The Spike-and-Slab LASSO

By V. Ročková and E.I. George

University of Chicago and University of Pennsylvania

Supplemental Material

A Proofs of Section 6

A.1 Proofs of Section 6.2.1

Throughout this section, we denote by $Q(\boldsymbol{\beta}) = -\frac{1}{2}||\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}||^2 + pen_S(\boldsymbol{\beta}\,|\theta)$ the log-posterior under the separable *SSL* penalty.

A.1.1 Proof of Theorem 6.1

Proof. Denote by $\boldsymbol{\Theta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$. Because $0 \geq Q(\boldsymbol{\beta}^0) - Q(\widehat{\boldsymbol{\beta}})$, we can write

$$0 \geq ||\boldsymbol{X}\boldsymbol{\Theta}||^2 - 2\boldsymbol{\varepsilon}'\boldsymbol{X}\boldsymbol{\Theta} + 2\log\left[\frac{\pi(\boldsymbol{\beta}_0\,|\,\theta)}{\pi(\widehat{\boldsymbol{\beta}}\,|\,\theta)}\right]. \tag{A.1}$$

Using the fact $p^\star_\theta(\widehat{\beta}_j) > c_+$ when $\widehat{\beta}_j \neq 0$, we can write

$$\begin{aligned}\log\left[\frac{\pi(\boldsymbol{\beta}_0\,|\,\theta)}{\pi(\widehat{\boldsymbol{\beta}}\,|\,\theta)}\right] &\geq -\lambda_1|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_0| + \sum_{j=1}^{p}\log\left[\frac{p^\star_\theta(\widehat{\beta}_j)}{p^\star_\theta(0)}\right] + \sum_{j=1}^{p}\log\left[\frac{p^\star_\theta(0)}{p^\star_\theta(\beta_{0j})}\right] \\ &\geq -\lambda_1|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_0| + \widehat{q}\,b + (\widehat{q}-q)\log[1/p^\star_\theta(0)],\end{aligned}$$

where $0 > b = \log c_+ > \log 0.5$ is a constant very close to 0. Because $||\boldsymbol{X}'\boldsymbol{\varepsilon}||_\infty \leq \eta\Delta$ under η-NC condition, we can use the Hölder inequality $|\boldsymbol{\alpha}'\boldsymbol{\beta}| \leq |\boldsymbol{\alpha}|_\infty|\boldsymbol{\beta}|$ to find that

$$0 \geq ||\boldsymbol{X}\boldsymbol{\Theta}||^2 - 2(\eta\Delta+\lambda_1)|\boldsymbol{\Theta}| + 2\,\widehat{q}\,b + 2(\widehat{q}-q)\log[1/p^\star_\theta(0)]. \tag{A.2}$$

From Lemma 1 we know that $\boldsymbol{\Theta}$ lives inside the cone $C(\eta;\boldsymbol{\beta}_0)$. Thus, we can use Definition 1 to find that $||\boldsymbol{X}\boldsymbol{\Theta}||^2 \geq c(\eta;\boldsymbol{\beta}_0)^2||\boldsymbol{\Theta}||^2\,||\boldsymbol{X}||^2$. Denote by $c = c(\eta;\boldsymbol{\beta}_0)$. Using the fact $|\boldsymbol{\Theta}| \leq ||\boldsymbol{\Theta}||\,||\boldsymbol{\Theta}||_0^{1/2}$, we have

$$0 \geq c^2||\boldsymbol{\Theta}||^2||\boldsymbol{X}||^2 - 2(\eta\Delta+\lambda_1)||\boldsymbol{\Theta}||\,||\boldsymbol{\Theta}||_0^{1/2} + 2\,\widehat{q}\,b + 2(\widehat{q}-q)\log[1/p^\star_\theta(0)],$$

which is equivalent to writing

$$\left[c||\boldsymbol{\Theta}||\,||\boldsymbol{X}|| - \frac{(\eta\Delta+\lambda_1)}{c||\boldsymbol{X}||}||\boldsymbol{\Theta}||_0^{1/2}\right]^2 - \frac{(\eta\Delta+\lambda_1)^2}{c^2||\boldsymbol{X}||^2}||\boldsymbol{\Theta}||_0 + 2\,\widehat{q} + 2(\widehat{q}-q)\log[1/p^\star_\theta(0)] \leq 0.$$

This yields

$$(\widehat{q}-q)\log[1/p^{\star}_{\theta}(0)]+\widehat{q}\,b\leq\frac{(\eta\Delta+\lambda_1)^2}{2c^2||\boldsymbol{X}||^2}||\boldsymbol{\Theta}||_0.$$

By noting $||\boldsymbol{\Theta}||_0\leq\widehat{q}+q$ and $||\boldsymbol{X}||^2=n$, we can write

$$\widehat{q}\leq q\left(1+\frac{2A-b}{B+b-A}\right),$$

where $A=\frac{(\eta\Delta+\lambda_1)^2}{2c^2||\boldsymbol{X}||^2}$ and $B=\log[1/p^{\star}_{\theta}(0)]$. Assume, for simplicity, that $(1-\theta)/\theta=C_1p^a$, $\lambda_0=C_2p^d$ and $\lambda_1<4\sqrt{n\log p}<4p$ we have $B=\log\left(1+\frac{1-\theta}{\theta}\frac{\lambda_0}{\lambda_1}\right)>\log(C_1\,C_2/4)+(a+d-1)\log p$. With $C_1C_2/4>0$ we obtain $\lambda_1<4\sqrt{n\log p}<4\sqrt{n\,B/(a+d-1)}$ and $\frac{A}{B}<\left(\frac{\eta}{c}+\frac{(\eta+1)2\sqrt{2}}{c\sqrt{a+d-1}}\right)^2\equiv D$. We can then write $\widehat{q}\leq q\left(1+M\frac{D}{1-D}\right)$. □

A.1.2 Proof of Theorem 6.2

Proof. With $\boldsymbol{\Theta}=\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0$ and by noting $\log\left[\frac{\pi(\boldsymbol{\beta}_0\,|\,\theta)}{\pi(\boldsymbol{\beta}\,|\,\theta)}\right]>-\lambda_1|\boldsymbol{\Theta}|+q\log p^{\star}_{\theta}(0)$ we can write

$$0\geq||\boldsymbol{X}\boldsymbol{\Theta}||^2-2(\eta\Delta+\lambda_1)|\boldsymbol{\Theta}|+2\,q\log[p^{\star}_{\theta}(0)],\tag{A.3}$$

where Δ is the selection threshold. From Theorem 6.1 we have $||\boldsymbol{\Theta}||_0\leq(K+1)\,q$ under the η-NC condition. Denote by $\phi=\phi[C(\eta;\boldsymbol{\beta}_0)]$. From the definition of the compatibility number ϕ, and using $4uv\leq u^2+4v^2$, we find that

$$\begin{aligned}2(\eta\Delta+\lambda_1)|\boldsymbol{\Theta}|&\leq 3(\eta\Delta+\lambda_1)\frac{||\boldsymbol{X}\boldsymbol{\Theta}||\sqrt{(K+1)\,q}}{||\boldsymbol{X}||\phi}-(\eta\Delta+\lambda_1)|\boldsymbol{\Theta}|\\&\leq\frac{||\boldsymbol{X}\boldsymbol{\Theta}||^2}{2}+\frac{5(K+1)q(\eta\Delta+\lambda_1)^2}{||\boldsymbol{X}||^2\phi^2}-(\eta\Delta+\lambda_1)|\boldsymbol{\Theta}|.\end{aligned}$$

Thus it follows from (A.3) that

$$\frac{1}{2}||\boldsymbol{X}\boldsymbol{\Theta}||^2+(\eta\Delta+\lambda_1)|\boldsymbol{\Theta}|\leq\frac{5(K+1)q(\eta\Delta+\lambda_1)^2}{||\boldsymbol{X}||^2\phi^2}+2\,q\log[1/p^{\star}_{\theta}(0)].\tag{A.4}$$

With $(1-\theta)/\theta=C_1p^a$, $\lambda_0=C_2p^d$ and $\sqrt{n}/p<\lambda_1<4\sqrt{n\log p}$ we have $(\eta\Delta+\lambda_1)<C_3\eta\sqrt{n\log p}$ and $\log[1/p^{\star}_{\theta}(0)]<C_4\log p$. With $||\boldsymbol{X}||^2=n$, the first two statements of the theorem follow directly from (A.4). Let $c=c(\eta,\boldsymbol{\beta}_0)$ be the minimal restricted eigenvalue. Then the last statement is obtained from $||\boldsymbol{X}\boldsymbol{\Theta}||>c||\boldsymbol{X}||||\boldsymbol{\Theta}||$. □

A.2 Proofs of Section 6.2.2

The construction of the proof follows Castillo et al. (2015), where suitable modifications are required when using the notion of generalized dimensionality. Before proceeding, we

need to introduce some more notation. Let

$$\Lambda_{n,\boldsymbol{\beta},\boldsymbol{\beta}_0} = \mathrm{e}^{-\frac{1}{2}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2+(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_0)'\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}$$

and

$$\Pi(\boldsymbol{\beta} \,|\, \theta) = \prod_{i=1}^{p} [\theta\psi_1(\beta_i) + (1-\theta)\psi_0(\beta_i)].$$

Throughout this section we will denote by $\bar{\lambda} = 2\sqrt{n \log p}$ the universal threshold. The rates in this section will be expressed in terms of slightly different compatibility and minimal eigenvalue numbers. Following Castillo (2015), for $S \subset \{1, \ldots, p\}$, we define: the compatibility number $\widetilde{\phi}(S)$ of a model S by

$$\widetilde{\phi}(S) = \inf \left\{ \frac{||\boldsymbol{X}\boldsymbol{\beta}||\,|S|^{1/2}}{||\boldsymbol{X}||\,|\boldsymbol{\beta}_S|} : |\boldsymbol{\beta}_{S^c}| \leq 5|\boldsymbol{\beta}_S|, \boldsymbol{\beta}_S \neq 0 \right\}, \tag{A.5}$$

the compatibility in vectors of generalized dimension s by

$$\bar{\phi}(s) = \inf \left\{ \frac{||\boldsymbol{X}\boldsymbol{\beta}||\,s^{1/2}}{||\boldsymbol{X}||\,|\boldsymbol{\beta}|} : 0 < |\boldsymbol{\gamma}(\boldsymbol{\beta})| \leq s \right\} \tag{A.6}$$

and the minimal eigenvalue restricted to vectors $\boldsymbol{\beta}$ of generalized dimensionality at most s by

$$\bar{c}(s) = \inf \left\{ \frac{||\boldsymbol{X}\boldsymbol{\beta}||}{||\boldsymbol{X}||\,||\boldsymbol{\beta}||} : 0 < |\boldsymbol{\gamma}(\boldsymbol{\beta})| \leq s \right\} \tag{A.7}$$

For $S \subset \{1, \ldots, p\}$, let $\boldsymbol{\beta}_S \in \mathbb{R}^p$ be a subset of $\boldsymbol{\beta}$ with coordinates in S. Denote by $\Pi_S(\boldsymbol{\beta} \,|\, \theta)$ the marginal prior confined to coordinates in S. Denote by δ the intersection point between *SSL* densities, by $\pi \equiv \mathsf{P}(|\beta_1| \leq \delta)$ and by $\pi(s \,|\, \theta) = \binom{p}{s}\pi^s(1-\pi)^{p-s} = \mathsf{P}[|\boldsymbol{\gamma}(\boldsymbol{\beta})| = s \,|\, \theta]$ the prior distribution on the effective dimensionality. By assumption, we have $||\boldsymbol{X}||^2 = \max_{1 \leq i \leq p} ||\boldsymbol{X}_i||^2 = n$.

We will need the following analogue of Lemma 2 of Castillo et al. (2015) for the separable *SSL* prior.

Lemma A.1. *Assume $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ has a support $S_0 \subset \{1, \ldots, p\}$, where $|S_0| = q$. Assume $\lambda_0 = (1-\theta)/\theta = Cp^a$, where $a \geq 2$ and $C > 0$, and $\sqrt{n}/p < \lambda_1 \leq 4\sqrt{n \log p}$. Assume $p > n$. Then*

$$\int \Lambda_{n,\boldsymbol{\beta},\boldsymbol{\beta}_0} \Pi(\boldsymbol{\beta} \,|\, \theta) \mathrm{d}\,\boldsymbol{\beta} \geq \frac{\pi(q \,|\, \theta)}{p^{2\,q}} \mathrm{e}^{-1-D-\lambda_1|\boldsymbol{\beta}|_1},$$

where $D > 0$.

Proof. Denote by $g(\boldsymbol{\beta}) = \mathrm{e}^{-||\boldsymbol{X}\boldsymbol{\beta}||^2+(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_0)'\boldsymbol{X}\boldsymbol{\beta}}$. Using the fact $||\boldsymbol{X}\boldsymbol{\beta}||^2 \leq 2||\boldsymbol{X}\boldsymbol{\beta}_{S_0}||^2 + 2||\boldsymbol{X}\boldsymbol{\beta}_{S_0^c}||^2$, we can write

$$\Lambda_{n,\boldsymbol{\beta},\boldsymbol{\beta}_0} > g(\boldsymbol{\beta}_{S_0^c})g(\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0}).$$

By the Jensen's inequality we have

$$\int g(\boldsymbol{\beta})\Pi(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta} \geq \int \mathrm{e}^{-||\boldsymbol{X}\boldsymbol{\beta}||^2}\Pi(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta}.$$

Conditionally on θ, the SSL prior is separable, implying $\Pi(\boldsymbol{\beta}\,|\,\theta) = \Pi_{S_0}(\boldsymbol{\beta}\,|\,\theta)\Pi_{S_0^c}(\boldsymbol{\beta}\,|\,\theta)$. Changing variables $\boldsymbol{b} \to (\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ and noting $\Pi_{S_0}(\boldsymbol{\beta}\,|\,\theta) > \theta^q \left(\frac{\lambda_1}{2}\right)^q \mathrm{e}^{-\lambda_1|\boldsymbol{\beta}_{S_0}|}$, we can write

$$\int \Lambda_{n,\boldsymbol{\beta},\boldsymbol{\beta}_0}\Pi(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta} > \int \mathrm{e}^{-||\boldsymbol{X}\boldsymbol{\beta}_{S_0^c}||^2}\Pi_{S_0^c}(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta}_{S_0^c} \tag{A.8}$$

$$\times\, \theta^q \mathrm{e}^{-\lambda_1|\boldsymbol{\beta}_0|} \int \mathrm{e}^{-||\boldsymbol{X}\boldsymbol{b}_{S_0}||^2} \left(\frac{\lambda_1}{2}\right)^q \mathrm{e}^{-\lambda_1|\boldsymbol{b}_{S_0}|}\mathrm{d}\,\boldsymbol{b}_{S_0}. \tag{A.9}$$

To simplify the integral in (A.9) we use arguments of Castillo (2015) in the proof of Lemma 2. Under the assumption $||\boldsymbol{X}||/p < \lambda_1 < 4||\boldsymbol{X}||\sqrt{\log p}$, we obtain

$$\int \mathrm{e}^{-||\boldsymbol{X}\boldsymbol{b}_{S_0}||^2} \left(\frac{\lambda_1}{2}\right)^q \mathrm{e}^{-\lambda_1|\boldsymbol{b}_{S_0}|}\mathrm{d}\,\boldsymbol{b}_{S_0} > \mathrm{e}^{-1} \left(\frac{\lambda_1}{||\boldsymbol{X}||}\right)^q \frac{\mathrm{e}^{-\lambda_1/||\boldsymbol{X}||}}{q!} > \frac{\mathrm{e}^{-1}}{p^q q!} \tag{A.10}$$

To simplify the integral in (A.8), we use $||\boldsymbol{X}\boldsymbol{\beta}|| \leq ||\boldsymbol{X}||\,|\boldsymbol{\beta}|$ to find that

$$\int \mathrm{e}^{-||\boldsymbol{X}\boldsymbol{\beta}_{S_0^c}||^2}\Pi_{S_0^c}(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta}_{S_0^c} > \int_{|\beta_i|\leq\delta; i\notin S_0} \mathrm{e}^{-||\boldsymbol{X}||^2|\boldsymbol{\beta}_{S_0^c}|^2}\Pi_{S_0^c}(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta}_{S_0^c} \tag{A.11}$$

$$\geq \mathrm{e}^{-(p-q)^2||\boldsymbol{X}||^2\delta^2}\mathsf{P}(|\beta_1| \leq \delta)^{p-q}. \tag{A.12}$$

Combining (A.10) and (A.12) and noting $\theta > \pi \equiv \mathsf{P}(|\beta_1| > \delta\,|\,\theta)$, we obtain

$$\int \Lambda_{n,\boldsymbol{\beta},\boldsymbol{\beta}_0}\Pi(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta} > \mathrm{e}^{-(p-q)^2||\boldsymbol{X}||^2\delta^2}\mathrm{e}^{-1-\lambda_1|\boldsymbol{\beta}_0|}\pi^q(1-\pi)^{p-q}\frac{1}{p^q q!}.$$

Recall that $\pi(q\,|\,\theta) = \binom{p}{q}\pi^q(1-\pi)^{p-q}$ is the prior probability of the effective dimensionality q. Since $\binom{p}{q}q! \leq p^q$, we can write

$$\int \Lambda_{n,\boldsymbol{\beta},\boldsymbol{\beta}_0}\Pi(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta} > \mathrm{e}^{-(p-q)^2||\boldsymbol{X}||^2\delta^2}\mathrm{e}^{-1-\lambda_1|\boldsymbol{\beta}_0|}\frac{\pi(q\,|\,\theta)}{p^{2\,q}}.$$

Using the fact $\delta = \frac{1}{\lambda_0-\lambda_1}\log[1/p^\star(0) - 1]$, we obtain

$$(p-q)^2||\boldsymbol{X}||^2\delta^2 = \frac{(p-q)^2||\boldsymbol{X}||^2}{(\lambda_0-\lambda_1)^2}\log^2[1/p^\star(0) - 1] \tag{A.13}$$

Since $||\boldsymbol{X}|| = \sqrt{n} < \sqrt{p}$ we have $\lambda_1 \leq 4||\boldsymbol{X}||\sqrt{\log p} < 4p$. Because $\lambda_0 \succeq p^d$ with $d \geq 2$, we have $\frac{(p-q)^2}{\lambda_0-\lambda_1} \preceq \frac{1}{p^{d-2}}$. Because $(1-\theta)/\theta \sim p^a$, we have

$$\log^2[1/p^\star(0) - 1] \preceq \log^2 p + \log^2 \lambda_0.$$

Therefore, with $p > n$ and $\lambda_0 \succeq p^d$ with $d \geq 2$ we obtain

$$(p-q)^2||\boldsymbol{X}||^2\delta^2 \preceq \frac{n\,(\log^2 p + \log^2 \lambda_0)}{p^{d-2}\lambda_0} \to 0. \quad \square$$

A.2.1 Proof of Theorem 6.3

Proof. Denote by $\mathcal{B} = \{\boldsymbol{\beta} : |\boldsymbol{\gamma}(\boldsymbol{\beta})| > R\}$. Then $\mathsf{E}_{\boldsymbol{\beta}_0}\mathsf{P}(\mathcal{B}\,|\,\boldsymbol{Y},\theta) \leq \mathsf{E}_{\boldsymbol{\beta}_0}\mathsf{P}(\mathcal{B}\,|\,\boldsymbol{Y},\theta)\mathbb{I}_{\tau_0} + \frac{2}{p}$, where $\tau_0 = \{||\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0)||_\infty \leq \bar{\lambda}\}$ and $\bar{\lambda} = 2\sqrt{n\log p}$. Then

$$\mathsf{P}(\mathcal{B}\,|\boldsymbol{Y},\theta) = \frac{\int_{\mathcal{B}} \Lambda_{n,\boldsymbol{\beta},\boldsymbol{\beta}_0}\Pi(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta}}{\int \Lambda_{n,\boldsymbol{\beta},\boldsymbol{\beta}_0}\Pi(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta}} \leq A\mathrm{e}^{\lambda_1|\boldsymbol{\beta}_0|}\int_{\mathcal{B}} \mathrm{e}^{-\frac{1}{2}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2 + (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0)'\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}\Pi(\boldsymbol{\beta}\,|\theta)\mathrm{d}\,\boldsymbol{\beta}, \tag{A.14}$$

where $A = \frac{p^{2\,q}}{\pi(q\,|\,\theta)}\mathrm{e}^{1+D}$. Similarly as in the proof of Theorem 12 of Castillo et al. (2015), we use Hölder's inequality to obtain on τ_0

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0)'\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \leq \bar{\lambda}|\boldsymbol{\beta} - \boldsymbol{\beta}_0|. \tag{A.15}$$

Therefore, the expectation under $\boldsymbol{\beta}_0$ of the integrand satisfies

$$\mathrm{e}^{-\frac{1}{2}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2}\mathsf{E}_{\boldsymbol{\beta}_0}\left[\mathrm{e}^{\left(1-\frac{\lambda_1}{2\,\bar{\lambda}}\right)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}\mathbb{I}_{\tau_0}\right]\mathrm{e}^{\frac{\lambda_1}{2}|\boldsymbol{\beta}-\boldsymbol{\beta}_0|} \tag{A.16}$$

$$\leq \mathrm{e}^{-\frac{1}{2}\left[1-\left(1-\frac{\lambda_1}{2\,\bar{\lambda}}\right)^2\right]||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2}\mathrm{e}^{\frac{\lambda_1}{2}|\boldsymbol{\beta}-\boldsymbol{\beta}_0|} \tag{A.17}$$

$$\leq \mathrm{e}^{-\frac{\lambda_1}{4\bar{\lambda}}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2}\mathrm{e}^{\frac{\lambda_1}{2}|\boldsymbol{\beta}-\boldsymbol{\beta}_0|}, \tag{A.18}$$

where we used $\lambda_1 \leq 2\bar{\lambda}$ and invoked the expectation of a log-normally distributed r.v. Thus,

$$\mathsf{E}_{\boldsymbol{\beta}_0}\mathsf{P}(\mathcal{B}\,|\,\boldsymbol{Y},\theta)\,\mathbb{I}_{\tau_0} \leq A\,\mathrm{e}^{\lambda_1|\boldsymbol{\beta}_0|}\int_B \mathrm{e}^{-\frac{\lambda_1}{4\bar{\lambda}}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2}\mathrm{e}^{\frac{\lambda_1}{2}|\boldsymbol{\beta}-\boldsymbol{\beta}_0|}\mathrm{d}\,\Pi(\boldsymbol{\beta}\,|\,\theta). \tag{A.19}$$

Now, when $5|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_0| \leq |\boldsymbol{\beta}_{S_0^c}|$, then

$$|\boldsymbol{\beta}_0| + \frac{1}{2}|\boldsymbol{\beta} - \boldsymbol{\beta}_0| \leq |\boldsymbol{\beta}_{S_0}| + \frac{5}{4}|\boldsymbol{\beta}_S - \boldsymbol{\beta}_0| + \frac{3}{4}|\boldsymbol{\beta}_{S_0^c}| - \frac{1}{4}|\boldsymbol{\beta} - \boldsymbol{\beta}_0| < -\frac{1}{4}|\boldsymbol{\beta} - \boldsymbol{\beta}_0| + |\boldsymbol{\beta}| \tag{A.20}$$

$$< -\frac{1}{4}|\boldsymbol{\beta} - \boldsymbol{\beta}_0| + |\boldsymbol{\beta}| + \frac{1}{4\bar{\lambda}}||\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)||^2 + 2\frac{\bar{\lambda}|S_0|}{||\boldsymbol{X}||^2\widetilde{\phi}(S_0)^2}. \tag{A.21}$$

When $5|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_0| > |\boldsymbol{\beta}_{S_0^c}|$, we use the definition of the compatibility number to find that

$$|\boldsymbol{\beta}_{S_0}| + \frac{5}{4}|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_0| + \frac{3}{4}|\boldsymbol{\beta}_{S_0^c}| - \frac{1}{4}|\boldsymbol{\beta} - \boldsymbol{\beta}_0| < -\frac{1}{4}|\boldsymbol{\beta} - \boldsymbol{\beta}_0| + |\boldsymbol{\beta}| + \frac{5}{4}\frac{||\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)||\,|S_0|^{1/2}}{||\boldsymbol{X}||\widetilde{\phi}(S_0)}.$$

Invoking the inequality $4uv \leq u^2 + 4v^2$, we can bound the last display from above by

$$-\frac{1}{4}|\boldsymbol{\beta} - \boldsymbol{\beta}_0| + |\boldsymbol{\beta}| + \frac{1}{4\bar{\lambda}}||\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)||^2 + 2\frac{\bar{\lambda}\,q}{||\boldsymbol{X}||^2\widetilde{\phi}(S_0)^2}.$$

Thus (A.19) can be bounded by

$$A\,\mathrm{e}^{\frac{2\lambda_1\bar{\lambda}\,q}{||\boldsymbol{X}||^2\widetilde{\phi}(S_0)^2}}\int_{\mathcal{B}}\mathrm{e}^{\lambda_1|\boldsymbol{\beta}|-\frac{\lambda_1}{4}|\boldsymbol{\beta}-\boldsymbol{\beta}_0|}\Pi(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta}.$$

Note that $\pi(\beta\,|\,\theta) < \theta\lambda_1\mathrm{e}^{-\lambda_1|\beta|}$ when $|\beta| > \delta$. For $\mathcal{B} = \{\boldsymbol{\beta} : |\boldsymbol{\gamma}(\boldsymbol{\beta})| > R\}$, we can write

$$\int_{\mathcal{B}}\mathrm{e}^{\lambda_1|\boldsymbol{\beta}|-\frac{\lambda_1}{4}|\boldsymbol{\beta}-\boldsymbol{\beta}_0|}\Pi(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta} \le \sum_{S:|S|>R}\theta^{|S|}\lambda_1^{|S|}\int_{|\beta_i|>\delta;i\in S}\mathrm{e}^{-\frac{\lambda_1}{4}|\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S}|}\mathrm{d}\,\boldsymbol{\beta}_S \tag{A.22}$$

$$\times\int_{|\beta_i|\le\delta;i\in S^c}\mathrm{e}^{\lambda_1|\boldsymbol{\beta}_{S^c}|-\frac{\lambda_1}{4}|\boldsymbol{\beta}_{S^c}-\boldsymbol{\beta}_{0S^c}|}\Pi_{S^c}(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta}_{S^c} \tag{A.23}$$

$$<\sum_{k=R+1}^{p}\binom{p}{k}(8\,\theta)^k\mathrm{e}^{\lambda_1\delta(p-k)}(1-\pi)^{p-k}. \tag{A.24}$$

Recall that $\pi \equiv \mathsf{P}(|\beta_1| > \delta\,|\,\theta) = \theta\,\mathrm{e}^{-\delta\lambda_1}\left(1+\frac{\lambda_1}{\lambda_0}\right) < \theta$. Because $\theta < \pi\mathrm{e}^{\delta\lambda_1}$, we can bound the last display by

$$\mathrm{e}^{\lambda_1 p\,\delta}\sum_{k=R+1}^{p}8^k\binom{p}{k}\pi^k(1-\pi)^{p-k} = \mathrm{e}^{\lambda_1 p\,\delta}\sum_{k=R+1}^{p}8^k\pi(k\,|\,\theta).$$

Because $\pi/(1-\pi) < \theta/(1-\theta) \preceq 1/p^a$ for $a \ge 2$, we have

$$\pi(k\,|\,\theta) \le \left(\frac{1}{p}\right)^{a-1}\pi(k-1\,|\,\theta) \quad \text{for} \quad k \ge 1.$$

Thereby, we can write for $R > q$

$$\mathrm{e}^{\lambda_1 p\,\delta}\sum_{k=R+1}^{p}8^k\pi(k\,|\,\theta) < \mathrm{e}^{\lambda_1 p\,\delta}\,8^q\,\pi(q\,|\,\theta)\left(\frac{8}{p^{a-1}}\right)^{R+1-q}\sum_{k=0}^{\infty}\left(\frac{8}{p^{a-1}}\right)^k.$$

With $(1-\theta)/\theta \sim p^a$ and $\lambda_0 \succeq p^d$ and $||\boldsymbol{X}||^2 = n$, we have $\lambda_1\delta p \preceq 1$. Altogether

$$\mathsf{P}(\mathcal{B}\,|\,\boldsymbol{Y},\theta) \preceq \mathrm{e}^{2q\log p+\lambda_1 p\,\delta+\frac{2\lambda_1\bar{\lambda}\,q}{n\widetilde{\phi}(S_0)^2}}\left(\frac{8}{p^{a-1}}\right)^{R+1-q}+\frac{2}{p}$$

$$\preceq \mathrm{e}^{(R+1-q)\log 8+2q\log p[1+4\lambda_1/(\bar{\lambda}\widetilde{\phi}(S_0)^2)]-(R+1-q)(a-1)\log p}+\frac{2}{p}.$$

The right side of the above display goes to zero when $R > q\left[1+\frac{M}{a-1}\left(1+\frac{4\lambda_1}{\bar{\lambda}\widetilde{\phi}(S_0)^2}\right)\right]$ for some $M > 2$. □

A.2.2 Proof of Theorem 6.4

Proof. By Theorem 6.3, the posterior distribution is asymptotically supported on the event $E = \{\boldsymbol{\beta} : |\boldsymbol{\gamma}(\boldsymbol{\beta})| \le q(1+K)|\}$, where $K = \frac{M}{a-1}\left(1+\frac{4\lambda_1}{\bar{\lambda}\widetilde{\phi}(S_0)^2}\right)$. Thus, we confine attention to

$E^\star = E \cap \tau_0$, where τ_0 was defined in the proof of Theorem 6.3. From (A.14) and (A.15), we can see that

$$\Pi(\mathcal{B} \,|\, \boldsymbol{Y}, \theta)\mathbb{I}_{\tau_0} \leq \frac{p^{2q}\mathrm{e}^{1+D}}{\pi(q \,|\, \theta)} \int_{\mathcal{B}} \mathrm{e}^{-\frac{1}{2}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2 + 3\bar{\lambda}|\boldsymbol{\beta}-\boldsymbol{\beta}_0| + \lambda_1|\boldsymbol{\beta}|} \Pi(\boldsymbol{\beta} \,|\, \theta) \mathrm{d}\,\boldsymbol{\beta}. \tag{A.25}$$

We now use the definition of the compatibility number in vectors of generalized dimensionality (A.6). With the inequality $4uv \leq u^2 + 4v^2$, we can then write

$$\begin{aligned}(4-1)\bar{\lambda}|\boldsymbol{\beta}-\boldsymbol{\beta}_0| &\leq \frac{4\bar{\lambda}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||\,|\boldsymbol{\gamma}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|^{1/2}}{\sqrt{n}\,\bar{\phi}(|\boldsymbol{\gamma}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|)} - \bar{\lambda}|\boldsymbol{\beta}-\boldsymbol{\beta}_0| \\ &\leq \frac{1}{4}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2 + \frac{16\bar{\lambda}^2|\boldsymbol{\gamma}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|}{n\,\bar{\phi}(|\boldsymbol{\gamma}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|)^2} - \bar{\lambda}|\boldsymbol{\beta}-\boldsymbol{\beta}_0|\end{aligned}$$

Thus,

$$\Pi(\mathcal{B} \,|\, \boldsymbol{Y}, \theta)\mathbb{I}_{\tau_0} \leq \frac{p^{2q}\mathrm{e}^{1+D}}{\pi(q \,|\, \theta)} \mathrm{e}^{\frac{16\bar{\lambda}q(2+K)}{n[\bar{\phi}(2q+Kq)]^2}} \int_{\mathcal{B}} \mathrm{e}^{-\frac{1}{4}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2 - \bar{\lambda}|\boldsymbol{\beta}-\boldsymbol{\beta}_0| + \lambda_1|\boldsymbol{\beta}|} \Pi(\boldsymbol{\beta} \,|\, \theta) \mathrm{d}\,\boldsymbol{\beta}.$$

Denote now $\mathcal{B} = \{\boldsymbol{\beta} \in E^\star : ||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|| > R\}$. Then

$$\begin{aligned}\Pi(\mathcal{B} \,|\, \boldsymbol{Y}, \theta)\mathbb{I}_{\tau_0} &\leq \frac{p^{2q}\mathrm{e}^{1+D}}{\pi(q \,|\, \theta)} \mathrm{e}^{\frac{16\bar{\lambda}q(2+K)}{n[\bar{\phi}(2q+Kq)]^2}} \mathrm{e}^{-\frac{R^2}{4}} \int_{\mathcal{B}} \mathrm{e}^{-\bar{\lambda}|\boldsymbol{\beta}-\boldsymbol{\beta}_0| + \lambda_1|\boldsymbol{\beta}|} \Pi(\boldsymbol{\beta} \,|\, \theta) \mathrm{d}\,\boldsymbol{\beta} \\ &\leq \frac{p^{2q}\mathrm{e}^{1+D}}{\pi(q \,|\, \theta)} \mathrm{e}^{\frac{16\bar{\lambda}q(1+K)}{n[\bar{\phi}(2q+Kq)]^2}} \mathrm{e}^{-\frac{R^2}{4}} \mathrm{e}^{\lambda_1\delta p} \sum_{k=0}^{p} 8^k \pi(k \,|\, \theta)\end{aligned}$$

Now, because $\pi > \theta \mathrm{e}^{-\delta\lambda_1}$ and $\theta \sim 1/p^a$ we can write

$$\pi(q \,|\, \theta) \succeq \pi(q-1 \,|\, \theta) \frac{\mathrm{e}^{-\delta\lambda_1}}{p^a} \succeq \pi(q-1 \,|\, \theta) \frac{C_2}{p^a},$$

where we used the fact $\mathrm{e}^{-\delta\lambda_1} > C_2$. Thus, $\pi(q \,|\, \theta) \succeq C_2^q / p^{a\,q} \pi(0 \,|\, \theta)$. Thereby,

$$\Pi(\mathcal{B} \,|\, \boldsymbol{Y}, \theta)\mathbb{I}_{\tau_0} \leq C_2^{-q} p^{q(2+a)} \mathrm{e}^{1+D} \mathrm{e}^{\frac{16\bar{\lambda}^2 q(2+K)}{n[\bar{\phi}(2q+Kq)]^2}} \mathrm{e}^{-\frac{R^2}{4}} \mathrm{e}^{\lambda_1\delta p} \sum_{k=0}^{p} \left(\frac{8}{p^{a-1}}\right)^k.$$

This quantity will tend to zero for

$$R^2 \succeq 4q(2+a)\log p + \frac{16\bar{\lambda}^2 q(2+K)}{n[\bar{\phi}(2q+Kq)]^2} \succeq \frac{q(2+K)\log p}{[\bar{\phi}(2q+Kq)]^2}.$$

This proofs the first assertion. The second assertion follows from

$$\bar{\lambda}|\boldsymbol{\beta}-\boldsymbol{\beta}_0| \leq ||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2 + \frac{\bar{\lambda}^2 q(2+K)}{2n\bar{\phi}(2q+Kq)^2}$$

and the last one from the definition of a minimal eigenvalue restricted to $\boldsymbol{\beta}$ of generalized dimensionality at most s in (A.7), which yields $||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|| > \bar{c}(2q+Kq)||\boldsymbol{X}||\,||\boldsymbol{\beta}-\boldsymbol{\beta}_0||$.

□

B Proof of Lemma 6.2

Proof. Denote by $\boldsymbol{e}_j$ the j^{th} canonical vector. The global optimality of $\widehat{\boldsymbol{\beta}}$ yields

$$-||\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}||/2 + pen_{NS}(\widehat{\boldsymbol{\beta}}) \geq -||\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - t\boldsymbol{X}_j||/2 + pen_{NS}(\widehat{\boldsymbol{\beta}} + t\boldsymbol{e}_j)$$

for any $t \in \mathbb{R}$. Because $||\boldsymbol{X}_j||^2 = n$, this is equivalent to

$$t\boldsymbol{X}_j'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \leq t^2 n/2 + \log\left[\frac{\pi(\widehat{\boldsymbol{\beta}})}{\pi(\widehat{\boldsymbol{\beta}} + t\boldsymbol{e}_j)}\right] < nt^2/2 - \widetilde{\rho}(t; \widehat{\boldsymbol{\beta}}_{\setminus j}),$$

where we used the definition of the conditional singleton (3.10) together with (3.8) and (3.9). The statement of the lemma follows from the definition Δ_j. □

C Simulation Study: Initialization

In order to assess the sensitivity of the *Spike-and-Slab LASSO* to initialization, we repeated the simulation study from Section 5 (with the correlated block design), considering different starting vectors $\boldsymbol{\beta}^0$ as well as different spike penalty sequences $\{\lambda_0^1, \ldots, \lambda_0^L\}$. In particular, we generated 100 datasets and initialized *SSL* at (a) the zero vector, (b) a random vector from $\mathcal{N}_p(0, I)$ and (c) a random vector from $\mathcal{N}_p(0, 10 \times I)$. Different random starting vectors were used for each dataset. The results are summarized in Table 1. With a fine grid of λ_0 values (e.g. choosing $\lambda_0 \in \{1, 2, \ldots, 100\}$), we do not observe any sensitivity to initialization! With a crude ladder of λ_0 values (e.g. choosing $\lambda_0 \in \{10, 20, 30, \ldots, 100\}$), there is indeed sensitivity to initialization. However, we observed that the zero starting vector performed the best regardless of the choice of λ_0 values.

D Further Implementations

D.1 Implementation via EMVS

The coordinate-wise optimization based on the univariate soft-thresholding operator (as explained in Section 4)

$$S(z, \lambda) = \frac{1}{n}(|z| - \lambda)_+ \text{sign}(z).$$

resembles the LLA algorithm (**??**), which iterates over joint updates

$$\boldsymbol{\beta}^{(k+1)} = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ -\frac{1}{2}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 - \sum_{j=1}^{p} \lambda_\theta^\star(\beta_j^{(k)})|\beta_j| \right\}. \tag{D.1}$$

λ_0	λ_1	θ	MSE	FDR	FNR	$\widehat{q}$	TRUE	TIME	HAM
$\boldsymbol{\beta}^0 = \mathbf{0}_p$									
$\{1, 2, \ldots, 100\}$	1	$\mathcal{B}(1, p)$	3.64	0.288	0.288	6	13	0.63	3.46
$\{1, 12, 23 \ldots, 100\}$	1	$\mathcal{B}(1, p)$	4.65	0.358	0.357	6.01	5	0.11	4.29
$\{10, 20, 30 \ldots, 100\}$	1	$\mathcal{B}(1, p)$	1.96	0.193	0.192	6.01	22	0.02	2.31
$\boldsymbol{\beta}^0 \sim \mathcal{N}_p(\mathbf{0}_p, \mathrm{I}_p)$									
$\{1, 2, \ldots, 100\}$	1	$\mathcal{B}(1, p)$	3.64	0.288	0.288	6	13	0.74	3.46
$\{1, 12, 23 \ldots, 100\}$	1	$\mathcal{B}(1, p)$	5.78	0.427	0.42	6.07	2	0.14	5.11
$\{10, 20, 30 \ldots, 100\}$	1	$\mathcal{B}(1, p)$	5.22	0.38	0.377	6.03	8	0.09	4.55
$\boldsymbol{\beta}^0 \sim \mathcal{N}_p(\mathbf{0}_p, 10 \times \mathrm{I}_p)$									
$\{1, 2, \ldots, 100\}$	1	$\mathcal{B}(1, p)$	3.64	0.288	0.288	6	13	0.78	3.46
$\{1, 12, 23 \ldots, 100\}$	1	$\mathcal{B}(1, p)$	7.74	0.524	0.517	6.09	2	0.18	6.29
$\{10, 20, 30 \ldots, 100\}$	1	$\mathcal{B}(1, p)$	14.38	0.88	0.842	7.93	0	0.16	12.03

Table 1: Simulation study using 100 repetitions; MSE (average mean squared error), FDR (false discovery rate), FNR (false non-discovery rate), DIM (average size of the model), TRUE (# true model detected), TIME (average execution time in seconds), HAM (average Hamming distance)

From another point of view, (D.1) coincides with the M-step of a Bayesian EM algorithm for posterior mode detection under continuous spike-and-slab priors, which treats $\boldsymbol{\gamma}$ as missing data and keeps θ fixed. This connection is made apparent by the fact $\lambda_\theta^\star(\beta_j^{(k)}) = \lambda_1 p_\theta^\star(\beta_j^{(k)}) + \lambda_0[1 - p_\theta^\star(\beta_j^{(k)})]$, and by noting $p_\theta^\star(\beta_j) = \mathsf{E}\left(\gamma_j \,|\beta_j^{(k)}, \theta\right)$ (the E-step calculation). A similar strategy was implemented for a mixture of two Gaussian distributions in the EMVS procedure by **?**. Whereas their approach was based on iteratively solving adaptive ridge regressions, here it entails solving weighted LASSO regressions. The advantages of using EMVS with the LASSO updates are (a) automatic variable selection through thresholding, (b) faster speed of convergence (follows from considerations of **?**).

Extending LLA to the case of a non-separable penalty is achieved naturally within the Bayesian EM framework by treating θ as an additional model parameter. Instead of carrying forward the same fixed value, one now simply updates θ throughout the algorithm. The non-separable variant of the M-step thus uses $\theta = \theta^{(k)}$ to obtain $\boldsymbol{\beta}^{(k+1)}$ from (D.1). This step is followed by a new update $\theta^{(k+1)}$ according to

$$\theta^{(k+1)} = \frac{\sum_{j=1}^{p} p_{\theta^{(k)}}^\star(\beta_j^{(k+1)}) + a - 1}{a + b + p - 2}. \tag{D.2}$$

The calculation (D.2) follows directly from equation (3.12) of **?**. A variant of this strategy was implemented for sparse factor analysis by **?**, where more details on this algorithm can be found.

EMVS Algorithm: *The Spike-and-Slab LASSO*
Input a grid of increasing λ_0 values $I = \{\lambda_0^1, \ldots, \lambda_0^L\}$
For each value $l \in \{1, \ldots, L\}$
Set $k = 0$
(a) Initialize: $\boldsymbol{\beta}_l^{(k)} = \boldsymbol{\beta}^\star$,$\theta^{(0)} = \theta^\star$
(b) While $\mathsf{diff} > \varepsilon$
(i) Increment k
(ii) Update $\boldsymbol{\beta}_l^{(k)}$ according to (D.1) with $\theta = \theta^{(k)}$
(iii) Update $\theta^{(k)}$ according to (D.2)
(iv) $\mathsf{diff} = \|\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}\|\|_2$
(c) Return $\boldsymbol{\beta}_l^{(k)}$
(d) Assign $\boldsymbol{\beta}^\star = \boldsymbol{\beta}_l^{(k)}$

Table 2: The EMVS implementation of the Spike-and-Slab LASSO procedure

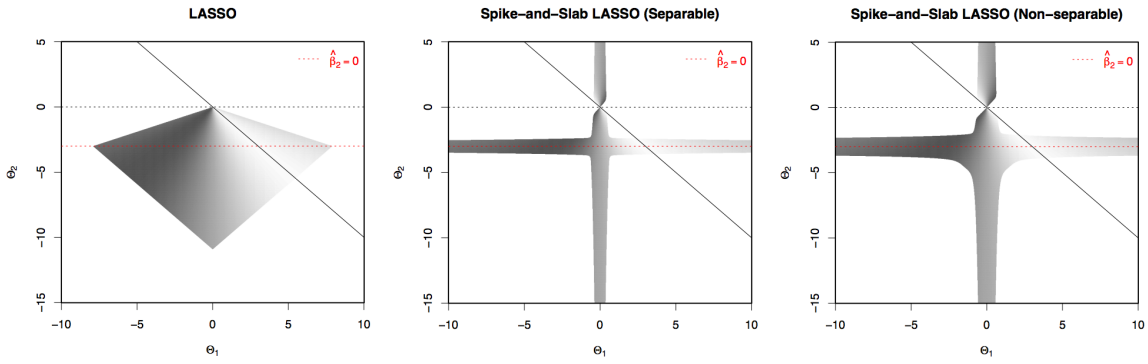
D.2 Posterior Simulation

The *SSL* and *NSSL* priors are also amenable to posterior simulation. Direct Gibbs sampling is available through the exponential scale mixture representation of the Laplace distribution (**?**), applying the SSVS strategy (**?**). Alternatively, one could deploy a variant of an orthant sampler developed for the Bayesian LASSO by **?**. Whereas simulating from the full-dimensional posterior $\pi(\boldsymbol{\beta} \,|\, \boldsymbol{Y})$ will only be practical when p is not overwhelmingly big, initiating the sampler at a posterior mode can save burn-in time and provide a quick insight into uncertainty surrounding the mode. Alternatively, one could confine the simulation to a lower-dimensional subspace, sampling only from active coordinates identified by the mode hunting strategies.

E Identifiability Considerations

As a followup on Lemma 6.1, it is instructive to compare the geometry of the region $C(\eta; \boldsymbol{\beta}_0)$ under different penalty functions. This region captures the discrepancy $\boldsymbol{\Theta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ between the true vector and its shrunken estimate. The shape of this region thus provides useful insights about the nature of the shrinkage.

For the sake of illustration, we assume $\boldsymbol{\beta}_0 = (0, 3)'$ and $\eta = 0.45$. Under the LASSO penalty (where $\lambda_1 = \lambda_0$), the set $C(\eta; \boldsymbol{\beta}_0)$ has a diamond shape (Figure 1(a)), embedded within a cone $\left\{\boldsymbol{\Theta} \in \mathbb{R}^p : |\boldsymbol{\Theta}_{S^c}| \leq \frac{1+\eta}{1-\eta} |\boldsymbol{\Theta}_S|\right\}$. On the other hand, for the other limiting case $\lambda_0 = \infty$ and $\lambda_1 = 0$, the *SSL* penalty corresponds to the ℓ_0 penalty. The set $C(\eta; \boldsymbol{\beta}_0)$ then consists of those values $\boldsymbol{\Theta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ for which $||\widehat{\boldsymbol{\beta}}||_0 < 2$, as marked by the two solid lines corresponding to $\Theta_1 = 0$ and $\Theta_2 = -3$. The *SSL* penalty yields a compromise between



(a) $\lambda_1 = \lambda_2$ (b) $\lambda_1 = 0.1, \lambda_0 = 10, \theta = 0.1$ (c) $\lambda_1 = 0.1, \lambda_0 = 10, a = 1, b = 9$

Figure 1: Plots of the feasible regions for $\boldsymbol{\Theta}$ under the LASSO penalty and the *SSL* penalty.

these two extremes. With $1 < \lambda_0 < \infty$ and $0 < \lambda_1 < 1$, $C(\eta; \boldsymbol{\beta}_0)$ is a star-shaped wrap around the set $\{\boldsymbol{\Theta} : ||\widehat{\boldsymbol{\beta}}||_0 < 2\}$. With larger λ_1, the set begins to resemble a diamond. The non-separable penalty ties the coordinates together, making the set larger in the center.

Figures 1(a), 1(b) and 1(c) are actual heat-maps of the restricted eigenvalues $\frac{||\boldsymbol{X}\boldsymbol{\Theta}||}{||\boldsymbol{X}||||\boldsymbol{\Theta}||}$ inside $C(\eta; \boldsymbol{\beta}_0)$; the darker the shade of grey, the larger the value. Here, $\boldsymbol{X}$ contains $n = 100$ observations on 2 highly collinear variables (correlation $\rho = 0.96$). The diamond in Figure 1(a) is seen as a continuum of rays of equal eigenvalues, the minimum attained on the ray $\Theta_2 = -\Theta_1$ (marked by a solid line). This ray dissects all three sets in Figure 1. Under the ℓ_0 penalty, the intersection occurs at $\widehat{\boldsymbol{\beta}} = (3, 0)'$, the "opposite" of $\boldsymbol{\beta}_0 = (0, 3)$ where the correct variable was mistaken for its "knockoff". This unfavorable case is assigned a very small $c(\eta; \boldsymbol{\beta}_0)$ value, an indication that the true variable is not easily distinguishable. Interestingly, compared to the LASSO set, the *SSL* feasible regions indicate that at least one of the coordinates in $\widehat{\boldsymbol{\beta}}$ must be negligible, acknowledging the sparsity of the true $\boldsymbol{\beta}_0$. However, despite their different geometries, the minimal eigenvalue taken over these different sets is the same. Thus, regardless of the penalty, the same identifiability condition has to be imposed in this example.