Statistica Sinica: Supplement

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

North Carolina State University and University of Melbourne

Supplementary Material

This document includes the proofs of the theorems presented in Section 3 of the article "Bayesian inference in high-dimensional linear models using an empirical correlation-adaptive prior," along with some details about the choice of tuning parameter ϕ and some additional simulation results.

S1 Preliminary lemmas

Before getting to the proofs of the main theorems, we need to suitably lower-bound the posterior denominator and upper-bound the posterior numerator, the latter depending on the type of neighborhood being considered. In particular, for a generic measurable subset A of the parameter space, write

$$\Pi^n(A) = \frac{N_n(A)}{D_n} = \frac{\int_A \sum_S \pi(S) R_n(\beta_{S+}, \beta^*)^\alpha \pi_\lambda(\beta_S \mid S)\, d\beta_S}{\int \sum_S \pi(S) R_n(\beta_{S+}, \beta^*)^\alpha \pi_\lambda(\beta_S \mid S)\, d\beta_S},$$

where $R_n(\beta_{S+}, \beta^*) = L_n(\beta_{S+})/L_n(\beta^\star)$ is the likelihood ratio, with β_{S+} the p-vector corresponding to β_S with zeros filled in around the indices in S. Recall that $\beta^\star$ denotes the true and sparse coefficient vector in $\mathbb{R}^p$, with $S^\star = S_{\beta^\star}$ the corresponding configuration of complexity $s^\star = |S^\star|$. Also, recall that the condition number, $\kappa(S^\star)$, of the matrix $n^{-1} X_{S^\star}^\top X_{S^\star}$ is $O(1)$ as $n \to \infty$ by Assumption 3 in the main text. Lemma 1 gives a general lower bound on the denominator D_n.

Lemma 1. *Given the inputs (α, g, λ) to our proposed posterior, define*

$$b = b(\alpha, g, \lambda, S^*) = \begin{cases} \frac{1}{2}\log\{1 + \alpha g \kappa(S^\star)^{1+\lambda}\}, & \text{if } \lambda \in [0, \infty) \\ \frac{1}{2}\log\{1 + \alpha g \kappa(S^\star)\}, & \text{if } \lambda \in [-1, 0) \\ \frac{1}{2}\log\{1 + \alpha g \kappa(S^\star)^{-\lambda}\}, & \text{if } \lambda \in (-\infty, -1). \end{cases} \tag{S1.1}$$

Then under Assumptions 1–3, the denominator D_n of the posterior satisfies $\mathsf{P}_{\beta^\star}\{D_n < \pi(S^\star) e^{-bs^\star}\} \to 0$ as $n \to \infty$.

Proof. Since D_n is a sum of non-negative terms, we get the trivial bound

$$\begin{aligned} D_n &> \pi(S^\star) \int R_n(\beta_{S^\star}, \beta^\star_{S^\star})^\alpha \mathsf{N}\big(d\beta_{S^\star} \mid \phi\hat\beta_{S^\star}, \sigma^2 g k_{S^\star} (X_{S^\star}^\top X_{S^\star})^\lambda\big) \\ &= \pi(S^\star) \prod_{i=1}^{s^\star} (1 + \alpha g k_{S^\star} d_{S^\star,i}^{1+\lambda})^{-1/2} \exp\Big\{ \frac{\alpha}{2\sigma^2}(A_n - B_n)\Big\}, \end{aligned}$$

where

$$\begin{aligned} A_n &= n(\hat\beta_{S^\star} - \beta^\star_{S^\star})^\top \big(n^{-1} X_{S^\star}^\top X_{S^\star}\big)(\hat\beta_{S^\star} - \beta^\star_{S^\star}) \\ B_n &= (1-\phi)^2 \hat\beta_{S^\star}^\top Q_{S^\star}^{-1} \hat\beta_{S^\star} \end{aligned}$$

are both non-negative, and the Q_S matrix is defined as

$$Q_S = (X_S^\top X_S)^{-1} + \alpha g k_S (X_S^\top X_S)^\lambda, \quad S \subseteq \{1, 2, \ldots, p\}. \tag{S1.2}$$

From the well-known sampling distribution of $\hat\beta_{S^\star}$, we have

$$A_n \sim \sigma^2 \,\mathsf{ChiSq}(s^\star) = O_p(s^\star).$$

Next, for B_n, under Assumption 3, it can be verified that the maximal eigenvalue of $Q_{S^\star}^{-1}$ is $O(n)$. Therefore, $B_n \lesssim n(1-\phi)^2 \|\hat\beta_{S^\star}\|^2$. Write

$$\|\hat\beta_{S^\star}\|^2 \lesssim \|\hat\beta_{S^\star} - \beta^\star_{S^\star}\|^2 + \|\beta^\star_{S^\star}\|^2$$

Since the first term is $O_p(s^\star n^{-1})$ and $(1-\phi)^2 \le o(1)$, we have

$$n(1-\phi)^2 \|\hat\beta_{S^\star} - \beta^\star_{S^\star}\|^2 = o_p(s^\star).$$

Similarly, by Assumption 2, for the second term we have

$$n(1-\phi)^2 \|\beta^\star_{S^\star}\|^2 = o(s^\star).$$

This implies $B_n = o_p(s^\star)$ and, hence, $\mathsf{P}_{\beta^\star}(A_n \leq B_n) \to 0$ which, in turn, implies that the exponential term in the lower bound for D_n is no smaller than 1. Therefore,

$$D_n > \pi(S^\star) \prod_{i=1}^{s^\star} (1 + \alpha g k_{S^\star} d_{S^\star,i}^{1+\lambda})^{-1/2}$$

and the product can be lower-bounded by $e^{-bs^\star}$ for b as defined in (S1.1). Finally, we have that $D_n \leq \pi(S^\star) e^{-bs^\star}$ with vanishing $\mathsf{P}_{\beta^\star}$-probability, as was to be shown. □

Next, consider the subset B_ε of $\mathbb{R}^p$ given by

$$B_\varepsilon = \{\beta \in \mathbb{R}^p : \|X\beta - X\beta^\star\|^2 > \varepsilon\}, \quad \varepsilon > 0.$$

For the sequence ε_n and constant $M > 0$ as in the statement of Theorem 1 in the main text, set $N_n = N_n(B_{M\varepsilon_n})$. The following lemma gives an upper bound on N_n.

Lemma 2. *There exists a constant $d = d(\alpha, \sigma^2)$ such that*

$$\sup_{\beta^\star} \mathsf{E}_{\beta^\star}(N_n) \leq e^{-d\varepsilon_n} \sum_{S:|S| \leq R} \psi(|S|)^{|S|} \pi_\lambda(S),$$

where

$$\psi(s)^2 = \begin{cases} \omega(s)^{2(\lambda+1)} \left[1 + \frac{q\phi^2}{g} \omega(s)^{-(\lambda+1)}\right]^{1-\frac{1}{q}} & \text{if } \lambda \in [0, \infty) \\ \omega(s)^2 \left[1 + \frac{q\phi^2}{g} \omega(s)^{-1}\right]^{1-\frac{1}{q}} & \text{if } \lambda \in [-1, 0) \\ \omega(s)^{-2\lambda} \left[1 + \frac{q\phi^2}{g} \omega(s)^{\lambda}\right]^{1-\frac{1}{q}} & \text{if } \lambda \in (-\infty, -1), \end{cases} \tag{S1.3}$$

where $q = h(h-1)^{-1}$ and $h \in (1, \alpha^{-1})$ is a constant.

Proof. Let us consider the expectation of N_n, given the true distribution of Y, i.e. $Y \sim \mathsf{N}(X\beta^\star, \sigma^2 I)$. Then by Hölder's inequality, for constants $h > 1$ and $q = h(h-1)^{-1}$, we can find an upper bound for $\mathsf{E}_{\beta^\star}(N_n)$,

$$\mathsf{E}_{\beta^\star}(N_n) \leq \sum_S \pi(S) \int_{B_{\epsilon_n}} J_n(\beta_S)^{\frac{1}{h}} K_n(\beta_S)^{\frac{1}{q}} d\beta_S, \tag{S1.4}$$

where

$$\begin{aligned} J_n(\beta_S) &= \mathsf{E}_{\beta^\star}\Big[\Big\{\frac{\mathsf{N}(y \mid X_S\beta_S, \sigma^2 I)}{\mathsf{N}(y \mid X_{S^\star}\beta^\star_{S^\star}, \sigma^2 I)}\Big\}^{h\alpha}\Big] \\ K_n(\beta_S) &= \mathsf{E}_{\beta^\star}\big[\mathsf{N}^q(\beta_S \mid \phi\hat{\beta}_S, \sigma^2 g k_S (X_S^\top X_S)^\lambda)\big]. \end{aligned}$$

If $h\alpha < 1$, then upon taking expectation with respect to $y \sim \mathsf{N}(X_{S^*}\beta^*_{S^*}, \sigma^2 I)$, we get

$$J_n(\beta_S) = e^{-\frac{\alpha(1-h\alpha)}{2\sigma^2}\|X\beta_{S+} - X\beta^\star\|^2} \leq e^{-[\alpha(1-h\alpha)/2\sigma^2]\epsilon_n}, \quad \forall\, \beta_{S+} \in B_{\varepsilon_n}. \tag{S1.5}$$

Next, for K_n, after factoring out the non-stochastic terms in the multivariate normal density, there is an expectation of exponential quadratic form to be dealt with, i.e.,

$$\mathsf{E}_{\beta^\star}\Big[\exp\Big\{-\frac{q}{2\sigma^2 g k_S} Z\Big\}\Big], \tag{S1.6}$$

where $Z = (\beta_S - \phi\hat{\beta}_S)^\top (X_S^\top X_S)^{-\lambda}(\beta_S - \phi\hat{\beta}_S)$ and $\hat{\beta}_S$ is the least square estimator under configuration S. Let $\beta^\star_S$ denote the mean of $\hat{\beta}_S$, then $\beta^\star_S =$

$(X_S^\top X_S)^{-1} X_S^T X_{S^\star} \beta^\star_{S^\star}$. Applying a spectral decomposition on $X_S^\top X_S$ in Z, we have,

$$Z/(\sigma^2\phi^2) = \nu_S^\top \Lambda_S^{-(\lambda+1)} \nu_S,$$

where $\nu_S = \Lambda_S^{1/2}\Gamma_S^\top(\beta_S - \phi\hat{\beta}_S)/(\sigma\phi)$. Λ_S is a diagonal matrix whose diagonal elements are the corresponding eigenvalues for $X_S^\top X_S$, and Γ_S is a matrix with columns being corresponding eigenvectors. It is not difficult to show,

$$\nu_S \sim \mathsf{N}\Big(\Lambda_S^{1/2}\Gamma_S^\top(\beta_S - \phi\beta_S^\star)/(\sigma\phi), I\Big),$$

which implies $\nu_{S,i}$ are iid $\mathsf{N}(d_{S,i}^{1/2}\Gamma_{S,i}^\top(\beta_S - \phi\beta_S^\star)/(\sigma\phi), 1)$, where $d_{S,i}$ is the i^{th} eigenvalue of $X_S^\top X_S$ and $\Gamma_{S,i}$ is the i^{th} eigenvector. Hence, $\nu_{S,i}^2$ has a non-central chi-square distribution with df $= 1$ and non-centrality $\mu_{S,i} = \frac{1}{\sigma^2\phi^2}(\beta_S - \phi\beta_S^\star)^\top\Gamma_{S,i} d_{S,i}\Gamma_{S,i}^\top(\beta_S - \phi\beta_S^\star)$. By taking advantage of the independence of the $\nu_{S,i}^2$s and using the moment generating function of the non-central chi-square distribution, (S1.6) can be written as

$$\prod_{i=1}^{s}(1-2t_{S,i})^{-\frac{1}{2q}} \exp\Big\{\frac{\mu_{S,i}}{2q}\Big(\frac{1}{1-2t_{S,i}} - 1\Big)\Big\}, \tag{S1.7}$$

where $t_{S,i} = -q\phi^2 d_{S,i}^{-(1+\lambda)}/2gk_S < 0 < 1/2$. It is clear that (S1.7) is a non-decreasing function with respect to $t_{S,i}$. And when $\lambda \geq 0$, $t_{S,i} \leq t_S := -\frac{\phi^2 q}{2g}\omega(s)^{-(\lambda+1)}$. Then, for all $\beta_S \in \mathbb{R}^{|S|}$, we can obtain an upper bound for

(S1.6) by replacing $t_{S,i}$ with t_S in (S1.7),

$$U(\beta_S) = (1-2t_S)^{-\frac{s}{2q}} \exp\Big\{ \Big(\frac{2t_S}{1-2t_S} \Big) \frac{1}{2q\sigma^2\phi^2} (\beta_S - \phi\beta_S^\star)^\top (X_S^\top X_S)(\beta_S - \phi\beta_S^\star) \Big\} \tag{S1.8}$$

For $\lambda < -1$ and $-1 \leq \lambda < 0$, we can get the same expression of $U_n(\beta_S)$ with $t_S = -\phi^2 q\omega(s)/2g$ and $t_S = -\phi^2 q\omega(s)^\lambda/2g$ respectively. Therefore,

$$K_n(\beta_S) \leq (2\pi\sigma^2 g k_S)^{-\frac{s}{2}} D(S)^{-\frac{\lambda}{2}} U_n(\beta_S) \tag{S1.9}$$

Since $J_n(\beta_S)$ and $K_n(\beta_S)$ are non-negative, we upper-bound the integral in (S1.4) by

$$\int_{\mathbb{R}^{|S|}} J_n(\beta_S)^{\frac{1}{h}} K_n(\beta_S)^{\frac{1}{q}} d\beta_S \tag{S1.10}$$

By plugging (S1.9) and (S1.5) into (S1.10), and integrating out β_S , we bound $\mathsf{E}_{\beta^\star}(N_n)$ by,

$$e^{-[\alpha(1-h\alpha)/2\sigma^2]\epsilon_n} \sum_{S:|S|\leq R} \psi(|S|)^s \pi(S)$$

where function $\psi(s)$ is defined (S1.3). □

S2 Proofs of theorems

S2.1 Proof of Theorem 1

With Lemmas 1 and 2, the expectation of posterior probability of event $B_n = B_{M\varepsilon_n}$ can be bounded by,

$$\begin{aligned}\mathsf{E}_{\beta^\star}\big\{\Pi^n(B_n)\big\} &\leq e^{bs^\star - dM\varepsilon_n} \frac{\sum_{S:s\leq R}\psi(s)^s\pi(S)}{\pi(S^\star)} \\ &= e^{bs^\star - dM\varepsilon_n} \frac{\sum_{s=1}^{R}\psi(s)^s f_n(s)}{\pi(S^\star)}.\end{aligned}$$

While for the prior, if $\lambda \geq 0$, we can get,

$$\pi(S^\star) \geq \omega(s^\star)^{-\lambda/2} f_n(s^\star) \binom{p}{s^\star}^{-1}.$$

Therefore the upper bound for the posterior probability can be written as,

$$\mathsf{E}_{\beta^\star}\big\{\Pi^n(B_n)\big\} \leq e^{bs^\star - dM\varepsilon_n}\xi_n,$$

where

$$\xi_n = \frac{\omega(s^\star)^{\lambda/2}\binom{p}{s^\star}}{f_n(s^\star)} \sum_{s=1}^{R} \psi(R)^s f_n(s).$$

Taking logarithm on both sides, we get

$$\log \mathsf{E}_{\beta^\star}\big\{\Pi^n(B_n)\big\} \leq \Big(\frac{bs^\star}{\varepsilon_n} - Md + \frac{\log\xi_n}{\varepsilon_n}\Big)\varepsilon_n. \tag{S2.1}$$

We require ε_n to have a certain rate such that the upper bound for posterior probability can vanish. A preliminary requirement for ε_n is $s^\star/\varepsilon_n \to 0$, in order make $e^{bs^\star - d\varepsilon_n}$ as $o(1)$. In addition, ε_n should satisfy $\log\xi_n = O(\varepsilon_n)$.

Therefore, as $n \to \infty$, $bs^\star/\varepsilon_n \to 0$ and $\log(\xi_n)/\varepsilon_n \to K$. Thus, for any M satisfying $Md > K$, we will have

$$\log \mathsf{E}_{\beta^\star}\big\{\Pi^n(B_n)\big\} \to -\infty.$$

Next, we establish the rate for $\log \xi_n$. Under Assumptions 1 and 2, $\omega(s^\star)$ is bounded with probability 1. By Stirling's formula, we have that

$$\log \binom{p}{s^\star} \leq s^\star \log(p/s^\star)\{1+o(1)\}.$$

Given that $f_n(s) \propto c^{-s}p^{-as}$, we can also have

$$-\log f_n(s^\star) \leq s^\star \log(cs^\star) + as^\star \log(p/s^\star) = O(s^\star \log(p/s^\star)).$$

Since we have ruled out cases with extremely ill-conditioned $X_S^\top X_S$, $\omega(s)$ is bounded above by Cp^r. Thus, for the nonnegative λ case,

$$\sum_{s=1}^{R} \psi(R)^s f_n(s) \lesssim p^{R\big[r(1+\lambda)-a\big]}.$$

Therefore, when $\lambda \geq 0$, the rate of ε_n should be

$$\max\{R\big[r(1+\lambda)-a\big]\log p, s^\star \log(p/s^\star)\}.$$

The proofs for $\lambda < 0$ are similar. Therefore, the rate ε_n then can be rewritten as

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

where function q is defined in the Theorem 1 statement.

S2.2 Proof of Theorem 2

Followed by Lemma 2 and Theorem 1, we can get

$$\mathsf{E}_{\beta^\star}^{1/h}\Big[\Big\{\frac{\mathsf{N}(y \mid X_S\beta_{S+}, \sigma^2 I)}{\mathsf{N}(y \mid X_S\beta_{S^\star}, \sigma^2 I)}\Big\}^{h\alpha}\Big] = \exp\big\{-\tfrac{\alpha(1-h\alpha)}{2\sigma^2}\|X\beta_{S+} - X\beta_{S^\star}\|^2\big\} \le 1.$$

If $U_n = \{\beta \in \mathbb{R}^p : |S_\beta| \ge \rho s^\star\}$, then it is not difficult to show that,

$$\mathsf{E}_{\beta^\star}\{N_n(U_n)\} \le \sum_{s=\rho s^\star}^{R} \psi(s)^s f_n(s).$$

With the help of Lemma 1, the posterior probability of event U_n can be bounded as,

$$\mathsf{E}_{\beta^\star}\{\Pi^n(U_n)\} \le e^{cs^\star}\frac{\omega(s^\star)^{\lambda/2}\binom{p}{s^\star}}{f_n(s^\star)}\sum_{s=\rho s^\star}^{R} \psi(s)^s f_n(s) \tag{S2.1}$$

From Theorem 1, we have

$$\log\frac{\omega(s^\star)^{\lambda/2}\binom{p}{s^\star}}{f_n(s^\star)} \le (a+1+o(1))s^\star \log(p/s^\star).$$

In addition, for $\lambda \ge 0$, if $a > r(1+\lambda)$, we can get

$$\sum_{s=\rho s^\star}^{R} \psi(s)^s f_n(s) \lesssim \exp\{-\rho s^\star[a - r(1+\lambda)]\log p\}.$$

When $\lambda \ge 0$, $\rho > \rho_0 = (a+1)\{a-r(1+\lambda)\}^{-1}$, then $\sum_{s=\rho s^\star}^{R}\psi(s)^s f_n(s)$ dominates the other two terms in (S2.1). Therefore, $\mathsf{E}_{\beta^\star}\{\Pi^n(U_n)\}$ will vanish as $n \to \infty$. Similarly for $\lambda < 0$, we can get the same result if $\rho > \rho_0$.

S2.3 Proof of Theorem 3

Let $|S_{\beta-\beta^\star}|$ be the number of non-zero entries of $(\beta-\beta^\star)$. Then,

$$\|X(\beta-\beta^\star)\|^2 > n\ell(|S_{\beta-\beta^\star}|)\|\beta-\beta^\star\|^2.$$

If $a > \max\{1+\lambda, 1, 1-\lambda\}$ and $\rho > \rho_0$ in Equation (3.1) in the main text, by Theorem 2 and monotonicity of $\ell(s)$, we can get

$$\mathsf{E}_{\beta^\star}\Pi^n(\{\beta : \ell(|S_{\beta-\beta^\star}|) \geq \ell(\rho s^\star + s^\star)\}) \to 1.$$

If we set δ_n as in the theorem's statement and apply Theorem 1, we get

$$\mathsf{E}_{\beta^\star}\{\Pi^n(\beta : \|\beta-\beta^\star\|^2 > M\delta_n)\} \leq \mathsf{E}_{\beta^\star}\{\Pi^n(B_{M\delta_n\ell(\rho s^\star+s^\star)})\} \to 0,$$

as was to be shown.

S2.4 Proof of Theorem 4

We segment the proof of Theorem 4 into two parts. First, under Assumptions 1 and 3, we aim to show that, $\mathsf{E}_{\beta^\star}[\Pi^n(\{\beta : S_\beta \supset S_{\beta^\star}\})] \to 0$. Second, under the beta-min condition, we prove that $\mathsf{E}_{\beta^\star}[\Pi^n(\{\beta : S_\beta \not\supseteq S_{\beta^\star}\})] \to 0$. We only consider positive λ case here. The proofs in Parts 1 and 2 below can go through the same way when $\lambda < 0$.

Note that some of the arguments presented here refer to Martin et al. (2017), but there are some oversights in the selection consistency results presented in the published version of that paper. The version available at `arXiv:1406.7718` contains corrections of those arguments.

Part 1 Let S be any configuration containing but not equal to the true model $S^\star$, i.e. $S \supset S^\star$. Then the posterior for S can be written as,

$$\begin{aligned}\pi^n(S) &\leq \Pi^n(S)/\Pi^n(S^\star) \\ &= F_S R_S \exp\big[-\tfrac{\alpha}{2\sigma^2}\big\{y^\top(P_{S^\star} - P_S)y + (1-\phi)^2\hat\beta_S^\top Q_S^{-1}\hat\beta_S - (1-\phi)^2\hat\beta_{S^\star}^\top Q_{S^\star}^{-1}\hat\beta_{S^\star}\big\}\big] \\ &\leq F_S R_S \exp\big[\tfrac{\alpha}{2\sigma^2}\big\{y^\top(P_S - P_{S^\star})y + (1-\phi)^2\hat\beta_{S^\star}^\top Q_{S^\star}^{-1}\hat\beta_{S^\star}\big\}\big],\end{aligned}$$

where $F_S = \pi(S)/\pi(S^\star)$, $P_S = X_S(X_S^\top X_S)^{-1}X_S^\top$, Q_S is as in (S1.2), and

$$R_S = \frac{\prod_{i=1}^{s}(1+\alpha g k_S d_{S,i}^{\lambda+1})^{-\frac{1}{2}}}{\prod_{i=1}^{s^\star}(1+\alpha g k_{S^\star} d_{S^\star,i}^{\lambda+1})^{-\frac{1}{2}}}.$$

Applying Hölder's inequality with the same constants in (S1.4), $h>1$ and $q=(h-1)/h$, we get that $\mathsf{E}_{\beta^\star}\{\pi^n(S)\}$ is upper-bounded by

$$F_S R_S \mathsf{E}_{\beta^\star}^{1/h}\big[\exp\{\tfrac{\alpha h}{2\sigma^2}y^\top(P_S - P_{S^\star})y\}\big]\mathsf{E}_{\beta^\star}^{1/q}\big[\exp\{\tfrac{\alpha q(1-\phi)^2}{2\sigma^2}\hat\beta_{S^\star}^\top Q_{S^\star}^{-1}\hat\beta_{S^\star}\}\big]. \quad \text{(S2.1)}$$

First, given that $S \supset S^\star$, $(P_S - P_S^\star)$ is idempotent and $(P_S - P_S^\star)X_{S^\star} = 0$. Therefore, $y^\top(P_S - P_{S^\star})y/\sigma^2$ has a chi-square distribution, $\mathsf{ChiSq}(s-s^\star)$. If $h\alpha < 1$, using moment generating function of central chi-square, the first expectation term in (S2.1) can be written as,

$$\mathsf{E}_{\beta^\star}^{1/h}\big[\exp\{\tfrac{h\alpha}{2\sigma^2}y^\top(P_S - P_{S^\star})y\}\big] = (1-h\alpha)^{-(s-s^\star)/2h}.$$

Second, from the spectral decomposition $X_{S^\star}^\top X_{S^\star} = \Gamma_{S^\star}\Lambda_{S^\star}\Gamma_{S^\star}^\top$, if $u = \Lambda_{S^\star}^{1/2}\Gamma_{S^\star}^\top\hat{\beta}_{S^\star}/\sigma$, then,

$$\hat{\beta}_{S^\star}^\top Q_{S^\star}^{-1}\hat{\beta}_{S^\star}/\sigma^2 = u^\top\left(I_s + \alpha g k_{S^\star}\Lambda_{S^\star}^{\lambda+1}\right)^{-1}u = \sum_{i=1}^{s^\star}\frac{1}{1+\alpha g k_{S^\star} d_{S^\star,i}^{\lambda+1}}u_i^2.$$

Obviously, $u \sim \mathsf{N}(\Lambda_{S^\star}^{1/2}\Gamma_{S^\star}^\top\beta_{S^\star}^\star/\sigma, I_{s^\star})$, and this implies $u_i \overset{iid}{\sim} \mathsf{N}(d_{S^\star,i}^{1/2}\Gamma_{S^\star,i}^\top\beta_{S^\star}^\star/\sigma, 1)$. It follows that the u_i^2s are independent non-central chi-square random variables, with non-centrality parameter $\mu_i = d_{S^\star,i}^{1/2}\Gamma_{S^\star,i}^\top\beta_{S^\star}^\star/\sigma$ and degrees of freedom of 1. Using the same argument as in (S1.7)–(S1.8), we get,

$$\mathsf{E}_{\beta^\star}^{1/q}\left[\exp\{\tfrac{\alpha q(1-\phi)^2}{2\sigma^2}\hat{\beta}_{S^\star}^\top Q_{S^\star}^{-1}\hat{\beta}_{S^\star}\}\right] \le (1-t)^{-\frac{s^\star}{2q}}\exp\left\{\alpha n\|\beta_{S^\star}^\star\|^2\lambda_{\max}(S^\star)\tfrac{t}{1-t}\right\},$$

where $0 < t \lesssim (1-\phi)^2$. Since $t = o(s^\star\{n\|\beta_{S^\star}^\star\|^2\}^{-1}) \le o(1)$, by Assumption 2, we have $1 - t > \frac{1}{2}$ for large n; in addition, by Assumption 3, we have that the expression inside $\exp\{\cdot\}$ on the right-hand side above is $o(s^\star)$. Therefore, since $q^{-1} < 1-\alpha$, we can conclude that the second expectation term in (S2.1) is asymptotically upper-bounded by

$$\exp[\{\tfrac{1}{2}(1-\alpha)\log 2 + o(1)\}s^\star\}] \le \exp\{(1-\alpha)(\log 2)s^\star\}.$$

Next, it is clear that

$$R_S \le \prod_{i=1}^{s^\star}(1+\alpha g k_{S^\star} d_{S^\star,i}^{\lambda+1})^{1/2}.$$

Since $k_S d_{S,i}^{\lambda+1}$ satisfies

$$\kappa(S)^{-(\lambda+1)} \le k_S d_{S,i}^{\lambda+1} \le \kappa(S)^{\lambda+1},$$

and the lower and upper bounds are stable according to Assumption 3, it follows that R_S is upper-bounded by $e^{ms^\star}$, where $m = \frac{1}{2}\log(1+\alpha g \kappa(S^\star)^{\lambda+1})$.

For F_S defined above, we have

$$\pi(S) \leq \binom{p}{s}^{-1} \omega(s)^{\frac{\lambda}{2}} f_n(s) \quad \text{and} \quad \pi(S^\star) \geq \binom{p}{s^\star}^{-1} \omega(s^\star)^{-\frac{\lambda}{2}} f_n(s^\star).$$

Now we can bound $\mathsf{E}_{\beta^\star}\{\pi^n(S)\}$ above by,

$$e^{Gs^\star}\omega(s)^\lambda \binom{p}{s}^{-1} \binom{p}{s^\star} \big(\frac{z}{cp^a}\big)^{s-s^\star} \times O(1),$$

where $G = (1-\alpha)(\log 2) + m$ and $z > 0$ is a constant. By Theorem 2, we only need to consider S of size no more than $\rho s^\star$, where $\rho > \rho_0$ in (3.2) and $a > r(1+\lambda)$. Therefore,

$$\begin{aligned} \mathsf{E}_{\beta^\star}\{\Pi^n(\beta : S_\beta \supset S_{\beta^\star})\} &= \sum_{S \supset S^\star} \mathsf{E}_{\beta^\star}\{\pi^n(S)\} \\ &\leq e^{Gs^\star}\omega(\rho s^\star)^\lambda \sum_{s=s^\star+1}^{\rho s^\star} \frac{\binom{p-s^\star}{p-s}\binom{p}{s^\star}}{\binom{p}{s^\star}} \big(\frac{z}{cp^a}\big)^{s-s^\star} \qquad \text{(S2.2)} \end{aligned}$$

Under Assumption 3, $\omega(\rho s^\star)$ is bounded, so following the results in Martin et al. (2017), (S2.2) turns out to be,

$$\mathsf{E}_{\beta^\star}\{\Pi^n(\beta : S_\beta \supset S_{\beta^\star})\} \leq \frac{e^{Gs^\star}s^\star}{p^a} \times O(1).$$

So, if p and a are such that $p^a \gg s^\star e^{Gs^\star}$, then $\mathsf{E}\{\Pi^n(S \supset S^\star)\}$ will go to 0 as $n \to \infty$.

Part 2 Consider a configuration S satisfying $S \not\supseteq S^\star$, in Part 1, we already have (S2.1), which is

$$\mathsf{E}_{\beta^\star}\{\pi^n(S)\} \le F_S R_S \mathsf{E}_{\beta^\star}^{1/h}\big[\exp\{\tfrac{\alpha h}{2\sigma^2} y^\top (P_S - P_{S^\star})y\}\big]\mathsf{E}_{\beta^\star}^{1/q}\big[\exp\{\tfrac{\alpha q(1-\phi)^2}{2\sigma^2}\hat\beta_{S^\star}^\top Q_{S^\star}^{-1}\hat\beta_{S^\star}\}\big].$$

For the first expectation term, if we plug in $y = X_{S^\star}\beta_{S^\star} + \sigma z$, where $z \sim \mathsf{N}(0, I_n)$, then according to the results in Martin et al. (2017)

$$\begin{aligned} y^\top (P_S - P_{S^\star})y &= -\|(I-P_S)X_{S^\star}\beta^\star_{S^\star}\|^2 - 2\sigma z^\top (I-P_S)X_{S^\star}\beta^\star_{S^\star} + \sigma^2 z^\top (P_S - P_{S^\star})z \\ &\le -\|(I-P_S)X_{S^\star}\beta^\star_{S^\star}\|^2 - 2\sigma z^\top (I-P_S)X_{S^\star}\beta^\star_{S^\star} + \sigma^2 z^\top (P_S - P_{S\cap S^\star})z. \end{aligned}$$

Since $z^\top(I-P_S)X_{S^\star}\beta^\star_{S^\star}$ and $z^\top(P_S - P_{S\cap S^\star})z$ are independent, using the moment generating function of normal and chi-square distributions, we can get,

$$\begin{aligned} &\mathsf{E}_{\beta^\star}^{1/h}\big[\exp\{\tfrac{\alpha h}{2\sigma^2} y^\top (P_S - P_{S^\star})y\}\big] \\ &\quad \le (1-h\alpha)^{-\frac{1}{2h}(|S|-|S\cap S^\star|)}\exp\{-\tfrac{\alpha(1-h\alpha)}{2\sigma^2}\|(I-P_S)X_{S^\star}\beta^\star_{S^\star}\|^2\} \end{aligned} \tag{S2.3}$$

With some algebraic manipulation, we can show that

$$\|(I-P_S)X_{S^\star}\beta^\star_{S^\star}\|^2 = \|(I-P_S)X_{S^\star\cap S^c}\beta^\star_{S^\star\cap S^c}\|^2,$$

and then based on Lemma 5 of Arias-Castro and Lounici (2014), we get

$$\|(I-P_S)X_{S^\star\cap S^c}\beta^\star_{S^\star\cap S^c}\|^2 \ge n\ell(s^\star)\|\beta^\star_{S^\star\cap S^c}\|^2.$$

By the beta-min condition (3.5), it follows that $\|\beta^\star_{S^\star\cap S^c}\|^2 \geq \varrho_n^2(s^\star - |S^\star \cap S|)$, and, hence,

$$\exp\{-\tfrac{\alpha(1-h\alpha)}{2\sigma^2}\|(I-P_S)X_{S^\star}\beta^\star_{S^\star}\|^2\} \leq p^{-M(s^\star-|S^\star\cap S|)}.$$

Then following the results in Martin et al. (2017), we can get,

$$\mathsf{E}_{\beta^\star}\{\Pi^n(\beta : S_\beta \not\supseteq S_{\beta^\star})\} \lesssim e^{Gs^\star}\sum_{s=0}^{\rho s^\star}\sum_{t=1}^{\min(s,s^\star)}\frac{\binom{s^\star}{t}\binom{p-s^\star}{s-t}\binom{p}{s^\star}}{\binom{p}{s}}(cp^a)^{s^\star-s}\{(1-h\alpha)p^{-M}\}^{s^\star-t}.$$

When $s < s^\star$, we have,

$$\mathsf{E}_{\beta^\star}\{\Pi^n(\beta : S_\beta \not\supseteq S_{\beta^\star})\} \lesssim e^{Gs^\star}\sum_{s=0}^{s^\star-1}(cp^{1+a-M})^{s^\star-s} \lesssim \frac{e^{Gs^\star}}{p^{M-(a+1)}}, \quad M > a+1.$$

When $s \geq s^\star$, we have,

$$\mathsf{E}_{\beta^\star}\{\Pi^n(\beta : S_\beta \not\supseteq S_{\beta^\star})\} \lesssim e^{Gs^\star}\{(1-h\alpha)p^{1-M}\}\sum_{s=s^\star}^{\rho s^\star}(cp^{a-1})^{s^\star-s} \lesssim \frac{e^{Gs^\star}}{p^{M-(a+1)}}, \quad a > 1.$$

Furthermore, if $p^{M-(a+1)}$ satisfies $p^{M-(a+1)} \gg e^{Gs^\star}$, the two fractions above will finally go to 0 as n and p go to infinity.

S3 Choice of ϕ

Consider the following sequence

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

As claimed in Section 4.2.3 of the main text, this choice of ϕ_n satisfies the condition in Assumption 2, that is,

$$1 - \phi_n = O(s^\star\{n\|\beta^\star_{S^\star}\|^2\}^{-1}).$$

To see this, first note that

$$\mathsf{E}\|\hat{\beta}_{S^\star} - \beta^\star_{S^\star}\|^2 = \sigma^2 \mathrm{tr}\{(X_{S^*}^\top X_{S^*})^{-1}\} \simeq \sigma^2 n^{-1} s^* \lambda_{\min}^{-1}(S^*),$$

where $\lambda_{\min}(S^*)$ is the smallest eigenvalue of $n^{-1}X_{S^*}^\top X_{S^*}$, which is $O(1)$. Then it is not difficult to show that

$$1 - \phi_n \simeq \frac{s^*}{n\|\beta^*_{S^*}\|^2}.$$

Therefore,

$$n(1-\phi_n)^2\|\beta^\star_{S^\star}\|^2 \simeq s^* \frac{s^*}{n\|\beta^*_{S^*}\|^2},$$

and if $s^\star\{n\|\beta^\star_{S^\star}\|^2\}^{-1} \to 0$, we have

$$n(1-\phi_n)^2\|\beta^\star_{S^\star}\|^2 = o(s^*).$$

In practice, we cannot use ϕ_n because it depends on the true $\beta^\star$. Instead, we use an estimator, $\hat{\phi}_n$, of ϕ_n. Replacing $\mathsf{E}\|\hat{\beta}_{S^\star} - \beta^\star_{S^\star}\|^2$ and $\|\beta^\star_{S^\star}\|^2$ by $\hat{\sigma}\mathrm{tr}\{(X_{S^*}^\top X_{S^*})^{-1}\}$ and $\|\hat{\beta}^{\hat{S}}\|$ respectively, we have,

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

where $\hat{S}$ is obtained from adaptive lasso, $\hat{\sigma}$ and $\hat{\beta}_{\hat{S}}$ are from the least squares estimator given $\hat{S}$.

S4 Sensitivity analysis of choices of hyperparameters

In this section, we investigate the sensitivity of our method's performance to different choices of hyperparameters. In our ECAP model, hyperparameters λ and g are chosen based on maximum marginal likelihood method, and σ^2 is determined by adaptive lasso. Here we only consider various values of a in prior (2.1) and the upper bound of ϕ, given that although we use a James-Stein type estimator $\hat{\phi}$, in practice we still set the maximum value of ϕ to be 0.7 for "stable" performance of model selection. We will also briefly discuss α later.

Table 1: Simulation results for Cases 1–5.

Case	a	$\mathsf{P}(\hat{S} = S^\star)$	$\mathsf{P}(\hat{S} \supseteq S^\star)$	Average $\lvert\hat{S}\rvert$
1	0.05*	0.263	0.342	9.65(0.15)
	0.5	0.411	0.465	9.50(0.73)
	2	0	0	6.30(0.81)
2	0.05*	0.994	1	10.00(0)
	0.5	0.975	0.975	9.96(0.20)
	2	0.775	0.775	9.76(0.57)
3	0.05*	0.760	0.778	4.90(0.08)
	0.5	0.182	0.182	4.18(0.14)

	2	0	0	3.67(0.12)
4	0.05*	0.872	0.897	5.05(0.07)
	0.5	0.809	0.825	4.90(0.06)
	2	0.360	0.360	4.17(0.06)
5	0.05*	0.827	0.919	4.95(0.05)
	0.5	0.823	0.845	4.82(0.07)
	2	0.240	0.240	4.27(0.08)

Table 2: Simulation results for Cases 1–5.

Case	ϕ	$\mathsf{P}(\hat{S} = S^\star)$	$\mathsf{P}(\hat{S} \supseteq S^\star)$	Average $\lvert\hat{S}\rvert$
1	0.5	0.435	0.609	10.01(0.81)
	0.7*	0.263	0.342	9.65(0.15)
	0.9	0.288	0.589	10.71(1.22)
	0.99	0.285	0.679	10.57(1.26)
2	0.5	0.980	1	10.02(0.14)
	0.7*	0.994	1	10.00(0)
	0.9	0.919	1	10.05(0.23)
	0.99	0.826	1	10.59(0.86)

3	0.5	0.590	0.614	4.77(0.09)
	0.7*	0.760	0.778	4.90(0.08)
	0.9	0.692	0.781	4.96(0.10)
	0.99	0.620	0.796	4.94(0.13)
4	0.5	0.882	0.912	4.97(0.05)
	0.7*	0.861	0.940	5.05(0.07)
	0.9	0.810	0.967	5.17(0.09)
	0.99	0.844	0.969	5.13(0.13)
5	0.5	0.857	0.918	4.89(0.06)
	0.7*	0.827	0.919	4.95(0.05)
	0.9	0.752	0.950	5.04(0.10)
	0.99	0.714	0.939	5.04(0.12)

From the above tables, a take-away message we can get is that the "optimal" choice of hyperparameters can be different for different settings. In practice, e.g., in Sections 4 and 5, we tend to choose a hyperparameter that can give us stable model selections results. For example, we choose $\phi = 0.7$ instead of 0.9 or 0.99 because when ϕ is close to 1, we tend to get slow MCMC convergence or unstable SSS results. And we usually avoid very small values of a, like 0.01, given that small a can lead to convergence problem in choosing g by maximizing local marginal likelihood.

For sensitivity analysis of hyperparamter α, it is not difficult to understand that in practice, there would be little significant difference in terms of model selection performance between choice of 0.99 and 0.999, since we want our α close to 1. The main issue about α is from a theoretical perspective, which we have discussed in Sections 2 and 3.

References

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

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