Statistica Sinica **31** (2021), 2051-2072
doi:https://doi.org/10.5705/ss.202019.0133

BAYESIAN INFERENCE IN HIGH-DIMENSIONAL LINEAR MODELS USING AN EMPIRICAL CORRELATION-ADAPTIVE PRIOR

Chang Liu[1], Yue Yang[1], Howard Bondell[2], and Ryan Martin[1]

[1]*North Carolina State University and* [2]*University of Melbourne*

Abstract: In the context of a high-dimensional linear regression model, we propose an empirical correlation-adaptive prior that uses information in the observed predictor variable matrix to adaptively address high collinearity. We use this prior to determine whether the parameters associated with the correlated predictors should be shrunk together or kept apart. Under certain conditions, we prove that our empirical Bayes posterior concentrates at the optimal rate. Therefore the benefits of correlation-adaptation in finite samples can be achieved without sacrificing asymptotic optimality. A version of the shotgun stochastic search algorithm is employed to compute the posterior and facilitate variable selection. Finally we demonstrate our method's favorable performance compared with that of existing methods using real and simulated data examples, even in ultrahigh-dimensional settings.

Key words and phrases: Collinearity, empirical Bayes, posterior convergence rate, stochastic search, variable selection.

1. Introduction

Consider the standard linear regression model

$$Y = X\beta + \varepsilon,$$

where Y is an $n \times 1$ vector of response variables, X is an $n \times p$ matrix of predictor variables, β is a $p \times 1$ vector of regression coefficients, and ε is a vector of independent and identically distributed (i.i.d.) $\mathsf{N}(0, \sigma^2)$ errors. We are interested in the high-dimensional case where $p \gg n$. Furthermore, we assume that the true β is sparse in the sense that only a small subset of the β coefficients are nonzero.

There are a variety of methods available for estimating β under a sparsity constraint. These include regularization-based methods such as the Lasso (Tibshirani (1996)), adaptive Lasso (Zou (2006)), smoothly clipped absolute deviations(SCAD) penalty (Fan and Li (2001)), and minimax concave penalty(MCP)

Corresponding author: Howard Bondell, School of Mathematics and Statistics, University of Melbourne, Melbourne, Australia. E-mail: howard.bondell@unimelb.edu.au.

(Zhang (2010)); see Fan and Lv (2010) for a review. From a Bayesian point of view, a variety of priors for regression coefficients and the model space have been developed, leading to promising selection properties. For the regression coefficients, β, the normal mixture prior is specified in George and McCullogh (1993); George and Foster (2000) introduce empirical Bayes ideas; Ishwaran and Rao (2005) use spike-and-slab priors; Bondell and Reich (2012) estimate β as the "most sparse" among those in a suitable posterior credible region; Polson and Scott (2012) consider a horseshoe prior; Narisetty and He (2014) use shrinking and diffusing priors; and Martin, Mess and Walker (2017) consider an empirical Bayes version of the spike-and-slab.

Collinearity is unavoidable in high-dimensional settings. Methods such as the Lasso tend to smooth away the regression coefficients of highly collinear predictors and, hence, deter correlated covariates from being included in the model simultaneously. This motivated Krishna, Bondell and Ghosh (2009) to propose an adaptive-powered correlation prior that lets the data itself decide how the collinear predictors are to be handled. However, their suggested generalized Zellner's prior is not applicable in the $p > n$ scenario. To overcome this, we adopt an empirical Bayes approach based on an *empirical correlation-adaptive prior* (ECAP) that uses the data to decide how to shrink the coefficients associated with the correlated predictors. In Section 2, we present our empirical Bayes model and a motivating example illustrating the effect of the correlation-adaptation in the prior. Asymptotic posterior concentration properties are derived in Section 3. In particular, the minimax optimal concentration rates are established for the mean response, showing that the finite-sample benefits of correlation-adaptation lead to no loss of asymptotic optimality. In Section 4, we recommend a shotgun stochastic search approach to compute the posterior distribution over the model space. Simulation experiments are presented in Section 5. Here we demonstrate the benefits for variable selection of adaptively varying the correlation structure in the prior, as compared with existing methods. The real-data illustration in Section 6 highlights the improved predictive performance, even in ultrahigh-dimensional settings, of the proposed correlation-adaptive prior. All proofs are deferred to the Supplementary Material.

2. Model Specification

2.1. The prior

Under assumed sparsity, it is natural to decompose β as (S, β_S), where $S \subseteq \{1, 2, \ldots, p\}$ is the set of nonzero coefficients, called the *configuration* of β, and

β_S is the $|S|$-vector of nonzero values, with $|S|$ denoting the cardinality of S. We write X_S for the sub-matrix of X corresponding to the configuration S. With this decomposition of β, a hierarchical prior is convenient, that is, a prior for S and a conditional prior for β_S, given S.

First, for the prior $\pi(S)$ for S, we follow Martin, Mess and Walker (2017) and write

$$\pi(S) = \pi(S \mid |S| = s) f_n(s),$$

where $f_n(s)$ is a prior on $|S|$ and $\pi(S \mid |S| = s)$ is a conditional prior on S, given $|S|$. Based on the recommendation in Castillo, Schmidt-Hieber and van der Vaart (2015), we take

$$f_n(s) \propto c^{-s} p^{-as}, \quad s = 0, 1, \ldots, R, \tag{2.1}$$

where a and c are positive constants, and $R = \operatorname{rank}(X) \leq n$. It is common to take $\pi(S \mid |S| = s)$ to be uniform, but here we break from this trend to take collinearity into account. Let $D(S) = |X_S^\top X_S|$ denote the determinant of $X_S^\top X_S$, and consider the geometric mean of the eigenvalues, $D(S)^{1/|S|}$, as a measure of the "degree of collinearity" in model S. We set

$$\pi_\lambda(S \mid |S| = s) = \frac{D(S)^{-\lambda/(2s)} 1\{\kappa(S) < Cp^r\}}{\sum_{S:|S|=s} D(S)^{-\lambda/(2s)} 1\{\kappa(S) < Cp^r\}}, \quad \lambda \in \mathbb{R}, \tag{2.2}$$

where $\kappa(S)$ is the condition number of $X_S^\top X_S$, and r and C are positive constants, specified to exclude models with extremely ill-conditioned $X_S^\top X_S$. The constant λ is an important feature of the proposed model, and is discussed in more detail below. Because of the dependence on λ above, we henceforth write $\pi_\lambda(S)$ for the prior of S.

In these high-dimensional problems, the properties of the posterior distribution are highly sensitive to the choice of prior. For example, Castillo and van der Vaart (2012) show that, with thin-tailed Gaussian priors on the coefficients, the posterior distribution might concentrate at a sub-optimal rate. As such, they recommend using priors with heavier-than-Gaussian tails. However, these heavy-tailed priors lack the desirable conjugacy properties and, therefore, their use adds to the already substantial computational burden. This creates a dilemma: do we use a theoretically justified heavy-tailed prior that makes the computation more difficult, or do we use a computationally convenient thin-tailed prior with potentially sub-optimal posterior convergence properties? Martin, Mess and Walker (2017) observe that the prior tails are less relevant if the center is chosen appropriately. Therefore, to overcome the aforementioned dilemma, they propose using an *empirical prior* with a data-driven centering. Following their general idea, as the prior for β_S, given S, we take

$$(\beta_S \mid S, \lambda) \sim \mathsf{N}\big(\phi\hat{\beta}_S, \sigma^2 g k_S (X_S^\top X_S)^\lambda\big). \tag{2.3}$$

Here, $\hat{\beta}_S$ is the least squares estimator corresponding to configuration S and design matrix X_S, $\phi \in (0,1)$ is a shrinkage factor to be specified, g is a parameter controlling the prior spread, $(X_S^\top X_S)^\lambda$ is an adaptive powered correlation matrix, and

$$k_S = \frac{\operatorname{tr}\{(X_S^\top X_S)^{-1}\}}{\operatorname{tr}\{(X_S^\top X_S)^\lambda\}}$$

is a standardizing factor, as in Krishna, Bondell and Ghosh (2009), designed to control for the scale corresponding to different values of λ. Let $\pi_\lambda(\beta_S \mid S)$ denote this prior density for β_S, given S.

The power parameter λ on the prior covariance matrix can encourage or discourage the inclusion of correlated predictors. When $\lambda > 0$, the prior shrinks the coefficients of the correlated predictors toward each other; when $\lambda < 0$, they tend to be kept apart, with $\lambda = -1$ being the most familiar; and, finally, $\lambda = 0$ implies prior independence. Therefore, a positive λ would prefer larger models by capturing as many correlated predictors as possible, while a negative λ tends to select models with less collinearity; see Krishna, Bondell and Ghosh (2009) for a discussion of this phenomenon. Our data-driven choice of λ, along with that of the other tuning parameters introduced here and in the next subsection, is discussed in Section 4.2.

2.2. The posterior distribution

For this standard linear regression model, the likelihood function is

$$L_n(\beta) = (2\pi\sigma^2)^{-n/2}\, e^{-\|Y - X\beta\|^2/2\sigma^2}, \quad \beta \in \mathbb{R}^p.$$

It is straightforward to include σ^2 as an argument in this likelihood function, introduce a prior for σ^2, and obtain a full (β, σ^2) posterior; see Martin and Tang (2019). However, our intention is to use a plug-in estimator for σ^2 in what follows. Hence, we omit the error variance as an argument to the likelihood function.

Given a prior and the likelihood, we can combine the two using Bayes' formula to obtain a posterior distribution for (S, β_S) or, equivalently, for the p-vector β. However, the fact that our prior also depends on the data changes the way we think about the posterior construction. Specifically, updating the data-dependent prior using the full likelihood amounts to a double-use of the data, and hence a risk of over-fitting. To avoid this risk, some regularization is needed. While there are a number of ways to achieve this regularization (Martin and Walker (2019)), arguably the simplest is to apply Bayes' formula, but using only a (large) portion of the likelihood. As in the generalized Bayes literature (e.g., Martin and Walker (2014); Grünwald and van Ommen (2017); Syring and Martin (2019)), we use a power likelihood and define our posterior for (S, β_S) as

$$\pi_\lambda^n(S, \beta_S) \propto L_n(\beta_{S+})^\alpha \, \pi_\lambda(\beta_S \mid S) \, \pi_\lambda(S),$$

where β_{S+} is the p-vector obtained by entering zeros around β_S in the entries corresponding to S^c, and $\alpha \in (0,1)$ is a regularization factor, which can be taken arbitrarily close to one. It may be possible to handle the case $\alpha = 1$, making appropriate adjustments elsewhere. However, the proposed approach achieves the optimal posterior concentration rate (see Section 3), and hence will not be improved.

To summarize, the posterior distribution for β, denoted by Π_λ^n, is obtained by summing over all configurations S; that is,

$$\Pi_\lambda^n(A) \propto \sum_S \int_{\{\beta_S : \beta_{S+} \in A\}} \pi_\lambda^n(S, \beta_S) \, d\beta_S, \quad A \subseteq \mathbb{R}^p.$$

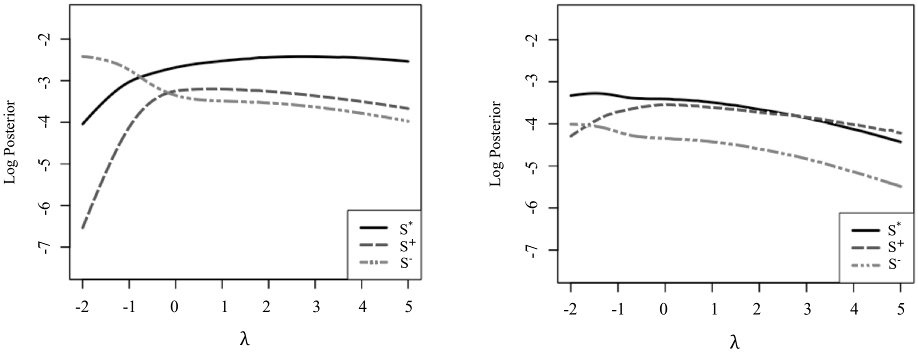
Because one of our primary objectives is variable selection, it is of interest that we can obtain a closed-form expression for the posterior distribution of S, up to a normalizing constant, a result of our use of a conjugate normal prior for β_S, given S. That is, we can integrate out β_S to obtain a marginal likelihood for Y; that is,

$$\begin{aligned} m_\lambda(Y \mid S) = (2\pi\sigma^2)^{-n\alpha/2} & \prod_{i=1}^{s} \bigl(1 + \alpha g k_S d_{S,i}^{\lambda+1}\bigr)^{-1/2} \\ & \times \exp\left[-\frac{\alpha}{2\sigma^2} \left\{ \|y - \hat{y}_S\|^2 + (1-\phi)^2 \sum_{i=1}^{s} \frac{d_{S,i}}{1 + \alpha g k_S d_{S,i}^{\lambda+1}} \theta_{S,i}^2 \right\} \right], \end{aligned} \tag{2.4}$$

where $\hat{y}_S$ is the least square estimate of y, given model S, $d_{S,i}$ is the ith eigenvalue of $X_S^\top X_S$, $\Gamma_S \Lambda_S \Gamma_S^\top$ is the spectral decomposition of $X_S^\top X_S$, with $\Lambda_S = \operatorname{diag}(d_{S,1}, \ldots, d_{S,s})$, and $\theta_{S,i}$ is the ith element of $\theta_S = \Gamma_S^\top \hat{\beta}_S$. Then, it is straightforward to obtain the posterior distribution for S, as follows:

$$\pi_\lambda^n(S) \propto m_\lambda(Y \mid S) \, \pi_\lambda(S). \tag{2.5}$$

The variable selection method described in Section 4 and illustrated in Sections 5–6 is based on this posterior distribution.



(a) $\rho = 0.8$

(b) $\rho = 0.1$

Figure 1. Plot of $\lambda \mapsto \log \pi_\lambda^n(S)$ for three different S and two different ρ.

2.3. A motivating example

We now give a simple example to illustrate the effects of incorporating λ into (2.2) and (2.3). Consider a case with $n = p = 5$, and let $X = X_{n\times p}$ have i.i.d. rows, each with a standard multivariate normal with first-order autoregressive dependence and correlation parameter ρ. Given X, the conditional distribution of the response is determined by the linear model

$$y_i = x_{i1} + 0.8x_{i2} + \varepsilon_i, \quad \text{where} \quad \varepsilon_1, \ldots, \varepsilon_5 \overset{i.i.d.}{\sim} \mathsf{N}(0,1).$$

The black, blue, and red curves in Figure 1 represent $\lambda \mapsto \log \pi_\lambda^n(S)$, for three different S configurations, namely, the true configuration $S^\star = \{1,2\}$, $S^- = \{1\}$, and $S^+ = \{1,2,3\}$. Panel (a) corresponds to a high correlation case, $\rho = 0.8$, and we see that the ECAP-based posterior prefers $S^\star$ for suitably large λ. Compare this to the choice $\lambda \equiv -1$ in Martin, Mess and Walker (2017), which prefers the smaller configuration S^-. On the other hand, when the correlation is relatively low, as in Panel (b), we see that a large positive λ encourages a larger configuration, while the true configuration is preferred for sufficiently large negative values of λ. The take-away message is that, by allowing λ to vary, the ECAP-based model has the ability to adjust to the correlation structure, which can be beneficial in identifying the relevant variables.

3. Posterior Convergence Properties

3.1. Setup and assumptions

We stick with the standard notation given previously; however, keep in mind that $Y^n = (Y_1^n, Y_2^n, \ldots, Y_n^n)$ and $X^n = ((X_{ij}^n))$ are better understood as triangular arrays. Therefore, we can have p, $s^\star = |S^\star|$, with $S^\star$ denoting the true configuration, and R all depend on n. We assume throughout that $s^\star \leq R \leq n \ll p$; more precise conditions are given below. We also assume that λ, g, and σ^2 are fixed constants in this setting, not parameters to be estimated/tuned. Therefore, to simplify the notation here and in the proofs, we drop the subscript λ, and simply write Π^n for the posterior for β, instead of Π_λ^n.

When estimating the mean response, the minimax rate does not depend on the correlation structure in X, so we cannot expect any improvements in the rate by incorporating this correlation structure in our prior distribution. Therefore, our goal here is simply to show that the minimax rates can still be achieved, while leaving room to adjust for collinearity in finite samples. The finite-sample benefits of the correlation-adaptive prior are shown in the numerical results presented in Section 5.

We start by stating the basic assumptions for all the results that follow, beginning with two assumptions about the asymptotic regime. In particular, relative to n, the true configuration, $S^\star$, is not too complex.

Assumption 1. *The true complexity satisfies $s^\star \to \infty$, with $s^\star = o(n)$.*

The next assumption puts a very mild size condition on $\beta_{S^\star}^\star$, that is, the nonzero regression coefficients of the true $\beta^\star$, and on the user-specified shrinkage factor $\phi = \phi_n$ in the prior.

Assumption 2. *The factor $\phi = \phi_n \in (0,1)$ satisfies $n(1-\phi_n)^2 \|\beta_{S^\star}^\star\|^2 = o(s^\star)$.*

Assumption 2 includes a very mild condition on the true $\beta^\star$, that is, that the "total signal" $\|\beta_{S^\star}^\star\|$ is not too small. There is, of course, no reason to think that the individual signals would be vanishing with n. If they do not, then we get $s^\star\{n\|\beta_{S^\star}^\star\|\}^{-1} \to 0$ automatically from Assumption 1. However, it is not required that *all* of the signals are bounded away from zero; the condition is related to the total signal; thus, it is enough that at least one of the signals is away from zero. Even if we require that all nonzero signals be lower-bounded, the condition above holds if $\min_{j \in S^\star} |\beta_j^\star| > n^{-1/2}$. In addition, an even stronger *beta-min* condition—see (3.3) in Section 3.5—is needed to establish variable selection consistency, both here and throughout the literature on high-dimensional inference (e.g., Bühlmann and van de Geer (2011); Arias-Castro and Lounici (2014)).

This also provides some insight into the connection between ϕ and the total signal; that is, ϕ controls the influence of the prior centering. When the total signal is large, this influence is more important than when the total signal is small. In Section 4.2.3, we present a data-driven choice of ϕ that adapts to the total signal size.

Finally, we need to make some assumptions on the $n \times p$ design matrix X. For a given configuration S, let $\lambda_{\min}(S)$ and $\lambda_{\max}(S)$ denote the smallest and the largest eigenvalues of $n^{-1}X_S^\top X_S$, respectively. Next, define

$$\ell(s) = \min_{S:|S|=s} \lambda_{\min}(S) \quad \text{and} \quad u(s) = \max_{S:|S|=s} \lambda_{\max}(S).$$

Recall that these depend (implicitly) on n because of the triangular array formulation. It is also clear that $\ell(s)$ and $u(s)$ are nonincreasing and nondecreasing functions, respectively, of the complexity s. If $\kappa(S) = \lambda_{\max}(S)/\lambda_{\min}(S)$ is the condition number of $n^{-1}X_S^\top X_S$, then we can define

$$\omega(s) = \max_{S:|S|=s} \kappa(S),$$

and obtain the relation $\omega(s) \leq u(s)/\ell(s)$.

Assumption 3. $0 < \liminf_n \ell(s^\star) < \limsup_n u(s^\star) < \infty$.

This assumption roughly states, that every submatrix X_S, for $|S| \leq s^\star$, is of full rank. This is implied by, for example, the sparse Riesz condition of order $s^\star$ in Zhang and Huang (2008).

3.2. Rates under prediction error loss

Ideally, we expect the posterior for β to concentrate asymptotically around values of β such that $\|X\beta - X\beta^\star\|$ is relatively small. The following theorem states this result precisely. Recall the definitions of the prior and, in particular, the quantities a and r.

Theorem 1. *Under Assumptions* 1–3*, there exists a constant M such that*

$$\sup \mathsf{E}_{\beta^\star}\{\Pi^n(\beta \in \mathbb{R}^p : \|X\beta - X\beta^\star\|^2 > M\varepsilon_n)\} \to 0, \quad n \to \infty,$$

where the supremum is over all $\beta^\star$, such that $|S_{\beta^\star}| = s^\star$,

$$\varepsilon_n = \max\left\{q(R, \lambda, r, a), s^\star \log\left(\frac{p}{s^\star}\right)\right\},$$

and

$$q(R, \lambda, r, a) = \begin{cases} R\{r(1+\lambda) - a\} \log p & \textit{if } \lambda \in [0, \infty); \\ R(r - a) \log p & \textit{if } \lambda \in [-1, 0); \\ R(-r\lambda - a) \log p & \textit{if } \lambda \in (-\infty, -1). \end{cases}$$

Proof. See Section S2.1 in the Supplementary Material.

In the so-called ordinary high-dimensional regime (e.g., Rigollet and Tsybakov (2012)), $s^\star \log(p/s^\star)$ is the minimax concentration rate. Thus, our proposed ECAP posterior attains the minimax optimal rate, as long as (a, r) in (2.1) and (2.2) are chosen such that $a > r \max\{1+\lambda, 1, -\lambda\}$.

3.3. Effective posterior dimension

Theorem 1 suggests that the posterior for β concentrates near the true $\beta^\star$, in a certain sense. However, because $\beta^\star$ is sparse, we might ask whether the posterior is also concentrated on a roughly $s^\star$-dimensional subset of $\mathbb{R}^p$. The following theorem gives an affirmative answer to this question. Aside from the economical benefits of having an effectively low-dimensional posterior, Theorem 2 aids in the proofs of the remaining results.

Theorem 2. *Suppose that the prior $\pi(S)$ has parameters (a, r) that satisfy the condition $a > r \max\{1+\lambda, 1, -\lambda\}$, and define*

$$\rho_0 = \frac{a+1}{a - r\max\{1+\lambda, 1, -\lambda\}} > 1. \tag{3.1}$$

Then, under Assumptions 1–3, for any $\rho > \rho_0$, we have

$$\sup \mathsf{E}_{\beta^\star}\{\Pi^n(\beta \in \mathbb{R}^p : |S_\beta| \geq \rho s^\star)\} \to 0, \quad \text{as } n \to \infty,$$

where the supremum is over all $s^\star$-sparse $\beta^\star$.

Proof. See Section S2.2 in the Supplementary Material.

3.4. Rates under the estimation error loss

Following on from the result in Section 3.2 on the posterior concentration with respect to the mean response difference, we might ask whether the concentration holds similarly with respect to a metric relevant to the estimation of β, namely, $\|\beta - \beta^\star\|$. The following theorem establishes this rate, which turns out to be optimal as well.

Theorem 3. *Suppose that the prior $\pi(S)$ has parameters (a, r) that satisfy the condition $a > r \max\{1+\lambda, 1, -\lambda\}$, and let ρ be greater than ρ_0 in (3.1). Under Assumptions 1–3, there exists a constant $M > 0$ such that*

$$\sup \mathsf{E}_{\beta^\star}\{\Pi^n(\beta \in \mathbb{R}^p : \|\beta - \beta^\star\|^2 > M\delta_n)\} \to 0, \quad \text{as } n \to \infty,$$

where the supremum is over all $s^\star$-sparse $\beta^\star$ and

$$\delta_n = \frac{s^\star \log(p/s^\star)}{n\ell(\rho s^\star + s^\star)}. \tag{3.2}$$

Proof. See Section S2.3 in the Supplementary Material.

Under Assumptions 1 and 3, $\ell(\rho s^\star + s^\star)$ is bounded with probability one. Hence, our rate, $n^{-1}s^\star \log(p/s^\star)$, is optimal in the so-called ordinary high-dimensional regime considered by Rigollet and Tsybakov (2012), where $s^\star \log(p/s^\star) < R$, with R the rank of X.

3.5. Variable selection consistency

One of our primary objectives in introducing the λ-dependent prior distribution to account for the collinearity structure in the design matrix is to achieve a more effective variable selection. Thus, it is imperative that we can show, at least asymptotically, that our posterior distribution concentrates around the correct configuration $S^\star$. The following theorem establishes this variable selection consistency property.

Theorem 4. *In addition to Assumptions* 1–3*, assume that the constant a in the prior $\pi(S)$ is such that $a > 1$ and $p^a \gg s^\star e^{Gs^\star}$, where $G = (1-\alpha)\log 2 + m$ and*

$$m = \frac{1}{2}\log\{1 + \alpha g \kappa(S^\star)^{\max\{\lambda+1,1,-\lambda\}}\} = O(1).$$

Then,

$$\sup \mathsf{E}_{\beta^\star}\{\Pi^n(\beta \in \mathbb{R}^p : S_\beta \supset S_{\beta^\star})\} \to 0, \quad n \to \infty,$$

where the supremum is over all $\beta^\star$ that are $s^\star$-sparse. Furthermore, if

$$\min_{j\in S^\star} |\beta_j^\star| \geq \varrho_n := \left\{\frac{2M\sigma^2}{n\ell(s^\star)\alpha(1-\alpha)}\log p\right\}^{1/2}, \tag{3.3}$$

where $M > a+1$ and $p^{M-(a+1)} \gg e^{Gs^\star}$, then

$$\mathsf{E}_{\beta^\star}\{\Pi^n(\beta \in \mathbb{R}^p : S_\beta \not\supseteq S_{\beta^\star})\} \to 0, \quad n \to \infty.$$

If both sets of conditions hold, then variable selection consistency holds; that is,

$$\mathsf{E}_{\beta^\star}\left[\Pi^n(\beta \in \mathbb{R}^p : S_\beta = S_{\beta^\star})\right\} \to 1, \quad n \to \infty.$$

Proof. See Section S2.4 in the Supplementary Material.

The extra conditions on $(p, s^\star)$ in Theorem 4 effectively require that the true configuration size, $s^\star$, is small relative to $\log p$ and, furthermore, that the constant a in (2.1) is large enough that $f_n(s)$ concentrates around comparatively small configurations. In addition, the nonzero $\beta^\star$ values are more difficult to detect if their magnitudes are small. This is intuitively clear, and shows up in our simulation results for Cases 1–2 in Section 5. Theorem 4 gives a mathematical explanation for this intuition, stating that the variable selection based on our empirical Bayes posterior is correct asymptotically if condition (3.3) is satisfied.

4. Implementation Details

4.1. Stochastic search of the configuration space

In order to compute the posterior probability for a configuration S, we need to evaluate $\pi_\lambda(S \mid |S| = s)$ in (2.2), which can be rewritten as

$$\frac{D(S)^{-\lambda/(2s)}1\{\kappa(S) < Cp^r\}}{\binom{p}{s}}\left\{\binom{p}{s}^{-1}\sum_{S:|S|=s} D(S)^{-\lambda/(2s)}1\{\kappa(S) < Cp^r\}\right\}^{-1}.$$

The difficulty comes from the term in curly braces, namely,

$$\binom{p}{s}^{-1}\sum_{S:|S|=s} D(S)^{-\lambda/2s}1\{\kappa(S) < Cp^r\},$$

where, again, $D(S) = |X_S^\top X_S|$ is the determinant. Here C and r can be chosen sufficiently large that only the few extremely ill-conditioned cases are excluded. This leaves approximately $\binom{p}{s}$ terms in the above summation, making brute-force computation a challenge. Given that the eigenvalues of $X_S^\top X_S$, for S with $|S| \approx s^\star$, are assumed to be bounded from above and below, the geometric mean, $D(S)^{1/s}$, of those eigenvalues should depend on the particular X_S, but not on s. Therefore, the quantity in the above expression, the average of these geometric means, is roughly constant in both S and s. As such, it is not unreasonable to approximate $\pi_\lambda(S \mid |S| = s)$ in (2.2) with

$$\frac{D(S)^{-\lambda/(2s)}1\{\kappa(S) < Cp^r\}}{\binom{p}{s}}.$$

This approximation is exact in the case of $\lambda = 0$ if all S are included, and our numerical experiments suggest that it is stable across a range of p, s, and λ. Using this approximation, the posterior distribution for S we use is given by

$$\pi_\lambda^n(S) \propto m_\lambda(Y \mid S)\, D(S)^{-\lambda/2|S|}\binom{p}{|S|}^{-1} f_n(|S|)1\{\kappa(S) < Cp^r\}. \tag{4.1}$$

In practice, C is chosen to be large enough that no configurations, S, are excluded, which effectively removes the indicator function.

Markov chain Monte Carlo (MCMC) methods can be used to compute this posterior, but this tends to be inefficient in high-dimensional problems. As an alternative, we employ a version of the shotgun stochastic search algorithm (SSS, Hans, Dobra and West (2007)) to explore our posterior distribution. In contrast to the traditional MCMC method, the SSS does not attempt to approximate the posterior distribution of S; instead, it only tries to explore high posterior probability regions as thoroughly as possible.

Our SSS algorithm is summarized as follows. Let S be a configuration of size s, with $\pi_\lambda^n(S)$ its corresponding (unnormalized) posterior. Define the neighborhood of S as $\text{nbd}(S) = \{S^+, S^0, S^-\}$, where S^+ is the set containing all $(s+1)$-dimensional configurations that include S, S^0 is the set containing all s-dimensional configurations that have only one variable different from those in S, and S^- is the set containing all $(s-1)$-dimensional configurations nested in S. The tth iteration of the SSS goes as follows:

1. Given S^t, compute $\pi_\lambda^n(S)$, for all $S \in \text{nbd}(S^t) = \{S^{t+}, S^{t_0}, S^{t-}\}$.
2. Sample S_1^t, S_2^t and S_3^t respectively from S^{t+}, S^{t_0} and S^{t-}, with probabilities $\propto \pi_\lambda^n(S_\cdot^t)$.
3. Sample S^{t+1} from $\{S_1^t, S_2^t, S_3^t\}$, with probabilities proportional to $\pi_\lambda^n(S^{t+})$, $\pi_\lambda^n(S^{t_0})$, and $\pi_\lambda^n(S^{t-})$, obtained by summing.

All visited configurations are recorded. The final chosen configuration can be the maximum *a posteriori* model, median probability model (the model that includes those variables with a marginal inclusion probability not less than 0.5), or something else. For our simulations in Section 5, the selected configuration $\hat{S}$ is the median probability model.

Although the SSS can explore many more high posterior configurations than the MCMC can, it is still computationally expensive, especially in high-dimensional cases. When p, the number of candidate predictors, is large and the true dimension $s^\star$ is small, the cost of exhausting all possible configurations in S^+ can be tremendous. Therefore, we adopt the simplified SSS algorithm with screening of Shin, Bhattacharya and Johnson (S5, 2018), which uses a screening technique to significantly decrease the computational cost. More specifically, when considering candidate models with an additional predictor, instead of calculating the posterior probabilities for all possible configurations, we first calculate the partial correlation between response Y and each of the remaining $p - s$ predictors, conditioning on all variables in the current model S^t. Then, we select only the top

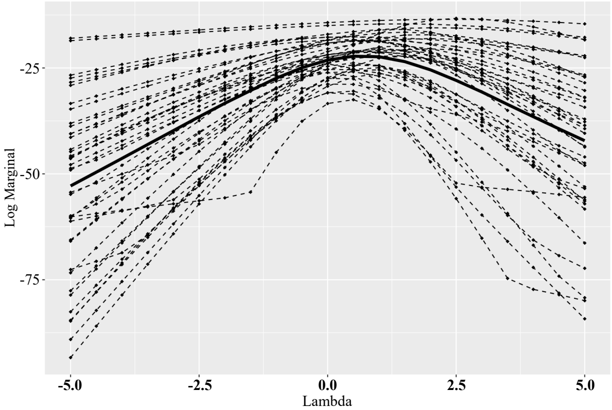


Figure 2. Dotted lines are $\lambda \mapsto \log m_\lambda(Y)$ for different Y samples, and the solid line is the point-wise average, which approximates $\lambda \mapsto \mathsf{E}\{\log m_\lambda(Y)\}$.

K predictors with the highest correlations to form S^+ and S^0. In the simulation, we choose $K = 20$.

4.2. Choice of tuning parameters

4.2.1. Choice of λ

An "ideal" value $\lambda^\star$ of λ is one that minimizes the Kullback–Leibler divergence of the marginal distribution $m_\lambda(y) = \sum_S m_\lambda(y \mid S)\pi_\lambda(S)$ from the true distribution of Y or, equivalently, one that maximizes the expected log marginal likelihood; that is,

$$\lambda^\star = \underset{\lambda}{\operatorname{argmax}}\, \mathsf{E}\{\log m_\lambda(Y)\}.$$

Unfortunately, the ideal value $\lambda^\star$ is not available, because we do not know the true distribution of Y, nor can we estimate it using an empirical distribution. However, a reasonable estimate of the ideal λ is

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \log m_\lambda(Y).$$

Indeed, Figure 2 shows $\log m_\lambda(Y)$ for several different Y samples, along with an approximation of $\mathsf{E}\{\log m_\lambda(Y)\}$ based on point-wise averaging. Note that the individual log marginal likelihoods are maximized very close to where the expectation is maximized.

There is still one more obstacle in obtaining $\hat{\lambda}$, namely, that we cannot directly compute the summation involved in $m_\lambda(Y)$, owing to the large number of configurations S. Fortunately, we can employ an importance sampling strategy to overcome this. Specifically, we have

$$\begin{aligned} m_\lambda(Y) &= \frac{\sum_S m_\lambda(Y \mid S) D(S)^{-\lambda/2|S|} f_n(|S|) \binom{p}{|S|}^{-1}}{\sum_S D(S)^{-\lambda/2|S|} f_n(|S|) \binom{p}{|S|}^{-1}} \\ &\approx \frac{\sum_{\ell=1}^N m_\lambda(Y \mid S_\ell) D(S_\ell)^{-\lambda/2|S_\ell|}}{\sum_{\ell=1}^N D(S_\ell)^{-\lambda/2|S_\ell|}}, \end{aligned}$$

where $\{S_\ell : \ell = 1, \ldots, N\}$ are samples from $\pi_0(S) \propto f_n(|S|)\binom{p}{|S|}^{-1}$. In our numerical results, we use this $m_\lambda(Y)$ to estimate $\hat{\lambda}$.

As discussed in Section 2.3, λ plays an important role in both the model prior and the coefficient prior. That is, for a fixed size s, a positive λ favors models that include predictors with relatively high correlations, and a negative λ favors models that include predictors with relatively low correlations. When λ is equal to zero, the models are treated equally, regardless of their predictors' correlation structure. The λ in the conditional prior for β_S, given S, has a similar effect; see Krishna, Bondell and Ghosh (2009). Thus, a "good" estimate of λ should be such that it reflects the correlation structure in X.

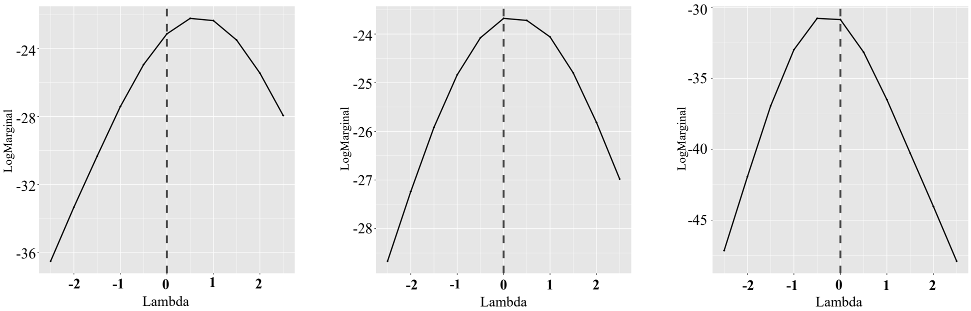
To help see this, consider a few examples, each with X of dimension $n = 100$ and $p = 500$, having an AR(1) correlation structure with varying correlation ρ and true configuration $S^\star$. In particular, we consider two configurations:

$$\begin{aligned} S_1^\star &= \{11, \ldots, 15, 31, \ldots, 35\} \\ S_2^\star &= \{1, 51, 100, 151, 200, 251, 300, 351, 400, 451\}. \end{aligned}$$

Figure 3 shows $\hat{\lambda}$ chosen by maximizing the marginal likelihood in three different cases, and we argue that $\hat{\lambda}$ is at least in the "right direction." In particular, when the true predictors are highly correlated, as in Panel (a), $\hat{\lambda}$ tends to be positive, which encourages the selection of highly correlated predictors. When the true predictors have low correlation, as in Panel (b), the estimate of λ is close to zero; hence, we have a nearly uniform prior for S. The situation in Panel (c) is different because the true predictors are minimally correlated, while unimportant predictors are highly correlated. In this case, $\hat{\lambda}$ tends to be negative, which discourages the selection of the highly correlated ones that are likely unimportant.

4.2.2. Choice of *g*

Now, recall that g determines the magnitude of the prior variance of β_S. If g is sufficiently large, the conditional prior for β_S is effectively flat; if g is extremely tiny, then the posterior probability for β_S will concentrate around the prior center $\phi_n \hat{\beta}_S$. Kass and Wasserman (1995) proposed the unit information criterion, which amounts to taking $g = n$ in the regression setting with Zellner's



(a) $\rho = 0.8$, $S^\star = S_1^\star$ (b) $\rho = 0.1$, $S^\star = S_1^\star$ (c) $\rho = 0.8$, $S^\star = S_2^\star$

Figure 3. Expected log marginal likelihood versus λ, for $\phi = 0$, under different correlation structures of true configurations $S^\star$; see the text for definitions of $S_1^\star$ and $S_2^\star$.

prior. Foster and George (1994) suggest a choice of $g = p^2$. Here, we use a local empirical Bayes estimator for g. That is, for given S and λ, we choose a g that maximizes the local marginal likelihood; that is,

$$\hat{g}_S = \underset{g}{\operatorname{argmax}}\, m_\lambda(y \mid S).$$

In the special case where $\phi_n = 0$ and $\lambda = -1$, and there is a conjugate prior for σ^2, Feng et al. (2008) showed that $\hat{g}_S = \max\{F_S - 1, 0\}$, where F_S is the usual F statistic under model S used to test $\beta_S = 0$. In general, our estimator, $\hat{g}_S$ must be computed numerically.

4.2.3. Choice of ϕ

In our choice of $\phi = \phi_n$, we seek to employ a meaningful amount of shrinkage, while still maintaining the condition in Assumption 2. To this end, if we view $\phi\hat{\beta}_{S^\star}$ as a shrinkage estimator, then it is possible to choose ϕ_n so that the corresponding James–Stein-type estimate has a smaller mean squared error. In particular, this is achieved by

$$\phi_n = 1 - \frac{2\mathsf{E}\|\hat{\beta}_{S^\star} - \beta^\star_{S^\star}\|^2}{\|\beta^\star_{S^\star}\|^2 + \mathsf{E}\|\hat{\beta}_{S^\star} - \beta^\star_{S^\star}\|^2},$$

and, moreover, it can be shown that $1 - \phi_n = O(s^\star\{n\|\beta^\star_{S^\star}\|^2\}^{-1})$; see Section S3 in the Supplementary Material for details. Unfortunately, this ϕ_n still depends on $S^\star$ and $\beta^\star_{S^\star}$, so we need to use a data-driven proxy. We recommend first estimating $S^\star$ using $\hat{S}$ from the adaptive Lasso, with $\hat{\beta}_{\hat{S}}$ and $\hat{\sigma}^2$ the corresponding least squares estimators, and then setting

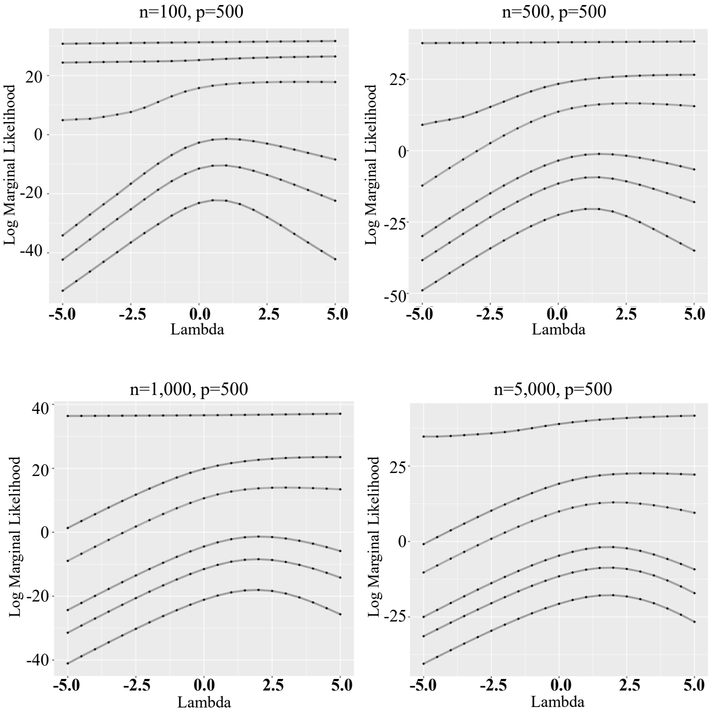


Figure 4. Approximated log marginal likelihood for different values of ϕ, with sample size $n = 100, 500, 1000, 5000$ and $p = 500$, under Scenario 2, as described in Section 5. The value of ϕ is $0.99, 0.95, 0.9, 0.7, 0.5, 0$ from top to bottom for all of the four plots above.

$$\hat{\phi}_n = \left[1 - \frac{2\hat{\sigma}^2 \mathrm{tr}\{(X_{\hat{S}}^\top X_{\hat{S}})^{-1}\}}{\|\hat{\beta}_{\hat{S}}\|^2 + \hat{\sigma}^2 \mathrm{tr}\{(X_{\hat{S}}^\top X_{\hat{S}})^{-1}\}}\right]^+ .$$

In practice, the variable selection results are not sensitive to the choice of ϕ, unless it is too close to one. That is, according to Figure 4, we see good curvature in the log marginal likelihood for λ, with roughly the same maximizer, for a range of ϕ. The curves flatten out when ϕ is too close to one, but that “too close” cutoff gets larger with n. To ensure identifiability of λ, we manually keep our estimate of ϕ away from one, taking $\tilde{\phi}_n = \min\{\hat{\phi}_n, 0.7\}$.

4.2.4. Specification of remaining parameters

It remains to specify the likelihood power α, the tuning parameters (a, c), specifying the prior on the configuration size, the tuning parameter (C, r), specifying the prior on the collinearity of the configurations, given a fixed size, and a plug-in estimator for the error variance σ^2. As in Martin, Mess and Walker (2017), we take $\alpha = 0.999$, $a = 0.05$, and $c = 1$. We let C and r be sufficiently large, so that, in practice, no models are excluded owing to the ill-conditioness. For the error variance, we use the adaptive Lasso to select a configuration, and set $\hat{\sigma}^2$ equal to the mean squared error of the selected configuration. A prior for σ^2 was used by Martin and Tang (2019) in this empirical Bayes framework for a simpler model formulation; their results were similar to those of the plug-in approach adopted here.

5. Simulation Experiments

Here, we investigate the variable selection performance of different methods in five simulated data settings. In each setting, $n = 100$ and $p = 500$, and the error variance σ^2 is set to one. The first two settings have severe collinearity. We employ the first-order autoregressive structure with $\rho = 0.8$ as the covariance structure of the $n \times p$ design matrix X. The true configuration $S^\star$ includes two blocks of variables; the first block contains the 11th to the 15th variables, and the second block contains the 31st to the 35th variables. We explored both large and small signal cases, as follows:

1. $\beta_{S^\star} = (0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95)^\top$
2. $\beta_{S^\star} = (1, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5)^\top$
3. In this case, we consider a block covariance setting, which is the same as Case 4 in Narisetty and He (2014). In this setting, interesting variables have common correlation $\rho_1 = 0.25$; uninteresting variables have common correlation $\rho_2 = 0.75$ and the common correlation between the interesting and uninteresting variables is $\rho_3 = 0.5$. The coefficients of the interesting variables are $\beta_{S^\star} = (0.6, 1.2, 1.8, 2.4, 3.0)^\top$.
4. This case is similar to Case 3, but let $\rho_1 = 0.75$, $\rho_2 = 0.25$, and $\rho_3 = 0.4$. In addition, a larger $\beta_{S^\star} = (1, 1.5, 2.0, 2.5, 3.0)^\top$ is adopted.
5. This is a low correlation case, set the same as Case 2 in Narisetty and He (2014). All variables are set to have common correlation $\rho = 0.25$, and the coefficients of the interesting variables are $\beta_{S^\star} = (0.6, 1.2, 1.8, 2.4, 3.0)^\top$.

For each case, 1,000 data sets are generated. Denoting the chosen configuration as $\hat{S}$, we compute $\mathsf{P}(\hat{S} = S^\star)$ and $\mathsf{P}(\hat{S} \supseteq S^\star)$ in these 1,000 iterations to measure the performance of our method, denoted by ECAP. For comparison purposes, we also consider the Lasso (Tibshirani (1996)), the adaptive Lasso (Zou (2006)), the SCAD (Fan and Li (2001)), the elastic net (EN, Zou and Hastie

Table 1. Simulation results for Cases 1–5. (The best score among the six methods is shown in bold.)

Case	Method	$\mathsf{P}(\hat{S} = S^\star)$	$\mathsf{P}(\hat{S} \supseteq S^\star)$	Average $\lvert\hat{S}\rvert$
1	lasso	0.082	0.996	13.61 (0.09)
	alasso	**0.397**	0.930	10.73 (0.04)
	EN	0.133	0.983	13.24 (0.20)
	SCAD	0	0.001	12.36 (0.15)
	EB	0.165	0.215	9.56 (0.17)
	ECAP	0.263	0.342	9.65 (0.15)
2	lasso	0.297	1	11.65 (0.05)
	alasso	0.356	0.412	9.33 (0.03)
	EN	0.557	0.816	10.25 (0.07)
	SCAD	0	0	7.93 (0.04)
	EB	0.815	1	11.27 (0.91)
	ECAP	**0.994**	1	10.00 (0.00)
3	lasso	0	0.874	18.67 (0.12)
	alasso	0.002	0.277	11.26 (0.10)
	EN	0	0.945	19.82 (0.22)
	SCAD	**0.882**	0.958	5.05 (0.01)
	EB	0.560	0.670	4.69 (0.05)
	ECAP	0.760	0.778	4.90 (0.08)
4	lasso	0.135	1	8.08 (0.09)
	alasso	0.701	0.940	5.34 (0.03)
	EN	0.327	0.997	7.33 (0.13)
	SCAD	0.070	0.148	4.45 (0.04)
	EB	0.793	0.822	4.87 (0.04)
	ECAP	**0.861**	0.940	5.05 (0.07)
5	lasso	0.001	0.990	17.55 (0.15)
	alasso	0.057	0.693	8.63 (0.11)
	EN	0.005	0.991	17.04 (0.28)
	SCAD	0.419	0.908	5.88 (0.04)
	EB	0.680	0.795	4.82 (0.04)
	ECAP	**0.827**	0.919	4.95 (0.05)

(2005)), and an empirical Bayes approach (EB, Martin, Mess and Walker (2017)). The tuning parameters in the first four methods are chosen using the BIC. The results are summarized in Table 1.

According to these results, ECAP performs significantly better than the Lasso, SCAD, and EN in terms of the probability of choosing the true configuration. It also has uniformly better performance compared with that of the EB, which is expected because the ECAP method takes the correlation information into account. However, when considering $\mathsf{P}(\hat{S} \supseteq S^\star)$, the ECAP is not always the highest(e.g., Case 1). Note that $\mathsf{P}(\hat{S} = S^\star)$ and $\mathsf{P}(\hat{S} \supseteq S^\star)$ for the ECAP are always close to each other, which is not the case for the Lasso or EN. This is because the ECAP method is more likely to shrink the coefficients of unimportant predictors to zero, which is desirable if the goal is to find the true $S^\star$.

6. Real-Data Illustration

Here, we examine our method in a real, data example to evaluate its performance against that of other prevalent approaches, including the Lasso, SCAD, and penalized credible region approach in Bondell and Reich (2012). We use data from an experiment conducted by Lan et al. (2006) that studies the genetics of two inbred mouse populations (B6 and BTBR). The data include 22,575 gene expressions of 31 female and 29 male mice. Some phenotypes, including phosphoenopiruvate (PEPCK) and glycerol-3-phosphate acyltransferase (GPAT), were also measured using quatitative real-time PCR. The data are available at the Gene Expression Omnibus data repository (`http://www.ncbi.nlm.nih.gov/geo`; accession number GSE3330).

We choose PEPCK and GPAT as the response variables. Given that this is an ultrahigh-dimensional problem, we use the marginal correlation-based screening method to screen down from 22,575 genes to 1,999 genes. Combining the screened 1,999 genes with the sex variable, the final dimension of the predictor matrix is $p = 2{,}000$. After screening, we apply our method to the data, and select the best subset of predictors $\hat{S}$. Then, we use the posterior mean of β_S as the estimator for β, for given $\hat{S}$ and y. The posterior distribution for β_S is normal, with

$$\begin{aligned} \text{mean} &= \big(X_{\hat{S}}^\top X_{\hat{S}} + V_{\hat{S}}^{-1}\big)^{-1}\big(X_{\hat{S}}^\top y + \phi V_{\hat{S}}^{-1}\hat{\beta}_{\hat{S}}\big) \\ \text{covariance} &= \sigma^2\big(X_{\hat{S}}^\top X_{\hat{S}} + V_{\hat{S}}^{-1}\big)^{-1}, \end{aligned}$$

where $V_{\hat{S}} = g k_{\hat{S}}\big(X_{\hat{S}}^\top X_{\hat{S}}\big)^\lambda$. For the hyperparameters λ, ϕ, and g, we can plug in their corresponding estimators, given in Section 4.

In order to evaluate the performance of our approach, we randomly split the sample into a training data set of size 55 and a test set of size five. First, we apply our variable selection method to the training set and obtain the selected variables. Then, conditioning on this model, we estimate the regression coefficients using the above method. Based on the estimated regression coefficient, we predict the remaining five observations and calculate the prediction loss. This process is repeated 100 times, and we can compute an estimated mean squared prediction

Table 2. Mean squared prediction error (MSPE) and average configuration size in the real-data example of Section 6; numbers in parentheses are standard errors. The results except for the ECAP are from Bondell and Reich (2012).

	PEPCK		GPAT	
Method	MSPE	Model Size	MSPE	Model Size
ECAP ($p = 2{,}000$)	1.02 (0.07)	5.04 (0.19)	2.26 (0.18)	8.34 (0.33)
lasso ($p = 2{,}000$)	3.03 (0.19)	7.70 (0.96)	5.03 (0.42)	3.30 (0.79)
BCR.joint ($p = 2{,}000$)	2.03 (0.14)	9.60 (0.46)	3.83 (0.34)	4.20 (0.43)
BCR.marginal ($p = 2{,}000$)	1.84 (0.14)	23.3 (0.67)	5.33 (0.41)	21.8 (0.72)
SIS+SCAD ($p = 22{,}575$)	2.82 (0.18)	2.30 (0.09)	5.88 (0.44)	2.60 (0.10)
ECAP ($p = 22{,}575$)	**0.72 (0.07)**	4.93 (0.30)	**1.66 (0.52)**	7.92 (0.73)

error (MSPE), along with its standard error; see Table 2.

In Table 2, BCR.joint and BCR.marginal denote methods using joint credible sets and marginal credible sets, respectively, for details, see Bondell and Reich (2012). The first four rows correspond to the ECAP, the Lasso, BCR.joint, and BCR.marginal, applied to the screened data with dimension $p = 2{,}000$. The fifth row shows sure independence screening (SIS) combined with the SCAD, applied to the full data $p = 22{,}575$, and the last row is based on directly applying the ECAP to the unscreened data. The stopping rules for the Lasso, the SCAD, BCR.joint, and BCR.marginal are based on the BIC.

In terms of the MSPE, the ECAP outperforms the other methods significantly in both the PEPCK and the GPAT cases, given the estimated standard errors. Moreover, the MSPE from the ECAP is even smaller for the full data set than it is for the screened data. For the model size, on average, the ECAP, the Lasso, BCR.joint, and the SIS+SCAD select models with comparable sizes, while BCR.marginal always chooses larger models. Overall, the ECAP performs very well in this real data example compared with these other methods in terms of both the MSPE and the model size.

Supplementary Material

The online Supplementary Material contains proofs of the theorems presented in Section 3, along with details about our choice of ϕ and some additional simulation experiments.

Acknowledgments

The authors thank the editor, associate editor, and two reviewers for their helpful feedback. The work presented herein was partially supported by the

U.S. National Science Foundation, grant DMS–1737933.

References

Arias-Castro, E. and Lounici, K. (2014). Estimation and variable selection with exponential weights. *Electron. J. Statist.* **8**, 328–354.

Bondell, H. D. and Reich, B. J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *J. Amer. Statist. Assoc.* **107**, 1610–1624.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer Science & Business Media.

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40**, 2069–2101.

Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43**, 1986–2018.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20**, 101–148.

Feng, L., Rui, P., German, M., Merlise, A. and Jim, O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103**, 410–423.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947–1975.

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

George, E. I. and McCullogh, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881–889.

Grünwald, P. and van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal.* **12**, 1069–1103.

Hans, C., Dobra, A. and West, M. (2007). Shotgun stochastic search for large p regression. *J. Amer. Statist. Assoc.* **102**, 507–517.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab gene selection for multigroup microarray data. *J. Amer. Statist. Assoc.* **100**, 764–780.

Kass, R. E. and Wasserman,L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the schwarz criterion. *J. Amer. Statist. Assoc.* **90**, 928–934.

Krishna, A., Bondell, H. and Ghosh, S. K. (2009). Bayesian variable selection using an adaptive powered correlation prior. *J. Statist. Plann. Inference* **139**, 2665–2674.

Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M. et al. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics* **2**, e6.

Martin, R., Mess, R. and Walker, S. G. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli* **23**, 1822–1847.

Martin, R. and Tang, Y. (2019). Empirical priors for prediction in sparse high-dimensional linear regression. *arXiv preprint arXiv:1903.00961.*

Martin, R. and Walker, S. G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electron. J. Stat.* **8**, 2188–2206.

Martin, R. and Walker, S. G. (2019). Data-driven priors and their posterior concentration rates.

Electron. J. Stat. **13**, 3049–3081.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.* **42**, 789–817.

Polson, N. G. and Scott, J. G. (2012). Local shrinkage rules, Lévy processes and regularized regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74**, 287–311.

Rigollet, P. and Tsybakov, A. B. (2012). Sparse estimation by exponential weighting. *Statist. Sci.* **27**, 558–575.

Shin, M., Bhattacharya, A. and Johnson, V. E. (2018). Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statist. Sinica* **28**, 1053–1078.

Syring, N. and Martin, R. (2019). Calibrating general posterior credible regions. *Biometrika* **106**, 479–486.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol* **58**, 267–288.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567–1594.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **67**, 301–320.

Chang Liu
Department of Statistics, North Carolina State University, North Carolina, USA.
E-mail: cliu22@ncsu.edu
Yue Yang
Department of Statistics, North Carolina State University, North Carolina, USA.
E-mail: yyang44@ncsu.edu
Howard Bondell
School of Mathematics and Statistics, University of Melbourne, Melbourne, Australia.
E-mail: howard.bondell@unimelb.edu.au
Ryan Martin
Department of Statistics, North Carolina State University, North Carolina, USA.
E-mail: rgmarti3@ncsu.edu

(Received April 2019; accepted March 2020)