Statistica Sinica Preprint No: SS-2019-0133	
Title	Bayesian inference in high-dimensional linear models using an empirical correlation-adaptive prior
Manuscript ID	SS-2019-0133
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0133
Complete List of Authors	Chang Liu, Yue Yang, Howard Bondell and Ryan Martin
Corresponding Author	Howard Bondell
E-mail	howard.bondell@unimelb.edu.au
Notice: Accepted version subject to English editing.	

Statistica Sinica

Bayesian inference in high-dimensional linear models using an empirical correlation-adaptive prior

Chang Liu, Yue Yang, Howard Bondell, and Ryan Martin

North Carolina State University and University of Melbourne

Abstract: In the context of a high-dimensional linear regression model, we propose the use of an empirical correlation-adaptive prior that makes use of information in the observed predictor variable matrix to adaptively address high collinearity, determining if parameters associated with correlated predictors should be shrunk together or kept apart. Under certain conditions, we prove that our empirical Bayes posterior concentrates at the optimal rate, so 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, and we demonstrate our method's favorable performance compared to existing methods in real- and simulated data examples, even in ultra-high 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 a $n \times 1$ vector of response variables, X is a $n \times p$ matrix of predictor variables, β is a $p \times 1$ vector of regression coefficients, and ε is a vector of iid $\mathsf{N}(0, \sigma^2)$ errors. We are interested in the high-dimensional case, where $p \gg n$. Furthermore, it is assumed 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. This includes regularization-based methods like the lasso (Tibshirani, 1996), the adaptive lasso (Zou, 2006), the SCAD (Fan and Li, 2001), and MCP (Zhang, 2010); see Fan and Lv (2010) for a review. From a Bayesian point of view, varieties 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 Mar-

tin et al. (2017) consider an empirical Bayes version of spike-and-slab.

Collinearity is unavoidable in high-dimensional settings, and methods such as lasso tend to smooth away the regression coefficients of highly collinear predictors, and hence deter correlated covariates to be included in the model simultaneously. This motivated Krishna et al. (2009) to propose an adaptive powered correlation prior that lets the data itself weigh in on how 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 correlated predictors. In Section 2, we present our empirical Bayes model and a motivating example illustrating the effect of correlation-adaptation in the prior. Asymptotic posterior concentration properties are derived in Section 3, in particular, minimax optimal concentration rates are established for the mean response, showing that the finite-sample benefits of correlation-adaptation leads to no loss of asymptotic optimality. In Section 4, we recommend a shotgun stochastic search approach for computation of the posterior distribution over the model space. Simulation experiments are presented in Section 5, and we demonstrate the benefits of adaptively varying the correlation structure in

the prior for variable selection compared to existing methods. The real-data illustration in Section 6 highlights the improved predictive performance even in ultra-high dimensional settings that can be achieved using the proposed correlation-adaptive prior. 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 non-zero coefficients, called the *configuration* of β, and β_S is the $|S|$-vector of non-zero values, with $|S|$ denoting the cardinality of S. We will write X_S for the sub-matrix of X corresponding to the configuration S. With this decomposition of β, a hierarchical prior is convenient, i.e., a prior for S and a conditional prior for β_S, given S.

First, for the prior $\pi(S)$ for S, we follow Martin et al. (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 et al. (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 $R = \text{rank}(X) \leq n$. 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

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

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 will be discussed in more detail below. Because of the dependence on λ above, we will henceforth write $\pi_\lambda(S)$ for the prior of S.

In these high-dimensional problems, 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 concentration at a sub-optimal rate, so they recommend the use of 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: use a theoretically justified heavy-tailed prior that makes computation more difficult, or a use a computationally conve-

nient thin-tailed prior with potentially sub-optimal posterior convergence properties? Martin et al. (2017) observed that the the prior tails are less relevant if the center is appropriately chosen so, to overcome the aforementioned dilemma, they proposed the use of 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

$$k_S = \mathrm{tr}\{(X_S^\top X_S)^{-1}\}/\mathrm{tr}\{(X_S^\top X_S)^\lambda\}$$

is a standardizing factor as in Krishna et al. (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 correlated predictors towards 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, positive λ would prefer larger models by capturing as many correlated predictors as possible, while negative λ tends to select models with less collinearity; see Krishna et al. (2009) for additional 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, will be discussed in Section 4.2.

2.2 The posterior distribution

For this standard linear regression model, the likelihood function is

$$L_n(\beta) = (2\pi\sigma^2)^{-\frac{n}{2}} \, e^{-\frac{1}{2\sigma^2}\|Y - X\beta\|^2}, \quad \beta \in \mathbb{R}^p.$$

It would be straightforward to include σ^2 as an argument in this likelihood function, introduce a prior for σ^2, and get 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 and, hence, we omit the error variance as an argument to the likelihood function.

Given a prior and the likelihood, we can combine the two via 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 data changes the way we think about the posterior construction. Specifically, updating the data-dependent prior with the full likelihood amounts to a

double-use of the data, hence a risk of over-fitting and, to avoid that 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 with only a (large) portion of the likelihood. As is done 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 filling in around β_S with zeros in the entries corresponding to S^c, and $\alpha \in (0, 1)$ is a regularization factor, which can be taken arbitrarily close to 1. 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, i.e.,

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

Since 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 above to get a marginal likelihood for Y, i.e.,

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

where $\hat{y}_S$ is the least square estimate of y given model S, $d_{S,i}$ is the i^{th} 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 $\Lambda_S = \text{diag}(d_{S,1}, \ldots, d_{S,s})$, and $\theta_{S,i}$ is the i^{th} element of $\theta_S = \Gamma_S^\top \hat{\beta}_S$. Then it is straightforward to get the posterior distribution for S:

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

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

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 iid 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

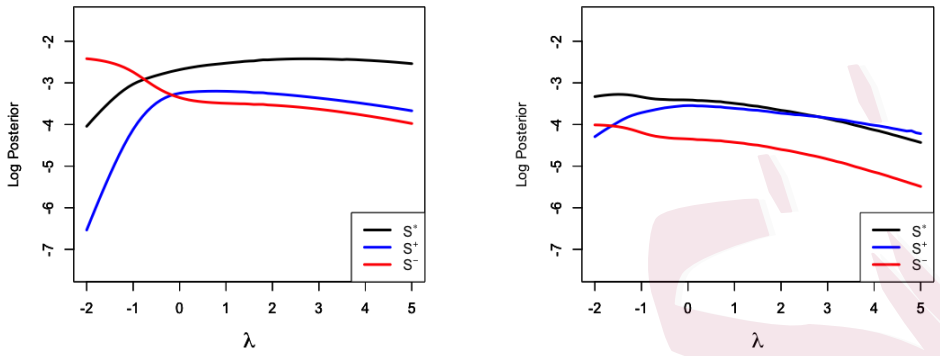
$$y_i = x_{i1} + 0.8x_{i2} + \varepsilon_i, \quad \text{where} \quad \varepsilon_1, \ldots, \varepsilon_5 \overset{\text{iid}}{\sim} \mathsf{N}(0,1).$$

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 $S^+ = \{1,2,3\}$, respectively. Panel (a) corresponds to a high correlation case, $\rho = 0.8$, and we see that the ECAP-based posterior would prefer $S^\star$ for suitably large λ; compare this to the choice $\lambda \equiv -1$ in Martin et al. (2017) which would prefer the smaller configuration S^-. On the other hand, when the correlation is relatively low, as in Panel (b), we see that large positive λ would encourage larger configuration while the true configuration would be 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 will stick with the standard notation given previously, but it will help to keep in mind that $Y^n = (Y_1^n, Y_2^n, ..., Y_n^n)$ and $X^n = ((X_{ij}^n))$ are better



(a) $\rho = 0.8$ (b) $\rho = 0.1$

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

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 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 will drop the subscript λ and simply write Π^n for the posterior for β instead of Π^n_λ.

For 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 will be seen 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 $\beta^\star_{S^\star}$, i.e., the non-zero regression coefficients of the true $\beta^\star$ as well as on the user-specified shrinkage factor $\phi = \phi_n$ in the prior.

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

Inside of Assumption 2 is a very mild condition on the true $\beta^\star$, i.e., that the "total signal" $\|\beta^\star_{S^\star}\|$ is not too small. There is, of course, no reason to think that the individual signals would be vanishing with n and, if not, then we get $s^\star\{n\|\beta^\star_{S^\star}\|\}^{-1} \to 0$ automatically from Assumption 1. But it is not required that *all* of the signals are bounded away from 0; the condition is about the total signal so it is enough that at least one of the signals is away from 0. Even if we ask for all the non-zero signals

to be lower-bounded, the condition above would hold if $\min_{j \in S^\star} |\beta_j^\star| > n^{-1/2}$, and an even stronger *beta-min* condition—see (3.3) in Section 3.5—is needed for establishing variable selection consistency here and throughout the literature on high-dimensional inference (e.g., Bühlmann and van de Geer, 2011; Arias-Castro and Lounici, 2014).

There is also some insight in the connection between ϕ and the total signal. That is, ϕ controls the influence of the prior centering and, when the total signal is large, that 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 non-increasing and non-decreasing functions of the complexity s, respectively. 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 get 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 says, roughly, that every submatrix X_S, for $|S| \leq s^\star$, is full rank, and 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, and the following theorem states this result this 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 $|S_{\beta^\star}| = s^\star$,

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

and

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

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. So the take-away message here is that our proposed ECAP posterior attains the minimax optimal rate as long as the (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, since $\beta^\star$ is sparse, we might ask if the posterior 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 estimation error loss

Following up on the result in Section 3.2 on the posterior concentration with respect to the mean response difference, we might ask if 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; see below.

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 1. 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 that accounts for collinearity structure in the design matrix is for the purpose of more effective variable selection. So it is imperative that we can show, at least asymptotically, that our posterior distribution will concentrate on 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*

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

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

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

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, i.e.,

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

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 on comparatively small configurations. Also, the non-zero $\beta^\star$ values are more difficult to detect if their magnitudes are small. This is intuitively clear, and also shows up in our simulation results for Cases 1–2 in Section 5. Theorem 4 gives a mathematical explanation of this intuition, stating that variable selection based on our empirical Bayes posterior will be 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

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

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. Since C and r can be chosen large enough so that only the few extremely ill-conditioned cases would be excluded, that 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 display, an average of these geometric means, is roughly constant in both S and s, so it is not unreasonable to approximate $\pi_\lambda(S \mid |S| = s)$ in (2.2) by

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

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

$$\pi_\lambda^n(S) \propto m_\lambda(Y \mid S)\, D(S)^{-\frac{\lambda}{2|S|}} \binom{p}{|S|}^{-1} f_n(|S|) 1\{\kappa(S) < Cp^r\}. \tag{4.4}$$

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

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 et al., 2007), to explore our posterior distribution. Different from traditional MCMC method, 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 can be 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 only have one variable different from variables in S, and S^- is the set containing all $(s-1)$-

dimensional configurations that are nested in S. The t^{th} iteration of SSS goes as follows:

1. Given S^t, compute $\pi_\lambda^n(S)$ for all $S \in \text{nbd}(S^t) = \{S^{t+}, S^{t0}, S^{t-}\}$.

2. Sample S_1^t, S_2^t, S_3^t respectively from S^{t+}, S^{t0}, 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 $\pi_\lambda^n(S^{t+})$, $\pi_\lambda^n(S^{t0})$, and $\pi_\lambda^n(S^{t-})$, obtained by summing.

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

While SSS has the ability to explore many more high posterior configurations than MCMC, it is still computationally expensive, especially in high-dimensional case. When p, the number of candidate predictors, is large and the true dimension $s^\star$ is small, cost of exhausting all possible configurations in S^+ can be tremendous. For this reason, we adopt the simplified SSS algorithm with screening in Shin et al. (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 the possible configurations, we first calculate the partial correlation between response Y and each of the rest $p - s$ predictors conditioning on all the variables in the current model S^t. Then we only select the top K predictors with highest correlation 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, i.e.,

$$\lambda^\star = \arg\max_\lambda \mathsf{E}\{\log m_\lambda(Y)\}.$$

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

$$\hat{\lambda} = \arg\max_\lambda \log m_\lambda(Y).$$

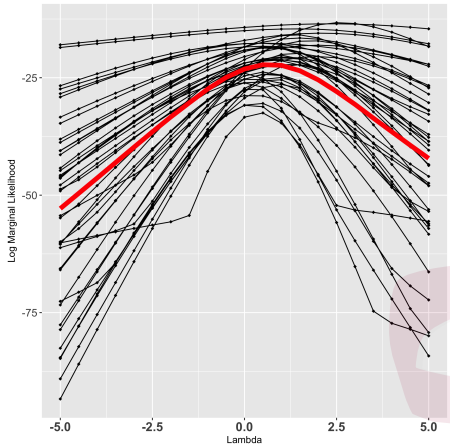


Figure 2: Black lines are $\lambda \mapsto \log m_\lambda(Y)$ for different Y samples, and the red line is the point-wise average, which approximates $\lambda \mapsto \mathsf{E}\{\log m_\lambda(Y)\}$.

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. Notice that the individual log marginal likelihoods are maximized very close by 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 $m_\lambda(Y)$ due 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 $\hat{\lambda}$.

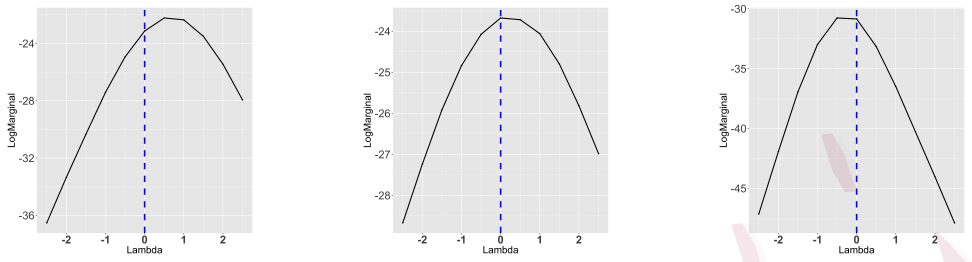
As discussed in Section 2.3, λ plays an important role in both model prior and coefficient prior. That is, for a fixed size s, a positive λ favors model including predictors with relatively high correlation; a negative λ favors model including predictors with relatively low correlation; λ equals zero actually put equal mass on each model regardless of their predictors' correlation structure. The λ in the conditional prior for β_S, given S, has a similar effect; see Krishna et al. (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:

$$S_1^\star = \{11, \ldots, 15, 31, \ldots, 35\}$$

$$S_2^\star = \{1, 51, 100, 151, 200, 251, 300, 351, 400, 451\}.$$

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 highly correlated predictors to be



(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 structure of true configurations $S^\star$; see definition of $S_1^\star$ and $S_2^\star$ in the text.

selected; and when the true predictors have low correlation, as in Panel (b), the estimate of λ is close to 0 hence a nearly uniform prior for S. The situation in Panel (c) is different since the true predictors are minimally correlated while unimportant predictors are highly correlated. In this case, $\hat{\lambda}$ tends to be negative which discourages selecting 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

on the prior center $\phi_n\hat{\beta}_S$. Kass and Wasserman (1995) proposed the unit information criteria, which amounts to taking $g = n$ in the regression setting with Zellner's 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,

$$\hat{g}_S = \arg\max_g m_\lambda(y \mid S).$$

In the special case where $\phi_n = 0$ and $\lambda = -1$, and 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 for testing $\beta_S = 0$. 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. Towards this, 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 smaller mean square error. In particular, a ϕ_n that achieves this is

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

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 some data-driven proxy for this. We recommend first estimating $S^\star$ by $\hat{S}$ from the adaptive lasso, with $\hat{\beta}_{\hat{S}}$ and $\hat{\sigma}^2$ the corresponding least squares estimators, and then setting

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

In practice, variable selection results are not sensitive to the choice of ϕ unless it is too close to 1. 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 1, but that "too close" cutoff gets larger with n. To ensure identifiability of λ, we manually keep our estimate of ϕ away from 1, in particular, we take $\tilde{\phi}_n = \min\{\hat{\phi}_n, 0.7\}$.

4.2.4 Specification of remaining parameters

It remains to specify the likelihood power α, the tuning parameters (a, c), specifying the prior on configuration size, the tuning parameter (C, r), specifying the prior on collinearity of configurations given a fixed size, and to specify a plug-in estimator for the error variance σ^2. As in Martin et al. (2017), we take $\alpha = 0.999$, $a = 0.05$, and $c = 1$. We let C and r be suffi-

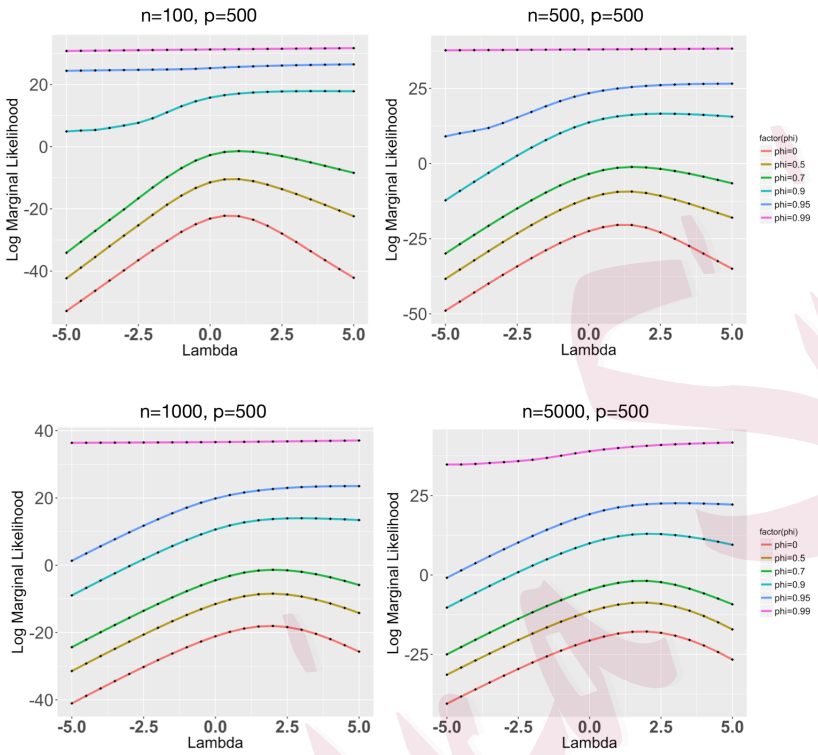


Figure 4: Approximated log marginal likelihood for different values of ϕ with sample size $n = 100, 500, 1000, 5000$ and $p = 500$, under Scenario 2 as is described in Section 5.

ciently large, so that in practice no models would be excluded due 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 square error for that selected configuration. A prior for σ^2 was used by Martin and Tang (2019) in this empirical Bayes framework for a simpler model formulation, and the results were similar compared to 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 1. 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; and the true configuration $S^\star$ includes two blocks of variables that the first block contains the 11th to the 15th variable and the second block contains the 31st to the 35th variable. 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 the Case 4 in Narisetty and He (2014). In this setting, interesting variables have common correlation $\rho_1 = 0.25$; uninteresting variables have common correlation $\rho_2 = 0.75$; the common correlation between interesting and uninteresting ones 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 $\rho_3 = 0.4$. Also, a larger $\beta_{S^\star} = (1, 1.5, 2.0, 2.5, 3.0)^\top$ is adopted.

5. This case is a low correlation case, which is 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 for interesting variables are $\beta_{S^\star} = (0.6, 1.2, 1.8, 2.4, 3.0)^\top$.

For each case, 1000 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 1000 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 smoothly clipped absolute deviation (SCAD, Fan and Li, 2001), the elastic net (EN, Zou and Hastie, 2005) and an empirical Bayes approach (EB, Martin et al., 2017). Tuning parameters in the first four methods are chosen by BIC. The results are summarized in Table 1.

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)

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

According to these results, ECAP performs significantly better than lasso, SCAD, and EN in terms of the probability of choosing the true configuration. It also has uniformly better performance compared with EB, which is expected since the new ECAP method takes the correlation information into account. However, when considering $\mathsf{P}(\hat{S} \supseteq S^\star)$, 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 ECAP are always close to each other, which is not the case for 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 and evaluate its performance against other prevalent approaches including lasso, SCAD and

the penalized credible region approach in Bondell and Reich (2012). We use the 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 22575 gene expressions of 31 female and 29 male mice. Some phenotypes, including phosphoenopiruvate (PEPCK) and glycerol-3-phosphate acyltransferase (GPAT) were also measured by quatitative real-time PCR. The data are available at Gene Expression Omnibus data repository (`http://www.ncbi.nlm.nih.gov/geo`; accession number GSE3330).

We choose PEPCK and GPAT as response variables. Given that this is an ultra-high dimensional problem, we use marginal correlation based screening method, to screen down from 22575 genes to 1999 genes. Combining the screened 1999 genes with the sex variable, the final dimension of the predictor matrix is $p = 2000$. After screening, we apply our method to the data and select a 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

$$\text{mean} = \left(X_{\hat{S}}^\top X_{\hat{S}} + V_{\hat{S}}^{-1}\right)^{-1}\left(X_{\hat{S}}^\top y + \phi V_{\hat{S}}^{-1}\hat{\beta}_{\hat{S}}\right)$$

$$\text{covariance} = \sigma^2\left(X_{\hat{S}}^\top X_{\hat{S}} + V_{\hat{S}}^{-1}\right)^{-1},$$

where $V_{\hat{S}} = g k_{\hat{S}}\left(X_{\hat{S}}^\top X_{\hat{S}}\right)^{\lambda}$. For 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 5. 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 5 observations and calculate the prediction loss. This process is repeated 100 times, and an estimated mean square prediction error (MSPE) along with its standard error can be computed; 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 ECAP, lasso, BCR.joint and BCR.marginal applied to the screened data with dimension $p = 2000$. The fifth row is for sure independence screening (SIS) combined with SCAD applied to the full data $p = 22575$, and the last row is based on directly applying ECAP to the unscreened data. The stopping rules for lasso, SCAD, BCR.joint and BCR.marginal are based on BIC.

In terms of MSPE, ECAP outperforms all the other methods significantly in both PEPCK and GPAT cases, given the estimated standard errors. Moreover, the MSPE from ECAP is even smaller for the full dataset

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

	PEPCK		GPAT	
Method	MSPE	Model Size	MSPE	Model Size
ECAP ($p = 2000$)	1.02 (0.07)	5.04 (0.19)	2.26 (0.18)	8.34 (0.33)
lasso ($p = 2000$)	3.03 (0.19)	7.70 (0.96)	5.03 (0.42)	3.30 (0.79)
BCR.joint ($p = 2000$)	2.03 (0.14)	9.60 (0.46)	3.83 (0.34)	4.20 (0.43)
BCR.marginal ($p = 2000$)	1.84 (0.14)	23.3 (0.67)	5.33 (0.41)	21.8 (0.72)
SIS+SCAD ($p = 22575$)	2.82 (0.18)	2.30 (0.09)	5.88 (0.44)	2.60 (0.10)
ECAP ($p = 22575$)	**0.72 (0.07)**	4.93 (0.30)	**1.66 (0.52)**	7.92 (0.73)

compared with the screened data. And for the model size, on average, ECAP, lasso, BCR.joint and SIS+SCAD select models with comparable sizes while BCR.marginal always chooses larger models. Overall, ECAP performs very well in this real data example compared to these other methods in terms of both MSPE and model size.

Supplementary Materials

Proofs of the theorems presented in Section 3 are included in the supplementary materials, along with details about our choice of ϕ and some additional simulation experiments.

Acknowledgements

The authors thank the editor, associate editor, and two reviewers for their helpful feedback. The work presented herein is partially supported by the U.S. National Science Foundation, grant DMS–1737933.

References

Arias-Castro, E. and K. Lounici (2014). Estimation and variable selection with exponential weights. *Electron. J. Statist. 8*(1), 328–354.

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

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

Castillo, I. and A. van der Vaart (2012). Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics. 40*(4), 2069–2101.

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

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

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

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

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

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

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

Grünwald, P. and T. van Ommen (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal. 12*(4), 1069–1103.

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

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

Kass, R. E. and L. Wasserman (1995). A reference Bayesian test for nested hypotheses and its

relationship to the schwarz criterion. *J. Amer. Statist. Assoc. 90*, 928–934.

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

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

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

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

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

Martin, R. and S. G. Walker (2019). Data-driven priors and their posterior concentration rates. *Electron. J. Stat. 13*, 3049–3081.

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

Polson, N. G. and J. G. Scott (2012). Local shrinkage rules, Lévy processes and regularized regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 74*(2), 287–311.

Rigollet, P. and A. B. Tsybakov (2012). Sparse estimation by exponential weighting. *Statist.*

Sci., 558–575.

Shin, M., A. Bhattacharya, and V. E. Johnson (2018). Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statist. Sinica 28*(2), 1053–1078.

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

Department of Statistics, North Carolina State University, North Carolina, United States of America.

E-mail: cliu22@ncsu.edu

Department of Statistics, North Carolina State University, North Carolina, United States of

America.

E-mail: yyang44@ncsu.edu

School of Mathematics and Statistics, University of Melbourne, Melbourne, Australia

E-mail: howard.bondell@unimelb.edu.au

Department of Statistics, North Carolina State University, North Carolina, United States of America.

E-mail: rgmarti3@ncsu.edu