A Proofs

A.1 Proof of Theorem 4.1

For \( y_i \) arising from model (15), from Ročková (2018) theorem 3.1 we have

\[
\hat{\beta}_i = \begin{cases} 
0, & \text{if } |Y_i| \leq \Delta. \\
[|Y_i| - \sigma^2 \lambda^*(\hat{\beta}_i)]_+, \text{sign}(Y_i), & \text{otherwise.}
\end{cases}
\]

After re-weighting, our objective function becomes

\[
\tilde{\beta} = \arg\max_{\beta \in \mathbb{R}^p} \left\{ -\sum_{i=1}^{n} \frac{w_i}{2\sigma^2} (Y_i - \beta_i)^2 + \sum_{j=1}^{p} \log \pi(\beta_j - \mu_j | \theta) \right\}
\]

\[
= \arg\max_{\beta \in \mathbb{R}^p} \left\{ -\sum_{i=1}^{n} \frac{1}{2\sigma^2} (Y_i^* - \beta^*_i)^2 + \sum_{j=1}^{p} \log \pi^*(\beta^*_j | \theta) \right\}
\]

where \( Y_i^* = \sqrt{w_i}(Y_i - \mu_i) \), \( \beta_i^* = \sqrt{w_i}(\beta_i - \mu_i) \), and we are imposing prior \( \pi^* \) with \( \lambda'_0 = \lambda_0/\sqrt{w_i} \) and \( \lambda'_1 = \lambda_1/\sqrt{w_i} \). So from (1), the mode estimator \( \hat{\beta}_i^* \) for \( \beta_i^* \) satisfies

\[
\hat{\beta}_i^* = \begin{cases} 
0, & \text{if } |Y_i^*| \leq \Delta. \\
[|Y_i^*| - \sigma^2 \lambda^*(\hat{\beta}_i^*)]_+, \text{sign}(Y_i^*), & \text{otherwise.}
\end{cases}
\]

From the chain rule, \( \lambda^*(\hat{\beta}_i - \mu_i) = \lambda^* \left( \sqrt{w_i} \left( \hat{\beta}_i - \mu_i \right) \right) \frac{\partial(\sqrt{w_i}(\hat{\beta}_i - \mu_i))}{\partial(\hat{\beta}_i - \mu_i)} \), and thus,

\[
\tilde{\beta}_i = \begin{cases} 
\mu_i, & \text{if } |\sqrt{w_i}(Y_i - \mu_i)| \leq \Delta. \\
\mu_i + [\sqrt{w_i}(\hat{\beta}_i - \mu_i)]_+ \text{sign}(\sqrt{w_i}(Y_i - \mu_i)), & \text{otherwise.}
\end{cases}
\]

For active coordinates,

\[
\mathbb{E}_Y|\mathbb{E}_{\mu_i,w_i}[((\hat{\beta}_i - \beta_i^0)^2 | Y)] = \mathbb{E}_Y|\mathbb{E}_{\mu_i,w_i}[((\hat{\beta}_i - Y_i + Y_i - \beta_i^0)^2 | Y)]
\]

\[
\leq 2\sigma^2 + 2\mathbb{E}_Y|\mathbb{E}_{\mu_i,w_i}[((\hat{\beta}_i - Y_i)^2 | Y]
\]

\[
= 2\sigma^2 + 2\mathbb{E}_Y|\mathbb{E}_{\mu_i,w_i}[((\hat{\beta}_i - Y_i)^2 | Y]|(\sqrt{w_i}|Y_i - \mu_i| \leq \Delta) + 2\mathbb{E}_Y|\mathbb{E}_{\mu_i,w_i}[((\hat{\beta}_i - Y_i)^2 | Y]|(\sqrt{w_i}|Y_i - \mu_i| > \Delta)
\]

where

\[
\mathbb{E}_Y|\mathbb{E}_{\mu_i,w_i}[((\hat{\beta}_i - Y_i)^2 | Y]|(\sqrt{w_i}|Y_i - \mu_i| \leq \Delta) = \mathbb{E}_Y|\mathbb{E}_{\mu_i,w_i}[((\mu_i - Y_i)^2 | Y]|(|Y_i - \mu_i| \leq \Delta/\sqrt{w_i}) \leq \Delta^2 \mathbb{E}_{w_i} \frac{1}{w_i}
\]

and

\[
\mathbb{E}_Y|\mathbb{E}_{\mu_i,w_i}[((\hat{\beta}_i - Y_i)^2 | Y]|(\sqrt{w_i}|Y_i - \mu_i| > \Delta) = \mathbb{E}_Y|\mathbb{E}_{\mu_i,w_i}[((1/w_i \lambda^*(\hat{\beta}_i - \mu_i))^2 | Y]|(|\sqrt{w_i}|Y_i - \mu_i| > \Delta) \leq 4\mathbb{E}_{w_i} \frac{1}{w_i^2}
\]
where the last inequality in (5) follows the same argument as Ročková (2018): \( |\beta^*_i| > \delta_{c+} \), so \( (p'_i)^*(\beta^*_i) < c_+ \) when \( |Y^*_i| > \Delta \) (here \( (p'_i)^* \) uses \( \lambda'_0 = \lambda_0/\sqrt{w_i} \) and \( \lambda'_1 = \lambda_1/\sqrt{w_i} \)), i.e., \( p^*(\beta_i - \mu_i) < c_+ \) when \( |\sqrt{w_i}(Y_i - \mu_i)| > \Delta \) (notice \( p^*_i \) uses \( \lambda_0 \) and \( \lambda_1 \)). Thus, since \( c_+(1 - c_+) = \frac{1}{\lambda_0 - \lambda_1^2} \), we have \( \lambda^*(\beta_i - \mu_i) < c_+(\lambda_1 - \lambda_0) + \lambda_0 = (1 - c_+)\lambda_0 - \lambda_1 < \frac{2}{\lambda_0 - \lambda_1^2} + 1 < \frac{2}{\lambda_0 - \lambda_1} < 2 \) when \( |\sqrt{w_i}(Y_i - \mu_i)| > \Delta \).

Thus, from (3), (4), (5) and condition (ii), we know that for active coordinates:

\[
E_{Y,t}E_{\mu_i,w_i}[(\hat{\beta}_i - \beta^0_i)^2 | Y] \leq 2\sigma^2 + 2\Delta^2 E_{w_i} \frac{1}{w_i} + 8E_{w_i} \frac{1}{w_i^2} \leq C_3 \Delta^2
\]

For inactive coordinates, for some constant \( t \) which satisfies condition (iii),

\[
E_{Y,t}E_{\mu_i,w_i}[(\hat{\beta}_i - \beta^0_i)^2 | Y] = E_{Y,t}E_{\mu_i,w_i}[(\hat{\beta}_i)^2 | Y]
\]

\[
\leq E_{Y,t}E_{\mu_i,w_i}[(Y_i - \mu_i)^2 | Y] I(|\sqrt{w_i}(Y_i - \mu_i)| \geq \Delta) I(w_i \leq t) I(\mu_i \leq \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}})
\]

\[
\leq 2 E_{Y,t}E_{\mu_i,w_i}[(Y_i - \mu_i)^2 + |\mu_i|^2] Y I(|\sqrt{w_i}(Y_i - \mu_i)| \geq \Delta) I(w_i \leq t) I(\mu_i \leq \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}})
\]

\[
\leq 2 E_{Y,t}E_{\mu_i,w_i}[(|Y_i - \mu_i|^2 + |\mu_i|^2] Y I(|\sqrt{w_i}(Y_i - \mu_i)| \geq \Delta) I(w_i \leq t) I(\mu_i \leq \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}})
\]

\[
= U_1 + U_2 + U_3
\]

For the first term \( U_1 \) in (6),

\[
U_1 = 2 E_{Y,t}E_{\mu_i,w_i} [(Y_i - \mu_i)^2 + |\mu_i|^2] Y I(|\sqrt{w_i}(Y_i - \mu_i)| \geq \Delta) I(w_i \leq t) I(\mu_i \leq \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}})
\]

\[
\leq 2 E_{Y,t}E_{\mu_i} [(Y_i - \mu_i)^2 Y] I(Y_i - \mu_i \geq \Delta) I(w_i \leq t) I(\mu_i \leq \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}}) + 2 E_{Y,t}E_{\mu_i,w_i} Y^2 I(Y_i \geq \Delta) I(w_i \leq t)
\]

\[
= 2 \sigma^2 \int E_{\mu_i} \left[ \left( \frac{\Delta}{\sigma \sqrt{t}} - \frac{\mu_i}{\sigma} \right) \phi \left( \frac{\mu_i}{\sigma} + \frac{\Delta}{\sigma \sqrt{t}} \right) + \left( \frac{\mu_i}{\sigma} + \frac{\Delta}{\sigma \sqrt{t}} \right) - \phi \left( \frac{\mu_i}{\sigma} - \frac{\Delta}{\sigma \sqrt{t}} \right) + (1 + \mu_i^2) \phi \left( \frac{\mu_i}{\sigma} - \frac{\Delta}{\sigma \sqrt{t}} \right) \right] I(\mu_i \leq \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}})
\]

\[
\leq 4 \sigma^2 \int E_{\mu_i} \left[ \left( \frac{\Delta}{\sigma \sqrt{t}} + \frac{\mu_i}{\sigma} \right) \phi \left( \frac{\mu_i}{\sigma} - \frac{\Delta}{\sigma \sqrt{t}} \right) + (1 + \mu_i^2) \phi \left( \frac{\Delta}{\sigma \sqrt{t}} \right) \right] I(\mu_i \leq \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}}) + \frac{16}{\lambda_0} \left( - \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}} + \frac{\Delta}{\sigma \sqrt{t}} \right)
\]

\[
\leq 4 \sigma^2 \int E_{\mu_i} \left[ \left( \frac{\Delta}{\sigma \sqrt{t}} + \frac{\mu_i}{\sigma} \right) \phi \left( \frac{1}{\sigma \lambda_0} + \frac{1}{\sigma \sqrt{\lambda_0}} - \frac{\Delta}{\sigma \sqrt{t}} \right) + (1 + \mu_i^2) \phi \left( \frac{1}{\sigma \lambda_0} + \frac{1}{\sigma \sqrt{\lambda_0}} - \frac{\Delta}{\sigma \sqrt{t}} \right) \right] I(\mu_i \leq \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}})
\]

\[
+ 0.5 \phi \left( \frac{1}{\sigma \lambda_0} + \frac{1}{\sigma \sqrt{\lambda_0}} + \frac{\Delta}{\sigma \sqrt{t}} \right)
\]

\[
= 4 \sigma^2 \left[ \left( \frac{\Delta}{\sigma \sqrt{t}} + \frac{1}{\sigma \lambda_0} \right) \phi \left( \frac{1}{\sigma \lambda_0} + \frac{1}{\sigma \sqrt{\lambda_0}} - \frac{\Delta}{\sigma \sqrt{t}} \right) + \left( 1 + \frac{2}{\lambda_0^2} \right) \phi \left( \frac{1}{\sigma \lambda_0} + \frac{1}{\sigma \sqrt{\lambda_0}} - \frac{\Delta}{\sigma \sqrt{t}} \right) \right] + 0.5 \phi \left( \frac{1}{\sigma \lambda_0} + \frac{1}{\sigma \sqrt{\lambda_0}} + \frac{\Delta}{\sigma \sqrt{t}} \right)
\]

\[
\leq 5 \sigma^2 \phi \left( \frac{\Delta}{\sigma \sqrt{t}} - \frac{1}{\sigma \lambda_0} + \frac{1}{\sigma \sqrt{\lambda_0}} \right)
\]

(7)
where (a) uses the following result (which basically utilizes $y\phi(y) = -\phi'(y)$ and integration by parts)

$$
\mathbb{E}_{Y_i} \left( (|Y_i - \mu_i|^2 - |Y_i|) \mathbb{I}(|Y_i - \mu_i| \geq \sqrt{t}) \right)
\approx^{(b)} \int_{Y_i - \mu_i > \frac{\Delta}{\sigma \sqrt{t}}} (Y_i - \mu_i)^2 \phi(Y_i) dY_i + \int_{Y_i - \mu_i < -\frac{\Delta}{\sigma \sqrt{t}}} (Y_i - \mu_i)^2 \phi(Y_i) dY_i
$$

$$
= \int_{Y_i - \mu_i > \frac{\Delta}{\sigma \sqrt{t}}} Y_i^2 \phi(Y_i) dY_i - 2\mu_i \int_{Y_i - \mu_i > \frac{\Delta}{\sigma \sqrt{t}}} Y_i \phi(Y_i) dY_i + \int_{Y_i - \mu_i < -\frac{\Delta}{\sigma \sqrt{t}}} Y_i \phi(Y_i) dY_i + \mu_i^2 \int_{Y_i - \mu_i < -\frac{\Delta}{\sigma \sqrt{t}}} \phi(Y_i) dY_i
$$

$$
= -\sigma^2 \int_{Y_i - \mu_i > \frac{\Delta}{\sigma \sqrt{t}}} Y_i^2 \phi(Y_i) dY_i + 2\mu_i \sigma \phi(Y_i) dY_i + \mu_i^2 \left[ 1 - \Phi(\mu_i / \sigma + \frac{\Delta}{\sigma \sqrt{t}}) \right]
$$

For the second term in (6),

$$
U_2 = 2\mathbb{E}_{Y_i} \mathbb{E}_{\mu_i, w_i} \left[ (|Y_i - \mu_i|^2 + |\mu_i|^2) | Y \right] \mathbb{I}(\sqrt{w_i} | Y_i - \mu_i| \geq \Delta) \mathbb{I}(w_i \leq t) \mathbb{I}(\mu_i > \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}})
$$

$$
\leq 2\mathbb{E}_{Y_i} \mathbb{E}_{\mu_i, w_i} \left[ (|Y_i - \mu_i|^2 + |\mu_i|^2) | Y \right] \mathbb{I}(\mu_i > \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}})
$$

$$
\leq 2\mathbb{E}_{Y_i} \mathbb{E}_{\mu_i, w_i} \left[ (2Y_i^2 + 3\mu_i^2) | Y \right] \mathbb{I}(\mu_i > \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}})
$$

$$
= 2\mathbb{E}_{\mu_i} (2\sigma^2 + 3\mu_i^2) \mathbb{I}(\mu_i > \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}})
$$

$$
= 4\sigma^2 \mathbb{P}(\mu_i > \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}}) + 6\mathbb{E}_{\mu_i} \mu_i^2 \mathbb{I}(\mu_i > \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}})
$$

$$
= 4\sigma^2 \mathbb{P}(\mu_i > \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}}) + 6\mathbb{E}_{\mu_i} \mu_i^2 \mathbb{I}(\mu_i > \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}})
$$

$$
\leq 4\sigma^2 e^{-1 - \sqrt{\lambda_0}} + 6 \left[ \left( \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}} \right)^2 + 2 \left( \frac{2}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}} \right) \right] e^{-1 - \sqrt{\lambda_0}} < \frac{\Delta}{\sigma \sqrt{t}} \phi \left( \frac{\Delta}{\sigma \sqrt{t}} - \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}} \right)
$$

where (e) follows from integration by parts.
For the third term in (6), utilizing condition (iii), we have $\mathbb{P}(w_i > t) \leq \tilde{C}_3 \frac{\Delta}{\sqrt{t}} \phi \left( \frac{\Delta}{\sqrt{t}} \right)$, thus

$$U_3 = 2\mathbb{E}_{Y_i} \mathbb{E}_{\mu, w_i}[|Y_i| - \mu_i|^2 + |\mu_i|^2 | Y] \mathbb{I}(\sqrt{w_i}|Y_i| - \mu_i| \geq \Delta) \mathbb{I}(w_i > t)$$

$$\leq 2\mathbb{E}_{Y_i} \mathbb{E}_{\mu, w_i}[2|Y_i|^2 + 3|\mu_i|^2 | Y] \mathbb{I}(w_i > t) = \left( 4\sigma^2 + \frac{12}{\lambda_0^2} \right) \mathbb{E}_{w_i} \mathbb{I}(w_i > t) \leq C_2 \sigma \frac{\Delta}{\sqrt{t}} \phi \left( \frac{\Delta}{\sigma \sqrt{t}} \right)$$

So from (6), the risk for inactive coordinates will be bounded

$$\mathbb{E}_{Y_i} \mathbb{E}_{\mu, w_i}[(\tilde{\beta}_i - \beta_0^0)^2 | Y] = U_1 + U_2 + U_3 \leq 6\sigma \frac{\Delta}{\sqrt{t}} \phi \left( \frac{\Delta}{\sigma \sqrt{t}} - \frac{1}{\sigma \lambda_0} + \frac{1}{\sigma \sqrt{\lambda_0}} \right) + C_2 \sigma \frac{\Delta}{\sqrt{t}} \phi \left( \frac{\Delta}{\sigma \sqrt{t}} \right)$$

and the risk for all coordinates satisfies

$$\mathbb{E}_{Y_i} \mathbb{E}_{\mu, w_i}[(\tilde{\beta} - \beta_0)^2 | Y] \leq qC_1 \Delta^2 + (n - q) \sigma (6 + C_2) \frac{\Delta}{\sqrt{t}} \phi \left( \frac{\Delta}{\sigma \sqrt{t}} - \frac{1}{\sigma \lambda_0} + \frac{1}{\sigma \sqrt{\lambda_0}} \right)$$

$$= qC_1 \Delta^2 + \tilde{C}_2(n - q) \sigma \frac{\Delta}{\sqrt{t}} e^{-\frac{1}{2} \left( \frac{\Delta}{\sigma \sqrt{t}} - \frac{1}{\sigma \lambda_0} + \frac{1}{\sigma \sqrt{\lambda_0}} \right)^2} \leq qC_1 \Delta^2 + \tilde{C}_2(n - q) \frac{\Delta}{\sqrt{t}} e^{-\frac{1}{2} \left( \frac{\Delta}{\sigma \sqrt{t}} \right)^2}$$

$$\leq qC_1 \Delta^2 + \tilde{C}_2(n - q) \frac{\Delta}{\sqrt{t}} e^{-\frac{1}{2} \left( \frac{\Delta}{\sigma \sqrt{t}} \right)^2} \leq qC_1 \Delta^2 + \tilde{C}_2(n - q) \frac{\Delta U}{\sqrt{t}}$$

where $\Delta^U = \sqrt{2 \log[1/p^*(0)]} + \lambda_1$ and $\Delta^L = \sqrt{2 \log[1/p^*(0)]} - d + \lambda_1$. From Roˇckov´a (2018) Theorem 3.1, $\Delta^L < \Delta \leq \Delta^U$. From Roˇckov´a (2018) Lemma 1.2 (stated in Appendix), $d < 2$.

Set $t = \eta + \gamma_i$, $\mathbb{E}_{Y_i} \mathbb{E}_{\mu, w_i}[(\tilde{\beta} - \beta_0)^2 | Y] \leq qC_3 \Delta^2$. Then from Markov Inequality, for any $M_n \to \infty$,

$$\mathbb{E}_{Y_i} \mathbb{P}_{\mu, w_i} \left( \|\tilde{\beta} - \beta_0\|^2 > M_n q \log \left( \frac{n}{\eta} \right) | Y \right) \leq \frac{\mathbb{E}_{Y_i} \mathbb{E}_{\mu, w_i} \left[ \|\tilde{\beta} - \beta_0\|^2 | Y \right]}{M_n q \log \left( \frac{n}{\eta} \right)} \leq \frac{qC_3 \Delta^2}{M_n q \log \left( \frac{n}{\eta} \right)}$$

where $\frac{qC_3 \Delta^2}{M_n q \log \frac{n}{\eta}} \to 0$. This means for any $M_n \to \infty$,

$$\mathbb{E}_{Y_i} \mathbb{P}_{\mu, w_i} \left( \|\tilde{\beta} - \beta_0\|^2 > M_n q \log \left( \frac{n}{\eta} \right) | Y \right) \to 0$$

**A.2 Proof of Corollary 4.1**

Condition (i) of theorem 4.1 is satisfied. With $\alpha \geq 2$, condition (ii) also holds because when $w_i \sim \frac{1}{\alpha} \text{Gamma}(\alpha, 1)$, $\frac{1}{w_i} \sim \alpha \times \text{Inverse-Gamma}(\alpha, 1)$ and when $w \sim n \text{Dir}(\alpha, \cdots, \alpha)$, $w_i \sim n \times \text{Beta}(\alpha, (n - 1)\alpha)$. Both of them satisfy condition (ii). So we only need to check
condition (iii). Write $B(\cdot)$ as the Beta function. When $\mathbf{w} \sim n\text{Dir}(\alpha, \cdots, \alpha)$, then for any $t \geq \frac{\alpha + 1}{\alpha}$, the following equation holds

\[
\mathbb{P}_{w_i}(w_i > t) = \mathbb{P}(\text{Beta}(\alpha, n\alpha - \alpha) > \frac{t}{n}) = \frac{\int_{t/n}^{1} v_i^{\alpha-1}(1 - v_i)^{n\alpha - \alpha - 1} dv_i}{B(\alpha, n\alpha - \alpha)}
\]

\[
= \frac{1}{B(\alpha, n\alpha - \alpha)(an)^{\alpha}} \int_{at}^{an} z^{\alpha-1}[(1 - \frac{1}{an/z})^{-a \alpha} - 1]^{\frac{\alpha - 1}{1 - \alpha/n}} dz
\]

\[
\leq \frac{1}{B(\alpha, n\alpha - \alpha)(an)^{\alpha}} \int_{at}^{an} z^{\alpha-1} \frac{\Gamma(n\alpha)}{\Gamma(n\alpha - \alpha)(an)^{\alpha}} \int_{at}^{an} z^{\alpha-1} dz
\]

\[
\leq \frac{1}{B(\alpha, n\alpha - \alpha)(an)^{\alpha}} \int_{at}^{an} z^{\alpha-1} \frac{\Gamma(n\alpha)}{\Gamma(n\alpha - \alpha)(an)^{\alpha}} \mathbb{P}(\text{Gamma}(\alpha, 1) > at) \leq \mathbb{P}(\text{Gamma}(\alpha, 1) > at)
\]

\[
\leq \frac{1}{e^{\alpha tx}} = e^{\alpha(-\log(1-x)-tx)}
\]

where (a) uses the fact $(1 - \frac{1}{t})^{-\alpha + 1} \leq e$ for any $x > 0$. Inequality (b) uses $t \geq \frac{\alpha + 1}{\alpha}$. Inequality (c) uses Chernoff bound and we need $x \in (0, 1)$ for the above inequality to hold.

Notice that when $w_i \sim \frac{1}{\alpha} \text{Gamma}(\alpha, 1)$, we can directly get

\[
\mathbb{P}_{w_i}(w_i > t) = \mathbb{P}(\text{Gamma}(\alpha, 1) > at) \leq e^{\alpha(-\log(1-x)-tx)}
\]

Set $x = 1 - \frac{1}{t}$ and $\alpha = \frac{\Delta^2}{2(t-1-\log t)}$, then

\[
\mathbb{P}_{w_i}(w_i > t) \leq e^{\alpha(-\log(1-x)-tx)} = e^{\alpha(\log(t)+1-t)} = e^{\frac{\Delta^2}{2t}} \leq e^{\frac{(\Delta^2)}{2nt}} \leq C_4 \left( \frac{q}{n} \right)^{\sigma^2(\eta + \gamma)/t} \tag{8}
\]

So condition (iii) holds for any $t \in (0, (\eta + \gamma)\sigma^2]$. We can get Lemma 4.1 by applying Theorem 4.1.

A.3 An example for Remark under 4.1

Here we want to show that for $\mathbf{w} \sim n \times \text{Dir}(\alpha, \cdots, \alpha)$ where $\alpha < 2$, the risk for BB-SSL arising from (15) can be arbitrarily large.

Now $\mathbf{w} = (w_1, w_2, \cdots, w_n) \sim n\text{Dir}(\alpha, \alpha, \cdots, \alpha)$. One important property we will use
below is that $nBeta(\alpha, n\alpha - \alpha) \xrightarrow{d} \frac{1}{\alpha} Gamma(\alpha, 1)$ for any fixed $\alpha$.

$$\mathbb{E}_Y \mathbb{E}_{\mu, w_i}[(\tilde{\beta}_i - Y_i)^2 | Y] \mathbb{P}(\sqrt{w_i}(Y_i - \mu_i) > \Delta) = \mathbb{E}_Y \mathbb{E}_{\mu, w_i}[(\frac{1}{w_i} \lambda^*(\tilde{\beta}_i - \mu_i))^2 | Y] \mathbb{P}(\sqrt{w_i}Y_i - \mu_i | \Delta)$$

$$\geq \lambda_i^2 \mathbb{E}_Y \mathbb{E}_{w_i} \frac{1}{w_i^2} \mathbb{P}(w_i > \frac{\Delta^2}{(Y_i - \mu_i)^2})$$

$$\equiv \lambda_i^2 \mathbb{E}_Y \mathbb{E}_{w_i} \mathbb{E}_{v_i \sim \Gamma(\alpha, 1)} \frac{\alpha^2}{v_i^2} \mathbb{P}(v_i > \frac{\alpha \Delta^2}{(Y_i - \mu_i)^2}) + o(1)$$

$$= \lambda_i^2 \alpha^2 \mathbb{E}_Y \mathbb{E}_{w_i} \mathbb{E}_{z_i \sim \Gamma(\alpha, 1)} \frac{z_i^2}{z_i^2} \mathbb{P}(z_i < \frac{(Y_i - \mu_i)^2}{\alpha \Delta^2}) + o(1)$$

where (a) follows from the following derivations:

$$w_i \xrightarrow{d} \frac{1}{\alpha} Gamma(\alpha, 1),$$

so for any function $f(\cdot)$ which is $X$-continuous and bounded, we have $\mathbb{E}_f(w_i) \to \mathbb{E}_f(v_i/\alpha)$ where $v_i \sim Gamma(\alpha, 1)$. Since $f(w_i) = \frac{1}{w_i^2} \mathbb{P}(w_i > \frac{\Delta^2}{(Y_i - \mu_i)^2})$ is $X$-continuous and bounded by $\frac{(Y_i - \mu_i)^2}{\alpha}$, so $\mathbb{E}_f(w_i) \to \mathbb{E}_f(v_i/\alpha)$. It means that if we let $f_n(Y_i, \mu_i) = \mathbb{E}_{w_i \sim \Gamma(\alpha, 1)} \frac{1}{w_i^2} \mathbb{P}(w_i > \frac{\Delta^2}{(Y_i - \mu_i)^2})$ and $f(Y_i, \mu_i) = \mathbb{E}_{v_i \sim Gamma(\alpha, 1)} \frac{\alpha^2}{v_i^2} \mathbb{P}(v_i > \frac{\alpha \Delta^2}{(Y_i - \mu_i)^2})$, then $f_n(Y_i, \mu_i) \to f(Y_i, \mu_i)$ for any $Y_i$ and $\mu_i$. Also notice that $|f_n(Y_i, \mu_i)| \leq \frac{(Y_i - \mu_i)^2}{\alpha \Delta^2}$ and $\frac{(Y_i - \mu_i)^2}{\alpha \Delta^2}$ is integrable. So from Dominated Convergence Theorem, $\mathbb{E}_{Y, \mu_i} f_n(Y_i, \mu_i) \to \mathbb{E}_{Y, \mu_i} f(Y_i, \mu_i)$, i.e., $\mathbb{E}_{Y, \mu_i} \frac{1}{w_i^2} \mathbb{P}(w_i > \frac{\Delta^2}{(Y_i - \mu_i)^2}) \to \mathbb{E}_{Y, \mu_i} \mathbb{E}_{v_i \sim \Gamma(\alpha, 1)} \frac{\alpha^2}{v_i^2} \mathbb{P}(v_i > \frac{\alpha \Delta^2}{(Y_i - \mu_i)^2})$.

Since

$$\mathbb{E}_z \mathbb{E}_{z_i \sim \Gamma(\alpha, 1)} \frac{z_i^2}{z_i^2} \mathbb{P}(z_i < \frac{\alpha \Delta^2}{(Y_i - \mu_i)^2}) = \int_0^M z_i^2 \frac{1}{z_i^{\alpha+1}} e^{-1/z_i} dz_i = \int_0^M z_i^{1-\alpha} e^{-1/z_i} dz_i$$

$$= \frac{1}{2 - \alpha} \int_0^M e^{-1/z_i} dz_i^{2-\alpha} = \frac{1}{2 - \alpha} (e^{-1/z_i} z_i^{2-\alpha} | z_i^0 - \int_0^M z_i^{2-\alpha} e^{-1/z_i} dz_i$$

$$= \frac{1}{2 - \alpha} (e^{-1/M} M^{2-\alpha} + \int_0^M \frac{1}{z_i^{2-\alpha}} e^{-1/z_i} dz_i) > \frac{1}{2 - \alpha} e^{-1/M} M^{2-\alpha}$$

So

$$\mathbb{E}_Y \mathbb{E}_{\mu, w_i}[(\tilde{\beta}_i - Y_i)^2 | Y] \mathbb{P}(\sqrt{w_i}(Y_i - \mu_i) > \Delta) \geq \lambda_i^2 \alpha^2 \mathbb{E}_Y \mathbb{E}_{w_i} \mathbb{E}_{z_i \sim \Gamma(\alpha, 1)} \frac{z_i^2}{z_i^2} \mathbb{P}(z_i < \frac{(Y_i - \mu_i)^2}{\alpha \Delta^2}) + o(1)$$

$$\geq \lambda_i^2 \alpha^2 \mathbb{E}_Y \mathbb{E}_{w_i} \mathbb{E}_{z_i \sim \Gamma(\alpha, 1)} \frac{z_i^2}{z_i^2} \mathbb{P}(z_i < \frac{(Y_i - \mu_i)^2}{\alpha \Delta^2}) + o(1)$$

When $\alpha < 2$, if we want $\lambda_i^2 \alpha^2 \mathbb{E}_Y \mathbb{E}_{w_i} \mathbb{E}_{z_i \sim \Gamma(\alpha, 1)} \frac{z_i^2}{z_i^2} \mathbb{P}(z_i < \frac{(Y_i - \mu_i)^2}{\alpha \Delta^2}) \leq \Delta^2$ for any $\beta_i^0$, we need

$$\mathbb{E}_Y \mathbb{E}_{\mu_i} e^{-\alpha \Delta^2 / (Y_i - \mu_i)^2} \frac{(Y_i - \mu_i)^2}{\alpha \Delta^2} \leq \Delta^2$$
to hold for any $\beta_i^0$. But if we set $\beta_i^0 = 2\Delta^t$ where $t = \frac{C}{2-\alpha} + 1$, then

$$
E_{Y, \mu} e^{-\alpha \Delta^2/(Y_i - \mu_i)^2} \left( \frac{(Y_i - \mu_i)^2}{\alpha \Delta^2} \right)^{2-\alpha} 
$$

$$
\geq E_{Y, \mu} e^{-\alpha \Delta^2/(Y_i - \mu_i)^2} \left[ \frac{3}{2} \Delta^t \right]^{2-\alpha} \mathbb{I}(|Y_i| \geq \frac{3}{2} \Delta^t \mathbb{I}(|\mu_i| < 1) 
$$

$$
\geq E_{Y, \mu} e^{-\alpha \Delta^2/(Y_i - \mu_i)^2} \left[ \frac{3}{2} \Delta^t \right]^{2-\alpha} \mathbb{I}(|Y_i| \geq \frac{3}{2} \Delta^t \mathbb{I}(|\mu_i| < 1) 
$$

$$
g(\bar{C} \Delta^{(2t-2)(2-\alpha)\alpha-2} \geq C_5 \Delta^{2\bar{C}} 
$$

where (b) follows from the fact that $\mathbb{P}(|Y_i| \geq \frac{3}{2} \beta_i^0) \geq C_1$ and $\mathbb{P}(|\mu_i| < 1) \geq C_2$ for some $C_1, C_2 > 0$ when $n$ is sufficiently large. Since here $\bar{C}$ can be arbitrarily large, the risk (depending on the truth $\beta_i^0$) can be of arbitrarily large order.

### A.4 Definitions and Lemmas Used In Theorem 4.2 and Theorem 4.3

We write model (1) in matrix form $Y = X\beta + \epsilon$. We denote $\text{pen}(\beta | \theta) = \log \frac{\pi(\beta | \theta)}{\pi(0_p | \theta)}$. We write $W = \text{diag}(\sqrt{w_1}, \ldots, \sqrt{w_n})$. We denote by $Q(\beta) = -\frac{1}{2}||WY - WX\mu - WX\beta||^2 + \text{pen}(\beta | \theta)$ and $\hat{\beta} = \arg \max_{\beta} Q(\beta)$. Notice that BB-SSL solution $\hat{\beta}$ is $\hat{\beta} = \hat{\beta} + \mu$. The matrix norm $|| \cdot ||_a$ is defined as $||X||_a = \sup_{\beta} \frac{||X\beta||}{||\beta||_a}$ where $|| \cdot ||_a$ is the vector $a$-norm. Write $\Theta = \hat{\beta} - \beta_0$. We use $|\beta|$ and $||\beta||_1$ equivalently to denote vector 1-norm. The proof here uses similar ideas and techniques from Ročková and George (2018).

**Definition A.1.** Let $\tilde{\eta} \in (0, 1]$. We say that $X$ with penalty $\text{pen}(\beta)$ satisfies the $\tilde{\eta}$-null consistency ($\tilde{\eta}$-NC) condition if

$$
\arg \max_{\beta \in \mathbb{R}^p} \left\{ -\frac{1}{2} ||\epsilon/\tilde{\eta} - X\beta||^2 + \text{pen}(\beta) \right\} = 0_p
$$

**Lemma A.1.** Under condition (5) in Theorem 4.2, we have

$$
\lim_{n \to \infty} \mathbb{P}_\beta \left( \arg \max_{\beta \in \mathbb{R}^p} \left\{ -\frac{1}{2} ||\epsilon/\tilde{\eta} - X\beta||^2 + \text{pen}(\beta | \theta) \right\} = 0_p \right) = 1
$$

i.e., $X$ satisfies the $\tilde{\eta}$-NC condition with probability approaching 1.

**Lemma A.2.** Under conditions (1)-(5) in Theorem 4.2, on condition that $||\epsilon||_\infty \lesssim \sqrt{\log n}$ and $\tilde{\eta}$-NC condition holds, we have

$$
\lim_{n \to \infty} \mathbb{P}_{\mu, \epsilon} \left( \arg \max_{\beta \in \mathbb{R}^p} \left\{ -\frac{1}{2} ||W(\epsilon - X\mu)/\tilde{\eta} - WX\beta||^2 + \text{pen}(\beta | \theta) \right\} = 0_p | X, \epsilon \right) = 1
$$
where $\eta^* = \max \left\{ \tilde{\eta} + C_n \frac{\|X\|}{\lambda_1}, \frac{\eta}{m} \right\}$ and $\tilde{C}_n$ is any sequence that satisfies $C_n \to \infty$.

Lemma A.3. If $\arg \max_{\beta \in \mathbb{R}^p} \left\{ -\frac{1}{2} \| W(\epsilon - X\mu)/\eta^* - WX\beta \|^2 + \text{pen}(\beta | \theta) \right\} = 0_p$ where $\eta^* \in (0, 1]$, then BB-SSL estimate $\tilde{\beta}$ lies in the cone $C(\eta^*; \beta) = \{ \Theta \in \mathbb{R}^p : (\eta^* + 1)\text{pen}(\Theta_S | \theta) \leq (1 - \eta^*)\text{pen}(\Theta_{SC} | \theta) \}$ with high probability, where $S$ is the active set of $\beta_j$'s.

Lemma A.4. If $\arg \max_{\beta \in \mathbb{R}^p} \left\{ -\frac{1}{2} \| W(\epsilon - X\mu)/\eta^* - WX\beta \|^2 + \text{pen}(\beta | \theta) \right\} = 0_p$ and $\max w_i \leq M$, then $\| X^T W^2 (\epsilon - X\mu) \|_\infty \leq M \Delta \eta^*$.

Definition A.2. The minimal restricted eigenvalue is defined as

$$c(\eta^*; \beta) = \inf_{\Theta \in \mathbb{R}^p} \left\{ \frac{\|X\Theta\|}{\|X\| \cdot \|\Theta\|} : \Theta \in C(\eta^*; \beta) \right\}$$

Definition A.3. The compatibility number $\phi(C)$ of vectors in cone $C \subset \mathbb{R}^p$ is defined as

$$\phi(C) = \inf_{\Theta \in \mathbb{R}^p} \left\{ \frac{\|X\Theta\| \cdot \|\Theta\|_0^{1/2}}{\|X\| \cdot \|\Theta\|_1} : \Theta \in C(\eta^*; \beta) \right\}$$

A.5 Proof of Lemma A.1

This Lemma is a direct consequence of Proposition 3 of Zhang and Zhang (2012). Notice that in Zhang and Zhang (2012), they divide likelihood by $n$, but we do not divide it by $n$. Our setting is equivalent to Zhang and Zhang (2012)'s by setting $\text{pen}'(\beta) = \text{pen}(\beta)/n$ and thus, $\rho'(t) = \rho(t)/n$. Notice that $\rho'$ satisfies $\rho'(t) \geq \lambda_1(|t|)/n$. From condition (5), $\lambda/n \geq (1 + \xi_0)\frac{\eta}{n} \sqrt{n}(1 + \sqrt{2\log(2p/\delta)})$ by setting $\delta = 2/p^{1/4}$. Then from condition (5) in Theorem 4.2 and Proposition 3 of Zhang and Zhang (2012), we know that

$$\mathbb{P}_\beta \left( \arg \max_{\beta \in \mathbb{R}^p} \left\{ -\frac{1}{2} \| \epsilon/\tilde{\eta} - X\beta \|^2 + \text{pen}(\beta | \theta) \right\} = 0_p \right) \leq 2 - e^{1/p^{1/4} - n(1-1/\sqrt{2})^2}$$

and thus

$$\lim_{n \to \infty} \mathbb{P}_\beta \left( \arg \max_{\beta \in \mathbb{R}^p} \left\{ -\frac{1}{2} \| \epsilon/\tilde{\eta} - X\beta \|^2 + \text{pen}(\beta | \theta) \right\} = 0_p \right) = 1$$
A.6 Proof of Lemma A.2

On condition that $\|\epsilon\|_\infty \lesssim \sqrt{\log n}$, we have

$$
\begin{align*}
\text{Var}_{\mu,w}(\epsilon - X\mu)^T W^2 X\beta | \epsilon) &= \text{Var}_{\mu,w} \left( \sum_i (\epsilon_i - x_i^T \mu) w_i x_i^T \beta | \epsilon \right) \\
&= \mathbb{E}_{\mu,w} \left[ \sum_i (w_i - 1) \epsilon_i x_i^T \beta - \sum_i w_i x_i^T \mu x_i^T \beta \right]^2 \\
&= \mathbb{E}_w \left[ \sum_i (w_i - 1) \epsilon_i x_i^T \beta \right]^2 + \mathbb{E}_{\mu,w} \left[ \sum_i w_i x_i^T \mu x_i^T \beta \right]^2 \\
&= \sum_i \text{Var}(w_i)(\epsilon_i x_i^T \beta)^2 + \sum_{i \neq j} \text{Cov}(w_i, w_j) \epsilon_i x_i^T \beta \epsilon_j x_j^T \beta + \sum_{i,j} x_i^T \beta x_j^T \beta x_i^T x_j \\
&\leq \frac{C_1}{\log n} \sum_i (\epsilon_i x_i^T \beta)^2 + \frac{C_2}{n \log n} \sum_{i \neq j} \frac{1}{2} [(\epsilon_i x_i^T \beta)^2 + (\epsilon_j x_j^T \beta)^2] + \mathbb{E}_{w_i, w_j} \frac{2n}{\lambda_0^2} \sum_i (\epsilon_i x_i^T \beta)^2 \\
&\leq \frac{C_1}{\log n} \|\epsilon\|_\infty^2 \|X\beta\|^2 + C_3 \frac{2n^2}{\lambda_0^2} \sum_i (\epsilon_i x_i^T \beta)^2 \\
&\leq \tilde{C}_3 \|X\beta\|^2
\end{align*}
$$

where (a) uses assumption (2) in Theorem 4.2, the fact $ab \leq \frac{1}{2} (a^2 + b^2)$ and $\mu \sim \text{Spike}$. Inequality (b) uses the property $x_i^T x_j \leq \|x_i\| \times \|x_j\| = n$. Inequality (c) follows from the fact that $(\sum_i x_i^T \beta)^2 \leq n \sum_i (x_i^T \beta)^2$. Inequality (d) follows from the fact that $\lambda_0 \approx p^\gamma$ where $\gamma \geq 1$ and $\|\epsilon\|_\infty \lesssim \sqrt{\log n}$.

Thus, from Markov Inequality, on condition that $\|\epsilon\|_\infty \lesssim \sqrt{\log n},$ we have

$$
\mathbb{P}_{\mu,w}(\|(\epsilon - X\mu)^T W^2 X\beta - \epsilon^T X\beta\| > t | X, \epsilon) \leq \frac{\text{Var}_{\mu,w}(\epsilon - X\mu)^T W^2 X\beta}{t^2} \leq \frac{\tilde{C}_3 \|X\beta\|^2}{t^2}
$$

Set $t = C_n \|X\beta\|$ where $C_n \to \infty$, we have

$$
\lim_{n \to \infty} \mathbb{P}_{\mu,w}(\|(\epsilon - X\mu)^T W^2 X\beta - \epsilon^T X\beta\| > C_n \|X\beta\| | X, \epsilon) = 0 \tag{9}
$$

When $\|(\epsilon - X\mu)^T W^2 X\beta - \epsilon^T X\beta\| \leq C_n \|X\beta\|$, we have

$$
(\epsilon - X\mu)^T W^2 X\beta \leq |\epsilon^T X\beta| + C_n \|X\beta\|. \tag{10}
$$

Notice that

$$
\|X\beta\| + \|X\beta\| \text{pen}(\beta | \theta) \leq \|X\beta\| + \|X\beta\| (-\lambda_1 \beta) \leq \|X\| \times |\beta| - \|X\| \times |\beta| = 0
$$
where (e) follows from $\text{pen}(\beta \mid \theta) = -\lambda_1|\beta| + \sum_j \log \frac{p_j^*(0)}{p_j^*(\beta)} \leq -\lambda_1|\beta|$. Plus this into (10), we have

$$(\epsilon - X\mu)^T W^2 X \beta \leq |\epsilon^T X \beta| - C_n \frac{\|X\|}{\lambda_1} \text{pen}(\beta \mid \theta) \quad (11)$$

When $\tilde{\eta}$-NC condition holds, we have

$$-\frac{\tilde{\eta}}{2} \|X\beta\|^2 + \epsilon^T X \beta + \tilde{\eta} \text{pen}(\beta \mid \theta) \leq 0, \forall \beta \quad (12)$$

Thus, if we choose $\eta^* = \max \left\{ \tilde{\eta} + C_n \frac{\|X\|}{\lambda_1}, \frac{\tilde{\eta}}{m} \right\}$, we have $\forall \beta$,

$$-\frac{\eta^*}{2} \|WX\beta\|^2 + (\epsilon - X\mu)^T W^2 X \beta + \eta^* \text{pen}(\beta \mid \theta) \leq -\frac{\eta^*}{2} \|X\beta\|^2 + |\epsilon^T X \beta| + (\eta^* - C_n \frac{\|X\|}{\lambda_1}) \text{pen}(\beta \mid \theta) \quad (13)$$

where (f) follows from the definition of $\eta^*$ and the fact that $\text{pen}(\beta \mid \theta) \leq 0$ for any $\beta$. We set $\zeta = \beta$ if $\epsilon^T X \beta \geq 0$ and $\zeta = -\beta$ if $\epsilon^T X \beta < 0$. The last inequality directly follows from (12).

This implies that under conditions (1)-(5) in Theorem 4.2, whenever $|\epsilon - X\mu)^T W^2 X \beta - \epsilon^T X \beta| \leq C_n \|X\beta\|$ holds, (13) holds. So on condition that $\|\epsilon\|_\infty \lesssim \sqrt{\log n}$ and $\tilde{\eta}$-NC condition holds, we have

$$P_{\mu,w} \left( |(\epsilon - X\mu)^T W^2 X \beta - \epsilon^T X \beta| \leq C_n \|X\beta\| \mid X, \epsilon \right) \leq P_{\mu,w} \left( -\frac{\eta^*}{2} \|WX\beta\|^2 + (\epsilon - X\mu)^T W^2 X \beta + \eta^* \text{pen}(\beta \mid \theta) \leq 0 \mid X, \epsilon \right)$$

Combined with (9), we know: on condition that $\|\epsilon\|_\infty \lesssim \sqrt{\log n}$ and $\tilde{\eta}$-NC condition holds, the following holds

$$\lim_{n \to \infty} P_{\mu,w} \left( -\frac{\eta^*}{2} \|WX\beta\|^2 + (\epsilon - X\mu)^T W^2 X \beta + \eta^* \text{pen}(\beta \mid \theta) \leq 0 \mid X, \epsilon \right) = 1$$

### A.7 Proof of Lemma A.3

Starting from basic inequality $Q(\tilde{\beta}) \geq Q(\beta_0)$, we get

$$\|WX\Theta\|^2 - 2(W\epsilon - WX\mu)^T WX\Theta + 2\text{pen}(\beta_0 \mid \theta) - 2\text{pen}(\tilde{\beta} \mid \theta) \leq 0 \quad (14)$$

From $\arg\max_{\beta \in \mathbb{R}} \left\{ -\frac{1}{2} \|W(e - X\mu)\| \eta^* - WX\beta\|^2 + \text{pen}(\beta \mid \theta) \right\} = 0_p$, we have

$$-2\Theta^T X^T W^2 (\epsilon - X\mu) \geq -\eta^* \|WX\Theta\|^2 + 2\eta^* \text{pen}(\Theta \mid \theta) \quad (15)$$
Plug (15) into (14), using the fact that \( \text{pen}(\cdot \mid \theta) \) is super-additive, we get

\[
(1 - \eta^*)||WX\theta||^2 \leq -2\eta^*\text{pen}(\Theta \mid \theta) - 2\text{pen}(\beta \mid \theta) + 2\text{pen}(\Theta + \beta \mid \theta)
\]

\[
= -2\eta^*\text{pen}(\Theta_S \mid \theta) - 2\eta^*\text{pen}(\Theta_{SC} \mid \theta) + (-2\text{pen}(\beta_S \mid \theta) + 2\text{pen}(\Theta_S + \beta \mid \theta)) + 2\text{pen}(\Theta_{SC} \mid \theta)
\]

\[
\leq -2\eta^*\text{pen}(\Theta_S \mid \theta) - 2\eta^*\text{pen}(\Theta_{SC} \mid \theta) - 2\text{pen}(\Theta_S \mid \Theta) + 2\text{pen}(\Theta_{SC} \mid \theta)
\]

\[
= -2(\eta^* + 1)\text{pen}(\Theta_S \mid \theta) - 2(\eta^* - 1)\text{pen}(\Theta_{SC} \mid \theta)
\]

Since \((1 - \eta^*)||WX\theta||^2 \geq 0\), we get the desired conclusion.

**A.8 Proof of Lemma A.4**

The proof follows from proof of Lemma 1 in Zhang and Zhang (2012). For any \( t \) and \( j \), since \( Q(\hat{\beta}) \geq Q(\hat{\beta} + t1_j) \) where \( 1_j \) is the vector where the \( j \)-th element is 1 and all the other elements are 0, using the fact that \( \text{pen}(\cdot) \) is super-additive and \( \text{pen}(\beta) = \sum_j \rho(\beta_j) \), we have

\[
t(WX_j)^T(W(Y - X\mu) - WX\hat{\beta}) \leq \frac{t^2}{2} ||WX_j||^2 + \text{pen}(\hat{\beta} - \text{pen}(\hat{\beta} + t1_j)) 
\]

\[
\leq \frac{t^2M}{2} ||X_j||^2 + \rho(t) = \frac{t^2Mn}{2} + M\rho(t)
\]

Thus, for any \( t \),

\[
||X_j^TW^2(Y - X\mu - X\hat{\beta})|| \leq M\left(\frac{1}{2}nt + \rho(t)/t\right)
\]

From the definition of \( \Delta \) in Equation (5), we have

\[
||X_j^TW^2(Y - X\mu - X\hat{\beta})||_\infty \leq M\Delta
\]

When \( Y = X\mu + \frac{\epsilon - X\mu}{\eta^*} \), from \( \arg\max_{\beta \in \mathbb{R}^p} \{-\frac{1}{2}||W(\epsilon - X\mu)/\eta^* - WX\beta||^2 + \text{pen}(\beta \mid \theta)\} = 0_\rho \), we know \( \beta = 0 \) and thus

\[
||X_j^TW^2(\epsilon - X\mu)||_\infty \leq \eta^*M\Delta
\]

**A.9 Proof of Theorem 4.2**

The proof follows from proof of Theorem 7 in Ročková and George (2018). We denote \( c^w_+ = 0.5\left(1 + \sqrt{1 - \frac{4||WX||^2}{(\lambda_0 - \lambda_1)^2}}\right) \) and \( \delta c^w_+ = \frac{1}{\lambda_0 - \lambda_1} \log(\frac{1-\theta}{\lambda_1 c^w_+ 1-c^w_+}) \), which is the inflection point.

Using the fact that \( p^*_\theta(\hat{\beta}_j) > c^w_+ \) when \( \hat{\beta}_j \neq 0 \) and the basic inequality \( 0 \geq Q(\beta_0) - Q(\hat{\beta}) \),
we get

\[
0 \geq \|WX\Theta\|^2 - 2(W\epsilon - WX\mu)^TWX\Theta + 2\log \frac{\pi(\beta_0 | \theta)}{\pi(\beta | \theta)}
\]

\[
\geq \|WX\Theta\|^2 - 2(W\epsilon - WX\mu)^TWX\Theta + 2\left[ -\lambda_1|\beta_0 - \tilde{\beta}| + \sum_{j=1}^{p} \log \frac{p_{\theta}^* (\beta_j)}{p_{\theta}^* (\beta_0)} + \sum_{j=1}^{p} \log \frac{p_{\theta}^* (0)}{p_{\theta}^* (\beta_0j)} \right]
\]

\[
\geq \|WX\Theta\|^2 - 2(W\epsilon - WX\mu)^TWX\Theta + 2\left[ -\lambda_1|\beta_0 - \tilde{\beta}| + \tilde{q}b^w + (\tilde{q} - q) \log \frac{1}{p_{\theta}^* (0)} \right]
\]

\[
\geq \|WX\Theta\|^2 - 2\|(W\epsilon - WX\mu)^TWX\|_{\infty} \times \|\Theta\|_1 - 2\lambda_1\|\beta_0 - \tilde{\beta}\|_1 + 2\tilde{q}b^w + 2(\tilde{q} - q) \log \frac{1}{p_{\theta}^* (0)}
\]

(16)

where \(0 > b^w = \log c^w > \log 0.5\). From Lemma A.4, we know that: if (i) \(\arg\max_{\beta \in R^p} \{-\frac{1}{2} \|W(\epsilon - X\mu)\|_{\eta^*}^2 - \text{pen}(\beta | \theta)\} = 0_p\); (ii) \(\max w_i \leq M\), then we have:

\[
\|X^TW^2(\epsilon - W\mu)\|_{\infty} \leq M\eta^*\Delta
\]

So, from definition of \(c(\eta^*; \theta)\) and Equation (16), we have: if (i), (ii), plus (iii) \(\min w_i \geq m\) and (iv) \(\|\epsilon\|_{\infty} \leq \sqrt{\log n}\) holds, then the following holds:

\[
0 \geq \|WX\Theta\|^2 - 2(\eta^*\Delta + \lambda_1)\|\Theta\| + 2\tilde{q}b^w + 2(\tilde{q} - q) \log \frac{1}{p_{\theta}^* (0)}
\]

\[
\geq mc^2\|\Theta\|^2\|X\|^2 - 2(\eta^*\Delta + \lambda_1)\|\Theta\| + 2\tilde{q}b^w + 2(\tilde{q} - q) \log \frac{1}{p_{\theta}^* (0)}
\]

\[
\geq mc^2\|\Theta\|^2\|X\|^2 - 2(\eta^*\Delta + \lambda_1)\|\Theta\|_2\|\Theta\|_0^{1/2} + 2\tilde{q}b^w + 2(\tilde{q} - q) \log \frac{1}{p_{\theta}^* (0)}
\]

This is equivalent to: if (i), (ii), (iii), (iv) holds, we have

\[
\left(\sqrt{mc\|\Theta\| \times \|X\|} - \frac{M\eta^*\Delta + \lambda_1}{\sqrt{mc\|X\|}} \|\Theta\|_0^{1/2}\right)^2 - \frac{(M\eta^*\Delta + \lambda_1)^2}{mc^2\|X\|^2} \|\Theta\|_0 + 2\tilde{q}b^w + 2(\tilde{q} - q) \log \frac{1}{p_{\theta}^* (0)} \leq 0
\]

So when (i), (ii), (iii), (iv) holds, we have

\[
(\tilde{q} - q) \log \frac{1}{p_{\theta}^* (0)} + \tilde{q}b^w \leq \frac{(M\eta^*\Delta + \lambda_1)^2}{2mc^2\|X\|^2} \|\Theta\|_0 \leq \frac{(M\eta^*\Delta + \lambda_1)^2}{2mc^2n} (\tilde{q} + q)
\]

Thus, when (i)-(iv) holds, we have

\[
\tilde{q} \leq q \frac{A + B}{B + b^w - A} = q(1 + \frac{2A - b^w}{B + b^w - A}) \leq q(1 + \frac{2r}{1 - r})
\]

where \(A = \frac{(M\eta^*\Delta + \lambda_1)^2}{2mc^2n}, \ B = \log \frac{1}{p_{\theta}^* (0)}, \ b^w = \log c^w \in (\log 0.5, 0), \ r = \frac{A}{B}\). For simplicity assume that \(\frac{1-\theta}{\theta} = C_1p^n, \lambda_0 = C_2p^n\) with \(C_1C_2 > 4\). Then \(B = \log(1 + \frac{1-\theta}{\theta} \lambda_0) > 12\).
log(C_1C_2/4) + (\eta + \gamma - 1) \log p$, so $\lambda_1 < 4\sqrt{n \log p} < 4\sqrt{nB/(\eta + \gamma - 1)}$ and $r = \frac{4}{\eta} < \left( \frac{M \eta^* \Delta}{\sqrt{2mBc}} + \frac{\lambda_1}{\sqrt{2mBc}} \right)^2 < \left( \frac{M \eta^*}{\sqrt{Bc}} \sqrt{\frac{\eta + \gamma - 1}{c^2/(\eta + \gamma - 1)}} \right)^2 = D$.

Notice that

$$\lim_{n \to \infty} \mathbb{P}_{\beta, \mu, \nu} \text{(condition (iv) holds)} = \lim_{n \to \infty} \mathbb{P}_\beta \left( \|\epsilon\|_\infty \leq \sqrt{C_1 \log n} \right) = 1$$

From Lemma A.1 and Lemma A.2, we have

$$\lim_{n \to \infty} \mathbb{P}_{\beta, \mu, \nu} \left( \arg \max_{\beta \in \mathbb{R}^p} \left\{ -\frac{1}{2} \|W(\epsilon - \mu')/\eta - WX\beta\|^2 + \text{pen}(\beta | \theta) \right\} = 0_p | X \right) = 1$$

So $\lim_{n \to \infty} \mathbb{P}_{\beta, \mu, \nu} \text{(condition (i) holds)} = 1$. Also $\lim_{n \to \infty} \mathbb{P}_{\beta, \mu, \nu} \text{(condition (ii) holds)} = 1$ and $\lim_{n \to \infty} \mathbb{P}_{\beta, \mu, \nu} \text{(condition (iii) holds)} = 1$.

So from union bound, $\lim_{n \to \infty} \mathbb{P}_{\beta, \mu, \nu} \text{(condition (i), (ii), (iii) and (iv) all holds)} = 1$.

And since

$$\lim_{n \to \infty} \mathbb{E}_{\beta, \mu, \nu} \left( \hat{q} \leq q(1 + \frac{2r}{1-r}) | Y^{(n)} \right) = \lim_{n \to \infty} \mathbb{P}_{\beta, \mu, \nu} \left( \hat{q} \leq q(1 + \frac{2r}{1-r}) \right) \geq \lim_{n \to \infty} \mathbb{P}_{\beta, \mu, \nu} \text{(condition (i), (ii), (iii) and (iv) all holds)} = 1$$

We have

$$\lim_{n \to \infty} \mathbb{E}_{\beta, \mu, \nu} \left( \hat{q} \leq q(1 + \frac{2r}{1-r}) | Y^{(n)} \right) = 1$$

### A.10  Proof of Theorem 4.3

The proof follows from proof of Theorem 8 in Ročková and George (2018). First we prove the high probability bound for $\|\Theta\|$. Since $\log \frac{\pi(\beta_i | \theta)}{\pi(\beta_0 | \theta)} \geq -\lambda_1 |\Theta| + q \log p^0(0)$, from $Q(\hat{\beta}) \geq Q(\beta_0)$ and Lemma A.4, we know if (i) $\arg \max_{\beta \in \mathbb{R}^p} \left\{ -\frac{1}{2} \|W(\epsilon - \mu')/\eta - WX\beta\|^2 + \text{pen}(\beta | \theta) \right\} = 0_p$; (ii) $\max w_i \leq M$, then

$$0 \geq \|WX\Theta\|^2 - 2(W\epsilon - WX\mu)^T WX\Theta + 2\log \frac{\pi(\beta_0 | \theta)}{\pi(\beta_i | \theta)} \geq \|WX\Theta\|^2 - 2(M \eta^* \Delta + \lambda_1) |\Theta| + 2q \log p^0(0) \quad (17)$$

From Theorem 4.2, $\|\Theta\|_0 \leq (1 + K)q$, and using $4uv \leq u^2 + 4v^2$, we have

$$2(M \eta^* \Delta + \lambda_1) |\Theta| \leq 3(M \eta^* \Delta + \lambda_1) \frac{\|X\Theta\| \sqrt{(K + 1)q}}{\|X\|_\infty |\phi|} - (M \eta^* \Delta + \lambda_1) |\Theta|$$

$$\leq \frac{m\|X\Theta\|^2}{2} + \frac{5(K + 1)q(M \eta^* \Delta + \lambda_1)^2}{m\|X\|^2 \phi^2} - (M \eta^* \Delta + \lambda_1) |\Theta|$$

$$\leq \frac{m\|X\Theta\|^2}{2} + \frac{5(K + 1)q(M \eta^* \Delta + \lambda_1)^2}{m\|X\|^2 \phi^2} - (M \eta^* \Delta + \lambda_1) |\Theta|$$

$$\leq \frac{m\|X\Theta\|^2}{2} + \frac{5(K + 1)q(M \eta^* \Delta + \lambda_1)^2}{m\|X\|^2 \phi^2} - (M \eta^* \Delta + \lambda_1) |\Theta|$$

$$\leq \frac{m\|X\Theta\|^2}{2} + \frac{5(K + 1)q(M \eta^* \Delta + \lambda_1)^2}{m\|X\|^2 \phi^2} - (M \eta^* \Delta + \lambda_1) |\Theta|$$
Plug into (17), we know (i), (ii), plus (iii) min $w_i \geq m$ and (iv) $\|e\|_\infty \lesssim \sqrt{\log n}$ implies the following:

$$0 \geq \|WX\Theta\|^2 - \frac{m\|X\Theta\|^2}{2} - \frac{5(K + 1)q(M\eta^*\Delta + \lambda_1)^2}{m\|X\|^2\phi^2} + (M\eta^*\Delta + \lambda_1)|\Theta| + 2q \log(p^*_n)$$

$$\geq \frac{m}{2}\|X\Theta\|^2 - \frac{5(K + 1)q(M\eta^*\Delta + \lambda_1)^2}{m\|X\|^2\phi^2} + (M\eta^*\Delta + \lambda_1)|\Theta| + 2q \log(p^*_n(0))$$

Thus, whenever (i)-(iv) holds,

$$\frac{m}{2}\|X\Theta\|^2 + (M\eta^*\Delta + \lambda_1)|\Theta| \leq \frac{5(K + 1)q(C_3M\eta^*\sqrt{n\log p})^2}{m\|X\|^2\phi^2} + 2qC_4 \log p$$

$$< \frac{C_5^2M^2(\eta^*)^2}{m\phi^2}q(1 + K) \log p$$

Thus, whenever (i)-(iv) holds,

$$\|X\Theta\| \leq \frac{C_5\eta^*}{\sqrt{m\phi}} \sqrt{q(1 + K) \log p}$$

It follows from definition of $c$ that

$$\lim_{n \to \infty} \mathbb{P}_{w,\mu,\beta_0}(\|\Theta\| \leq \frac{C_5\eta^*}{\sqrt{m\phi}} \sqrt{q(1 + K) \log p}) = \lim_{n \to \infty} \mathbb{P} (\text{condition (i)-(iv) holds}) = 1$$

Notice that the difference between $\tilde{\beta}$ and $\bar{\beta}$ only depends on $\mu$ and satisfies

$$\mathbb{P}_\mu\left(\|\tilde{\beta} - \bar{\beta}\|^2 > u \mid Y^{(n)}\right) \leq \frac{\mathbb{E}\sum_{j=1}^p (\tilde{\beta}_j - \bar{\beta}_j)^2}{u} = \frac{1}{u} \sum_{j=1}^p \left(\frac{1}{\lambda_j^2} + \frac{2}{\lambda_0^2}\right) = \frac{13p}{u \lambda_0^2}$$

Set $u = \frac{C_5^2(\eta^*)^2M^2}{m\phi^2}q(1 + K)\log \frac{p}{n}$, then $\frac{13p}{u \lambda_0^2} \to 0$ when $n, p \to \infty$, thus, from the tower law and triangle inequality,

$$\mathbb{E}_{\beta_0}\mathbb{P}_{w,\mu}\left(\|\tilde{\beta} - \beta_0\| > \frac{C_5\eta^* M}{\sqrt{m\phi}} \sqrt{q(1 + K) \log p} \|Y^{(n)}\|\right)$$

$$\leq \mathbb{E}_{\beta_0}\mathbb{P}_{w,\mu}\left(\|\tilde{\beta} - \bar{\beta}\|^2 > \frac{C_5\eta^* M}{2\sqrt{m\phi}} \sqrt{q(1 + K) \log p} \|Y^{(n)}\| \right) + \mathbb{E}_{\beta_0}\mathbb{P}_{w,\mu}\left(\|\Theta\| > \frac{C_5\eta^* M}{2\sqrt{m\phi}} \sqrt{q(1 + K) \log p} \|Y^{(n)}\| \right)$$

$$= \mathbb{E}_{\beta_0,\mu}\mathbb{P}_w\left(\|\tilde{\beta} - \beta_0\|^2 > \frac{C_5\eta^* M}{2\sqrt{m\phi}} \sqrt{q(1 + K) \log p} \|Y^{(n)}\| \right) + \mathbb{E}_{\beta_0,\mu}\mathbb{P}_w\left(\|\Theta\| > \frac{C_5\eta^* M}{2\sqrt{m\phi}} \sqrt{q(1 + K) \log p} \|Y^{(n)}\| \right)$$

$$\leq \frac{13p}{u \lambda_0^2} + \mathbb{P}_{\beta_0,\mu}\left(\|\Theta\| > \frac{C_5\eta^* M}{2\sqrt{m\phi}} \sqrt{q(1 + K) \log p} \|Y^{(n)}\| \right)$$
where the right hand side
\[\frac{3p}{u_0^2} + \mathbb{P}_{\beta_0, u, \nu} \left( ||\Theta|| > \frac{C_5 \eta M}{2\sqrt{m} \rho c} \sqrt{q(1 + K) \log p \over n} | Y^{(n)} \right) \to 0\]
Thus, we have
\[\lim_{n \to \infty} \mathbb{E}_{\beta} \mathbb{P}_{w, \mu} \left( ||\beta - \beta_0|| > C_5 \eta \sqrt{m} \sqrt{\rho c} \sqrt{q(1 + K) \log p \over n} | Y^{(n)} \right) = 0\]

### A.11 Proof of Corollary 4.1

We notice that if \( w_i \sim \Gamma(\alpha, 1) \), we have \( \mathbb{E}w_i = 1, \text{Var}(w_i) = \frac{1}{\alpha} \lesssim \frac{1}{\log n}, \text{Cov}(w_i, w_j) = 0 \). So we only need to prove the two high probability bounds on order statistics for \( w_i \). Let \( \alpha = 2(\eta + \gamma) \log p \), from union bound,

\[\mathbb{P}(\min w_i < \frac{1}{e}) \leq n \mathbb{P}(w_i < \frac{1}{e}) = n \mathbb{P}_{\nu \sim \Gamma(\alpha, 1)}(v < \alpha/e) = ne^{-\alpha/e} \sum_{i=0}^{\infty} \left( \frac{\alpha}{e} \right)^i \frac{1}{i!}\]
\[\leq ne^{-\alpha/e} \sum_{i=0}^{\infty} \frac{(\alpha/e)^i}{\sqrt{2\pi}i^{1/2}e^{-i}} = n e^{-\alpha/e} \sum_{i=0}^{\infty} \frac{(\alpha/e)^i}{\sqrt{2\pi}i^{1/2}e^{-i}} \leq n e^{-\alpha/e} \sum_{i=0}^{\infty} \frac{(\alpha/e)^i}{\sqrt{2\pi}i^{1/2}e^{-i}} \leq C_0 \sqrt{2\pi} \left( \alpha + 1 \right)^{1/2} \left( \alpha + 1 \right) \leq C_0 \sqrt{2\pi} \left( \alpha + 1 \right)^{1/2} \left( \alpha + 1 \right) \leq C_1 (\log p) n p^{-2(\eta + \gamma)}\]

So \( \lim_{n \to \infty} \mathbb{P}(\min w_i < \frac{1}{e}) = 0 \). From the proof of 4.1 and union bound,

\[\mathbb{P}(\max w_i > t) = n e^{-\alpha/(\log(t) + 1) - t + 1} \leq C n (p)^{-\frac{\eta + \gamma}{e}}\]

Set \( t = \frac{2}{3}(\eta + \gamma) > \frac{4}{3} \), then \( \lim_{n \to \infty} \mathbb{P}(\max w_i > t) = 0 \).

If \( w_i \sim n \text{Dir}(\alpha, \cdots, \alpha) \), we have \( \mathbb{E}w_i = 1, \text{Var}(w_i) = n^2 \left[ \frac{1}{n(\alpha + 1)} \right] \lesssim \frac{1}{\log n}, \text{Cov}(w_i, w_j) = -n^2 \left[ \frac{1}{n(\alpha + 1)} \right] \lesssim \frac{1}{n \log n} \). And using the fact that \( nw_i \overset{d}{\to} \frac{1}{\alpha} \Gamma(\alpha, 1) \), we can prove

\[\mathbb{P}(\min w_i < \frac{1}{e}) \leq C_1 (\log p) n p^{-2(\eta + \gamma)}\]

and

\[\mathbb{P}(\max w_i > t) \leq C n (p)^{-\frac{\eta + \gamma}{e}}\]

Thus, condition (1)-(4) hold for both \( w \sim n \text{Dir}(\alpha, \cdots, \alpha) \) and \( w_i \sim \frac{1}{\alpha} \Gamma(\alpha, 1) \) where \( \alpha \gtrsim \sigma^2 \log p \).
A.12 Derivation of observations in section 3.1

In this section we prove some statements in section 3.1.

A.12.1 Notation

Define

\[ c_0^{(-)} = (1 - \theta)\lambda_0 e^{-y_i \lambda_0 + \lambda_0^2/2n}, \quad c_0^{(+)} = (1 - \theta)\lambda_0 e^{y_i \lambda_0 + \lambda_0^2/2n}. \]

Next, define

\[ \phi_0^{(-)}(x) = \phi(x; y_i - \lambda_0/n, 1/n), \quad \phi_0^{(+)}(x) = \phi(x; y_i + \lambda_0/n, 1/n), \]

where \( \phi(x; \mu, \sigma^2) \) is the Gaussian density with mean \( \mu \) and variance \( \sigma^2 \). The quantities \( c_1^{(-)}, c_1^{(+)} \) are defined in a similar way as in 3.1. \( \Delta_j \) is defined as in (5). Define \( c_+ = 0.5 \left(1 + \frac{n}{\sqrt{1 - 4/(\lambda_0 - \lambda_1)^2}}\right) \) and \( \delta_{c+} = 1/(\lambda_0 - \lambda_1) \log \left[\frac{1-\theta}{\lambda_1} \frac{c+}{1-c+}\right] \). Also \( p^*(x) \) and \( \lambda^*(x) \) are as defined in (6).

Consider Gaussian sequence model (8) If multiplying both sides of (8) by \( \sqrt{n} \), we get the linear model with noise variance equal to 1 and \( ||X_j|| = \sqrt{n} \). So here we have \( \Delta^L < \Delta = \inf_{t>0} \left[nt/2 - \rho(t|/t)\right] \equiv \Delta < \Delta^U \) where \( \Delta^L = \sqrt{2n \log 1/p^*(0) - d + \lambda_1} \) and \( \Delta^U = \sqrt{2n \log 1/p^*(0) + \lambda_1}, \theta = -[x(\delta_{c+}) - \lambda_1]^2 - 2n \log p^*(\delta_{c+}) \). Throughout this section we assume that the parameter \( \lambda_0, \lambda_1 \) satisfies: \( (1 - \theta)/\theta \approx n^a, \lambda_0 > n^d \) where \( a,d \geq 2 \) and \( 1/\sqrt{n} \leq \lambda_1 \leq c_0, c_0 \) is a constant.

A.12.2 Active Coordinates

Proposition 1. For the true posterior defined in (9), for active coordinates, conditioning on the event that \( |y_i| > |\beta_1|/2 > 0 \), we have \( w_0 \to 0, w_1 \to 1 \).

Proof. Without loss of generality, we assume \( y_i > \beta_1/2 > 0 \).

\[
\begin{align*}
w_1 &= \pi(\gamma_i = 1 | y_i) = \frac{\pi(y_i | \gamma_i = 1)\pi(\gamma_i = 1)}{\pi(y_i | \gamma_i = 1)\pi(\gamma_i = 1) + \pi(y_i | \gamma_i = 0)\pi(\gamma_i = 0)} \\
&= \frac{\int_{\beta_1} \pi(y_i | \beta_i)\pi(\beta_i | \gamma_i = 1)\pi(\gamma_i = 1) d\beta_i}{\int_{\beta_1} \pi(y_i | \beta_i)\pi(\beta_i | \gamma_i = 1)\pi(\gamma_i = 1) d\beta_i + \int_{\beta_1} \pi(y_i | \beta_i)\pi(\beta_i | \gamma_i = 0)\pi(\gamma_i = 0) d\beta_i} \\
&= \frac{\int_0^\infty c_1(-)\phi_1(-)(\beta_i)d\beta_i + \int_0^\infty c_1(+)(\beta_i)d\beta_i}{\int_0^\infty c_1(-)\phi_1(-)(\beta_i)d\beta_i + \int_0^\infty c_1(+)(\beta_i)d\beta_i + \int_0^\infty c_0(-)\phi_0(-)(\beta_i)d\beta_i + \int_0^\infty c_0(+)(\beta_i)d\beta_i} \\
&= \frac{1}{\theta \lambda_1} \int_0^\infty c_1(-)\phi_1(-)(\beta_i)d\beta_i = e^{-y_i \lambda_1 + \lambda_1^2/2n} \left(1 - \Phi \left(-\sqrt{n}(y_i - \lambda_1/n)\right)\right) \to e^{-y_i \lambda_1}
\end{align*}
\]
For the second term in denominator, from Mills ratio, we have:

\[
\frac{1}{\theta \lambda_1} \int_0^\infty c_0(-) \phi_0(-)(\beta_i) d\beta_i = \frac{1}{\theta \lambda_1} c_0(-) \Phi(-\sqrt{n}(y_i - \frac{\lambda_0}{n}))
\]

\[
\leq C \frac{(1 - \theta) \lambda_0}{\theta \lambda_1} e^{-2\sqrt{n}(y_i - \frac{\lambda_0}{n})} = C \frac{(1 - \theta) \lambda_0}{\sqrt{n}(y_i - \frac{\lambda_0}{n}) \theta \lambda_1} \frac{1}{2\pi} e^{-\frac{2}{n} y_i^2}
\]

where \(\frac{(1 - \theta) \lambda_0}{\sqrt{n}(y_i - \frac{\lambda_0}{n}) \theta \lambda_1} \frac{1}{2\pi} e^{-\frac{2}{n} y_i^2} \to 0\). Thus,

\[
\frac{1}{\theta \lambda_1} \int_0^\infty c_0(-) \phi_0(-)(\beta_i) d\beta_i \to 0
\]  

(20)

For the third term in denominator:

\[
\frac{1}{\theta \lambda_1} \int_{-\infty}^0 c_1(+) \phi_1(+) (\beta_i) d\beta_i = e^{y_i + \frac{\lambda_1}{2n}} \Phi \left(-\sqrt{n} \left(y_i - \frac{\lambda_1}{n}\right) \right) \to 0
\]  

(21)

For the fourth term in denominator:

\[
\frac{1}{\theta \lambda_1} \int_{-\infty}^0 c_0(+) \phi_0(+) (\beta_i) d\beta_i = \frac{1}{\theta \lambda_1} \int_{-\infty}^0 (1 - \theta) \lambda_0 \sqrt{n} \frac{2\pi}{2\pi} e^{-\frac{n}{n} (\beta_i - y_i)^2 + \beta_i \lambda_0} d\beta_i
\]

\[
\leq \frac{1}{\theta \lambda_1} \int_{-\infty}^0 (1 - \theta) \lambda_0 \sqrt{n} \frac{2\pi}{2\pi} e^{-\frac{n}{n} y_i^2 + \beta_i \lambda_0} d\beta_i = \frac{1}{\theta \lambda_1} (1 - \theta) \sqrt{n} \frac{2\pi}{2\pi} e^{-\frac{n}{n} y_i^2}
\]

where \(\frac{1}{\theta \lambda_1} (1 - \theta) \sqrt{n} \frac{2\pi}{2\pi} e^{-\frac{n}{n} y_i^2} \to 0\). Thus,

\[
\frac{1}{\theta \lambda_1} \int_{-\infty}^0 c_0(+) \phi_0(+) (\beta_i) d\beta_i \to 0
\]  

(22)

From (19), (20), (21) and (22), we know that \(w_0 \to \frac{e^{-\frac{n}{n} y_i}}{e^{-\frac{n}{n} y_i}} = 1\). Thus, \(w_1 = 1 - w_0 \to 0\). \(\square\)

**Proposition 2.** For the true posterior (9), \(\pi(\sqrt{n}(\beta_i - y_i) | y_i, \gamma_i = 1) \to \phi(\sqrt{n}(\beta_i - y_i); 0, 1)\).

**Proof.** Set \(u_i = \sqrt{n}(\beta_i - y_i)\), we have

\[
\pi(u_i | y_i, \gamma_i = 1) = \frac{1}{\sqrt{n}} \frac{\mathbb{I}(u_i \geq \sqrt{n} y_i) c_1(-) \phi_1(-)(u_i/\sqrt{n} + y_i) + \mathbb{I}(u_i < -\sqrt{n} y_i) c_1(+) \phi_1(+) (u_i/\sqrt{n} + y_i)}{\int_0^\infty c_1(-) \phi_1(-)(\beta_i) d\beta_i + \int_{-\infty}^0 c_1(+) \phi_1(+) (\beta_i) d\beta_i}
\]  

(23)

Notice that both

\[
\frac{1}{\sqrt{n}} \phi_1(-)(u_i/\sqrt{n} + y_i) \to \phi(u_i; 0, 1) \quad \text{and} \quad \frac{1}{\sqrt{n}} \phi_1(+) (u_i/\sqrt{n} + y_i) \to \phi(u_i; 0, 1)
\]

For any \(u_i\), only one of \(\mathbb{I}(u_i < -\sqrt{n} y_i)\) and \(\mathbb{I}(u_i \geq \sqrt{n} y_i)\) holds, and the denominator in (23) does not depend on \(u_i\), so \(\pi(u_i | y_i, \gamma_i = 1) \to \phi(u_i; 0, 1)\). \(\square\)
Proposition 3. For fixed WBB estimate $\hat{\beta}_i$, conditioning on the event that $|y_i| > \frac{|\beta_0|}{2} > 0$, when $\hat{\beta}_i \neq 0$, $n(\hat{\beta}_i - y_i) \mid y_i \to -\frac{1}{w_i}\lambda_1$.

Proof. The objective function for each fixed WBB sample is

$$\hat{\beta}_i = \arg \max_{\beta_i \in \mathbb{R}} \left\{ -\frac{w_i n}{2}(y_i - \beta_i)^2 + \log \pi(\beta_i \mid \theta) \right\} = \arg \max_{\beta_i \in \mathbb{R}} \left\{ -\frac{1}{2}(\sqrt{w_i n}y_i - \sqrt{w_i n}\beta_i)^2 + \log \pi(\beta_i \mid \theta) \right\}$$

Thus $\hat{\beta}_i$ satisfies (12). Without loss of generality, we assume $y_i > \frac{|\beta_0|}{2} > 0$. When $\hat{\beta}_i \neq 0$,

$$(n(\hat{\beta}_i - y_i) \mid Y_i = y_i, w_i) = \frac{1}{w_i}\lambda_1 - \frac{1}{w_i}(\lambda_0 - \lambda_1)(1 - p^*(\hat{\beta}_i))$$

$$= \frac{1}{w_i}\lambda_1 - \frac{1}{w_i}\frac{(1-\theta)\lambda_0}{\theta\lambda_1}e^{-|\hat{\beta}_i|/(\lambda_0 - \lambda_1)} \to \frac{1}{w_i}\lambda_1$$

Proposition 4. For fixed WBB estimate $\hat{\beta}_i$, conditioning on the event that $|y_i| > \frac{|\beta_0|}{2} > 0$, $\mathbb{P}_{w_i}(\hat{\beta}_i = 0 \mid y_i) \to 0$.

Proof. From Markov Inequality,

$$\mathbb{P}_{w_i}(\hat{\beta}_i = 0 \mid y_i) \leq \mathbb{P}((\hat{\beta}_i - y_i)^2 > y_i^2) \leq \frac{\mathbb{E}(\hat{\beta}_i - y_i)^2}{y_i^2}$$

$$= \frac{1}{y_i} \left\{ \mathbb{E}\left[y_i^2 1(|\sqrt{w_i}y_i| \leq \frac{\Delta}{n})\right] + \mathbb{E}\left[\left(\frac{1}{w_i} \lambda^*(\hat{\beta}_i)\right)^2 1(|\sqrt{w_i}y_i| > \frac{\Delta}{n})\right]\right\}$$

$$\leq \mathbb{P}(w_i \leq \frac{\Delta^2}{n^2 y_i^2}) + \frac{1}{y_i} \mathbb{E}\left[\left(\frac{c_+(\lambda_1 - \lambda_0) + \lambda_0}{\sqrt{n}w_i}\right)^2 1(|\sqrt{w_i}y_i| > \frac{\Delta}{n}y_i)\right]$$

$$\leq \mathbb{P}(\frac{1}{w_i} \geq \frac{n^2 y_i^2}{\Delta^2}) + \frac{4}{y_i^2 n} \mathbb{E}\left(\sqrt{w_i} > \frac{\Delta}{n y_i}\right)$$

$$\leq \frac{\mathbb{E} \frac{1}{w_i}}{n^2 y_i^2 \Delta^2} + \frac{4}{y_i^2 n} \frac{1}{w_i}$$

where the third row of (24) follows from Ročková (2018): $|\sqrt{nw_i}\hat{\beta}_i| > \delta_{c_+}, p^*(\sqrt{nw_i}\hat{\beta}_i) > c_+, \lambda^*(\sqrt{nw_i}\hat{\beta}_i) < c_+(\lambda_1 - \lambda_0) + \lambda_0$ when $|\sqrt{w_i}y_i| > \frac{\Delta}{n}$ and thus

$$\lambda^*(\hat{\beta}_i) = \sqrt{nw_i}\lambda^*(\sqrt{nw_i}\hat{\beta}_i) < \sqrt{nw_i}(c_+(\lambda_1 - \lambda_0) + \lambda_0)$$

and the fourth row of (24) follows directly from the definition of $c_+, \Delta$ and the conditions on $\lambda_0, \lambda_1$ as described at the beginning of this section. The right hand size of (24) satisfies

$$\frac{\mathbb{E} \frac{1}{w_i}}{n^2 y_i^2 \Delta^2} + \frac{4}{y_i^2 n} \frac{1}{w_i} \to 0$$

Thus $\mathbb{P}_{w_i}(\hat{\beta}_i = 0 \mid y_i) \to 0$ holds. \(\square\)
**A.12.3 Inactive coordinates**

**Proposition 5.** For \( w_0 \) and \( w_1 \) in (9), when conditioning on \( y_i \simeq \frac{1}{\sqrt{n}} \), we have \( w_1 \to 0 \) and \( w_0 \to 1 \).

**Proof.** The expression of \( w_1 \) is in (18). Again we consider the four terms in denominator separately. And we also divide each of them by \( \theta \lambda \). For the first term:

\[
\frac{1}{\theta \lambda_1} \int_0^\infty c_1^{(-)} \phi_1^{(-)}(\beta_i) d\beta_i = e^{-y_i \lambda_1 + \lambda_1^2/2n} \left( 1 - \Phi \left( -\sqrt{n}(y_i - \frac{\lambda_1}{n}) \right) \right) \to 1 - \Phi (-\sqrt{n}y_i) \tag{25}
\]

For the second term in denominator, for any fixed \( \epsilon > 0 \),

\[
\frac{1}{\theta \lambda_1} \int_0^\infty c_0^{(-)} \phi_0^{(-)}(\beta_i) d\beta_i = \frac{1}{\theta \lambda_1} \int_0^\infty (1 - \theta) \lambda_0 \sqrt{\frac{n}{2\pi}} e^{-\frac{n}{2}(\beta_i-y_i)^2 - \beta_i \lambda_0} d\beta_i \\
\geq \frac{1}{\theta \lambda_1} \int_{\epsilon y_i}^\infty (1 - \theta) \lambda_0 \sqrt{\frac{n}{2\pi}} e^{-\frac{n(\epsilon^2-1)}{2} y_i^2 - \beta_i \lambda_0} d\beta_i \\
= \frac{1}{\theta \lambda_1} (1 - \theta) \sqrt{\frac{n}{2\pi}} e^{-\frac{n}{2}(\epsilon^2-1) y_i^2} (1 - e^{-\epsilon y_i \lambda_0})
\]

where \( e^{-\epsilon y_i \lambda_0} \to 0 \). Since the right hand size term

\[
\frac{1}{\theta \lambda_1} (1 - \theta) \sqrt{\frac{n}{2\pi}} e^{-\frac{n(\epsilon^2-1)}{2} y_i^2} (1 - e^{-\epsilon y_i \lambda_0}) \to \infty
\]

we know

\[
\frac{1}{\theta \lambda_1} \int_0^\infty c_0^{(-)} \phi_0^{(-)}(\beta_i) d\beta_i \to \infty \tag{26}
\]

For the third term in denominator,

\[
\frac{1}{\theta \lambda_1} \int_{-\infty}^0 c_1^{(+)} \phi_1^{(+)}(\beta_i) d\beta_i = e^{y_i \lambda_1 + \lambda_1^2/2n} \Phi \left( -\sqrt{n}(y_i + \frac{\lambda_1}{n}) \right) \to \Phi (-\sqrt{n}y_i) \tag{27}
\]

For the fourth term in denominator, for any fixed \( \epsilon > 0 \),

\[
\frac{1}{\theta \lambda_1} \int_{-\infty}^0 c_0^{(+)} \phi_0^{(+)}(\beta_i) d\beta_i = \frac{1}{\theta \lambda_1} \int_{-\infty}^0 (1 - \theta) \lambda_0 \sqrt{\frac{n}{2\pi}} e^{-\frac{n}{2}(\beta_i-y_i)^2 + \beta_i \lambda_0} d\beta_i \\
\geq \frac{1}{\theta \lambda_1} \int_{-\epsilon y_i}^0 (1 - \theta) \lambda_0 \sqrt{\frac{n}{2\pi}} e^{-\frac{n(\epsilon^2+1)}{2} y_i^2 + \beta_i \lambda_0} d\beta_i \\
= \frac{1}{\theta \lambda_1} (1 - \theta) \sqrt{\frac{n}{2\pi}} e^{-\frac{n(\epsilon^2+1)}{2} y_i^2} (1 - e^{-\epsilon y_i \lambda_0})
\]

where \( e^{-\epsilon y_i \lambda_0} \to 0 \). Since the right hand side term

\[
\frac{1}{\theta \lambda_1} (1 - \theta) \sqrt{\frac{n}{2\pi}} e^{-\frac{n(\epsilon^2+1)}{2} y_i^2} (1 - e^{-\epsilon y_i \lambda_0}) \to \infty
\]
we know
\[ \frac{1}{\theta \lambda_1} \int_{-\infty}^{0} e_0^{(+)} \phi_0^{(+)}(\beta_i) d\beta_i \to \infty \]  
(28)

Combining (25), (26), (27) and (28), we know that in (18), denominator \( \to \infty \) and numerator \( \to 1 \). Thus \( w_1 \to 0 \) and \( w_0 = 1 - w_1 \to 1 \).

**Proposition 6.** When \( n \) is sufficiently large, conditioning on \( |y_i| \simeq O(\frac{1}{\sqrt{n}}) \), we have
\[ \pi(\lambda_0 \beta_i | y_i, \gamma_i = 0) \to \frac{1}{2} e^{-|\lambda_0 \beta_i|}. \]

*Proof.* Notice that
\[ \pi(\beta_i | y_i, \gamma_i = 0) \propto \pi(y_i | \beta_i, \gamma_i = 0) \pi(\beta_i | \gamma_i = 0) \propto e^{-\frac{1}{2}(\beta_i - y_i)^2} e^{-|\beta_i| \lambda_0} \]
Thus, let \( \beta_i' = \lambda_0 \beta_i \), since \( \lambda_0 \to \infty \), for any fixed \( \beta_i' \), we have
\[ \pi(\beta_i' | y_i, \gamma_i = 0) \propto e^{-\frac{1}{2}(\beta_i'/\lambda_0 - y_i)^2} e^{-|\beta_i'|} \to e^{-\frac{1}{2}(y_i)^2} e^{-|\beta_i'|} \]
which implies that
\[ \pi(\beta_i' | y_i, \gamma_i = 0) \to e^{-|\beta_i'|} \]
From Scheffé (1947), we know that \( \beta_i' \) converges in total variation to \( e^{-|\beta_i'|} \).

**Proposition 7.** For fixed WBB estimate \( \hat{\beta}_i \), conditioning on \( |y_i| \simeq O(\frac{1}{\sqrt{n}}) \), if \( \exists t > 0 \) s.t. \( Ec^{t w_i} \leq C \), \( \forall i \), we have \( P_{w_i}(\hat{\beta}_i = 0 | y_i) \to 1 \).

*Proof.* The definition of the fixed WBB sample is in \( \hat{\beta}_i \) (12). From Chernoff bound,
\[ P_{w_i} \left( \frac{\sqrt{w_i}}{n |y_i|} | y_i \right) \leq \frac{E e^{t w_i}}{e^{t \Delta^2/(n^2 y_i^2)}} \leq \frac{C}{e^{t \Delta^2/(n^2 y_i^2)}} = \frac{C}{e^{t (2 \log 1/p^* (0)-d/n)/(ny_i^2)}} = \frac{\tilde{C}}{n^2(\eta+\gamma)^{t/(ny_i^2)}} \]
where \( \frac{\tilde{C}}{n^2(\eta+\gamma)^{t/(ny_i^2)}} \to 0 \). Thus, \( P_{w_i}(\hat{\beta}_i = 0 | y_i) \to 1 \).

**Remark A.1.** Random WBB is equivalent to fixed WBB by setting weights to be \( w/w_p \) where \( w_p \) is the weight put on prior term. Thus, using exactly the same arguments as fixed WBB, we can prove: if \( \exists t > 0 \) s.t. \( E_{w_i, w_p} e^{t w_i/w_p} \leq C \), \( \forall i \), we have \( P_{w_i}(\beta_i^{\text{random}} = 0 | y_i) \to 1 \).

**B Details of Connection to NPL in Section 4.2**

We define the loss function \( l(\cdot) \) as
\[ l(x_i, y_i, \beta) = - \frac{1}{2 \sigma^2} (y_i - x_i^T \beta)^2 + \frac{1}{n} \log \left[ \prod_{j=1}^{p} \pi(\hat{\beta}_j | \theta) d\pi(\theta) \right] \]  
(29)
Motivation for the Prior For paired data $(x_i, y_i)$, Fong et al. (2019) uses independent prior which assumes that $y_i$ does not depend on $x_i$:

$$\text{Prior 1: } \tilde{x}_k \sim \hat{F}_{n}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i), \quad \tilde{y}_k | \tilde{x}_k \sim N(0, \sigma^2)$$

However, this choice of $F_n$ might be problematic when the sample size $n$ is small. When $n$ is small, a well-specified prior can help us better estimate $\beta$ but this independent prior shrinks all coefficients towards zero and will result in bias (see Figure 1).

One possible solution is to use $y = \mathbf{x}^T \tilde{\beta} + \epsilon$ where $\tilde{\beta}$ is the MAP of $\beta$ under SSL penalty. This choice of $f_\pi(y|x)$ has some flavor of Empirical Bayes (Martin and Walker (2014)). However, it includes information only about the posterior mode, but ignores all the other information like posterior variance. We can consider adding back that information through some additive noise $\mu$ on $\tilde{\beta}$. We want $\mu$ to be centered at origin and not too far away from origin. One choice that comes to mind is the Spike. So the prior $F_\pi$ becomes

$$\tilde{x}_k \sim \hat{F}_{n}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i), \quad \tilde{y}_k | \tilde{x}_k = \mathbf{x}_k^T (\tilde{\beta} + \mu) + \epsilon$$

where $\mu \sim \text{Spike}$ and $\epsilon \sim N(0, \sigma^2)$. If $\tilde{\beta}$ is close enough to truth, we have $\tilde{x}_k^T \tilde{\beta} \approx \tilde{x}_k^T \beta_0$.

Since $y_i = x_i^T \beta_0 + \epsilon_i$ and $\epsilon_i \overset{d}{=} \epsilon$, we can set $\tilde{y}_k | \tilde{x}_k = y_i + x_i^T \mu$ where $i$ satisfies $\tilde{x}_k = x_i$.

Then the above prior becomes

$$\text{Prior 2: } \tilde{x}_k \sim \hat{F}_{n}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i), \quad \tilde{y}_k | \tilde{x}_k = Y_i + x_i^T \mu \text{ where } i \text{ satisfies } \tilde{x}_k = x_i$$

Derivation for Equation (18) If choosing $m = n$ in Algorithm 3, the NPL posterior of Fong et al. (2019) using Prior 2 becomes

$$\hat{\beta}^* = \arg\max_{\beta \in \mathbb{R}^p} \left\{-\frac{1}{2} \sum_{i=1}^{n} w_i(Y_i - x_i^T \beta)^2 - \frac{1}{2} \sum_{i=1}^{n} \tilde{w}_i(Y_i + x_i^T \mu - x_i^T \beta)^2 + \frac{1}{n} \log \left( \prod_{j=1}^{p} \pi(\beta_j | \theta) d\pi(\theta) \right) \right\}$$

$$= \arg\max_{\beta \in \mathbb{R}^p} \left\{-\frac{1}{2} \sum_{i=1}^{n} (w_i + \tilde{w}_i)(Y_i + \frac{\tilde{w}_i}{w_i + \tilde{w}_i} x_i^T \mu - x_i^T \beta)^2 + \frac{1}{n} \log \left( \prod_{j=1}^{p} \pi(\beta_j | \theta) d\pi(\theta) \right) \right\}$$

$$\approx \arg\max_{\beta \in \mathbb{R}^p} \left\{-\frac{1}{2} \sum_{i=1}^{n} (w_i + \tilde{w}_i)(Y_i + \frac{c}{c + n} x_i^T \mu - x_i^T \beta)^2 + \frac{1}{n} \log \left( \prod_{j=1}^{p} \pi(\beta_j | \theta) d\pi(\theta) \right) \right\}$$

$$\tilde{\beta} = \beta - c/(c+n) \mu \Rightarrow \arg\max_{\beta \in \mathbb{R}^p} \left\{-\frac{1}{2} \sum_{i=1}^{n} w_i^* (Y_i - x_i^T \tilde{\beta})^2 + \log \left( \prod_{j=1}^{p} \pi(\tilde{\beta}_j + \frac{c}{c + n} \mu_j | \theta) d\pi(\theta) \right) \right\} + \frac{c}{c + n} \mu$$

where $w_i^* = n(w_i + \tilde{w}_i)$. Since $\mu$ and $-\mu$ follows the same distribution, define $\mu^* = -\mu$, then

$$\hat{\beta}^D = \arg\max_{\beta \in \mathbb{R}^p} \left\{-\frac{1}{2} \sum_{i=1}^{n} w_i^* (Y_i - x_i^T \tilde{\beta})^2 + \log \left( \prod_{j=1}^{p} \pi(\tilde{\beta}_j - \frac{c}{c + n} \mu_j^* | \theta) d\pi(\theta) \right) \right\} - \frac{c}{c + n} \mu^*$$
Table 1: Running time (in seconds, based on average of 10 independent runs) of each algorithm (per 100 iterations). The number after ‘±’ is the standard deviation of 10 independent runs. BB-SSL are either fitted using a sequence of $\lambda_0$’s: an equal difference series of length $\lceil \lambda_0 \rceil$ starting at $\lambda_1 = 0.05$ and ends at $\lambda_0 = 200$, or a single $\lambda_0 = 200$. SSVS1 is using Woodbury matrix identity to calculate matrix inverse. Signals are $(1, 2, -2, 3)$. Predictors are grouped into blocks of size 10 where each block contains at most one signal.

\[
(w_{1:n}, \tilde{w}_{1:n}) \sim \text{Dir}(1, \cdots, 1, c/n, c/n, \cdots, c/n) \quad \text{and} \quad (w_1^*, w_2^*, \cdots, w_n^*) \sim n\text{Dir}(1 + c/n, \cdots, 1 + c/n).
\]

C Additional Results on Computational Considerations

Table 1 contains comparisons of the running times of different algorithms for varying $p$ and $n$. Below are details for the computational complexity analysis as shown in Table 1 of the main text.

BB-SSL, WBB1, WBB2 BB-SSL uses R package SSLASSO (Rocková and Moran (2017)) to implement the coordinate-descent algorithm in Ročková and George (2018). It iteratively updates $\beta$ until it converges. In each iteration we update first the active coordinates, then the candidate coordinates, and finally the inactive coordinates. The total number of iterations is limited to a pre-defined number. There are two ways to update each coordinate, one way is to keep track of a residual vector and for each coordinate we compute the inner product between the residual vector and $X_j$ – this takes $O(n)$; another way is to pre-compute the Gram matrix and for each coordinate, we calculate the inner product between $X_j^T X_j$ and $\beta$ – this takes $O(p)$. For both ways, we update $\theta$ every $c$ iterations where each update is $O(p)$. So for a single value of $\lambda_0$, SSLASSO is $O\left(\min\left(\maxiter \times p(n + \frac{p}{c_1}), (n + \maxiter) \times p^2\right)\right)$. For a sequence of $\lambda_0$’s, complexity is $O\left(L \times \min\left(\maxiter \times p(n + \frac{p}{c_1}), (n + \maxiter) \times p^2\right)\right)$ where $L$ is the length of $\lambda_0$’s. Usually if the biggest $\lambda_0$ is large enough, we would expect that the larger $\lambda_0$’s can reach...
(a) Active predictors, from top to bottom: $\beta_1, \beta_4, \beta_7, \beta_{10}$

(b) Inactive predictors, from top to bottom: $\beta_2, \beta_5, \beta_8, \beta_{11}$

Figure 1: Comparison of Fong et al. (2019)’s algorithm 2 with different choice of priors and loss function versus BB-SSL and SSVS under simulated dataset. We use $n = 50$, $p = 12$, $\beta = (1.3, 0, 0, 1.3, 0, 0, 1.3, 0, 0, 1.3, 0, 0)^T$, predictors are grouped into 4 blocks with correlation coefficient $\rho = 0.6$, $\alpha = 2$, $m = 50$, $c = 10$, $\lambda_0 = 7$, $\lambda_1 = 0.15$. 
convergence quickly using the estimated $\beta$ from previous $\lambda_0$’s, so usually it takes less than $O\left(L \times \min \left(\maxiter \times p(n + \frac{p}{c_1}), (n + \maxiter) \times p^2\right)\right)$.

Empirically, when $\lambda_0$ is small and does not provide enough shrinkage, SSL takes more iterations to converge. When choosing $\lambda_0$, we suggest plotting the regularization path from SSL and choose a large enough $\lambda_0$ after which further increasing $\lambda_0$ does not affect the chosen model. This is also reasonable from the goal of variable selection. When $\lambda_0$ is extremely small, a single value fitted SSL might have difficulty converging so we have to use a sequence of $\lambda_0$’s to burn-in. It usually takes much longer time than a single value fitted SSL. Note that the computational complexity of other methods - SSVS1, SSVS2 and Skinny Gibb - does not depend on $\lambda_0$.

**SSVS1** The computation complexity for Algorithm 1 is $O(p^3)$ per iteration when $p > n$. Under this setting, we use Woodbury matrix identity to simplify the matrix multiplication $(X^T X + D^{-1})^{-1} = D - D X T (I_n + X D_\tau X^T)^{-1} X D_\tau$, whose complexity is $O(p^2 n)$.

**SSVS2** Using Bhattacharya et al. (2016)’s matrix inversion formula to generate the $p$-dimensional multivariate Gaussian takes $O(n^2 p)$.

**Skinny Gibbs** We modified Skinny Gibbs to sample from the posterior using SSL prior. Theoretically it is of complexity $O(np)$. However, sometimes when $n$ is relatively small, we observe that the running time of Skinny Gibbs is slower than Bhattacharya’s method. This is because Skinny Gibbs involves an $O(n)$ matrix product for each coordinate and the update for each coordinate is implemented via for-loop, whereas in Bhattacharya’s method the $O(n^2 p)$ operation is one matrix product which is very efficiently optimized in R. We will see that as $n$ increases, this problem diminishes and Skinny Gibbs becomes faster than Bhattacharya’s method.

**WLB** Generating each WLB sample involves solving a least square problem whose complexity is $O(p^2 n)$ when $p \leq n$. WLB is not applicable when $p > n$.

### D Additional Experimental Results

#### D.1 Low Dimensional, Uncorrelated Setting

We first investigate the marginal density of $\beta_1^0$, as shown in figure 2. We find that all methods perform well for active $\beta_i$’s. WBB1 and WBB2 are doing poorly for inactive $\beta_i$’s. For the marginal mean of $\gamma_i$’s, as shown in figure 3, all methods perform pretty well. All methods can detect over 95% of models.
Figure 2: Density and credible interval of $\beta_i$'s in low-dimensional, independent case. $n = 50, p = 12,$ $\beta_{active} = (1.3, 1.3, 1.3, 1.3),$ $\lambda_0 = 13, \lambda_1 = 0.05$ and predictors are mutually independent. Each method is (thinned to) 1,000 sample points. SSLASSO is fitted using a single $\lambda_0$. Since WBB1 and WBB2 produce a point mass and do not fit into the $y$-axis, we exclude it in the density plot.

(a) Active predictors, from top to bottom: $\beta_1, \beta_4, \beta_7, \beta_{10}$

(b) Inactive predictors, from top to bottom: $\beta_2, \beta_5, \beta_8, \beta_{11}$
Figure 3: Mean of $\gamma_i, i = 1, 2, \cdots, 12$ in low-dimensional, independent case. $n = 50, p = 12, \beta_{active} = (1, 3, 1, 3, 1, 3)$ and predictors mutually independent. Each method is (thinned to) 1,000 sample points. $\lambda_0 = 13, \lambda_1 = 0.05$. SSLASSO is fitted using a single $\lambda_0$.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Method</th>
<th>Skinny</th>
<th>WBB1</th>
<th>WBB2</th>
<th>BB-SSL</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_i$</td>
<td>KL divergence</td>
<td>active</td>
<td>3.02</td>
<td>3.66</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td>inactive</td>
<td>1.66</td>
<td>3.49</td>
<td>3.49</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>active</td>
<td>5.05</td>
<td>2.15</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>inactive</td>
<td>6.05</td>
<td>2.15</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 2: Evaluation of posterior distribution in $n = 100, p = 1000, \rho = 0.99$ setting.

D.2 High Dimensional, $\rho = 0$ and $\rho = 0.9$

Figure 4 shows the posterior for $\beta_i$’s when $\rho = 0$ and Figure 5 is for $\rho = 0.9$. Skinny Gibbs has some problems with estimating the variance. WBB1 and WBB2 does poorly for inactive coordinates. In general BB-SSL does pretty well.

D.3 High Dimensional, $\rho = 0.99$

Here we have $n = 100, p = 1000, \rho = 0.99$. All the other settings are exactly the same as that for high dimensional designs in Section 5, except that here SSVS1 also falls into local trap for chains as long as 1 000 000. So we manually run two SSVS chains, with one initialized at origin and the other at the true $\beta$ and combine them as the truth posterior. Figure 6 and 7 show that no method performs perfect in this extreme setting. BB-SSL is good at detecting multi-modality but sometimes overshoot (for $\beta_2$ and $\beta_{12}$ it has a false mode at the correlated truth). Skinny Gibbs also falls into local trap as SSVS and has some problems with the variance. WBB1 and WBB2 does not work for inactive coordinates.

D.4 Influence of $\alpha$ On The Posterior

In this section we investigate the influence of $\alpha$ on BB-SSL posterior under (1) high-dimensional, moderately correlated ($\rho = 0.6$) setting, as shown in Figure 8, and (2) Durable Goods Marketing Data Set, as shown in Figure 9.

In both datasets, as we increase $\alpha$, the change in posterior variance reduces for each unit increase in $\alpha$. 

26
(a) Active predictors, from top to bottom: $\beta_1, \beta_{11}, \beta_{21}, \beta_{31}$  
(b) Inactive predictors, from top to bottom: $\beta_2, \beta_{12}, \beta_{22}, \beta_{32}$

Figure 4: Density and credible interval of active $\beta_i$'s in high-dimensional, independent case. $n = 100, p = 1000, \beta_{active} = (1, 2, -2, 3), \lambda_0 = 50, \lambda_1 = 0.05$ and predictors are mutually independent. Each method (is thinned to) 1,000 sample points. SSLASSO is fitted using a single $\lambda_0$. Since WBB1 and WBB2 produce a point mass and do not fit into the y-axis, we exclude it in the density plot.
Figure 5: Density and credible interval of $\beta_i$’s in high-dimensional, highly correlated case. $n = 100$, $p = 1000$, $\beta_{active} = (1, 2, -2, 3)$, $\lambda_0 = 50$, $\lambda_1 = 0.05$, $\rho = 0.9$ and predictors are grouped into 100 blocks. The correlation coefficient within each block is $\rho = 0.9$. Each method is thinned to 5,000 sample points. SSL is fitted with $\lambda_0$ being an equal difference sequence of length 50 starting at 0.05 and ending at 50. Since WBB1 and WBB2 produce a point mass and do not fit into the y-axis, we exclude it in the density plot.
Figure 6: Density and credible interval of $\beta_i$'s in high-dimensional, extremely correlated case. $n = 100, p = 1000, \beta_{\text{active}} = (1, 2, -2, 3), \lambda_0 = 50, \lambda_1 = 0.05$ and predictors are grouped into 100 blocks. The correlation coefficient within each block is $\rho = 0.99$. Each method has 10,000 sample points. SSLASSO is fitted using a sequence of $\lambda_0$'s.
Figure 7: Mean of $\gamma_i, i = 1, 2, \cdots, 1000$ in high-dimensional, correlated case. $n = 100, p = 1000, \beta_{active} = (1, 2, -2, 3), \lambda_0 = 50, \lambda_1 = 0.05$ and predictors are grouped into blocks of size 10 with $\rho = 0.99$. SSL is fitted using a sequence of $\lambda_0$’s.
Figure 8: Comparison of posterior density for active $\beta_i$’s when choosing different $\alpha$ in high-dimensional, moderately correlated setting ($\rho = 0.6$). SSL is fitted with $\lambda_0$ being an equal difference sequence of length 10 starting at 0.05 and ending at 50. We set $\lambda_1 = 0.05$. Since WBB1 and WBB2 produce a point mass and do not fit into the $y$-axis, we exclude it in the density plot.
Figure 9: Trace plots of $\beta_i$, $i = 1, 2, \cdots, 12$ under varying $\alpha$’s in model (20).
References


