian methods conduct variable selection either relying on the calculation of posterior inclusion probabilities for each predictor or each possible model, or a choice of posterior threshold.

Typical global-local shrinkage priors are represented as the class of global-local scale mixtures of normals (Polson and Scott, 2010),

$$\beta_j \sim N(0, w\xi_j),\ \xi_j \sim \pi(\xi_j),\ (w, \sigma^2) \sim \pi(w, \sigma^2), \tag{2}$$

where w controls the global shrinkage towards the origin, while ξ_j allows local deviations of shrinkage. Various options of shrinkage priors for $\boldsymbol{\beta}$, include normal-gamma (Griffin et al., 2010), Horseshoe prior (Carvalho et al., 2009, 2010), generalized double Pareto prior (Armagan et al., 2013a), Dirichlet–Laplace (DL) prior (Bhattacharya et al., 2015), Horseshoe+ prior (Bhadra et al., 2016), and others that can be represented as (2). The GL shrinkage priors usually shrink small coefficients greatly due to a tight peak at zero, and rarely shrink large coefficients due to the heavy tails. It has been shown that GL shrinkage priors have improved posterior concentrations (Bhattacharya et al., 2015). However, the shrinkage prior itself would not lead to variable selection, and to go further, some rules need to be set on the posteriors.

Bondell and Reich (2012) proposed a Bayesian variable selection method only based on posterior credible regions. However, the implementation and results of that paper depended on the use of conjugate normal priors. Due to the improved concentration, incorporating the global-local shrinkage priors into this framework can perform better, both in theory and practice. We show that the DL prior yields consistent posteriors in this regression setting, along with selection consistency.

Another difficulty in high dimensional data is the choice of hyperparameters, which can highly affect the results. In this paper, we also propose an intuitive default method to tune the hyperparameters in the prior distributions. By minimizing a discrepancy between the induced distribution of R^2 from the prior and the desired distribution (Beta distribution by default), one gets a default choice of hyperparameter value. For the choice of normal priors, the hyperparameter that minimizes the Kullback–Leibler (KL) divergence between the distributions is shown to have a closed form solution.

Overall, compared to other Bayesian methods, on the one hand, our method makes use of the advantage of global-local shrinkage priors, which can effectively shrink small coefficients and reliably estimate the coefficients of important variables simultaneously. On the other hand, by using the credible region variable selection approach, we can easily transform the non-sparse posterior estimators to sparse solutions. Compared to

the common frequentist method, our approach provides flexibility to estimate the tuning parameter jointly with the regression coefficients, allows easy incorporation of external information or hierarchical modeling into Bayesian regularization framework, and leads to straightforward computing through Gibbs sampling.

The remainder of the paper is organized as follows. Section 2 reviews the penalized credible region variable selection method. Section 3 details the proposed method which combines shrinkage priors and penalized credible region variable selection. Section 4 presents the posterior consistency under the choice of shrinkage priors, as well as the asymptotic behavior of the selection consistency for diverging p. Section 5 discusses a default method to tune the hyperparameters in the prior distributions based on the induced prior distribution on R^2. Section 6 reports the simulation results, and Section 7 gives the analysis of a real-time polymerase chain reaction (real-time PCR) dataset. All proofs are given in the supplementary material (Zhang and Bondell, 2017).

2 Background

Bondell and Reich (2012) proposed a penalized regression method based on Bayesian credible regions. First, the full model is fit using all predictors with a continuous prior. Then based on the posterior distribution, a sequence of joint credible regions are constructed, within which, one searches for the sparsest solution. The choice of a conjugate normal prior of

$$\boldsymbol{\beta}|\sigma^2, \gamma \sim N(0, \sigma^2/\gamma \boldsymbol{I}_p) \tag{3}$$

is used, where σ^2 is the error variance term as in (1), and γ is the ratio of prior precision to error precision. The variance, σ^2, is often given a diffuse inverse Gamma prior, while γ is the hyperparameter which is either chosen to be fixed or given a Gamma hyperprior.

The credible region is to find $\tilde{\boldsymbol{\beta}}$, such that

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} ||\boldsymbol{\beta}||_0 \text{ subject to } \boldsymbol{\beta} \in \mathcal{C}_\alpha, \tag{4}$$

where $||\boldsymbol{\beta}||_0$ is the L_0 norm of $\boldsymbol{\beta}$, i.e., the number of nonzero elements, and $\mathcal{C}_\alpha$ is the $(1-\alpha)\times 100\%$ posterior credible regions based on the particular prior distributions. The use of elliptical posterior credible regions yields the form $\mathcal{C}_\alpha = \{\boldsymbol{\beta} : (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq c_\alpha\}$, for some nonnegative c_α, where $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}$ are the posterior mean and covariance respectively. Then by replacing the L_0 penalization in (4) with a smooth homotopy between L_0 and L_1 proposed by Lv and Fan (2009) and linear approximation, the optimization problem in (4) becomes

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \lambda_\alpha \sum_{j=1}^{p} |\hat{\beta}_j|^{-2} |\beta_j|, \tag{5}$$

where there exists a one-to-one correspondence between c_α and λ_α. The sequence of solutions to (5) can be directly accomplished by plugging in the posterior mean and covariance and using the LARS algorithm (Efron et al., 2004). Note that a fixed α is not

used here, and the actual value of α is not needed to examine the sequence. To understand this method better, although Bondell and Reich (2012) start from constructing a posterior credible region, now to find the sparsest estimator within the credible region problem has become a penalized variable selection problem as shown in (5). By changing the value of λ_α, the variables will be selected in sequence into the model.

3 Penalized Credible Regions with Global-Local Shrinkage Priors

3.1 Motivation

Global-local shrinkage priors produce a posterior distribution with good empirical and theoretical properties. Compared to the usual normal prior, GL priors concentrate more along the regions with zero parameters. This leads to a better estimate of the uncertainty about parameters in the full model based on the posterior distribution. The penalized credible region variable selection approach separates model fitting and variable selection. So it seems natural to fit the model under a GL shrinkage prior, and then conduct variable selection through the penalized credible region method. The motivation is that within the same credible region level, GL shrinkage priors would lead to more concentrated posteriors, thus having better performance for variable selection, by finding sparse solutions more easily. Bondell and Reich (2012) demonstrated via simulations and real data examples that the credible region approach using the normal prior distribution improved on the performance of both Bayesian Stochastic Search and Frequentist approaches, such as Lasso, Dantzig Selector, and SCAD. The use of the GL shrinkage priors instead of the normal is a natural approach. In addition, although we do not have uncertainty about the model, the full posterior is obtained first, so that uncertainty about the parameters can be used based on the full model posterior distribution. Using the global-local shrinkage prior gives a more concentrated posterior even if we did not add the penalized credible region model selection step to choose an estimate of the model.

Although GL shrinkage priors would not lead to elliptical posterior distributions, valid credible regions can still be constructed using elliptical contours. These would no longer be the high density regions, but would remain valid regions. Elliptical contours would also be reasonable approximations to the high density regions, at least around the largest mode. Thus, the penalized credible region selection method can be feasibly performed by plugging the posterior mean and covariance matrix into the optimization algorithm (5). So given any GL prior, once Markov-Chain Monte-Carlo (MCMC) steps produce the posterior samples, the sample mean, $\hat{\boldsymbol{\beta}}$, and sample covariance, $\boldsymbol{\Sigma}$, would hence be obtained, then variable selection can be performed through the penalized credible region method. In this paper, we modify the Dirichlet–Laplace (DL) prior to implement in the regression setting. We also consider the Laplace prior, also referred as Bayesian Lasso, described in Park and Casella (2008) and Hans (2010), as

$$\beta_j \sim \mathrm{DE}(\sigma/\lambda) \ (j = 1 \cdots, p), \tag{6}$$

where λ is the Lasso parameter, controlling the global shrinkage.

3.2 Dirichlet–Laplace Priors

For the normal mean model, Bhattacharya et al. (2015) proposed a new class of Dirichlet–Laplace (DL) shrinkage priors, possessing the optimal posterior concentration property. We construct the generalization of the DL priors for the linear regression model. The proposed hierarchical DL prior is as follows: for $j = 1, \cdots, p$,

$$\begin{aligned} \beta_j|\sigma, \phi_j, \tau &\sim \text{DE}(\sigma\phi_j\tau), \\ (\phi_1, \cdots, \phi_p) &\sim \text{Dir}(a, \cdots, a), \\ \tau &\sim \text{Ga}(pa, 1/2), \end{aligned} \tag{7}$$

where $\text{DE}(b)$ denotes a zero mean Laplace kernel with density $f(y) = (2b)^{-1}\exp(-|y|/b)$ for $y \in \mathbb{R}$, $\text{Dir}(a, \cdots, a)$ is the Dirichlet distribution with concentration vector $(a, \cdots, a)$, and $\text{Ga}(pa, 1/2)$ denotes a Gamma distribution with shape pa and rate $1/2$. Here, small values of a would lead most of $(\phi_1, \cdots, \phi_p)$ to be close to zero and only few of them nonzero; while large values allow less singularity at zero, thus controlling the sparsity of regression coefficients. The ϕ_j's are the local scales, allowing deviations in the degree of shrinkage. As pointed out in Bhattacharya et al. (2015), τ controls global shrinkage towards the origin and to some extent determines the tail behaviors of the marginal distribution of β_j's. We also assume a common prior on the variance term σ^2, $\text{IG}(a_1, b_1)$, the inverse Gamma distribution with shape a_1 and scale b_1.

3.3 Computation of Posteriors

For posterior computation, the Gibbs sampling steps proposed in Bhattacharya et al. (2015) can be modified to accommodate the linear regression model. The DL prior (7) can be equivalently denoted as

$$\begin{aligned} \beta_j|\sigma^2, \phi_j, \psi_j, \tau &\sim N(0, \sigma^2\psi_j\phi_j^2\tau^2), \\ \psi_j &\sim \text{Exp}(1/2), \\ (\phi_1, \cdots, \phi_p) &\sim \text{Dir}(a, \cdots, a), \\ \tau &\sim \text{Ga}(pa, 1/2), \end{aligned} \tag{8}$$

where $\text{Exp}(\cdot)$ is the usual exponential distribution. Note that DL prior is also a global-local shrinkage prior as it is a particular form of (2). Gibbs sampling steps would be obtained based on (8). The derivation is similar as in Bhattacharya et al. (2015), hence omitted here. The parameterization of the three-parameter generalized inverse Gaussian (giG) distribution, $Y \sim \text{giG}(\chi, \rho, \lambda_0)$, means the density of Y is $f(y) \propto y^{\lambda_0 - 1}\exp\{-0.5(\rho y + \chi/y)\}$ for $y > 0$. Then the summary of the Gibbs sampling steps are as below:

(i) Sample $\sigma^2|\beta, \psi, \phi, \tau, y$. Draw σ^2 from an inverse Gamma distribution, $\text{IG}(a_1 + (n+p)/2, b_1 + \{\boldsymbol{\beta}^T\boldsymbol{S}^{-1}\boldsymbol{\beta} + (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\}/2)$, where $\boldsymbol{S} = \text{diag}(\psi_1\phi_1^2\tau^2, \cdots, \psi_p\phi_p^2\tau^2)$.

(ii) Sample $\beta|\psi, \phi, \tau, \sigma^2, y$. Draw $\boldsymbol{\beta}$ from a $N(\boldsymbol{\mu}, \sigma^2\boldsymbol{V})$, where $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{S}^{-1})^{-1}$ with the same $\boldsymbol{S}$ as above, and $\boldsymbol{\mu} = \boldsymbol{V}\boldsymbol{X}^T\boldsymbol{Y} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{S}^{-1})^{-1}(\boldsymbol{X}^T\boldsymbol{Y})$.

(iii) Sample $\psi_j|\phi_j,\tau,\beta,\sigma^2$. First draw $\psi_j^{-1}|\phi_j,\tau,\beta,\sigma^2$, $j=1,\cdots,p$, independently from the distribution InvGaussian($\mu_j = \sigma\phi_j\tau/|\beta_j|, \lambda_0 = 1$), where InvGaussian($\mu,\lambda_0$) denotes the inverse Gaussian with density $f(y) = \sqrt{\lambda_0/(2\pi y^3)}\exp\{-\lambda_0(y-\mu)^2/(2\mu^2 y)\}$ for $y>0$. Then take the reciprocal to get the draws of ψ_j ($j=1,\cdots,p$).

(iv) Sample $\tau|\phi,\beta,\sigma^2$. Draw τ from a giG($\chi = 2\sum_{j=1}^p |\beta_j|/(\phi_j\sigma), \rho=1, \lambda_0 = pa-p$).

(v) Sample $\phi_j|\beta,\sigma^2$. Draw $T_1,\cdots,T_p$ independently with $T_j \sim$ giG($\chi = 2|\beta_j|/\sigma, \rho = 1, \lambda_0 = a-1$), then set $\phi_j = T_j/T$ where $T=\sum_{j=1}^p T_j$.

4 Asymptotic Theory

In this section, we first study the posterior properties of the normal and DL prior, when both n and p_n go to infinity, and further investigate the selection consistency of the penalized variable selection method. Assume the true regression parameter is $\boldsymbol{\beta}_n^0$, and the estimated regression parameter is $\boldsymbol{\beta}_n$. Denote the true set of non-zero coefficients is $\mathcal{A}_n^0 = \{j_n : \beta_{nj_n}^0 \neq 0, j_n = 1,\cdots,p_n\}$, and the estimated set of non-zero coefficients is $\mathcal{A}_n = \{j_n : \beta_{nj_n} \neq 0, j_n = 1,\cdots,p_n\}$. Also let $q_n = |\mathcal{A}_n^0|$ denote the number of predictors with nonzero true coefficients. As $n\to\infty$, consider the sequence of credible sets of the form $\{\boldsymbol{\beta}_n : (\boldsymbol{\beta}_n - \hat{\boldsymbol{\beta}}_n)^T \boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\beta}_n - \hat{\boldsymbol{\beta}}_n) \le c_n\}$, where $\hat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\Sigma}_n$ are the posterior mean and covariance matrix respectively, and c_n is a sequence of non-negative constants. Let $\boldsymbol{\Gamma}_n$ denote the $p_n \times p_n$ matrix whose columns are eigenvectors of $\boldsymbol{X}_n^T\boldsymbol{X}_n/n$ ordered by decreasing eigenvalues, i.e., $d_1 \ge d_2 \ge \cdots \ge d_{p_n} \ge 0$. Then $\boldsymbol{X}_n^T\boldsymbol{X}_n/n = \boldsymbol{\Gamma}_n\boldsymbol{D}_n\boldsymbol{\Gamma}_n^T$ where $\boldsymbol{D}_n = \text{diag}\{d_1,\cdots,d_{p_n}\}$.

Assume the following regularity conditions throughout.

(A1) The error terms ε_i, $i=1,\cdots,n$, are independent and identically distributed (i.i.d.) with mean zero and finite variance σ^2;

(A2) $0 < d_{\min} < \liminf_{n\to\infty} d_{p_n} \le \limsup_{n\to\infty} d_1 < d_{\max} < \infty$, where $d_{\min}$ and $d_{\max}$ are fixed;

(A3) $\limsup_{n\to\infty} \max_{j_n=1,\cdots,p_n} |\beta_{nj_n}^0| < \infty$;

(A4) $p_n = o(n)$;

(A5) $(n/p_n)^{1/2} \min_{j_n\in\mathcal{A}_n^0} |\beta_{nj_n}^0| \to \infty$ as $n\to\infty$.

Assumption (A2) regarding the eigenvalues bounded away from 0 and ∞ is a sufficient condition for estimation consistency in the Bayesian methods, and also for the consistency of Ordinary Least Squares in the case of growing dimension but with $p_n = o(n)$. This is akin to the condition in the fixed dimension of $\boldsymbol{X}^T\boldsymbol{X}/n$ converging to a positive definite matrix (Assumption (A2) in Bondell and Reich (2012)). The basic intuition is that without a lower bound on the eigenvalue, there is an asymptotic singularity, which then leaves a linear combination of the regression parameters that is not identifiable,

i.e, it would have a variance that was infinite, hence could not be consistent. The upper bound, on the other hand, ensures that there is a proper covariance matrix for every p_n. If we assume that each row of $\boldsymbol{X}_n$ was a random draw from a p_n-dimensional probability distribution, the bounded eigenvalue condition is an assumption on the true sequence of covariance matrices, as for large n and $p_n = o(n)$, the sample covariance, $\boldsymbol{X}_n^T \boldsymbol{X}_n/n$, (assuming centered variables) will converge to the true covariance. As an example of a common covariance structure that would have the bounded eigenvalue property, consider an AR(1) covariance structure with parameter ρ among the p_n variables. Then, the results of Stroeker (1983), show that (for $\rho \geq 0$), as $p_n \to \infty$, the largest eigenvalue will converge to $(1+\rho)^2$ and the smallest will converge to $(1-\rho)^2$. For $\rho < 0$, the largest becomes $(1-\rho)^2$ and smallest becomes $(1+\rho)^2$. Hence, both the smallest and largest eigenvalues have bounded limits.

Also note that Assumption (A5) restricts the minimum signal size for the non-zero coefficients. This sufficient condition comes from the fact that the radius of the credible region is $O((p_n/n)^{1/2})$. Hence, in order for the region not to contain $\beta_{nj_n} = 0$ for those having $\beta^0_{nj_n} \neq 0$, the true $\beta^0_{nj_n}$ must be further away than $O((p_n/n)^{1/2})$.

4.1 Posterior Consistency: Normal and DL Priors

Armagan et al. (2013b) investigate the asymptotic behavior of posterior distributions of regression coefficients in the linear regression model (1) as p_n grows with n. They prove the posterior consistency under the assumption of a variety of priors, including the Laplace prior, Student's t prior, generalized double Pareto prior, and the Horseshoe-like priors. By definition, posterior consistency implies that the posterior distribution of $\boldsymbol{\beta}_n$ converges in probability to $\boldsymbol{\beta}^0_n$, i.e., for any $\epsilon > 0$, $P(\boldsymbol{\beta}_n : ||\boldsymbol{\beta}_n - \boldsymbol{\beta}^0_n|| > \epsilon | \boldsymbol{Y}_n) \to 0$ as $p_n, n \to \infty$. In this section, we show that the normal and Dirichlet–Laplace prior also yield consistent posteriors. However, the DL prior can yield consistent posteriors under weaker conditions on the signal.

Theorem 1. *Under Assumptions (A1)–(A4), if* $q_n = o(n^{1-\rho}/(p_n \log^2 n))$ *for* $\rho \in (0,1)$, *and* $(\sigma^2/\gamma_n)^{1/2} = C/({p_n}^{1/2} n^{\rho/2} \log n)$ *for finite* $C > 0$, *the normal prior (3) yields a consistent posterior.*

Theorem 2. *Under Assumptions (A1)–(A4), if* $q_n = o(n/\log n)$, *and* $a_n = C/(p_n n^{\rho} \log n)$ *for any finite* $\rho > 0$ *and finite* $C > 0$, *the Dirichlet–Laplace prior (7) yields a consistent posterior.*

Note that the difference in the above two theorems is the number of nonzero components, i.e., q_n. As $n/\log n > n^{1-\rho}/(p_n \log^2 n)$, the Dirichlet–Laplace prior leads to posterior consistency in a much broader domain, compared to the normal prior as well as compared to the Laplace prior who also yields consistent posteriors as shown in Theorem 2 in Armagan et al. (2013b). This strengthens the justification for replacing the normal prior with the DL prior theoretically. However, note that the theorems only give a sufficient condition for posterior consistency under each of the priors. The sufficient condition does have a broader domain for q_n in Theorem 2, for the Dirichlet–Laplace prior, than in Theorem 1 for the normal prior. However, it is not clear that these conditions are also necessary, so although we are able to prove the consistency for the DL

prior under a more general condition than the normal prior, there may be room to improve this condition in either or both of these cases.

4.2 Selection Consistency of Penalized Credible Regions

Bondell and Reich (2012) have shown that when p is fixed and $\boldsymbol{\beta}$ is given the normal prior in (3), the penalized credible region method is consistent in variable selection. In this paper, we show that the consistency of the posterior distribution under a global-local shrinkage prior also yields consistency in variable selection under the case of $p_n = o(n)$.

Theorem 3. *Under Assumptions (A1)–(A5), given the normal prior in (3), if* $\lim_{n\to\infty} c_n/p_n \to c$ *with* $1 \leq c < \infty$, *and the prior precision,* $\gamma_n = o(n)$, *then the penalized credible region method with optimization problem (4) is consistent in variable selection, i.e.* $P(\mathcal{A}_n = \mathcal{A}_n^0) \to 1$.

The proof is given in the Appendix. The selection consistency allows us to expect that the true model is contained in the credible regions with high probability, when the number of predictors increases together with the sample size. Such selection consistency is obtained under the normal prior. However, as reviewed in Section 1, since the GL shrinkage priors can be expressed as a scale mixture of normals, as long as the posterior distribution of the precision is $o(n)$ with probability 1 (analogous to $\gamma_n = o(n)$ in the normal prior), then the result can be directly applied to the GL shrinkage prior.

Theorem 4. *Under Assumptions (A1)–(A5), given any global-local shrinkage prior represented as (2), if the conditions of posterior consistency are satisfied, then the posterior distribution of the precision is* $o(n)$ *with probability 1 as* $n \to \infty$. *Furthermore, if* $\lim_{n\to\infty} c_n/p_n \to c$ *with* $1 \leq c < \infty$, *then the penalized credible region method with optimization problem (4) with the particular shrinkage prior is consistent in variable selection, i.e.* $P(\mathcal{A}_n = \mathcal{A}_n^0) \to 1$.

So given the conditions of posterior consistency under the global-local shrinkage prior, we automatically get the selection consistency of the credible region method. For example, for the DL prior in (7), we have the following result.

Corollary 1. *Under Assumptions (A1)–(A5), given the DL prior in (7),* $q_n = o(n/\log n)$, $a_n = C/(p_n n^{\rho} \log n)$ *for any finite* $\rho > 0$ *and finite* $C > 0$, *if* $\lim_{n\to\infty} c_n/p_n \to c$ *with* $1 \leq c < \infty$, *then the penalized credible region method with optimization problem (4) is consistent in variable selection, i.e.* $P(\mathcal{A}_n = \mathcal{A}_n^0) \to 1$.

Note that the variable selection consistency is derived based on the posterior consistency. However, Assumption (A3) is not necessary to ensure the variable selection consistency. If (A3) is not satisfied, i.e., $\boldsymbol{\beta}_n^0$ is truly unbounded, although it would not be possible to obtain a consistent estimator, or posterior, the credible region would become bounded away from zero in that direction, and hence will pick out that direction consistently as well.

5 Tuning Hyperparameters

The value of hyperparameters in the prior distribution plays an important role in the posteriors. For example, in the normal prior (3), γ is the hyperparameter, whose value controls the degree of shrinkage. This is often chosen to be fixed at a "large" value or given a hyperprior. However, the choice of the "large" value affects the results, as does the choice of hyperprior such as a gamma prior, particularly in the high dimensional case. Also, in the DL prior (7), the choice of a is critical. If a is too small, then the DL prior would shrink each dimension of $\boldsymbol{\beta}$ towards zero; while, if a is too large, there would be no strong concentration around the origin. Instead of fixing a, a discrete uniform prior can be given on a supported on some interval (for example, $[1/\max(n,p), 1/2]$), with several support points on the interval. However, introducing the hyperprior for the hyperparameters will not only arise new values to tune, but also increase the complexity of the MCMC sampling. In practice, although the specification of a p-dimensional prior on $\boldsymbol{\beta}$ may be difficult, some prior information on a univariate function may be easier. The motivation is to incorporate such prior information of the one-dimensional function into the priors on the p-dimensional $\boldsymbol{\beta}$.

In this paper, we propose an intuitive way to tune the values of hyperparameters, by incorporating a prior on R^2 (the coefficient of determination). Practically, a scientist may have information on R^2 from previous experiments, and this can be coerced into say a Beta(a, b) distribution. In this way, tuning hyperparameters is equivalent to searching for the hyperparameter which leads to the induced distribution of R^2 closest to the desired distribution. Intuitively, if we fix any value for b, as we increase a, then R^2 will approach 1, hence this controls the total size of the signal that is anticipated in the data. As we will see shortly, it is a prior on the value of the quadratic form $\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta}$. Combining this with the choice of prior, gives also the degree of sparsity. For example, with a Dirichlet–Laplace Prior, the parameter in the DL distribution then controls how this total signal is distributed to the coefficients, either to a few coefficients, giving a sparse model, or to many coefficients, giving a dense model. In many cases, a scientist may have done many similar experiments before and can look back and see the values of the sample coefficient of determination from all of these studies. Then treating this as a sample from a Beta distribution, the hyperparameters, a and b, can be obtained from this fit. Without any prior information for R^2, a uniform prior, Beta$(1, 1)$, may be used as default.

For the linear regression model (1), the population form can be represented as $y = \boldsymbol{x}^T \boldsymbol{\beta} + \varepsilon$, with $\boldsymbol{x}$ independent of ε. Let σ_y^2 be the marginal variance of y and σ^2 be the variance of the random error term. The definition of the POPULATION R^2 is given by:

$$pop\ R^2 = 1 - \frac{\sigma^2}{\sigma_y^2},$$

which is the proportion of the variation of y in the population explained by the independent variables. Furthermore, for fixed $\boldsymbol{\beta}$, it follows that $\sigma_y^2 = \boldsymbol{\beta}^T \mathrm{Cov}(\boldsymbol{x}) \boldsymbol{\beta} + \sigma^2$. Assume $E(\boldsymbol{x}) = 0$, then we can estimate $\mathrm{Cov}(\boldsymbol{x})$ by $\boldsymbol{X}^T \boldsymbol{X}/n$. So R^2 as a function of $\boldsymbol{\beta}$ and σ^2 is given by $R^2 = 1 - \sigma^2/(\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta}/n + \sigma^2)$. Given that the form of prior distributions

considered includes σ in the scale, it follows that $\boldsymbol{\beta} = \sigma\boldsymbol{\eta}$ for $\boldsymbol{\eta}$ having the distribution of the prior fixed with $\sigma^2 = 1$. Hence

$$R^2 = 1 - \frac{1}{1 + \boldsymbol{\eta}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\eta}/n}. \tag{9}$$

For a specified prior on $\boldsymbol{\eta}$, the induced distribution of R^2 can be derived based on (9). Then the hyperparameters which yield the induced distribution of R^2 closest to the desired distribution is the tuned value.

For a better understanding of the intuition here, we give a simple example. Suppose $\sigma^2 = 1$ and we have an intercept only model, i.e., model (1) is simplified as $\boldsymbol{Y} = \mathbf{1_n}\beta + \boldsymbol{\varepsilon}$ with $\mathbf{1_n}$ the n-dimensional vector with all elements of 1. Then (9) can be written as $R^2 = 1 - (1 + \beta^2)^{-1}$. Suppose the desired distribution for R^2 is Beta(a, b), then the corresponding induced distribution for β is

$$f_\beta(t) = \frac{2\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \left(\frac{t^2}{1+t^2}\right)^{a-1} \left(\frac{1}{1+t^2}\right)^{b+1} |t|,$$

where $\Gamma(.)$ denotes the gamma function. The left panel of Figure 1 shows the distribution on R^2 for 4 choices of hyperparameters in the Beta distribution, while the right panel shows the corresponding induced prior distribution on β. We see that for a uniform distribution on R^2, we obtain a distribution on β that puts its mass slightly skewed away from zero on each side. For a bathtub distribution ($a = b = 0.5$), we see it reduces to the Cauchy distribution, giving heavy tails to obtain the R^2 near one, and the peak around zero to obtain the R^2 near zero. We also see two other extremes, as $a \to 0$ for fixed $b = 1$, we obtain a distribution that decays very quickly and puts most of its mass around zero, as expected; while as $b \to 0$ and a fixed at 1, we obtain a density proportional to $|t|/(1 + t^2)$, allowing for larger values of β with high probability.

In practice, one can consider a grid of possible values of the hyperparameters. For each value, draw a vector $\boldsymbol{\eta}$. This is converted to a draw of R^2. Given this hyperparameter, a comparison between the sample of R^2 and the desired distribution is performed, for example, a Kolmogorov–Smirnov (KS) test. The best fit is then chosen. The whole tuning process only involves the prior distributions, no MCMC sampling, thus avoiding comprehensive computing.

However, given a specific prior for $\boldsymbol{\beta}$, based on (9), the exact induced distribution of R^2 can be derived, which relies on the value of hyperparameters. By minimizing the Kullback–Liebler directed divergence between such distribution and the desired distribution (Beta distribution by default), a default hyperparameter value can be found. For continuous random variables with density function f_1 and f_2, the KL divergence is defined as

$$D(f_1|f_2) = \int_{-\infty}^{\infty} f_1(x) \log(f_1(x)/f_2(x))\, dx.$$

For the choice of normal priors, the following theorem shows that there is a closed form solution for the hyperparameter to minimize the KL divergence for large p.

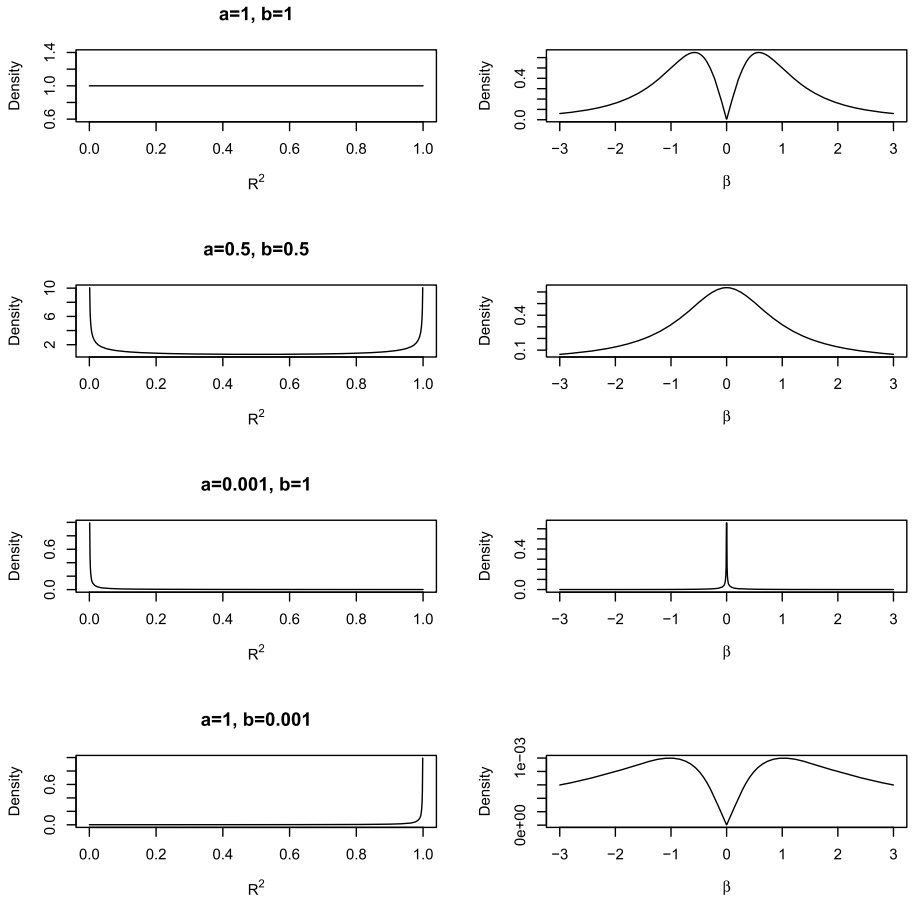

Figure 1: Beta(a, b) density for R^2 and the corresponding induced distribution density for β.

Theorem 5. *For the normal prior in (3), to minimize the KL directed divergence between the induced distribution of R^2 and the Beta(a, b) distribution, as $p \to \infty$, the hyperparameter, γ, is chosen to be $(A+\sqrt{B})^{1/3}+(A-\sqrt{B})^{1/3}-P/3$, where $P = (2a-b)\sum_{j=1}^{p} d_j/a$, $Q = 2(a+b)\sum_{j=1}^{p} d_j^2/a + (a-2b)(\sum_{j=1}^{p} d_j)^2/a$, $R = -b(\sum_{j=1}^{p} d_j)^3/a$, $C = P^2/9 - Q/3$, $A = PQ/6 - P^3/27 - R/2$, $B = A^2 - C^3 \geq 0$, and $d_1, \cdots, d_p$ denote the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}/n$.*

In theory, for other continuous priors, one can derive the optimal hyperparameters similarly. However, sometimes the calculation can be quite complex. In this case, the simulation-based approach discussed earlier can be implemented. However, since GL priors can be represented as mixture normal priors (see Section 1), by matching its prior precision with that of the normal prior, the derived default solution as shown in Theorem 5 can offer an intuitive idea for the hyperparameter values in the GL shrinkage priors.

6 Simulation Results

6.1 Comparisons of Different Priors

To compare the performance of the penalized credible region variable selection method using different shrinkage priors, including the normal prior (3), Laplace prior (6), and DL prior (7), a simulation study is conducted. Bondell and Reich (2012) demonstrated the improvement in performance of the credible region approach using the normal prior over both Bayesian and Frequentist approaches, such as SSVS, Lasso, adaptive Lasso, Dantzig Selector, and SCAD. Given the previous comparisons, the focus here is to see if replacing the normal prior with the global-local prior can even further improve the performance of the credible region variable selection approach.

We use a similar simulation setup as in Bondell and Reich (2012). In each setting, 200 datasets are simulated from the linear model (1) with $\sigma^2 = 1$, sample size $n = 60$, and the number of predictors p varying in $\{50, 500, 1000\}$. To represent different correlation settings, X_{ij} are generated from standard normal distribution, and the correlation between x_{ij_1} and x_{ij_2} is $\rho^{|j_1 - j_2|}$, with $\rho = 0.5$ and 0.9. The true coefficient $\boldsymbol{\beta}$ is $(\mathbf{0}_{10}^T, \boldsymbol{B}_1{}^T, \mathbf{0}_{20}^T, \boldsymbol{B}_2{}^T, \mathbf{0}_{p-40}^T)^T$ for $p \in \{50, 500, 1000\}$ in which $\mathbf{0}_k$ represents the k-dimensional zero vector, $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are both 5-dimensional vector generated component-wise and uniform from $(0, 1)$. For each case of shrinkage prior, the posterior mean and covariance can be obtained from the Gibbs samplers, and then plugged into the optimization algorithm (5) of the penalized credible region method to implement the variable selection.

For each method, the induced ordering of the predictors are created. We consider the resulting model at each ordering step to measure the performance. For each step on the ordering, true positives (TP) are defined as those selected variables which also appear in the true model. False positives (FP) are those selected variables which also do not appear in the true model. True negatives (TN) correspond to those not selected variables which are not in the true model. False negatives (FN) refer to variables which are not selected in the model, but indeed are in the true 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, where FPR is the fraction of FP's of the fitted model in the total number of irrelevant variables in the true model, and TPR is the fraction of TP's of the fitted model in the total number of important variables in the true model. The Precision-Recall (PRC) curve plots the precision on the y-axis, and the Recall (or TPR or Sensitivity) on x-axis, where precision is the ratio of true positives to the total declared positive number.

The compared credible set methods are listed as below:

- Method "Normal_hyper", refers to the normal prior, with "non-informative" hyperparameters, i.e., $N(0, \sigma_b^2)$ is the prior for β, and IG$(0.001, 0.001)$ prior is given for σ_b^2.
- Method "Normal_tune", refers to the normal prior (3), where γ is tuned through the R^2 method introduced in Section 5, with a target of uniform distribution.

- Method "Laplace_hyper", means Laplace prior (6), with λ given a Ga$(1, 1)$ prior.
- Method "Laplace_tune", means Laplace prior (6), and λ is tuned through the R^2 method introduced in Section 5, with a target of uniform distribution.
- Method "DL_hyper" is the DL prior (7), in which a is given a discrete uniform prior supported on the interval $[1/\max(n, p), 1/2]$ with 1000 support points in this interval.
- Method "DL_tune" is the DL prior (7), in which a is tuned through the R^2 method introduced in Section 5, with a target of uniform distribution.

In all above cases, the variance term σ^2 is given an IG$(0.001, 0.001)$ prior. In addition, we show the results from using the Lasso (Tibshirani, 1996) fit via the LARS algorithm (Efron et al., 2004).

	ROC Area		PRC Area	
	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.9$
Lasso	0.900 (0.0047)	0.815 (0.0052)	0.694 (0.0053)	0.628 (0.0068)
Normal_hyper	0.909 (0.0048)	0.899 (0.0041)	0.782 (0.0054)	0.749 (0.0058)
Normal_tune	0.949 (0.0037)	0.978 (0.0020)	0.830 (0.0043)	0.845 (0.0039)
Laplace_hyper	0.890 (0.0049)	0.859 (0.0052)	0.756 (0.0058)	0.691 (0.0069)
Laplace_tune	0.942 (0.0040)	0.976 (0.0020)	0.820 (0.0046)	0.844 (0.0039)
DL_hyper	0.917 (0.0044)	0.908 (0.0044)	0.786 (0.0052)	0.749 (0.0062)
DL_tune	0.939 (0.0039)	0.945 (0.0032)	0.811 (0.0048)	0.802 (0.0050)

Table 1: Mean area under the ROC Curve and the PRC curve for $p = 50$, $n = 60$, based on 200 datasets with standard errors in parentheses.

For the above priors (normal, Laplace and DL), we ran the MCMC chain (Gibbs sampling) for $15,000$ iterations, with the first $5,000$ for burn-in. Posterior mean and covariance were calculated based on the $10,000$ samples, which were then plugged into the penalized credible interval optimization algorithm (5), to conduct variable selection. Table 1 gives the mean and standard error for the area under the ROC and PRC curve for $p = 50$ with $\rho \in \{0.5, 0.9\}$. In addition, Figure 2 plots the mean ROC and PRC curves of the 200 datasets for the selected above methods to compare. Table 2 and Figure 3 give the results for the $p = 500$ case. Table 3 and Figure 4 show the results for the $p = 1000$ case. Since the Lasso estimator can select at most $\min\{n, p\}$ predictors, when $p = 500$ or 1000, the ROC and PRC curves cannot be fully constructed. So the area under the curves cannot be compared directly for Lasso with other methods, which are omitted in Tables 2 and 3, but partial ROC and PRC curves can still be plotted, which are shown in Figures 3 and 4.

On the one hand, in terms of whether given a hyperprior for the hyperparameter or tuning hyperparameters through the R^2 method proposed in Section 5 would lead to better posterior performance, one might compare each "$*$_hyper" and "$*$_tune" pair in Tables 1, 2 and 3. In general, for all three priors, the tuning method leads to significantly better posterior performance than the hyperprior method in all simulation setups.

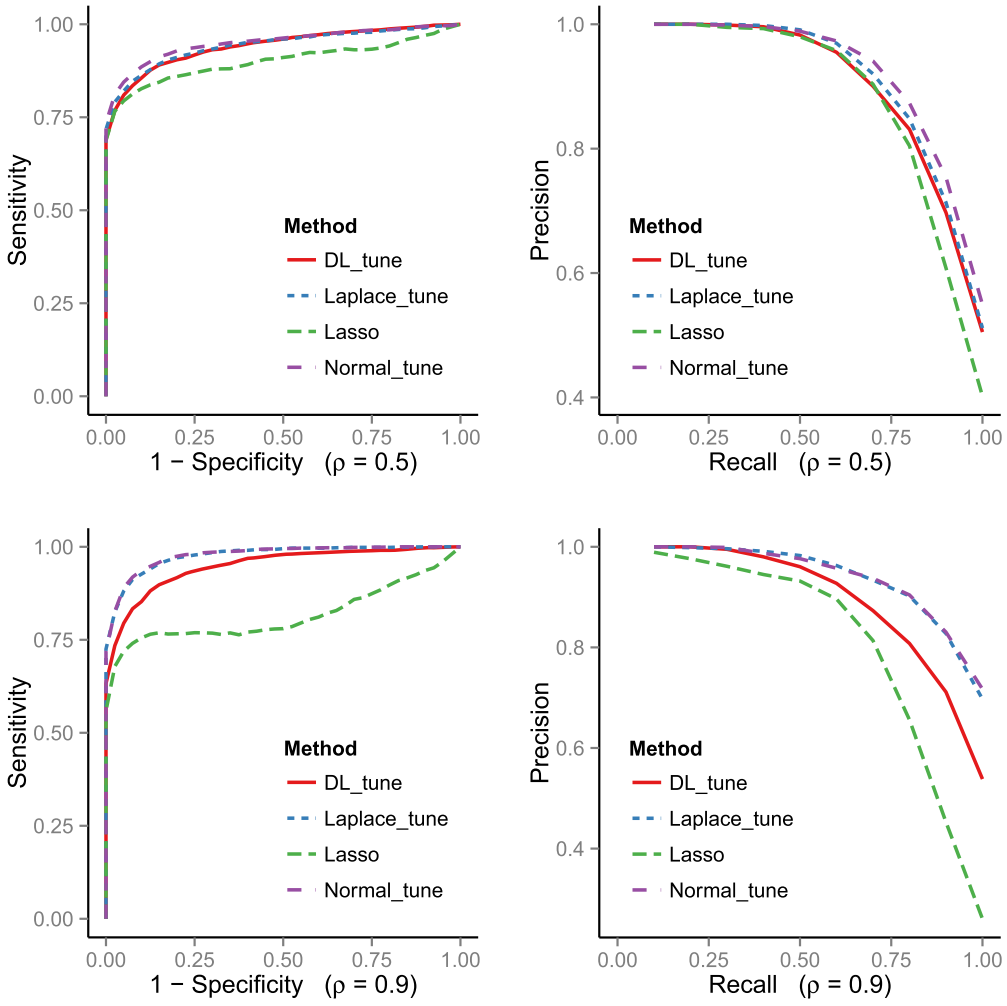

Figure 2: Plot of mean ROC and PRC curves when $\rho = 0.5$ and $\rho = 0.9$, over the 200 datasets for $p = 50$ predictors, $n = 60$ observations. The left column is the ROC curve, the right column is the PRC curve.

On the other hand, in terms of comparing performance of different priors applied on the penalized credible region variable selection, combining both the tables and figures, we have the following findings. When considering the Precision-Recall in particular, the DL and Laplace priors outperform the normal prior and Lasso. This is particularly true if the hyperparameters in them are tuned via a uniform distribution on R^2. We note that when there are only a few true and many unimportant variables, the Precision-Recall curve is a more appropriate measure than the ROC curve. For example, when $p = 1000$, in both $\rho = 0.5$ and 0.9 cases, in Figure 4, the PRC curve shows that the DL prior is significantly better than the normal prior; the ROC curve of the normal prior

	ROC Area		PRC Area	
	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.9$
Lasso	-	-	0.550 (0.0087)	0.550 (0.0089)
Normal_hyper	0.948 (0.0031)	0.990 (0.0013)	0.615 (0.0093)	0.784 (0.0062)
Normal_tune	0.950 (0.0029)	0.992 (0.0007)	0.610 (0.0091)	0.721 (0.0077)
Laplace_hyper	0.937 (0.0030)	0.969 (0.0020)	0.621 (0.0087)	0.680 (0.0087)
Laplace_tune	0.959 (0.0027)	0.995 (0.0004)	0.701 (0.0077)	0.822 (0.0055)
DL_hyper	0.927 (0.0038)	0.908 (0.0047)	0.651 (0.0092)	0.570 (0.0102)
DL_tune	0.949 (0.0027)	0.970 (0.0025)	0.717 (0.0085)	0.797 (0.0073)

Table 2: Mean area under the ROC Curve and the PRC curve for $p = 500$, $n = 60$, based on 200 datasets with standard errors in parentheses.

	ROC Area		PRC Area	
	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.9$
Lasso	-	-	0.507 (0.0093)	0.536 (0.0091)
Normal_hyper	0.942 (0.0039)	0.992 (0.0018)	0.515 (0.0101)	0.727 (0.0076)
Normal_tune	0.943 (0.0039)	0.991 (0.0018)	0.539 (0.0099)	0.680 (0.0083)
Laplace_hyper	0.914 (0.0041)	0.968 (0.0021)	0.444 (0.0093)	0.554 (0.0102)
Laplace_tune	0.951 (0.0038)	0.994 (0.0012)	0.638 (0.0092)	0.764 (0.0071)
DL_hyper	0.931 (0.0040)	0.943 (0.0034)	0.635 (0.0096)	0.623 (0.0094)
DL_tune	0.925 (0.0045)	0.967 (0.0025)	0.633 (0.0116)	0.768 (0.0092)

Table 3: Mean area under the ROC Curve and the PRC curve for $p = 1000$, $n = 60$, based on 200 datasets with standard errors in parentheses.

goes higher when FPR (or 1-Specificity) is large, however, when FPR is small (which is of more interest), DL prior still leads to significantly larger sensitivity than the normal prior. Overall, the DL prior outperforms the normal prior, as does the Laplace prior.

6.2 Additional Simulations on Hyperparameter Tuning

To examine the role of a in the DL prior, additional simulations were conducted. Table 4 gives the average squared error for the posterior mean based on the 200 same datasets as Section 6.1, for the DL priors with a fixed at $1/2$, $1/n$, and $1/p$. The results show that when p is large or there is strong correlation in the dataset, $a = 1/n$ is better than $a = 1/2$. When p is small and there is only moderate correlation for the data, $a = 1/2$ is recommended. Since the performance of different values of a varies relying on the dimension of predictors and the correlation structure of the predictors, fixing a is difficult. Thus either giving a hyperprior for a or using the R^2 method proposed in Section 5 to tune a is suggested.

Furthermore, to verify Theorem 5 described in Section 5, additional calculations were performed. For each of the above 200 datasets, 'Normal_tune" returns a "best" tuned γ through conducting the practical procedures as introduced in Section 5, and we name it as "Tuned". Also, by Theorem 5, the theoretic "best" γ can be derived based on

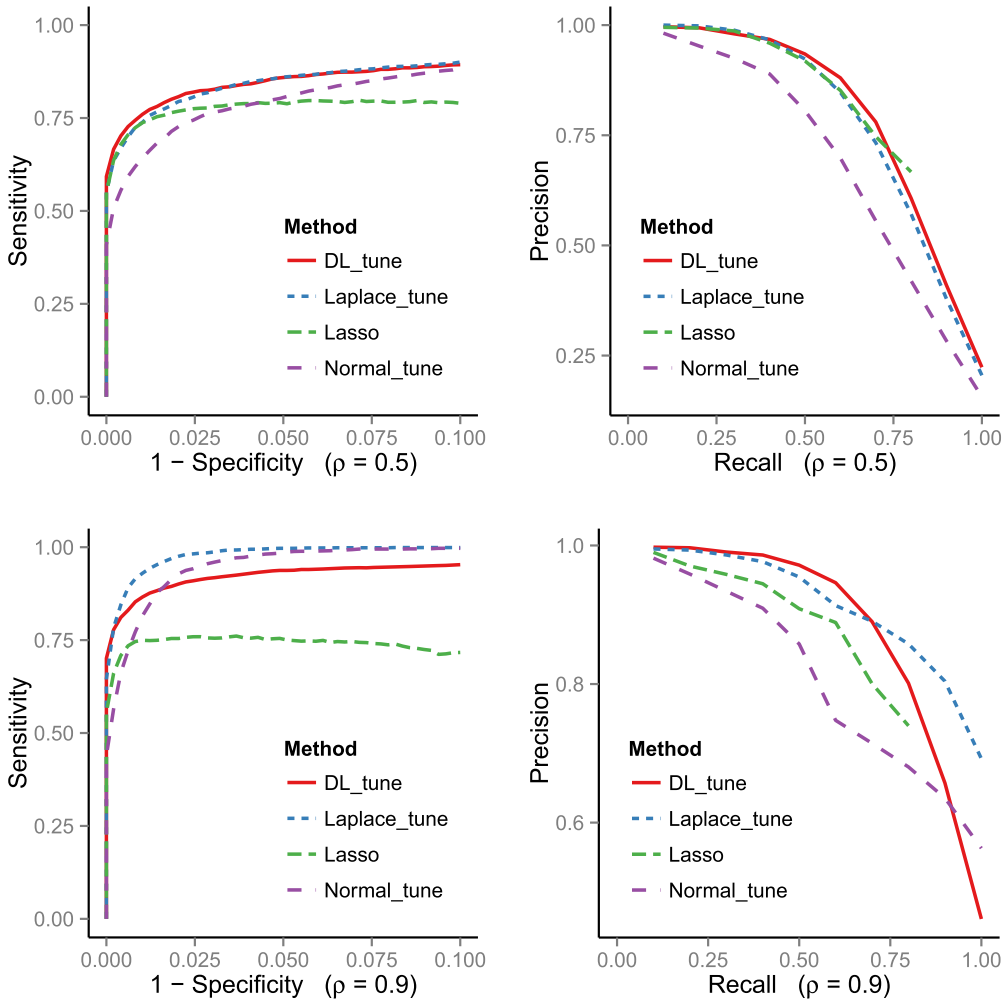

Figure 3: Plot of mean ROC and PRC curves when $\rho = 0.5$ and $\rho = 0.9$, over the 200 datasets for $p = 500$ predictors, $n = 60$ observations. The left column is the ROC curve, the right column is the PRC curve.

the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}/n$ for each dataset, and we name it as "Derived". In addition, for each of the above 200 datasets, the design matrix X is generated from a multivariate normal distribution with specific and fixed covariance structure. So the eigenvalues of such true covariance matrix, instead of $\boldsymbol{X}^T\boldsymbol{X}/n$, can be used to derive the theoretic "best" γ, and we name it as "Theoretic" value. Table 5 gives the "Theoretic" value, and the mean of "Derived" and "Tuned" value together with the standard error among the 200 datasets, for simulation setups $\rho = 0.5$ and 0.9. In general, the three values are similar and all of them are close to the value of p. So in practice, γ can be set as the "Derived" value based on the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}/n$, or for simplicity, $\gamma = p$ can also be used.

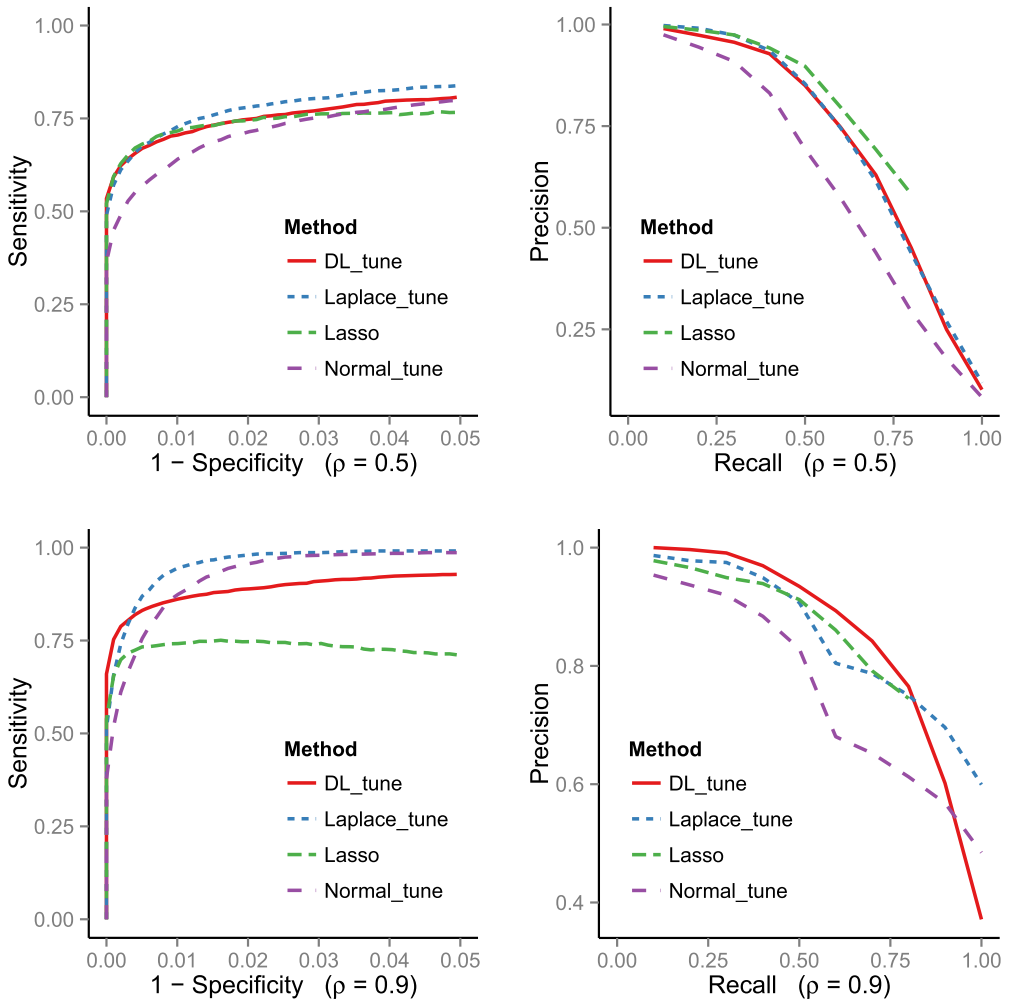

Figure 4: Plot of mean ROC and PRC curves when $\rho = 0.5$ and $\rho = 0.9$, over the 200 datasets for $p = 1000$ predictors, $n = 60$ observations. The left column is the ROC curve, the right column is the PRC curve.

7 Real Data Analysis

We now analyze data on mouse gene expression from the experiment conducted by Lan et al. (2006). There were 60 arrays to monitor the expression levels of 22,575 genes consisting of 31 female and 29 male mice. Quantitative real-time PCR were used to measure some physiological phenotypes, including numbers of phosphoenopyruvate carboxykinase (PEPCK), glycerol-3-phosphate acyltransferase (GPAT), and stearoyl-CoA desaturase 1 (SCD1). The gene expression data and the phenotypic data can be found at GEO (http://www.ncbi.nlm.nih.gov/geo; accession number GSE3330).

	$p = 50$			$p = 500$			$p = 1000$		
a	$\frac{1}{2}$	$\frac{1}{n}$	$\frac{1}{p}$	$\frac{1}{2}$	$\frac{1}{n}$	$\frac{1}{p}$	$\frac{1}{2}$	$\frac{1}{n}$	$\frac{1}{p}$
$\rho = 0.5$	0.772	0.877	0.874	1.292	1.400	1.953	1.470	1.434	2.196
	(0.0234)	(0.0325)	(0.0329)	(0.0421)	(0.0519)	(0.0576)	(0.0451)	(0.1070)	(0.1196)
$\rho = 0.9$	1.989	1.751	1.715	2.193	2.142	2.546	2.299	2.247	2.426
	(0.0559)	(0.0737)	(0.0739)	(0.0767)	(0.0981)	(0.1180)	(0.1101)	(0.1178)	(0.1186)

Table 4: Average squared error for the posterior mean, given Dirichlet–Laplace prior with a fixed at $1/2$, $1/n$ and $1/p$, based on 200 datasets with standard errors in parentheses.

	$\rho = 0.5$			$\rho = 0.9$		
	Theoretic	Derived	Tuned	Theoretic	Derived	Tuned
$p = 50$	47.6	46.6 (0.11)	48.6 (0.24)	40.8	39.9 (0.18)	41.3 (0.25)
$p = 500$	490.0	481.8 (0.35)	474.3 (0.83)	482.4	474.1 (0.81)	471.9 (1.23)
$p = 1000$	981.7	965.1 (0.51)	947.1 (1.50)	974.0	956.9 (1.18)	944.3 (1.82)

Table 5: Theoretic γ in the normal prior (3) based on Theorem 5, together with mean of the derived and tuned γ through methods proposed in Section 5, based on 200 datasets with standard errors in parentheses.

First, by ordering the magnitude of marginal correlation between the genes with the three responses from the largest to the smallest, $22,575$ genes were screened down to the 999 genes, thus reducing the number of candidate predictors of the three linear regressions. Note that the top 999 genes were not the same for the 3 responses. Then for each of the 3 regressions, the dataset is composed of $n = 60$ observations and $p = 1,000$ predictors (gender along with the 999 genes). After the screening, the Lasso estimator and the penalized credible region method applied on the normal, Laplace and DL priors were used. The hyperparameters in those prior distributions are tuned through the R^2 method introduced in Section 5, with a target of uniform distribution.

To evaluate the performance of the proposed approach, the first step was to randomly split the sample size 60 into a training set of size 55 and a testing set of size 5. The stopping rule was BIC. To be more specific, the selected model was the one with smallest BIC among all models in which the number of predictors is less than 30. Then the selected model was used to predict the remaining 5 observations, and the prediction error was then obtained. We repeated this for 100 replicates in order to compare the prediction errors. Table 6 shows the mean squared prediction error (with its standard error) based on the 100 random splits of the data. The mean selected model size (with its standard error) is also included.

Overall, the results show that the proposed penalized credible region selection method using global-local shrinkage priors such as DL prior performs well. For all 3 responses, the penalized credible region approach with DL prior performs better than the Lasso estimator and has a smaller number of predictors. For PEPCK and SCD1, the DL prior has significant better performance than the normal prior and Laplace prior. For GPAT, there is no significant difference between normal and DL prior. In all, for this dataset,

	PEPCK		GPAT		SCD1	
	MSPE	Model Size	MSPE	Model Size	MSPE	Model Size
Lasso	0.54 (0.026)	25.8 (0.34)	1.43 (0.082)	24.4 (0.56)	0.55 (0.052)	26.1 (0.33)
Normal	0.66 (0.033)	16.8 (0.67)	1.30 (0.099)	16.3 (0.66)	0.71 (0.059)	10.8 (0.60)
Laplace	0.70 (0.037)	17.0 (0.78)	1.19 (0.086)	21.4 (0.56)	0.69 (0.054)	14.8 (0.82)
DL	0.49 (0.032)	18.4 (0.73)	1.37 (0.102)	13.1 (0.68)	0.54 (0.037)	14.0 (0.59)

Table 6: Mean squared prediction error and model size, with standard errors in parenthesis, based on 100 random splits of the real data.

the proposed approach generally improves the performance by replacing the normal prior with the DL prior.

8 Discussion

In this paper, we extend the penalized credible variable selection approach by using global-local shrinkage priors. Simulation studies show that the GL shrinkage priors outperform the original normal prior. Our main result also includes modifying the Dirichlet–Laplace prior to accommodate the linear regression model instead of the simple normal mean problem as in Bhattacharya et al. (2015).

In theory, we obtain the selection consistency for the penalized credible region method using the global-local shrinkage priors when $p_n = o(n)$. Posterior consistency for the normal and DL priors are also shown. However, since the proposed method considers the posterior credible region for the full p_n-dimensional parameter, the proposed approach is sub-optimal compared to the results in (Arias-Castro et al., 2014; Castillo et al., 2015; Martin et al., 2017). This comes from the fact that the full design matrix $\mathbf{X}_n^T\mathbf{X}_n/n$ is used instead of submatrices as in the above references. Since exact sparsity is not exploited, some of the conditions for selection consistency are stronger than these approaches.

Furthermore, this paper introduces a new default method to tune the hyperparameters in prior distributions based on the induced prior distribution of R^2. The hyperparameter is chosen to minimize a discrepancy between the induced distribution of R^2 and a default Beta distribution. For the normal prior, a closed form of the hyperparameters is derived. This method is straightforward and efficient as it only involves the prior distributions. A simulation study illustrates that our proposed tuning method improves upon the usual hyperprior method.

Supplementary Material

Supplementary material of "Variable selection via penalized credible regions with Dirichlet–Laplace global-local shrinkage priors" (DOI: 10.1214/17-BA1076SUPP; .pdf).

References

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle." In *Selected Papers of Hirotugu Akaike*, 199–213. Springer. MR1486823. 823

Arias-Castro, E., Lounici, K., et al. (2014). "Estimation and variable selection with exponential weights." *Electronic Journal of Statistics*, 8(1): 328–354. MR3195119. doi: https://doi.org/10.1214/14-EJS883. 841

Armagan, A., Dunson, D. B., and Lee, J. (2013a). "Generalized double Pareto shrinkage." *Statistica Sinica*, 23(1): 119. MR3076161. 824

Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U., and Strawn, N. (2013b). "Posterior consistency in linear models under shrinkage priors." *Biometrika*, 100(4): 1011–1018. MR3142348. doi: https://doi.org/10.1093/biomet/ast028. 829

Bhadra, A., Datta, J., Polson, N. G., Willard, B., et al. (2016). "The horseshoe+ estimator of ultra-sparse signals." *Bayesian Analysis*. 824

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). "Dirichlet–Laplace priors for optimal shrinkage." *Journal of the American Statistical Association*, 110(512): 1479–1490. MR3449048. doi: https://doi.org/10.1080/01621459.2014.960967. 824, 827, 841

Bondell, H. D. and Reich, B. J. (2008). "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR." *Biometrics*, 64(1): 115–123. MR2422825. doi: https://doi.org/10.1111/j.1541-0420.2007.00843.x. 824

Bondell, H. D. and Reich, B. J. (2012). "Consistent high-dimensional Bayesian variable selection via penalized credible regions." *Journal of the American Statistical Association*, 107(500): 1610–1624. MR3036420. doi: https://doi.org/10.1080/01621459.2012.716344. 824, 825, 826, 828, 830, 834

Candes, E. and Tao, T. (2007). "The Dantzig selector: Statistical estimation when p is much larger than n." *The Annals of Statistics*, 2313–2351. MR2382644. doi: https://doi.org/10.1214/009053606000001523. 824

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). "Handling sparsity via the horseshoe." In *International Conference on Artificial Intelligence and Statistics*, 73–80. 824

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The Horseshoe estimator for sparse signals." *Biometrika*, asq017. 824

Castillo, I., Schmidt-Hieber, J., Van der Vaart, A., et al. (2015). "Bayesian linear regression with sparse priors." *The Annals of Statistics*, 43(5): 1986–2018. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 841

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). "Least angle regression." *The Annals of Statistics*, 32(2): 407–499. MR2060166. doi: https://doi.org/10.1214/009053604000000067. 825, 835

Fan, J. and Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American Statistical Association*, 96(456): 1348–1360. MR1946581. doi: https://doi.org/10.1198/016214501753382273. 823

George, E. and Foster, D. P. (2000). "Calibration and empirical Bayes variable selection." *Biometrika*, 87(4): 731–747. MR1813972. doi: https://doi.org/10.1093/biomet/87.4.731. 824

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 824

Griffin, J. E., Brown, P. J., et al. (2010). "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis*, 5(1): 171–188. MR2596440. doi: https://doi.org/10.1214/10-BA507. 824

Hans, C. (2010). "Model uncertainty and variable selection in Bayesian lasso regression." *Statistics and Computing*, 20(2): 221–229. MR2610774. doi: https://doi.org/10.1007/s11222-009-9160-9. 826

Ishwaran, H. and Rao, J. S. (2005). "Spike and slab variable selection: frequentist and Bayesian strategies." *Annals of Statistics*, 730–773. MR2163158. doi: https://doi.org/10.1214/009053604000001147. 824

Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E., Flowers, M. T., Schueler, K. L., Manly, K. F., et al. (2006). "Combined expression trait correlations and expression quantitative trait locus mapping." *PLoS Genet*, 2(1): e6. MR2709393. 839

Leng, C., Tran, M.-N., and Nott, D. (2014). "Bayesian adaptive lasso." *Annals of the Institute of Statistical Mathematics*, 66(2): 221–244. MR3171404. doi: https://doi.org/10.1007/s10463-013-0429-6. 824

Li, Q., Lin, N., et al. (2010). "The Bayesian elastic net." *Bayesian Analysis*, 5(1): 151–170. MR2596439. doi: https://doi.org/10.1214/10-BA506. 824

Lv, J. and Fan, Y. (2009). "A unified approach to model selection and sparse recovery using regularized least squares." *The Annals of Statistics*, 3498–3528. MR2549567. doi: https://doi.org/10.1214/09-AOS683. 825

Martin, R., Mess, R., Walker, S. G., et al. (2017). "Empirical Bayes posterior concentration in sparse high-dimensional linear models." *Bernoulli*, 23(3): 1822–1847. MR3624879. doi: https://doi.org/10.3150/15-BEJ797. 841

Park, T. and Casella, G. (2008). "The bayesian lasso." *Journal of the American Statistical Association*, 103(482): 681–686. MR2524001. doi: https://doi.org/10.1198/016214508000000337. 824, 826

Polson, N. G. and Scott, J. G. (2010). "Shrink globally, act locally: Sparse Bayesian regularization and prediction." *Bayesian Statistics*, 9: 501–538. MR3204017. doi: https://doi.org/10.1093/acprof:oso/9780199694587.003.0017. 824

Polson, N. G., Scott, J. G., and Windle, J. (2013). "The Bayesian Bridge." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. MR3248673. doi: https://doi.org/10.1111/rssb.12042. 824

Schwarz, G. et al. (1978). "Estimating the dimension of a model." *The annals of statistics*, 6(2): 461–464. MR0468014. 823

Stroeker, R. (1983). "Approximations of the eigenvalues of the covariance matrix of a first-order autoregressive process." *Journal of Econometrics*, 22(3): 269–279. MR0714099. doi: https://doi.org/10.1016/0304-4076(83)90103-3. 829

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288. MR1379242. 823, 835

Zhang, Y. and Bondell, H. D. (2017). "Supplementary material of "Variable selection via penalized credible regions with Dirichlet–Laplace global-local shrinkage priors"." *Bayesian Analysis*. doi: https://doi.org/10.1214/17-BA1076SUPP. 825

Zou, H. (2006). "The adaptive lasso and its oracle properties." *Journal of the American Statistical Association*, 101(476): 1418–1429. MR2279469. doi: https://doi.org/10.1198/016214506000000735. 824

Zou, H. and Hastie, T. (2005). "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320. MR2137327. doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x. 824